1 **orthoCapture: Facilitating Gene Capture Probe Design for Non-Model Species**

2

3 Running title: Bait capture design for non-model species

4

5 Corresponding author:

6 M. Elise Lauterbur¹

7 Department of Ecology and Evolution

8 Stony Brook University

9 Stony Brook, NY 11794

1 Address as of August 2019: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721

10 **Abstract**

11 In non-model species, targeted gene capture (selective enrichment of specific genomic regions of
12 interest) applications in molecular ecology have been limited by the practicalities of capture design.
13 Currently, the minimal requirement for designing capture probes is a transcriptome, or established
14 reference genome for the species of interest. When an established, annotated reference genome is
15 unavailable, one common approach is to design probes from annotated reference genomes (or
16 transcriptomes) of related species. Unfortunately, as divergence between probes and the genome of
17 interest increases, such as occurs during directional selection, capture performance decreases. Here I
18 introduce orthoCapture, a tool to overcome such limitations by mining unannotated whole-genome
19 sequence (WGS) data from non-model species and/or their close relatives to allow probe design using
20 multiple genomic sources. orthoCapture finds orthologs in WGS data from multiple related species to
21 create a set of exon sequences that encompasses the diversity of the exons of interest. These “design
22 sequences” can then be used to design capture probes for the species of interest. orthoCapture thus
23 eliminates the need for transcriptome or whole-genome sequencing for bait capture experiments,
24 making this technique accessible for molecular ecology and conservation studies. Use of
25 orthoCapture is via command-line interface on Unix systems, and requires the input of a gene
26 sequence from an unrelated annotated genome and a fasta database from a target, unannotated
27 genome (e.g., whole-genome shotgun contigs). The output, sequence templates from the non-
28 annotated genomic data, allows probe creation by any commercial company providing gene capture
29 services.

30

31 **Keywords:** gene capture, orthologs, molecular ecology, conservation genetics, molecular evolution,
32 non-model species

32 Introduction

33 Targeted gene capture (also called targeted enrichment, exon capture, hybrid enrichment, and
 34 sequence capture) has become an important method for facilitating the sequencing of targeted sets of
 35 genes and conserved elements for evolutionary studies. This method allows the selective enrichment
 36 and sequencing of genomic regions of interest by targeting these regions with a set of DNA or RNA
 37 probes prior to sequencing. Gene capture balances cost with the number of genes to be sequenced,
 38 allowing comparisons among many species or individuals when it is not necessary to acquire whole-
 39 genome data (Jones and Good 2016). This is especially relevant since it remains expensive to create
 40 whole-genome assemblies for many individual species in a single study, despite recent advances in
 41 Next-Generation Sequencing technology and associated cost reductions. Thus, targeted gene capture
 42 is especially useful for comparative phylogenetic studies, and it holds potential for broad molecular
 43 ecology and conservation applications.

44

45 Gene capture applications in molecular ecology and conservation have so far been more limited than
 46 in phylogenetics because of the challenge of designing appropriate probes when a high quality,
 47 annotated reference genome does not already exist (Jones and Good 2016). Since the probes (also
 48 called baits) are intended to hybridize with, and thus “capture,” the regions of interest, they must have
 49 nucleotide sequences that complement these regions. When the nucleotide sequence of the regions
 50 of interest is known, for example when an annotated reference genome or transcriptome for the
 51 species exists, probe design is trivial. However, many species of evolutionary and conservation
 52 interest still lack well-annotated reference genomes.

53

54 One common approach is to design probes from annotated reference genomes (or transcriptomes) of
 55 related species (eg. Cosart et al. 2011; George et al. 2011; Nadeau et al. 2012; Ilves and López-

56 Fernández 2014; Bragg et al. 2016; Jones and Gomulkiewicz 2012; Portik, Smith, and Bi 2016).

57 Unfortunately, as divergence between the probes and the genome of interest increases, capture

58 performance decreases, such that probes perform poorly on regions of interest that are more than 5-

59 12% diverged from the genome used to create the probes (Vallender 2011; Bi et al. 2012; Hancock-

60 Hanser et al. 2013; Hedtke et al. 2013). For example, Vallender (2011) found that capture efficiency in

61 non-human primates dropped off when the genome of interest exceeded 4% sequence divergence

62 from the probes. Hancock-Hanser et al. (2013) found that three-quarters of sequencing gaps (areas

63 with no base calls) occurred in regions of less than 88% sequence identity. It is thus preferable to

64 custom-design probe sets for species of interest (Kadlec et al. 2017).

65

66 When well-annotated genomes of the study species are lacking, increasing the diversity of the bait set

67 has been recommended to increase the likelihood of capturing genomic regions of interest (Bi et al.

68 2012; Schott et al. 2017). This is achieved by designing baits from the reference genomes of multiple,

69 related species. With multiple baits tiled across each region of interest, a bait designed from one

70 species may capture the region even if baits from other species do not (figure 1).

71

72 Methods to find and develop bait sets from conserved or ultra-conserved regions for phylogenetic

73 analyses are well-established (Bi et al. 2012; Faircloth et al. 2012; Lemmon, Emme, and Lemmon

74 2012; Faircloth 2017). These vary in both taxonomic scale and the size of the conserved region to be

75 captured, but the focus remains on producing data for phylogenetic studies rather than studies of

76 molecular evolution. Thus, these methods optimize for easily captured targets instead of coding

77 regions of interest. But in studies of molecular evolution, increased divergence between species is

78 expected, and conserved elements are uninformative about genes under selection. The challenge is

79 to find a balance between baits that are sufficiently similar to the regions of interest, and capturing the

80 sequence diversity that can result from differential selection across species (Schott et al. 2017).

81

82 Currently, the minimal requirement for custom-designing bait sets without the reference genome is a
83 transcriptome from the species of interest (Matz 2018; Puritz and Lotterhos 2017). A number of well-
84 tested methods exist to design bait sets from transcriptomes, including Hyb-Seq (Weitemier et al.
85 2014), MarkerMiner (Chamala et al. 2015), and others (Yang and Smith 2014; Kadlec et al. 2017).
86 Many studies have designed custom methods (eg. Ilves and López-Fernández 2014; Mandel et al.
87 2014; Hugall et al. 2016; Teasdale et al. 2016; Schmickl et al. 2016; Schott et al. 2017; Johnson et al.
88 2018) unique to the individual study. Transcriptome-based methods, however, require either an
89 existing transcriptome, or flash-frozen fresh tissue or RNA, and funding to generate a transcriptome.
90 This is frequently infeasible, particularly when the study includes many species without existing
91 transcriptomic resources (thus requiring new transcriptomes for each species), and when the species
92 of interest are elusive or highly endangered, thus sources of RNA are hard to come by (Schott et al.
93 2017). In these cases, it would be preferable to have a bait design method that used available
94 genomic sequences.

95

96 The method required must, first, design baits from existing un-annotated genomic sequences (whole-
97 genome contigs) from the species of interest or, when whole-genome contigs are not available, from
98 multiple related species; second, Use annotations from other species regardless of if they are highly
99 (>10%) diverged from the species of interest; and third, provide a set of sequences that can be
100 curated by standard bait design methods.

101

102 Here I introduce orthoCapture, a tool to facilitate gene capture probe design for non-model organisms.
103 It works by mining un-annotated genomic data from non-model species and/or their close relatives. By
104 mining multiple whole-genome shotgun assemblies it can increase the diversity of the resulting probe
105 set, increasing the likelihood of capturing regions of interest. To my knowledge, this is the only such

106 publicly-available tool that can prepare sequences for probe design of non-conserved regions that
 107 does not rely on transcriptome data for any of its steps. orthoCapture thus eliminates the need for
 108 costly transcriptome creation, or whole-genome sequencing and annotation for bait capture
 109 experiments, making this technique more accessible for molecular ecology and conservation studies.
 110
 111 Use of orthoCapture is via command-line interface using Python 3, and requires the input of a gene
 112 sequence from an unrelated, annotated genome and a fasta database from a target, unannotated
 113 genome (whole-genome shotgun contigs, for example). The output, sequence templates from the non-
 114 annotated genomic data, allows probe creation by any commercial company providing gene capture
 115 services.

116 **Methods**

117 *Description*

118 orthoCapture uses six steps to find target sequences for probe creation (figure 2). Here, “target”
 119 sequence refers to the putative gene sequence from the whole-genome shotgun (WGS) assembly,
 120 whereas “design” sequence is used to refer to the sequence from a well-annotated genome used to
 121 mine the WGS assembly. orthoCapture requires a list of genes to be captured identified by gene
 122 name (e.g., ND1), as well as design sequences of their individual exons based on annotations.

123

124 orthoCapture is intended to run on multiple WGS assemblies to maximize the diversity of the probe
 125 set. The following steps are repeated each time orthoCapture runs:

126

- 127 1. orthoCapture uses tblastx from BLAST (v. 2.29, Altschul 1997) on default settings to search
 128 the WGS assembly for regions similar to the exon design sequences. BLAST v. 2.29 is used in
 129 this step to retrieve nucleotide sequences because subsequent versions of BLAST output
 130 alignments as amino acid sequences rather than the necessary nucleotide sequences.
- 131 2. For each sequence retrieved from the WGS assembly, it then uses blastn from BLAST (v.
 132 2.7.1, Camacho et al. 2009) on default settings to reciprocally match the returned sequences
 133 to the full NCBI nucleotide database, which is a collection of sequences from GenBank,
 134 RefSeq, the Third-Party Annotation (TPA) database, and the Protein Data Bank (PDB). This
 135 step verifies that the returned sequences are similar to sequences previously identified as the
 136 gene of interest.
- 137 4. After retrieving orthologous sequences from the WGS assembly, orthoCapture attempts to
 138 merge partial matches and collapse duplicates. Partial sequences are merged when they

overlap perfectly by at least 30 nucleotides, and perfect matches are collapsed as duplicate records. 30 nucleotides was chosen as the default minimum overlap to minimize the number of incorrectly merged sequences (Magoc and Salzberg 2011). While paired-end read merging programs typically use a minimum overlap of 10 nucleotides (Magoc and Salzberg 2011; Liu et al. 2012; Zhang et al. 2014), this overlap length is still prone to incorrectly merging reads (Magoc and Salzberg 2011) and the longer length of WGS contigs allows a larger minimum to reduce incorrect merges that would result in poor probe design.

5. To facilitate probe design, all target sequences are then oriented as the forward or reverse complement to match the orientation of the original design sequences using seqOrient.pl (Takebayashi 2013).
6. Lastly, all target sequences are checked for minimum length. Any sequences $\leq$ 120 nucleotides long are extended equally on both 5' and 3' ends to reach this minimum probe length.

Evaluation

orthoCapture's reliability at discovering orthologues from WGS assemblies was evaluated by running it on WGS assemblies of species with annotated genomes. A set of 1035 random exons was generated using molbiotools Random Gene Set Generator (<http://www.molbiotools.com/randomgenesetgenerator.html>) and their sequences were retrieved from *Homo sapiens* based on genome annotations using BioMart from ensembl (Zerbino et al. 2018). Sequences for these genes were also retrieved from five species with annotated genomes, *Pan paniscus*, *P. troglodytes*, *Mus musculus*, *Rattus norvegicus*, and *Canis familiaris*. orthoCapture was run on WGS assemblies of each of these five species using *H. sapiens* design sequences of the 1035

162 exons. The orthologs retrieved by orthoCapture were then compared to the exons identified by
163 annotation in the corresponding genomes (figure 3).

164

165 For each species, the target sequences discovered with orthoCapture were then compared to the
166 sequences from the corresponding annotated exons, as well as from the NCBI nucleotide database. If
167 orthoCapture discovered the correct sequence for an exon from a particular WGS assembly, the
168 orthoCapture target sequence should match with 100% sequence identity for at least the length of the
169 exon from the corresponding annotated genome and/or the gene as recorded for that species in the
170 NCBI genome database. An orthoCapture-generated target sequence that is not a perfect match to,
171 and/or shorter than the exon sequence from the annotated genome would indicate that orthoCapture
172 returned a partial or erroneous sequence.

173

174 Target and annotation exon sequences were compared by local alignment and calculation of percent
175 sequence identity and length using exonerate (Slater and Birney 2005). Those exon sequences
176 retrieved by orthoCapture that were shorter than the respective annotated sequences were examined
177 further by comparison with the design sequence. When orthoCapture retrieved target sequences not
178 present in the genome annotation, these sequences were BLASTed against the NCBI nucleotide
179 database (NCBI Resource Coordinators et al. 2018), as well as compared with the design sequence,
180 and considered correct if matching sequences were identified in the database as the gene of interest.

181

182 **Evaluation Results**

183 orthoCapture retrieved sequences for 99-100% as many exons from WGS contigs as were identified
184 by standard annotation protocols (figure 4). The length of the sequence retrieved by orthoCapture
185 varied by species, between $85 \pm 27\%$ (*Mus musculus*, maximum 100%) and $94 \pm 17\%$ (*Pan paniscus*,

186 maximum 100%) (figure 5). Some exons were consistently absent across species in both
187 orthoCapture and annotated results (figure 6), suggesting that some genes were not present in the
188 assembly.

189

190 The false positive rate for orthoCapture (retrieved an erroneous sequence) was less than 1% for all
191 species, and it retrieved at least a partial sequence for all annotated exons.

192 Discussion

193 orthoCapture reliably and accurately retrieves orthologous target sequences from whole-genome
194 shotgun assemblies. It is capable of retrieving nearly the full length of most exons for which there are
195 annotations. However because it is not always able to retrieve the entire exon, it is unsuitable for use
196 as an alternative when a well-annotated genome is available. This tool makes gene capture practical
197 for species lacking a well-annotated reference genome or one from a closely-related species.

198

199 Target capture has been a useful tool for phylogenetic analyses because it is possible to discover and
200 design probes from (ultra-)conserved elements across diverse taxa (Bi et al. 2012; Faircloth et al.
201 2012; Lemmon, Emme, and Lemmon 2012; Faircloth 2017). Conserved elements are inadequate
202 when sequence divergence is targeted, so the application of target capture has so far been limited in
203 molecular ecology studies to species with well-annotated reference genomes or transcriptomes (or
204 close relatives with such).

205

206 Other methods to prepare sequences for target capture probe design include HybSeq (Weitemier et
207 al. 2014), MarkerMiner (Chamala et al. 2015), methods described by Yang and Smith (2014) and
208 Kadlec et al. (2017), and a number of others designed for individual studies (including Ilves and
209 López-Fernández 2014; Mandel et al. 2014; Hugall et al. 2016; Teasdale et al. 2016; Schmickl et al.
210 2016; Schott et al. 2017; Johnson et al. 2018). These methods routinely capture an average of 93% of
211 target loci (Vatanparast et al. 2018). However these methods all require transcriptomes (or expressed
212 mRNA for full exome capture, e.g., Eec-seq (Puritz and Lotterhos 2017)) from the species of interest
213 or a close relative, which restricts their application.

214

215 orthoCapture extends target capture to species without well-annotated reference genomes or
216 transcriptomes, thus expanding its application to taxa with few existing genomic resources that may
217 otherwise be understudied. It does so by taking advantage of broad orthology between species of
218 interest and species with annotated reference genomes, as well as the increased likelihood of
219 successful capture with increased probe diversity (Bi et al. 2012; Schott et al. 2017) (figure 1).

220

221 This increased diversity presents a functional drawback to the application of orthoCapture: higher
222 probe diversity means that more probes must be synthesized per locus, which in turn exacerbates the
223 trade-off between number of loci covered and cost. orthoCapture would be complemented by methods
224 such as BaitFisher (Mayer et al. 2016) or the k-medoid clustering method developed by Johnson et al.
225 (2018), which minimize the number of probes required to capture target loci in multiple species.

226 **System Requirements**

227 orthoCapture runs with a command-line interface and requires Python version 3.5 or later (see manual
228 for required modules), as well as Perl 5.24, Bioperl, and bioperl-run. It also requires BLAST version
229 2.2.29 and BLAST version 2.7.1, which are linked with orthoCapture on github. A set of 1,000 exons,
230 each less than 1,500 bp long, takes approximately 3 days running on 10 nodes with 500 gb RAM. It is
231 recommended to break up large sets of exons into smaller batches (10-100) and run as an array if
232 using a job scheduler (e.g., PBS or SLURM).

233 **Resources**

234 Source code and a detailed manual are available at <https://github.com/lauderbur/orthocapture>.

235 **Acknowledgements**

236 Thanks go to Liliana M. Dávalos and Krishna R. Veeramah for constructive discussions and
237 comments. Development and testing was performed on the Stony Brook University Seawulf HPC
238 cluster and Indiana University HPC Mason and Carbonate clusters. This research was supported, in
239 part, by an American Association of University Women Dissertation Fellowship to M. Elise Lauterbur,
240 and NSF-DEB1456455 to Liliana M. Dávalos.

References

- Altschul, S. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. doi:10.1093/nar/25.17.3389.
- Bi, Ke, Dan Vanderpool, Sonal Singhal, Tyler Linderorth, Craig Moritz, and Jeffrey M Good. 2012. "Transcriptome-Based Exon Capture Enables Highly Cost-Effective Comparative Genomic Data Collection at Moderate Evolutionary Scales." *BMC Genomics* 13 (1): 403. doi:10.1186/1471-2164-13-403.
- Bragg, Jason G., Sally Potter, Ke Bi, and Craig Moritz. 2016. "Exon Capture Phylogenomics: Efficacy across Scales of Divergence." *Molecular Ecology Resources* 16 (5): 1059–68. doi:10.1111/1755-0998.12449.
- Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. doi:10.1186/1471-2105-10-421.
- Chamala, Srikar, Nicolás García, Grant T. Godden, Vivek Krishnakumar, Ingrid E. Jordon-Thaden, Riet De Smet, W. Brad Barbazuk, Douglas E. Soltis, and Pamela S. Soltis. 2015. "MarkerMiner 1.0: A New Application for Phylogenetic Marker Development Using Angiosperm Transcriptomes." *Applications in Plant Sciences* 3 (4): 1400115. doi:10.3732/apps.1400115.
- Cosart, Ted, Albano Beja-Pereira, Shanyuan Chen, Sarah B Ng, Jay Shendure, and Gordon Luikart. 2011. "Exome-Wide DNA Capture and next Generation Sequencing in Domestic and Wild Species." *BMC Genomics* 12 (1). doi:10.1186/1471-2164-12-347.
- Faircloth, Brant C. 2017. "Identifying Conserved Genomic Elements and Designing Universal Bait Sets to Enrich Them." Edited by M. Gilbert. *Methods in Ecology and Evolution* 8 (9): 1103–12. doi:10.1111/2041-210X.12754.
- Faircloth, Brant C., John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, and Travis C. Glenn. 2012. "Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales." *Systematic Biology* 61 (5): 717–26. doi:10.1093/sysbio/sys004.
- George, Renee D., Graham McVicker, Rachel Diederich, Sarah B. Ng, Alexandra P. MacKenzie, Willie J. Swanson, Jay Shendure, and James H. Thomas. 2011. "Trans Genomic Capture and Sequencing of Primate Exomes Reveals New Targets of Positive Selection." *Genome Research*, July. doi:10.1101/gr.121327.111.
- Hancock-Hanser, Brittany L., Amy Frey, Matthew S. Leslie, Peter H. Dutton, Frederick I. Archer, and Phillip A. Morin. 2013. "Targeted Multiplex Next-Generation Sequencing: Advances in Techniques of Mitochondrial and Nuclear DNA Sequencing for Population Genomics." *Molecular Ecology Resources* 13 (2): 254–68. doi:10.1111/1755-0998.12059.
- Hedtke, Shannon M., Matthew J. Morgan, David C. Cannatella, and David M. Hillis. 2013. "Targeted Enrichment: Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time." Edited by Ulrich Joger. *PLoS ONE* 8 (7): e67908. doi:10.1371/journal.pone.0067908.
- Hugall, Andrew F., Timothy D. O'Hara, Sumitha Hunjan, Roger Nilsen, and Adnan Moussalli. 2016. "An Exon-Capture System for the Entire Class Ophiuroidea." *Molecular Biology and Evolution* 33 (1): 281–94. doi:10.1093/molbev/msv216.
- Ilves, Katriina L., and Hernán López-Fernández. 2014. "A Targeted Next-Generation Sequencing Toolkit for Exon-Based Cichlid Phylogenomics." *Molecular Ecology Resources* 14 (4): 802–11. doi:10.1111/1755-0998.12222.
- Johnson, Matthew, Lisa Pokorný, Steven Dodsworth, Laura R Botigue, Robyn S Cowan, Alison Devault, Wolf L Eiserhardt, et al. 2018. "A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using K-Medoids Clustering," July.

doi:10.1101/361618.

Jones, Emily I, and Richard Gomulkiewicz. 2012. "Biotic Interactions, Rapid Evolution, and the Establishment of Introduced Species." *The American Naturalist* 179 (2): E28–36. doi:10.1086/663678.

Jones, Matthew R., and Jeffrey M. Good. 2016. "Targeted Capture in Evolutionary and Ecological Genomics." *Molecular Ecology* 25 (1): 185–202. doi:10.1111/mec.13304.

Kadlec, Malvina, Dirk U. Bellstedt, Nicholas C. Le Maitre, and Michael D. Pirie. 2017. "Targeted NGS for Species Level Phylogenomics: 'Made to Measure' or 'One Size Fits All'?" *PeerJ* 5 (July): e3569. doi:10.7717/peerj.3569.

Lemmon, Alan R., Sandra A. Emme, and Emily Moriarty Lemmon. 2012. "Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics." *Systematic Biology* 61 (5): 727–44. doi:10.1093/sysbio/sys049.

Liu, B., J. Yuan, S.-M. Yiu, Z. Li, Y. Xie, Y. Chen, Y. Shi, et al. 2012. "COPE: An Accurate k-Mer-Based Pair-End Reads Connection Tool to Facilitate Genome Assembly." *Bioinformatics* 28 (22): 2870–74. doi:10.1093/bioinformatics/bts563.

Magoc, T., and S. L. Salzberg. 2011. "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies." *Bioinformatics* 27 (21): 2957–63. doi:10.1093/bioinformatics/btr507.

Mandel, Jennifer R., Rebecca B. Dikow, Vicki A. Funk, Rishi R. Masalia, S. Evan Staton, Alex Kozik, Richard W. Michelmore, Loren H. Rieseberg, and John M. Burke. 2014. "A Target Enrichment Method for Gathering Phylogenetic Information from Hundreds of Loci: An Example from the Compositae." *Applications in Plant Sciences* 2 (2): 1300085. doi:10.3732/apps.1300085.

Matz, Mikhail V. 2018. "Fantastic Beasts and How To Sequence Them: Ecological Genomics for Obscure Model Organisms." *Trends in Genetics* 34 (2): 121–32. doi:10.1016/j.tig.2017.11.002.

Mayer, Christoph, Manuela Sann, Alexander Donath, Martin Meixner, Lars Podsiadlowski, Ralph S. Peters, Malte Petersen, et al. 2016. "BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design." *Molecular Biology and Evolution* 33 (7): 1875–86. doi:10.1093/molbev/msw056.

Nadeau, Nicola J., Annabel Whibley, Robert T. Jones, John W. Davey, Kanchon K. Dasmahapatra, Simon W. Baxter, Michael A. Quail, et al. 2012. "Genomic Islands of Divergence in Hybridizing *Heliconius* Butterflies Identified by Large-Scale Targeted Sequencing." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1587): 343–53. doi:10.1098/rstb.2011.0198.

NCBI Resource Coordinators, Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A Benson, Colleen Bollin, Evan Bolton, et al. 2018. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 46 (D1): D8–13. doi:10.1093/nar/gkx1095.

Portik, Daniel M., Lydia L. Smith, and Ke Bi. 2016. "An Evaluation of Transcriptome-Based Exon Capture for Frog Phylogenomics across Multiple Scales of Divergence (Class: Amphibia, Order: Anura)." *Molecular Ecology Resources* 16 (5): 1069–83. doi:10.1111/1755-0998.12541.

Puritz, Jonathan B, and Kathleen E Lotterhos. 2017. "Expressed Exome Capture Sequencing (EecSeq): A Method for Cost-Effective Exome Sequencing for All Organisms with or without Genomic Resources," December. doi:10.1101/223735.

Schmickl, Roswitha, Aaron Liston, Vojtěch Zeisek, Kenneth Oberlander, Kevin Weitemier, Shannon C. K. Straub, Richard C. Cronn, Léanne L. Dreyer, and Jan Suda. 2016. "Phylogenetic Marker Development for Target Enrichment from Transcriptome and Genome Skim Data: The Pipeline and Its Application in Southern African *Oxalis* (Oxalidaceae)." *Molecular Ecology Resources* 16 (5): 1124–35. doi:10.1111/1755-0998.12487.

Schott, Ryan K, Bhawandeep Panesar, Daren C Card, Matthew Preston, Todd A Castoe, and Belinda S W Chang. 2017. "Targeted Capture of Complete Coding Regions across Divergent Species." *Genome Biol. Evol.*, 17.

337 Slater, Guy St C, and Ewan Birney. 2005. "Automated Generation of Heuristics for Biological
338 Sequence Comparison." *BMC Bioinformatics*, 11.

339 Takebayashi, Naoki. 2013. *SeqOrient.Pl*.
340 <http://raven.wrrb.uaf.edu/~ntakebay/teaching/programming/perl-scripts/seqOrient.pl>.

341 Teasdale, Luisa C., Frank Köhler, Kevin D. Murray, Tim O'Hara, and Adnan Moussalli. 2016.
342 "Identification and Qualification of 500 Nuclear, Single-Copy, Orthologous Genes for the
343 Eupulmonata (Gastropoda) Using Transcriptome Sequencing and Exon Capture." *Molecular*
344 *Ecology Resources* 16 (5): 1107–23. doi:10.1111/1755-0998.12552.

345 Vallender, Eric J. 2011. "Expanding Whole Exome Resequencing into Non-Human Primates."
346 *Genome Biology* 12 (9): R87. doi:10.1186/gb-2011-12-9-r87.

347 Vatanparast, Mohammad, Adrian Powell, Jeff J. Doyle, and Ashley N. Egan. 2018. "Targeting Legume
348 Loci: A Comparison of Three Methods for Target Enrichment Bait Design in Leguminosae
349 Phylogenomics." *Applications in Plant Sciences* 6 (3): e1036. doi:10.1002/aps3.1036.

350 Weitemier, Kevin, Shannon C. K. Straub, Richard C. Cronn, Mark Fishbein, Roswitha Schmickl,
351 Angela McDonnell, and Aaron Liston. 2014. "Hyb-Seq: Combining Target Enrichment and
352 Genome Skimming for Plant Phylogenomics." *Applications in Plant Sciences* 2 (9): 1400042.
353 doi:10.3732/apps.1400042.

354 Yang, Ya, and Stephen A. Smith. 2014. "Orthology Inference in Nonmodel Organisms Using
355 Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for
356 Phylogenomics." *Molecular Biology and Evolution* 31 (11): 3081–92.
357 doi:10.1093/molbev/msu245.

358 Zerbino, Daniel R, Premanand Achuthan, Wasii Akanni, M Ridwan Amode, Daniel Barrell, Jyothish
359 Bhai, Konstantinos Billis, et al. 2018. "Ensembl 2018." *Nucleic Acids Research* 46 (D1): D754–
360 61. doi:10.1093/nar/gkx1098.

361 Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. "PEAR: A Fast and Accurate Illumina Paired-
362 End ReAd MergeR." *Bioinformatics* 30 (5): 614–20. doi:10.1093/bioinformatics/btt593.

Data Accessibility

The data that support the evaluation are available at <https://github.com/lauderbur/orthocapture>. These data were derived from the following resources available in the public domain: BioMart in ensemble at <https://www.ensembl.org/biomart/martview> and the molbiotools Random Gene Set Generator at <http://www.molbiotools.com/randomgenesetgenerator.html>

Author Contributions

M. Elise Lauterbur designed, coded, and evaluated the program, and wrote the paper.

Figures

Figure 1. Increased bait (probe) diversity (resulting from baits designed based on multiple related species) increases the likelihood of successful capture of regions of interest from the study species. Dark blue represents the region of interest in the study species. Light blue and purple indicate sequences in the related species orthologous to the region of interest in the study species. Red indicates regions of divergence from the study species. When baits from two or more related species are included in the capture pool, there is increased likelihood of at least one bait matching the sequence in the study species well enough to capture it.

381
382 Figure 2. Steps of orthoCapture.

383

385 Figure 3. Evaluation steps for orthoCapture. Silhouettes from phylopic.org, *Canis familiaris* by Tracy A.
386 Heath; *Mus musculus* by David Liao (CC BY-SA 3.0); *Pan troglodytes* by T. Michael Keesey
387 (vectorization) and Tony Hisgett (photography) (CC BY-SA 3.0); *P. paniscus* by T. Michael Keesey;
388 *Rattus norvegicus* by Rebecca Groom (CC BY-SA 3.0).

390 Figure 4. Number of exons retrieved by annotation vs orthoCapture for each genome, out of 1035
391 genes with annotated sequences in *Homo sapiens*. Grey bars are exons correctly retrieved by
392 orthoCapture (true positives), stacked black bars are exons incorrectly retrieved by orthoCapture
393 (false positives), and hatched bars are exons identified by annotation. (All exons identified by
394 annotation are assumed to be correct.)

396 Figure 5. Percent of exon recovered by orthoCapture compared to annotated version. Percent
397 recovery is shown only for those exons which were both recovered correctly by orthoCapture and
398 annotated. Annotations were always considered to be correct and complete.

399

400 Figure 6. Heat maps of sequences retrieved by orthoCapture vs. from the annotation for *Homo*
401 *sapiens*-based comparisons. Exons recovered by orthoCapture were considered to be correct if they
402 had > 80% identity match with the annotated version.