Gathering genomic-scale data efficiently is challenging for nonmodel species with large, complex genomes. Transcriptome sequencing is accessible for organisms with large genomes, and sequence capture probes can be designed from such mRNA sequences to enrich and sequence exonic regions. Maximizing enrichment efficiency is important to reduce sequencing costs, but relatively few data exist for exon capture experiments in nonmodel organisms with large genomes. Here, we conducted a replicated factorial experiment to explore the effects of several modifications to standard protocols that might increase sequence capture efficiency for amphibians and other taxa with large, complex genomes. Increasing the amounts of c0-t-1 repetitive sequence blocker and individual input DNA used in target enrichment reactions reduced the rates of PCR duplication. This reduction led to an increase in the percentage of unique reads mapping to target sequences, essentially doubling overall efficiency of the target capture from 10.4% to nearly 19.9% and rendering target capture experiments more efficient and affordable. Our results indicate that target capture protocols can be modified to efficiently screen vertebrates with large genomes, including amphibians.

Keywords: amphibian, exon capture, large genome, target enrichment

Received 15 September 2015; revision received 13 April 2016; accepted 6 May 2016
fragments from samples of interest. After hybridization, the biotin on these probes is bound to streptavidin molecules attached to paramagnetic beads, allowing the target sequences to be magnetically captured, and nonhybridized DNA is washed away. Unfortunately, capture of off-target DNA can happen for several reasons and can drastically reduce the efficiency of sequencing (Hodges et al. 2009). Because library fragments are often longer than the probe sequences, part of the hybridized library fragment is usually free to bind to other molecules in the pool. Repetitive DNA sequences are often present at high concentrations in large-genome organisms; if this exposed region is from a repetitive element, it has a high probability of binding to another such fragment, pulling the entire construct through to the final library pool. Adapter sequences are also present at very high concentrations, presenting another opportunity for molecules to bind to captured fragments, creating ‘daisy chains’ of random library molecules (Hodges et al. 2009; Nijman et al. 2010). To mitigate these factors, several ‘blockers’, designed to hybridize to high-copy-number regions early in the protocol and prevent daisy chaining, are typically added to target capture reactions. One such blocker, cdt-1, is a solution of high-copy repetitive DNA fragments that hybridizes with repetitive library fragments and blocks them from attaching to captured fragments. Amphibian genomes contain complex patterns of repetitive elements that may be present at an even higher concentration than normal (Straus 1971; Sun & Mueller 2014; Keinath et al. 2015). To hypothesize that increasing the amount of cdt-1 in solution may improve hybridization efficiency.

Amphibians have large genomes, ranging up to 117 gigabases (Gregory 2002), currently rendering full-genome sequencing approaches untenable, but exon capture is well suited to bridge the gap between single-locus comparative studies and whole-genome analyses for these and other diploid species with large genomes. Several amphibian species have large collections of EST sequences available (Abdulayev et al. 2013; Robertson & Cornman 2014), and transcriptome sequencing and de novo assembly are becoming increasingly accessible for species that lack such resources. Despite the available resources, few exon capture studies have been performed in amphibians (but see Hedtke et al. 2013). This likely reflects limited success of those who have tried and reticence of others to attempt sequence capture approaches with large, highly repetitive genomes.

Beyond the initial purchase price of custom exon probe sets, laboratory costs of exon capture experiments primarily hinge on the efficiency of the enrichment process. Increasing the percentage of reads ‘on target’ (sequence reads that align to regions targeted in the capture array) directly reduces the amount of sequencing required to attain a desired coverage level. Off-target reads may be present for several reasons, including non-specific hybridization of capture probes to off-target regions, hybridization of off-target DNA to the ends of captured target fragments, and failure to wash away DNA not hybridized to capture probes following enrichment (Hodges et al. 2009). The ratio of off-target reads may be particularly problematic in amphibians because their large genome size often reflects a massive increase in the amount of repetitive DNA (Straus 1971; Sun & Mueller 2014; Keinath et al. 2015), which leads to a greatly increased concentration of off-target DNA in solution relative to on-target fragments.

We conducted a series of experiments to optimize existing protocols for exon capture experiments for amphibians and other taxa with large genomes. Our focus is on three different Ambystoma salamanders—the California tiger salamander (Ambystoma californiense), the barred tiger salamander (Ambystoma mavortium) and an F1 hybrid between the two (Ambystoma californiense × mavortium, referred to as F1). Given the enormous size of their genomes, estimated at about 32 gigabases (Keinath et al. 2015), and the observation that they, like many amphibians, have genomes rich in repetitive DNA, we altered the amount of cdt-1 blocker, under the assumption that highly repetitive genomes may benefit from an increased amount of repetitive sequence blocker. We also manipulated the amount of individual input and total DNA in sequence capture reactions to manipulate the total number of copies of the genome, estimating trade-offs among multiplexibility and enrichment efficiency to maximize the number of individuals that can be sequenced for each sequence capture reaction.

Materials and methods

Array design and laboratory methods

We designed an array from 8706 putative exons (8706 distinct genes) using EST sequences from the closely related Mexican axolotl (Ambystoma mexicanum) (Smith et al. 2005). Mitochondrial sequence divergence between the California tiger salamander and the Mexican axolotl is approximately 6.4% (Samuels et al. 2005), and is approximately 1.2% between the barred tiger salamander and Mexican axolotl (Shaffer and McKight 1996), suggesting that less-diverged nuclear exons from the axolotl should serve as appropriate targets for our species. In our design, we attempted to avoid targeting regions that span exon/intron boundaries, as these targets have been found to be much less efficient (Neves et al. 2013). Exon boundaries can be found by mapping EST sequences to a reference genome while allowing for long gaps that represent introns. However, no salamander genome is
currently available, and the two available frog genomes [Xenopus tropicalis (Hellsten et al. 2010) and Nanorana parkeri (Sun et al. 2015)] last shared a common ancestor with salamanders approximately 290 million years ago (San Mauro 2010). To account for this, we developed a comparative method for conservatively predicting intron splice sites within EST sequences (E. McCartney-Melstad & H. B. Shaffer, unpublished data). Target sequences were an average of 290 bp in length (minimum length = 88 bp, maximum = 450 bp, standard deviation = 71 bp), for a total target region length of 2.53 megabases. A total of 39 984 100-bp probe sequences were tiled across these target regions at an average of 1.8× tiling density. These probes were synthesized as biotinylated RNA oligos in a mybaits kit (MYcroarray, Ann Arbor, MI).

We used a salt extraction protocol (Sambrook & Russell 2001) to extract genomic DNA from three individual salamanders: one California tiger salamander (Ambystoma californiense #HBS127160—CTS), one barred tiger salamander (Ambystoma mavortium #HBS127161—BTS) and one F1 hybrid between the two species (#HBS109668). Multiple independent extractions were performed for each individual to attain the amount needed for preparing several libraries. Extractions were then combined into pools and used for library preparations. Two of these pools consisted of pure California tiger salamander DNA or pure F1 DNA and are labelled CTS and F1, respectively. The third pool, which was intended to be pure BTS, was found to consist of approximately 70% barred tiger salamander DNA and 30% California tiger salamander DNA, apparently due to a pooling error, which was later verified through reextraction of the original tissues and Sanger sequencing. We refer to this pool as BTS* and treat it as a third sample in our experimental design. DNA was diluted to 20 ng/µL and sheared to approximately 500 bp on a BioRuptor (Diagenode, Denville, NJ). For each of the 53 individual library preparations (Table S1, Supporting information), we used approximately 450 ng of DNA for library preparations. Standard illumina library preparations (end repair, A-tailing and adapter ligation) were performed using Kapa LTP library preparation kits (Kapa Biosystems, Wilmington, MA). Samples were dual-indexed with 8-bp indices that were added via PCR (adapters from Travis Glenn, University of Georgia). Following library preparation, we performed a double-sided size selection with SPRI beads (Bronner et al. 2009) to attain a fragment size distribution centred around 400 bp and ranging from 200 bp to 1000 bp. Species-specific c9t-1 was prepared using DNA extracted from a California tiger salamander and a single-strand nuclease as follows: First, extracted DNA was treated with RNase and brought to 500 µL at 1000 ng/µL in 1.2× SSC. This DNA was then sheared on a BioRuptor (Diagenode, Denville, NJ) to approximately 300 bp. Next, the solution was denatured at 95 °C for 10 min, partially renatured at 60 °C for 5 min and 45 s, placed on ice for 2 min and then put in a 42 °C incubator. A preheated 250 µL aliquot of S1 nuclease (in buffer) was then added to the partially renatured DNA and incubated for 1 h at 42 °C. The DNA was then precipitated with 75 µL of 3 M sodium acetate and 750 µL isopropanol and centrifuged for 20 min at 10 000 g at 4 °C. Isopropanol was then removed and the pellet was washed with 500 µL of cold 70% ethanol, centrifuged again at 10 000 g for 10 min (4 °C) and dried following ethanol removal. We rehydrated this pellet with 50 µL of 10 m Tris-HCl, pH 8 and dried it down to the appropriate concentration (for 1× c9t-1, 500 ng/µL; for 6× and 12× c9t-1, 1000 ng/µL).

We then multiplexed prepared libraries into capture reactions (Table S1, Supporting information). Total DNA input into the sequence capture was either 500 ng or 1000 ng, and individual library input DNA for multiplexing ranged from 20 to 1000 ng (Table S1, Supporting information). The repetitive DNA blocker c9t-1 was added to the 24 different capture reactions in one of three amounts—2500, 15 000 or 30 000 ng, corresponding to 1×, 6× and 12× protocol recommendation. Libraries were enriched using the mybaits protocol (version 2.3.1), hybridizing probes for 24.5 h and implementing the optional high-stringency washes. Following the three wash steps in the mybaits protocol, we amplified the remaining enriched DNA (with streptavidin beads still in solution) using 14 cycles of PCR. Multiple separate PCR reactions were performed for each capture reaction, which were subsequently pooled after amplification to reduce PCR amplification bias (Barnard et al. 1998).

Postcapture, post-PCR libraries were quantitated and characterized with qPCR using the Kapa Illumina library quantification kit (PicoGreen® Life Technologies, Grand Island, NY and Kapa Biosystems, Wilmington, MA) on a LightCycler 480 (Roche, Basel, Switzerland). We also visualized fragment size distributions using a BioAnalyzer 2100 DNA HS chip (Agilent, Santa Clara, CA). All capture reactions were tested for preliminary evidence of enrichment via qPCR. We developed five primer pairs for different test loci chosen from our targets as positive controls, and one primer pair from a mitochondrial locus we were not targeting as a negative control. We used these to measure the relative concentrations of target molecules in solution by calculating the mean number of cycles required for qPCR reactions to reach the crossing point (Cp) in libraries pre- and post-enrichment. Changes in (Cp) were measured for each test locus for all samples and averaged across all five test loci. For targeted loci, we expected that the number of cycles needed to reach this point would decrease, because target sequences
would be present in higher concentrations. Conversely, we expected the number of cycles for the mitochondrial DNA locus to increase after enrichment, because that sequence was not targeted and we expected its concentration to decrease.

All capture reactions were combined together for sequencing on an Illumina HiSeq 2500 with 150 bp paired-end reads. Reactions were pooled such that all individual libraries would receive at least 1.5 million reads (Table S1, Supporting information). Sample pooling and sequencing was performed at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Genetic data analysis
Demultiplexed reads were checked for adapter contamination and quality trimmed using Trimmomatic 0.32 (Bolger et al. 2014). Quality trimming was performed using several criteria. First, leading base pairs with a phred score < 5 were removed. Next, trailing (3') base pairs with a phred score < 15 were removed. Finally, we used a four base pair sliding window (5' to 3'), trimming all trailing bases when the average phred score within that window dropped below 20. We discarded all reads under 40 bp after trimming, and overlapping reads were merged using fastq-join (Aronesty 2013).

Genetic data from all of the California tiger salamander libraries were combined for assembly to create the most complete possible single-species de novo assembly of our target regions. Targets were de novo assembled using the Assembly by Reduced Complexity (ARC) pipeline (Hunter et al. 2015). This assembly pipeline separates reads that align to target regions and performs small, target-specific de novo assemblies on these read pools. Each assembled contig then replaces its original target sequence, and the process is repeated iteratively. Within ARC, read mapping was performed using Bowtie2 (Langmead & Salzberg 2012), error correction with BayesHammer (Nikolenko et al. 2013), and assemblies were generated using SPAdes (Bankevich et al. 2012).

The ARC pipeline was run for six iterations, which was enough to exhaust all of the reads assignable to most targets.

Following assembly, all contigs were compared against the original target sequences using blastn (Camacho et al. 2009), and reciprocal best blast hits (RBBHs) were found (Rivera et al. 1998). Chimeric assemblies are pervasive and problematic for studies involving de novo assembly of target sequences, because they can insert repetitive sequences into the contigs, making it appear that many reads are mapping to a target when those reads are actually from repetitive regions in the genome (for instance, see the coverage across the example contig in Fig. 1). To attempt to reduce the presence of chimeric assemblies and repetitive sequences in our data, the RBBHs were blasted to themselves (blastn e-value of \(1 \times 10^{-20}\)), and base pairs in sequence regions that positively matched other targets were replaced with N's. These chimera-masked RBBHs served as our final assembled target set.

After assembly, reads from each individual were mapped against the chimera-masked RBBH target set using BWA-MEM (Li 2013). BAM file conversion, sorting, and merging was done using SAMtools v1.0 (Li et al. 2009). PCR duplicates were marked using picard tools v. 1.119 (http://broadinstitute.github.io/picard) which finds fragments that have identical 5' and 3' mapping coordinates, under the assumption that two different chromosomal copies are unlikely to shear in the exact same positions with random sonication. Under this assumption, fragments with identical 5' and 3' mapping coordinates most likely are the result of sequencing multiple amplified copies of the same original DNA molecule, which is undesirable. Finally, mapping rates and PCR duplication rates were inferred by counting the relevant SAM flags using SAMtools flagstat (Li et al. 2009).

In addition to measuring the total percentage of unique reads that mapped to target regions, target-level performance was also evaluated. Because most targets showed a characteristic peak of read depth centred over
the middle of the target where probes were tiled, and because a few targets maintained confounding repetitive sequences at the periphery of the assembled contigs, we characterized the read depths of targets over bases that had direct overlap with our target probes. That is, for target-level metrics, we did not consider read depth for the flanking regions that are naturally appended to the ends of each target during the assembly process. For each individual library preparation, we calculated the average unique read sequencing depth across (i) entire target regions and (ii) across the 100-bp window within each target that had the highest average coverage. For all read depth comparisons, depths were corrected for the total number of reads a library received in sequencing by multiplying by a scaling factor $n_i/n_r$, where $n_i$ is the fewest number of reads received by any individual in the experiment and $n_r$ is the number of reads received by the individual under consideration. Assembled target sequences <100 bp were not included in read depth calculations because 100 bp is significantly less than the average read length and these targets tended to recruit very few reads.

Assessing the importance of c0t-1 and individual input DNA amounts

Linear regression was used to quantify the relationships between c0t-1 and individual input DNA to the percentage of unique reads that mapped to targets. Because three different biological individuals were used for library preparations in this experiment, we also included the identity of the individual as a possible source of variation to explain enrichment efficiency. Models were built that included different combinations of c0t-1, individual input DNA and the identity of the individual (CTS, BTS* or F1) as predictor variables, and unique reads mapping to targets as the response variable. A similar approach was used to model the average sequencing depths across all targets. All models were evaluated by examining the regression coefficients, adjusted $R^2$ and AIC values.

Results

Presequencing library quantitation

DNA concentration yields for post-enrichment, post-PCR samples were lower than anticipated. After 14 PCR cycles, amplified enrichment pools contained an average of 279.5 ng of DNA (after amplifying 15 μL of a total 33 μL in the post-enrichment pools with a 50 μL PCR reaction). One capture reaction (Library #18, see Table S2, Supporting information) had a much higher yield after post-enrichment PCR (2150 ng). Mean $C_p$ in qPCR enrichment verification reactions decreased by an average of 9.1 cycles across the five test loci after enrichment, while the number of cycles required for amplification of a nontargeted negative control locus increased by an average of 2.17 cycles. We found a positive correlation between the mean change in $C_p$ averaged across the five test loci and the raw percentage of reads on target after sequencing for each library (Fig. S1, Supporting information, adjusted $R^2 = 0.1136$, $P = 0.00784$), although the relationship was stronger between post-enrichment, post-PCR DNA concentration and raw mapping rate (Fig. 2, adjusted $R^2 = 0.224$, $P = 0.000204$).

Sequence data

We generated 45 641 469 300 base pairs of sequence data in the form of 150 bp paired-end reads. All libraries received at least 1 207 605 read pairs passing filter (mean = 2 766 149 read pairs, SD = 1 582 161 read pairs). Average base quality phred scores for samples ranged from 33.6 to 34.8 (mean = 34.4, SD = 0.29). An average of 93% of all read pairs both passed the Trimomatic filter, whereas 5.2% of all read pairs had either the forward or reverse read removed, and 1.8% had both members removed. Because our insert size was mostly larger than 300 bp (which is two times the read length), fastq-join did not merge most reads—percentages of
joined reads ranged from 24.0% to 35.1% for the different samples. Nuclear sequence divergence between the Mexican axolotl (the species from which probes were designed) and California tiger salamander in the exon targets averaged 1.84%.

**Reference assembly and read mapping**

A total of 78,674,304 reads (all of the reads from the CTS individual) representing 11,960,279,114 bp were supplied to ARC for de novo assembly of targets. An average of 905 reads in iteration 1, 1496 reads in iteration 2, 1999 reads in iteration 3, 4485 reads in iteration 4, 8132 reads in iteration 5 and 11,199 reads in iteration 6 were assigned to each target for de novo assembly. The final assembly, after six iterations of the ARC assembly pipeline, contained 120,617 sequences for a total of 8,386,921 bp. After blasting the target sequences to the line, contained 120,617 sequences for a total of 2.5 million bp of total target sequence. Among all approximately 2.4 million 150 bp reads against just over 1 blocker were both associated with higher percentages of unique reads on target and higher realized sequence depth across targets (Tables 1 and 2, Fig. 4). Linear regression recovered positive and significant slopes for both variables separately and when combined in multiple linear regression. Models predict an extra 1% of unique reads on target for every 166 ng of extra individual DNA pool underperformed (higher AIC value) nested models that did not incorporate information regarding the identity of the input DNA. Because of this, and because slope coefficients for the identity term in all models were never significant ($P = 0.44$ or greater), the identity of the individual did not significantly improve capture efficiency or read mapping, and models including this variable are not included in the summary tables.

Increasing amounts of individual input DNA and cdt-1 blocker were both associated with higher percentages of unique reads on target and higher realized sequence depth across targets (Tables 1 and 2, Fig. 4). Linear regression recovered positive and significant slopes for both variables separately and when combined in multiple linear regression. Models predict an extra 1% of unique reads on target for every 166 ng of extra individual DNA (hs $P = 0.00672$) or every 6750 ng of extra cdt-1 blocker ($P = 0.00896$) used in enrichment reactions. Regression coefficients for models that contained both
individual input DNA and \( c_{0t-1} \) were quite similar to the single-variable model, differing by <3%. Individual input DNA and \( c_{0t-1} \) did a better job predicting the percentage of unique reads on target than the average depth across target regions (adjusted \( R^2 \) of 0.325 vs. 0.252 for the combined models). Finally, the models that contained both input DNA and \( c_{0t-1} \) as variables had better (lower) AIC scores and higher \( R^2 \) values than the nested single-variable models (see Fig. 5), and within the single-variable tests, individual input DNA models outperformed \( c_{0t-1} \) models for both success measures in AIC and \( R^2 \) (Tables 1 and 2).

**Discussion**

Perhaps the most important conclusion from this experiment is that target capture experiments can indeed be successful in amphibians (and other taxa) with large genomes. This was not at all obvious based on prior work on these organisms, and our hope is that others will use these results to bring amphibians more fully into the

### Table 1 Model comparison predicting percentage of unique reads on target, sorted by AIC values

<table>
<thead>
<tr>
<th>Model</th>
<th>( R^2 )</th>
<th>Adj. ( R^2 )</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_{0t-1} ) + inputDNA***</td>
<td>0.3252</td>
<td>0.2982</td>
<td>-193.6057</td>
</tr>
<tr>
<td>inputDNA***</td>
<td>0.2046</td>
<td>0.189</td>
<td>-186.8963</td>
</tr>
<tr>
<td>( c_{0t-1} )**</td>
<td>0.1265</td>
<td>0.1094</td>
<td>-181.9297</td>
</tr>
</tbody>
</table>

***Signifies \( P < 0.001 \), **signifies 0.001 < \( P < 0.01 \), *signifies 0.01 < \( P < 0.05 \).

### Table 2 Model comparison predicting average depth across target region, sorted by AIC values

<table>
<thead>
<tr>
<th>Model</th>
<th>( R^2 )</th>
<th>Adj. ( R^2 )</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_{0t-1} ) + inputDNA**</td>
<td>0.252</td>
<td>0.222</td>
<td>-269.6817</td>
</tr>
<tr>
<td>inputDNA**</td>
<td>0.1676</td>
<td>0.1513</td>
<td>-273.3437</td>
</tr>
<tr>
<td>( c_{0t-1} )*</td>
<td>0.08887</td>
<td>0.07101</td>
<td>-278.1344</td>
</tr>
</tbody>
</table>

***Signifies \( P < 0.001 \), **signifies 0.001 < \( P < 0.01 \), *signifies 0.01 < \( P < 0.05 \).

Fig. 4 Relationship between individual input DNA and \( c_{0t-1} \) amounts and PCR duplication rates and enrichment efficiency. Each dot is an individual library: square = CTS, triangle = F1, circle = BTS. \( P \)-values for slope coefficients in the four panels are as follows: top left \( P = 1.39 \times 10^{-7} \), top right \( P = 9.28 \times 10^{-6} \), bottom left \( P = 0.000672 \), bottom right \( P = 0.00896 \).
realm of population and phylogenomic analyses. The percentage of unique reads on target is the most important summary metric for enrichment. Our average percentage of unique reads on target across all library treatments was 14%; only three libraries were under 9%, while our four best-performing libraries were all over 20%. These numbers suggest that it is reasonable to sequence 50–100 samples on a single HiSeq lane for a capture array size similar to ours (2.5 megabases), depending on array configuration and coverage requirements, which vary based on the particular application.

Our percentages of unique reads on target are in line with several other nonmodel exon capture studies for species with smaller genomes. For instance, Hedtke et al. (2013) designed Agilent probes from the Xenopus tropicalis genome and enriched libraries from two smaller-genome frogs, achieving rates of 7.4% unique reads on target in Pipa pipa and 47.8% in Xenopus tropicalis. Bi et al. (2012) recovered 25.6% to 29.1% unique reads on target for an exon capture study in chipmunks. Similarly, Cosart et al. (2011) designed an Agilent exon capture microarray from the bovine (Bos taurus) genome and attained 20–29% unique read mapping percentages in Bos taurus, Bos indicus and Bison bison for a similarly sized target array as this study. Finally, Neves et al. (2013) reached 50% raw mapping rates in multiplexed exon capture experiments in Pinus taeda, a pine species with an approximately 21 Gb genome (approximately 2/3 of the size of the salamander genomes in this study), although they did not report percentages of unique reads on target or levels of PCR duplication. Several factors may be important in explaining these results, including a potential negative relationship between the phylogenetic distance to the species from which the capture array was developed and the percentage of unique reads on target, and the size of the genome under investigation. As more target capture studies are reported across diverse nonmodel taxa, we will better understand the relationship between genome size and enrichment efficiency, as well as the effects of designing capture probes from divergent taxa.

Human exome capture studies, which typically use predesigned sequence capture arrays across one of several different technologies (e.g. Truseq, Nimblegen, Agilent or Nextera exome capture kits) often attain percentages of unique reads on target in the range of 40% to 70% or higher (Chilamakuri et al. 2014). However, the high numbers in human experiments are likely a function of many iterations of probe set optimization experiments that have been conducted, which is generally not feasible in non-human systems.

We found evidence that increasing c0t-1 and individual input DNA into sequence capture reactions increased the percentage of unique reads mapping to targets. This effect was driven largely by the correlation of these two variables with the reduction in PCR duplication rates (Fig. 4). Because duplicate reads (reads with the same 5′ and 3′ mapping coordinates) are typically removed prior to genotyping analyses, lowering duplication rates as much as possible is critical for increasing the efficiency, and therefore reducing the sequencing costs of target enrichment studies. Researchers are generally encouraged to use paired-end sequencing whenever possible in exon capture studies, as single-end reads have a much higher false identification rate of PCR duplication (Bainbridge et al. 2010).

The low yields of DNA after enrichment and PCR are interesting. We speculate that they may be a consequence of libraries prepared from large genomes containing relatively low absolute numbers of on-target fragments in the pools during enrichment, so that a higher percentage of the pool is washed away. Comparing the results of qPCR from pre- and post-enrichment libraries using primers meant to amplify targeted regions is a common way to qualitatively assess enrichment efficiency, but we found post-enrichment DNA concentrations to be a better predictor of enrichment success with our protocol (Figs 2 and S1, Supporting information). Also, we note that Library #18, which had a very high post-enrichment post-PCR DNA concentration, showed correspondingly low performance in terms of percentage of raw and unique reads on target (5.4% unique read mapping rate). This suggests that off-target fragments may not have
been efficiently removed during the post-enrichment washing steps in that library.

After duplicate removal, we observed a greater than fivefold difference in unique read mapping percentages (from 5.4% to 30.8%) among the samples tested in this experiment. Given that the targeted region represents approximately 0.008% of the 32 Gb Ambystoma genome (Keinath et al. 2015), this is a striking improvement over whole-genome shotgun sequencing. While even the low end of our enrichment efficiency values are encouraging for future exon capture studies in large-genome amphibians, regularly attaining percentages of unique reads on target at the upper end of our success rate would lead to a concurrent fivefold reduction in sequencing costs for a given target coverage depth.

Based on these experiments, we recommend using at least 30 000 ng of species-specific ct-1 blocker, and as much input DNA as possible for each individual multiplexed into a capture reaction when working with large-genome species. One drawback of this recommendation is that producing this much ct-1 is challenging in species where it is difficult or impossible to extract large amounts of DNA, in which case preparing ct-1 from a closely related species or performing whole-genome amplification on a smaller starting quantity of species-specific ct-1 may suffice. The threshold at which the addition of more blocker DNA ceases to improve (and may potentially inhibit) capture efficiency is not yet known. Additional experiments should attempt to define this limit and should also seek to understand whether additional ct-1 blocker enhances target enrichment in more modestly sized amphibian genomes.

Amounts of individual input DNA are constrained by the total amount of DNA in the capture reaction divided by the number of samples. This means that for a set amount of total input DNA, increasing the number of individuals multiplexed into a single capture reaction will decrease the percentage of unique reads on target. Ongoing research in our laboratory to further optimize target capture for taxa with large genomes is focusing on RNA-seq, which can then be used to design probes. Another important variable is hybridization temperature. Probes with more mismatches hybridize more readily to fragments at lower temperatures, and one promising strategy is touch-down hybridization, where temperatures are sequentially lowered during hybridization (Li et al. 2013). Finally, the tiling density of probes across target sequences is another source of variation in target enrichment experiments. This variable is a function of the total desired target length and the total number of unique probes in the probe set, which varies according to the manufacturer and product purchased. Although information regarding the impacts of this parameter is sparse, Ávila-Arcos et al. (2011) found no clear differences between 5x and 10.9x tiling densities for enriching ancient plant DNA, suggesting that it may not strongly affect enrichment efficiency.

As large-scale sequencing projects become the norm for data acquisition in nonmodel systems, it is crucial to build a body of literature with standard reporting metrics for both laboratory procedures and data filtering and analysis. At a minimum, we suggest researchers report raw mapping rates to target sequences, PCR duplication rates (ideally based on paired-end reads) and average depths across the different targets, including standard deviations, for a given sequencing effort. Standardized metrics will allow researchers to evaluate whether a particular probe set may work in their study system and how much sequencing may be needed. We hope this study can help set a precedent for such reporting on successful laboratory procedures, including a thorough discussion of efficiency and success of target capture in nonmodel organisms.

Acknowledgements

We thank Randal Voss for the Ambystoma mexicanum sequences used to design the capture array and Brant Faircloth for input on experimental design and laboratory troubleshooting. Animal work was conducted under California Department of Fish and Wildlife permit #SC-2480 and associated MOU, USFWS permit #TE-094642-9 and UCLA IACUC protocol #2013-011. This experiment used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation.
EXON CAPTURE OPTIMIZATION WITH LARGE GENOMES 1093

Grants S10RR209668 and S10RR207303. EMM and HBS are supported by NSF-DEB 1257648 and NSF-DEB 1457832.

References


Avisio-Arcos MC, Cappellini E, Romero-Navarro JA et al. (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Scientific Reports, 1, 74.


E.M.M. contributed to the design of the study, performed some of the molecular work, analysed the results and wrote the manuscript. G.G.M. contributed to the design of the study, performed most of the laboratory work and revised the manuscript. H.B.S. contributed to the design of the study and interpretation of results and revised the manuscript. All authors read and approved the final manuscript.

Data accessibility

The data set supporting the results of this article is available at [NCBI SRA:PRJNA285335, SRA:SRP058854]. Accession numbers for individual libraries are shown in Table S1 (Supporting information). The target sequences used for this study, the corresponding Ambystoma mexicanum-derived capture probes and the source code used to analyse the data from this experiment are available at http://dx.doi.org/10.5281/zenodo.18587 (McCartney-Melstad 2015).

Supporting Information

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

Fig. S1. The change in raw mapping rate as a function of post-enrichment qPCR cycle number.

Table S1. Individual libraries (1–53), their treatment levels, and description of yields and sequencing statistics.

Table S2. Post-enrichment concentrations and sequencing efficiency results.