MOLECULAR ECOLOGY
RESOURCES WILEY

Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA

Patricia L. M. Lang[1,2] | Clemens L. Weiß[1,3] | Sonja Kersten[4] | Sergio M. Latorre[1] | Sarah Nagel[5] | Birgit Nickel[5] | Matthias Meyer[5] | Hernán A. Burbano[1,6]

[1]Research Group for Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

[2]Department of Biology, Stanford University, Stanford, CA, USA

[3]Department of Genetics, Stanford University, Stanford, CA, USA

[4]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany

[5]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[6]Centre for Life's Origins and Evolution, Department of Genetics, Evolution, and Environment, University College London, London, UK

Correspondence
Hernán A. Burbano, Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK.
Email: h.burbano@ucl.ac.uk

Abstract

Species' responses at the genetic level are key to understanding the long-term consequences of anthropogenic global change. Herbaria document such responses, and, with contemporary sampling, provide high-resolution time-series of plant evolutionary change. Characterizing genetic diversity is straightforward for model species with small genomes and a reference sequence. For nonmodel species—with small or large genomes—diversity is traditionally assessed using restriction-enzyme-based sequencing. However, age-related DNA damage and fragmentation preclude the use of this approach for ancient herbarium DNA. Here, we combine reduced-representation sequencing and hybridization-capture to overcome this challenge and efficiently compare contemporary and historical specimens. Specifically, we describe how homemade DNA baits can be produced from reduced-representation libraries of fresh samples, and used to efficiently enrich historical libraries for the same fraction of the genome to produce compatible sets of sequence data from both types of material. Applying this approach to both *Arabidopsis thaliana* and the nonmodel plant *Cardamine bulbifera*, we discovered polymorphisms de novo in an unbiased, reference-free manner. We show that the recovered genetic variation recapitulates known genetic diversity in *A. thaliana*, and recovers geographical origin in both species and over time, independent of bait diversity. Hence, our method enables fast, cost-efficient, large-scale integration of contemporary and historical specimens for assessment of genome-wide genetic trends over time, independent of genome size and presence of a reference genome.

KEYWORDS
ancient DNA, capture, herbarium, hybridization double-digest RADseq, nonmodel species

1 | INTRODUCTION

Evolutionary studies have over recent years moved from focusing on the effects of various evolutionary forces on genetic variation at single loci (McDonald & Kreitman, 1991) to investigating whole genome sequencing data (Mackay et al., 2012). With the continuous development of high-throughput next-generation sequencing (NGS) technologies (e.g., short-read Illumina sequencing: HiSeq4000, NovaSeq [Bentley et al., 2008]), such questions can now in principle be addressed at the population scale, covering large geographical distributions (The 1001 Genomes Consortium 2016), or densely sampled phylogenetic space (Zhang et al., 2014). A limiting factor especially for phylogenetic studies, both in terms of sequencing cost and regarding downstream analysis, are species that lack reference genomes, have large genomes, or both. However, this is true for the majority of species, excluding a few well-studied model organisms such as *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000) or the genus *Drosophila* (Drosophila 12 Genomes Consortium et al., 2007). Most population-scale studies in molecular ecology or evolutionary and conservation genomics circumvent this bottleneck using a variety of reduced-representation approaches such as restriction-enzyme associated DNA sequencing (RADseq) (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Baird et al., 2008; Catchen et al., 2017; Miller, Dunham, Amores, Cresko, & Johnson, 2007; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012; Puritz et al., 2014) or exome sequencing (De Wit, Pespeni, & Palumbi, 2015). This trades large amounts of shallowly sequenced genomes, which are difficult to analyse without a reference genome, for sequence data of higher quality and depth, which can be readily analysed with dedicated bioinformatics pipelines (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011), independent of a reference genome.

Despite their reduced view on the genome, these approaches serve to infer evolutionary processes based on contemporary sequence variation (Andrews et al., 2016). With the advent of ancient DNA sequencing, however, we now have the opportunity to study evolution in real time (Gutaker & Burbano, 2017; Shapiro & Hofreiter, 2014). This is particularly relevant in the context of anthropogenic global change, which has been affecting the environment at a rapid pace for the last +200 years (Lang, Willems, Scheepens, Burbano, & Bossdorf, 2018). To date, largely uncharacterized species responses to this selective force are key to understanding the long-term consequences of global change, and to promoting species survival (Aitken & Bemmels, 2016)—a key challenge of our time. In the case of plants, dense time-series that document plant responses to environmental change are stored in herbaria. This largely untapped resource provides a global collection of specimens that, especially combined with contemporary sampling, allows for studying plant morphological and molecular change over the last ~200 years in minute detail (Bieker & Martin, 2018; Lang et al., 2018; Meineke et al., 2018).

However, the specific molecular characteristics of DNA retrieved from historical specimens, so called ancient DNA (aDNA; Pääbo et al., 2004), complicate using such samples at large scale, as they do not allow the use of RADseq. The most limiting characteristic in the context of reduced-representation methods is the age-related breakdown of aDNA fragments to median lengths of 50–80 bp (Sawyer, Krause, Guschanski, Savolainen, & Pääbo, 2012; Weiß et al., 2016). Enzymatic restriction used in RADseq approaches would further shorten these fragments, thereby reducing their mappability (Figure S1) and thus the overall information content historical samples can provide. In addition, fragmentation is likely to reduce the number of available RAD sites over time, thereby also reducing the information that can be retrieved, and the overlap between time-series samples. These problems would be even more pronounced in double-digest RADseq (ddRADseq), which uses two restriction enzymes with different cutting sequences (Peterson et al., 2012).

The combination of historical and modern samples is thus difficult when RADseq approaches are the only feasible option, for example when working with large genome sizes, or population-scale sampling. Joint analyses of the different sample types require high sequence overlap, which in this situation cannot be achieved by employing the same method across samples. For historical samples, deep whole genome sequencing can be used to retrieve the sites recovered with RADseq of modern samples—a costly and unrealistic solution for large genomes and sample sizes, especially considering the lower quality and metagenomic nature of aDNA (Gutaker & Burbano, 2017; Poinar et al., 2006). To enrich historical samples for specific genomic subsets, many studies therefore employ hybridization-based captures where biotinylated baits target particular regions of the genome. The resulting complexes are immobilized on streptavidin-coated beads, and washing steps remove unassociated "background" DNA prior to sequencing of the thus enriched targeted DNA. These protocols often use commercially synthesized baits (Gnirke et al., 2009). Because such baits need to be designed in silico, which requires genomic resources, this is both time-intense and bioinformatically demanding, particularly in nonmodel species. In addition, commercial bait synthesis is very expensive, especially for large sample sizes.

Protocols for home-made baits derived from RNA, DNA- or exome-based RAD libraries try to address these issues (e.g., hybridization RADseq or hyRAD, and exome-based hyRAD-x; Suchan et al., 2016; Schmid et al., 2017; Sánchez Barreiro et al., 2017; Linck, Hanna, Sellas, & Dumbacher, 2017), but do not explicitly address the challenge of combining modern and historical samples at large scale for joint population genetics analyses. Furthermore, current protocols depend on enzymatic removal of sequencing adapters from bait-pools to avoid mix-ups between baits and sequencing libraries. They produce only a limited, and as result of adapter-removal not amplifiable amount of bait, and rely on commercial kits for bait biotinylation (Suchan et al., 2016). Here, we present extensive modifications of current hyRAD protocols and a combined ddRAD–hyRAD approach that allows standardized generation of reduced-representation sequencing data with population-scale historical and modern plant specimens. Using parallel processing of ddRAD libraries and hyRAD baits with individual

adapter pairs, we produce highly overlapping modern and historical fragment libraries for joint analyses (Fu et al., 2013; Slon et al., 2017). Their specific adapters "immortalize" our baits for unlimited amplifications and captures of libraries, while requiring minimal input DNA during primary bait production. Biotinylation based on a biotinylated primer and linear amplification of bait libraries keeps costs at a minimum, while simultaneously increasing the diversity of our captures.

With the approach described here, sequence data generation and subsequent analyses do not depend on the presence of a published reference genome, as the use of a customized bioinformatic pipeline allows a largely identical processing of historical and modern sequencing data, and reference-independent de novo discovery of polymorphic sites across both data types. To evaluate this strategy, we compare our ddRAD and hyRAD-based data to a "gold standard" of whole-genome shotgun sequencing data mapped to a reference genome and show that the method can faithfully recapitulate known genetic relationships in a geographically broad set of historical and modern *A. thaliana* samples. Using three different bait pools based on genetically distinct *A. thaliana* populations, we also show that recapitulation of this genetic diversity is independent of the geographical origin and thus of genetic relatedness of the baits with the captured historical samples. As a proof-of-principle, we then analyse historical and modern *Cardamine bulbifera* specimens, a nonmodel species that lacks a reference genome, and identify genetic variation that recapitulates the geographical and temporal distribution of the investigated samples.

2 | METHODS

2.1 | Fresh plant samples

Arabidopsis thaliana seeds of the North American HPG1 lineage (H2081 and H1943) and two Moroccan accessions (Arb-0, Elh-2) were surface sterilized with 10% bleach, 0.5% sodium dodecylsulphate (SDS) and stratified for 2 days at 4°C. Plants were grown at 16°C or 23°C in soil under either short-day (8 hr light/16 hr dark) or long-day conditions (16 hr light/8 hr dark) in growth chambers with 65% humidity. A mixture of Cool White and Gro-Lux Wide Spectrum fluorescent lights with a fluence rate of 125–175 µmol/m^2 s^{-1} was used. HPG1 and Moroccan accessions have been described and were obtained from colleagues (Table S1; Durvasula et al., 2017; Platt et al., 2010). For DNA extraction, leaves of six single plants per accession were collected. Leaves of flowering specimens of *Cardamine bulbifera* were sampled in forest plots of the southern (Schwäbische Alb) and central (Hainich) German biodiversity exploratory (www.biodiversity-exploratories.de) (Table S2; Fischer et al., 2010), and kept on ice or at 4°C for a maximum of 2 weeks until transportation back to the laboratory. Samples for DNA extraction were kept at −80°C until further use.

Frozen plant tissue was thoroughly ground using two metal beads (KGM, Brammer) per sample and a TissueLyser II (Qiagen).

Because incomplete grinding was a major factor limiting extraction efficiency, samples were ground in several (five or six) rounds (1 min, 20 s^{-1}), including re-freezing in-between rounds (>15 min at −80°C). Extractions were performed using CTAB. In brief, DNA was extracted with preheated CTAB (NaCl 1.4 m, Tris pH 8 10 mm, EDTA 2 mm, CTAB 2%, PVP 1% and freshly added beta-mercaptoethanol 0.2% v/v), subsequent phase separation with chloroform/isoamylalcohol (24:1), precipitation with isopropanol and final washing with EtOH 70%. DNA concentrations of eluted samples (buffer: Tris-HCl pH 9 10 mm, EDTA 0.5 mm, with 0.5 µl RNAse A per sample) were measured with the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific) and an Infinite M200 Pro plate reader (TECAN). DNA was stored at −20°C or −80°C until further use.

For a detailed version of the protocol, see Supporting Information and Table S3.

2.2 | Herbarium samples

Ancient DNA libraries of *A. thaliana* lineages and shotgun sequencing data for these libraries have been published (PRJEB19780 and PRJEB15366; Durvasula et al., 2017; Gutaker, Reiter, Furtwängler, Schuenemann, & Burbano, 2017). Previously prepared *A. thaliana* aDNA libraries were PCR-amplified using primers IS5 and IS6 (Meyer & Kircher, 2010) to obtain ~1 µg input per capture reaction (Table S1).

Herbarium specimens of *C. bulbifera* collected between 1798 and 1995 were sampled at, and with the kind permission of, the herbaria of Jena, Stuttgart and Tübingen, Germany (Table S2). We conducted sampling as minimally destructive as possible, collecting a maximum of ~1 cm^2 of leaf tissue, preferably of leaves that were either already damaged, or of leaves hidden at the specimens' back, to preserve overall specimen morphology and phenotype. Each sampled specimen was photographed in its entirety (see Figure 5b), and a note with contact information and the purpose of the sampling was attached to the sampled sheets to enable tracking of the samples. Until further use, samples were kept in tubes and stored in boxes with silica gel to reduce humidity.

Historical *C. bulbifera* samples were extracted in a cleanroom at the University of Tübingen as published previously (Gutaker et al., 2017). Briefly, tissue was ground and incubated in PTB lysis buffer at 37°C overnight. After transfer of the solution to a QIAShredder column, extraction mainly followed the DNEasy kit (Qiagen) protocol (Gutaker et al., 2017). Single-stranded DNA libraries were constructed as published (Gansauge et al., 2017), employing a Bravo Automated Liquid Handling Platform (Agilent; Slon et al., 2017) and using 10 µl of DNA extract as input. In brief, library preparation encompassed dephosphorylation and heat denaturation of the sample DNA, ligation of biotinylated adapters to the 3' ends of the single-stranded molecules and their immobilization on streptavidin-coated magnetic beads. Second strand synthesis and the ligation of the second adapter were performed on solid support before the final library was recovered from the beads by heat denaturation.

2.3 | Flow cytometry

We collected plant and leaf samples of multiple *C. bulbifera* individuals at the Tübingen Botanical Garden and sent them to Plant Cytometry Services (J. G. Schijndel, The Netherlands) for genome size estimation. *Vinca major* and *Ophiogon planiscapus* "Nigrescens" were used as internal standards, and flow cytometric measurements were conducted at two instances, for a total of five individuals. Gbp/1C was calculated from pg/2C using a conversion factor of 1 pg = 978 Mbp and dividing the resulting value by 2, resulting in an estimated genome size reported in Table S4 (Doležel, Bartoš, Voglmayr, & Greilhuber, 2003). Genome ploidy has been estimated to be up to 12× (Carlsen, Bleeker, Hurka, Elven, & Brochmann, 2009; Kučera, Valko, & Marhold, 2005).

2.4 | ddRAD library and bait generation

ddRAD libraries were prepared using a modified and optimized version of previously published protocols (Meyer & Kircher, 2010;

Peterson et al., 2012; Suchan et al., 2016). Major differences to the published hyRAD protocols included the parallel generation of ddRAD libraries and digestion-based capture baits, biotinylation of the home-made baits with a 5′-biotinylated primer through linear amplification (Fu et al., 2013), double-indexing of fresh tissue libraries (Kircher, Sawyer, & Meyer, 2012), and the use of different adapter sequences for libraries and baits (Figures 1 and 2; Figure S2). Deviating from the hyRAD method of Suchan et al. (2016), bait-adapters are not enzymatically removed from the baits, which allows a nearly unlimited production of baits through re-amplification of the bait library. Also, the use of different adapters for libraries and baits allows for their specific amplification even after both have been mixed for capture, and prevents baits from being sequenced, which may occur when enzymatic adapter-removal is incomplete.

Briefly, for bait generation, we selected 10 freshly collected samples per species, with the samples covering the extremes of our geographical sampling to maximize the genetic diversity represented in the baits. All samples were processed individually until pooling for size selection. For library and bait samples alike, we digested ~200 ng input DNA per sample with *Eco*RI (methylation sensitive, Thermo
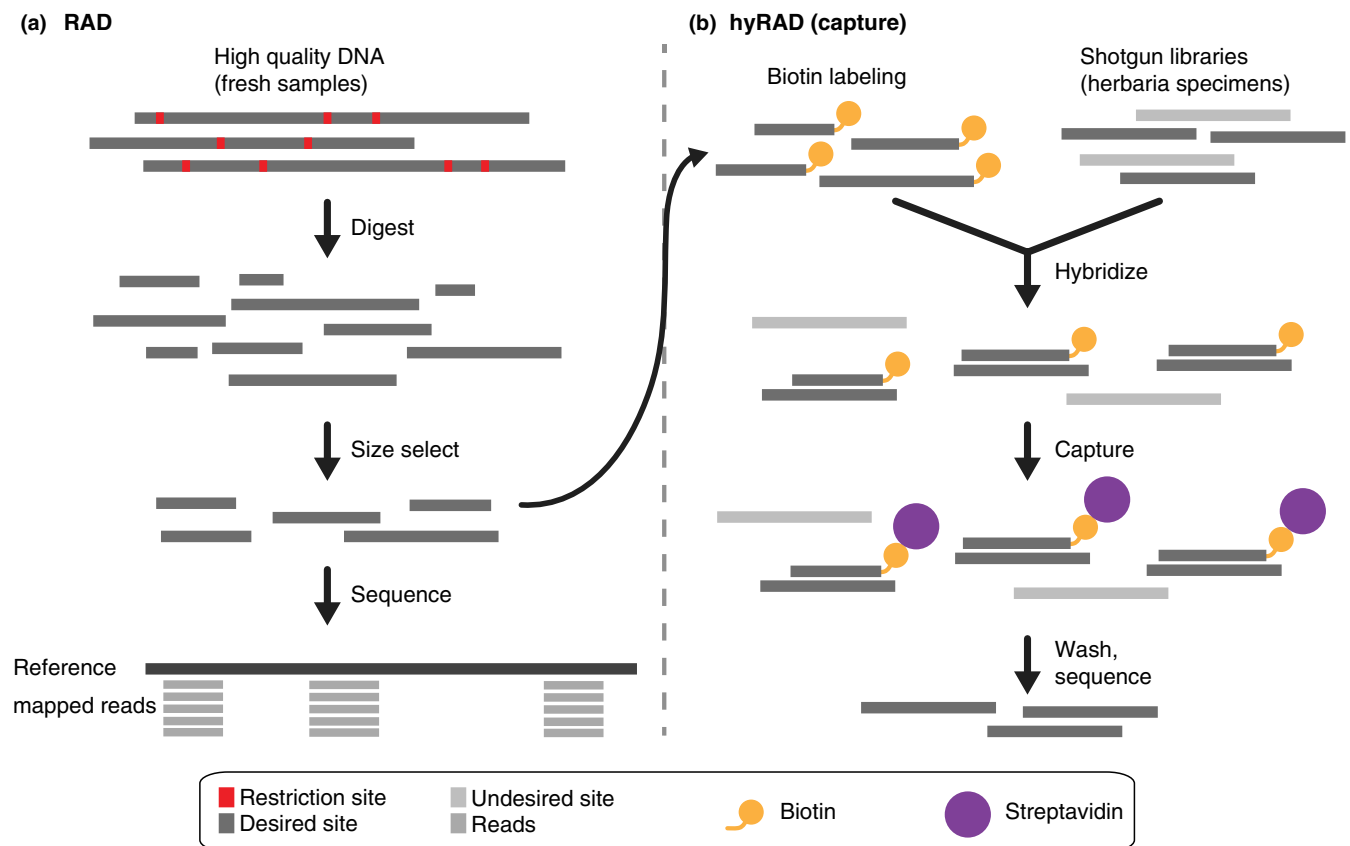
FIGURE 1 RAD and hyRAD method overview. (a) RAD: restriction enzyme(s) (one for RAD, two for double-digest RAD, ddRAD) cut the DNA. Prior to sequencing, the fraction of the genome that will be part of the RAD library is reduced using size-selection, reducing the complexity of the library (reduced-representation method). The sequenced fragments cover a fraction of the genome at high coverage and quality. (b) hyRAD: after digestion and size-selection of the fresh DNA, a subset of samples are processed to become baits for capture of ancient DNA (aDNA) libraries. They are biotinylated and mixed with aDNA libraries for sequence similarity-based hybridization. Streptavidin, through strong affinity to biotin, captures the hybridized double-strands. Nonhybridized library fragments are washed off, and the targeted fraction is sequenced

Fisher, FD0274) and *Nsi*I/*Mph*1103I (Thermo Fisher, FD0734; 37°C, 3 hr) and ligated double-stranded custom-adapters to the fragments' "sticky" restriction ends (restriction sites 5′–3′: *Eco*RI, G'AATTC; *Nsi*I, ATGCA'T). The adapter sequences contained primer sequences specific for either the library or the bait samples, to allow their independent amplification when pooled. In addition, we generated four different pairs of adapters, containing between zero and three additional base-pairs between the generic adapter sequence (where the sequencing primer binds) and the restriction site, which we call shift-bases (see Supporting Information and Figure S2a for details). Addition of these shift-bases avoids problems with base calling during the sequencing of the ddRAD libraries, which always start with the same nucleotides (the restriction sites). We thus ligated one-quarter of all ddRAD libraries with one of the four (shifted) adapter-pairings each.

Before and after adapter ligation, homemade magnetic SPRI-beads (Rohland & Reich, 2012) were used to clean samples and remove fragments above ~500 bp in length. Libraries were amplified and double-indexed via PCR, using Nextera-based primers and unique index combinations for each sample (Kircher et al., 2012). P5 and P7 indexing primers were designed for hybridization to the restriction-site-independent parts of the adapters that ligate to the *Eco*RI/*Nsi*I-based sticky ends (see Table S5, Figure S2a), respectively. This ensures exclusive indexing and amplification of fragments with one *Eco*RI and one *Nsi*I cutting site. In parallel, bait samples were amplified with APL5 and APL6, to keep sample concentrations at similar levels (Figure S2b).

For size selection, library and bait samples are ideally run as one pool in one single lane of a Blue Pippin (Sage Science). Therefore, based on Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific) DNA quantifications, library and bait samples were pooled at equal concentrations. Subsequently, the pool volume was reduced and cleaned via column purification (EconoSpin, Epoch Life Science). With a Blue Pippin (Sage Science), the pool was restricted to fragment sizes (including adapters and indices) of 300–500 bp. To disentangle libraries for sequencing and baits for hybridization capture, we amplified fractions of the pool with primers specific for either the library (IS5 and IS6, Table S5; Meyer & Kircher, 2010) or

FIGURE 2 Method overview for ddRAD and parallel bait preparation. (1) Fresh high-quality DNA samples are digested with two restriction enzymes (*Eco*RI and *Nsi*I/Mph1103I). Depending on whether the samples will be (2) sequenced or (3) become baits for hybridization capture, different sets of double-stranded adapters are ligated to the sticky ends of the fragments (IS7- and IS8-based for sequencing, APL5- and APL6-based for baits). In parallel with (4) double-indexing of the sequencing libraries (P5 + P7 primers), (5) baits get to similar concentration levels through amplification with APL5 and APL6. Either amplification works only on fragments that have one of each restriction site, and not on fragments that have the same site at both their 5′ and 3′ end. (6) To avoid biases introduced with size selection of individual library and bait pools, all libraries and baits are combined into one pool for size selection. Subsequently, the specific adapters allow disentangling of the different sample types using PCR—IS5 and IS6 for libraries (7), APL5 and APL6 for baits (8). While libraries are now ready for sequencing, (9) baits are amplified further (APL5 and APL6), to reach high concentrations for the final linear amplification with APL2, a biotinylated version of APL5. After this, the biotinylated baits can be used for hybridization capture
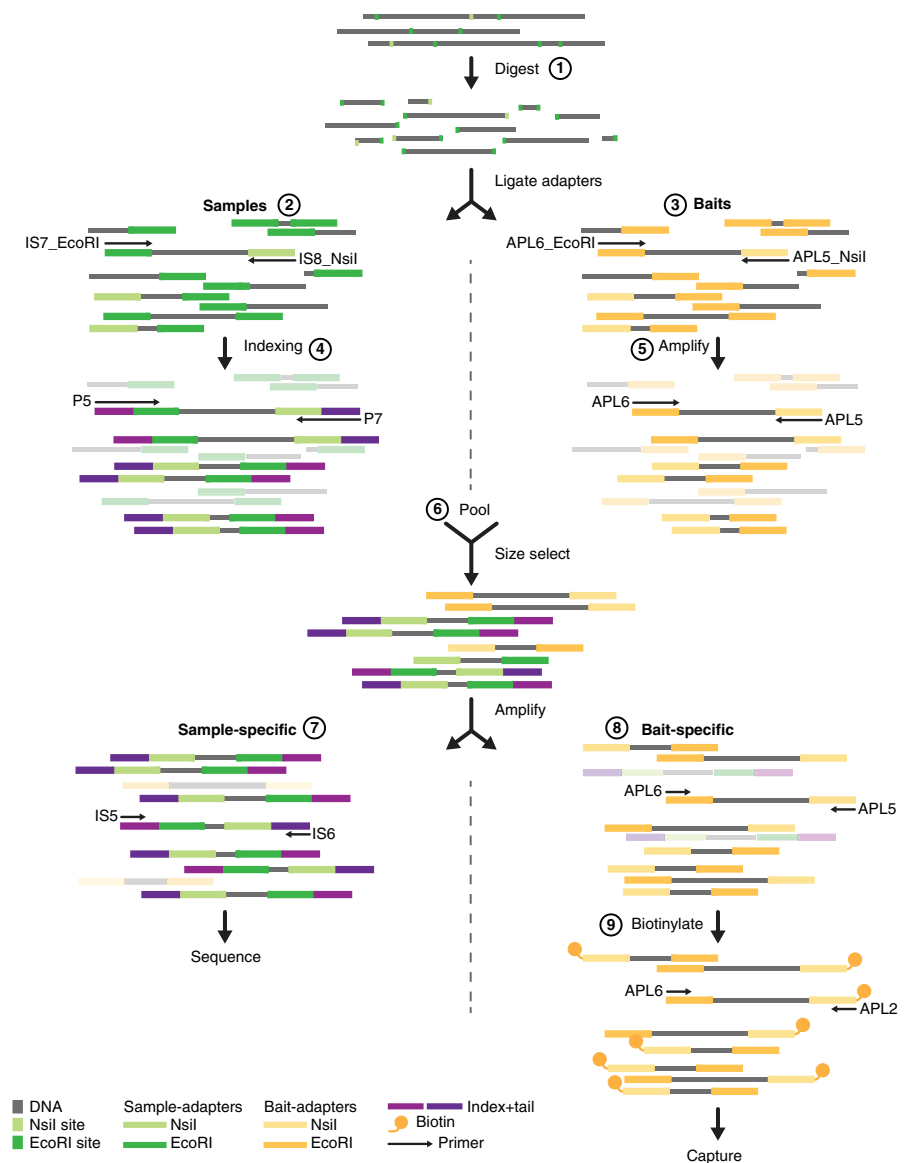
the bait adapters (APL5 and APL6, Table S5; Fu et al., 2013). Final library pools were sequenced, alone or pooled with libraries from other projects, in paired-end 150-bp runs on a HiSeq 3000 (Illumina) at the MPI for Developmental Biology in Tübingen, Germany. Bait pools were stored at −20°C until further use.

For a detailed version of the protocol, optimized for large sample sizes, see Supporting Information and Table S3.

2.4.1 | Bait generation

Bait generation is a two-step process of regular exponential PCR amplification followed by a linear, single-primer biotinylation reaction, starting with ~8 ng bait pool, and then using ~200 ng PCR product from the first amplification reaction. Depending on the number of samples to be captured (i.e., the final amount of baits needed), we ran multiple reactions of each step in parallel (see detailed protocol, Table S3 and online, for expected yields and calculations). The volume and concentration of the originally obtained bait pool in principle do not limit the amount of bait that can be generated, as the bait pool—both before and after the first amplification reaction, but always before biotinylation—can be amplified almost indefinitely using APL2/5 and APL6 (Table S5, Figure S2b; Fu et al., 2013). After the first amplification, PCRs were pooled and cleaned with the MinElute PCR purification kit (Qiagen), and concentrations were measured using a Nanodrop. Subsequent linear amplification with the 5′-biotinylated APL2 primer and SPRI-bead based cleanup of the pooled reactions results in the final biotinylated baits. The primer-mediated biotinylation—as opposed to insertion of biotinylated nucleotides with a nick-translation-based commercial kit—is cheap and easy. In addition, the linear PCR enriches specifically for one strand only, leading to improved capture efficiency.

2.4.2 | *Arabidopsis thaliana* pilot baits

To compare bait libraries generated with plants of different genetic diversity levels for the *A. thaliana* capture pilot experiment, namely of low (US HPG1 lineage) or high (African accessions) genetic diversity, fresh sample libraries were produced in technical replicates to obtain sufficient amounts of DNA. We pooled technical replicates for each sample, measured the concentration of those pools, and equimolarly joined bait libraries for the HPG1 lineage or for the Moroccan accessions to generate the separate low- and high-diversity pools. Each bait pool was combined with a volume of the library pool and cleaned via column purification (EconoSpin, Epoch Life Science). The combined library–bait pools were then run in parallel in one of two Blue Pippin lanes. After size-selection, we amplified the pools for five or eight cycles with library- or bait-specific primers in four replicates each. We combined the libraries for sequencing equimolarly, whereas further bait amplification was done separately for the US (pUS) or Moroccan (pMA) pool. In addition to the US low-, and African high-diversity bait pool, we mixed both at equal volumes to generate a third bait pool (pMix, Figure 4a).

For a detailed version of the protocol, see the Supporting Information and Table S3.

2.5 | Hybridization RADseq

To capture double-indexed historical libraries (single-strand libraries for *C. bulbifera*, double-strand libraries for *A. thaliana*; Gutaker et al., 2017), we used ~1 µg of input library per sample and a hybridization capture protocol adapted from Fu et al. (2013). In brief, after heat denaturation, blocking of the library-specific adapter sequences using blocking oligos was done to prevent rehybridization of the library double strands, which would otherwise reduce the specificity of the capture reaction through the formation of daisy chains between target and nontarget library molecules. Libraries and baits (~500 ng per sample) were then mixed and incubated for 24 hr (up to 72 hr) at 65°C. Hybridized library–bait duplexes were then immobilized on streptavidin beads, and free library molecules washed off over multiple steps. Incubation in NaOH-based melt solution dissociated the nonbiotinylated library strands, which were then precipitated and bound to magnetic SPRI beads, washed and eluted. For qPCR of the capture eluate, we compared the concentration of a 1:10 dilution of this final eluate to a home-made standard dilution series using qPCR with Illumina-specific IS7 and IS8 primers (Meyer & Kircher, 2010). Enriched libraries were then amplified (IS5 and IS6, Table S5, Figure S2c; Meyer & Kircher, 2010), cleaned and pooled at equal volumes for sequencing.

Because the individual hybridization captures for the three different bait sets (pUS, pMA, pMix, Figure 4) and the two capture rounds in *A. thaliana* were all based on the same double-stranded aDNA libraries and hence had identical indices, each of the six captures was sequenced in ~10% of different flow-cell lanes. The single-strand-based aDNA *C. bulbifera* library captures were sequenced in entire lanes, supplying the single-strand library-specific shorter second sequencing primer (CL72, ACACTCTTTCCCTACACGACGCTCTTCC; Gansauge & Meyer, 2013) in the respective lane of the HiSeq 3000 flow-cell. The first *C. bulbifera* capture was sequenced in a paired-end 75-bp run at the MPI for Evolutionary Anthropology in Leipzig, whereas all other sequencing for both *A. thaliana* and *C. bulbifera* libraries was conducted in paired-end 150-bp runs at the MPI for Developmental Biology in Tübingen, Germany.

2.6 | Sequencing data processing

Unless mentioned otherwise, all software was used with default options.

2.6.1 | Fresh samples

After demultiplexing, sequences of fresh ddRAD samples for both *C. bulbifera* and *A. thaliana* were trimmed for adapters and shift-bases

(see sequences below, and Table S5) with CUTADAPT version 1.12 (Martin, 2011). While adapter-trimming was sequence-based, shift-bases were removed only using the information of how many bases were added. Due to different numbers of shift-bases in the fragments' 5' and 3' ends (between 0 and 3, Figure S2), those bases were trimmed in individual steps for the forward and reverse read ("*cutadapt --cut [#bases_fwd] -o [read_cut_R1_.fastq.gz] [read_R1.fastq.gz]*" and "*cutadapt --cut [#bases_rev] -o [read_cut_R2_.fastq.gz] [read_R2.fastq.gz]*"), before trimming of 5' low-quality bases and adapter sequences ("*cutadapt -q 15 -b TAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -b TGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -b TGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -b TGCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -B CAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -B CGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -B CGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -B CACTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --trim-n --minimum-length 35 -o read_cutadapt_R1_.fastq.gz --paired-output read_cutadapt_R2_.fastq.gz read_cut_R1_.fastq.gz read_cut_R2_.fastq.gz*") and quality-control using FASTQC version 0.11.5 (Andrews 2010). FASTQC, a quality control tool for high-throughput sequence data, is available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc). After merging data from independent sequencing runs, ddRAD-related restriction sites at the fragment ends were removed with CUTADAPT version 1.12 (Martin, 2011), and paired-end reads were merged using FLASH v1.2.11 ("extended --max-overlap = 100"; Magoč & Salzberg, 2011). Merged and remaining unmerged reads of all fresh samples were then used to build a pseudo-reference with MEGAHIT version 1.1.3 ("*megahit -r [merged] −1 [unmerged_fwd] −2 [unmerged_rvs] -m 400,000,000,000 --num-cpu-threads 40 --min-contig-len 50*"; Li, Liu, Luo, Sadakane, & Lam, 2015; Li & Luo, 2016). Removal of restriction sites prior to de novo assembly of the sequenced regions around the restriction sites resulted in better mapping quality of reads against the assembly, and inclusion of the unmerged read fraction reduced the mapping error.

We then independently mapped merged and remaining unmerged reads to the corresponding MEGAHIT reference (BWA MEM 0.7.15-r1142-dirty; Li, 2013), subsequently combining the resulting. bam-files for each sample, and finally for all samples, generating a multi-bam for downstream analyses (SAMTOOLS version 1.4.1, SAMTOOLSmerge; Li et al., 2009). Mapping statistics were assessed based on SAMTOOLS stats (input bp, mapped bp, mapping error) of those combined files, whereas sizes of mapped fragments were retrieved individually, either as fragment sizes (merged reads, SAMTOOLS view) or insert sizes (unmerged, i.e., paired reads, SAMTOOLS stats).

A. thaliana shotgun sequencing data for fresh samples of the accessions Arb-0, Elh-2 (Morocco) and Tanz-1 (Tanzania) were downloaded from the European Nucleotide Sequence Archive (ENA, study PRJEB19780, samples ERS1575068 [Arb-0], ERS1575074 [Elh-2], ERS1575132 [Tanz-1]; Durvasula et al., 2017), while reads for HPG1-2081 (North America) were provided by G. Shirsekar (personal communication, Table S1). Forward and reverse reads for the samples were merged using FLASH, mapped independently to TAIR10 (Berardini et al., 2015), combining mappings of merged and unmerged

reads afterwards in one file per sample, and into a final multi-bam for all samples. Overview analyses were done as described above.

2.6.2 | Historical samples

Raw reads of the first *C. bulbifera* capture sequenced in Leipzig were reformatted from bam to fastq using BEDTOOLS version 2.28.0 (Quinlan & Hall, 2010). With ADAPTERREMOVAL version 2.2.1a, we then trimmed adapters and merged forward and reverse reads for all *C. bulbifera* and *A. thaliana* historical sequencing (sslibrary: "*AdapterRemoval --file1 R1_.fastq.gz --file2 R2_fastq.gz --basename [samplename] --gzip --adapter2 GGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT --collapse --minlength 30*", dslibrary: "*AdapterRemoval --file1 R1_.fastq.gz --file2 R2_.fastq.gz --basename [samplename] --gzip --collapse --minlength 30*"; Schubert, Lindgreen, & Orlando, 2016). The resulting files were mapped to the MEGAHIT reference as described above for the fresh samples, as well as to either TAIR10 or the *C. hirsuta* reference genome (BWA MEM 0.7.15-r1142-dirty; Berardini et al., 2015; Gan et al., 2016; Li, 2013), and cleaned of PCR duplicates with DEDUP version 0.12.0 (Peltzer et al., 2016), with mapping statistics assessed by SAMTOOLS before and after deduplication for quality control. For subsequent analyses, all mapped files were combined into a single multi-bam (SAMTOOLS version 1.4.1; H. Li et al., 2009).

To authenticate the historical nature of the DNA retrieved from herbarium specimens, we investigated the aDNA-associated patterns of deamination and fragmentation (Figure S4; Briggs et al., 2007; Weiß et al., 2016). Deamination of cytosines (C) to uracils, recognized as thymines (T) in the sequencing process and hence reported as the ratio or fraction of C-to-T changes, was assessed with MAPDAMAGE version 2.0.8 ("*mapDamage -i sample_dedup.bam --merge-reference-sequences -r megahit_reference.fa*"; Jónsson, Ginolhac, Schubert, Johnson, & Orlando, 2013). Fragmentation patterns of merged reads were, as described above, determined with SAMTOOLS.

A. thaliana historical shotgun sequences for samples from Africa (AH0011 [Algeria], AH0004 and AH0006 [South Africa], AH0007 and AH0008 [Tanzania]) and North America (HB0001, 3, 5, 7, 9; Table S2) were downloaded from ENA (African samples: study PRJEB19780, accession nos. ERS1575137 [AH004], ERS1575138 [AH006], ERS1575139 [AH007], ERS1575140 [AH008], ERS1575142 [AH011], Durvasula et al., 2017; NA samples: study PRJEB15366, accession nos. ERS1342420 [HB0001], ERS1342418 [HB0003], ERS1342416 [HB0005], ERS1342414 [HB0007], ERS1342412 [HB0009], Gutaker et al., 2017). Reads of these samples were merged, mapped to TAIR, deduplicated and authenticated as described above.

2.7 | Evaluation of captures and bait types

For biological samples with very low DNA contents, such as highly degraded historical samples, two subsequent captures can increase the amount of retrieved sample-specific DNA (Avila-Arcos et al., 2011). To assess the efficiency of two versus one capture, we performed

(a) *Arabidopsis thaliana*

(b) *Cardamine bulbifera*
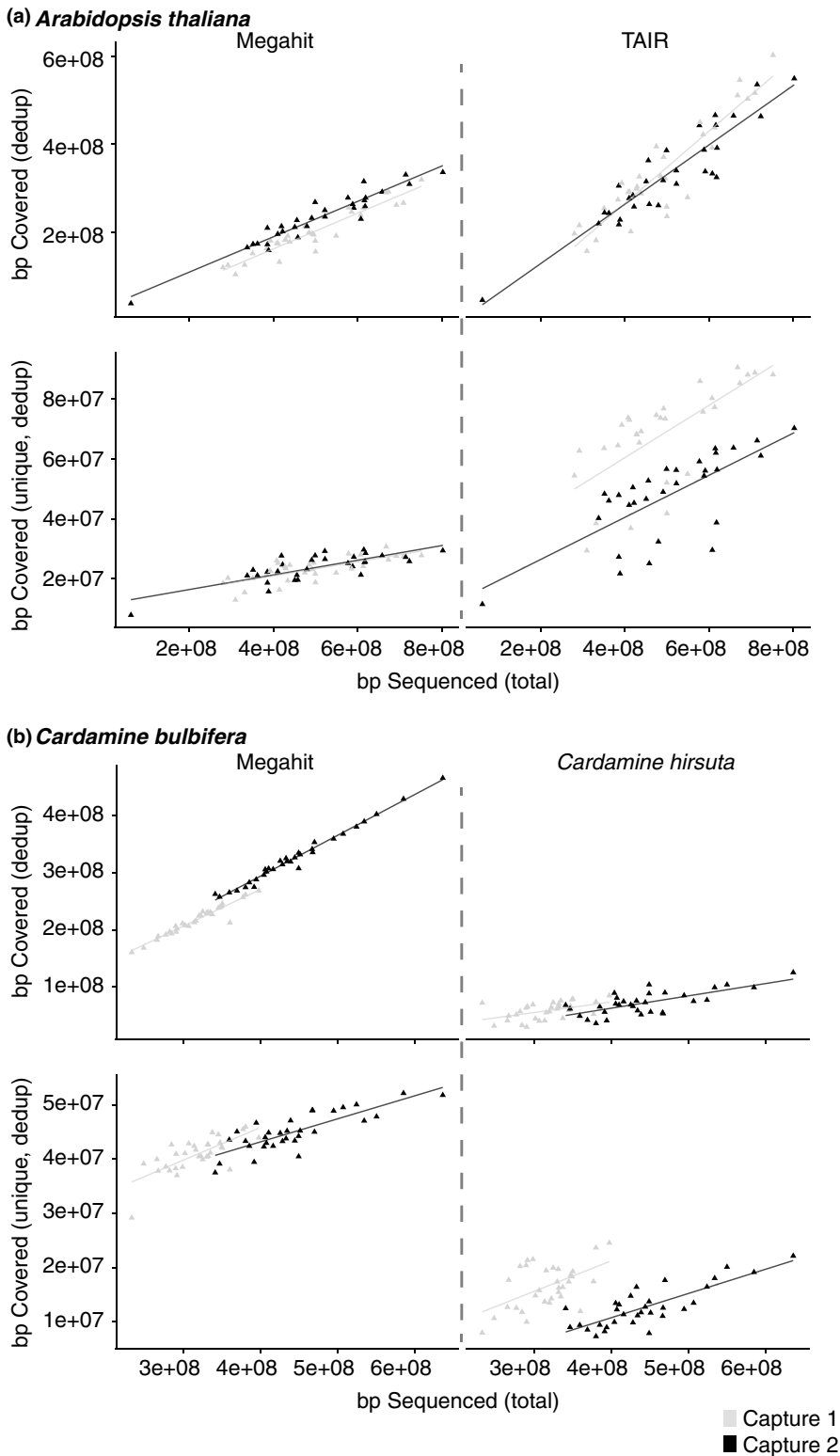
FIGURE 3 Efficiency of a single versus two subsequent rounds of hyRAD captures. Comparison of total and unique base pairs covered per base pair sequenced, mapping aDNA sequences against either a published whole-genome reference or the ddRAD-based MEGAHIT assembly. (a) *Arabidopsis thaliana* samples, mapped against the MEGAHIT reference or *A. thaliana* reference genome TAIR10 (Berardini et al., 2015), and (b) *Cardamine bulbifera* samples, mapped against the MEGAHIT reference or the closest published reference, the *Cardamine hirsuta* genome (Gan et al., 2016)

subsequent captures for the entire sample sets of both of our species, *A. thaliana* and *C. bulbifera*. All retrieved historical sequences were trimmed, merged and mapped as described above, and their historical authenticity was evaluated (Figure S4). For each sample, we determined the overall sequencing effort (bp sequenced) with SAMTOOLS stats (SAMTOOLS version 1.4.1; Li et al., 2009) and the genome-wide coverage depth using BEDTOOLS GENOMECOV version 2.26.0

("*bedtools genomecov -bga -ibam [file.bam]> [name outfile]*"; Quinlan & Hall, 2010). Based on this, we then used R (see below) to calculate the total coverage, as well as unique coverage in base pairs, and to plot both values in relation to the total sequencing effort (Figure 3).

To compare bait-sets of variable genetic diversity, and their ability to capture genetic diversity, we captured the same historical samples with three different bait sets, based on either fresh Moroccan

FIGURE 4 *Arabidopsis thaliana* pilot capture. (a) Pilot design with biotinylated baits made from fresh tissue of US (blue) and Moroccan plants (MA, red) that are used to capture ancient DNA libraries from the same geographical locations. For each library, three captures were performed, with either the geographically corresponding baits, a mix of both bait types, or the opposite baits. Sample clustering based on pairwise genetic distances, for (b) previously published historical and fresh whole genome shotgun sequence samples mapped to TAIR (Durvasula et al., 2017; G. Shirsekar, personal communication), and (c) fresh ddRAD and historical hyRAD samples produced in this study and mapped to MEGAHIT, which are recoloured in (d) based on the bait-type that was used for capture (pUS, pMA or pMix). Sample sets were filtered before clustering for data completeness (m75/95, indicating 25% or 5% missing data, respectively), the minimum in-sample a variant must have within a sample to be considered (a6, i.e., 60%), and for the minimum required number of samples representing the minor allele (f2/3, indicating two or three samples, respectively)
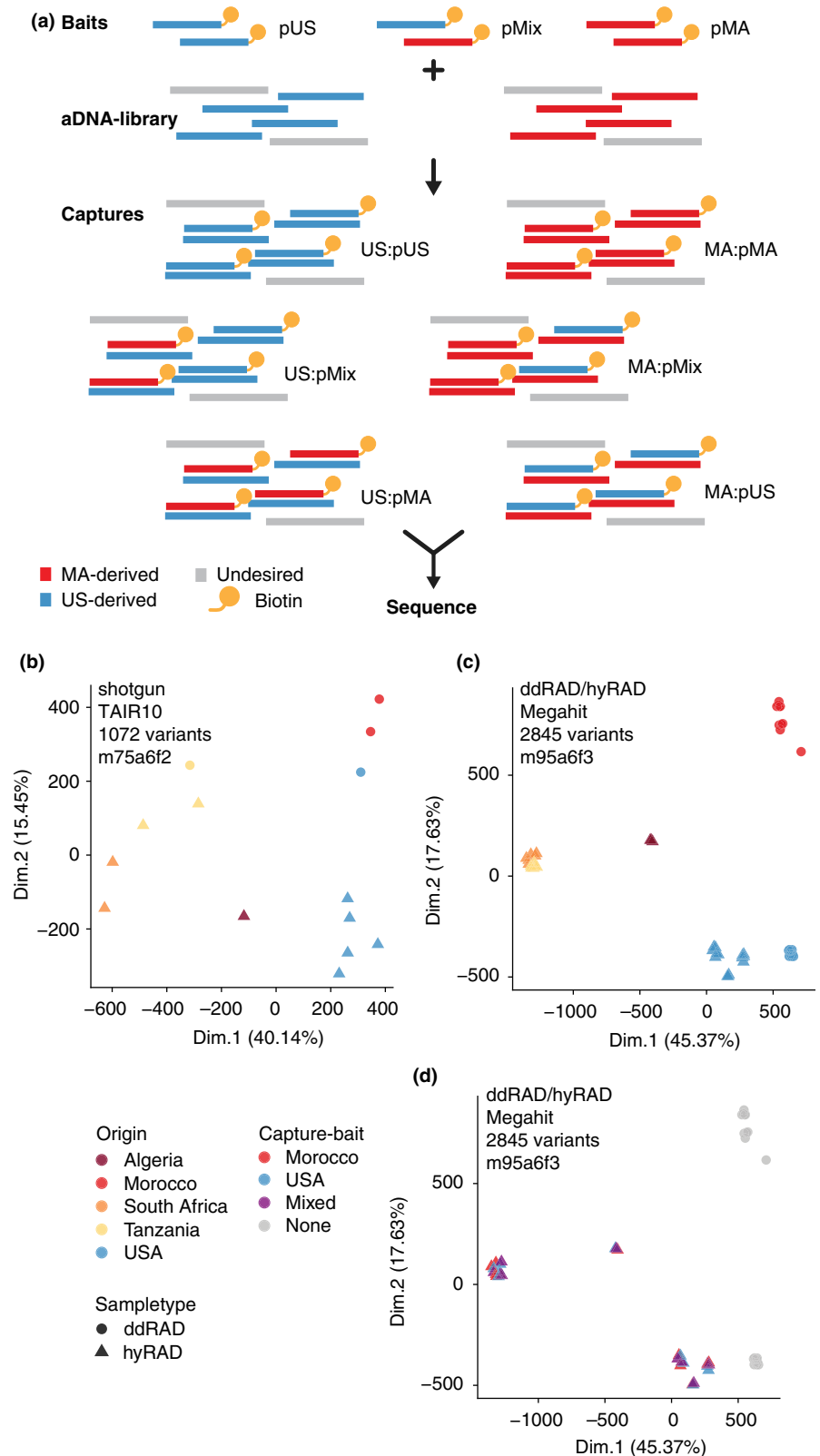
accessions (pMA), fresh HPG1 (USA) accessions (pUS), or a mix of the two (pMix, Figure 4a). Based on sites discovered using BSH-DENOVO (https://github.com/clwgg/bsh-denovo, rev. 30c95ab) on the entire set of historical hyRAD samples, filtered as described below, we calculated Identity-by-State distances using PLINK version 1.90b5.3

("*plink --memory 32,000 --file [name of filtered map/ped] --distance square ibs allele-ct --out [name outputfile]*"; Purcell et al., 2007; Chang et al., 2015). Focusing on the two main clusters of historical samples (i.e., excluding the Algerian sample [AH0011]), we then grouped these genetic distances for samples captured with the same bait, examining
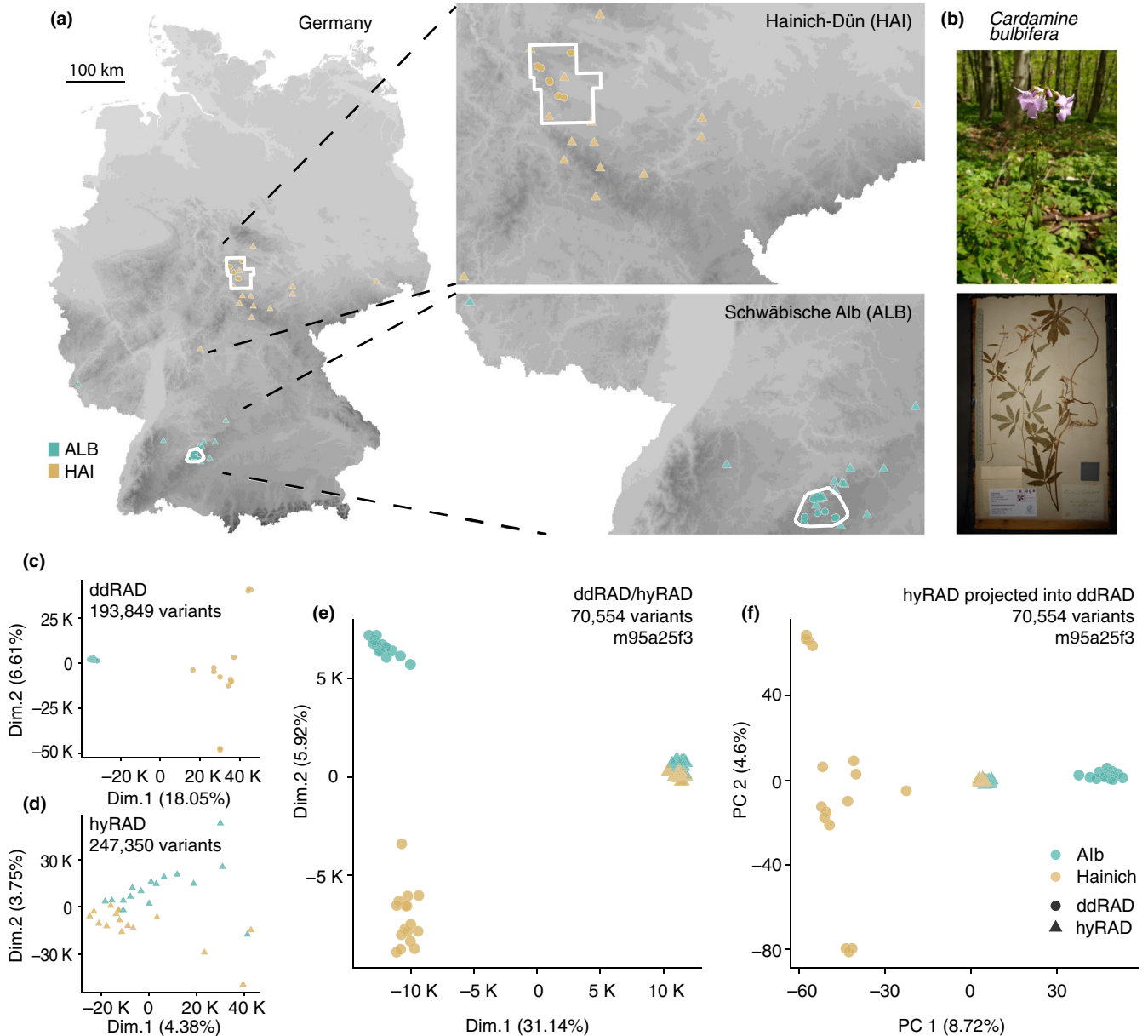
FIGURE 5 ddRAD and hyRAD with the nonmodel species *Cardamine bulbifera*. (a) Overview and zoomed maps of Germany showing the geographical origin of samples; circles represent contemporary samples, triangles historical samples, turquoise colour for samples from Schwäbische Alb, beige from Hainich, and exploratory circumferences are marked by white lines. (b) Contemporary and historical *C. bulbifera* plants at the reproductive stage. (c) MDS of fresh and (d) historical samples, separately and (e) combined. (f) PCA of fresh and historical samples, with historical samples projected into the modern PC space. Sample sets were filtered before clustering for data completeness (m95, indicating 5% missing data), the minimum frequency a variant must have within a sample to be considered (a25, i.e., 25%), and for the minimum required number of samples representing the minor allele (f3, indicating three samples)

sample distances within the African and the North American cluster, and between both clusters. Neither distances within nor between clusters varied significantly for the different bait sets (Figure S3).

2.8 | Analysis of genetic distances

The lack of a reference genome when working with nonmodel species such as *C. bulbifera* complicates reliable calling of genetic variation, and thus population genetics analyses. We do not have, besides

our own sequencing, any data detailing the genetic diversity of the species. Therefore, we were unable to estimate to what extent our sampling and the assembly we made based on our data represent the true genetic diversity of the species. This applies both to geographical genetic variation, but also to temporal variation—the latter being true also for sequenced model organisms, where reference genomes traditionally are generated using sequencing of contemporary specimens. To avoid any such bias of the genetic diversity that we retrieved and analysed based on our MEGAHIT-generated reference, we used BSH-DENOVO for variant discovery (https://github.

com/clwgg/bsh-denovo, rev. 30c95ab). BSH-DENOVO "discovers" variable sites solely based on the samples' reads. The reference used for mapping only provides a common coordinate system to align the reads, but is not taken into account for variant discovery. After identification of variable sites, one base per sample and site is randomly sampled (pseudo-haploidization). This approach has been successfully used in aDNA to estimate relatedness between samples from low-coverage data (Green et al., 2010; Malaspinas et al., 2014). We adjusted variant discovery to the size and type of our data sets: For the diploid selfer *A. thaliana* with low expected heterozygosity, a base was required to have a frequency of at least 0.6 within a sample for the site to be considered ("*-a 0.6*"), thus excluding sites where low heterozygosity and sequencing error can be confounded. In contrast, for *C. bulbifera*, dodecaploid and reproducing mostly vegetatively, site discovery was extended with a required base frequency of at least 0.25, which takes the plant's expected higher heterozygosity into account. Depending on the number of samples present in a data set, we further filtered for a minimum minor-allele count of at least 2 or 3 ("*-f 2*" or "*-f 3*"), and required data completeness to be 75% or 95% for a site to be considered ("*-m 0.75*" or "*-m 0.95*"; full command: "*bsh-denovo -o [name outfile] -m 0.95 -a 0.6 -f 3 [input_multi.bam]*"). Only when all filters (-a, -m, -f) are passed do we identify a site as polymorphic across samples. Subsequently, a base is sampled randomly from all bases observed in a given sample. We chose to use minor-allele count cutoffs instead of minor-allele frequencies to control for sample size inequalities, as the latter could result in biases due to different site missingness across data sets (Linck & Battey, 2019).

The resulting .map and .ped files were then filtered with PLINK, removing tri- and quadrallelic positions. For combined data sets, we then created two separate files for either only modern or only historical data, filtering both separately for 95% (or 75%) full information per site to avoid biases in missingness towards either of the two data sets. Afterwards, we remerged the filtered sets ("*plink --file [filtered modern .map/.ped] --merge [filtered historical .map/.ped] --recode ped --out [name merged outfile]*"), and again filtered for 95% (or 75%) full information per site as well as for sites with at least three (or two) individuals carrying the minor allele, to avoid the inclusion of sites filtered out only in one of the data sets, because such sites will artifactually increase the number of differences between historical and modern samples ("*plink --file [name of .map/.ped] --mac 3 --geno 0.05 --recode ped --out [name outfile]*"). Analysing potential biases between the modern and historical *C. bulbifera* data set, after general filtering, we filtered again to retain only variable sites with full information in all samples ("*plink --file [name of .map/.ped] --geno 0 --recode ped --out [name outfile]*"). For only historical or only modern data sets, we filtered only once for missingness and minor allele counts.

Using the resulting data sets, we calculated Identity-by-State distances ("*plink --memory 32,000 --file [name of filtered .map/.ped] --distance square ibs allele-ct --out [name outputfile]*"). The resulting matrix of pairwise genetic distances was loaded into R, and translated with classical multidimensional scaling ("*stats::cmdscale(data, eig = T)*") to

enable plotting of the individuals, relative to their genetic distances, in a Cartesian space.

2.9 | Data processing and plotting using R

Complex data processing and all plotting was done with R version 3.4.4 for command line processing, otherwise version 3.6.1 combined with RSTUDIO version 1.2.1335 (R Core Team, 2019; RStudio Team, 2018).

For data manipulation, we used the packages TIDYVERSE (Wickham (2017)) and MOIR (Exposito-Alonso, 2019).

For general plotting, we used the packages GGPLOT (Wickham, 2016), COWPLOT (Wilke, 2019), RCOLORBREWER (Neuwrith, 2014), and QUANTREG (Koenker, 2019).

For plotting of geographical maps, we used the packages SP (Pebesma & Bivand, 2005) and RASTER (Bivand, Pebesma, Gomez-Rubio, & Hijmans, 2019).

3 | RESULTS

3.1 | ddRAD for nonmodel species

DNA extractions of about 1 cm^2 of *Arabidopsis thaliana* and *Cardamine bulbifera* leaf tissue using CTAB yielded average DNA concentrations of 34.0 (12.2–60.1) and 37.2 (6.2–98.9) ng/µl, respectively. For size selection, we combined library and bait samples at the same concentrations (mean sample concentrations: *A. thaliana* pUS-pool: 4.2 ng/µl, pMA-pool: 4.0 ng/µl; *C. bulbifera* pool: 4.1 ng/µl). Size selection with a Blue Pippin resulted in fragment size ranges of ~250–450 bp (as measured with a Bioanalzyer), in ~40 µl per sample, with concentrations of about 0.96 ng/µl (*A. thaliana* pUS-pool: 0.496 ng/µl, pMA-pool: 0.478 ng/µl; *C. bulbifera* pool: 1.92 ng/µl). We separated bait and library pools with adapter-based PCR amplification, reaching final concentrations of ~3.7 ng/µl for sequencing libraries (*A. thaliana* pUS-pool: 2.32 ng/µl, pMA-pool: 2.76 ng/µl; *C. bulbifera* pool: 5.88 ng/µl), and ~6.9 ng/ul for bait pools (*A. thaliana* pUS-pool: 8.56 ng/µl, pMA-pool: 6 ng/µl; *C. bulbifera* pool: 6.22 ng/µl).

For *A. thaliana*, we sequenced an average of 1.37×10^6 reads per sample (6,674 to 2.68×10^6). In total, 78% (61.5%–80.5%) of the obtained paired-end reads were merged. In comparison, of the 3.46×10^6 reads per *C. bulbifera* sample sequenced (1.75–7.43 × 10^6), 73.3% were merged (65.5%–77.3%).

3.2 | Fast and powerful pseudo-reference

Our method is intended to work for both model organisms with a high-quality reference genome, as well as for nonmodel species that entirely lack a reference assembly. We therefore used the processed ddRAD sequencing data for all fresh samples (i.e., after trimming of adapter sequences and restriction sites, and merging

of the paired-end reads), combining the merged and unmerged fraction of reads, to generate de novo references for both *A. thaliana* and *C. bulbifera* (assembly stats: *A. thaliana*—464,854 contigs, total 83,583,168 bp, min. 142 bp, max. 3,130 bp, avg. 180 bp, N50 175 bp; *C. bulbifera*—916,529 contigs, total 163,631,378 bp, min. 142 bp, max. 3,347 bp, avg. 179 bp, N50 171 bp). Because these MEGAHIT references are based on the generated fresh ddRAD sequencing data, despite their "unpolished" nature, the mapping fraction to those pseudo-assemblies is comparable with (or better than) mapping to published reference genomes (*A. thaliana* TAIR10 [Berardini et al., 2015] and the *C. hirsuta* genome [Gan et al., 2016]—for *C. bulbifera* the phylogenetically closest published reference). In total, 81.75%–96.42% of the *A. thaliana* reads, and 93.79%–95.30% of the *C. bulbifera* sequenced bases map to the respective MEGAHIT assemblies. In comparison, on average 94.84% of *A. thaliana* sequenced bases mapped to TAIR, and 44.45% of *C. bulbifera* bases to the *C. hirsuta* reference genome.

3.3 | hyRAD capture

3.3.1 | Bait generation and capture

To generate the biotinylated baits used for the hyRAD capture, we amplified the size-selected baits in a regular exponential PCR to concentrations of ~517 ng/µl (*A. thaliana* pUS: 483.6 ng/µl, pMA: 554.6 ng/µl; *C. bulbifera* 515 ng/µl). We then used this PCR product in a linear biotinylation-PCR to obtain the final bait pools (*A. thaliana* pUS: 34 µg in 110 µl, pMA: 38 µg in 110 µl; *C. bulbifera* 128 µg in 110 µl). As an example, to capture 40 samples with the *A. thaliana* pUS bait, (11 for capture 1 and 2 each, and six for captures with the pMix bait, which consists of 50% of pUS, results in 34 samples to capture, rounded up to 40), we used 32 ng in a total of four PCRs for the first amplification, and then ~0.5 ng/µl in each of 10 linear reactions to obtain a final bait pool of 34 ng and a concentration of ~326 ng/µl.

Dilutions of successful captures of ancient DNA libraries—independent of library type (i.e., single stranded [*C. bulbifera*] or double stranded [*A. thaliana*])—amplified in a standard qPCR to ~10^8 molecules, while captures of DNA extraction blanks and library blanks reached rarely more than ~10^6 molecules. Qubit measurements of captured libraries prior to final amplification did not produce meaningful concentrations, and were hence not indicative of capture success or failure. Post-qPCR amplification of successfully captured samples after the first capture resulted in average final library concentrations of ~190–230 ng/µl, with results being similar for both species, and both captures.

3.3.2 | Large mapped in-target fraction of authentic aDNA

For the first capture of the historical *A. thaliana* libraries, we sequenced on average 7.77×10^6 reads for the pUS-capture (6.54–9.87×10^6), 5.13×10^6 reads for the pMA capture (4.36–5.80×10^6),

7.34×10^6 reads for the capture using equal volumes of pUS and pMA bait (pMix, 6.21–8.79×10^6), and 6.89×10^6 reads for *C. bulbifera* libraries (5.04–8.51×10^6). On average 94% of reads were merged (*A. thaliana* pUS: mean 94.3%, 90.9%–97.1%; pMA mean 94.2%, 90.8%–97.1%; pMix mean 94.4%, 91.0%–97.1%; *C. bulbifera* mean 82.6%, 77.7%–85.9%).

The MEGAHIT reference represents only the genomic fraction that is selected with the ddRAD protocol. Therefore, the amount of reads mapping to MEGAHIT does not entirely reflect the fraction of the library that is plant endogenous DNA (*A. thaliana* or *C. bulbifera*). However, successful mapping does indicate the amount of historical DNA that was successfully captured with the ddRAD-based baits and is on-target. In all three captures, the fraction of sequenced base pairs that mapped to the respective MEGAHIT reference (and that thus is on-target) was above 55%, indicating a highly successful capture (*A. thaliana* pUS: mean 65.30%, 55.58%–73.48%; pMA mean 64.48%, 55.71%–71.26%; pMix mean 65.30%, 55.99%–72.75%; *C. bulbifera* mean ~75.9%, 65.41%–80.43%). PCR-duplicated reads accounted on average for ~21% of all mapped reads (*A. thaliana* pUS: mean 26.19%, 20.49%–35.47%; pMA mean 23.49%, 18.18%–29.49%; pMix mean 26.09%, 19.73%–33.09%; *C. bulbifera* mean 7.46%, 5.15%–11.40%), and were removed prior to further analysis.

Analysis of fragment sizes and cytosine deamination authenticated the historical nature of all hyRAD captured libraries (Figure S4). All libraries displayed the aDNA-typical increase of cytosine-to-thymine substitutions in their 5' ends, which ranged from 1.6% to 4.3% (*A. thaliana* pUS: mean 2.7%, 1.6%–4.2%; pMA mean 2.7%, 1.6%–4.3%; pMix mean 2.7%, 1.6%–4.2%; *C. bulbifera* mean 4%, 2.8%–6.2%; Figure S4a, b). On average, the *A. thaliana* libraries had a median fragment length of 70 bp (capture 1: overall median 69.1 bp; pUS: 68.6 bp; pMA 69.5 bp; pMix 69.2 bp; capture 2: overall median 70.87 bp; pUS: 70 bp; pMA 71.5 bp; pMix 71.1 bp; Figure S4d, f), slightly longer than the *C. bulbifera* measured medians of 50 bp for capture 1 and 53.5 bp for capture 2 (Figure S4c, e).

3.3.3 | One versus two captures

All samples of both species, and using all bait sets, were subjected to two rounds of capture to assess the in-target gain after performing sequential capture. In the second capture, an average of 6.79×10^6 reads were sequenced per sample (*A. thaliana* pUS: mean 6.23×10^6, 1.2–7.89×10^6; pMA mean 5.56×10^6, 4.80–7.32×10^6; pMix mean 8.57×10^6, 7.66–9.68×10^6; *C. bulbifera* mean 8.26×10^6, 5.36–9.83×10^6), of which about 95% were merged prior to mapping (*A. thaliana* pUS: mean 94.6%, 89.7%–97.2%; pMA mean 94.9%, 92.8%–97.3%; pMix mean 94.9%, 93.0%–97.0%; *C. bulbifera* mean 88.8%, 86.2%–91.7%). Both the first and the second capture round were then mapped against TAIR and the *C. hirsuta* reference as well as against the respective MEGAHIT assemblies. For both captures, a larger fraction of reads could be mapped against the published references than against MEGAHIT, a tendency that was less obvious for the second round of capture. Using TAIR, an average of ~87% bp of the

first capture could be mapped (pUS: mean 87.17%, 65.36%–96.85%; pMA mean 87.1%, 65.59%–96.75%; pMix mean 87.25%, 65.77%–96.9%), which for the second capture had increased to ~93% (pUS: mean 92.92%, 81.61%–97.87%; pMA mean 92.84%, 81.82%–97.8%; pMix mean 92.94%, 81.64%–97.87%). In comparison, ~65% of the first capture mapped to MEGAHIT (see above), and ~79% of the second capture (pUS: mean 79.62%, 74.31%–82.71%; pMA mean 78.8%, 74.35%–81.85%; pMix mean 79.71%, 74.44%–84.19%). The *C. bulbifera* samples displayed a smaller effect of first versus second capture, in part probably resulting from the divergence between *C. bulbifera* and the published *C. hirsuta* genome: ~22.6% of the first, and ~ 23.8% of the second capture mapped to *C. hirsuta* (13.94%–40.28% and 16.03%–39.16%, respectively), compared to a mapped 75.9% and 86.9% (81.80%–89.10%) to MEGAHIT for the first and second capture, respectively. Overall, independent of capture and reference, a very high fraction of all reads could be mapped in all samples, indicating highly successful capture and a high proportion of in-target reads.

Deduplication (i.e., the removal of PCR duplicates after mapping) reduced the amount of reads per sample by a similar fraction as seen for the first capture mapped against MEGAHIT only. TAIR-mapped *A. thaliana* contained on average 19.45% (capture 1) and 27.98% (capture 2) duplicated reads, and mapped against MEGAHIT 25.26% or 33.9% (capture 1 and 2, respectively). Historical *C. bulbifera* samples, in comparison, lost about 4.73% and 8.88% (capture 1 and 2) reads to deduplication when mapped to *C. hirsuta*, and 7.46% (capture 1) and 14.25% (capture 2) when mapped to the MEGAHIT reference.

To estimate the information gain resulting from the second capture, we compared how many base pairs (after deduplication) were covered per sample and capture, relative to the invested sequencing effort (Figure 3). When mapped to MEGAHIT, the second capture slightly increased the sequencing efficiency, with lower sequencing efforts resulting in on average more base pairs covered—an effect barely seen when samples were mapped against the published references (Figure 3a *A. thaliana* and Figure 3b *C. bulbifera*, upper panels). When restricting the analysis to unique base pairs covered, however, sequencing efficiency when mapping to the published full genome references was pronouncedly different between the two captures (Figure 3a *A. thaliana* and Figure 3b *C. bulbifera*, right lower panels). For both *A. thaliana* and *C. bulbifera*, the second capture resulted in a distinct decrease in unique mapped sites—an effect that was not recapitulated when mapping samples to MEGAHIT (Figure 3a *A. thaliana* and Figure 3b *B. bulbifera*, left lower panels). Taken together, while a second round of capture does increase the overall number of covered base pairs, it does not increase the number of unique sites mapped in the MEGAHIT reference—representing the targeted genome fraction.

The largest effect the second capture has is a reduction of unique sites mapping to the published reference genomes. Because a similar pattern is not observed when samples are mapped to MEGAHIT, these reads or sites probably represent off-target regions that can be mapped in the more extensive TAIR and *C. hirsuta* genomes, but are not part of the ddRAD fraction, and hence cannot

be mapped to the MEGAHIT assembly. The second capture further reduces this background variation, and enriches the samples for the targeted—MEGAHIT-mappable—fraction.

3.4 | Recapitulation of expected genetic diversity

Reduced representation analysis of population samples with ddRAD and hyRAD is only useful when these—compared to targeted SNP-chip sequencing or similar methods—unguided methods succeed in recapitulating existing genetic diversity, and thus are representative of the genetic diversity present at the whole-genome level. We took advantage of the extensively studied *A. thaliana* diversity to assess this. For comparison, we mapped published historical and modern shotgun sequencing data of African and North American *A. thaliana* samples to the TAIR reference, retrieving 1,362 variants de novo, of which 1,072 were left after filtering. Plotting the first two dimensions of a multidimensional scaling (MDS) analysis that was based on the pairwise genetic distances between these samples recapitulated, as expected, the geographical origins of the samples (Figure 4b), with samples clustering primarily based on geography, and not based on sample type (historical or modern). Parallel analysis of similar, ddRAD and hyRAD processed and sequenced samples retrieved 2,845 variable sites (of 3,616 prior to filtering). These sites recovered an almost identical distribution of samples (Figure 4c), emphasizing the ability of our reduced-representation approaches to recapitulate known genetic diversity, across both historical and modern samples, by targeting and sequencing the same, small fraction of the genome in two highly different types of samples.

3.5 | Negligible effects of bait diversity

Depending on the stringency of the capture conditions, the genetic diversity of the baits used for capture could influence how much, or which, genetic diversity can be captured. To investigate this, we generated bait sets with genetically distant North American and Moroccan lineages, as well as with a mix of the two. Independent captures of the same historical libraries were then analysed in a data set combined with the modern ddRAD data. As described, captured and modern samples recapitulate the geographical origins of the samples, and known genetic diversity. Recolouring of the hyRAD samples to indicate the bait set used for their capture does not show an obvious bias of the baits driving recovered genetic diversity, and thus of biasing the location of the samples in the first two dimensions of the MDS (Figure 4d). We formally tested this, removing the Algerian sample from the historical sample set, thus reducing the set to the distinct two clusters of North American and African samples. Of these, we calculated pairwise distances between samples that were captured with the same bait sets, both within and between clusters. For all comparisons (within the North American or the African cluster, and between clusters), the genetic distances recovered did not

differ significantly between the different bait sets, and as expected were largest for the comparisons between clusters (Figure S3).

3.6 | Genetic diversity along temporal and geographical scales in *C. bulbifera's* large, not referenced genome

De novo discovery of variants in the modern, historical and the joint data set retrieved 245,951, 365,780 and 159,989 variants, respectively. Filtering for missingness and minor allele count, 193,849, 247,350 and 70,554 variants were left for subsequent analyses. Individual MDS plots for the modern and historical data sets separated the samples based on their geographical origin into two clusters, for samples originating from Hainich and Schwäbische Alb, reflected in dimension 1 in the modern data, and dimension 2 in the historical data (18.05% and 3.75% of variance explained, Figure 5d,e). Especially the modern data set not only recapitulates this larger geographical pattern, but also reflects the different spatial locations of the samples within the two exploratories: samples from the Schwäbische Alb, the smaller and less latitudinally extended exploratory, cluster more tightly than those from Hainich (Figure 5a,c,e,f).

Joint analysis of both data sets first separates historical and modern samples in dimension 1 of the MDS (31.14% of variance explained), before also reflecting their geographical origin (dimension 2, 5.92% of variance). This separation persists when only variants that have full information across the whole sample set are used for the analysis (31,387 variants, dimension 1 32.58% variance, dimension 2 6.15% variance, Figure S5a). Similarly, it persists when removing all variants that might originate from deamination, and hence might not reflect true genetic variation (CT/TC, and AG/GA; 23,385 variants, 32.54% of variance explained in dimension 1, 6.79% in dimension 2, Figure S5b). Finally, analysing the GC content of modern and historical sample reads (merged reads, prior to mapping), we do not find evidence of a (bacterial) contamination in the historical samples that could cause the large difference between the two sample types (Figure S5c), but do see a slightly overall increased GC content in historical samples, a previously documented side-effect of hybridization capture (White et al., 2019).

To assess the relationship among historical and modern samples without taking into account historical-specific diversity, which could be driven by aDNA-associated damage, we used the same data to project the historical samples into the principal components analysis (PCA) space of the modern samples. The projected historical samples positioned in the centre (at coordinates close to 0,0) between the fresh samples from Hainich on one side and Schwäbische Alb on the other side (Figure 5f).

4 | DISCUSSION

We modified ddRADseq to enable parallel production of modern-sample-based sequencing libraries and re-amplifiable baits used for hybridization capture of historical libraries (hyRAD; Linck et al., 2017; Suchan et al., 2016). Generating data from two plant species—one the referenced model plant *A. thaliana*, the other the estimated dodecaploid nonmodel *C. bulbifera* that lacks a reference sequence for its ~2-Gbp genome—we investigate how many captures are sufficient for efficient retrieval of historical data. We analyse how our method recapitulates known genetic diversity across historical and modern samples, and how this is affected by the genetic relatedness of baits with the captured historical samples, using whole-genome sequenced samples mapped to a published reference genome as quality comparison. Finally, we show that our strategy uncovers new genetic diversity that recapitulates the geographical and temporal distribution of the investigated *C. bulbifera* samples.

4.1 | Improved ddRAD and hyRAD for (non-) model species

4.1.1 | Parallel production of "immortal" ddRAD-based capture baits

The main improvement in comparison with previously published methods for homemade hyRAD baits (Linck et al., 2017; Suchan et al., 2016) is the introduction of bait-specific adapters, which brings multiple advantages. In other protocols, RAD-based (or exome-based) baits are initially processed following regular library protocols. Conventional library-adapters are then removed enzymatically to avoid hybridization of capture libraries with the baits based on the adapter sequences, and to prevent unwanted amplification of baits (Puritz & Lotterhos, 2018; Suchan et al., 2016). However, it is unclear how efficient and complete this removal is, which is particularly problematic when baits also contain sequencing indices and can thus be sequenced alongside the captured libraries (Puritz & Lotterhos, 2018; Suchan et al., 2016). While such erroneously sequenced baits may potentially be identified as contaminants based on their index sequences, sequencing will be lost on uninformative and unwanted bait sequences.

In addition, removal of adapter sequences simultaneously eliminates the possibility of further amplification of the baits, a serious limitation for the number of possible captures, and for future additional captures or experimental replication. In contrast, and also unlike costly commercial products, our baits with their unique, retained adapters are theoretically "immortal," as they can be almost indefinitely amplified for cheap and flexible capture of large amounts of libraries (Fu et al., 2013).

Furthermore, specific adapters for hyRAD baits that are different from ddRAD library adapters enable the combination of highly overlapping hyRAD and ddRAD sequencing data for joint analysis. With ddRAD libraries and hyRAD baits being separately amplifiable, both can be pooled together for joint size-selection—a main variation-inducing step for separately processed libraries. Subsequent PCR-based amplification faithfully separates them again for further processing. Such parallel processing of fresh ddRAD libraries

and ddRAD-based capture baits ensures high similarity of the final fragment pools. It maximizes the overlap of modern, ddRAD-based genetic diversity, with historical genetic diversity retrieved by hybridization-based capture (hyRAD), and saves sample processing costs as well as time by circumventing the need to capture both fresh and historical libraries as a means to ensure compatibility (as done for example by Suchan et al., 2016). Ultimately, this allows joint population genetics analysis across geographical and temporal gradients.

In contrast to Suchan et al. (2016), during bait production, instead of employing a commercial biotinylation kit that randomly introduces biotinylated nucleotides into the bait sequences, we used a 5′-biotinylated primer in a linear amplification to generate biotinylated baits (Fu et al., 2013). A primer is cheaper and hence more scalable for high-throughput bait production than commercial kits. Also, linear amplification enriches specifically for a single strand, increasing bait and, thus ultimately, capture diversity.

Finally, we consistently use double-indexing for both the fresh ddRAD and the historical aDNA hyRAD-captured libraries, increasing the reliability of demultiplexing and reducing the probability of faulty read assignments (Kircher et al., 2012).

4.1.2 | Efficiency and sequential rounds of capture

Our capture and read mapping results confirm the efficiency of our baits and captures, and the high overlap with the fresh ddRAD libraries. On average, about ~70% of all historical reads map to our MEGAHIT pseudo-references (e.g., Figure 5b). Because those pseudo-references are based on the ddRAD sequences, and thus correspond to the genome fraction accessible with our RAD protocol, mapping of historical reads to the assembly can be interpreted as reads being successfully captured and "in target."

Further validating the efficiency of our protocol, we show that a single round of capture is sufficient to retrieve a majority of informative historical fragments. A subsequent second capture barely increases the number of new, unique sites mapped from the historical data (Figure 3), and serves mostly to increase sequencing depth of already captured sites. This is true for both *A. thaliana* and *C. bulbifera*, independent of their largely different genome sizes (135 Mbp versus. >2 Gbp, Table S4) and ploidy levels (diploid versus estimated dodecaploid; Carlsen et al., 2009; Kučera et al., 2005). In addition, with multiple captures the number of PCR cycles and thus of PCR-duplicated reads increases, which ultimately results in an overall decrease of library complexity.

Achieving a target coverage within given cost and time constraints will of course require balancing the number of captures and the invested sequencing effort. However, given the high in-target fraction of historical sequences already after one round of capture, we expect a single capture to be sufficient at least for historical samples with similar DNA properties as seen here: samples with a reasonably high endogenous DNA content (in our case at least ~70%, only taking into account the fraction of in-target reads, without

remaining bycatch that does not map to the RAD-based pseudo-reference), and a median fragment size of at least 50 bp—properties that are commonly seen in the majority of reasonably well-conserved herbarium specimens (Exposito-Alonso et al., 2018; Gutaker et al., 2019; Weiß et al., 2016). Re-evaluation of capture efficiency may be required for archaeobotanical samples, whose properties are closer to those encountered in ancient human remains (da Fonseca et al., 2015; Ramos-Madrigal et al., 2019), where a second capture has been shown to substantially increase the on-target fraction of reads (Ávila-Arcos et al., 2015; Burbano et al., 2010). In accordance, a recent study of faecal samples, which have similarly low DNA contents, also found one round of capture to be sufficient for samples with >2%–3% of endogenous DNA, but predicted two rounds of capture to be beneficial for samples of lower DNA content (White et al., 2019).

4.2 | Uncovering known and novel genetic diversity

4.2.1 | De novo site discovery without reference bias

Traditionally, analysis of RADseq data can be done de novo (i.e., without a reference genome), with popular pipelines such as STACKS or IPYRAD (Catchen et al., 2011; Eaton & Ree, 2013; https://ipyrad.readthedocs.io/). However, these approaches naturally only work for RADseq data, not for our associated hyRAD sequencing. To seamlessly combine true RADseq data and historical hyRAD sequencing, we therefore assembled a new, modern RAD-based pseudo-reference for mapping both historical and modern reads, and subsequent joint de novo polymorphism discovery.

As discussed above, despite the lack of a "true" reference genome, our pseudo-reference allows us to define the fraction of historical capture that is "in target." Apart from this, because the pseudo-reference comprises only a small part of the genome, it cannot be used to define the amount of total endogenous plant DNA present in our historical samples (see Section 4.1.2). It therefore also does not allow polarization of variable sites into "reference" and "alternative" alleles. Importantly, however, the pseudo-reference provides a reference for read mapping, thus establishing a shared coordinate system. With this, genetic variation can be aligned and compared for the same sites across all samples. This information is sufficient for de novo discovery of polymorphic sites independent of the reference using BSH-DENOVO (https://github.com/clwgg/bsh-denovo, rev. 30c95ab). By nature, this thus avoids ascertainment bias (Clark, Hubisz, Bustamante, Williamson, & Nielsen, 2005), a common problem in particular also for historical samples. Choosing the variable sites de novo, based on the genetic variation present in the entire investigated sample set, allows optimal use of all available sequencing information. It thereby maximizes the amount of retrieved polymorphic sites that can explain the genetic relationships within our sample set, and that may be used for further in-depth population genetics analyses.

4.2.2 | *A. thaliana* RADseq and pseudo-assembly recover known genetic diversity

Indeed, while RAD methods by nature only recover a small fraction of the genome, we show that this fraction is sufficient to recapitulate known genetic diversity in highly geographically dispersed and genetically different *A. thaliana* samples (Figure 4b,c; Platt et al., 2010; The, 1001 Genomes Consortium 2016). The genetic relationship of *A. thaliana* from the African continent (Durvasula et al., 2017) and from Northern America (Platt et al., 2010), identified with ddRAD, hyRAD and a ddRAD-based MEGAHIT reference (Figure 4c), recapitulates the clustering patterns that are generated using reference-genome-mapped whole-genome shotgun sequenced historical and modern samples from the same geographical areas (Figure 4b). Combining the two sample types and methods thus succeeds in retrieving not only overlapping, but also informative genetic variation, without the need for a high-quality reference genome. This also distinguishes our approach from, for example, exome-based captures (Puritz & Lotterhos, 2018; Schmid et al., 2017; White et al., 2019) that have been used for historical samples. An exome-based RAD-capture of ancient DNA, hyRAD-X, was recently presented as an alternative to genome-based hyRAD (Schmid et al., 2017). The focus of exome-based baits on transcribed regions of the genome may compromise population history analyses, because the roles of genetic drift and selection are more difficult to disentangle in exome-based data, whereas RAD-based data sets are more suitable for looking at genetic-drift-driven population differentiation. RADseq-based hyRAD offers a less biased, but still reduced-representation view of the genome, and at the same time is cheaper and probably faster. Most importantly, however, it allows a facile and straightforward DNA-based comparison of fresh with historical material. In comparison, exome-based methods require first the assembly of a reference transcriptome using fresh samples. If not done carefully to cover the variability of the transcriptome, this might create a biased view of the genome, making exome-based methods susceptible to (environmentally induced) expression fluctuations and associated dropout that not necessarily reflect true genetic differences.

4.2.3 | Bait diversity

Working with capture, a much-discussed subject is the necessary and sufficient genetic diversity of capture baits (Bi et al., 2012; Good et al., 2013). Predesigned commercial capture baits are generally designed based on reference genomes. Most—especially nonmodel—species lack such resources. Generation of informative, unbiased baits is therefore particularly problematic for nonmodel species that typically lack a referenced genome or prior sequencing information required for guided bait design. To investigate potential biases in how different baits recover population differentiation, we captured the same historical *A. thaliana* samples with three different bait sets, representing either of the two major geographical clusters within our samples (African and North American, pMA and pUS), and a mix of both (pMix, Figure 4a). Visually assessing the resulting clustering

of samples based on MDS, we did not find an effect of the different bait sets (Figure 4c,d). Furthermore, baits did not have an effect on patterns of population differentiation in comparisons of IBS-based (Identity-by-state) pairwise genetic distances among samples (Figure S3) both within and between the major genetic/geographical clusters.

This supports the suggestion that hybridization capture is to a certain extent resistant to sequence variation, and can thus be used for species with unknown genetic diversity, where fresh material for the baits is by necessity selected "blindly."

4.3 | Temporal and geographical genetic diversity in *C. bulbifera*

We show that combined ddRAD and hyRAD data can be used to genetically characterize populations of nonmodel organisms across both geographical and temporal gradients, using the analysis of historical and modern *C. bulbifera* as a test case (Figure 5). Our results indicate that the two populations sampled in Germany and close to the biodiversity exploratories (Fischer et al., 2010) Schwäbische Alb (ALB) and Hainich (HAI, Figure 5a) are genetically distinct. This reflects their—in fact rather small—geographical separation of ~300 km, and holds true for both historical and modern populations.

Interestingly, this separation reflecting geographical origin is more pronounced in modern samples (dimension 1, 18.05% variation explained), where even the difference between the latitudinally extended HAI and the smaller and more compressed ALB exploratory is recapitulated (Figure 5c). In contrast, in the historical MDS, it is the second dimension (with 3.75% variance explained) that reflects (weaker than in modern samples) the geographical origin of the samples (Figure 5d). This absence of strong geographical population structure is partially due to the geographically spread origin of the historical samples, which were selected based on their proximity to the exploratories, but do cover a wider geographical range than the fresh samples that originate exclusively from within the exploratories (Figure 5a). In addition, it is possible that the genetic variation found over the sampling period of 197 years (Table S2) is greater than the geography-related variation. This could also explain the strong separation between historical and modern samples when looking at the full data set, where dimension 1 separates both sample types and explains 31.14% of the observed variance (Figure 5e; as opposed to the very subtle separation of modern and historical samples seen for *A. thaliana*, Figure 4c). We could not attribute this variation to a bias in missingness in either sample set (Figure S5a), nor to age-related variation (Figure S5b), to a contamination of historical samples (Figure S5c) or biased mapping of reads in the two exploratories in only one of the two data sets (Figure S5d). In addition, when we project the historical samples into the PCA-space of the modern samples, the geographical origin of the samples is again reflected in PC1. Placement of the historical samples in between the two clusters of modern samples suggests the presence of independent structured populations at different time points. Therefore, it is likely that the observed pronounced

separation illustrates true genetic differences between historical and modern samples, potentially related to a changing climate or generally changing environmental conditions over time, or simply due to population structure. Correlation of genetic changes with historical climate data, or analysis of allele frequency changes over time (and space) could serve to further investigate this possibility.

5 | CONCLUSIONS

The strategy presented here substantially improves published ddRAD and hyRAD methods, adding to a growing repertoire of reduced-representation methods for either historical or modern (nonmodel species) samples. We explicitly use the method for the joint analysis of historical and modern samples, showing that it is possible to obtain reduced-representation overlapping genetic variation from both, despite the large differences in DNA preservation and quality in the two sample types, and entirely independent of a sequenced reference genome.

This method further opens the door to the richnesses of herbaria (Lang et al., 2018; Meineke et al., 2018) and of museum collections in general, for example to vast collections of insect species. With it, studies of nonmodel species lacking references or large genomes become broadly accessible even for analyses at the population scale. The method allows comparisons of historical and modern diversity, for example to investigate responses of species to anthropogenic global change, evidenced in changes in genetic diversity and population structure over time until today. Molecular analyses of historical collections thus pave the way to move past the mostly descriptive analyses of, for example, species declines (Shaffer, Fisher, & Davidson, 1998), to start understanding how genome-scale processes such as eroding genetic diversity are related to species declines and biodiversity loss.

CONFLICT OF INTERESTS
The authors declare no competing or financial interests.

AUTHOR CONTRIBUTIONS
P.L.M.L. and H.A.B. designed the project, H.A.B. supervised research. P.L.M.L. and S.L. extracted historical DNA. P.L.M.L. and S.K. developed the fresh sample protocol with input from C.L.W., M.M. and H.A.B. S.N. prepared *Cardamine bulbifera* aDNA libraries. B.N. and M.M. gave input for hyRAD protocol development and guided aDNA captures done by P.L.M.L. P.L.M.L. analysed results with input from C.L.W., S.L. and H.A.B. C.L.W. developed the polymorphism de novo sampling tool. P.L.M.L. wrote the first version of the manuscript with input from H.A.B. and C.L.W. The manuscript was finalized with input from all authors.

DATA AVAILABILITY STATEMENT
DNA sequencing data are deposited in the European Nucleotide Archive (ENA), with accession no. PRJEB36294. Published shotgun sequences of modern African *Arabidopsis thaliana* were downloaded from ENA, from study PRJEB19780 (accession nos. ERS1575068 [Arb-0], ERS1575074 [Elh-2], ERS1575132 [Tanz-1]; Durvasula et al., 2017), modern HPG1 shotgun data were obtained from G. Shirsekar (personal communication). Historical shotgun sequencing data for African *A. thaliana* are available in study PRJEB19780 (ERS1575137 [AH0004], ERS1575138 [AH0006], ERS1575139 [AH0007], ERS1575140 [AH0008], ERS1575142 [AH0011]), and historical North American (HGP1) samples were published before at ENA under study PRJEB15366 (accession nos. ERS1342420 [HB0001], ERS1342418 [HB0003], ERS1342416 [HB0005], ERS1342414 [HB0007], ERS1342412 [HB0009]; Gutaker et al., 2017).

ORCID
Patricia L. M. Lang https://orcid.org/0000-0001-6648-8721
Clemens L. Weiß https://orcid.org/0000-0003-3321-3902
Sonja Kersten https://orcid.org/0000-0002-9096-0448
Sergio M. Latorre https://orcid.org/0000-0002-5889-0670
Matthias Meyer https://orcid.org/0000-0002-4760-558X
Hernán A. Burbano https://orcid.org/0000-0003-3433-719X

REFERENCES

Aitken, S. N., & Bemmels, J. B. (2016). Time to get moving: Assisted gene flow of forest trees. *Evolutionary Applications*, 9(1), 271–290. https://doi.org/10.1111/eva.12293

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews, Genetics*, 17(2), 81–92.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814), 796–815.

Ávila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V., Rasmussen, M., ... Gilbert, M. T. P. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports*, 1(August), 74. https://doi.org/10.1038/srep00074

Ávila-Arcos, M. C., Sandoval-Velasco, M., Schroeder, H., Carpenter, M. L., Malaspinas, A.-S., Wales, N., ... Gilbert, P. (2015). Comparative performance of two whole-genome capture methodologies on ancient DNA illumina libraries. *Methods in Ecology and Evolution/British Ecological Society*, 6(6), 725–734.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Barreiro, S., Fátima, F. G., Vieira, M. D., Martin, J. H., Thomas, M., Gilbert, P., & Wales, N. (2017). Characterizing restriction enzyme-associated loci in historic ragweed (Ambrosia Artemisiifolia) voucher specimens using custom-designed RNA probes. *Molecular Ecology Resources*, 17(2), 209–220.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53–59. https://doi.org/10.1038/nature07517

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the 'Gold Standard' annotated reference plant genome. *Genesis*, 53(8), 474–485.

Bi, K. E., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, 13(8), 403. https://doi.org/10.1186/1471-2164-13-403

Bieker, V. C., & Martin, M. D. (2018). Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. *Botany Letters*, 165(3-4), 409–418. https://doi.org/10.1080/23818107.2018.1458651

Bivand, R., Pebesma, E., Gomez-Rubio, V., & Hijmans, R. J. (2019). RASTER: Geographic Data Analysis and Modeling. R package version 3.0-2, https://CRAN.R-project.org/package=raster

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prufer, K., ... Paabo, S. (2007). Patterns of damage in genomic DNA sequences from a neandertal. *Proceedings of the National Academy of Sciences of USA*, 104(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Burbano, H. A., Hodges, E., Green, R. E., Briggs, A. W., Krause, J., Meyer, M., ... Paabo, S. (2010). Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, 328(5979), 723–725. https://doi.org/10.1126/science.1188046

Carlsen, T., Bleeker, W., Hurka, H., Elven, R., & Brochmann, C. (2009). Biogeography and phylogeny of cardamine (Brassicaceae) 1. *Annals of the Missouri Botanical Garden*, 96(2), 215–236.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). STACKS: Building and genotyping loci de novo from short-read sequences. *G3*, 1(3), 171–182.

Catchen, J. M., Hohenlohe, P. A., Louis Bernatchez, W., Funk, C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, 17(3), 362–365. https://doi.org/10.1111/1755-0998.12669

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(2), 7. https://doi.org/10.1186/s13742-015-0047-8

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15(11), 1496–1502. https://doi.org/10.1101/gr.4107905

da Fonseca, R. R., Smith, B. D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M., ... Gilbert, M. T. P. (2015). The origin and evolution of maize in the Southwestern United States. *Nature Plants*, 1(1), 14003. https://doi.org/10.1038/nplants.2014.3

De Wit, P., Pespeni, M. H., & Palumbi, S. R. (2015). SNP Genotyping and population genomics from expressed sequences - current advances and future possibilities. *Molecular Ecology*, 24(10), 2310–2323. https://doi.org/10.1111/mec.13165

Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry*, 51A(2), 127–128.

Drosophila 12 Genomes Consortium, Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., & ... MacCallum, I. 2007. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167), 203–218.

Durvasula, A., Fulgione, A., Gutaker, R. M., Alacakaptan, S. I., Flood, P. J., Neto, C., ... Hancock, A. M. (2017). African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of USA*, 114(20), 5213–5218.

Eaton, D. A. R., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology*, 62(5), 689–706. https://doi.org/10.1093/sysbio/syt032

Exposito-Alonso, M. (2019). MOIR: A set of R functions for an easy life and analyses. R package version 0.0.1. https://github.com/MoisesExpositoAlonso/moiR

Exposito-Alonso, M., Becker, C., Schuenemann, V. J., Reiter, E., Setzer, C., Slovak, R., ... Weigel, D. (2018). The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genetics*, 14(2), e1007155. https://doi.org/10.1371/journal.pgen.1007155

Fischer, M., Bossdorf, O., Gockel, S., Hänsel, F., Hemp, A., Hessenmöller, D., ... Weisser, W. W. (2010). Implementing large-scale and long-term functional biodiversity research: The biodiversity exploratories. *Basic and Applied Ecology*, 11(6), 473–485. https://doi.org/10.1016/j.baae.2010.07.009

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., & Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of USA*, 110(6), 2223–2227. https://doi.org/10.1073/pnas.1221359110

Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R. D., ... Tsiantis, M. (2016). The Cardamine Hirsuta genome offers insight into the evolution of morphological diversity. *Nature Plants*, 2(11), 16167. https://doi.org/10.1038/nplants.2016.167

Gansauge, M.-T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., ... Meyer, M. (2017). Single-stranded DNA Library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Research*, 45(10), e79. https://doi.org/10.1093/nar/gkx033

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, 8(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., ... Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2), 182–189. https://doi.org/10.1038/nbt.1523

Good, J. M., Wiebe, V., Albert, F. W., Burbano, H. A., Kircher, M., Green, R. E., ... Pääbo, S. (2013). Comparative population genomics of the ejaculate in humans and the great apes. *Molecular Biology and Evolution*, 30(4), 964–976. https://doi.org/10.1093/molbev/mst005

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... Paabo, S. (2010). A draft sequence of the neandertal genome. *Science*, 328(5979), 710–722. https://doi.org/10.1126/science.1188021

Gutaker, R. M., & Burbano, H. A. (2017). Reinforcing plant evolutionary genomics using ancient DNA. *Current Opinion in Plant Biology*, 36(2), 38–45. https://doi.org/10.1016/j.pbi.2017.01.002

Gutaker, R. M., Reiter, E., Furtwängler, A., Schuenemann, V. J., & Burbano, H. A. (2017). Extraction of ultrashort DNA molecules from herbarium specimens. *BioTechniques*, 62(2), 76–79. https://doi.org/10.2144/000114517

Gutaker, R. M., Weiß, C. L., Ellis, D., Anglin, N. L., Knapp, S., Fernández-Alonso, J. L., ... Burbano, H. A. (2019). The origins and adaptation of European potatoes reconstructed from historical genomes. *Nature Ecology & Evolution*, 3(7), 1093–1101. https://doi.org/10.1038/s41559-019-0921-3

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). MAPDAMAGE2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the illumina platform. *Nucleic Acids Research*, 40(1), e3. https://doi.org/10.1093/nar/gkr771

Koenker, R. (2019). QUANTREG: Quantile Regression. R Package version 5.51. https://CRAN.R-project.org/package=quantreg

Kučera, J., Valko, I., & Marhold, K. (2005). On-line database of the chromosome numbers of the genus Cardamine (Brassicaceae). *Biologia*, 60(4), 473–476.

Lang, P. L. M., Willems, F. M., Scheepens, J. F., Burbano, H. A., & Bossdorf, O. (2018). Using herbaria to study global environmental change. *The New Phytologist*, 221(1), 110–122. https://doi.org/10.1111/nph.15401

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn Graph. *Bioinformatics*, 31(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., ... Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 0, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.

Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. arXiv [q-bio.GN]. arXiv. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMTOOLS. *Bioinformatics*, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3), 639–647. https://doi.org/10.1111/1755-0998.12995

Linck, E. B., Hanna, Z. R., Sellas, A., & Dumbacher, J. P. (2017). Evaluating hybridization capture with RAD probes as a tool for museum genomics with historical bird specimens. *Ecology and Evolution*, 7(13), 4755–4767. https://doi.org/10.1002/ece3.3065

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., ... Gibbs, R. A. (2012). The *Drosophila melanogaster* genetic reference panel. *Nature*, 482(7384), 173–178. https://doi.org/10.1038/nature10811

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Malaspinas, A.-S., Tange, O., Moreno-Mayar, J. V., Rasmussen, M., DeGiorgio, M., Wang, Y., ... Nielsen, R. (2014). BAMMDS: A tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics*, 30(20), 2962–2964. https://doi.org/10.1093/bioinformatics/btu410

Martin, M.(2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 10. https://doi.org/10.14806/ej.17.1.200.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the adh locus in Drosophila. *Nature*, 351(6328), 652–654. https://doi.org/10.1038/351652a0

Meineke, E. K., Davis, C. C., & Jonathan Davies, T. (2018). The unrealized potential of herbaria for global change biology. *Ecological Monographs*, 165(6), 351.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6), pdb.prot5448–pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2), 240–248. https://doi.org/10.1101/gr.5681207

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., ... Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, 38, 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. R News, 5(2). https://cran.r-project.org/doc/Rnews/

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., & Nieselt, K. (2016). EAGER: Efficient ancient genome reconstruction. *Genome Biology*, 17(March), 60. https://doi.org/10.1186/s13059-016-0918-z

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., ... Borevitz, J. O. (2010). The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, 6(2), e1000843. https://doi.org/10.1371/journal.pgen.1000843

Poinar, H. N., Schwarz, C., Qi, J. I., Shapiro, B., MacPhee, R. D. E., Buigues, B., ... Schuster, S. C. (2006). Metagenomics to Paleogenomics: Large-scale sequencing of mammoth DNA. *Science*, 311(5759), 392–394. https://doi.org/10.1126/science.1123360

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. https://doi.org/10.1086/519795

Puritz, J. B., & Lotterhos, K. E. (2018). Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Molecular Ecology Resources*, 18(6), 1209–1222. https://doi.org/10.1111/1755-0998.12905

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD Fad. *Molecular Ecology*, 23(24), 5937–5942. https://doi.org/10.1111/mec.12965

Quinlan, A. R., & Hall, I. M. (2010). BEDTOOLS: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ramos-Madrigal, J., Runge, A. K. W., Bouby, L., Lacombe, T., Samaniego Castruita, J. A., Adam-Blondon, A.-F., … Wales, N. (2019). Palaeogenomic insights into the origins of french grapevine diversity. *Nature Plants*, *5*(6), 595–603. https://doi.org/10.1038/s41477-019-0437-5

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946. https://doi.org/10.1101/gr.128124.111

RStudio Team. (2018). RSTUDIO: *Integrated Development for R*. Boston, MA: RStudio Inc. http://www.rstudio.com/

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, *7*(3), e34131. https://doi.org/10.1371/journal.pone.0034131

Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., & Alvarez, N. (2017). HyRAD-X, a Versatile Method Combining Exome Capture and RAD Sequencing to Extract Genomic Information from Ancient DNA. Edited by M. Gilbert. *Methods in Ecology and Evolution/ British Ecological Society*, *8*(10), 1374–1388.

Schubert, M., Lindgreen, S., & Orlando, L. (2016). ADAPTERREMOVAL v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(2), 88.

Shaffer, H. B., Fisher, R. N., & Davidson, C. (1998). The role of natural history collections in documenting species declines. *Trends in Ecology & Evolution*, *13*(1), 27–30. https://doi.org/10.1016/S0169-5347(97)01177-4

Shapiro, B., & Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: New insights from ancient DNA. *Science*, *343*(6169), 1236573. https://doi.org/10.1126/science.1236573

Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., … Meyer, M. (2017). Neandertal and Denisovan DNA from Pleistocene Sediments. *Science*, *356*(6338), 605–608. https://doi.org/10.1126/science.aam9695

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., … Alvarez, N. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE*, *11*(3), e0151651.

The 1001 Genomes Consortium (2016). 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, *166*(2), 481–491. https://doi.org/10.1016/j.cell.2016.05.063

Weiß, C. L., Schuenemann, V. J., Devos, J., Shirsekar, G., Reiter, E., Gould, B. A., … Burbano, H. A. (2016). Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *Royal Society Open Science*, *3*(6), 160239. https://doi.org/10.1098/rsos.160239

White, L. C., Fontsere, C., Lizano, E., Hughes, D. A., Angedakin, S., Arandjelovic, M., … Vigilant, L. (2019). A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. *Molecular Ecology Resources*, *19*(3), 609–622. https://doi.org/10.1111/1755-0998.12993

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Cham, Switzerland: Springer.

Wickham, W. (2017). TIDYVERSE: Easily Install and Load the "Tidyverse". R package version 1.2.1, https://CRAN.R-project.org/package=tidyverse

Wilke, C. O. (2019). cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2". R package version 1.0.0.https://CRAN.R-project.org/package=cowplot

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320. https://doi.org/10.1126/science.1251385

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.