Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens

Tomasz Suchan[1‡*], Camille Pitteloud[1‡], Nadezhda S. Gerasimova[2,3], Anna Kostikova[3], Sarah Schmid[1], Nils Arrigo[1], Mila Pajkovic[1], Michał Ronikier[4], Nadir Alvarez[1*]

1 Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland,

2 Biology Faculty, Lomonosov Moscow State University, Moscow, Russia,

3 InsideDNA Ltd., London, United Kingdom,

4 Institute of Botany, Polish Academy of Sciences, Kraków, Poland

‡ These authors are joint co-first authors on this work.

* tomasz.suchan@unil.ch (TS); nadir.alvarez@unil.ch (N. Alvarez)

In the recent years, many protocols aimed at reproducibly sequencing reduced-genome subsets in non-model organisms have been published. Among them, RAD-sequencing is one of the most widely used. It relies on digesting DNA with specific restriction enzymes and performing size selection on the resulting fragments. Despite its acknowledged utility, this method is of limited use with degraded DNA samples, such as those isolated from museum specimens, as these samples are less likely to harbor fragments long enough to comprise two restriction sites making possible ligation of the adapter sequences (in the case of double-digest RAD) or performing size selection of the resulting fragments (in the case of single-digest RAD). Here, we address these limitations by presenting a novel method called hybridization RAD (hyRAD). In this approach, biotinylated RAD fragments, covering a random fraction of the genome, are used as baits for capturing homologous fragments from genomic shotgun sequencing libraries. This simple and cost-effective approach allows sequencing of orthologous loci even from highly degraded DNA samples, opening new avenues of research in the field of museum genomics. Not relying on the restriction site presence, it improves among-sample loci coverage. In a trial study, hyRAD allowed us to obtain a large set of orthologous loci from fresh and museum samples from a non-model butterfly species, with a high proportion of single nucleotide polymorphisms present in all eight analyzed specimens, including 58-year-old museum samples. The utility of the method was further validated using 49 museum and fresh samples of a Palearctic grasshopper species for which the spatial genetic structure was previously assessed using mtDNA amplicons. The application of the method is eventually discussed in a wider context. As it does not rely on the restriction site presence, it is therefore not sensitive to among-sample loci polymorphisms in the restriction sites that usually causes loci dropout. This should enable the application of hyRAD to analyses at broader evolutionary scales.

Introduction

With the advent of next-generation sequencing, conducting genomic-scale studies on non-model species has become a reality [1]. The cost of genome sequencing has substantially dropped over the last decade and repositories now encompass an incredible amount of genomic data, which has opened avenues for the emerging field of ecological genomics. However, when working at the population level—at least in eukaryotes—sequencing whole genomes still lies beyond the capacities of most laboratories, and a number of techniques targeting a subset of the genome have been developed [2, 3]. Among the most popular are approaches relying on hybridization capture of exome [4] or conserved fragments of the genome [5], RNA sequencing (RNAseq [6]), and Restriction-Associated-DNA sequencing (RADseq [7, 8]). The latter has been developed in many different versions, but generally relies on specific enzymatic digestion and further selection of a range of DNA fragment sizes. RAD-sequencing has proved to be a cost- and time-effective method of SNP (single nucleotide polymorphisms) discovery, and currently represents the best tool available to tackle questions in the field of molecular ecology. The wide utility of RAD-sequencing in ecological, phylogenetic and phylogeographic studies is however limited by two main factors: i) the quality of the starting genomic DNA; ii) the degree of divergence among the studied specimens, that translates into DNA sequence polymorphism at the restriction sites targeted by the RAD protocols.

Sequence polymorphism at the DNA restriction site causes a progressive loss of shared restriction sites among diverging clades and results in null alleles for which sequence data cannot be obtained. This limitation critically reduces the number of orthologous loci that can be surveyed across the complete set of analyzed specimens and leads to biased genetic diversity estimates [9-11]. This phenomenon, combined with other technical issues – such as polymerase chain reaction (PCR) competition effects – is a serious limitation of most classic RAD-sequencing protocols that needs to be addressed.

66 In addition, RAD-sequencing protocols rely on relatively high molecular weight DNA
67 (especially for the ddRAD protocol [12]), notably because enzyme digestion and size selection of
68 resulting fragments are the key steps for retrieving sequence data across orthologous loci.
69 Therefore, RAD-sequencing cannot be applied to degraded DNA samples, a limitation also
70 shared by classical genotyping methods and amplicon sequencing. Museum collections,
71 although encompassing samples covering large spatial areas and broad temporal scales, have
72 not necessarily ensured optimal conditions for DNA preservation. As a result, many museum
73 specimens yield highly fragmented DNA – even for relatively recently collected samples [13-
74 15], limiting their use for molecular ecology, conservation genetics, phylogeographic and
75 phylogenetic studies [16, 17]. A cost-effective and widely applied approach for genomic
76 analyses on museum specimens would allow exploring often unique biological collections, e.g.,
77 encompassing rare or now extinct taxa/lineages or organisms occurring ephemerally in natural
78 habitats and posing problems for sampling. It would also allow studying temporal shifts in
79 genetic diversity using historical collections, now applied only in a handful of cases at a genomic
80 scale [18].
81 Hybridization-capture methods have been acknowledged as a promising way to address
82 both the allele representation and DNA quality limitations [19, 20]. Such approaches however
83 usually rely on prior genome/transcriptome knowledge and until recently have been largely
84 confined to model organisms. Addressing this limitation, the recent development of
85 UltraConserved Elements (UCE [3, 5]) or anchored hybrid enrichment [21] capture-based
86 methods allowed targeting homologous loci at broad phylogenetic scales using one set of
87 probes. It however requires a time-consuming design and costly synthesis of the probes for
88 capturing the DNA sequences of interest. Similarly, exon capture techniques, recently applied in
89 the field of museum genomics [18], require fresh specimens for RNA extraction or synthesizing
90 the probes based on the known transcriptome.

91 Here, we present an approach we called 'hybridization RAD' (hyRAD), in which DNA
92 fragments, generated using double digestion RAD protocol (ddRAD [8]) applied to fresh
93 samples, are used as hybridization-capture probes to enrich shotgun libraries in the fragments
94 of interest. Our method thus combines the simplicity and relatively low cost of developing RAD-
95 sequencing libraries with the power and accuracy of hybridization-capture methods. This
96 enables the effective use of low quality DNA and limits the problems caused by sequence
97 polymorphisms at the restriction site. Moreover, utilizing standard ddRAD and shotgun
98 sequencing protocols allows application of the hyRAD protocol in laboratories already utilizing
99 the abovementioned methods, for little cost.
100 In short, the hyRAD approach consists of the following steps (Fig 1):
101 1) generation of a ddRAD library based on high-quality DNA samples, narrow size
102 selection of the resulting fragments and removing adapter sequences;
103 2) biotinylation of the resulting fragments, hereafter called the probes;
104 3) construction of a shotgun sequencing library from DNA samples (either fresh or
105 degraded as in museum specimens);
106 4) hybridization capture of the resulting shotgun libraries on the probes;
107 5) sequencing of enriched shotgun libraries and optionally of the ddRAD library
108 (probes precursor) for further use as a reference;
109 6) bioinformatic treatment (Fig 2): assembly of the reads into contigs, alignment to the
110 sequenced ddRAD library or *de novo* assembly, SNP calling.

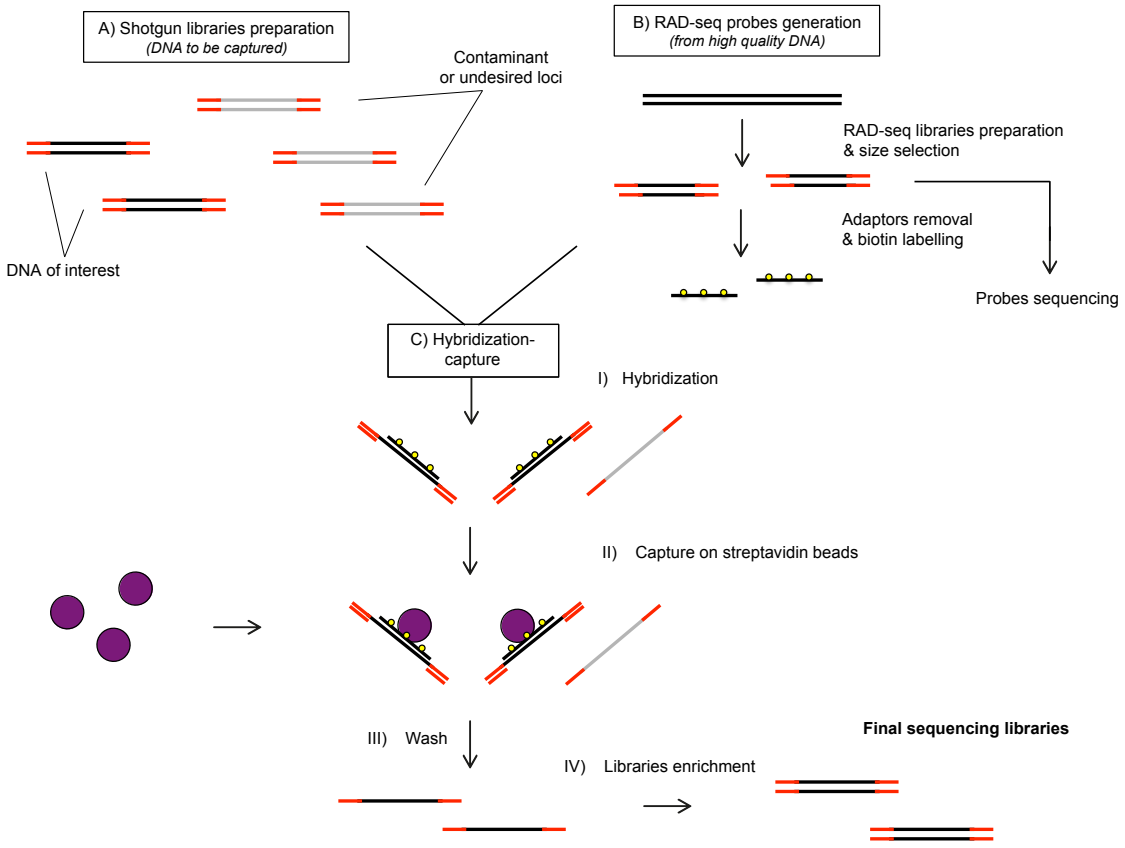


Fig 1. Lab-work procedure used for hyRAD development. Homologous reads from shotgun genomic libraries are captured through hybridization on random RAD-based probes. These fragments are then separated using streptavidin-coated beads and sequenced.

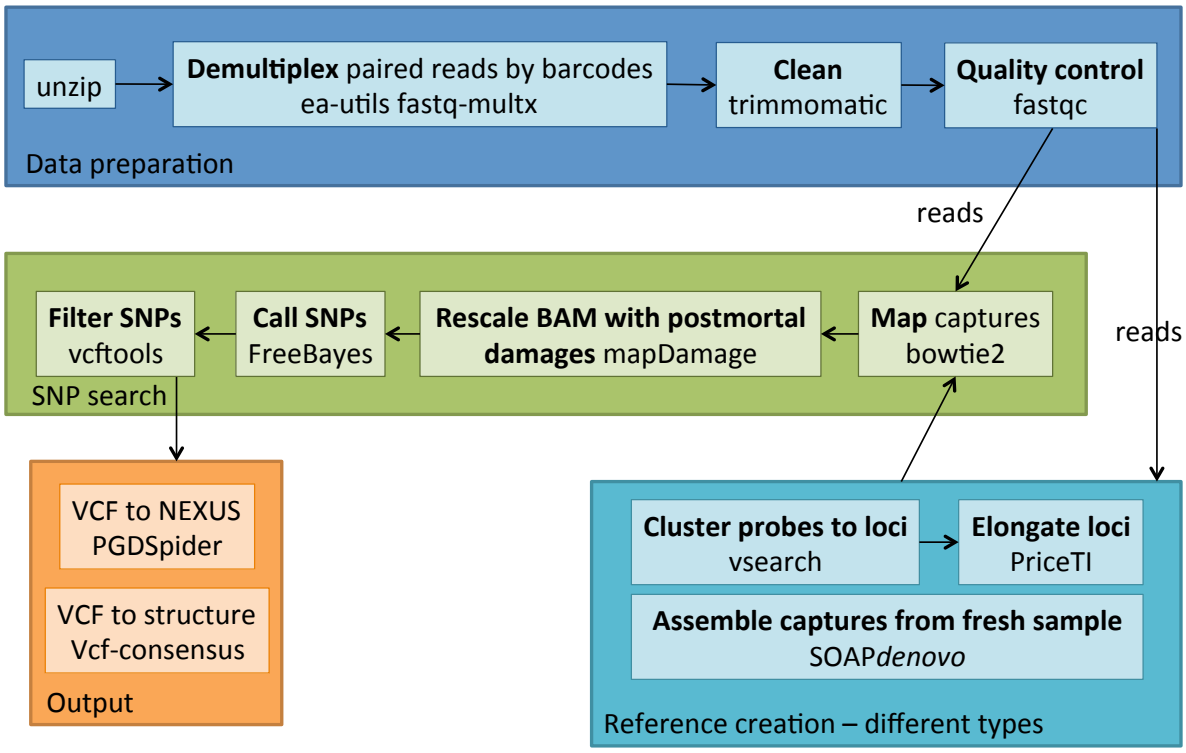


Fig 2. Bioinformatic pipeline used for processing hyRAD sequences. First, the reads are demultiplexed and cleaned. Different types of references were built and the captured fragments were mapped on the reference. The SNPs are then called after correcting for post-mortem DNA damages.

In this paper, we describe the laboratory and bioinformatic pipelines for obtaining hyRAD data, as well as validate the usefulness of the method on two empirical datasets. We first test the method on the DNA obtained from museum and fresh specimens of *Lycaena helle* butterflies. We explore different bioinformatic approaches for assembling loci out of the hybridization-capture libraries, namely: i) mapping captured libraries on previously sequenced RAD loci from fresh samples; ii) using RAD loci as seeds and the captured libraries' reads to extend the RAD loci in order to obtain longer loci for mapping; iii) *de novo* assembly of the

128 captured reads from a single, well-preserved butterfly specimen for the reference. Secondly, as a
129 proof of concept, we apply the protocol to museum and fresh samples of *Oedaleus decorus*, a
130 Palearctic grasshopper species for which a marked east-west spatial genetic structure has been
131 identified in a previous study [22].

132 Materials and Methods

133

134 Study species and study design

135 For the first step of the method development, we used samples of the butterfly *Lycaena helle*
136 (Lepidoptera, Lycaenidae) (Table 1). Three recently collected, ethanol-preserved samples from
137 Romania, France and Kazakhstan were used for generating the RAD probes, in order to cover
138 variation within the full species range. Genomic libraries to be enriched by sequence capture
139 were built using eight samples which included seven museum dry-pinned specimens from
140 Finland (4 collected in 1985 and 3 collected in 1957) and one recently collected and ethanol-
141 preserved specimen from Romania. Using these eight samples we compared the outputs
142 between fresh and historical DNA of different age, and tested the importance of DNA sonication
143 in each case. The museum samples were loaned from the Finnish Museum of Natural History in
144 Helsinki, and the ethanol-preserved samples were obtained from Roger Vila's Butterfly
145 Diversity and Evolution lab (Institute of Evolutionary Biology, CSIC, Barcelona, Spain).

146
147

148 Table 1. *Lycaena helle* samples used in the study.

Type of preservation	Year of collection	DNA concentration [ng/ul]	Locality
dry (pin-mounted)	1957	2.12	Kuusamo, Finland
dry (pin-mounted)	1957	3.02	Kuusamo, Finland
dry (pin-mounted)	1957	1.48	Kuusamo, Finland
dry (pin-mounted)	1985	29.2	Kuusamo, Finland
dry (pin-mounted)	1985	19.7	Kuusamo, Finland
dry (pin-mounted)	1985	17.5	Kuusamo, Finland
dry (pin-mounted)	1985	8.84	Kuusamo, Finland
ethanol	2007	2.12	Dumbrava Vadului, Romania

149

150 For the method validation, we used 53 samples of the grasshopper *Oedaleus decorus*,
151 including 49 samples of both fresh and museum collection specimens for constructing genomic
152 libraries for the capture (see Table 2), and four fresh specimens spanning the species'
153 distribution (Switzerland, Spain, Hungary, Russia) for generating the RAD probes. The museum
154 samples were on average 64-years-old, with the oldest sample dating back to 1908 and were
155 provided by the National History Museum of London (UK), the Natural History Museum of Bern
156 (Switzerland), the Zoological Museum of Lausanne (Switzerland), the ETH Entomological
157 Collection (Switzerland), the Natural History Museum of Basel (Switzerland), the Natural
158 History Museum of Geneva (Switzerland), and the Natural History Museum of Zurich
159 (Switzerland). The four fresh grasshopper samples were provided by G. Heckel (University of
160 Bern).

161

162

163 Table 2. Summary of *Oedaleus decorus* samples used in the study.

Type of preservation	Mean year of collection (range)	Mean DNA concentration [ng/ul] ± SD (range)	Localities
ethanol	2006 (2005-2009)	28.34 ± 28.04 (3.2-105.7)	Croatia, France, Italy, Russia, Spain, Switzerland
dry (pin-mounted)	1952 (1908-1997)	18.31 ± 21.73 (0.3-121.4)	Algeria, France, Greece, Italy, Madeira,, Portugal (mainland and Madeira), Spain (mainland and Canary islands), Switzerland, Turkey

164

165 **DNA extraction**

166 DNA was extracted from insect legs for all the samples. As museum specimens are
167 usually characterized by low-content of degraded DNA, the isolation protocol was optimized
168 accordingly. The samples were extracted using QIAamp DNA Micro kit (Qiagen, Hombrechtikon,
169 Switzerland) in a laboratory dedicated to low-DNA content samples at the University of
170 Lausanne, Switzerland. For these samples, DNA recovery was improved by prolonged sample
171 grinding, overnight incubation in the lysis buffer for 14h and final DNA elution in 20 µl of the
172 buffer with gradual column centrifugation. Extraction of fresh samples was performed using
173 DNeasy Blood & Tissue Kit (Qiagen). DNA extraction and library preparation using museum
174 specimens was performed using consumables dedicated to the museum specimens only.
175 Benches were thoroughly cleaned with bleach and filter tips were used at all stages of lab work.

176 **RAD probes preparation**

177 The probe precursors were prepared using a double-digestion RAD protocol [8, 23],
178 with further modifications.

179 Total genomic DNA was digested at 37°C for 3 hours in a 9 µl reaction, containing 6 µl of
180 DNA, 1x CutSmart buffer (New England Biololabs - NEB, Ipswich, MA, USA), 1 U MseI (NEB) and
181 2 U of SbfI-HF (NEB). The reaction products were purified using AMPure XP (Beckman Coulter,
182 Brea, USA), with a ratio of 2:1 with the sample, according to the manufacturer's instructions,

183 and resuspended in 10 µl of 10 mM Tris buffer. Subsequently, adapters were ligated to the
184 purified restriction-digested DNA in a 20 µl reaction containing 10 µl of the insert, 0.5 µM of
185 RAD-P1 adapter, 0.5 µM of universal RAD-P2 adapter, 1x T4 ligase buffer, and 400 U of T4 DNA
186 ligase (NEB). Adapter sequences are shown in Table 3; single strand adapter oligonucleotides
187 are annealed before use by heating to 95°C and gradual cooling. Ligation was performed at 16°C
188 for 3 hours. The reaction products were purified using an AMPure XP ratio 1:1 with the sample,
189 and resuspended in 10 mM Tris buffer. The ligation product was size-selected using the Pippin
190 Prep electrophoresis platform (Sage Science, Beverly, USA) with a peak at 270 bp and 'tight' size
191 selection range.

192
193 Table 3. Oligonucleotides used in the protocol. x = barcode sequence in the adapters; barcode
194 sequences can be designed using published scripts [24], available at:
195 https://bioinf.eva.mpg.de/multiplex/; I = inosine in the region complementary to the barcode in
196 blocking oligonucleotides sequences.

RAD probes P1 adapters, SbfI-compatible (RAD-P1)	
RAD-P1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxCCTGCA
RAD-P1.2	GGxxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
RAD probes P2 adapter, MseI-compatible (RAD-P2)	
RAD-P2.1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RAD-P2.2	TAAGATCGGAAGAGCGAGAACAA
Shotgun library P1 adapters	
P1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxx
P1.2	xxxxxxAGATCGGAAGAGC

Shotgun library P2 oligonucleotide	
P2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCCCC
PCR primers	
ILLPCR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
ILLPCR2_01	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_02	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_03	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_04	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_05	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_06	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_07	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_08	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_09	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGC
Blocking oligonucleotides	
BO1.P5.F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
BO2.P5.R	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
BO3.P7.F	CAAGCAGAAGACGGCATACGAGATIIIIIIGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BO4.P7.R	AGATCGGAAGAGCACACGTCTGAACTCCAGTCACIIIIIIATCTCGTATGCCGTCTTCTGCTTG

The resulting template was amplified by PCR in a 10 µl mix consisting of 1x Q5 buffer, 0.2 mM of each dNTP, 0.6 µM of each primer (Table 3), and 0.2 U Q5 hot-start polymerase (NEB). The thermocycler program included initial denaturation for 30 sec at 98°C; 30 PCR cycles of 20 sec at 98°C, 30 sec at 60°C, and 40 sec at 72°C; followed by a final extension for 10 min at 72°C. In order to obtain sufficient amount of material for the probes, the reaction had to be run in replicates. The necessary number of replicates has to be determined empirically to reach in total 500-1000 ng of the amplified product required for each capture. DNA amounts

205 were assessed using a Qubit fluorometer (Waltham, MA, USA). Success of size selection and of
206 PCR reactions was confirmed by running them on a Fragment Analyzer (see S1 Fig; Advanced
207 Analytical, Ankeny, IA, USA; see S1 Fig). Afterwards, the PCR products were pooled and purified
208 using AMPure XP, with a ratio of 1:1 with the amplified DNA volume.

209 An aliquot of the resulting library was sequenced and the rest of the library was
210 converted into probes by removing adapter sequences by enzymatic restriction followed by
211 biotinylation. The probe precursors were incubated at 37°C for 3 hours in a 50 µl reaction
212 containing 30 µl of DNA, 1x CutSmart buffer (NEB), 5 U of MseI (NEB) and 10 U of SbfI-HF
213 (NEB), replicated as required by the amount of the amplified product. The reaction was ended
214 with 20 min enzyme inactivation at 65°C for 20 min and the resulting fragments were purified
215 using AMPure XP:reaction volume ratio of 1.5:1. Purified fragments were biotin nick-labelled
216 using BioNick DNA Labeling System (Thermo Fisher, Waltham, MA, USA) according to the
217 supplier's instructions and purified using AMPure XP:reaction volume ratio of 1.5:1. The
218 resulting fragments will thereafter be referred to as probes.

219 **Shotgun library preparation**

220 Shotgun libraries were prepared from the fresh and museum specimens based on a
221 published protocol for degraded DNA samples [15], modified in order to incorporate adapter
222 design of Meyer & Kircher [24]. The approach used for library preparation, utilizing barcoded
223 P1 adapter and 12 indexed P2 PCR primers, allows a high sample multiplexing on a single
224 sequencing lane (see Table 3 [24]).

225 For *L. helle,* DNA from each individual was divided in two aliquots. One aliquot was kept
226 intact (i.e. high molecular weight DNA in the fresh sample and naturally degraded DNA in
227 museum specimens) and the second was sonicated using Covaris focused ultrasonicator
228 (Woburn, MA, USA) with a peak at 300 bp. Both aliquots were processed in parallel during

229 subsequent steps of libraries preparation. All the libraries from *O. decorus* were prepared
230 without sonication, based on the test results obtained from the *L. helle* libraries.
231 DNA samples were first 5'-phosphorylated in order to allow adapter ligation in the next
232 steps of the protocol. 8 µl of DNA was denatured at 95°C for 10 minutes and quickly chilled on
233 ice. The 10 µl reaction consisting of denatured DNA, 1x PNK buffer and 10U of T4 polynucleotide
234 kinase (NEB) was incubated at 37°C for 30min and heat-inactivated at 65°C for 20 min. The
235 DNA was then purified using an AMPure:reaction volume ratio of 2:1 and resuspended in 10 µl
236 of 10 mM Tris buffer.
237 A guanidine tailing reaction of the 3'-terminus was performed after heat denaturation of
238 DNA at 95°C for 10 minutes and quickly chilling on ice. The reaction composed of 1x buffer 4
239 (NEB), 0.25 mM cobalt chloride (NEB), 4 mM GTP (Life Technologies), 10 U TdT (NEB) and 10 µl
240 of denatured DNA in 20 µl reaction volume was incubated at 37°C for 30 min and heat-
241 inactivated at 70°C for 10 min.
242 The second DNA strand was synthesized using Klenow Fragment (3' → 5' exo-), with a
243 primer consisting of the Illumina P2 sequence and a poly-C sequence homologous to the poly-G
244 tail (see Table 3) added to the DNA strand in the previous reaction. A 10 µl reaction mix
245 consisting of 1 µl of NEBuffer 4 (10x), 0.6 µl of dNTP mix (25 mM each), 1 µl of the P2
246 oligonucleotide (15 mM), 5.4 µl of water, and 2 µl of Klenow Fragment (3' → 5' exo-; NEB, 5
247 U/µl) was added to the 20 µl of the TdT reaction mix, incubated at 23°C for 3 h, and heat-
248 inactivated at 75°C for 20 min. The double stranded product was then blunt-ended by adding a
249 mix consisting of 0.5 µl of NEBuffer 4 (10x), 0.35 µl of BSA (10 mg/ml), 0.2 µl of T4 DNA
250 polymerase (NEB, 3 U/µl) and 3.95 µl of water, and incubated at 12°C for 15 min. The resulting
251 product was purified using AMPure XP:reaction ratio of 2:1 and resuspended in 10 µl of 10 mM
252 Tris buffer.
253 Barcoded P1 adapters (see Table 3) were ligated to the 5'-phosphorylated end of the
254 double-stranded product in a 20 µl reaction consisting of 10 µl of the double-stranded DNA, 1 µl

255 of the 25 uM adapters, 1x T4 DNA ligase buffer, and 400 U of T4 DNA ligase (NEB). Adapters
256 have to be annealed before use in the RAD probes protocol. The reaction was incubated at 16°C
257 overnight. The resulting product was purified using an AMPure:reaction ratio of 1:1 and
258 resuspended in 20 µl of 10 mM Tris buffer. Ligated P1 adapters were filled-in in a 40 µl reaction
259 consisting of 20 µl of purified ligation product, 1x ThermoPol reaction buffer (NEB), 12 U of Bst
260 polymerase (NEB), and dNTPs (0.25 mM each), and incubated at 37°C for 20 min.

261 The resulting template was amplified by PCR adding 15 µl of a mix consisting of 5 µl of
262 Q5 reaction buffer (5x), 0.2 µl of dNTPs (25 mM each), 2.5 µl of the PCR primer mix (5 µM each),
263 and 0.5 U of Q5 Hot Start High-Fidelity DNA polymerase (NEB) to the 10 µl of the template. The
264 program started with 20 sec at 98°C, followed by 25 cycles of 10 sec at 98°C, 20 sec at 60°C, and
265 25 sec at 72°C, followed by a final extension for 2 min at 72°C. Success of each PCR reaction was
266 checked using gel electrophoresis, and the resulting products were purified using AMPure
267 XP:reaction ratio of 0.7:1. Samples were then pooled in equimolar ratios.

268 **In solution hybridization capture, library reamplification and sequencing**

269 The hybridization capture and library enrichment steps described below are based on
270 previously published protocols [13, 25] with some modifications. The hybridization mix
271 consisted of 6x SSC, 50 mM EDTA, 1% SDS, 2x Denhardt's solution, 2 µM of each blocking
272 oligonucleotide (to prevent hybridization of adapter sequences; see Table 3), 500 to 1000 ng of
273 the probes and 500 to 1000 ng of the shotgun libraries, in a total volume of 40 µl. On account of
274 grasshopper larger genome size and preliminary results indicating low signal-noise ratio in the
275 butterfly libraries, 500 ng of human Cot-1 DNA (Thermo Fisher Scientific, Switzerland) was
276 added to the *O. decorus* hybridization mix in order to prevent non-specific hybridizations caused
277 by repetitive sequences. The mix was denatured at 95°C for 10 min and subsequently incubated
278 at 65°C for 48 hours. The probes, hybridized with targeted fragments of the library, were then
279 separated on streptavidin beads (Dynabeads M-280, Life Technologies). 10 µl of the beads

280 solution was washed three times on the magnet with 200 µl of TEN buffer (10 mM Tris-HCl 7.5,
281 1 mM EDTA, 1 M NaCl) and resuspended in 200 µl of TEN. 40 µl of the hybridization mix was
282 added to the 200 µl of the beads solution and incubated for 30 min at room temperature. After
283 separating the beads with the magnet, the supernatant was removed and the beads were
284 washed four times under different stringency conditions as follows. The beads were
285 resuspended in 200 µl of 65°C 1x SSC/0.1% SDS wash buffer, incubated for 15 min at 65°C,
286 separated on the magnet and the supernatant was removed. The above step was performed
287 again with 1x SSC/0.1% SDS, followed by 0.5x SSC/0.1% SDS and 0.1x SSC/0.1% SDS. Finally,
288 the hybridization-enriched product was washed-off from the probes by adding 30 µl of 80°C
289 water and incubating at 80°C for 10 min.

290 Enrichment of the captured libraries was performed in a 50 µl PCR reaction containing
291 1x Q5 reaction buffer (NEB), 0.2 mM dNTPs, 0.5 µM of each PCR primer (the P1 universal primer
292 and one of the 12 P2 indexed primers, see Table 3), 1U of Q5 Hot Start High-Fidelity DNA
293 Polymerase (NEB), and 15 µl of the template. The program started with 20 sec initial
294 denaturation at 98°C; followed by 25 PCR cycles of 10 sec at 98°C, 20 sec at 60°C, and 25 sec at
295 72°C; and a final extension for 2 min at 72°C. The enriched-captured libraries were purified
296 using an AMPure XP:reaction ratio 1:1 and pooled in equimolar ratios for sequencing (see S2 Fig
297 for a profile example of the re-amplified capture library after AMPure purification).

298 The probes precursors (RAD library) for the butterfly libraries were sequenced on one
299 lane of Illumina MiSeq 300 bp single-end. Butterfly capture-enriched libraries were sequenced
300 on one lane of MiSeq 150 bp paired-end, and grasshopper capture-enriched libraries were
301 sequenced on one lane of Illumina HiSeq 100 bp paired-end.

302 Updated versions of the lab protocol can be found at
303 https://github.com/chiasto/hyRAD.
304

305 **Data analysis**

306 The hyRAD datasets correspond to target-enriched libraries and cannot be analysed with the
307 usual RAD pipelines [26, 27] Indeed, although they were generated using RAD loci, the obtained
308 sequences are not flanked by the restriction sites and instead may not overlap completely
309 and/or extend before and after the RAD locus. As a result, the analysis pipeline must include the
310 following steps:

311 1) demultiplexing and cleaning of raw reads;
312 2) building of reference sequences for each RAD locus;
313 3) alignment of reads against the obtained references;
314 4) SNP calling.

315 All bioinformatic steps of the hyRAD pipeline can be run at https://insidedna.me.

316 **Demultiplexing and data preparation**

317 The obtained reads were demultiplexed using the fastx_barcode_splitter tool from the
318 FASTX-Toolkit package [28]. RAD-seq sequences (probe precursors) were processed by Trim
319 Galore! [29] and cleaned with the fastq-mcf tool from ea-utils package [30] to remove low
320 quality nucleotides and adapter sequences. The PCR duplicates were removed from RAD-seq
321 probes precursors and hyRAD datasets using the MarkDuplicates tool of Picard toolkit [31].
322 Reads from hyRAD libraries were tested for exogeneous DNA contamination using BLAST
323 against NCBI nucleotide database (50,000 reads for sonicated or non-sonicated fresh or
324 museum DNA samples).

325 **Exploring the methods of reference creation on *Lycaena helle* libraries**

326 Paired-end reads obtained from the hybridization-capture library for each sample were
327 mapped onto three references: (1) consensus sequences for the clustered RAD-seq reads (RAD-
328 ref), (2) RAD-seq reads extended using hybridization-captured reads (RAD-ref-ext), and (3)
329 contigs assembled from the reads of hybridization-captured samples (assembly-ref). As we had

330 no reference genome available, we focused on checking the numbers of loci/SNPs obtained by
331 mapping the reads on each reference and the overlap of the loci obtained using different
332 methods. Theoretically, we should be able to retrieve all the loci from the sequenced probe
333 precursors (RAD-seq library) in the captured libraries. We thus evaluate the number of
334 captured loci homologous to those retrieved by sequencing the RAD-seq libraries.However,
335 different factors affect signal to noise ratio (i.e. the proportion of fragments homologous with
336 the probe) in the captured libraries and can decrease the numbers of homologous fragments
337 retrieved.

338 **Vsearch RAD loci clustering (RAD-ref).** High quality reads of RAD probes were
339 clustered by similarity using Vsearch [32] to obtain loci for further mapping the reads from the
340 hybridization-capture libraries. Before Vsearch run, we converted cleaned fastq files into fasta
341 format using the fasta_to_fastq tool from the the FASTX-Toolkit package [28]. To obtain the most
342 reliable contigs across samples, Vsearch was run in two iterations. During the first iteration, we
343 obtained consensus clusters at the within-individual level (i.e. clustering of the raw reads for
344 each sample independently). During the second iteration, Vsearch was run on the consensus
345 clusters obtained from the first iteration. The second iteration allowed us to obtain consensus
346 clusters at the among-individual level. For both iterations we ran Vsearch with various identity
347 thresholds (0.51, 0.61, 0.71, 0.81, 0.83, 0.91, 0.93, 0.96, 0.98 for the within-individual level and
348 0.51, 0.61, 0.71, 0.81, 0.85, 0.88 for the among-individual level) in order to identify an optimal
349 identity threshold for clustering, i.e. a threshold that maximizes the number of clusters with a
350 minimal coverage of 2x and 3x, respectively. In all cases, we used the cluster_fast option for
351 clustering. The consensus sequences of each secondary cluster were then used as locus
352 references in subsequent alignment and SNP calling steps.

353 **Vsearch RAD loci clustering and extension using captured reads (RAD-ref-ext).** To
354 obtain RAD-ref-ext, we iteratively extended contigs of the RAD-ref using reads from the
355 hybridization-capture library (by pooling reads contributed by all the analysed specimens)

356 using PriceTI [33] with 30 cycles of extension and a minimum overlap of a sequence match to
357 30. The obtained references were trimmed by 60 bp at each end in order to remove sequences
358 with putatively low-quality ends. We applied this tough threshold for RAD-ref-ext only, as
359 probes extension can be performed on very low-coverage data, and we therefore wanted to
360 keep the error rate (usually higher on both sequence ends) at the minimum.

361 **De-novo assembly from captured reads only (assembly-ref).** Assembly was
362 performed on the hybridization-captured reads of one good quality ethanol-preserved,
363 sonicated, sample. Only sequences obtained from the single fresh sample were used, as
364 stringent cleaning parameters in Trimmomatic [34], used for the reference construction, led to a
365 large loss (up to 80%) of the reads from historical samples (and such data was therefore less
366 optimal than that from the fresh specimen for producing a reference). Moreover, using one
367 individual allows obtaining a more reliable reference, as any among-sample divergence can
368 result in bubbles in the contigs and bias the assembly by oversplitting alleles. As a result we
369 could have obtained chimeric duplicated loci presented in different contigs. We used
370 SOAP*denovo* V2.04 to assemble cleaned reads into contigs [35].

371 **Mapping and SNP calling**

372 Reads from the hyRAD library were cleaned by Trimmomatic with milder parameters
373 than reads used in the *de novo* assembly (keeping 60-85% of the reads in historical specimens).
374 Read mapping was performed using bowtie2-build for the reference indexing and bowtie2 for
375 mapping [36], PCR duplicates were removed using the MarkDuplicates tool from the Picard
376 toolkit [31] and SNPs were called with FreeBayes using the default parameters [37, 38]. To
377 evaluate the level of DNA damage in museum DNA samples, we used mapDamage2.0 that
378 rescales base quality scores of putatively post-mortem damaged bases [39] in order to minimize
379 presence of post-mortem conversions in the resulting SNPs. Datasets for replicates were
380 merged and analysed for SNPs with FreeBayes. Obtained VCF files were filtered by the vcffilter

381 tool from the vcflib [40] and VCFtools [41]. In the resulting set only biallelic sites with high
382 quality (PHRED>30), minor allele count larger than 1/6 of all, present in at least 50% of the
383 samples and with a minimum depth of 6 were kept, indels were removed. Potential paralogs and
384 multi-copy sites were removed based on a coverage of a standard-deviation three times higher
385 than the mean [42]. VCF format files [43] were converted to SNP-based NEXUS files using
386 PGDSpider converter [44] and to structure data files for every individual using the vcf-
387 consensus tool (see also Fig 2). Updated versions of the bioinformatic pipeline can be found at
388 https://github.com/chiasto/hyRAD.

389 **Genetic structure**

390 In order to check whether data was reflecting genetic structure, we applied
391 fastStructure, a Structure-like algorithm adapted to large SNP genotype data [45] to the six final
392 datasets (RAD-ref, RAD-ref-ext and assembly-ref, both for samples with and without
393 sonication). The analyses were conducted using a *simple prior* and assuming two groups (*k*=2).

394 **Overlap between assembly references**

395 To evaluate the level of overlap among the three assembly references for *L. helle* (RAD-
396 ref, RAD-ref-ext, and assembly-ref) we used the OrthoMCL [46] pipeline for orthology detection.
397 Most of the pipeline was run with the default parameters, except for Blastall and MCL clustering
398 steps. Here, we used more stringent parameter values (e-value of 0.0001 and MCL was run with
399 an inflation parameter of 2.0) in order to reduce chances of detecting false orthology groups. As
400 a result, we obtained clusters of contigs being contributed by the three assembly references. We
401 then counted how many of these clusters – presumably corresponding to homologous loci –
402 were shared among the available reference assembly approaches. Eventually, to reveal the
403 number of RAD loci present in the references, reads of the raw RAD library were mapped on
404 RAD-ref and assembly-ref using bowtie2 [36] and levels of mapping were compared.

405 **Proof of concept: application of hyRAD to *Oedaleus decorus***

406 The utility of the method was further validated using 49 museum and fresh samples of a
407 Palearctic grasshopper species for which a marked east-west spatial genetic structure was
408 identified previously [22]. The catalog was built based on eight specimens from the captured
409 library that showed the largest number of reads and spanned the species' distribution area
410 (Switzerland, Italy, Spain, Russia), using the method that yielded the highest number of contigs
411 and produced consistent genetic structure in *L. helle* (assembly-ref, i.e., *denovo* reference built
412 using SOAP*denovo*, see Results and Discussion section and Table 4). The generated contigs were
413 blasted against GenBank databases for bacterial, fungal and technical sequences with a
414 minimum E-value threshold of 0.1. Endogeneous contigs of each samples were assembled to
415 generate the final reference using Geneious V9.0.2 [47]. Filters were subsequently applied to
416 keep only high quality and informative SNPs as for the *L. helle* libraries. The final VCF matrix
417 was converted into Structure format using PGDSpider V2.0.9.0 [44] and population structure
418 was inferred using fastStructure [45] using a *simple prior* and assuming two groups (*k*=2).
419 Quantum GIS V2.4.0 was used to present the geographic distribution of the genetic clusters [48].
420

421 Table 4. Data on obtained references for *L. helle* (RAD-ref, RAD-ref-ext, assembly-ref) and *O.*
422 *decorus* (assembly-ref).

Reference	Number of contigs	Largest contig (bp)	Total length (bp)	N50
RAD-ref (*L. helle*)	25 478	544	5 445 942	209
RAD-ref-ext (*L. helle*)	24 820	851	2 613 024	98
assembly-ref (*L. helle*)	304 161	2 352	35 579 979	666
assembly-ref (*O. decorus*)	408 851	13 103	119 789 911	321

423

424 ## Results and Discussion

425

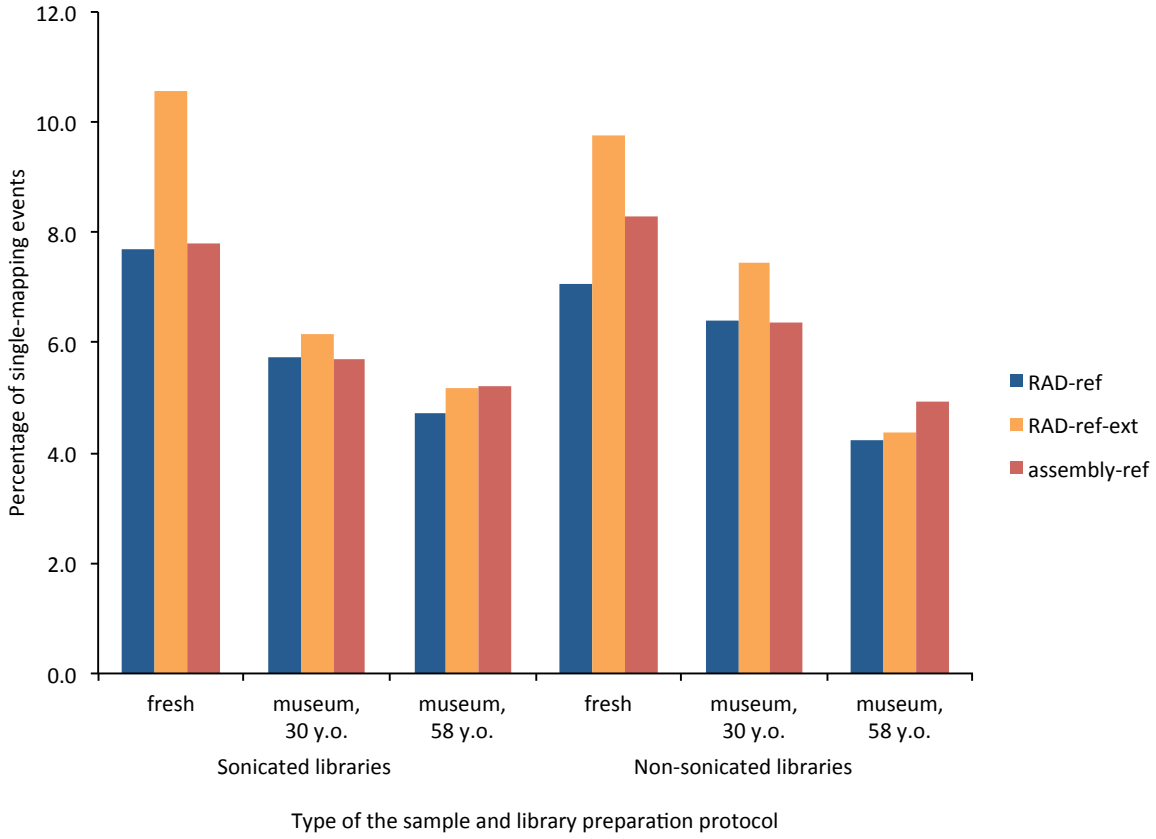
426 ### Sequencing and data quality

427 ***Lycaena helle* libraries.** RAD-seq libraries sequencing yielded 14,188,023 and
428 hybridization-capture libraries 16,636,502 raw reads: 8,217,522 for sonicated and 8,418,980
429 for non-sonicated samples. Additional sequencing of the reference library from the ethanol-
430 preserved specimen (used for the RAD-ref creation) yielded 4,703,744 reads. The proportion of
431 reads kept after quality filtering varied with sample age and preparation method. For the
432 ethanol-preserved sample, 89.8% of reads from sonicated and 89.4% from non-sonicated
433 sample were retained. For the 30 years old samples the mean was 74.3% and 79.6%, and for 58
434 years old samples 70.8% and 73.8% for sonicated and non-sonicated samples, respectively.
435 ***Oedaleus decorus* libraries.** Hybridization-capture libraries yielded a total of
436 69,306,042 raw reads. After quality filtering, 80.3% of reads were kept among all the samples.

437 **Comparison of references obtained for *L. helle* libraries**

438 **Proportion of single-hit alignments and SNP numbers.** Consensus clustering of the
439 RAD-based reference within individuals produced the largest number of clusters with 2x and 3x
440 coverage with a clustering identity threshold of 0.91, compared to other threshold values.
441 Consensus clustering among individuals produced the best results with a clustering identity
442 threshold of 0.71 (see S3 Fig).
443 The highest number, length and the total length of reference contigs were obtained
444 using *de novo* assembly with SOAP*denovo* (assembly-ref; Table 4). Both RAD-based assemblies
445 produced an order of magnitude lower number of contigs. The extension performed on the
446 obtained RAD reference followed by trimming of adapter sequences resulted in references with
447 an average shorter length (lower N50) than the starting contigs—whereas priceTI extended a
448 large number of probes, this did not reflect in a substantially higher average loci length because
449 of further trimming of obtained contigs.
450 The highest levels of single-hit alignments for most of the samples, except the oldest
451 ones, for both preparation methods (sonicated and not sonicated) were obtained when mapped

452 on the sequenced RAD loci extended using PriceTI (RAD-ref-ext). This method was followed by
453 *de novo* assembly using reads from the hybridization-capture library from a single fresh
454 specimen (assembly-ref) and mapping on the RAD loci (RAD-ref); although the difference
455 between the last two approaches was not large (Fig 3). Only for the oldest samples as well as in
456 the non-sonicated fresh sample, *de novo* reference provided slightly better results.



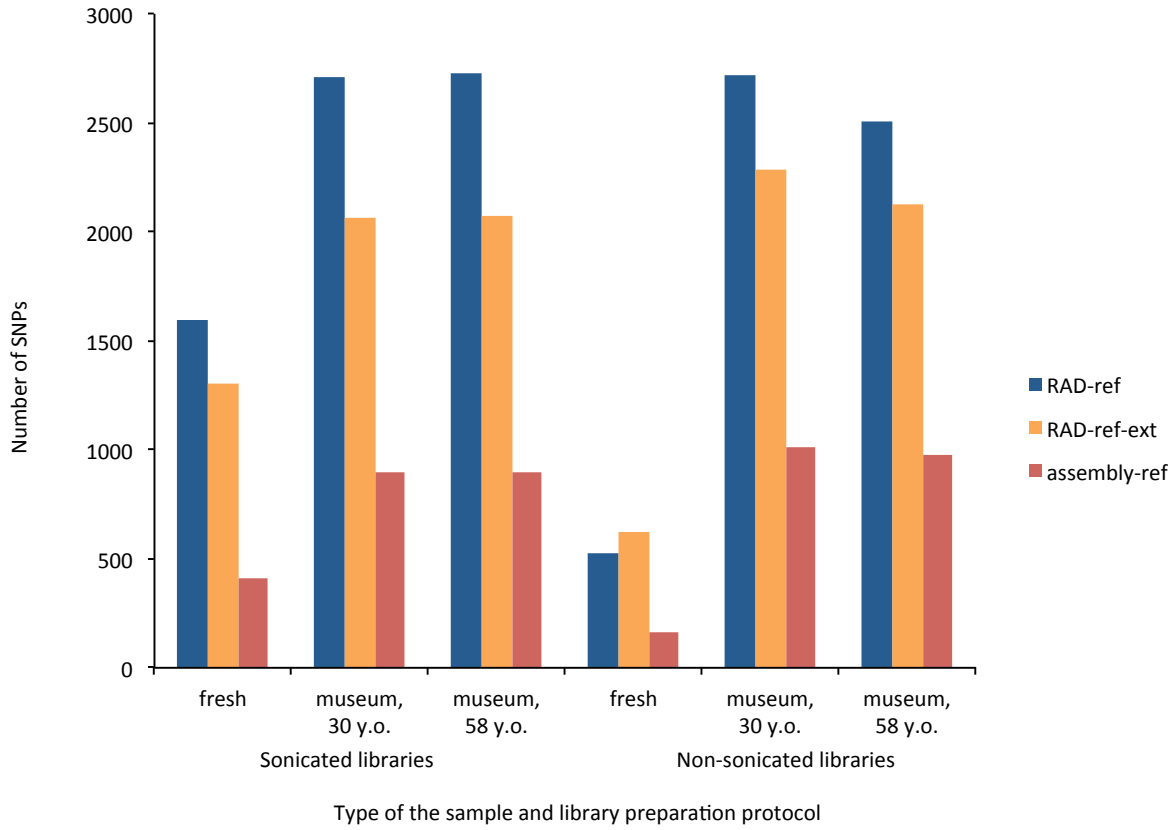
457

458 Fig 3. Percentage of the captured reads showing unique mapping events for different
459 types of DNA preparations and bioinformatic pipelines.

460

461 In terms of the number of SNPs retained after coverage, paralogs and among-samples
462 overlap filtering, the RAD-ref pipeline detected the highest numbers of loci, regardless of sample
463 age and preparation (Fig 4). No clear correlation with sample age could be observed, although
464 all the methods provided the lowest SNP numbers in the fresh samples – most likely an effect of

465 small genetic distance between the reference sample and the fresh specimens (as the fresh
466 sample was used both for the reference and the aligned sample, only heterozygote sites account
467 for SNPs here). A higher number of SNPs was detected in the sonicated library only for the fresh
468 specimens.



469

470 Fig 4. Mean number of SNPs per sample obtained for different types of DNA
471 preparations and bioinformatic pipelines.

472

473 RAD-sequencing datasets depend on the presence of the restriction sites and therefore
474 any polymorphism in such sites leads to either missing loci or alleles. As our method does not
475 depend on the restriction site presence, combined with the high number of gathered SNPs, this
476 allows obtaining largely filled data matrices. Matrix fullness was >50% in all cases:

477 - assembly-ref, non-sonicated: 66.1%

478 - assembly-ref, sonicated: 63.4%

479 - RAD-ref, non-sonicated: 52.0%

480 - RAD-ref, sonicated: 69.1%

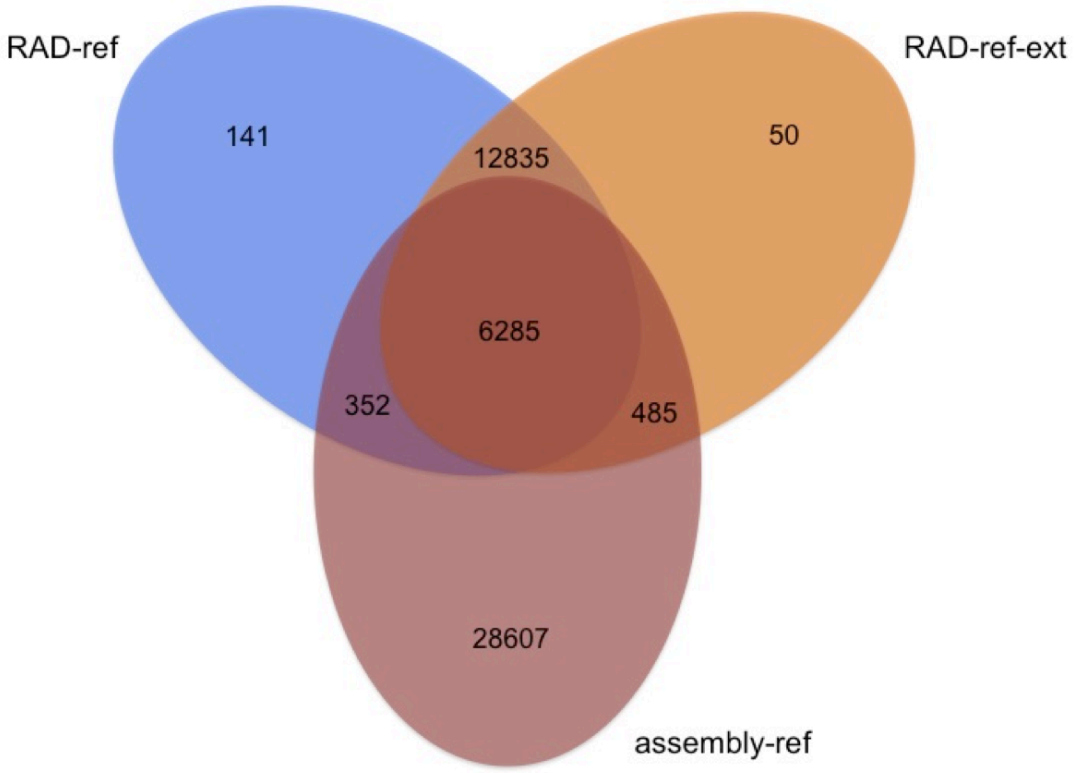
481 - RAD-ref-ext, non-sonicated: 63.4%

482 - RAD-ref-ext, sonicated: 72.5%

483 Lacking an objective criterion for assessing the 'best' performing method for building
484 the catalog, we tested which of the three references created the dataset producing the expected
485 spatial division of genetic structure between Finnish and Romanian *L. helle* samples. The spatial
486 genetic structure inferred by fastStructure revealed that the eight samples are divided into two
487 clusters of, respectively, seven Finnish *vs.* one Romanian sample only when using the non-
488 sonicated library mapped to the assembly-ref catalog. This result is in agreement with the
489 hypothesis that when using sonicated DNA, we are at high risk of incorporating contaminant
490 DNA which can blur the signal (Matthias Meyer, Max Planck Institute for Evolutionary
491 Anthropology, Liepzig, DE ; personal communication). This is an interesting result as BLAST
492 analyses on the six catalogs did not retrieve differences in the level of known contaminants (see
493 below).

494 **Loci overlap among the assembly methods.** About 15.2% of the reference loci
495 obtained via *de novo* assembly were shared with those based on the clustering of RAD-seq reads
496 (either RAD-ref or RAD-ref-ext; Fig 5). In contrast, an appreciable fraction of the obtained
497 reference loci (59.7%) were unique to the assembly-ref approach. When mapping raw RAD
498 reads on RAD-ref, 26.3% did not map, 42.4% mapped once and 31.3% mapped more than once.
499 This shows that RAD-ref summarizes ca. ¾ of the reads from the RAD dataset. In contrast, when
500 mapping raw RAD reads on assembly-ref, 78,3% of reads did not map, 21,6% mapped once and
501 0,1% mapped more than once. This result shows that assembly-ref possibly contains three
502 quarters of all loci that are not homologous to the RAD probes. Such low signal to noise ratio
503 (targeted reads to the number of total reads) is most likely a result of background carryover in

504 the hybridization capture step, a phenomenon which can have many sources., It could result
505 from 'daisy-chaining' of the captured fragments [49, 50], where partially complementary DNA
506 molecules hybridize with the other fragments that are already hybridized to the probes. We can
507 however discard this explanation as a primary reason for the background carryover as
508 extending of RAD probes did not produce longer contigs (in RAD-ref-ext assemblies). Another
509 likely reason could be carryover of random DNA fragments with repetitive sequences. The
510 extent of such process can be significantly reduced by adding blocking agents to the
511 hybridization mix (typically Cot-1 as was performed for the *O.decorus* dataset [51] or salmon
512 sperm DNA), as well as optimizing hybridization temperature and wash stringency to increase
513 capture efficiency. Such developments are desirable because they increase the percentage of
514 reads matching the loci of interest and eventually improve the overall sequencing coverage.
515



516
517 Fig 5. Number of loci obtained using different bioinformatic approaches, identified using
518 the OrthoMCL [42] pipeline for orthology detection.

519

520 **Effects of sample preparation and age on the numbers of SNPs obtained and the**
521 **exogeneous DNA content**

522 Reads can be mapped on a reference either once with a highest score (i.e., single
523 mapping) or on more than one region of reference with close scores (i.e., multi mapping). The
524 reasons for multi-mapping events can be biological (e.g., paralog sequences) or technical
525 (splitting single loci into more reference loci), nevertheless these mapping events cannot be
526 used for SNP calling and offer another benchmark for the assembly methods used. Differences in
527 the number of single mapping events and in the numbers of SNPs obtained were not substantial
528 between sonicated and non-sonicated samples, and depended on the sample age and the
529 bioinformatic pipeline used (Figs 3 and 4). We expected that museum specimens should
530 perform better without sonication, as the DNA was already visibly fragmented, and sonication of
531 museum specimens may increase the levels of exogenous DNA contamination (by fragmenting
532 intact fungal or bacterial DNA contaminating museum samples; M. Meyer, pers. comm.). In
533 terms of mapping events, whereas the fresh sample usually showed a better ratio of single- to
534 multi-mapping events when sonicated, there was a trend towards a higher percentage of unique
535 mapping events for non-sonicated 30-years old museum specimens, on average 1.3 times
536 higher, irrespective of the reference used (the difference was less clear for 58-years old museum
537 specimens, and depended on the reference used). We would therefore advise not to sonicate
538 DNA obtained from the museum specimens, which significantly cuts down the price and time
539 required for library preparation, except in cases when no signs of degradation are observable
540 on the DNA profile. As levels of DNA degradation of contemporary samples may vary, one may
541 consider that the sonication step should be advisable when working with well-preserved DNA.
542 However, this is still an open question, as whereas BLAST searches did not retrieve higher
543 fractions of contaminants in sonicated *vs*. non-sonicated libraries, the expected population
544 structure was retrieved was the non-sonicated one mapped on assembly-ref.

545 As one of the main types of post-mortem DNA degradation is deamination of cytosines,
546 highly damaged ancient or museum DNA samples are usually characterized by higher uracil
547 content [52-54]. In classical library preparation protocols, the usage of a proofreading
548 polymerases should stall the chain elongation in the presence of uracil and thus reduce the
549 misincorporation errors in the final dataset. On the other hand this approach might not be
550 optimal for highly degraded ancient DNA samples, where a large fraction of DNA fragments may
551 carry cytosine to uracil misincorporations [53]. Moreover the usage of a proofreading
552 polymerase does not prevent the misincorporations caused by direct deamination of
553 methylated cytosine to thymine, or less common deamination of guanidine to adenine [52, 54].
554 In the protocol used above [15], the second strand synthesis was performed using Klenow
555 Fragment (3' → 5' exo-), lacking proofreading ability, and thus approximately half of the
556 resulting DNA fragments should have cytosine to uracil misincorporation substituted for
557 thymine, amplifiable by proofreading DNA polymerase. We thus opted for a bioinformatic post-
558 processing way of filtering-out such bases. Post-mortem damage in the sequenced samples was
559 assessed by mapDamage2.0, which rescales sequence files by downscaling quality scores of
560 likely post-mortem damaged bases. As some SNPs became filtered by lower quality scores after
561 the rescaling, the number of SNPs is decreased after mapDamage2.0. We expected higher
562 number of discarded SNPs in the oldest samples, because of a higher proportion of DNA damage
563 occurring with time. The proportion of SNPs discarded after applying mapDamage2.0 was the
564 highest among the 58 years old samples (1.46% for RAD-ref; 1.89% for RAD-ref-ext; 2.36% for
565 assembly-ref), although relatively low, given the samples' age and preservation type (see S1
566 Table).
567 It is worth mentioning, that the higher number of SNPs detected in libraries from
568 museum specimens, comparing to the fresh samples, is not an effect of post-mortem damage
569 (an opposite trend was detected with the higest proportion of type II transitions to transitions
570 [55] and transversions present in the fresh specimens; S4 Fig)

571 **Application: spatial genetic structure of *Oedaleus decorus***

572 The *de novo* reference catalog was composed of 408,851 reference contigs. The N50
573 length was 321 bp and the total length was 119,789,911 bp. Among the total number of contigs,
574 9% were shown to be of exogeneous origin by the BLAST search, either against fungi and
575 bacteria GenBank databases or against technical sequences. Such a level of contaminants is
576 expected here, as in contrast to the *L. helle* references, which were built from fresh samples, the
577 *O. decorus* assembly-ref was based on eight specimens from the captured library —either fresh
578 or pin-mounted—that showed the largest number of reads. A total of 4,783,774 informative
579 sites were retrieved after SNP calling. After removing indels and low quality sites, 125,890 sites
580 were conserved. Keeping only biallelic loci with a minor allele count of at least 6, with data
581 fullness higher or equal to 50% of the samples, we obtained 6,046 loci. Finally, we conserved
582 2,979 SNPs after the removal of potential paralogous sites. The median depth for each SNP was
583 10. On average, each of the 49 samples were characterized by 1864 SNPs and each SNP was
584 found in 32 individuals (62.7% of matrix fullness).

585 The spatial genetic structure inferred by fastStructure revealed two geographically
586 distinct clades in the west and the east of the Palearctic (Fig 6). This result supports the eastern-
587 western split previously highlighted in *Oedaleus decorus* based on mtDNA amplicons [22]. This
588 demonstrates that hyRAD is a reliable technique to infer spatial genetic structure from both
589 fresh samples and museum samples collected at various time points in the past.

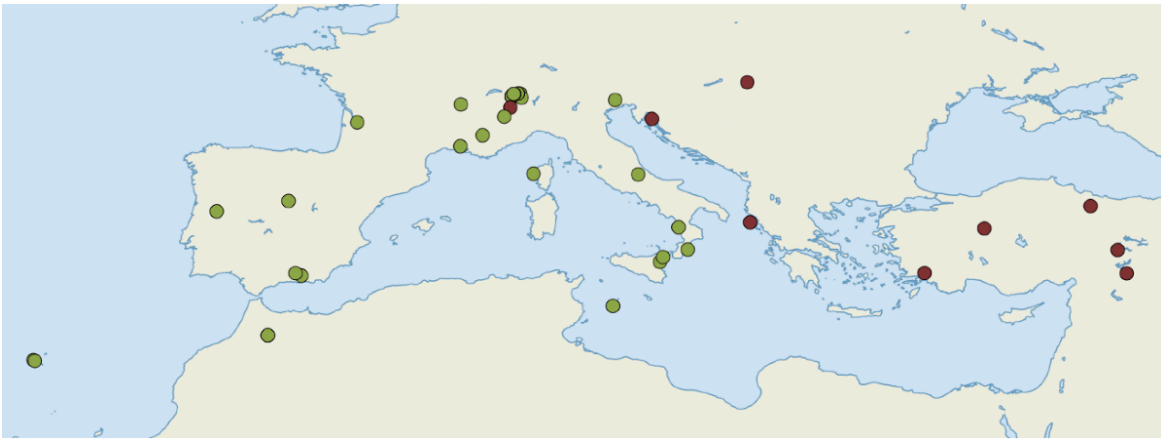


Fig 6. Spatial genetic structure of *O. decorus* inferred using fastStructure with *k*=2 and simple priors. Colours denote the two different genetic groups supported by a previous study relying on mtDNA markers [22].

Conclusions

Here, we present a method for obtaining large sets of homologous loci from museum specimens, without any a priori genome information. Despite the differences in single-mapping events among samples of different ages were retrieved, the obtained numbers of SNPs were not significantly different for the two age classes of the *L. helle* museum specimens. We obtained around a thousand of SNPs from *L. helle* samples up to 58 years old, confirming that it can be successfully applied in the field of museum genomics. The application of the catalog-building method that was the most promising for resolving population genetic structure (i.e., non-sonicated library mapped to assembly-ref) to the grasshopper *O. decorus*, confirmed the usefulness of hyRAD to retrieve phylogeographic data using museum samples up to one hundred years old. Our method does not require time-consuming and costly probes design and synthesis, nor access to fresh samples for RNA extraction, making it one of the simplest and most straightforward technique for obtaining orthologous loci from degraded museum samples.

609 In the protocol, we applied a modified shotgun library preparation method, optimized
610 for degraded DNA from museum specimens [15]. However, the capture protocol presented here
611 can be applied to any type of library preparation, including commercial ones, simplifying the
612 workflow and cutting down the preparation time.
613 We also explored several bioinformatic approaches for loci assembly from the captured
614 libraries, a crucial step when working on organisms without a reference genome. Identifying the
615 most appropriate catalog-building method may depend on the goals of each study. In our case,
616 the pipeline that was the best at identifying population structure in the butterfly was relying on
617 a non-sonicated library mapped to the *de novo* reference assembly from captured reads from a
618 single ethanol-preserved specimen, using SOAP*denovo* assembler (assembly-ref). Despite the
619 fact that a maximum of 26% of the obtained sequences mapped to the references and the
620 proportion of single mapping events were not higher than 10% on average (Fig 3), we could
621 successfully call around a thousand of loci in each case (Fig 4), with high coverage across the
622 samples.
623 Importantly from the wetlab protocol perspective, in the hybridization step, we have
624 used blocking oligonucleotides to prevent 'daisy-chaining' of captured sequences by adapter
625 sequences' homology. Using a blocking agent preventing similar chaining caused by repetitive
626 sequences (Cot-1 DNA, applied to the grasshopper libraries, see below [51]) as well as
627 optimizing the conditions of hybridization and capture reactions for increased stringency (e.g.,
628 by decreasing hybridization temperature and the stringency of the washes using higher
629 concentration of SDS and/or lower concentration of SSC) may further increase the
630 hybridization efficiency and thus the numbers of reads mapping on the reference and reduce
631 the number of low-coverage loci.
632 The *de novo* assembly building pipeline produced the largest contig of 2,352 bp for the
633 butterfly and 13,103 bp for the grasshopper dataset. Although the mean length of the assembled
634 contigs was much smaller, our method also allows retrieving longer sequences than the length

635 of the probes used. The reason for this is that captured sequences hybridize with other DNA
636 fragments with homologous sequences, flanking the probe sequence (i.e., 'daisy-chaining' [49,
637 50]). This may lead to enrichment across larger fractions of genome, a side effect of our method,
638 that can be utilized for assembly of larger contigs by using longer probes and capturing longer
639 targets.

640 The method presented here, although based on the restriction enzyme digestion of DNA
641 to create the random genomic probes, does not depend on the restriction site presence in the
642 captured library. This represents a significant improvement over classical RAD-sequencing
643 datasets, in which increase in the phylogenetic distance among samples is correlated with an
644 increase in the number of missing sites [56-61], sometimes leading to conflicting signals
645 between RAD- and capture-based datasets [62], or are characterized by the presence of null
646 alleles that lead to heterozygosity or F_{ST} underestimation [9, 10]. In this aspect, our approach is
647 similar to other capture-enrichment protocols, such as UltraConserved Elements [5] or exome-
648 capture [4], with the benefit of much simpler and less expensive probe generation, without
649 access to genome information or fresh specimens for RNA isolation. Not relying on the presence
650 of restriction site, the method presented here should be also useful for broader phylogenetic
651 scales, allowing sequencing homologous loci from more divergent taxa, which would not be
652 possible to retrieve using classical RAD-seq approaches.

653

654 **Acknowledgments**

655

656 We thank Alan Brelsford, Alicia Mastretta-Yanes and Pawel Rosikiewicz for their help
657 with developing RAD-sequencing protocols. Jairo Patiño tested early versions of the protocol
658 and provided valuable feedback. Roger Vila and Gerald Heckel kindly provided fresh samples for
659 the study. We thank the following museum curators for providing collection samples: Hannes
660 Baur (Natural History Museum, Bern, Switzerland), Anne Freitag (Zoological Museum,

Lausanne, Switzerland), Rod Eastwood (ETH Entomological Collection, Zurich, Switzerland), Daniel Burckhardt (Natural History Museum, Basel, Switzerland), Peter Schwendinger (Natural History Museum, Geneva, Switzerland), Barabara Oberholzer (Natural History Museum, Zurich, Switzerland), George Beccaloni (Natural History Museum, London, UK), Lauri Kaila (Finnish Museum of Natural History, Helsinki, Finland). We also thank Brent Emerson for his constant support during the development of this method.

669 **References**

670

671 1. Ellegren H. Genome sequencing and population genomics in non-model organisms.
672 Trends in Ecology & Evolution 2014;29: 51-63.
673 2. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide
674 genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews
675 Genetics 2011;12: 499-510.
676 3. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-
677 generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and
678 Evolution 2013;66: 526-538.
679 4. Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, et al. Comparison of
680 solution-based exome capture methods for next generation sequencing. Genome Biology
681 2011;12: R94.
682 5. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC.
683 Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary
684 timescales. Systematic Biology 2012;61: 717-726. doi:10.1093/sysbio/sys004.
685 6. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in
686 expressed exons of the human genome using RNA-Seq. Nucleic Acids Research 2009;37: e106.
687 7. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP
688 discovery and genetic mapping using sequenced RAD markers. PLoS ONE 2008;3: e3376.
689 doi:10.1371/journal.pone.0003376.
690 8. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an
691 inexpensive method for de novo SNP discovery and genotyping in model and non-model
692 species. PLoS ONE 2012;7: e37135.

693 9. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. RADseq underestimates diversity and
694 introduces genealogical biases due to nonrandom haplotype sampling. Molecular Ecology
695 2013;22: 3179–3190.
696 10. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, et al. The effect of
697 RAD allele dropout on the estimation of genetic variation within and between populations.
698 Molecular Ecology 2013;22: 3165-3178. doi: 10.1111/mec.12089.
699 11. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. Special features
700 of RAD Sequencing data: implications for genotyping. Molecular Ecology 2013;22(11), 3151-
701 3164. doi: 10.1111/mec.12084.
702 12. Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE. Demystifying the RAD
703 fad. Molecular Ecology 2014;23: 5937–5942. doi: 10.1111/mec.12965.
704 13. Mason VC, Li G, Helgen KM, Murphy WJ. Efficient cross-species capture hybridization
705 and next-generation sequencing of mitochondrial genomes from noninvasively sampled
706 museum specimens. Genome Research 2011;21: 1695-1704.
707 14. Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, et al.
708 DNA Damage in Plant Herbarium Tissue. PLoS ONE 2011;6: e28448. doi:
709 10.1371/journal.pone.0028448.
710 15. Tin MM-Y, Economo EP, Mikheyev AS. Sequencing degraded DNA from non-
711 destructively sampled museum specimens for RAD-tagging and low-coverage shotgun
712 phylogenetics. PLoS ONE 2014;9: e96793. doi:10.1371/journal.pone.0096793.
713 16. Wandeler P, Hoeck PE, Keller LF. Back to the future: museum specimens in
714 population genetics. Trends in Ecology & Evolution 2007;22: 634-642.
715 17. Rowe KC, Singhal S, MacManes MD, Ayroles JF, Morelli TL, Rubidge EM, et al.
716 Museum genomics: low-cost and high-accuracy genetic data from historical specimens.
717 Molecular Ecology Resources 2011;11: 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x.

718 18. Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. Unlocking the vault:
719 next generation museum population genomics. Molecular Ecology 2013;22: 6018–6032.
720 19. Jones MR, Good JM. Targeted capture in evolutionary and ecological genomics.
721 Molecular Ecology 2015. doi: 10.1111/mec.13304.
722 20. Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and
723 epigenomes. Nature Reviews Genetics 2015;16: 395-408. doi: 10.1038/nrg3935.
724 21. Lemmon A R, Emme SA, Lemmon EM. Anchored hybrid enrichment for massively
725 high-throughput phylogenomics. Systematic Biology 2012;sys049.
726 22. Kindler E, Arlettaz R, Heckel G. Deep phylogeographic divergence and cytonuclear
727 discordance in the grasshopper Oedaleus decorus. Molecular phylogenetics and evolution
728 2012;65(2), 695-704.
729 23. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC.
730 Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly
731 optimization for population genetic inference. Molecular Ecology Resources 2015;15: 28-41.
732 doi: 10.1111/1755-0998.12291.
733 24. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed
734 target capture and sequencing. Cold Spring Harbor Protocols 2010;2010: t5448. doi:
735 10.1101/pdb.prot5448.
736 25. OpenWetWare contributors 'Hyb Seq Prep'. OpenWetWare 2015;
737 http://openwetware.org/index.php?title=Hyb_Seq_Prep&oldid=553025.
738 26. Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W. Stacks: an analysis tool set
739 for population genomics. Molecular Ecology 2013; 22: 3124-3140.
740 27. Eaton DA. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
741 Bioinformatics 2014;30: 844-1849. doi:10.1093/bioinformatics/btu121.
742 28. FASTX-Toolkit. 2015. Database: GitHub [Internet]. Available:
743 https://github.com/agordon/fastx_toolkit.

744 29. Krueger F. Trim Galore: A wrapper tool around Cutadapt and FastQC to consistently
745 apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-
746 digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. 2015. Available:
747 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
748 30. Aronesty E. ea-utils: command-line tools for processing biological sequencing data;
749 2011. Database: Google Code [Internet] Available: http://code.google.com/p/ea-utils
750 31. Picard tools. 2015. Database: GitHub [Internet]. Available:
751 http://broadinstitute.github.io/picard/.
752 32. Flouri T, Ijaz UZ, Mahé F, Nichols B, Quince C, Rognes T. VSEARCH GitHub repository.
753 Release 1.0.16; 2015. Database: GitHub [Internet]. Available:
754 https://github.com/torognes/vsearch. doi: 10.5281/zenodo.15524.
755 33. Ruby JG, Bellare P, DeRisi JL. PRICE: software for the targeted assembly of
756 components of (meta) genomic sequence data. G3: Genes, Genomes, Genetics 2013;3: 865-880.
757 34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
758 sequence data. Bioinformatics 2014;30: 2114-2120. doi:10.1093/bioinformatics/btu170.
759 35. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically
760 improved memory-efficient short-read de novo assembler. GigaScience 2012;1(1):18. doi:
761 10.1186/2047-217X-1-18.
762 36. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods
763 2012;9: 357-359.
764 37. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing.
765 arXiv 2012;arXiv:1207.3907.
766 38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
767 alignment/map (SAM) format and SAMtools. Bioinformatics 2009;25: 2078-2079.

768 39. Jónsson H, Ginolhac A, Schubert M, Johnson P, Orlando L. mapDamage2.0: fast
769 approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 2013;29:
770 1682-1684. doi: 10.1093/bioinformatics/btt193.
771 40. Garrison E. vcflib. 2015. Database: GitHub [Internet]. Available:
772 https://github.com/ekg/vcflib.
773 41. Auton A, Danecek P, Marcketta A. VCFtools. 2015. Database: GitHub [Internet].
774 Available: https://vcftools.github.io/.
775 42. Puritz JB, Hollenbeck CM, Gold JR. dDocent: a RADseq, variant-calling pipeline
776 designed for population genomics of non-model organisms. PeerJ 2014;10.7717/peerj.431.
777 43. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
778 call format and VCFtools. Bioinformatics 2011;27: 2156-2158.
779 44. Lischer HEL, Excoffier L. PGDSpider: An automated data conversion tool for
780 connecting population genetics and genomics programs. Bioinformatics 2012;28: 298-299.
781 45. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of Population
782 Structure in Large SNP Data Sets. Genetics 2014;197:573-589; doi:
783 10.1534/genetics.114.164350.
784 46. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for
785 eukaryotic genomes. Genome Research 2003;13: 2178-2189.
786 47. Kearse M, Moi R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious
787 Basic: an integrated and extendable desktop software platform for the organization and analysis
788 of sequence data. Bioinformatics 2012;28(12),1647-1649;doi: 10.1093/bioinformatics/bts199.
789 48. QGIS Development Team. QGIS Geographic Information System. Open Source
790 Geospatial Foundation Project. 2014; http://qgis.osgeo.org.
791 49. Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, et al. Targeted
792 enrichment strategies for next-generation plant biology. American Journal of Botany 2012;99:
793 291-311.

794 50. Tsangaras K, Wales N, Sicheritz-Pontén T, Rasmussen S, Michaux J, Ishida Y, et al.
795 Hybridization capture using short PCR products enriches small genomes by Capturing Flanking
796 sequences (CapFlank). PLoS ONE 2014;9: e109101.
797 51. Faircloth BC, Branstetter MG, White ND, Brady SG. Target enrichment of
798 ultraconserved elements from arthropods provides a genomic perspective on relationships
799 among Hymenoptera. Molecular Ecology Resources 2015;15: 489–501. doi: 10.1111/1755-
800 0998.12328.
801 52. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated
802 cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 2010;38:
803 e87.
804 53. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, et al. Patterns of
805 damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of
806 Sciences USA 2007;104: 14616-14621.
807 54. Stiller M, Green RE, Ronan M, Simons JF, Du L, He W, et al. Patterns of nucleotide
808 misincorporations during enzymatic amplification and direct large-scale sequencing of ancient
809 DNA. Proceedings of the National Academy of Sciences USA, 2006;103:13578-13584.
810 55. Gilbert MTP, Hansen AJ, Willerslev E, Rudbeck L, Barnes I, Lynnerup N, et al.
811 Characterization of genetic miscoding lesions caused by postmortem damage. The American
812 Journal of Human Genetics, 2003, 72:48-61.
813 56. Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, et al. Empirical
814 assessment of RAD sequencing for interspecific phylogeny. Molecular Biology and Evolution
815 2014;31: 1272-1274.
816 57. Eaton DA, Ree RH. Inferring phylogeny and introgression using RADseq data: an
817 example from flowering plants (Pedicularis: Orobanchaceae). Systematic Biology 2013;62: 689-
818 706.

819 58. Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, et al. A Framework
820 Phylogeny of the American Oak Clade Based on Sequenced RAD Data. PLoS ONE 2014;9:
821 e93975.
822 59. Jones JC, Fan S, Franchini P, Schartl M, Meyer A. The evolutionary history of
823 Xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction
824 site-associated DNA sequencing. Molecular Ecology 2013;22: 2986-3001.
825 60. Rubin BE, Ree RH, Moreau CS. Inferring phylogenies from RAD sequence data. PLoS
826 ONE 2012;7: e33394.
827 61. Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, et al. Genome-wide
828 RAD sequence data provide unprecedented resolution of species boundaries and relationships
829 in the Lake Victoria cichlid adaptive radiation. Molecular Ecology 2013;22: 787-798.
830 62. Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW.
831 Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus
832 restriction site associated DNA sequencing. Genome Biology and Evolution 2015;7: 706–719.
833

834 **Supporting Information**

835

836 S1 Fig. Profile of the RAD-probes precursor, the RAD-seq library. Left panel, *X*-axis:
837 fragment size (semi-*log* scale); *Y*-axis: fragment density (Relative Fluorescent Units). Right
838 panel, gel-like representation of the left panel.

839

840 S2 Fig. Profile of the re-amplified capture library after AMPure purification. Left panel,
841 *X*-axis: fragment size (semi-*log* scale); *Y*-axis: fragment density (Relative Fluorescent Units).
842 Right panel, gel-like representation of the left panel.

843

844 S3 Fig. Illustration of the clustering optimization of RAD-ref assembly clustering
845 thresholds using Vsearch. *X*-axis: clustering threshold; *Y*-axis: number of clusters with 2x (red)
846 or 3x (green line) coverage. The top panel shows within-sample, whereas the bottom panel
847 shows among-sample clustering results. The optimal threshold optimizes the number of the
848 clusters with 2x and 3x coverage.

849

850 S4 Fig. Proportion of type II transitions to all the transitions and transversions for each
851 of the reference catalog, with and without post-mortem bias correction. The data shown are for
852 the fresh sample with DNA sonication and the museum samples without sonication. The left plot
853 shows values without and the right plot with mapDamage2.0 correction.

854

855 S1 Table. mapDamage2.0 results based on each of the three reference catalogs for *L.*
856 *helle* analyses, with the number of obtained SNPs (with and without application of
857 mapDamage2.0).