Received: 23 December 2017 | Revised: 30 March 2018 | Accepted: 8 May 2018

DOI: 10.1111/1755-0998.12905

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms

Jonathan B. Puritz | Katie E. Lotterhos

Department of Marine and Environmental Sciences, Northeastern Marine Science Center, Nahant, Massachusetts

Correspondence
Jonathan B. Puritz, Department of Biological Sciences, University of Rhode Island, 120 Flagg RD, Kingston, 02881, Rhode Island.
Email: jpuritz@uri.edu

Present address
Jonathan B. Puritz, Department of Biological Sciences, University of Rhode Island, 120 Flagg RD, Kingston, 02881, Rhode Island.

Funding information
Division of Biological Infrastructure, Grant/Award Number: 1722553; Division of Environmental Biology, Grant/Award Number: 1635423; KEL from Northeastern University

Abstract

Exome capture is an effective tool for surveying the genome for loci under selection. However, traditional methods require annotated genomic resources. Here, we present a method for creating cDNA probes from expressed mRNA, which are then used to enrich and capture genomic DNA for exon regions. This approach, called "EecSeq," eliminates the need for costly probe design and synthesis. We tested EecSeq in the eastern oyster, *Crassostrea virginica*, using a controlled exposure experiment. Four adult oysters were heat shocked at 36°C for 1 hr along with four control oysters kept at 14°C. Stranded mRNA libraries were prepared for two individuals from each treatment and pooled. Half of the combined library was used for probe synthesis, and half was sequenced to evaluate capture efficiency. Genomic DNA was extracted from all individuals, enriched via captured probes, and sequenced directly. We found that EecSeq had an average capture sensitivity of 86.8% across all known exons and had over 99.4% sensitivity for exons with detectable levels of expression in the mRNA library. For all mapped reads, over 47.9% mapped to exons and 37.0% mapped to expressed targets, which is similar to previously published exon capture studies. EecSeq displayed relatively even coverage within exons (i.e., minor "edge effects") and even coverage across exon GC content. We discovered 5,951 SNPs with a minimum average coverage of 80×, with 3,508 SNPs appearing in exonic regions. We show that EecSeq provides comparable, if not superior, specificity and capture efficiency compared to costly, traditional methods.

KEYWORDS
exome capture, population genomics, selection

1 | INTRODUCTION

The invention of next-generation sequencing has made it possible to obtain massive amounts of sequence data. These data have given insight into classical problems in evolutionary biology, including the repeatability of evolution (e.g., Jones et al., 2012), the degree of convergent evolution across distant taxa (e.g., Yeaman et al., 2016) and whether selection is driving changes in existing genetic variation or new mutations (e.g., Reid et al., 2016). Despite this rapid progress, it is still cost prohibitive to sequence dozens or hundreds of full genomes. This limits our ability to study the genomic basis of local adaptation, which requires large sample sizes for statistical power (De Mita et al., 2013; Hoban et al., 2016; Lotterhos & Whitlock, 2015). This leads to an inherent trade-off between sample size and genomic coverage, leading investigators to make decisions about whether to sequence more individuals (for higher power and precision) versus more of the genome (for making more accurate statements about the genetic basis of adaptation).

Reduced representation library preparation methods offer various kinds of random or targeted genome reduction, but the available

 © 2018 John Wiley & Sons Ltd

approaches have contrasting advantages and limitations. RADseq uses restriction enzymes to randomly sample the genome and is appropriate for linkage mapping and studying neutral processes like gene flow and drift (Puritz, Matz et al., 2014), but the data can be limited for understanding the genetic basis of adaptation (Catchen et al., 2017; Lowry et al., 2017a, 2017b; McKinney, Larson, Seeb, & Seeb, 2017). To focus on coding regions, some investigators have used RNAseq (De Wit, Pespeni, & Palumbi, 2015); however, only about a dozen individuals can be sequenced per lane because of log-fold differences in transcript abundance amongst loci. In addition, allele-specific expression limits the confidence in genotypes derived from RNAseq data (Pastinen, 2010), especially in pooled samples. Genomic DNA can also be pooled (Pool-seq), and allele frequencies for species or populations inferred directly from read counts in a single library (Schlötterer, Tobler, Kofler, & Nolte, 2014). Another increasingly popular option for increasing precision with larger samples while still maintaining coverage of the entire genome is low-coverage sequencing, which sequences every individual to very low (1×) coverage and uses genotype likelihoods instead of called genotypes to impute allele frequencies while still preserving information about individuals (Buerkle & Gompert, 2013; Therkildsen & Palumbi, 2017). Both Pool-seq and low-coverage sequencing cannot be used to understand the fitness of heterozygotes, and the types of statistical analyses that can be performed are limited, due to difficulty in determining haplotypes (e.g., Fariello, Boitard, Naya, SanCristobal, & Servin, 2013).

To overcome some of these limitations, many investigators have used capture approaches with biotinylated probes (Jones & Good, 2016). Capture approaches have the advantage of enriching the data for sequences of interest—allowing for individual-level data and a large number of individuals to be sequenced—but require the investigator to have genomic resources for probe design and then to purchase the probes from a company. For nonmodel species, the development of these resources takes time and a significant amount of bioinformatics expertise. In addition, for a population-level genomic study with 100s of individuals, probes may cost several tens of thousands of dollars, depending on how much sequence is captured. Overall, what is needed is a cost-effective approach to subsample genomes for coding regions, without previously developed genomic resources. Such an approach would allow for the assessment of rapid adaptation to environmental disasters such as Deepwater Horizon Oil Spill (Lee et al., 2017), and would also be useful for a variety of traditional molecular ecological and evolutionary applications such as investigating natural selection in wild and captive populations (Charlesworth & Charlesworth, 2017) and examining ecological speciation (Nosil & Schluter, 2011; Schluter & Conte, 2009).

Here, we present a novel, cost-effective method of exome capture that synthesizes probes in situ from expressed mRNA sequences. Expressed exome capture sequencing (EecSeq) builds upon existing approaches for in situ probe synthesis that rely on restriction enzymes (Schmid et al., 2017; Suchan et al., 2016). To improve capture efficiency, we developed a novel library preparation procedure that uses standardized procedures to synthesize cDNA from expressed RNA (without template reduction via restriction digest) and then create biotinylated probes from cDNA (see Figure 1 for a conceptual diagram). The EecSeq design includes custom RNA library adapters that offer several major advantages. The custom adapters are fully compatible with duplex-specific nuclease normalization, which is included in the protocol to reduce log-fold differences in expression—resulting in more even coverage across high- and low-expressed transcripts. The custom adapters also allow for probe sequencing—before normalization if differential expression data is desired, or after normalization if probe abundance data is desired. Moreover, the adapters are easily removed with a single enzymatic treatment before biotinylation, preventing any interference during hybridization.

Our approach is cost-effective and does not require any prior genomic resources, making it a good choice for studies seeking to understand adaptation in exomes. The approach, however, is limited in the sense that the probes are designed from expressed RNA, and so investigators should be careful to choose which tissues and life stages would be relevant. Here, we show proof-of-concept of the approach in the eastern oyster (*Crassostrea virginica*) and find that the performance of the approach is comparable, if not superior, to the performance of published exome capture data sets where probes were designed from sequence data and purchased from a company.

2 | METHODS

2.1 | Experimental overview

Expressed exome capture sequencing is designed with two specific goals: (a) to eliminate the need for expensive exome capture probe design and synthesis and (b) to focus exon enrichment of genes that are being expressed relevant to tissue(s) and condition(s) of interest. To illustrate this conceptually, we exposed adult oysters to a stressor (extreme heat) that would generate a predictable gene and protein expression profile (expression of heat shock proteins). Having a predictable coverage profile in the probes allowed us to evaluate whether the genomic DNA in these exons were captured by the probes. Note, however, that this experiment is not specifically part of the EecSeq method and that the investigator can choose appropriate tissue(s) and condition(s) of interest. The steps to probe synthesis and capture are visualized in Figure 1.

2.2 | Heat shock exposure, tissue collection and nucleic acid extraction

Eight adult *Crassostrea virginica* individuals were collected and acclimated to a flow-through seawater system for 24 hr. After acclimation, individuals were randomly assigned to two treatments, control and heat shock (HS). HS individuals were placed a small aquaria filled with 36°C filtered seawater for 1 hr while control individuals were kept in an identical aquarium filled with 14°C (ambient) filtered seawater. Immediately after the exposure period, all individuals were shucked and mantle tissue was extracted and frozen in liquid

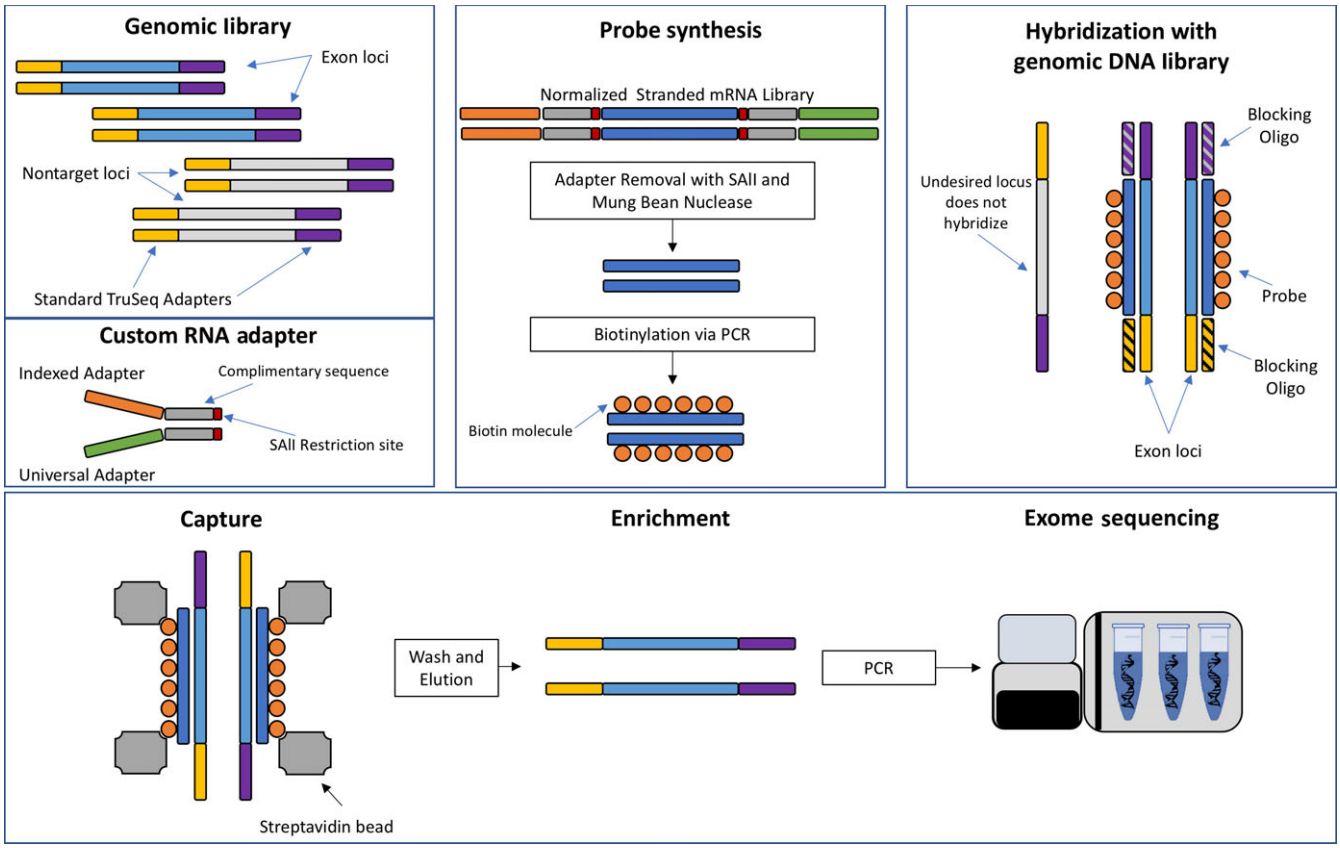

FIGURE 1 Conceptual diagram of expressed exome capture sequencing. Upper left panel: The shotgun genomic DNA library that will be captured with probes. Middle left panel: EecSeq relies on custom RNA adapters that contains a SAII restriction site. Middle upper panel: The adapters are incorporated into a mRNA library preparation that is normalized with duplex-specific nuclease. Adapters are then removed with a SA1I restriction digest, cDNA probes are subsequently blunted with mung bean nuclease and biotinylated via a PCR product. Upper right panel: The probes are then hybridized to the shotgun genomic library with TruSeq style adapters. Exon loci bind to the cDNA probes. Lower panel: Hybridized exon loci and probes are then captured with magnetic Streptavidin beads. The captured exome fragments are washed several times, eluted, enriched with PCR and then sequenced

nitrogen in duplicate. DNA was extracted using the DNeasy kit (Qiagen) and RNA was extracted using TRI Reagent Solution (Applied Biosystems) using included, standard protocols. DNA was visualized on an agarose gel and quantified using the Qubit DNA Broad Range kit (Invitrogen). RNA was visualized on an Agilent BioAnalyzer using the RNA 6000 Nano kit and was quantified using the Qubit High Sensitivity Assay Kit (Invitrogen).

2.3 | Expressed exome capture sequencing

A complete and updated EecSeq protocol can be found at https://github.com/jpuritz/EecSeq.

2.3.1 | RNA Adapters

Custom RNA adapters were used in this protocol. The RNA adapters were similar to the Illumina TruSeq design, but include the SAII restriction site at the 3′ end of the "Universal adapter" and at 5′ end of the "Indexed adapter." The presence of this restriction site allows the Illumina sequence to be removed before hybridization to prevent interference. Note that the adapters used in this study had an erroneous deletion of a Thymine in position 58 of "Universal_SAI1_Adapter" and in position 8 of all four indexed adapters (the corrected versions are shown in Table 1, and erroneous version used in this study are shown in Supporting Information Table S1). Adapters were annealed in equal parts in a solution of Tris-HCl (pH 8.0), NaCl and EDTA; heated to 97.5°C for 2.5 min; and then cooled at a rate of 3°C/min until the solution reached a temperature of 21°C.

2.3.2 | mRNA library preparation and normalization

Probes were made from two (of four) control individuals and two (of four) exposed individuals. The first step for this subset of individuals was to prepare stranded mRNA libraries using the Kapa Stranded mRNA-Seq Kit (KAPA Biosystems) with the following modifications: Custom adapters were used, 4 micrograms of RNA per individual was used as starting material, half volume reactions were used for all steps, adapters were used at a final reaction concentration of 50 nM during ligation, and 12 cycles of PCR were used for enrichment. Complete libraries were visualized on a BioAnalyzer using the DNA

TABLE 1 Corrected adapter sequences for mRNA library preparation

Oligo name	Sequence
Universal_SAI1_Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCGACT*T
Indexed_Adapter_SAI1_I5	P*AGTCGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
Indexed_Adapter_SAI1_I8	P*AGTCGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG
Indexed_Adapter_SAI1_I9	P*AGTCGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG
Indexed_Adapter_SAI1_I11	P*AGTCGACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG

Note. Oligos are listed in a 5′ to 3′ orientation with "P" indicates a phosphorylation modification to enable ligation.

1000 kit and quantified using fluorometry, and then, 125 ng of each library was taken and pooled to single library of 500 ng.

To reduce the abundance of highly expressed transcripts in our final probe set, complete libraries were normalized following Illumina's standard protocol for DSN normalization. First, the cDNA library was heat denatured and slowly allowed to reanneal. Next, the library was treated with duplex-specific nuclease (DSN), which will remove abundant DNA molecules that have properly annealed. After DSN treatment, the library was solid phase reversible immobilization (SPRI) purified and enriched via 12 cycles of PCR. A subsample of probes was exposed to an additional 12 cycles of PCR to test for PCR artefacts in probe synthesis. The normalized cDNA library was visualized on a BioAnalyzer using the DNA 1000 kit, quantified with a Qubit DNA Broad Range kit (Invitrogen) and then split into two equal volume tubes, one to be saved for sequencing and one for probe synthesis. The DNS-normalized libraries were sequenced on one half lane of HiSeq 4000 by GENEWIZ (www.genewiz.com).

2.3.3 | Probe synthesis

To remove the sequencing adapters, the cDNA library was treated with 100 units of SalI-HF restriction enzyme (New England Biolabs) in a total volume of 40 μl at 37°C for 16 hr. After digestion, the digested library was kept in the same tube, and 4.5 μl of 10× Mung Bean Nuclease Buffer and 5 units of Mung Bean Nuclease (New England Biolabs) were added to remove overhangs. The reaction was then incubated at 30°C for 30 min. An SPRI clean-up using AMPure XP (Agencourt) was completed with an initial ratio of 1.8×. After visualization of the library on an Agilent BioAnalyzer, a subsequent SPRI clean-up of 1.5× was completed to remove all digested adapters. The clean, digested cDNA fragments were then biotin labelled using the DecaLabel Biotin DNA labeling kit (Thermo Scientific) using the included protocol. The labelling reaction was then cleaned using a 1.5× SPRI clean-up and fluorometrically quantified. To test the effects of additional PCR cycles on probe effectiveness, 40 ng of the original, normalized cDNA library was subjected to an additional 12 cycles of PCR, and then converted to probes as described above.

2.3.4 | Genomic DNA library preparation

Capture was performed on a standard genomic DNA library. 500 ng of genomic DNA from all eight individuals was sheared to a modal peak of 150 base pairs using a Covaris M220 Focused-ultrasonicator. The sheared DNA was inserted directly into step 2.1 of the KAPA HyperPlus kit with the following modifications: Half reaction volumes were used, and a final adapter:insert molar ratio of 50:1 was used with custom TruSeq style, barcoded adapters (note: the adapters contained erroneous mismatches in the barcodes between the top and bottom oligos; the original oligonucleotide sequences can be found in Supporting Information Table S2 and corrected versions in Supporting Information Table S3). After adapter ligation, individuals were pooled into one single library, and libraries were enriched with six cycles of PCR using primers that complemented the Illumina P5 adapter and Indexed P7 (Supporting Information Table S2). The final library was fluorometrically quantified and analysed on an Agilent BioAnalyzer.

2.3.5 | Hybridization

Three replicate captures were performed using the set of original probes and the set of probes with 12 extra cycles of PCR. The hybridization protocol closely followed that of Suchan et al. (2016). 500 ng of probes and 500 ng of genomic DNA library were hybridized along with blocking oligonucleotides (Table 2) at a final concentration of 20 μM in a solution of 6× SSC, 5 mM EDTA, 0.1% SDS, 2× Denhardt's solution, and 500 ng c_0t-1 DNA. The hybridization mixture was incubated at 95°C for 10 min and then 65°C for 48 hr in a thermocycler. The solution was gently vortexed every few hours, although not overnight.

2.3.6 | Exome capture

Forty microlitre of hybridization mixture was added to 200 μl of DynaBeads M-280 Streptavidin beads (Thermo Fisher Scientific). The beads and hybridization mixture were then incubated for 30 min at room temperature. The mixture was then placed on a magnetic stand until clear, and the supernatant was removed. This was followed by four bead washes under slightly different conditions. First, the beads were washed with 200 μl 1× SSC and 0.1% SSC solution, incubated at 65°C for 15 min and placed on the magnet stand, and the supernatant was removed. Second, the beads were washed with 200 μl 1× SSC and 0.1% SSC solution incubated at 65°C for 10 min, placed on the magnet stand, and the supernatant was removed. Third, the beads were washed with 200 μl 0.5 SSX and 0.1% SDS solution,

TABLE 2 Exome capture sequencing, filtering, and mapping statistics

Replicate capture pool	Raw reads	Filtered reads	Mapped reads	% Duplicates	Filtered mapped reads	% mapping to mitochondrial genome
EC_2	53,493,950	53,118,952	42,403,525	5.8	35,955,539	1.7
EC_4	44,935,340	44,651,228	35,275,663	6.1	29,519,347	1.6
EC_7	43,745,614	43,448,296	35,007,184	5.6	29,723,437	2.2
EC_1	41,402,996	41,145,724	32,668,750	4.5	27,940,717	1.8
EC_3	56,127,536	55,753,268	44,605,960	5.0	38,103,268	1.9
EC_12	46,068,760	45,750,394	37,227,067	6.2	31,497,298	2.2

Note. EC_2, EC_4 and EC_7 are the three replicate captures with the original probe pool, and EC_1, EC_3 and EC_12 are the replicate captures with the probe pool exposed to 12 extra rounds of PCR.

incubated at 65°C for 10 min, placed on the magnet stand, and the supernatant was removed. At last, the beads were washed with 200 μl 0.1× SSC and 0.1% SDS, incubated at 65°C for 10 min, placed on the magnet stand, and the supernatant was removed. In conclusion, DNA was eluted from the beads in 22 μl of molecular grade water heated to 80°C for 10 min. The solution was placed on the magnet and the supernatant was saved. The hybridized fragments were then enriched with 12 cycles of PCR using the appropriate P5 and P7 PCR primers and cleaned with 1× AMPure XP with a final elution in 10 mM Tris-HCl (pH 8.0). The six replicate captures—three with the original probes and three with probes exposed to 12 additional cycles of PCR—each containing eight uniquely barcoded individuals, were sequenced on one half lane (separate from the RNA libraries) on the HiSeq 4000 platform by GENEWIZ (www.genewiz.com).

2.4 | Bioinformatic analysis

All bioinformatic code, including custom scripts and a script to repeat all analyses, can be found at https://github.com/jpuritz/EecSeq/tree/master/Bioinformatics.

2.4.1 | RNA libraries

RNA reads were first trimmed for quality and custom adapter sequences were searched for with Trimmomatic (Bolger, Lohse, & Usadel, 2014) as implemented in the DDOCENT PIPELINE (version 2.2.20; Puritz, Hollenbeck, & Gold, 2014). Reads were then aligned to release 3.0 of the *Crassostrea virginica* genome (Accession: GCA_002022765.4) using the program STAR (Dobin et al., 2013). The genome index was created using NCBI gene annotations for splice junctions. Reads were aligned in a two-step process, first using the splice junctions in the genome index, and then again using both the splice junctions in the index and additional splice junctions found during the first alignment. Alignment files from the four libraries were then merged with SAMTOOLS (version 1.4; Li et al., 2009) and filtered for MAPQ >4, only primary alignments, and reads that were hard/soft clipped at less than 75 bp. SAMTOOLS (Li et al., 2009) and BEDTOOLS (Quinlan, 2014) were used to calculate read and per bp coverage levels for exons, introns and intergenic regions.

2.4.2 | EecSeq libraries

Raw reads were first trimmed using the standard methods in the DDOCENT PIPELINE (version 2.2.20; Puritz, Hollenbeck et al., 2014). The DNA adapters contained erroneous mismatches between the top and bottom oligos in the barcode (original oligonucleotide sequences can be found in Supporting Information Table S2 and corrected versions in Supporting Information Table S3). These differences prevented demultiplexing beyond the capture pool level and also lead to potentially erroneous base calls within the first 7 bp of sequencing. To remove these artefacts, the first 7 bp of every forward read was clipped. In addition, adapter sequences were searched for in the paired-end sequences using custom scripts. After trimming, reads were aligned to the reference genome using BWA (Li & Durbin, 2009) with the mismatch parameter lowered from 4 to 3, and the gap opening penalty lowered from 6 to 5. PCR duplicates were marked using the *MarkDuplicatesWithMateCigar* module of Picard (http://broadinstitute.github.io/picard), and then, SAMTOOLS (Li et al., 2009) was used to remove duplicates, secondary alignments, mappings with a quality score less than ten and reads with more than 80 bp clipped. SAMTOOLS (Li et al., 2009) and BEDTOOLS (Quinlan, 2014) were used to calculate read and per bp coverage levels for exons, introns and intergenic regions. FreeBayes (Garrison & Marth, 2012) was used for variant calling. Variants were decamped into SNPs and InDels using vcflib (https://github.com/vcflib/vcflib). InDels were then discarded, SNPs below a minimum quality score of 20 were filtered out using VCFTOOLS (Danecek et al., 2011). SNPs were then filtered by various levels of minimum mean depth across captures.

2.4.3 | Calculating capture efficiency

EecSeq is unique amongst exome capture methods because the probes are not designed directly; that is, there is no set of a priori targets. In addition, EecSeq is designed to capture exons that are expressed in the samples used to create probes—not the entire exome. To compare EecSeq to other capture methods, capture targets were defined as exons that had more than 35× coverage in the RNAseq (probe) data and confidence intervals were generated by defining capture targets as 20× RNAseq coverage and 50× RNAseq coverage. We also calculated a conservative, near-target range of

150 bp on either side of the defined targets. This range corresponds to the modal DNA fragment length used for the capture libraries with the expectation that exon probes could capture reads that far from the original target.

3 | RESULTS

3.1 | Probe synthesis

After normalization and subsequent 12 cycles of PCR enrichment, the cDNA library consisted of ~2,500 ng. For the original probe set synthesis, one microgram of the original cDNA library yielded 2,298 ng of probes, as the biotinylation occurs via a DNA polymerase. In contrast for the second probe set, 40 ng of the original normalized cDNA library was subjected to 12 cycles of PCR and then probe synthesis, yielding approximately 1,650 ng of probes. For each capture, 500 ng of probes was hybridized with 500 ng of genomic DNA library. This means that the original probe set could be used for a little over four captures but taking advantage of additional PCR cycles (which did not affect the results, see below), 1 microgram of cDNA library could generate over 40,000 ng of probes, enough for 800 captures. Successful captures were also performed with as little as 250 ng (data not shown), potentially increasing efficiency further.

3.2 | RNA sequencing results

RNA sequencing, filtering and mapping statistics can be found in Supporting Information Table S4. After filtering, a total of 21,990,025 RNA reads were mapped uniquely to the eastern oyster genome. Of the total RNA reads, 78% mapped to genic regions of the genome, and 58% mapped to annotated exon regions. Across all exonic bases in the genome, less than 5% had more than 50× coverage; however, over 16% had at least 20× coverage and over 45% had at least 5× coverage (Figure 2).

3.3 | Exome capture sequencing results

Six replicate capture pools of the same eight individuals were sequenced on half a lane of Illumina HiSeq (three replicates from probes that had been enriched via 12 cycles of PCR and three replicates from probes that had been enriched via 24 cycles of PCR). A summary of exome capture sequencing, filtering and mapping statistics are shown in Table 2. On average, there were 47,629,033 raw reads (forward and paired-end) per capture pool and an average of 32,123,268 mapped reads per capture pool after filtration. Across the entire oyster genome, RNA sequencing coverage and exome sequencing coverage was highly correlated (Supporting Information Figure S1), and across all exon regions total RNA coverage predicted 72.6% of the variation in exome capture coverage (Figure 3; log-log transformation, $R^2 = 0.72619$, $p < 0.0001$). Coverage across all exons and expressed exon targets was highly correlated ($0.984 < r < 0.996$) across all replicate captures, and the average capture of pools with standard probes and the average capture of pools with probes with extra PCR was virtually identical ($R^2 = 99.1$; $p < 0.0001$).

3.4 | Exome capture efficiency

Capture sensitivity, or the percentage of targets covered by at least one read (1×), was high across all replicate pools, regardless of target set (Table 3). Across all known exons, sensitivity was on average 86.8% across replicate capture pools, and across all defined target sets, sensitivity was over 99.4%. Increasing the sensitivity threshold from 1× to 10× lowers the sensitivity across all exons but has little effect on sensitivity across defined target sets (Supporting Information Table S5). Sensitivity can also be measured at the per bp level instead of per exon. The per cent of target bases captured is shown as a function of sensitivity threshold (read depth of capture libraries) in Figure 4.

Capture specificity is the percentage of mapped reads that fall within target regions. Across all exons, capture pools averaged 47.9% reads on target, 6.8% of reads near target (falling within

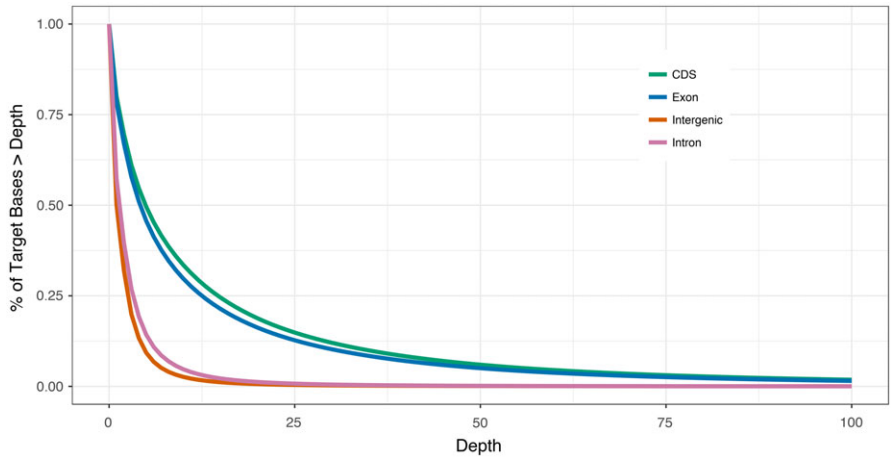

FIGURE 2 Distribution of RNA reads across regions of the oyster genome. The percentage of bases for different genomic regions (for entire exons including the untranslated regions [Exon], for just the coding sequences [CDS] within exons, for intergenic regions, and for introns) are shown at various coverage levels

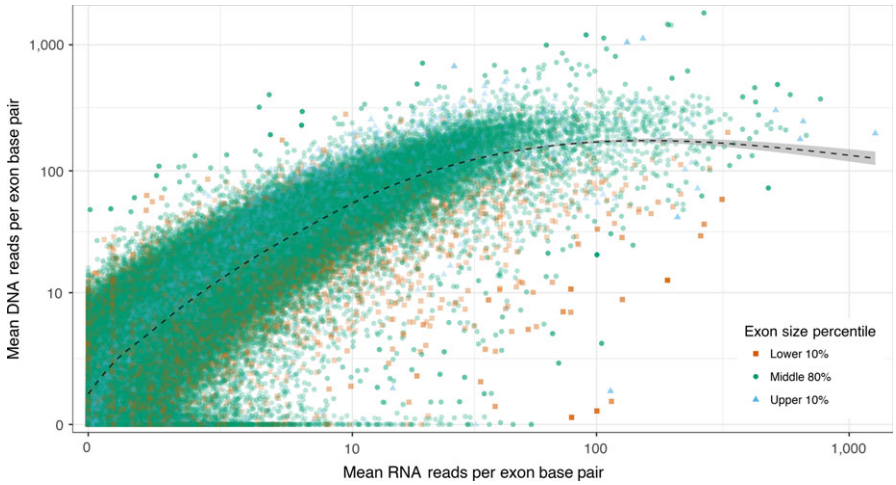

FIGURE 3 Mean DNA and RNA coverage per basepair across all exons. DNA depth, or mean reads per exon basepair, was calculated by taking the average of the mean per base pair coverage for each exon across all six captures. RNA depth, or mean reads per exon basepair, was calculated by taking the average of the mean per base pair coverage for each exon across all four RNA libraries. The shape and colour of each point was determined by the percentile size of the respective exon (lower 10% <59 bp, upper 10% >517 bp, and the middle 80% was between 57 and 517 bp). The dashed line is a general additive model smoother

TABLE 3 Exome capture sensitivity with a 1× threshold

	Capture pool					
Targets	**EC_2 (%)**	**EC_4 (%)**	**EC_7 (%)**	**EC_1 (%)**	**EC_3 (%)**	**EC_12 (%)**
All Exons	88.0	86.0	85.8	86.5	87.9	86.4
20×R Exons	99.5	99.4	99.4	99.4	99.5	99.4
35×R Exons	99.6	99.6	99.6	99.6	99.6	99.6
50×R Exons	99.7	99.7	99.7	99.7	99.7	99.7

Note. Sensitivity is the percentage of target bp with at least one read mapping successfully. Here, targets are broken up into subsets: All annotated exons, exons with at least 20× coverage from the RNA library, exons with at least 35× coverage from the RNA library and exons with at least 50× coverage from the RNA library. EC_2, EC_4 and EC_7 are the three replicate captures with the original probe pool, and EC_1, EC_3 and EC_12 are the replicate captures with the probe pool exposed to 12 extra rounds of PCR.

150 bp of an exon, one modal read length) and 45.3% of reads off target (more than 150 bp away from an exon). Across defined expressed exon targets (exons that sequenced to 35× read depth), capture pools averaged 37.1% (CI 33.6%–41.4%) reads on target, 3.55% (CI 3.0%–4.4%) of reads near target, and 59.38% (CI 54.2%–63.4%) reads off target.

For all exons, between the 10th and 90th percentile of exon length (59–517 bp), the mean per basepair coverage averaged 17.75× ± 0.06× for each capture pool of eight individuals. When considering target exons (35× coverage in RNA-derived probes), the mean per basepair coverage increased to 61.22× ± 0.23× on average for each capture pool. This breaks down to approximately 7.66 reads on average per individual per bp within expressed exome targets. Within exons, mean per basepair coverage was evenly distributed across all base pairs with only slightly lower coverage at the 5′ or 3′ edges of exons compared to the middle of exons (Figure 5; Supporting Information Figure S2).

Mean capture coverage also did not appear to relate to the GC content of the target exon (Figure 6), although it did appear to peak near the mean GC content of 43.57%. To test this, we calculated the reciprocal of the absolute value of the difference between each exon GC content and the average GC content, and then tested for a linear relationship to mean coverage. Although we found this relationship to be significant ($p < 0.0008$), it explained only the 0.0033% of the variance in coverage, confirming that exon GC content did not affect exon capture in a meaningful way.

Coverage did vary significantly between untranslated regions (UTR) within exons and coding sequence (CDS) within exons (Welch's test $t = 40.063$; $df = 135580$; $p < 0.0001$) with a mean coverage for UTR equalling 11.59× ± 0.0864 and a mean coverage for CDS equalling 17.71× ± 0.1261. This small but significant coverage difference was also evident as the per cent of target bases greater than a given read depth (Supporting Information Figure S3). This pattern was not surprising, however, because the same pattern was observed for the RNA reads (CDS mean coverage = 13.65× ± 0.2011; UTR mean coverage = 8.25 ± 0.1275; Welch's test $t = 22.677$; $df = 129{,}300$; $p < 0.0001$), indicating that the probes also had lower coverage in UTR compared to CDS.

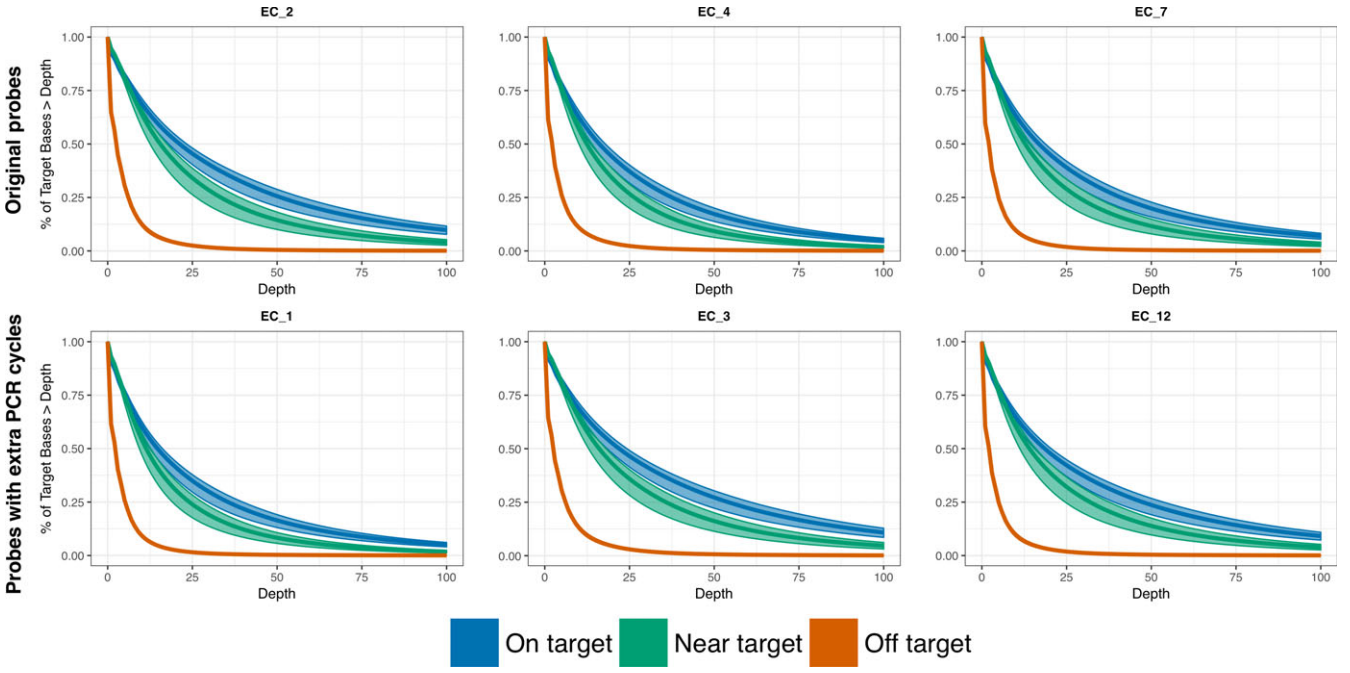

FIGURE 4 Per base pair EecSeq capture sensitivity. To measure EecSeq capture (DNA) sensitivity, capture targets were defined as exons that had more than 35× coverage in the RNAseq (probe) data. Confidence intervals were generated by defining capture targets between 20× RNAseq coverage and 50× RNAseq coverage. Near-target mapping were 150 bp on either side of the defined targets. This range corresponds to the modal DNA fragment length used for the capture libraries with the expectation that exon probes could capture reads that far from the original target. EC_2, EC_4 and EC_7 are the three replicate captures with the original probe pool, and EC_1, EC_3 and EC_12 are the replicate captures with the probe pool exposed to 12 extra rounds of PCR. Depth in this figure is the depth of DNA reads from EecSeq captures

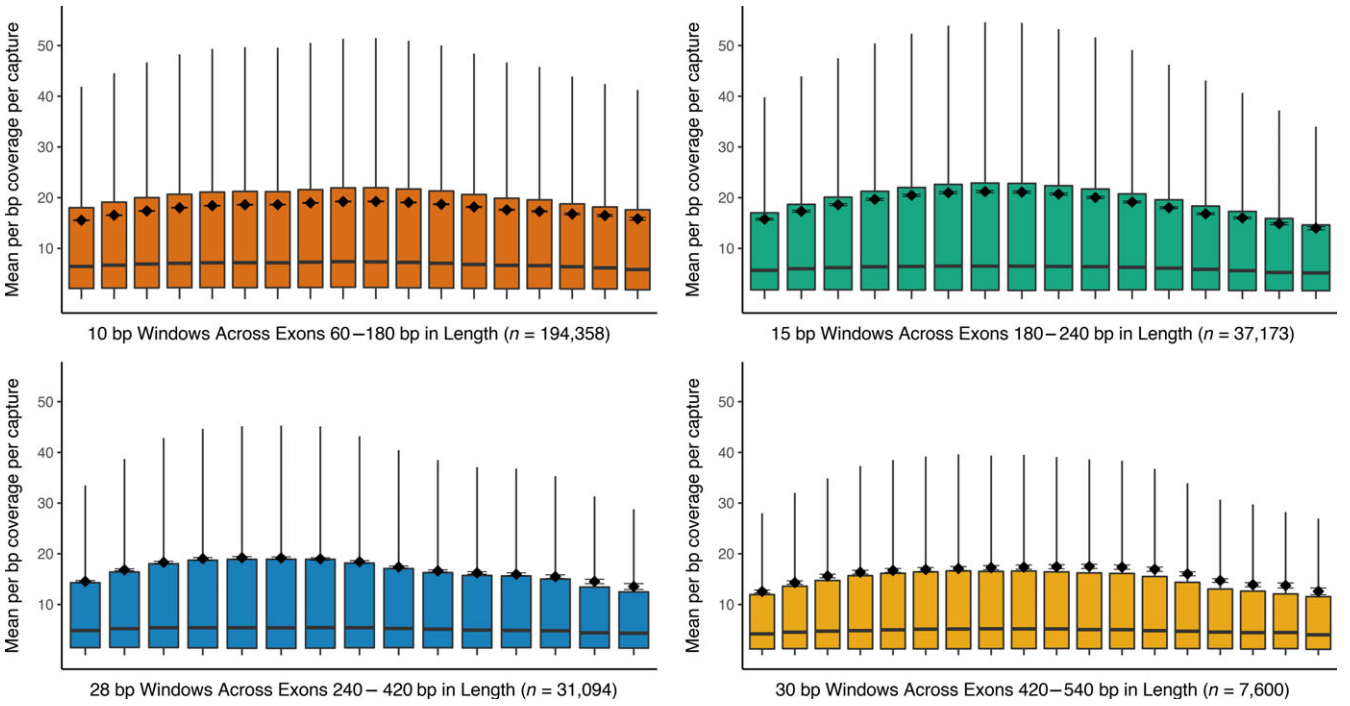

FIGURE 5 Boxplots of mean per basepair coverage levels plotted across exons size windows. All annotated exons were broken into 10–30 bp windows depending on overall size, and the mean per basepair coverage per capture was calculated for each window size. The line each box represents the median of mean coverage values and the box surrounds the 25th and 75th percentiles. The mean of each bin class is plotted as a black diamond with standard error bars around it. Outlier points were not plotted. Note that the data for this graph are for all annotated exons, regardless of expected capture. See Supporting Information Figure S3 for a similar plot focused on an expressed target set

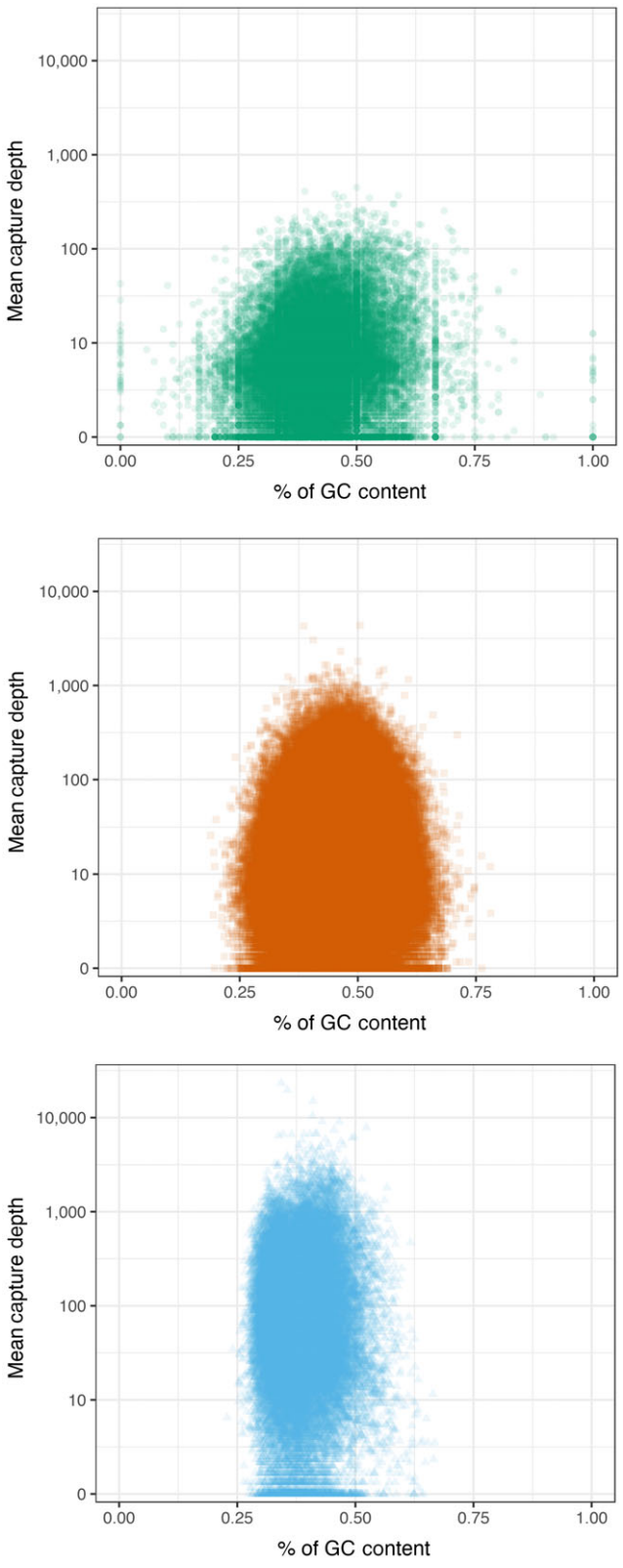

FIGURE 6 Mean capture depth plotted against exon GC content. Exons were broken up into three size windows: (1) Lower 10%—exons less than 57 bp, (2) Middle 80%—exons greater than 56 bp and less than 518, (3) Upper 10%—exons greater than 517 bp

3.5 | Expressed exon capture

To visualize the relationship between coverage and an expected expressed target, we plotted coverage of the six capture pools along two heat shock proteins, Heat Shock cognate 71 kDa (NCBI Reference Sequence: XM_022472393.1, Figure 7) and Heat Shock 70 kDa protein 12B-like (NCBI Reference Sequence: XM_022468697.1; Supporting Information Figure S4). As expected, exons in both genes show elevated coverage that corresponded to the coverage of the mRNA-derived probes, especially along regions with corresponding CDS with few reads mapping to intronic or intergenic regions.

3.6 | SNP discovery

A total of 1,011,107 raw SNPs were discovered with 909,792 SNPs having a quality score higher than 20. A total of 99,169 high-quality SNPs were found within known exons. Of these, 31,579 exome SNPs had at least an average of 16× coverage, 15,760 exome SNPs had at least an average of 32× coverage, 8,837 exome SNPs had at least an average of 48× coverage, and 3,508 exome SNPs had at least an average of 80× coverage with an additional 2,443 80×-SNPs found outside of exon regions.

4 | DISCUSSION

Expressed exome capture sequencing is a novel design for exome capture that uses in situ synthesized biotinylated cDNA probes to enrich for exon sequences, thereby removing the requirement of a priori genomic resources, costly exon probe design and synthesis. Here, we showed that EecSeq target enrichment had high levels of sensitivity, with comparable if not superior performance and specificity to traditional methods (see summary of comparisons in Table 4). EecSeq exon enrichment showed even coverage levels with exons and across exons with differing levels of GC content. In conclusion, we showed that EecSeq can quickly and cheaply generate thousands of exon SNPs.

4.1 | Benefits of EecSeq

4.1.1 | Diverse probes

With EecSeq, cDNA exon probes are constructed in situ from extracted mRNA, and this allows for the design of a high-diversity probe pool. Traditional sequence capture probes are typically designed from a single reference genome or individual, and this may limit capture efficiency on individuals with different SNPs, insertions or deletions than the reference. While probes been successfully used to capture sequences in quite divergent species (less than 5% sequence divergence, Jones & Good, 2016), there is

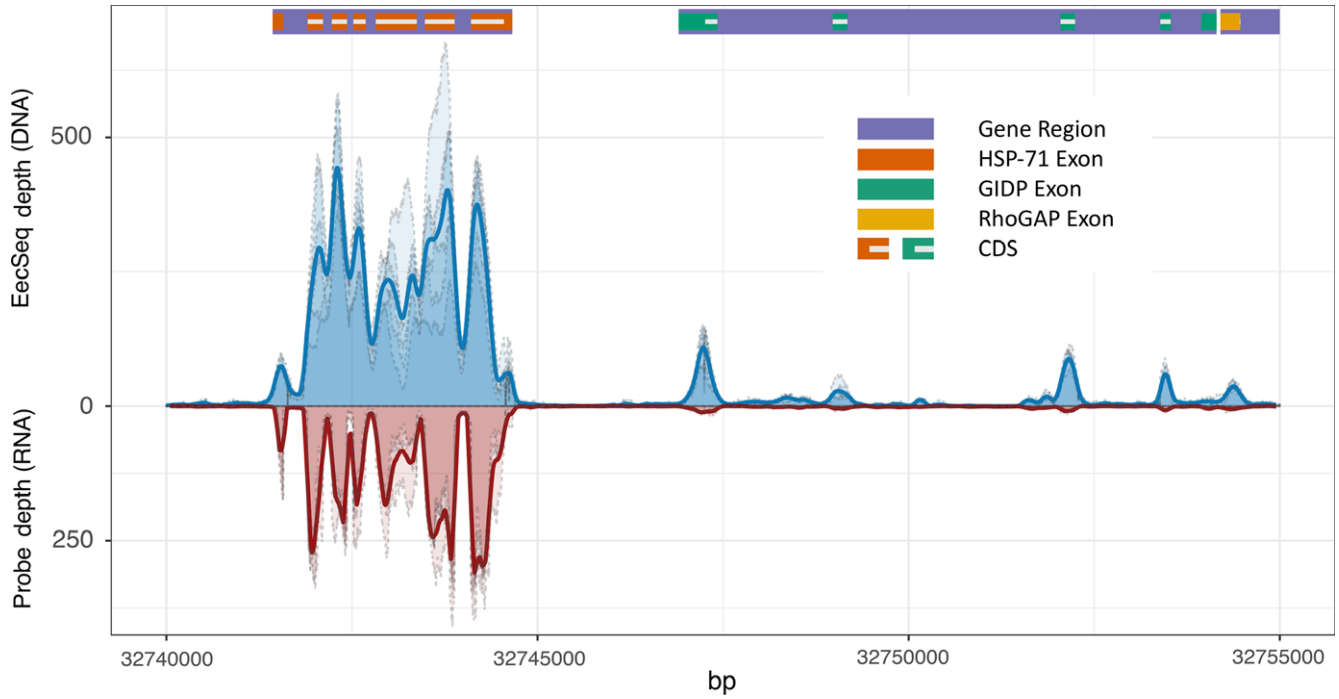

FIGURE 7 EecSeq capture and probe coverage across Heat Shock cognate 71 kDa. Coverage for each replicate capture pools is plotted along base pairs 32,740,000 to 32,755,000 of reference Chromosome NC_035780.1 containing the full gene region of Heat Shock cognate 71 kDa (NCBI Reference Sequence: XM_022472393.1), predicted glucose-induced degradation protein 8 homolog (NCBI Reference Sequence: XM_022486802) and a partial gene region for rho GTPase-activating protein 39-like (NCBI Reference Sequence: XM_022486743.1). Each exome capture pool coverage is plotted in light blue with dashed grey border, and a rolling 100-bp window average across all pools is plotted in dark blue. Each RNAseq (probe) sample coverage is plotted in light red with dashed grey border and a rolling 100-bp window average across all pools is plotted in dark red. Gene regions are marked in purple with exons colour coded by gene. Coding sequence (CDS) is marked by a white bar within exon markers

evidence that capture success declines as sequences become less related to the reference. Portik, Smith, and Bi (2016) found that for each per cent increase of pairwise divergence, missing data increased 4.76%, sensitivity decreased 4.57%, and specificity decreased 3.26%. Even with well-designed, commercially available capture kits for human exon capture, Sulonen et al. (2011) found that allele balances for heterozygous variants tended to have more reference bases than variant bases in the heterozygous variant position across all methods for probe development. Insertions and deletions (InDels) are arguably an even larger problem, as these would decrease hybridization with a probe due to a frameshift.

4.1.2 | Longer probes

Traditional exome capture relies on synthesized RNA or DNA baits. These baits can be relatively small (60 bp; Bi et al., 2012) or range between 95 and 120 bp (Chilamakuri et al., 2014; Clark et al., 2011; Nadeau et al., 2012; Sulonen et al., 2011). In contrast, EecSeq probes have a modal length of 150 bp but also range up to over 400 bp (data not shown). The longer length of EecSeq probes likely helps to buffer against divergence between probes and targets. The longer probes may also be the reason why we observed relatively little GC bias in coverage across exons, and may help explain the uniformity of coverage within exons in EecSeq data.

4.1.3 | Cost

EecSeq provides significant cost and time savings over traditional exome capture and RNA sequencing (RNAseq). No a priori genomic information is necessary for EecSeq, saving substantial time and money for obtaining these data in nonmodel organisms. Likewise, the cost of synthesizing the probes is significantly reduced because probes can be made in-house and do not have to be designed by a company. On a per sample basis, EecSeq is also significantly cheaper than RNAseq because (a) commercial DNA library preps are cheaper than those for mRNA, and (b) more individuals can be multiplexed on a single lane. For example, the cost of RNA seq is $246 per sample (cost estimated using the same RNA kits used with EecSeq and ½ reactions) and assuming that 12 RNAseq libraries can be sequenced in a single lane of Illumina HiSeq (($1,008; cost of the kit Kapa Biosystems Stranded mRNA-Seq Kit with 24 reactions or 48 half-reactions)*(1/48; the amount used per sample) + $2,700/12 = $246 per sample). The equivalent cost per sample for EecSeq is $48.02 per sample (for 96 samples in one lane of HiSeq; Supporting Information Table S6) or $62.08 per sample if a more

TABLE 4 Comparing specificity and sensitivity across capture methods

Reference and species	Num. target genes or exons	Sensitivity % of targeted regions >10× depth	Specificity % of reads mapping to targeted bases	% of reads mapping near target	% of reads mapping off target	Notes
EecSeq (this study) eastern oyster *Crassostrea virginica*	71,105 (51,096–110,020)	All exons: 54.7% Expressed Exons: 98.8% (97.4%–99.1%)	All exons: 47.8% Expressed Exons: 37.0% (33.6%–41.4%)	All exons: 28.4% Expressed Exons: 23.6% (22.3%–25.2%)	All exons: 23.7% Expressed Exons: 39.3% (33.3%–44.1%)	
Suren et al. (2016) pine and spruce *Picea glauca x engelmanii* and *Pinus contorta*	26,824 genes (pine) 28,649 genes (spruce)	51% (spruce) and 59% (pine)	18.5% (spruce) and 21% (pine)	37% (spruce) 38% (pine)	44% (spruce) and 41% (pine)	Nonmodel species, large genomes, near target defined as 500 bp
Zhou & Holliday (2012) black cottonwood *Populus trichocarpa*	20.76 Mb (5%) of exons, regulatory regions	86.8% (at 100× coverage about 0%–8%)	~93%	On average, approximately 80 base pairs nearest the bait were sequenced at a depth of >10×	NR	Model species with good genome. Off target defined as >250 bp away.
Hebert, Renaut, and Bernatchez (2013) lake whitefish *Coregonus clupeaformis*	11,975 nuclear exons, and other genomic markers using 62,438 probes	NR	11.8%	NR	NR	98% of targeted genes (2,728) were successfully captured at a mean read depth of 31×
Bi et al. (2012) chipmunk *Tamias alpinus*	11,975 exons	40.3%	25%	NR	NR	% of exons that were covered by at least one read, >99%
Christmas, Biffin, Breed, and Lowe (2017) narrow-leaf hopbush *Dodonaea viscosa* ssp. *angustissima*	700 genes	NR	15.7%	NR	NR	Did not account for intron sites
Syring et al. (2016) whitebark pine *Pinus albicaulis*	7,849 distinct transcripts	NR	13%	NR	NR	
Müller, Freund, Wildhagen, and Schmid (2015) douglas-fir *Pseudotsuga menziesii*	57,110 exons	90%	32%–52% per individual	NR	NR	
Nadeau et al. (2012) butterflies	BAC loci (3.5 MB; 57,610 baits)	75.6%	33.5%	NR	NR	

Note. A summary of sensitivity and specificity of recent exome capture studies in which probes were designed from the same species.
NR: not reported.

conservative sequencing strategy is used (96 samples sequenced over 1.5 lanes of HiSeq; Supporting Information Table S6).

4.1.4 | No dependency on restriction sites

A recently published method, hyRAD-X (Schmid et al., 2017), is similar to EecSeq in that it uses in situ synthesized cDNA probes from expressed mRNA to capture exome sequences. However, the protocol relies on a restriction digest to fragment cDNA and ligate on probes. This may result in a reduced template of probes because not all cDNA fragments will have restriction sites on both ends. To evaluate the possibility that the hyRAD-X would produce a reduced template of probes, we performed crude calculations using SIMRAD in R (Lepais & Weir, 2014) on the *Crassostrea virginica* exome. Of the 31,383 known mRNA transcripts in the oyster genome (assuming 1 transcript variant), 29,555 contain at least 2 MseI cut sites (TTAA). However, there is an SPRI clean-up on the digestion (2×), meaning that at best, only fragments 100 bp and larger are getting through to biotinylation (http://www.keatslab.org/blog/pcrpurificationampureandsimple). SIMRAD estimates 220,184 of a possible 440,881 fragments. Therefore, at the absolute best hyRAD-X is only sampling (29,555/31,383)*(220,184/440,881) = 47% of the exome, although this number may increase slightly due to transcript variations. Relying on restriction digests may also produce skewed size distributions in probes which would be magnified in subsequent rounds of PCR. In Schmid et al. (2017), hyRAD-X generated 524 exome SNPs at a minimum of 6× coverage across 27 samples (compared to the 3,508 exome SNPs discovered at 80× coverage derived from only 8 effective samples in 6 replicate capture using EecSeq), but they were also studying ancient DNA and so whether

the hyRAD-X protocol results in limited coverage across exons remains to be tested.

4.2 | Caveats of EecSeq

Despite the demonstrated benefits of EecSeq, there are some potential caveats that should be considered before employing the method. First, there is no ability to filter out probes that belong to repetitive sequences, which are often present at high concentrations in large-genome organisms such as amphibians (Keinath et al., 2015) or conifers (Amanda et al., 2014). In one capture study from designed probes, a small proportion of the probes (unknowingly at the time of probe development) matched highly repetitive sequences (Syring et al., 2016). This resulted in an inordinate number of reads to these few probe sequences (Syring et al., 2016). However, the inclusion of known repetitive sequence blocker in hybridization, such as c_0t-1 that is used in the EecSeq protocol, has been shown to nearly double capture efficiency (McCartney-Melstad, Mount, & Bradley Shaffer, 2016). In general, repetitive elements, short repeats and low complexity regions are problematic for all types of probe design and capture.

Another caveat of using EecSeq is the need to obtain RNA from relevant samples, although capture designs or gene expression studies based on transcriptomes face the same challenge. Note, however, the advantage that EecSeq probes can be made from mRNA pooled from many individuals, tissues and conditions of interest. If genes of interest are expressed in tissues that are difficult to dissect or are in small abundances (such as neurons), then the RNA-based methods presented here would not be a feasible approach unless pooling multiple extractions. Although the probes are a limited resource, our results indicate that additional rounds of PCR on the probes have little effect on capture.

4.3 | Unique aspects of EecSeq

Our approach relies on expressed mRNA for probe synthesis and the abundance of particular mRNAs will vary depending on gene expression. EecSeq includes a normalization step to decrease the abundance of very common transcripts, but probe pools will still skew towards highly expressed genes, and therefore, capture coverage will be higher for those exons. This aspect of EecSeq can be customized for particular research questions. For projects focused on total exome capture, pools from multiple individuals, tissue types and environmental/laboratory exposures can be constructed to generate a robust probe set. On the contrary, if an investigator is focused on a subset of genes that are responding to a particular stressor, it is possible to make probes from organisms exposed that specific condition and then use those probes to capture other individuals. This reduced probe set may also allow for greater multiplexing, but this remains to be specifically tested. While we have only used mRNA to create probes, there may be possibilities to capture other types transcribed sequences such as long noncoding RNAs or possibly even miRNA.

Previous work on exome capture probe design has focused on intron/exon boundaries. In general, it is thought that capture probes that span exon boundaries will result in low coverage of these regions (Jones & Good, 2016) or that certain regions will not be covered at all (Neves, Davis, Barbazuk, & Kirst, 2013). Inclusion of too many boundaries may also lower overall capture performance by increasing off-target capture (Suren et al., 2016). EecSeq exome probes are derived from mature RNA, so some of the probes will inevitably span exon boundaries. Although exon/intron boundaries cannot be eliminated in EecSeq, both input mRNA and genomic DNA were fragmented down to a modal size of 150 base pairs, with the intention of making both smaller than the average exon size (~273 bp) of Eastern Oysters (note that this size is at the lower limit of what is possible with Illumina sequencing). We found that coverage within exons was fairly uniform, indicating a lack of "edge effects." We hypothesize that the relative long length of EecSeq probes (compared to commercially synthesized probes), the near matching length of genomic DNA fragments and the length distribution relative to actual exon size helped to ensure uniform exon coverage.

We compared our observed measures of sensitivity and specificity to other recently published studies in nonmodel species where probes were designed from bioinformatic resources for the same species. EecSeq capture efficiency performed as well as or outperformed almost all other previously published exome capture studies in nonmodel species (excluding mice and humans; Table 4) with the notable exception of black cottonwood (Zhou & Holliday, 2012), a species with exceptional genomic resources. Note, however, that we analysed capture efficiency across pools of 8 individuals, and there could be considerable variability at the individual level that remains to be quantified.

5 | CONCLUSIONS AND FUTURE DIRECTIONS

Here, we have shown that EecSeq effectively targets expressed exons, delivers consistent and efficient exome enrichment that is comparable to traditional methods of exome capture, and generates thousands of exome-derived SNPS cost effectively. Additional tests are needed to examine the efficiency of exome capture across individuals for different species, which should be coupled with sequencing of EecSeq probes to investigate the effects of probe pool diversity and sequence divergence between probes and targets on capture. Nonetheless, EecSeq holds substantial promise as a universally applicable and cost-effective method of exome sequencing for virtually any macroscopic organism.

ACKNOWLEDGEMENTS

The authors would like to thank Alan Downey-Wall and Sara Schaal for informative discussions and input throughout the duration of this project. The authors also thank Nadir Alvarez for thoughtful comments on a preprint of this manuscript. This work was funded with

funds provided to KEL from Northeastern University and NSF DEB-1635423 and DBI-1722553.

DATA ACCESSIBILITY

Raw, demultiplexed sequences are archived at the NCBI Short Read Archive (BioProject: PRJNA423022). A complete and updated EecSeq protocol can be found at (https://github.com/jpuritz/EecSeq) along with bioinformatic code to repeat all analyses described in this study.

AUTHOR CONTRIBUTIONS

J.P. conceived the original concept of this work and performed all laboratory and data analysis. K.L. contributed all reagents and experimental materials. J.P. and K.L. designed the research, experiments and data analysis, and wrote the manuscript.

ORCID

Jonathan B. Puritz http://orcid.org/0000-0003-1404-4680

REFERENCES

Amanda, R., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., Jones, S. J., … Street, N. (2014). Insights into conifer giga-genomes. *Plant Physiology*, *166*(4), 1724–1732.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*, 403. https://doi.org/10.1186/1471-2164-13-403

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology*, *22*, 3028–3035. https://doi.org/10.1111/mec.12105

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *17*, 362–365. https://doi.org/10.1111/1755-0998.12669

Charlesworth, B., & Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, *118*(1), 2.

Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., … Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, *15*, 449. https://doi.org/10.1186/1471-2164-15-449

Christmas, M. J., Biffin, E., Breed, M. F., & Lowe, A. J. (2017). Targeted capture to assess neutral genomic variation in the narrow-leaf hopbush across a continental biodiversity refugium. *Scientific Reports*, *7*, 41367. https://doi.org/10.1038/srep41367

Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Chen, R., Euskirchen, G., … Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013). Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, *22*, 1383–1399. https://doi.org/10.1111/mec.12182

De Wit, P., Pespeni, M. H., & Palumbi, S. R. (2015). SNP genotyping and population genomics from expressed sequences—current advances and future possibilities. *Molecular Ecology*, *24*, 2310–2323. https://doi.org/10.1111/mec.13165

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, *193*, 929–941. https://doi.org/10.1534/genetics.112.147231

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, arXiv:1207.3907

Hebert, F. O., Renaut, S., & Bernatchez, L. (2013). Targeted sequence capture and resequencing implies a predominant role of regulatory regions in the divergence of a sympatric lake whitefish species pair (*Coregonus clupeaformis*). *Molecular Ecology*, *22*, 4896–4914. https://doi.org/10.1111/mec.12447

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, *188*, 379–397. https://doi.org/10.1086/688018

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*, 185–202. https://doi.org/10.1111/mec.13304

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61. https://doi.org/10.1038/nature10944

Keinath, M. C., Timoshevskiy, V. A., Timoshevskaya, N. Y., Tsonis, P. A., Voss, S. R., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. *Scientific Reports*, *5*, 16413. https://doi.org/10.1038/srep16413

Lee, C. E., Remfert, J. L., Opgenorth, T., Lee, K. M., Stanford, E., Connolly, J. W., … Tomke, S. (2017). Evolutionary responses to crude oil from the Deepwater Horizon oil spill by the copepod *Eurytemora affinis*. *Evolutionary Applications*, *10*, 813–828. https://doi.org/10.1111/eva.12502

Lepais, O., & Weir, J. T. (2014). SimRAD: An R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, *14*, 1314–1321. https://doi.org/10.1111/1755-0998.12273

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*, 1031–1046. https://doi.org/10.1111/mec.13100

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017a). Breaking RAD: An evaluation of the

utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017b). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, *17*, 366–369. https://doi.org/10.1111/1755-0998.12677

McCartney-Melstad, E., Mount, G. G., & Bradley Shaffer, H. (2016). Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, *16*, 1084–1094. https://doi.org/10.1111/1755-0998.12538

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on Breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, *17*, 356–361. https://doi.org/10.1111/1755-0998.12649

Müller, T., Freund, F., Wildhagen, H., & Schmid, K. J. (2015). Targeted resequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection. *Tree Genetics & Genomes*, *11*(1), 816.

Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., … Jiggins, C. D. (2012). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 343–353. https://doi.org/10.1098/rstb.2011.0198

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal: for Cell and Molecular Biology*, *75*, 146–156. https://doi.org/10.1111/tpj.12193

Nosil, P., & Schluter, D. (2011). The genes underlying the process of speciation. *Trends in Ecology & Evolution*, *26*(4), 160–167.

Pastinen, T. (2010). Genome-wide allele-specific analysis: Insights into regulatory variation. *Nature Reviews. Genetics*, *11*, 533–538. https://doi.org/10.1038/nrg2815

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, *16*, 1069–1083. https://doi.org/10.1111/1755-0998.12541

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431. https://doi.org/10.7717/peerj.431

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014b). Demystifying the RAD fad. *Molecular Ecology*, *23*, 5937–5942. https://doi.org/10.1111/mec.12965

Quinlan, A. R. (2014). BEDTools: The Swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics/editorial board, Andreas D. Baxevanis… [et al.]*, *47*, 11.12.1–11.12.34.

Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R., … Whitehead, A. (2016). The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*, *354*, 1305–1308. https://doi.org/10.1126/science.aah4993

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. *Nature Reviews. Genetics*, *15*, 749–763. https://doi.org/10.1038/nrg3803

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(Supp. 1), 9955–9962.

Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., & Alvarez, N. (2017). HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution/British Ecological Society*, *8*(10), 1374–1388.

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., … Alvarez, N. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE*, *11*, e0151651. https://doi.org/10.1371/journal.pone.0151651

Sulonen, A.-M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., … Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biology*, *12*, R94. https://doi.org/10.1186/gb-2011-12-9-r94

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., … Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, *16*, 1136–1146. https://doi.org/10.1111/1755-0998.12570

Syring, J. V., Tennessen, J. A., Jennings, T. N., Wegrzyn, J., Scelfo-Dalbey, C., & Cronn, R. (2016). Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome. *Frontiers in Plant Science*, *7*, 484.

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*, 194–208. https://doi.org/10.1111/1755-0998.12593

Yeaman, S., Hodgins, K. A., Lotterhos, K. E., Suren, H., Nadeau, S., Degner, J. C., … Wang, T. (2016). Convergent local adaptation to climate in distantly related conifers. *Science*, *353*, 1431–1433. https://doi.org/10.1126/science.aaf7812

Zhou, L., & Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, *13*, 703. https://doi.org/10.1186/1471-2164-13-703

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Puritz JB, Lotterhos KE. Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Mol Ecol Resour*. 2018;00:1–14. https://doi.org/10.1111/1755-0998.12905