Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on museum collection specimens.

Authors:

Tomasz Suchan[1*], Camille Pitteloud[1*], Nadezhda Gerasimova[2,3], Anna Kostikova[3], 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., United Kingdom

[4] Władysław Szafer Institute of Botany, Polish Academy of Sciences, Kraków, Poland

Corresponding author: tomasz.suchan@unil.ch

[*] TS and CP equally participated and are considered as joint first authors

Funding: NA was funded by a Swiss National Science Foundation grant (PP00P3_144870). The work was financially supported by a grant from Switzerland through the Swiss Contribution to the enlarged European Union (Polish-Swiss Research Program, project no. PSPB-161/2010). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions: Conceived the original idea: TS, NAl, NAr. Designed the experiments: TS, CP. Performed the experiments: CP. Contributed to testing the protocol and sampled museum specimens: MP. Analyzed the data: AK, NG, NAl, NA, TS. Wrote the paper: TS, CP, NAl, AK, NG, NAr, MR.

Competing interests: AK is a founder of the insidedna.me platform used for analyzing the datasets. Other authors have declared that no competing interests exist.

Abstract

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 utility, this method is of a limited use with degraded DNA samples, such as those isolated from museum specimens, as these are either less likely to harbor fragments long enough to comprise two restriction sites making possible ligation of the technical sequences required or performing size selection of the resulting fragments. In addition, RAD-sequencing also reveals a suboptimal technique when applied to an evolutionary scale larger than the intra-specific level, as polymophisms in the restriction sites cause loci dropout. Here, we address both of these limitations by a novel method called hybridization RAD (hyRAD). In this method, biotinylated RAD fragments, covering a random fraction of the genome, are used as baits for capturing homologous fragments from samples processed through a classical genomic shotgun sequencing protocol. This simple and cost-effective approach allows sequencing orthologous sequences 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, and can be applied to broader phylogenetic scales. 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 over 10.000 single nucleotide polymorphisms present in all eight analyzed specimens, including 58 years old museum samples.

Keywords: RADseq, museum genomics, low-content DNA samples, hybridization capture

Introduction

With the advent of next-generation sequencing, conducting genomic-scale studies on non-model species has become a reality (Ellegren 2014). The cost of genome sequencing has dramatically dropped over the last decade and depositories 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 (Davey et al. 2011; McCormack et al. 2013). Among the most popular are approaches relying on hybridization capture of exome (Sulonen et al. 2011) or conserved fragments of the genome (Faircloth et al. 2012), RNA sequencing (RNAseq; Chepelev et al. 2009), and Restriction-Associated-DNA sequencing (RADseq; Baird et al. 2008; Peterson et al. 2012). The latter has been developed in many different versions, but generally relying 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) generation, and currently represents the best tool available to tackle questions related to ecological genomics *sensu lato*. The wide utility of RAD-sequencing in ecological, phylogenetic and phylogeographic studies is however limited by two main factors: i) the degree of divergence among the studied specimens, that translates into DNA sequence polymorphism at the restriction sites targeted by the RAD protocols; ii) the quality of the starting genomic DNA.

Sequence polymorphism at the DNA restriction site causes a progressive loss of shared restriction sites among diverging taxa and results in null alleles for which sequence data cannot be obtained. At the inter-specific level, this limitation critically reduces the number of orthologous loci that can be surveyed across the complete set of analyzed specimens (Rubin et al. 2012; Eaton & Ree 2013; Jones et al. 2013; Wagner et al. 2013; Cruaud et al. 2014; Hipp et al. 2014). At the intra-specific level, null alleles can lead to biased genetic diversity estimates (Arnold et al. 2013; Gautier et al. 2013). This phenomenon, combined to other technical issues –

such as PCR competition effects – is a serious limitation of most classic RAD-sequencing protocols that needs to be addressed.

In addition, RAD-sequencing protocols rely on relatively high molecular weight DNA (especially for the ddRAD protocol; Puritz et al. 2014), notably because enzyme digestion and size selection of resulting fragments are the key steps for retrieving sequence data across orthologous loci. As a result, RAD-sequencing cannot be applied to degraded DNA samples, a limitation also shared by the classical amplicon-sequencing methods. Museum collections, although encompassing samples covering large spatial areas and broad temporal scales, have not necessarily ensured optimal conditions to DNA preservation. As a result, many museum specimens yield highly fragmented DNA – even for relatively recently collected specimens (e.g. Mason et al. 2011; Staats et al. 2011; Tin et al. 2014), and have long remained of limited use for molecular ecology, conservation genetics, phylogeographic and phylogenetic studies (Wandeler et al 2007; Rowe et al. 2011; Bi et al. 2013). A cost-effective and widely applied approach overcoming this limitation would open capital perspectives of exploring biological collections, e.g., for rare or now extinct taxa/lineages or organisms occurring ephemerally in natural habitats (as many fungi or insects). It would also open new research avenues, allowing temporal comparisons of genetic diversity in natural populations – now applied only in a handful of cases (e.g. Thomas et al. 1990; Harper at al. 2006; Bi et al. 2013).

The development of hybridization-capture methods has recently been acknowledged as a promising way to address both of these limitations (Jones & Good 2015; Orlando et al. 2015). Such approaches however largely rely on prior genome/transcriptome knowledge and still remain confined to model organisms. Addressing this limitation, the recent development of hybridization-capture of Ultraconserved Elements (UCE) allows targeting homologous loci at broad phylogenetic scales using one set of probes. It however requires a time-consuming design and costly synthesis of the probes for capturing the DNA sequences of interest. Similarly, exon capture techniques, recently applied to the field of museum genomics (Bi et al. 2013),

require fresh specimens for RNA extraction or synthesizing the probes based on the known transcriptome.

Here, we present an approach we called 'hybridization RAD' (hyRAD), in which DNA fragments generated by double digestion RAD protocol (ddRAD; Peterson et al. 2012) are used as hybridization capture probes to enrich NGS library in the fragments of interest. Our protocol thus combines the simplicity and relatively low cost of developing RAD-sequencing libraries with the power and accuracy of hybridization-capture methods. This enables the effective use of low quality DNA and limits the problems caused by sequence polymorphisms at the restriction site. As a result, orthologous loci can be sequenced among divergent taxa and null alleles show increased detectability. Moreover, utilizing standard ddRAD and shotgun sequencing protocols allows application of the hyRAD protocol in laboratories already utilizing the abovementioned methods, for little cost.

In short, the approach consists of the following steps (Fig. 1):

1) generation of the ddRAD library based on high-quality DNA samples, narrow size selection of the resulting fragments and removing adaptor sequences,

2) biotinylation of the resulting fragments, hereafter called the probes,

3) constructing shotgun sequencing library from fresh DNA samples or degraded DNA from museum specimens,

4) hybridization capture of the resulting shotgun libraries on the probes, immobilization on streptavidin-coated magnetic beads, washing off non-binded sequences,

5) sequencing enriched shotgun libraries and ddRAD probes as a reference,

6) bioinformatic treatment (Fig. 2): assembly of the reads into contigs, alignment to the probes or *de novo* assembly.

In this paper, we describe the laboratory and bioinformatical pipelines for obtaining hyRAD data, as well as validate the usefulness of the method on the dataset obtained from museum and fresh specimens of *Lycaena helle* butterflies. We also 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 probes in order to obtain longer loci for mapping; iii) *de novo* assembly of the captured reads from a single well-preserved specimen for the reference (either keeping all or only large contigs in the catalog).

Materials and methods

Study species and study design

We used eight samples of *Lycaena helle* (Lepidoptera, Lycaenidae): seven historical dry-pinned and one recently collected and ethanol-preserved. Four of the dry-pinned samples were 30 years old (collected in 1985), and three were 58 years old (collected in 1957). All of the museum samples were loaned from the Finnish Museum of Natural History. The ethanol-preserved sample was obtained from Roger Vila's Butterfly Diversity and Evolution lab collection in Barcelona. In addition, ethanol-preserved samples from Romania, France and Kazakhstan (R. Vila collection) were used for generating the RAD probes, in order to cover variation within the full species range. For the detailed information on the samples used, see Table 1.

Using these eight samples we tested library preparation protocols differing in applying or not DNA fragmentation by sonication, in order to compare difference between fresh and historical DNA of different age, and test the importance of DNA sonication for each case. DNA extraction and library preparation using museum specimens was performed using consumables dedicated to the museum specimens only. Benches were cleaned with bleach and filter tips were used at all stages of lab work.

DNA extraction

DNA was extracted from butterfly legs from fresh and museum specimens (see Table 1). As museum specimens are characterized by low-content of degraded DNA, DNA isolation protocol was optimized accordingly. Museum samples were extracted using QIAamp DNA Micro kit (Qiagen, Hombrechtikon, Switzerland) in a laboratory dedicated to low-DNA content samples at

the University of Lausanne, Switzerland. For these samples, DNA recovery was improved by prolonged sample grinding, overnight incubation in the lysis buffer for 14h and elution in 20 µl with gradual column centrifugation. Extraction of fresh samples was performed using DNeasy Blood & Tissue Kit (Qiagen).

RAD probes preparation

Probes were prepared using double-digestion RAD approach (Peterson et al. 2012; Mastretta-Yanes et al. 2014; A. Brelsford, personal communication), with further modifications.

Total genomic DNA was digested at 37°C for 3 hours in a 10 µl reaction, containing 6 µl of DNA, 1x CutSmart buffer (New England Biololabs - NEB, Ipswich, USA), 1 U MseI (NEB) and 2 U of SbfI-HF (NEB). The reaction products were purified using AMPure XP (Beckman Coulter, Brea, USA), with a ratio of 2:1 with the sample, according to the manufacturer's instructions, and resuspended in 10 µl of 10 mM Tris buffer. Subsequently, adaptors were ligated to the purified restriction-digested DNA in a 20 µl reaction containing 10 µl of the insert, 1 µl of the barcoded 10uM RAD-P1 adaptor, 1 µl of 10 uM universal RAD-P2 adaptor, 1x T4 ligase buffer, and 400 U of T4 DNA ligase (NEB). Adaptor sequences are shown in Table 2; single strand adaptor oligonucleotides are annealed before use by heating to 95°C and gradual cooling. Ligation was performed at 16°C for 3 hours. The reaction products were purified using AMPure XP, ratio 1:1 with the sample, and resuspended in 10 mM Tris buffer. The ligation product was size-selected using Pippin Prep electrophoresis platform (Sage Science, Beverly, USA) with a peak at 270 bp and 'tight' size selection range (see Fig. S1 for the desired profile).

The resulting template was amplified in a 10 µl PCR reaction consisting of 1x Q5 buffer, 0.2 mM of each dNTP, 0.6 uM of each primer (Table 2), and 0.2 U Q5 hot-start polymerase (NEB). The program started with 30 sec at 98°C, followed by 30 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 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. Success of size selection and PCR reactions was confirmed by Bioanalyzer (Agilent Technologies, Santa Clara, USA) (see Fig. S2). Afterwards, the PCR products were pooled and purified using AMPure XP, with a ratio of 1:1 with the amplified DNA volume.

An aliquot of the resulting library was sequenced and the rest of the library was converted into probes by removing adaptor sequences by enzyme restriction at 37°C for 3 hours in a 50 µl reaction containing 30 µl of DNA, 1x CutSmart buffer (NEB), 5 U of MseI (NEB) and 10 U of SbfI-HF (NEB), replicated as required by the amount of the amplified product. The reaction was ended with 20 min enzyme inactivation at 65°C and the resulting fragments were purified using AMPure:reaction volume ratio of 1.5:1. Purified fragments were biotin nick-labelled using Biotin Nick Translation Mix (Life Technologies, Zug, Switzerland) according to the supplier's instructions and purified using AMPure:reaction volume ratio of 1.5:1. The resulting fragments will thereafter be referred to as probes.

Shotgun library preparation

Shotgun libraries were prepared from the fresh and museum specimens based on a published protocol for degraded DNA samples (Tin et al. 2014), modified in order to incorporate adaptor design of Meyer & Kircher (2010). The approach used for library preparation, utilizing barcoded P1 adaptor and 12 indexed P2 PCR primers, allows high sample multiplexing on a single sequencing lane (see Table 2; Meyer & Kircher 2010). For each fresh and museum sample, library was prepared from two aliquots of DNA: non-sonicated (i.e. high molecular weight DNA for the fresh sample and naturally degraded DNA from museum specimens; see Fig S1) and sonicated using Covaris instrument (Woburn, MA, USA) with a peak at 300 bp.

DNA samples were first 5'-phosphorylated in order to allow adaptor ligation in the next steps of the protocol. 8 µl of DNA was denatured in 95°C for 10 minutes and quickly chilled on ice. The 10 µl reaction consisting of denatured DNA, 1x PNK buffer and 10U of T4 polynucleotide kinase (NEB) was incubated at 37°C for 30min and heat inactivated at 65°C for 20 min. The

DNA was then purified using AMPure:reaction volume ratio of 2:1 and resuspended in 10 µl of 10 mM Tris buffer.

Guanidine tailing reaction of the 3'-terminus was performed after heat denaturation of DNA at 95°C for 10 minutes and quickly chilling on ice. The reaction composed of 1x buffer 4 (NEB), 0.25 mM cobalt chloride (NEB), 4 mM GTP (Life Technologies), 10 U TdT (NEB) and 10 µl of denatured DNA in 20 µl reaction volume was incubated at 37°C for 30 min and heat-inactivated at 70°C for 10 min.

Second DNA strand was synthesized using Klenow Fragment (3' → 5' exo-), with a primer consisting of the Illumina P2 sequence and a poly-C sequence homologous to the poly-G tail (see Table 2) added to the DNA strand in the previous reaction. A 10 µl reaction mix consisting of 1 µl of NEBuffer 4 (10x), 0.6 µl of dNTP mix (25 mM each), 1 µl of the P2 oligonucleotide (15 mM), 5.4 µl of water, and 2 µl of Klenow Fragment (3' → 5' exo-; NEB, 5 U/ul) was added to the 20 µl of the TdT reaction mix, incubated at 23°C for 3 h, and heat-inactivated at 75°C for 20 min. The double stranded product was blunt-ended by adding a mix consisting of 0.5 µl of NEBuffer 4 (10x), 0.35 µl of BSA (10 mg/ml), 0.2 µl of T4 DNA polymerase (NEB, 3 U/ul) and 3.95 µl of water, and incubated at 12°C for 15 min. The resulting product was purified using AMPure:reaction ratio of 2:1 and resuspended in 10 µl of 10 mM Tris buffer.

Barcoded P1 adaptors (see Table 2) were ligated to the 5'-phosphorylated end of the double-stranded product in a 20 µl reaction consisting of 10 µl of the double-stranded DNA, 1 µl of the 25 uM adaptors, 1x T4 DNA ligase buffer, and 400 U of T4 DNA ligase (NEB). Adaptors have to be annealed before use, as explained in the RAD probes protocol. The reaction was incubated at 16°C overnight. The resulting product was purified using AMPure:reaction ratio of 1:1 and resuspended in 20 µl of 10 mM Tris buffer. Ligated P1 adaptors were filled-in in a 40 µl reaction consisting of 20 µl of purified ligation product, 1x ThermoPol reaction buffer (NEB), 12 U of Bst polymerase (NEB), and dNTPs (0.25 mM each), and incubated at 37°C for 20 min.

The resulting template was amplified in a polymerase chain reaction (PCR), adding 15 µl of a mix consisting of 5 µl of Q5 reaction buffer (5x), 0.2 µl of dNTPs (25 mM each), 2.5 µl of the

PCR primer mix (5 uM each), and 0.5 U of Q5 Hot Start High-Fidelity DNA polymerase (NEB) to the 10 µl of the template. The program started with 20 sec at 98°C, followed by 25 cycles of 10 sec at 98°C, 20 sec at 60°C, and 25 sec at 72°C, followed by a final extension for 2 min at 72°C. Success of each PCR reaction was validated using gel electrophoresis, and the resulting products were purified using AMPure:reaction ratio of 0.7:1. Samples were then pooled in equimolar ratios.

In solution hybridization capture, library reamplification and NGS

The hybridization capture and library enrichment steps described below are based on previously published protocols (Mason et al. 2011; Parks et al. 2012) with some modifications. The hybridization mix consisted of 6x SSC, 50 mM EDTA, 1% SDS, 2x Denhardt's solution, 2 uM of each blocking oligonucleotide (to prevent hybridization of technical sequences; see Table 2), 500 to 1000 ng of the probes and 500 to 1000 ng of the library for the capture, in a total volume of 40 ul. The mix was denatured at 95°C for 10 min, followed by 48-hour incubation at 65°C. The probes with the hybridized fragments of the library were then separated by adding streptavidin beads (Dynabeads M-280, Life Technologies). 10 µl of the beads solution was washed three times on the magnet with 200 µl of TEN buffer (10 mM Tris-HCl 7.5, 1 mM EDTA, 1 M NaCl) and resuspended in 200 µl of TEN. 40 µl of the hybridization mix was added to the 200 µl of the beads solution and incubated for 30 min at room temperature. After separating the beads with the magnet, the supernatant was removed and the beads were washed by resuspending in 200 µl of 65°C 1x SSC/0.1% SDS mix, incubating for 15 min at 65°C, separating beads on the magnet and removing the supernatant. The above step was repeated three times with 1x SSC/0.1% SDS, 0.5x SSC/0.1% SDS and 0.1x SSC/0.1% SDS solutions. Finally, the hybridization-enriched product was washed off from the probes by adding 30 µl of 80°C water and 10 min incubation at 80°C.

Enrichment of the captured libraries was done in a 50 µl PCR reaction containing 1x Q5 reaction buffer (NEB), 0.2 mM dNTPs, 0.5 uM of each PCR primer (the P1 universal primer and

one of the 12 P2 indexed primers, see Table 2), 1U of Q5 Hot Start High-Fidelity DNA Polymerase (NEB), and 15 µl of the template. The program started with 20 sec at 98°C, followed by 25 cycles of 10 sec at 98°C, 20 sec at 60°C, and 25 sec at 72°C, followed by a final extension for 2 min at 72°C. The enriched-captured libraries were purified using AMPure XP:reaction ratio 1:1 and pooled in equimolar ratios for sequencing.

The libraries were sequenced on the Illumina MiSeq platform according to the manufacturer's' instructions. The probes were sequenced on one lane with 300 bp single-end sequencing protocol, and the captured libraries were sequenced on one lane with 150 bp paired-end protocol, along with the *de novo* assembly reference library.

Data analysis

The hyRAD datasets typically correspond to target-enriched libraries and cannot be analysed with the usual RAD pipelines (e.g. Catchen et al. 2013, Eaton 2014). Indeed, although they were generated using RAD loci, the obtained sequences are not flanked by the restriction sites and instead may not overlap completely or extend before and after the RAD locus. As a result, the analysis pipeline must include the following steps:

1) demultiplexing and cleaning of raw reads

2) building of reference sequences for each RAD locus

3) alignment of reads against the obtained references and SNP calling

Data analysis 1: demultiplexing and data preparation

The obtained reads were demultiplexed using fastx-multx tool from ea-utils package (Aronesty 2011). Reads of RAD-seq probes were processed by Trim Galore! Tool (Krueger 2015) and fastq-mcf from ea-utils package (Aronesty 2011). Reads for the probes were cleaned by Trimmomatic V0.30 tool (Bolger et al. 2014) to remove low quality nucleotides and technical sequences.

Data analysis 2: reference creation

Paired-end reads obtained from the hybridization-capture library for each sample were mapped onto three references: (1) consensus sequences for the clustered RAD-seq reads (RAD-ref), (2) RAD-seq probes' sequences extended using hybridization-captured reads (RAD-ref-ext), and (3) contigs assembled from the reads of hybridization-captured samples [further split into (3.1) assembly-ref and (3.2) assembly-long] as described below.

1) Vsearch RAD loci clustering (RAD-ref)

High quality reads of RAD probes were clustered by similarity using Vsearch (Flouri et al. 2015) to obtain loci for further mapping the reads from the hybridization-capture libraries. To obtain most reliable contigs across samples, Vsearch was run in two iterations. During the first iteration, we obtained consensus clusters on the within-individual level (i.e. clustering of hybridization-capture reads for each sample independently). During the second iteration, Vsearch was run on the consensus clusters obtained from the the first iteration. The second iteration allowed us to obtain consensus clusters at the among-individual level. For both iterations we ran Vsearch with various identity 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 0.51, 0.61, 0.71, 0.81, 0.85, 0.88 for the among-individual level) in order to identify an optimal identity threshold for clustering, i.e. a threshold that maximizes the number of clusters with a minimal coverage of 2x and 3x, respectively. In all cases, we used the cluster_fast option for clustering. The consensus sequences of each secondary cluster were then used as locus references in subsequent alignment and SNP calling steps.

2) Vsearch RAD loci clustering and extension using captured reads (RAD-ref-ext)

To obtain RAD-ref-ext, we iteratively extended contigs of the RAD-ref using reads from the hybridization-capture library (by pooling reads contributed by all the analysed specimens) by applying PriceTI tool (Ruby et al. 2013) with 30 cycles of extension and a minimum overlap of a

sequence match to 30. To avoid contamination with technical sequences, the obtained references were trimmed by 60 bp at each end.

3) Newbler assembly from captured reads only (assembly-ref, assembly-long)

Assembly was performed on the hybridization-captured reads of one ethanol-preserved sample. Only sequences obtained from the single fresh sample were used, as stringent cleaning parameters in Trimmomatic led to a large loss (up to 80%) of the reads from historical samples. Moreover, using one individual allows obtaining a more reliable reference, as any among-sample divergence can bias the assembly. We only used the first read of the paired-end reads due to considerably lower quality of the second read and higher quality of the resulting contigs when the paired read was dropped. We used Newbler *de novo* assembler V2.9 to assemble cleaned reads into contigs. In further analysis we tested two assemblies: one with all available resulting contigs (assembly-ref), and the second with only contigs larger than 500 bp (assembly-large).

Data analysis 3: mapping and SNP calling

Read mapping was performed using bowtie2 (Langmead & Salzberg 2012) and SNPs were called with samtools (Li et al 2009). To evaluate the level of DNA damage in museum DNA samples, we used mapDamage (Jónsson et al. 2013). Data sets for replicates were merged and analysed for SNPs with samtools package. VCF format files (Danecek et al. 2011) were converted to SNP-based NEXUS files using PGDSpider converter (Lischer & Excoffier 2012) and to structure data files for every individual using vcf-consensus tool (all used commands can be found in supplementary materials, see also Fig. 2).

Data analysis 4: Overlap detection between assembly references

To evaluate the level of overlap among the four assembly references (RAD-ref, RAD-ref-ext, assembly-ref and assembly-long) we used OrthoMCL (Li et al. 2003) pipeline for orthology detection. Most of the pipeline was run with the default parameters, except for Blastall and MCL

clustering steps. Here, we used stringent parameter values (e-value of 0.0001 and MCL was run with an inflation parameter of 2.0) in order to reduce chances of detecting false orthology groups. As a result, we obtain clusters of contigs being contributed by the four assembly references. We then counted how many of these clusters – presumably corresponding to homologous loci – were shared among the available reference assembly approaches.

Results and discussion

Sequencing and data quality

Probes sequencing yielded 14,188,023 and hybridization-capture libraries 16,636,502 raw reads: 8,217,522 for sonicated and 8,418,980 for non-sonicated samples, respectively. Additional sequencing of the reference library from the ethanol-preserved specimen (used for the assembly-ref creation) yielded 4,703,744 reads. The proportion of reads kept after quality filtering varied with sample age and preparation method. For the ethanol-preserved sample, 89.8% of reads from sonicated and 89.4% from non-sonicated sample were retained. For the 30 years old samples the mean was 74.3% and 79.6%, and for 58 years old samples 70.8% and 73.8% for sonicated and non-sonicated samples, respectively.

Reference creation

Consensus clustering of the RAD-based reference within individuals produced the best results with a clustering identity threshold of 0.91. This threshold resulted in the largest number of clusters with double and triple coverages compared to other threshold values. Consensus clustering among individuals produced the best results with a clustering identity threshold of 0.71 (see Fig. S3).

The highest number, length and the total length of reference contigs was obtained using *de novo* assembly with Newbler (assembly-ref; Table 3). In contrast, larger Newbler contigs alone produced rather low number of contigs. Both RAD-based assemblies produced similar numbers of contigs, but the extension performed on the obtained RAD reference followed by

trimming of technical sequences resulted in references with an average shorter length (lower N50) than the starting contigs—whereas priceTI extended a large number of probes, this did not reflect in a higher average loci length because of further trimming of obtained contigs.

Single-hit alignments, SNPs and matrix fullness

The highest level of single-hit alignments for each type of sample (fresh, 30 and 58 years old museum samples) were obtained with the reference loci from *de novo* assembly using reads from the hybridization-capture library from a single fresh specimen (assembly-ref; Fig. 3). This method provided the highest number of SNPs, 32,528 for non-sonicated and 35,236 for sonicated library, with respectively 30.5% and 29.7% of reads present in all eight specimens, and also the highest total number of SNPs present across all the samples (Fig. 5). Using only contigs larger than 500 bp (assembly-large) provided the most precise reference as shown by the ratio of multi-mapping to single-mapping events (ratio >1/1; Table S1), but also the lowest number of SNPs obtained and low coverage across the samples.

In contrast, references obtained via the clustering of RAD libraries (RAD-ref, RAD-ref-ext) produced the worst ratio of multi to single mapping events (Table S1). Although producing smaller numbers of reference contigs than the *de novo* assembly based approach, many of the captured reads mapped to more than one contig – even with fine-tuned parameters, Vsearch produced duplicated clusters of reads that in fact correspond to the same regions of genome. Despite that, this approach provided the highest percentage of SNPs present in all samples (54.8% for non-sonicated and 54.3% for sonicated), although the absolute number of SNPs present in all the samples was higher using the assembly-ref approach.

RAD-sequencing datasets depend on the presence of the restriction sites and therefore any polymorphism in such sites, leads to either missing loci or alleles. We observed that a large fraction of loci were successfully captured and sequenced across the 8 analysed specimens (right-hand side of Fig. 5). Combined with the high number of obtained SNPs, this allows obtaining largely filled data matrices. For instance, the best performing pipeline (assembly-ref)

produced 9,935 SNPs for non-sonicated and 10,451 SNPs for sonicated samples that were present in all sequenced individuals, including 58 years old museum specimens.

Loci representation

About 15% of the reference loci obtained via *de novo* assembly were shared with those based on the clustering of RAD probes and presumably contained the targeted restriction site. In contrast, an appreciable fraction of the obtained reference loci (40.5 %) were unique to the assembly-ref approach and were the largest contributors of low-coverage loci (i.e. those being sequenced in at most two specimens, left-hand side of Fig. 5). These loci most likely resulted from the co-hybridization of repetitive regions matching partially with the probes. This phenomenon, termed 'daisy-chaining' (Cronn et al. 2012; Tsangaras et al. 2014), can be significantly reduced by adding specific blocking agents to the hybridization mix (typically Cot-1, e.g. see Faircloth et al. 2015). Such developments are desirable because they increase the percentage of reads matching the loci of interest and eventually improve the overall sequencing coverage.

Effects of sample preparation and age

Differences in the number of single mapping events and in the numbers of SNPs obtained were not substantial between sonicated and non-sonicated samples, and depended on the sample age and the bioinformatic pipeline used (Table S1; Fig. 4 and 5). We expected that museum specimens should perform better without sonication, as the DNA was already visibly fragmented, and sonication of museum specimens may increase the levels of exogenous DNA contamination (by fragmenting intact fungal or bacterial DNA contaminating museum samples; L. Orlando, pers. comm.). Also contrary to our expectations, the effect of sonication on the quality of the library was also non-observable for the ethanol-preserved sample, suggesting that the sonication step might not be necessary even for the relatively well preserved samples with no obvious DNA degradation. In terms of mapping events (Table S1), the fresh sample showed a better ratio of multi- to single-mapping events when sonicated whereas museum samples

showed a better ratio when non-sonicated. We would therefore advise not to sonicate DNA obtained the museum specimens, which significantly cuts down the price and time required for library preparation, except in cases when no signs of degradation are observable on the DNA profile. As levels of DNA degradation of contemporary samples may vary, the sonication step is advisable when working with relatively well-preserved DNA.

The effects of sample age are observable at the level of mapping events, with a statistically significant decrease in single-mapping events in older museum samples (*t*-test, Bonferroni correction: $p = 0.010$ for RAD-ref, $p < 0.001$ for RAD-ref-ext, $p = 0.003$ for assembly-ref, and $p = 0.006$ for assembly-large; Fig. 3). On the contrary, the differences in the number of SNPs obtained between the 30 and 58 years old samples were not significant (*t*-test, Bonferroni correction: $p = 1.000$ for RAD-ref, $p = 1.000$ for RAD-ref-ext, $p = 0.804$ for assembly-ref, and $p = 0.726$ for assembly-large; Fig. 4).

As one of the main mechanisms of DNA degradation is deamination of cytosines, highly damaged ancient or museum DNA samples are characterized by high uracil content (Briggs et al. 2007; Briggs et al. 2010; Stiller et al. 2006). In classical library preparation protocols, the usage of a proofreading Q5 polymerase should stall in the presence of uracil and thus reduce the misincorporation errors in the final dataset. On the other hand this approach might not be optimal for highly degraded ancient DNA samples, with high uracil content, where large fraction of DNA fragments may carry cytosine to uracil misincorporations (Briggs et al. 2007). Moreover the usage of a proofreading polymerase does not prevent the misincorporations caused by direct deamination of methylated cytosine to thymine, or less common deamination of guanidine to adenine (Briggs et al. 2010; Stiller et al. 2006). In the protocol used above (Tin et al. 2014), the second strand synthesis was performed using Klenow Fragment (3’ → 5’ exo-), lacking proofreading ability, and thus approximately half of the resulting DNA fragments should have cytosine to uracil misincorporation substituted for thymine, amplifiable by proofreading DNA polymerase. We thus opted for a bioinformatic post-processing way of filtering-out such bases. Post-mortem damage in the sequenced samples was assessed by mapDamage, which

rescales sequence files by downscaling quality scores of likely post-mortem damaged bases. As some SNPs became filtered by lower quality scores after the rescaling, the number of SNPs is decreased after mapDamage. We expected higher number of discarded SNPs in the oldest samples, because of a higher proportion of DNA damage occurring with time. The proportion of SNPs discarded after applying mapDamage was not different however between ethanol-preserved, 30, and 58 years old museum samples: 6% for RAD-ref, 5% for RAD-ref-ext, 4% for assembly-ref, and 3% for assembly-large approach.

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, the obtained numbers of SNPs were not significantly different. Applying our protocol museum specimens from a Lycaenid butterfly, we obtained thousands of SNPs from samples up to 58 years old, confirming that it can be successfully applied in the field of museum genomics. Our method does not require time-consuming and costly probes design and synthesis, nor access to fresh samples for RNA extraction, making it the simplest and most straightforward technique for obtaining orthologous loci from degraded museum samples.

In the protocol, we applied a modified shotgun library preparation method, optimized for degraded DNA from museum specimens (Tin et al. 2014). However, the capture protocol presented here can be applied to any type of library preparation, including commercial ones, simplifying the workflow and cutting down the preparation time.

We also tested several bioinformatic approaches for loci assembly from the captured libraries, a crucial step when working on organisms without a reference genome. The best performing pipeline was relying on *de novo* reference assembly from captured reads from a single ethanol-preserved specimen, using Newbler assembler (assembly-ref). Despite maximum 26% of the obtained sequences mapped to the references (Table S1) and the proportion of single mapping events were not higher than 10% on average (Fig. 3), we could successfully call

more than ten thousand of loci in each case (Fig. 4), with very high coverage across the samples (Fig. 5). From the wetlab protocol perspective, in the hybridization step, we have used blocking oligonucleotides to prevent 'daisy-chaining' of captured sequences by technical sequences' homology. We suggest that using a blocking agent preventing similar chaining caused by repetitive sequences (Cot-1 DNA; Faircloth et al. 2015) can further increase the hybridization efficiency and thus the numbers of reads mapping on the reference and reduce the number of low-coverage loci (left-hand side of Fig. 5).

The best performing pipeline produced the largest contig of 5,233 bp, more than 27 times longer than the probes used. Although the mean length of the assembled contigs was much smaller, our method allows retrieving much longer sequences than the length of the probes used. The reason for this is that captured sequences hybridize with the other having homologous sequences, flanking the probe sequence (Cronn et al. 2012; Tsangaras et al. 2014). This may lead to enrichment across larger fractions of genome, a side effect of our method, that can be utilized for assembly of larger contigs by using longer probes and capturing longer targets.

The method presented here, although based on the restriction enzyme digestion of DNA to create the random genomic probes, does not depend on the restriction site presence in the captured library. This represents a serious improvement over classical RAD-sequencing datasets, whose increase in the number of missing sites is correlated with the increase in the phylogenetic distance among samples (Rubin et al. 2012; Eaton & Ree 2013; Jones et al. 2013; Wagner et al. 2013; Cruaud et al. 2014; Hipp et al. 2014), sometimes leading to conflicting signals between RAD- and capture-based datasets (Leaché et al. 2015) or are characterized by the presence of null alleles, leading to heterozygosity or F_{ST} underestimation (Arnold et al. 2013; Gautier et al. 2013). In this aspect, our approach is similar to other capture-enrichment protocols, such as ultraconserved elements (UCE; Faircloth et al. 2012) or exome-capture (Sulonen et al. 2011), with the benefit of much simpler and less expensive probe generation, without access to genome information or fresh specimens for RNA isolation. Not relying on the

presence of restriction site, the method presented here should be also useful for broader phylogenetic scales, allowing sequencing homologous loci from more divergent taxa, which would not be possible to retrieve using classical RAD-seq approaches.

Supplementary data

Updated versions of the lab protocol and bioinformatic pipeline can be found at https://github.com/chiasto/hyRAD. Moreover, automatized pipeline can be run at https://insidedna.me.

Acknowledgements

We thank A. Brelsford, A. Mastretta-Yanes and P. Rosikiewicz for their help with developing RAD-sequencing protocols. J. Patiño tested early versions of the protocol and provided valuable feedback. R. Vila and L. Kaila kindly provided the samples for the study. We also thank B. Emerson for his constant support during the development of this method.

References

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Molecular Ecology 2013;22: 3179–3190.

Aronesty E. ea-utils: command-line tools for processing biological sequencing data; 2011. Database: Google Code [Internet] Accessed: http://code.google.com/p/ea-utils

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 2008;3: e3376. doi:10.1371/journal.pone.0003376.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R & Moritz C. Unlocking the vault: next generation museum population genomics. Molecular Ecology 2013;22: 6018–6032.

Bolger AM, Lohse M & Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30: 2114-2120. doi:10.1093/bioinformatics/btu170

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences USA 2007;104: 14616-14621.

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 2010;38: e87.

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W. Stacks: an analysis tool set for population genomics. Molecular Ecology 2013; 22: 3124-3140.

Chepelev I, Wei G, Tang Q & Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Research 2009;37: e106.

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV & Udall J. Targeted enrichment strategies for next-generation plant biology. American Journal of Botany 2012;99: 291-311.

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G & Rasplus JY. Empirical assessment of RAD sequencing for interspecific phylogeny. Molecular Biology and Evolution 2014;31: 1272-1274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27: 2156-2158.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM & Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 2011;12: 499-510.

Eaton DA, & Ree RH. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Systematic Biology 2013;62: 689-706.

Eaton DA, PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 2014;30: 844-1849. doi:10.1093/bioinformatics/btu121

Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution 2014;29: 51-63.

Faircloth BC, Branstetter MG, White ND & Brady SG. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Molecular Ecology Resources 2015;15: 489–501. doi: 10.1111/1755-0998.12328

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology 2012;61: 717-726. doi:10.1093/sysbio/sys004.

Flouri T, Ijaz UZ, Mahé F, Nichols B, Quince C, Rognes T. VSEARCH GitHub repository. Release 1.0.16; 2015. Database: GitHub [Internet]. Accessed: https://github.com/torognes/vsearch. doi: 10.5281/zenodo.15524.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P et al. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology 2013;22: 3165-3178. doi: 10.1111/mec.12089

Harper GL, Maclean N, & Goulson D. Analysis of museum specimens suggests extreme genetic drift in the adonis blue butterfly (*Polyommatus bellargus*). Biological Journal of the Linnean Society, 2006;88: 447-452.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, et al. A Framework Phylogeny of the American Oak Clade Based on Sequenced RAD Data. PLoS ONE 2014;9: e93975.

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

Jones MR & Good JM. Targeted capture in evolutionary and ecological genomics. Molecular Ecology 2015. doi: 10.1111/mec.13304.

Jónsson H, Ginolhac A, Schubert M, Johnson P, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 2013;29: 1682-1684. doi: 10.1093/bioinformatics/btt193

Krueger F. Trim Galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. 2015 Available: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Langmead B & Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods 2012;9: 357-359.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD & Linkem CW. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biology and Evolution 2015;7: 706–719.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 2009;25: 2078-2079.

Li L, Stoeckert CJ & Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 2003;13: 2178-2189.

Lischer HEL and Excoffier L. PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 2012;28: 298-299.

Mason VC, Li G, Helgen KM & Murphy WJ. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Research 2011;21: 1695-1704.

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D & Emerson BC. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Molecular Ecology Resources 2015;15: 28-41. doi: 10.1111/1755-0998.12291

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

Meyer M & Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010;2010: t5448. doi: 10.1101/pdb.prot5448.

Orlando L, Gilbert MTP, & Willerslev E. Reconstructing ancient genomes and epigenomes. Nature Reviews Genetics 2015;16: 395-408. doi: 10.1038/nrg3935

Parks M, Cronn R & Liston A. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). BMC Evolutionary Biology 2012;12: 100. doi: 10.1186/1471-2148-12-100

Peterson BK, Weber JN, Kay EH, Fisher HS & Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 2012;7: e37135.

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI & Bird CE. Demystifying the RAD fad. Molecular Ecology 2014;23: 5937–5942. doi: 10.1111/mec.12965

Rowe KC, Singhal S, MacManes MD, Ayroles JF, Morelli TL, Rubidge EM, et al. Museum genomics: low-cost and high-accuracy genetic data from historical specimens. Molecular Ecology Resources 2011;11: 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x

Rubin BE, Ree RH & Moreau CS. Inferring phylogenies from RAD sequence data. PLoS ONE 2012;7: e33394.

Ruby JG, Bellare P & DeRisi JL. PRICE: software for the targeted assembly of components of (meta) genomic sequence data. G3: Genes, Genomes, Genetics 2013;3: 865-880.

Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, et al. DNA Damage in Plant Herbarium Tissue. PLoS ONE 2011;6: e28448. doi: 10.1371/journal.pone.0028448

Stiller M, Green RE, Ronan M, Simons JF, Du L, He W, et al. Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. Proceedings of the National Academy of Sciences USA, 2006;103: 13578-13584.

Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biology 2011;12: R94.

Thomas WK, Pääbo S, Villablanca FX, & Wilson AC. Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens. Journal of Molecular Evolution 1990;31: 101-112.

Tin MM-Y, Economo EP, Mikheyev AS. Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. PLoS ONE 2014;9: e96793. doi:10.1371/journal.pone.0096793

Tsangaras K, Wales N, Sicheritz-Pontén T, Rasmussen S, Michaux J, Ishida Y et al. Hybridization capture using short PCR products enriches small genomes by Capturing Flanking sequences (CapFlank). PLoS ONE 2014;9: e109101.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L et al. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Molecular Ecology 2013;22: 787-798.

Wandeler P, Hoeck PE & Keller LF Back to the future: museum specimens in population genetics. Trends in Ecology & Evolution 2007;22: 634-642.

Tables

Table 1: *Lycaena helle* samples used in the study

Sample name	Type of preservation	Year of collection	DNA concentration (ng/ul)	Locality
Capture library				
LHKS103	dry pinned	1957	2.12	Kuusamo, Finland
LHKS104	dry pinned	1957	3.02	Kuusamo, Finland
LHKS105	dry pinned	1957	1.48	Kuusamo, Finland
LHKS106	dry pinned	1985	29.2	Kuusamo, Finland
LHKS107	dry pinned	1985	19.7	Kuusamo, Finland
LHKS108	dry pinned	1985	17.5	Kuusamo, Finland
LHKS109	dry pinned	1985	8.84	Kuusamo, Finland
07-D251	ethanol	2007	2.12	Dumbrava Vadului, Romania
Probes				
07-D251	ethanol	2007		Dumbrava Vadului, Romania
09-V309	ethanol	2009		Porte Puymorens, France
11-G718	ethanol	2011		Uspenka, Kazakhstan

Table 2: Oligonucleotides used in the protocol. x = barcode sequence in the adaptors; barcode sequences can be designed using published scripts (Meyer & Kircher 2010), available at: https://bioinf.eva.mpg.de/multiplex/; I = inosine in the region complementary to te barcode in blocking oligonucleotides sequences.

RAD probes P1 adaptors, SbfI-compatible (RAD-P1)	
RAD-P1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxCCTGCA
RAD-P1.2	GGxxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
RAD probes P2 adaptor, MseI-compatible (RAD-P2)	
RAD-P2.1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RAD-P2.2	TAAGATCGGAAGAGCGAGAACAA
Shotgun library P1 adaptors	
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

Table 3: Data on obtained references.

	Number of contigs	Largest contig (bp)	Total length (bp)	N50
RAD-ref	25 478	544	5 445 942	209
RAD-ref-ext	24 820	851	2 613 024	98
assembly-ref	42 273	5 233	6 955 153	160
assembly-large	474	5 233	448 289	923

Figures

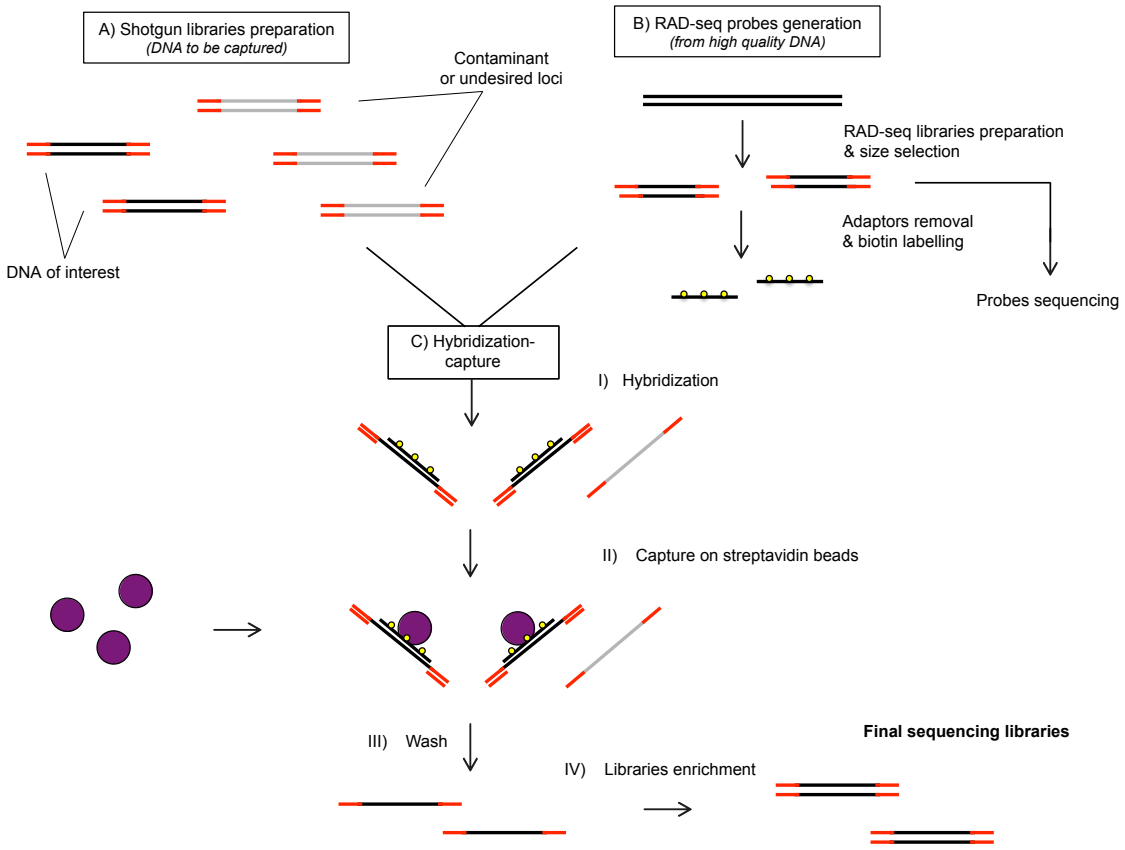

Fig. 1: Lab-work procedure used for hyRAD. Homologous reads from shotgun genomic libraries are captured through hybridization on random RAD-based probes. These fragments are then separated using streptavidin-coated beads and sequenced.

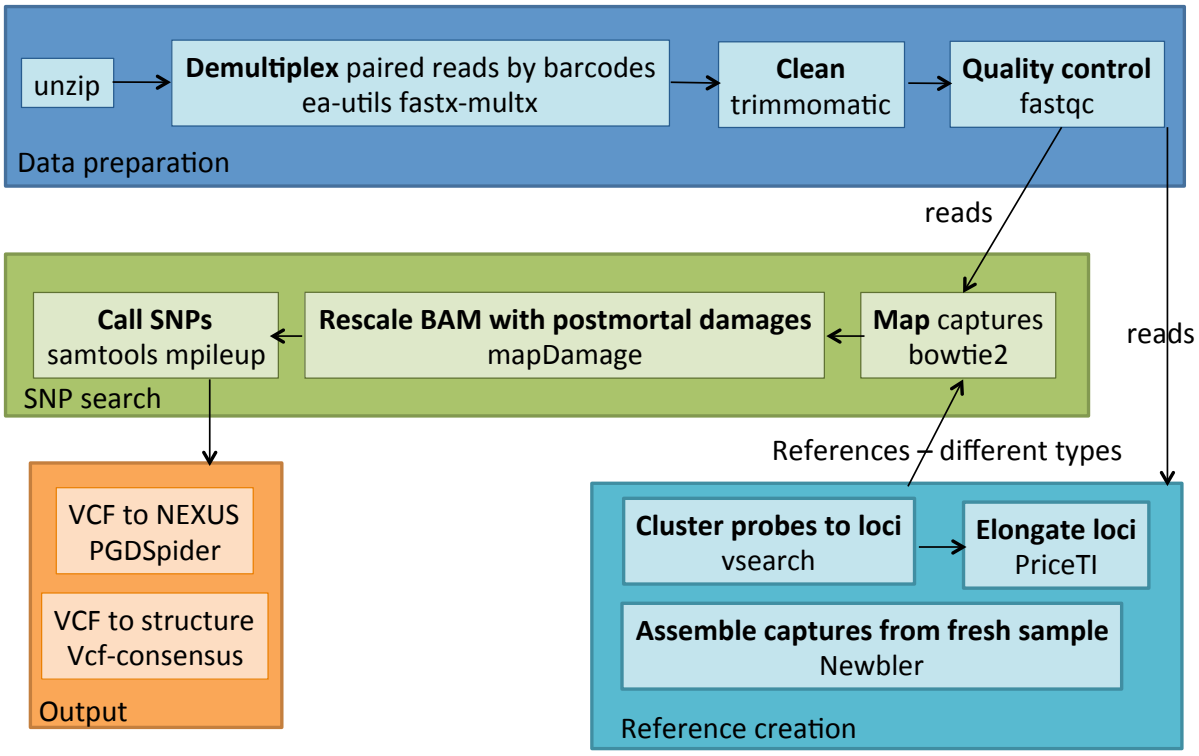

Fig. 2: Bioinformatical pipeline used for processing hyRAD sequences. First, the reads are demultiplexed and cleaned. Different types of references were build and the captured fragments were mapped on the reference. The SNPs are then called after correcting for post-mortem DNA damages.

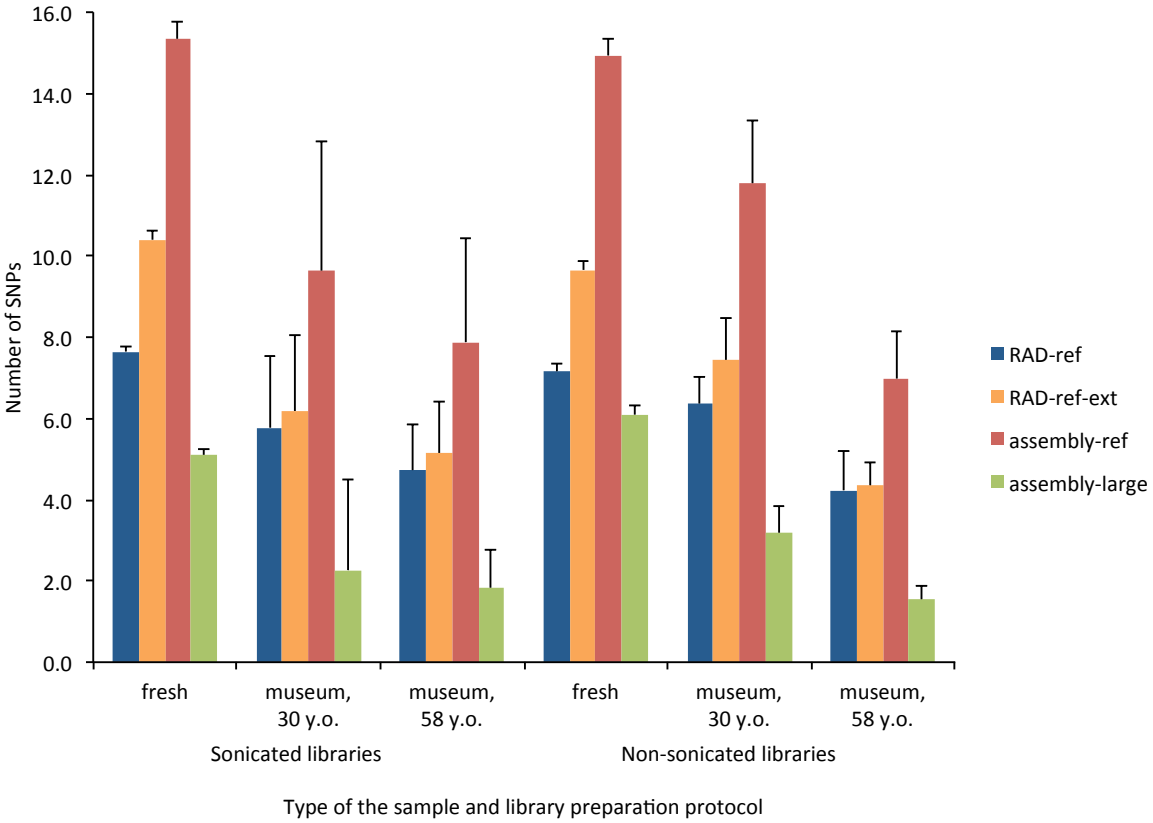

Fig. 3: Percentage of the captured reads showing unique mapping events for different types of DNA preparations and bioinformatical pipelines.

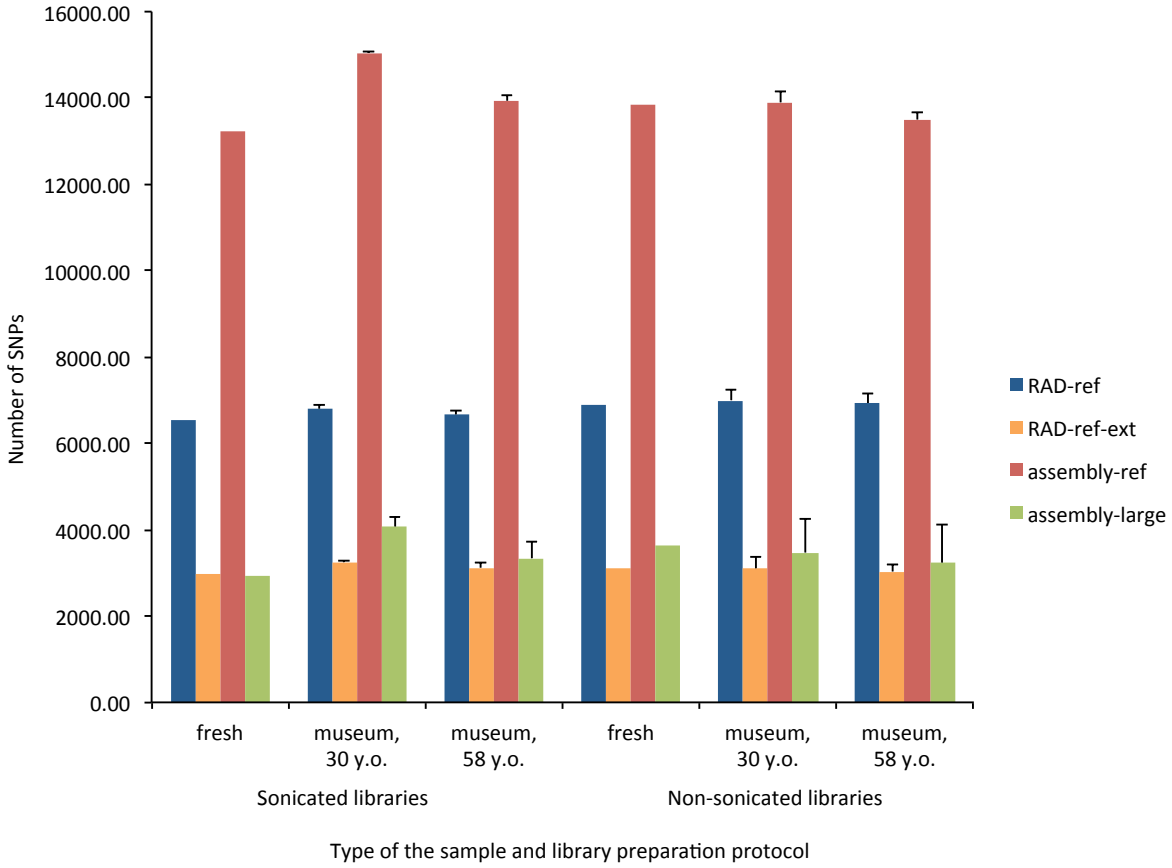

Fig. 4: Mean number of SNPs per sample obtained for different types of DNA preparations and bioinformatical pipelines.

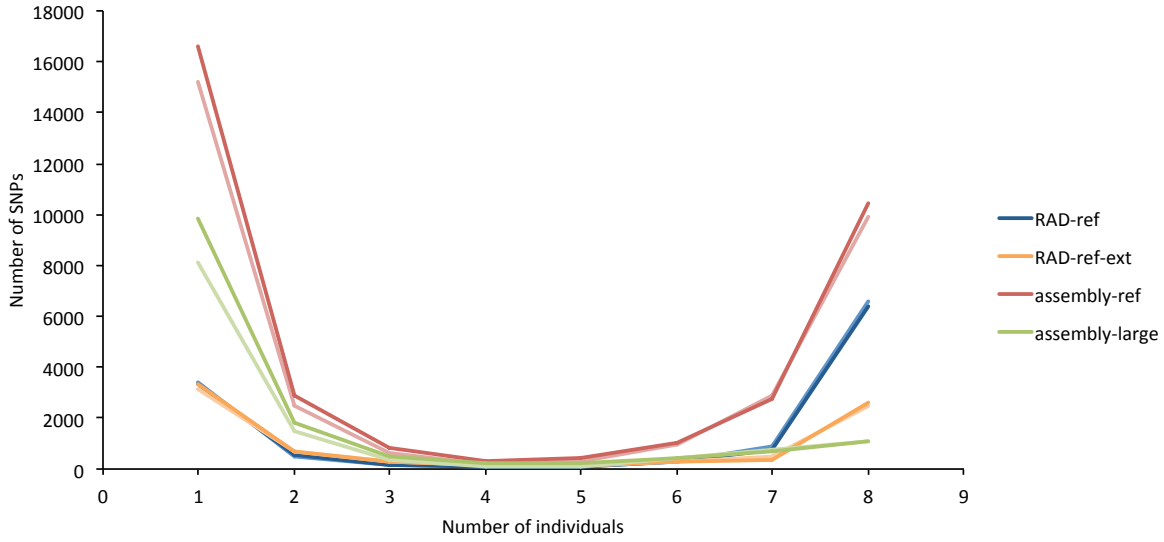

Fig. 5: Among-sample SNP coverage obtained with different bioinformatical pipelines. The horizontal axis represents the number of individuals, the vertical axis represents the number of SNPs. Darker color denotes sonicated, lighter colour non-sonicated samples. A large number of polymorphisms is found in only one or two samples (left-hand of the graph), most likely because suboptimal specific-sequence capture. Despite that, large numbers of SNPs are present in all the sequenced individuals (right-hand of the graph).

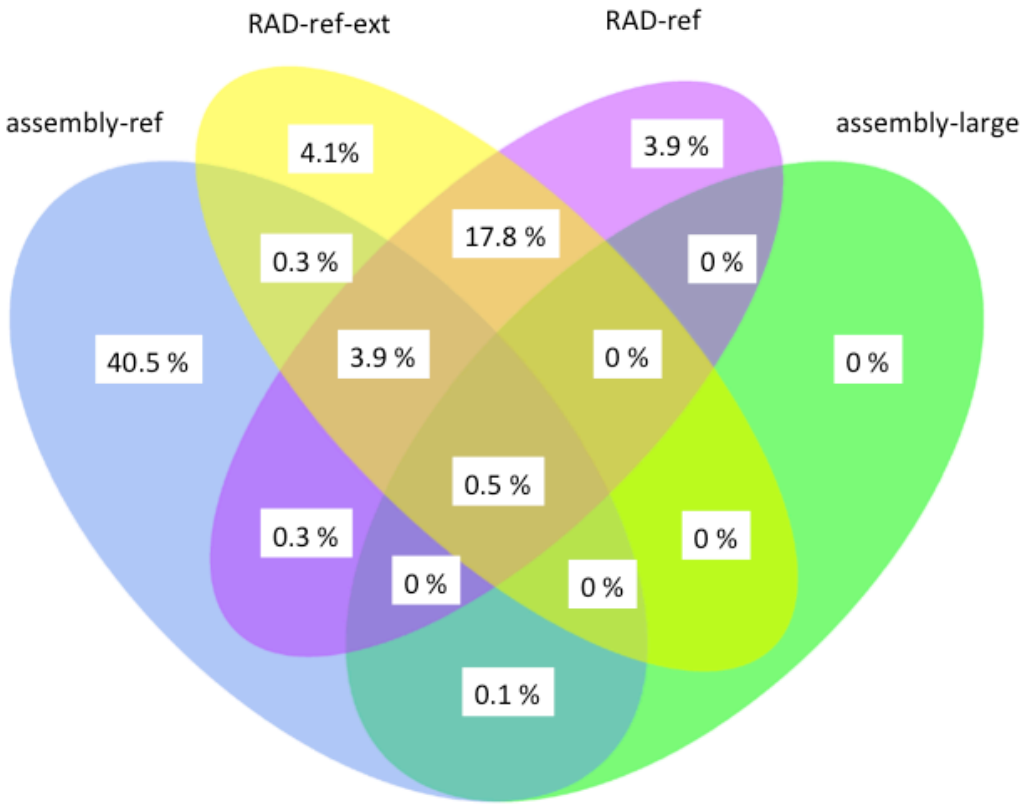

Fig. 6: Overlap percentage of the loci obtained using different bioinformatical approaches, identified using the OrthoMCL (Li et al. 2003) pipeline for orthology detection.