MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1303–1314 doi: 10.1111/1755-0998.12527

Development of a universal double-digest RAD sequencing approach for a group of nonmodel, ecologically and economically important insect and fish taxa

M. O. BURFORD REISKIND,*[1] K. COYLE,† H. V. DANIELS,* P. LABADIE,‡ M. H. REISKIND,‡ N. B. ROBERTS,† R. B. ROBERTS,† J. SCHAFF§ and E. L. VARGO¶

*Department of Applied Ecology, North Carolina State University, Campus Box 7617, Raleigh, NC 27695, USA, †Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA, ‡Department of Entomology, North Carolina State University, Raleigh, NC 27695, USA, ¶Department of Entomology, Texas A&M University, College Station, TX 77843, USA, §Genomic Sciences Laboratory, North Carolina State University, Raleigh, NC 27695, USA

Abstract

The generation of genome-scale data is critical for a wide range of questions in basic biology using model organisms, but also in questions of applied biology in nonmodel organisms (agriculture, natural resources, conservation and public health biology). Using a genome-scale approach on a diverse group of nonmodel organisms and with the goal of lowering costs of the method, we modified a multiplexed, high-throughput genomic scan technique utilizing two restriction enzymes. We analysed several pairs of restriction enzymes and completed double-digestion RAD sequencing libraries for nine different species and five genera of insects and fish. We found one particular enzyme pair produced consistently higher number of sequence-able fragments across all nine species. Building libraries off this enzyme pair, we found a range of usable SNPs between 4000 and 37 000 SNPS per species and we found a greater number of usable SNPs using reference genomes than de novo pipelines in STACKS. We also found fewer reads in the Read 2 fragments from the paired-end Illumina Hiseq run. Overall, the results of this study provide empirical evidence of the utility of this method for producing consistent data for diverse nonmodel species and suggest specific considerations for sequencing analysis strategies.

Keywords: double-digest RAD sequencing, high-throughput sequencing, reduced representation sequencing, restriction enzyme digests

Received 28 July 2015; revision received 3 March 2016; accepted 9 March 2016

Introduction

The promise of genomics is being realized primarily through work in model systems, including humans, mice and fruit flies (Venter *et al.* 2001; Gregory *et al.* 2002; Clark *et al.* 2007). The generation of genome-scale data in these organisms is critical for a wide range of questions in basic and applied biology. However, nonmodel species, including many organisms important in applied biology (agriculture, natural resources, conservation, public health biology), have only just begun to be scrutinized using these powerful techniques (Stinchcombe & Hoekstra 2008; Ekblom & Galindo 2011; Hand *et al.* 2015). With the advances of high-throughput sequencing and accessibility of reference genome sequences available for related species, we are now able to collect and utilize genome-scale sequence data of these species critical to human well-being. These data can be used to address a variety of questions on patterns of intraspecific phenotype variation, population genomics, mechanisms of evolutionary change, phylogenetic relationships and even the identification of the genetic basis of traits (e.g. Stinchcombe & Hoekstra 2008; Hohenlohe *et al.* 2010; Ellegren *et al.* 2012; McCormack *et al.* 2013; Wagner *et al.* 2013; Gamble *et al.* 2015; McCluskey & Postelthwait 2015). When used in conjunction with an annotated reference genome (either of the species of interest or related species), long-standing issues with the evolutionary relationship among species or the specific genes involved in any of these questions above can be rapidly determined (Roberts *et al.* 2011; Jones *et al.* 2013; Albertson *et al.* 2014; Brawand *et al.* 2014; Franchini *et al.* 2014; McCluskey & Postlethwait 2015). Moreover, these approaches

Correspondence: Martha O. Burford Reiskind, Fax: 919-515-5327; E-mail: mbreiski@ncsu.edu

[1]*Note*: After the first author, all authors were arranged in alphabetical order.

can pinpoint parts of an organism's genome responsible for species or population divergence and the response to natural or artificial selection.

In addition to being critical for understanding the mechanisms or genes responsible for population or species divergence, comparative genomics has important practical implications (Allendorf *et al.* 2010; Narum *et al.* 2013; Williams *et al.* 2014). By comparing parts of the genome showcasing these differences (polymorphisms) to annotated, published genomes of the same or related species, candidate functional genes may be identified (e.g. Roberts *et al.* 2011; Albertson *et al.* 2014). In a conservation context, we may better predict how environmental change affects growth, survival and fecundity of nonmodel and ecologically important species by identifying genes associated with these traits in different populations (e.g. Reitzel *et al.* 2013; Brawand *et al.* 2014; Reichwald *et al.* 2015; Yoshizawa *et al.* 2015). Additionally, genomic techniques can provide a comprehensive measure of genetic diversity, effective population size, and genetic background for management of threatened populations (e.g. Hess *et al.* 2015; Lew *et al.* 2015), providing information that will aid predictions of future population viability.

The limitations of genomic sequencing of nonmodel organisms stem from the high-cost, low-throughput of previous technologies, lack of draft genomes to align sequences and arguably the greatest limitations of computing and bioinformatic expertise. However, we are in a new era in applying genomics to nonmodel organisms with the advent of inexpensive approaches for assembling genomic data into reference genomes and methodologies that provide consistent and high coverage (breadth and depth) of individual genomes. These approaches, while relatively economical, still require the development of technological expertise.

Building on previous approaches using restriction enzymes to sample the genome (Miller *et al.* 2007; Baird *et al.* 2008; Hohenlohe *et al.* 2010), several research groups have utilized a double-digest restriction enzyme DNA sequencing (ddRAD sequencing) approach to sample the genome with a multiplexed, high-throughput method (Elshire *et al.* 2011; Peterson *et al.* 2012; Poland *et al.* 2012). This new method provides high quantity and consistency of genomic data while reducing the number of steps of the previous method. In addition, ligating barcodes onto both ends of the fragments in this new method and size selection using computerized, gel electrophoresis provides greater uniformity of fragment size, number and position across individuals and species. This uniformity is an important consideration for studies that analyse the genomes of multiple populations and species. In particular, if there is greater quantity of fragments from one genomic region (i.e. deeper stacks or read depth) within and among individuals, polymorphisms in this area are more likely due to true polymorphism as opposed to sequencing error. While it is true that in some studies in population genomics low coverage or stack depth is not as much a concern (i.e. Buerkle & Gompert 2013), we would argue for broad applications such as identifying outlier loci, greater coverage and stack depth will increase confidence.

To build this capacity and to lower costs, we modified a multiplexed, high-throughput genomic scan technique utilizing two restriction enzymes. We empirically tested multiple enzyme pairs on several ecologically or commercially important, nonmodel animal taxa with closely, distantly and no reference genomes. Our goal was to develop a broadly applicable approach to diverse taxa while maximizing quality and consistency of read locations within a genome and between closely related species' genomes. To further reduce costs of creating barcodes for each species individually, our goal was to use the same library building materials (same enzyme pairs and barcodes) while achieving high quality and quantity of sequence-able fragments across a diverse group of species. In addition, we wanted to generate critical exploratory data to compare our approach in several nonmodel systems: the commercially important fishery and aquaculture fish *Paralichthys lethostigma* (the southern flounder), the commercially important rockfish (*Sebastes*) of the groundfish fishery in the western USA, the cichlid fish and behavioural model *Astatotilapia burtoni*, the mosquito vector for several important human diseases *Aedes aegypti* (the yellow fever mosquito), the eastern subterranean termites (*Reticulitermes flavipes*) in its native eastern range and resurgent bedbug (*Cimex lectularius*) populations in the United States and Europe.

Methods

Enzyme experiment

DNA extraction, quantification and dilutions. For the southern flounder (*Paralichthys lethostigma*) and rockfish (*Sebastes* spp.), we extracted genomic DNA from fin clips. For the yellow fever mosquito (*Aedes aegypti*), we used the whole body of the mosquito. For the extractions of these three species, we used a Qiagen DNA Extraction Kit (Qiagen Inc., Valencia, CA, USA) and quantified template DNA using a fluorometer (Qubit 2.0; Invitrogen, Carlsbad, CA, USA) following both manufacturer's protocols, with the exception that we eluted in H_2O and not elution buffer to allow for subsequent concentration of DNA if needed. For the bedbugs (*Cimex lectularius*) and termites (*Reticulitermes flavipes*), we extracted DNA using the Qiagen DNA Extraction Kit used above but preceded by liquid nitrogen crushing and resuspending the

samples in molecular grade water. These samples were also quantified using the fluorometer as above. For *Astatotilapia burtoni*, we extracted genomic DNA from fin clips using Thermo GeneJet Genomic DNA Purification Kit (Thermo Fisher Scientific, Altham, MA, USA) and quantified DNA via Quant-iT PicoGreen dsDNA assay (Invitrogen, Carlsbad, CA, USA). We did not use *A. burtoni* in the enzyme pair experiment, but it was used for the full library building.

Enzyme digest. To analyse the number of sequence-able fragments using the double-digest method, we compared estimates of the number of fragments generated by parallel digestion of genomic DNA with different enzyme pairs and single enzymes (Table 1). This allowed us to determine which pairs provided both the greatest number and the most consistent results within and across the different taxonomic groups. We conducted the digestion with at least three replicates of each species (the replicate was three different individuals that were digested by all the enzyme pairs and single enzymes). For the *Ae. aegypti*, we pooled three different colony individuals for each of the three replicates to obtain enough DNA to run all the digests. We conducted nine digests per replicate individual or pool using four enzyme pairs: *SphI-EcoRI, EcoRI-MspI, SphI-MluCI and NlaIII-EcoRI* and five single enzyme digestions: *SphI, EcoRI, MspI, MluCI and NlaIII* (New England BioLabs, Inc., Ipswich, MA, USA). We conducted digests in a 30-μL reaction volume, with 26 μL of DNA (total of 200–300 ng DNA), 3 μL of CutSmart buffer, 0.5 μL of each enzyme diluted to 2 units per μL. We incubated all digests for 3 h at 37 °C and held the digests at 4 °C. To clean the digests, we used MagBead purification (1.5× the volume of the digest AMPure XP beads, Beckman Coulter, Inc., Brea, CA, USA), washed the beads one time with 70% EtOH, eluted the sample in 35 μL of H_20 and recovered 30 μL of the suspended sample for further analysis. For all digestions of individuals and individual enzyme pairs that appeared to fail, we reran those digestions two additional times to confirm that they did indeed fail.

Calculation of fragments generated by enzyme digests. Following a modified protocol from Peterson *et al.* (2012), we conducted an assay of the double and single enzyme digest on the Agilent 2100 Bioanalyzer using the high sensitivity DNA chip, using default conditions to confirm the size distribution of digested fragments. Using the Bioanalyzer quantitative output, we calculated the number of fragments for different base pair size groupings including 200 ± 20; 300 ± 30; 400 ± 40; and 500 ± 50 and calculated the proportion of fragments from the total. For the single-digest samples, we analysed the entire distribution (50–1000 bps), which we used to calculate the approximate number of fragments with enzyme 1 and enzyme 2 ends that would be amplified by the ddRAD sequencing method (Supporting information). To estimate the number of sequence-able fragments, we calculated the proportion of the total (Supporting information). To make the calculations, we included an estimate of the genome size (Table 2: *Sebastes*, Ojima & Yamamoto 1990; *P. lethostigma*, Ojima & Yamamoto 1990; *A. burtoni*, Brawand *et al.* 2014; *Ae. aegypti*, Nene *et al.* 2007; bedbugs, Benoit *et al.* 2016; termites, Koshikawa *et al.* 2008).

Double-digest RAD sequencing library building

After conducting the above experiment, we built ddRAD sequencing libraries using the enzyme pairs *SphI* and *MluCI*. We built six libraries, one of 26, three of 48, one of 93 and one of 96 individuals (Table 2). The libraries were as follows: *A. burtoni* family library ($N = 26$); *Sebastes* library ($N = 48$); *Ae. aegypti* library ($N = 93$); *P. lethostigma* family library ($N = 96$); *C. lectularius* library ($N = 48$); *R. flavipes* library ($N = 48$). We ran each library on a single HiSeq lane with the exception of *C. lectularius* and *R. flavipes* libraries, which we ran on the same lane.

Barcodes and indices. We designed 48 unique variable-length barcodes for the *SphI* cut side (P1 adapters) and a fixed-length y-adapter for the *MluCI* cut site (P2 adapters; see Supporting information for details). In our study, we chose 5-bp to 10-bp barcodes for a total of 48 unique barcodes per individual library. We also built two indices into the reverse PCR primer allowing us to multiplex two groups of 48 into one sequencing lane (for a maximum number 96 individuals per lane). We added a custom primer (5′CGGAAGAGCGGTTCAGCAGGAA TGCCGAGACCG 3′) to our sequence-ready library to pick up our custom indices within the y-adapter, allowing us to successfully de-index libraries of 48 individuals from the Illumina platform (see Supporting information).

Digestion to sequencing. The detailed protocol that includes the digestion, ligation and sequencing methods is found in the Supporting information. For consistency, we used 200 ng of template DNA per individual (there were no pooled samples) and found this amount to produce similar results to 300 ng of template DNA. With only 200 ng of template DNA and to prevent loss of sample using magnetic beads for DNA capture, we used columns for purification and pooling of the libraries (QIAquick PCR purification columns; Qiagen Inc., Valencia, CA, USA). In the case of bedbugs, we had 24 samples with pre-extracted DNA from a collaborator. Based on Qubit quantification of our 24 bedbug samples, we had

Table 1 Serial digest analysis of the number of sequence-able fragments based on empirical enzyme digest data and size selection. All digests included 300 ng DNA/reactions with the exception of *Aedes aegypti* replicate 1 and 2, which were 200 ng DNA/reaction. A replicate is a single individual replicate with the exception of *Aedes aegypti* (see text), such that *Sebastes melanops* 1 is the same individual's DNA for all enzyme pairs. For individual replicates or enzyme pairs that failed (e.g. *EcoRI-MspI Aedes* Replicate 3), the numbers in this table represent three digestion attempts and therefore are not due to operator error

Enzyme Pair	Species	Replicate	200 bp	300 bp	400 bp	500 bp
SphI-MluCI	*Sebastes mystinus* Type 1	1	70 605	75 312	71 139	56 711
		2	75 943	91 132	69 039	61 121
		3	74 947	80 211	68 134	54 288
	Sebastes mystinus Type 2	1	74 773	70 022	60 271	54 053
		2	68 174	72 719	61 977	49 283
		3	151 495	130 287	98 373	69 813
	Sebastes melanops	1	70 446	75 394	63 881	50 925
		2	72 885	78 004	66 092	52 688
	Sebastes serranoides	1	62 882	58 690	57 021	45 457
		2	67 743	72 501	61 430	54 522
		3	64 583	68 889	58 564	51 874
	Sebastes entomelas	1	80 072	74 733	72 680	64 314
		2	77 763	72 821	62 680	49 968
		3	89 701	90 001	67 784	54 037
	Aedes aegypti	1	174 681	104 037	58 699	39 014
		2	150 653	80 754	45 539	30 283
		3	172 480	114 795	77 436	31 260
		4	190 547	143 632	83 997	47 685
	Paralichthys lethostigma	1	49 055	59 160	49 423	39 440
		2	41 542	48 627	41 638	41 625
		3	68 194	68 194	51 402	40 998
		4	53 241	56 791	48 037	38 334
		5	61 381	61 586	51 279	41 003
		6	46 159	49 401	41 647	37 075
	Cimex lectularius	1	110 006	65 298	37 041	24 520
		2	178 140	92 524	49 986	23 872
		3	240 218	120 715	60 664	36 167
	Reticulitermes flavipes	1	175 604	105 895	53 216	28 239
		2	163 827	98 792	49 773	26 345
		3	133 634	79 584	44 993	23 875
NlaIII-EcoRI	*Sebastes mystinus* Type 1	1	76 053	76 307	57 616	45 815
		2	70 870	75 847	71 405	56 923
		3	68 953	73 796	62 527	49 846
	Sebastes mystinus Type 2	1	72 945	83 021	69 474	55 384
		2	76 990	92 697	77 572	61 840
		3	118 911	125 410	85 795	61 523
	Sebastes melanops	1	56 457	60 120	50 940	45 121
		2	56 574	60 244	51 045	40 692
	Sebastes serranoides	1	59 600	71 759	60 050	47 871
		2	60 705	64 644	60 858	48 418
		3	57 181	68 847	57 613	45 929
	Sebastes entomelas	1	58 581	62 382	52 856	46 818
		2	103 520	103 866	86 918	69 290
		3	58 805	70 802	59 250	47 233
	Aedes aegypti	1	106 421	113 895	96 503	85 307
		2	90 563	105 995	102 624	90 706
		3	71 906	111 854	120 446	105 886
		4	113 258	242 695	128 378	113 485
	Paralichthys lethostigma	1	47 979	51 349	48 464	46 245
		2	5886	5512	5930	6133
		3	87 066	65 738	54 966	43 796
		4	33 585	33 641	33 670	30 196

Table 1 (Continued)

Enzyme Pair	Species	Replicate	200 bp	300 bp	400 bp	500 bp
		5	41 678	50 601	42 310	40 211
		6	46 096	49 334	51 218	44 430
	Cimex lectularius	1	114 817	32 470	40 279	25 457
		2	241 428	77 034	68 532	51 745
		3	353 634	80 525	72 717	57 592
	Reticulitermes flavipes	1	101 444	108 569	91 990	73 333
		2	100 684	94 286	91 300	72 784
		3	0	12 367	9369	7405
EcoRI-MspI	*Sebastes mystinus* Type 1	1	22 757	30 444	34 307	27 363
		2	22 173	29 565	33 427	35 549
		3	22 277	14 851	22 389	17 786
	Sebastes mystinus Type 2	1	0	0	11 635	18 514
		2	0	15 318	22 977	18 345
		3	68 746	91 968	80 809	64 420
	Sebastes melanops	1	0	14 247	21 424	25 594
		2	20 469	13 623	20 469	16 294
	Sebastes serranoides	1	21 516	14 344	21 570	25 819
		2	21 426	14 332	21 534	25 711
		3	21 443	28 590	32 326	25 783
	Sebastes entomelas	1	21 256	28 294	31 884	25 431
		2	22 682	30 243	34 194	27 218
		3	21 477	43 098	32 541	25 824
	Aedes aegypti	1	60 622	60 825	76 543	60 622
		2	30 905	41 214	62 298	61 717
		3	0	0	79 527	39 164
		4	33 600	45 472	51 413	54 134
	Paralichthys lethostigma	1	0	4269	6408	5129
		2	0	10 991	8150	41 542
		3	34 741	34 857	35 004	27 849
		4	0	0	0	6736
		5	0	11 252	8418	13 475
		6	0	11 365	17 104	20 525
	Cimex lectularius	1	0	0	0	0
		2	20 524	13 728	20 575	24 580
		3	59 431	39 687	29 641	35 410
	Reticulitermes flavipes	1	0	18 864	28 367	45 183
		2	0	18 690	28 530	33 979
		3	0	18 740	28 275	33 710
SphI-EcoRI	*Sebastes mystinus* Type 1	1	24	16 100	11 994	9556
		2	23 737	15 878	23 917	19 028
		3	23 851	16 035	11 986	9 474
	Sebastes mystinus Type 2	1	0	0	0	0
		2	0	0	0	0
		3	23 799	63 147	47 839	28 645
	Aedes aegypti	1	0	0	0	13 510
		2	0	0	0	13 377
	Paralichthys lethostigma	1	0	0	0	0
		2	0	0	0	0
		3	17 730	23 640	17 864	21 361
		4	0	0	0	7098

suboptimal quantity of genomic material (likely due poor yield extractions), far below the recommended quantity of 200 ng of DNA for building libraries. Therefore, we conducted a whole-genome amplification (WGA) using Qiagen Repli-g Single Cell Kit (Qiagen Inc., Valencia, CA, USA) on these 24 samples. We conducted PCR amplification of four polymorphic microsatellites before and after WGA. We did this comparison to confirm whether we got the same alleles before and after WGA and to ascertain the fidelity of the whole-genome

Table 2 Details of the genomic libraries of the double-digest RAD sequencing using the enzyme pairs *Sph1* and *MluCI*. Note that there are two rows associated with *Cimex lectularius*, one for the individuals that have no whole-genome amplification and one for the individuals that do have whole-genome amplification (WGA). We calculated GC content using raw reads in GENIOUS 9.03 for all the read 1 reads

Library	*n*	Populations/Family	Genome size	GC content	% Cover	Average stack depth
*Sebastes**	48	16 *S. mystinus* Type 1, 16 *S. mystinus* Type 2, 15 *S. entomelas*, 1 *S. melanops*)	0.98 GB	45.0%	de novo = 2% reference = 6%	de novo = 8.4× reference = 7.4×
Paralichthys lethostigma	96	2 parents, 94 progeny	0.71 GB	43.9%	de novo = 0.3%	de novo = 12.0×
Astatotilapia burtoni	26	2 parents, 24 progeny	0.92 GB	44.6%	de novo = 1.1% reference = 5%	de novo = 10.1× reference = 9.6×
Aedes aegypti	93	Tucson, AZ (21 wild type & 24 selected), Key West, FL (24 wild type & 24 selected)	1.4 GB	42.9%	de novo = 0.2% reference = 3%	de novo = 8.4× reference = 2.5×
Cimex lectularius	48	24 Human host USA	0.87 GB	41.8%	de novo = 0.4% reference = 0.6%	de novo = 13.7× reference = 16.7×
		12 Human host Europe, 12 Bat host Europe with WGA	0.87 GB	41.8%	de novo = 0.7% reference = 1.2%	de novo = 17.4× reference = 23.2×
Reticulitermes flavipes	48	eastern subterranean termite USA	1.20 GB	44.9%	de novo = 0.4%	de novo = 12.9×

**Sebastes mystinus* Type 1 and Type 2 identified by Burford & Bernardi (2008).

replication. We obtained positive results through PCR and achieved yields of 200 ng of DNA to continue the library preparation for the bedbug samples.

To avoid the formation of among DNA fragments or adapter dimer due to either too little or too much adapter in the ligation step, we followed Elshire *et al.*'s (2011) adapter titration protocol and conducted a species-specific adapter titration experiment for three randomly chosen individual from each species. We used six adapter concentrations and ran all titrations on the Bioanalyzer after PCR amplification. We reviewed the output of the Bioanalyzer to first confirm any adapter dimer and then chose the concentration of adapters that resulted in the best library. We qualitatively determined the best library by the size curve that was the most consistent and cleanest (i.e. no breaks in the curve and no stutter below or above the curve). Once we confirmed the adapter concentration that yielded the best library for each species, we used that adapter concentration in the ligation step.

After ligation of individual samples, we pooled all individuals with the 48 unique P1 barcodes and purified this library for size selection on the BluePippin (Sage Science, Beverly, MA, USA). We targeted a 100-base pair read length; therefore, we size-selected the library fragments between 300 and 500 bps to account for barcode and y-adapters resulting in a peak of 400–base pair fragments with a target fragment size of over 250 base pairs depending on the size of the barcode. We wanted to avoid sequencing through the target fragment, and therefore, the range of 300–500 assured a gap between the read 1 and read 2 sequences as we were only sequencing 90 base pairs of the target sequence on either side. After confirming size selection on the Bioanalyzer and quantifying the remaining library, we ran eight PCRs of the same purified and size-selected libraries. The PCR protocol included a 25-μL reaction with 10 μL of size-selected library (at approximately 10 pg/μL) and used the following conditions: 72 °C (5 min); 98 °C (30 s); 18x (98 °C [10 s], 65 °C [30 s], 72 °C [30 s], 72 °C [5 min]); 4 °C (∞). We used the QIAquick PCR purification protocol and post-cleanup combined all eight PCRs into one sample (see Supporting information). For sequencing lanes of 96 individuals, which combined two libraries of 48 individuals with separate indices, we analysed post-PCR samples on the Bioanalyzer to equalize the two combined libraries for final sequencing. We submitted a final concentration of 15 nmol/L in 30 μL for sequencing.

For all libraries, we conducted paired-end sequencing of 100-bp fragments on the Illumina HiSeq 2000 at University of North Carolina Chapel Hill. Specifications for sequencing were 15 nM in 30 μL based on UNC Chapel Hill's protocol at that time.

ddRAD sequencing library analysis

Initial quality control. The Illumina platform de-multiplexed the two indices into separate FASTA files. We ran FASTQC (Babraham Bioinformatics; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and checked the quality of the reads prior to processing the barcodes. We then ran PROCESS_RADTAGS to filter and demultiplex

our variable-length barcodes and rename files with individual identifiers in STACKS (Catchen *et al.* 2011). Due to low phred scores at the beginning of read 2, we trimmed the first 4 base pairs off both reads, resulting in 90-base pair reads, to make all read lengths identical in length as required by the STACKS platform.

SNP detection. For SNP detection, we ran the de novo pipeline (DENOVO_MAP.PL) in STACKS for all species using the default settings with the exception of parameter *n*, the catalog error term (default settings in STACKS include: $M = 2$ (mismatches between loci within an individual), $m = 3$ (minimum stack depth), $n = 0$ (mismatches allowed between loci when combining them in a catalog), $N = M + 2$ (mismatches between secondary and primary stacks); Catchen *et al.* 2011). To optimize *n*, we conducted a serial set of runs with *n* set at the default of 0 to $n = 3$ with the *Sebastes* and *Ae. aegypti* data sets. We compared the number of post-analysis reads from the different *n*s with what we found using a reference genome, as a conservative estimate. For both data sets, we found that the default *n* of 0 was not as consistent with what we found using a reference genome when we filtered the reads as *n* of 1, 2, or 3. Therefore, we ran all de novo pipelines with $n = 2$, which allowed combination of stacks with two errors into catalogs and avoided the issue of allowing too many or too few errors per catalog (Table 3).

Four species had reference genomes (*Sebastes*, *A. burtoni*, *Ae. aegypti* and *C. lectularius*), and therefore, we used the reference pipeline (REF_MAP.PL) in STACKS after aligning sequences in BWA-short (Li & Durbin 2009) and Bowtie 2 (Langmead & Salzberg 2012; default settings: maximum and minimum mismatches allowed were MX = 6, MN = 2). Based on a comparable number of aligned reads to number of reads in the original FASTQ files, we opted for Bowtie 2. We ran all reference runs with $m = 3$ (minimum stack depth) in STACKS. For the *Sebastes* library, we used the *S. aleutianus* (rough eye rockfish; https://genomeevolution.org/CoGe/OrganismView.p|?oid=39830) draft genome; for *A. burtoni*, we used the published reference genome (Broad Institute Tilapia Genome project; Brawand *et al.* 2014); for *Ae. aegypti*, we used the liverpool3 reference genome found on Vectorbase (www.vectorbase.org, downloaded August 31, 2014); and for *C. lectularius*, we used the reference genome published by the Baylor College of Medicine and The Human Genome Sequencing Center i5K Pilot project.

Results

Enzyme experiment

The parallel digest worked well for all enzyme pairs with the exception of *SphI-EcoRI*, which produced inconsistent results and low estimates of sequence-able fragments (Table 1). Therefore, we abandoned this enzyme pair for *C. lectularius* and *R. flavipes*. *EcoRI-MspI* worked well for some species (*Sebastes)*, but it lacked consistency across the diverse taxonomic groups and the estimated number of sequences was lower than other enzyme pairs. We found two enzyme pairs *SphI-MluCI* and *NlaIII-EcoRI* that had similar quantity and consistency of estimated sequences. We conducted paired *t*-test between the *SphI-MluCI* and *NlaIII-EcoRI* enzyme pairs from the serial digest, and there were no significant differences in number of fragments, except at the largest base pair size (500 bp, $P < 0.05$; Table 1). However, there was considerably more variation in the number of fragments across all taxa with *NlaIII-EcoRI* than *SphI-MluCI* (SD at 200 bp 42 342 vs. 27,620; at 300 bp 27 608 vs. 13 399,

Table 3 Total number of loci over all individuals of the *Sebastes* and *Ae. aegypti* libraries using *SphI-MluCI* and the de novo pipeline run in STACKS and using different settings of the parameter *n* ($n = 0$–3), the number of mismatches allowed between loci when building the catalog. Also includes the results using the reference pipeline runs as a comparison. The comparison includes the number of catalog loci for default settings ($M = 2$, $m = 3$), and for two filtering settings: reads with 1–3 SNPs or reads with 1–5 SNPs with matching from at least three individuals to the number in the library (48 or 96 individuals depending)

Species/Pipeline	*n*	Unfiltered loci	Filtered loci by SNPs 1–3	Filtered loci by SNPs 1–5
Sebastes de novo	0	4 529 343	97 562	103 291
Sebastes de novo	1	3 751 429	237 143	253 772
Sebastes de novo	2	3 634 534	242 718	269 369
Sebastes de novo	3	3 427 296	236 078	288 212
Sebastes reference	Default	1 130 980	173 398	207 192
Aedes aegypti de novo	0	881 587	37 230	45 160
Aedes aegypti de novo	1	800 959	65 019	73 373
Aedes aegypti de novo	2	745 787	69 909	79 758
Aedes aegypti de novo	3	702 952	67 805	80 863
Aedes aegypti reference	Default	502 310	90 433	122 223

respectively), with a few replicates yielding very few fragments for *NlaIII-EcoRI* (Table 1: *P. lethostigma*). This result suggested that *SphI-MluCI* offered the best combination of consistent performance and high fragment number across taxa. Therefore, for this diverse set of taxonomic organisms, we chose to build the libraries with the *SphI-MluCI* enzyme digest and, appropriately, designed our variable-length barcodes to pick up the cut sites generated by these enzymes.

Double-digest RAD sequencing library building

Initial quality control. After the PROCESS_RADTAGS pipeline in STACKS, we found approximately 200 million reads per library and we had on average 2–4 million reads per individual. The results of the FASTQC showed excellent quality scores of the ddRAD sequencing reads, with mean phred scores of 33 or 31 or above, for read 1 and read 2, respectively. In the raw output from Illumina, we found a consistent number of reads per barcode, indicating there was no bias in particular barcodes across all seven species (data not shown).

Coverage. We found a range of per cent coverage for the different analyses of the de novo and reference pipelines in STACKS. To calculate the per cent coverage (breadth), we used the number of unique stacks output from the de novo and *reference* pipeline runs in STACKS, using the size of the reads in base pairs and the size of the genome (Table 2). The average per cent coverage ranged from 0.2% (*Ae. aegypti*) to 2% (*Sebastes*) and 3% (*Ae. aegypti*) to 6% (*Sebastes*) for de novo and reference runs, respectively (Table 2). Using the bam files generated per individual in BOWTIE2 and the DENOVO_MAP.LOG output from stacks, we found an average stack depth that ranged from 8.4× to 17.4× and 2.5× to 23.2× for de novo and reference runs, respectively (Table 2). Notably the *C. lectularius* from Europe with whole-genome amplification (WGA) had higher per cent cover and depth than the *C. lectularius* from the USA (Table 2).

Comparison of Read 1 and Read 2 sequences of the paired-end analysis. After we ran the de novo pipeline (DENOVO_MAP.PL) in STACKS and separated out read 1 and read 2, we found consistently lower number of unique stacks, polymorphic loci and SNPs in read 2 compared to read 1 (Table 4), with the exception of polymorphic loci and SNPs in read 2 of *A. burtoni*. However, the number of unique stacks was higher in read 1 than read 2 for *A. burtoni*. This suggested that read 2 sequences were more prone to sequencing errors or lower quality and were filtered during the *process-radtag* pipeline in STACKS, resulting in a lower quantity of acceptable sequences. However, those that made it through the filtering process had a high phred score.

SNP detection. In general for those groups in which we could compare the results of the reference and de novo pipelines, we found greater number of unique stacks, polymorphic loci and SNPs from the reference pipeline than de novo pipeline in STACKS, averaged across individuals for both read 1 and read 2 (Table 5). The library with the highest number of polymorphic loci was *Sebastes* reference with 18.2K, and the lowest was *C. lectularius* de novo with only 2.7K (Table 5). The number of SNPs ranged from 3K to 37K SNPs depending on the species, with the reference pipeline showing a higher number of SNPs and in most cases over three times the number of SNPs detected by the de novo pipeline (Table 5). The one exception was *A. burtoni*, with a

Table 4 Average number of unique stacks, polymorphic loci or SNPs found using *SphI-MluCI* and the de novo pipeline runs in STACKS using either read 1 or read 2 ($n = 2$). Each column includes the average across individuals and the parenthetical standard deviation. These include all individuals and all stacks without *post hoc* filtering

Species	Unique stacks	Polymorphic loci	SNPs Found
Sebastes R1	164 866 (±31 998)	9210 (±2465)	12 042 (±3012)
Sebastes R2	127 237 (±36 797)	4857 (±1258)	6445 (±1689)
Aedes aegypti R1	35 114 (±14 448)	3702 (±1365)	6286 (±2407)
Aedes aegypti R2	12 348 (±4389)	1282 (±451)	1984 (±715)
Cimex lectularius R1	42 060 (±19 206)	2338 (±1754)	3555 (±2706)
Cimex lectularius R2	34 405 (±21 827)	1643 (±1039)	2514 (±1642)
Cimex lectularius WGA R1	63 149 (±19 587)	3365 (±1653)	4769 (±2227)
Cimex lectularius WGA R2	90 587 (±32 185)	2034 (±894)	3111 (±1321)
Reticulitermes flavipes R1	56 782 (±24 354)	6163 (±2890)	8379 (±4002)
Reticulitermes flavipes R2	33 022 (±6914)	4088 (±1106)	5339 (±1478)
Paralichthys lethostigma R1	47 978 (±27 868)	4506 (±3124)	5990 (±4105)
Paralichthys lethostigma R2	25 664 (±14 555)	2553 (±1513)	3409 (±2009)
Astatotilapia burtoni R1	67 067 (±27 625)	2606 (±963)	3902 (±1434)
Astatotilapia burtoni R2	33 829 (±6968)	11 116 (±2703)	13 163 (±3279)

slightly higher number of SNPs detected in the de novo versus the reference pipeline (Table 5), but well within one standard deviation of each other. In addition, similar to the previous results from the coverage analysis, we found a slightly higher numbers of unique stacks, polymorphic loci and SNPs in the *C. lectularius* that were run with whole-genome amplification (Table 5). Differences in the number of catalog loci sequenced in the library were lower for the reference than the de novo runs when we compared *Sebastes* and *Ae. aegypti* unfiltered loci versus reference loci (first column Table 3), but when filtered for matching more than 2 individuals (at least 2 individuals had to have these loci), the number of SNPs were nearly equal, suggesting the assembly filtered out unusable reads (Table 3).

Discussion

We found a range of 4131–37 170 SNPs using a double-digest RAD sequencing (ddRAD sequencing) approach for a diverse group of nonmodel organisms. The paired enzymes *SphI* and *MluCI* cut sites worked consistently well for both vertebrate and invertebrate organisms from bedbugs to fish. This extends previous *in silico* reviews of this methodology by empirically testing multiple enzyme pairs and by providing derived estimates of sequence-able fragments for several species. Recently, we tested these enzyme pairs for sweet potato cultivars, the threatened North Carolina madtom (*Noturus furiosus*) and a marine seaweed (*Gacilaria sp.*) (M. O. Burford Reiskind, unpublished data) and found these enzymes produced the most consistent results in the serial digests, further confirming the general utility of this enzyme pair for detecting SNPs. This study provides an important step in generalizing a laboratory methodology for a wide range of projects using the same enzyme pair, which can further reduce both time and costs. There are three extensions to the present ddRAD sequencing methodology that we provided in this study: utilizing variable-length barcodes to reduce error and data loss, providing empirical evidence that some enzyme pairs work more consistently than others across a wide range of taxonomic groups and highlighting differences in SNP discovery using a reference genome for aligning reads versus de novo. In addition, we also provide empirical results of whole-genome amplification on the number of unique stacks, SNP discovery and coverage.

Previously, Peterson *et al.* (2012) provided a method for multiplexing ddRAD sequencing using fixed-length barcodes and Illumina indices to pool multiple libraries with the same barcodes. This provided greater capacity to multiplex individuals in a single sequencing lane, reducing costs substantially. However, in developing our protocol, we exchanged the fixed-length barcodes for variable-length barcodes, similar to Poland *et al.* (2012), while balancing base pairs to avoid phasing issues in the Illumina platform. Fixed-length barcodes are an issue because the cut site is identical in each fragment and fixed-length barcodes means that multiple reads will reach the same base pairs of the cut site at the same time. This is within the region (the first 20 + bps) where the Illumina platform is distinguishing whether nearby clonally amplified clusters are identical versus unique. If the platform pools clones that appear similar in this critical region yet are actually unique, it may subsequently reject this clone due to high sequence error downstream of this critical area. Unlike the fixed-length barcodes, the variable-length barcode allows for differences in the number of base pairs before identical cut sites and prevents grouping clones of different individuals on the platform or tossing out clones due to high sequencing error that are from different individuals. This is similar to the

Table 5 Comparison of average number of unique stacks, polymorphic loci and SNPs of the double-digest RAD sequencing SNP discovery using *SphI-MluCI* enzyme pairs across several taxonomic groups using both the de novo ($n = 2$) and reference pipelines in STACKS. Each column includes the average across individuals and the parenthetical standard deviation. These include all individuals and all stacks without *post hoc* filtering. Note that *Cimex lectularius* has a second set WGA

Species/Pipeline	Unique stacks	Polymorphic loci	SNPs Found
Sebastes de novo	146 051 (±39 158)	7033 (±2928)	9243 (±3715)
Sebastes reference	206 334 (±45 030)	18 223 (±4876)	37 170 (±9995)
Aedes aegypti de novo	23 708 (±15 561)	2490 (±1577)	4131 (±2783)
Aedes aegypti Reference	61 457 (±22 867)	6 863 (±3167)	14 651 (±6874)
Cimex lectularius de novo	38 233 (±20 703)	1990 (±1469)	3035 (±2276)
Cimex lectularius reference	48 649 (±19 473)	4961 (±2963)	10 764 (±6613)
Cimex lectularius de novo WGA	76 973 (±29 816)	2699 (±1477)	3940 (±1996)
Cimex lectularius reference WGA	105 712 (±31 164)	9259 (±3624)	19 171 (±7269)
Reticulitermes flavipes de novo	44 902 (±21 441)	5126 (±2414)	6859 (±3367)
Paralichthys lethostigma de novo	36 821 (±24 835)	3529 (±2637)	4699 (±3473)
Astatotilapia burtoni de novo	100 867 (±33 747)	13 685 (±3526)	16 990 (±4526)
Astatotilapia burtoni Reference	114 102 (±39 972)	7231 (±3141)	14 320 (±6176)

methodology used in GBS double digest and contributes to reduced error and data loss (Elshire *et al.* 2011; Poland *et al.* 2012). In addition, we modified how we attached the indices following Poland *et al.* (2012) to further reduce costs of the barcodes (see Supporting information).

Our study also found that many enzyme pairs did not work consistently well for different taxonomic groups. Given the many possible enzyme pairs suggested by previous research (Peterson *et al.* 2012) and to avoid bias towards a particular enzyme pair that works for one species and but may not be appropriate for another species, the results from this study suggest that some enzyme pairs perform more consistently across wide ranging taxonomic groups than others. In our study, two different enzyme pairs seemed to do equally well for disparate groups (e.g. the enzyme pair for this study and *NlaIII-EcoRI*), while other pairs performed well for some species but not others. Peterson *et al.* (2012) provided per individual read values from *in silico* estimates, but we were able to provide empirical estimates of sequenceable fragments for a variety of enzyme pairs. Moreover, the information gained from running this protocol with many different species will support general use in genomic facilities. It is difficult to know why certain enzymes were poor, or even failed, to cut some of the genomes. The temperature and buffer conditions may have not been optimal for some genomes, the particular specific cutting sequences may have been very rare for unknown reasons, or there may be unknown chemistry issues with certain enzyme pairs.

For three of four libraries, we found a reduced ability to detect SNPs in the absence of a reference genome (see Table 5), which suggests there are limitations when even a distant reference genome is not available. We found both higher number of unique stacks and SNPs in the reference runs compared to the de novo run, suggesting this is not just due to false positives generated when sequences are aligned to a distant reference genome. Whether this discrepancy is due to repetitive or duplicated regions overassembled in the de novo runs, or the de novo run generally combining unique stacks the result is a lower number of polymorphic loci or SNPs. This suggests an area for further investigation. Alternatively, the potential that alignments may produce false positives because the reference is more distantly related could explain the discrepancy. However, with a greater than two times, and in some cases four times, the number of SNPs between the two pipelines and the consistently greater number of unique stacks and not just SNPs make these other options possible but less likely. De novo and reference-based pipelines performed similarly in one species, *A. burtoni*, discovering a similar number of unique stacks and SNPs with both methods (Table 5). Notably, the *A. burtoni* individuals analysed came from a single family of a semi-inbred laboratory line, while individuals of the other three libraries included wild population sampling. Therefore, the *A. burtoni* samples contain no rare genetic variants, while samples in the other three libraries should have a significant number of rare SNPs. One hypothesis is that the SNP discovery discrepancy between the de novo and reference-based analysis in the three libraries with wild samples reflects increased power to detect rare variants in the reference-based pipeline.

Our analysis used the STACKS pipeline, and it is possible other de novo approaches may perform better. With this caveat, our results suggest that having a reference genome, even if it is from a related species, is better for the quantity of SNPs in SNP discovery than none at all. Therefore, it should be recognized that de novo analyses are likely conducted on a substantial subset of available data, and inferences from de novo analyses may be correspondingly weaker with regard to statistical power. For example, if the de novo pipeline in a comparative study found no differences between populations across a selection gradient, the lack of a positive finding may be due to low statistical power. We did not assess whether the reference or de novo pipelines made more or less valid SNP calls, only the number of SNPs.

An outstanding issue is whether this powerful SNP discovery tool can be used with small-yield samples given the quantity of DNA required for the ddRAD method. In particular, the question of whether whole-genome amplification (WGA) causes bias to specific parts of the genome or whether WGA introduces errors in the amplified regions (Pinard *et al.* 2006). If WGA subsamples the genome, we would expect a lower number of unique stacks, biased to the genomic regions that were amplified, and greater coverage of these regions (i.e. fewer stacks but with greater stack depth). Alternatively, if WGA introduces error to amplified regions, we would expect to find similar number of unique stacks between the WGA and non-WGA samples, with greater number of polymorphic loci and SNPs in the former. However, the results of the comparison between bedbugs subjected to WGA versus those not amplified suggested a slightly different result. In WGA samples, we found a greater number of unique stacks and greater stack depth, resulting in a greater number of SNPs. Given that our minimum stack depth, *m*, in both the de novo and reference pipeline was set at three, it is possible that fewer stacks met that requirement in the nonamplified genome than in the amplified one. Alternatively, a higher number of stacks could be produced by introduced error in some of the amplified regions. This result suggests that WGA in this species does result in bias or increased error or both

and should be further analysed to understand what that bias is to avoid erroneous conclusions.

In conclusion, we found consistent results using the ddRAD sequencing methodology using two enzyme pairs that worked for a broad range of taxonomic groups. Moreover, we further reduced costs and decreased error rates using variable-length barcodes and modified indices. Finally, we found that the use of a reference genome greatly increased the detection of usable SNPs for future analyses and that whole-genome amplification in some cases may over-represent parts of the genome or introduce error.

Acknowledgements

We thank the following for their assistance with collections of specimens: O. Blavín, C. Schal, W. Booth, A.A. Aguilar, A. Gill, B. Reading, P. Lounidos and I. Bargielowski. We thank C. Dashiell for assistance with the Bioanalyzer and sequencing logistics and E. Scholl for consultation on sequence alignment programs. R.W. Whetton for consultation on bioinformatic questions and T. Schultz for initial consultation on library building. Finally, we thank several anonymous reviewers for comments that improved this manuscript. This research was funded by the Wynne Innovation Grant from the CAL Dean's Enrichment Grant Programme at NCSU awarded to MO Burford Reiskind. Part of this research was also funded by the NSF, Grant IOS-1456765 awarded to RB Roberts. We also thank both the Applied Ecology and Entomology Departments at NCSU for providing matching funds to further fund this collaborative research.

References

Albertson RC, Powder KE, Hu Y, Coyle KP, Roberts RB, Parsons KJ (2014) Genetic basis of continuous variation in the levels and modular inheritance of pigmentation in cichlid fishes. *Molecular Ecology*, **23**, 5135–5150.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews*, **11**, 697–709.

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

Benoit JB, Adelman ZN, Reinhardt K *et al.* (2016) Unique features of the bed bug, a global human ectoparasite, identified through genome sequencing. *Nature Communications*, **7**, 1–10, doi:10.1038/ncomms 10165.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, **513**, 375–381.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Burford MO, Bernardi G (2008) Incipient speciation within a subgenus of rockfish (*Sebastosomus*) provides evidence of recent radiations within an ancient species flock. *Marine Biology*, **154**, 701–717.

Catchen J, 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.

Clark AG, Eisen MB, Smith DR *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Franchini P, Fruciano C, Speitzer ML *et al.* (2014) Genomic architecture of ecologically divergent body shape in a pair of sympatric Crater lake cichlid fishes. *Molecular Ecology*, **23**, 1828–1845.

Gamble T, Coryell J, Ezaz T, Lynch J, Scantlebury DP, Zarkower D (2015) Restriction site-associated DNA sequencing (RAD-seq) reveals an extraordinary number of transitions among gecko sex-determining systems. *Molecular Biology & Evolution*, **32**, 1296–1309.

Gregory S, Sekhon M, Schein J *et al.* (2002) A physical map of the mouse genome. *Nature*, **418**, 743–750.

Hand BK, Hether TD, Kovach RP *et al.* (2015) Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, **61**, 146–154.

Hess JE, Campbell NR, Docker MF *et al.* (2015) Use of genotyping by sequencing data to develop a high-throughput and multifunctional SNP panel for conservation applications in Pacific lamprey. *Molecular Ecology Resources*, **15**, 187–202.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Koshikawa S, Miyazaki S, Cornette R, Matsumoto T, Miura T (2008) Genome size of termites (Insecta, Dictyoptera, Isoptera) and wood roaches (Insecta, Dictyoptera, Cryptocercidae). *Naturwissenschaften*, **95**, 859–867.

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Lew RM, Finger AJ, Baerwald MR, Goodbla A, May B, Meek MH (2015) Using next-generation sequencing to assist a conservation hatchery: a single-nucleotide polymorphism panel for the genetic management of endangered delta smelt. *Transactions of the American Fisheries Society*, **144**, 767–779.

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

McCluskey BM, Postelthwait JH (2015) Phylogeny of zebrafish, a "model species", within Danio, a "model genus". *Molecular Biology & Evolution*, **32**, 635–652.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

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.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Nene V, Wortman JR, Lawson D *et al.* (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*, **316**, 1718–1723.

Ojima Y, Yamamoto K (1990) Cellular DNA contents of fishes determine by flow cytometry. *La Kromosomo II*, **57**, 1971–1888.

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.

Pinard R, de Winter A, Sarkis GJ *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.

Poland JA, Brown PJ, Sorrells ME, Jannick JL (2012) Development of high-density genetic maps for barley and what using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, **7**, e32253.

Reichwald K, Petzold A, Koch P *et al.* (2015) Insights into sex chromosome evolution and aging from the genome of short-lived fish. *Cell*, **163**, 1527–1538.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.

Roberts RB, Hu Y, Albertson RC, Kocher TD (2011) Craniofacial divergence and ongoing adaptation via the hedgehog pathway. *Proceedings of the National Academy of Sciences*, **108**, 13194–13199.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Venter J, Adams M, Myers E *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wagner CE, Keller I, Wittwer S *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Williams AV, Nevill PG, Krauss SL (2014) Next generation restoration genetics: applications and opportunities. *Trends in Plant Science*, **19**, 529–537.

Yoshizawa M, Robinson BG, Duboue ER *et al.* (2015) Distinct genetic architecture underlies the emergence of sleep loss and prey-seeking behavior in the Mexican cavefish. *BMC Biology*, **13**, 1–12.

MOBR designed the experiment, conducted the experiment, built the libraries, analyzed the data, and wrote the manuscript. PL and KC helped conduct the experiment and analyze the data. NBR performed fish husbandry of the cichlid fish and built libraries. MHR helped conduct the experiment and analyze the data, and provided biological samples. HVD performed fish husbandry and provided the flounder samples. ELV and RBR provided samples and helped analyze the data. JS provided logistical and technical support for post-library processing and sequencing of libraries. All authors edited the manuscript.

Data accessibility

Supporting information will be provided with the original submission for the online version of this manuscript. This will include the specific double-digest library building protocol that was used in the manuscript and the excel spreadsheet for estimating the number of sequence-able fragments from a serial digest. The reader can use this excel spreadsheet to estimate the number of sequence-able fragments from different enzyme pair digests.

Data from this manuscript will be available through Dr. Martha Burford Reiskind's DRYAD account. This will include the raw data used to generate the estimated number of sequence-able fragments for the serial digest, which is summarized in Table 1; the sequence of the 48 unique barcoded P1 adapters and the P2 adapters; the output from the STACKS program used to generate the number of unique stacks, polymorphic loci and SNPs for Table 5; and the data used to calculate the difference between the R1 and R2 reads for Table 4. In addition, the raw sequence data will be available upon request, as the data files far exceed the limits at DRYAD without addition payment.

Supporting Information

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

Appendix S1 Burford Reiskind Lab Double Digest RAD sequencing Protocol.

Appendix S2 Excel spreadsheet provided for designing a restriction enzyme experiment.