Population genomics of herbicide resistance in *Alopecurus myosuroides*

Dissertation to obtain the doctoral degree of Agricultural Sciences
(Dr. sc. agr.)

Faculty of Agricultural Sciences
University of Hohenheim

Institute of Plant Breeding, Seed Science and Population Genetics

submitted by
Sonja Maria Kersten
born in Tübingen, Germany

2022

Tag der mündlichen Prüfung: 09.12.2022

1. Prodekan:	Prof. Dr. U. Ludewig
Berichterstatter, 1. Prüfer:	Prof. Dr. K. Schmid
Mitberichterstatter, 2. Prüfer:	Prof. Dr. D. Weigel
3. Prüfer:	Prof. Dr. M. Hasselmann

For Friedrich

Contents

List of Figures

List of abbreviations

ABC - ATP-binding cassettes

ACCase - Acetyl-CoA carboxylase

ALS - Acetolactate synthase

ATP - Adenosine triphosphate

CCS - Circular consensus sequencing

cDNA - complementary DNA

CLR - Continuous long reads

CRISPR - Clustered regularly interspaced short palindromic repeats

ddRADseq - Double-digest restriction site-associated DNA sequencing

DENs - Phenylpyrazolines

DIMs - Cyclohexanediones

DSN - Duplex-specific nuclease

EPSPS - 5-enolpyruvylshikimate-3-phosphate synthase

FOPs - Aryloxyphenoxy-propionates

F_{ST} - Fixation index

GMO - Genetically modified organisms

GWAS - Genome-wide association study

HRAC - Herbicide resistance action committee

IAA - Indole-3-acetic acid

IMIs - Imidazolinones

IWM - Integrated weed management

MoA - Mode of action

NADPH - Nicotinamide adenine dinucleotide phosphate hydrogen

NGS - Next-generation-sequencing

NTSR - Non-target site resistance

ONT - Oxford Nanopore Technologies

PacBio - Pacific Biosciences

pbaa - Pacbio amplicon analysis

PCA - Principal component analysis

PS II - Photosystem II

RADseq - Restriction site-associated DNA sequencing

SMRT - Single-molecule real-time

SNP - Single nucleotide polymorphism

SSR - Simple sequence repeats

SUs - Sulfonylureas

TE - Transposable elements

TSR - Target-site resistance

WGS - Whole-genome shotgun

Summary

Over the past 50 years, herbicides have often replaced mechanical and manual human weed control, thus representing a major factor in yield productivity in modern agriculture. Herbicide applications, however, exert strong selection pressures on weeds. As a consequence, these species have developed herbicide resistance through adaptive, beneficial alleles that increase in number to ensure the persistence of the populations, a phenomenon known as evolutionary rescue. A major research question is whether herbicide resistance adaptation is more likely to arise from standing genetic variation that was present before the onset of herbicide selection or from *de novo* mutations that arose after herbicide selection began. To address this question, I focused on target-site resistance (TSR) point mutations, which cause a lower binding affinity to the target protein of the respective herbicides. I first investigated the diversity of TSR haplotypes in populations of the grass species *Alopecurus myosuroides* (common name: blackgrass), and compared it with the TSR diversity outcome of simulated populations under both evolutionary scenarios.

I first conducted a population genetics study of *A. myosuroides*, which is the most problematic weed in winter cereals across the European continent due to rapid resistance evolution. To obtain genome-wide polymorphic markers, I adapted a restriction site-associated DNA sequencing protocol to this species. I began by analyzing the diversity and population structure in a smaller local South German collection. The fact that I could differentiate populations on a local scale motivated me to extend the study to a European-wide collection, in which I found clear population structure, albeit with low differentiation and some evidence for admixture across Europe. In addition, I generated highly accurate long-read amplicons from single individuals of two loci, *ACETYL-COA CARBOXYLASE* (*ACCase*) and *ACETOLACTATE SYNTHASE* (*ALS*), which are the targets of the two main herbicide modes of action used in European cereal crops. I obtained completely phased haplotype information, supporting the analysis of haplotype diversity on a population level. I found a remarkable diversity of beneficial TSR mutations at the field level arising from multiple haplotypes of independent origin, so called soft sweeps. I used this information to perform forward simulations to investigate the evolutionary origin of these mutations. I found evidence that a majority of resistance mutations originated from standing genetic variation. While this at first may appear surprising, it is consistent with very large census and effective population sizes in blackgrass.

Since long-read amplicon sequencing of single individuals could be costly and time consuming, I extended the analysis to pools of 150 to 200 individuals from Germany,

Belgium, France, the Netherlands and the United Kingdom. By combining the power of a more stringent accuracy criterion in our long-reads and a novel clustering software (PacBio amplicon analysis), I was able to preserve individual haplotype information in pooled samples. Furthermore, in a proof of concept experiment, I was able to recover in our pools most haplotypes previously sequenced in individuals. The amplicon study provides a versatile workflow that can be easily adapted to any gene of interest in different species.

In conclusion, I found that many *A. myosuroides* populations likely already have the genetic prerequisites not only for rapid evolution of resistance to currently used herbicides, but also to herbicides that have not yet been brought to market.

Zusammenfassung

In den letzten 50 Jahren haben Herbizide größtenteils die mechanische und manuelle Unkrautentfernung durch den Menschen ersetzt und bilden damit einen wichtigen Beitrag zur Ertragsstabilität in der modernen Landwirtschaft. Der Einsatz von Herbiziden übt jedoch einen starken Selektionsdruck auf Unkräuter aus. Infolgedessen haben diese Arten eine Herbizidresistenz entwickelt, welche mit dem Anstieg von adaptiven, vorteilhaften Allelen einhergeht und so den Fortbestand der Populationen sichert - ein Phänomen, das als evolutionäre Rettung bekannt ist. Hier stellt sich die Frage, ob die Herbizidresistenz als bestehende genetische Variationen in den Populationen bereits vor Beginn der Herbizid-Selektion vorhanden ist oder aus de-novo Mutationen hervorgeht, die nach Beginn der Herbizid-Selektion entstanden sind. Um diese Frage zu klären, habe ich mich auf Punktmutationen der Target-Site-Resistenz (TSR) konzentriert, die eine geringere Bindungsaffinität zum Zielprotein des jeweiligen Herbizids bewirken. Dazu analysierte ich die Diversität der TSR-Haplotypen in Populationen der Grasart *Alopecurus myosuroides* (allgemein: Ackerfuchsschwanz) und verglich sie mit dem Ergebnis der TSR-Diversität simulierter Populationen unter beiden Evolutionsszenarien.

Als Erstes führte ich eine populationsgenetische Studie in *A. myosuroides* durch, welches aufgrund der schnellen Resistenzentwicklung das problematischste Unkraut beim Anbau von Wintergetreide in Europa ist. Um genomweite polymorphe Marker zu generieren, habe ich ein Protokoll zur Erzeugung von Restriktionsstellen-assoziierten DNA-Markern an diese Grasart angepasst. Zunächst analysierte ich die Diversität und Populationsstruktur in einer kleineren lokalen süddeutschen Sammlung. Die Tatsache, dass ich Populationen auf lokaler Ebene unterscheiden konnten, motivierte mich, die Studie auf eine europaweite Sammlung auszudehnen. Dort zeigte sich eine klare Populationsstruktur, wenn auch mit einer geringen Differenzierung, sowie Hinweise auf genetische Vermischungen in ganz Europa. Darüber hinaus generierte ich hochpräzise Long-Read Amplikons von einzelnen Individuen der beiden Gene *ACETYL-COA CARBOXYLASE (ACCase)* und *ACETOLACTAT SYNTHASE (ALS)*, welches die Zielgene der beiden wichtigsten herbiziden Wirkmechanismen im Getreideanbau sind. Ich erhielt vollständig phasierte Haplotyp Informationen, die es mir ermöglichten, die Haplotyp-Diversität auf Populationsebene zu analysieren. Ich fand eine bemerkenswerte Vielfalt an vorteilhaften TSR-Mutationen auf Feldebene, die aus mehreren Haplotypen unabhängigen Ursprungs, so genannten Soft Sweeps, hervorgingen. Diese Informationen konnte ich für Vorwärts-Simulationen nutzen, um den evolutionären Ursprung dieser Mutationen näher zu untersuchen. Ich fand Hinweise darauf, dass die Mehrzahl der Resistenzmutationen aus bestehenden genetischen Variationen hervorgegangen sind. Dies

mag auf den ersten Blick überraschen, steht aber im Einklang mit den hohen effektiven Populationsgrößen von Ackerfuchsschwanz.

Da die Long-Read Amplikon Sequenzierung von Einzelindividuen kostspielig und zeitaufwändig sein kann, habe ich meine Methode auf Pools von 150 bis 200 Individuen aus Deutschland, Belgien, Frankreich, den Niederlanden und dem Vereinigten Königreich ausgedehnt. Durch die Erhöhung der Read Genauigkeit und mit Hilfe der neuartigen Clustering-Software (PacBio amplicon analysis) war ich in der Lage, individuelle Haplotyp-Informationen in den gepoolten Proben zu erhalten. Darüber hinaus konnte ich die meisten Haplotypen, die zuvor als Einzelproben sequenziert worden waren, in unseren Pools identifizieren. Die Amplikon Pool Methode kann leicht an jedes Gen von Interesse in verschiedenen Arten angepasst werden.

Ich komme zu dem Schluss, dass in vielen *A. myosuroides* Populationen die genetischen Vorraussetzungen für eine rasche Resistenzentwicklung höchstwahrscheinlich bereits vorhanden sind. Dies betrifft nicht nur Herbizide, die derzeit verwendet werden, sondern auch Wirkstoffe, die noch nicht auf dem Markt sind.

1. General Introduction

1.1 Winning the race against extinction through resistance evolution

With the increasing demand for agricultural products in our growing world population and the associated need for higher and stable yields, the use of chemical plant protection in agriculture has become essential (Food and Agriculture Organization 2019; Gianessi 2013). However, this widespread use of pesticides carries, among other risks, the threat of rapid and repeated evolutionary adaptation through resistance development (REX Consortium 2013).

From Darwin's theory of evolution, we know that the best adapted individuals have the highest probability to produce the most offspring in a given environment (Darwin 1859). This "survival of the fittest", in which the less well adapted is gradually displaced, can be observed in a time-lapse rate when resistance builds up in populations. The application of pesticides represents a human-induced high selection pressure that triggers a race against extinction in populations subjected to selection. This resistance evolution occurs very rapidly and is based on the increase of beneficial adaptive alleles that emerge from the genetic pool of heritable variance in a population, a process also known as evolutionary rescue (Alexander et al. 2014).

To provide a quantitative framework to link Darwinian selection theory with Mendelian concepts of the inheritance of traits, Wright, Haldane and Fisher developed mathematical models and theories that became the foundations of population genetics (Wright 1931; Haldane 1932; Fisher 1930; Mendel 1866). Dobzhansky (1937) and Mayr (1942) shaped the so-called Neo-Darwinism, the fusion of Darwinian evolution and Mendelian genetics, which today is called evolutionary genetics (Huxley and Julian 1943). They defined five evolutionary factors that change the allele frequencies of a population and consequently affect the gene pool: mutation, recombination, selection, genetic drift and isolation, all of which play a central role in the type and nature of resistance evolution in different species (Dobzhansky 1937; Mayr 1942). In this context, selection plays the most important role in resistance adaptation, as it can lead to a rapid and targeted change in allele frequencies in response to the repeated application of pesticides.

The increasing resistance of agricultural pests and diseases greatly affects the use of insecticides, fungicides and herbicides. The first resistance observation was made in 1914 in the insect species *Quadraspidiotus perniciosus* against the active ingredient lime sulfur, and already at that time it was hypothesized that without preventive measures this would become an increasing problem (Melander 1914). Today, pesticide resistance is an essential problem that not only causes significant economic losses for farmers, but also threatens our food production (Varah et al. 2019; Palumbi 2001). Up to date, resistance to insecticides can be found in more than 625 species (Mota-Sanchez and Wise 2022). Similarly for fungicides, the most important mode of actions have been overcome by different resistance mechanisms in many economically relevant crops (Lucas et al. 2015). The same applies to herbicides, in which numerous modes of action lost their effectiveness quickly after their introduction (see further details in the next section) (Heap 2014a). New alternatives in pesticide development offer new opportunities for pest control. However, it is very likely that resistance development will occur faster than human innovations to counteract it.

1.2 Herbicide resistance in weeds as agricultural challenge

The highest potential agricultural yield losses are caused by weed growth and the associated competition for space, light and nutrients (up to 74%) (Oerke 2006). Therefore, the use of herbicides for weed control is a common practice. This has led to the unintended consequence of herbicide resistance in many different species in recent decades (mainly since 1980) with the broader introduction and increasing use of herbicides (Heap 2014a). The resistance cases are still rising today due to several reasons: herbicides have the highest weed control efficiencies, they are relatively easy to use and they are not costly (Gressel 2011). In rural areas in developing countries where hand weeding was common practice until today, workers are becoming increasingly scarce and herbicides are a cheap alternative (Gianessi 2013). In addition, agricultural best practices to maintain natural soil composition and reduce surface runoff, such as reduced tillage, require increased herbicide use (Moss 2017; Clarke et al. 2000). Furthermore, only few different modes of action are available in many crops and only one new active ingredient (cinmethylin) has been added in the last 30 years (Duke 2012; Campe et al. 2018).

According to the current Herbicide Resistance Action Committee (HRAC) classification code, 28 mode of action (MoA) groups are specified (Herbicide Resistance Action Committee 2022), of which 18 have been overcome by resistance up until the present time (Figure 1) (Heap 2022). The first cases of herbicide resistance were reported in 1957 in the species *Commelina diffusa* and *Daucus carota.* They evolved resistance against the synthetic auxin

2,4-D (MoA group 4), which has been primarily used to control broadleaf weeds in grass crops (Switzer 1957; Hilton 1957). To date, there are 512 unique cases of herbicide resistance in 266 species reported (Heap 2022).

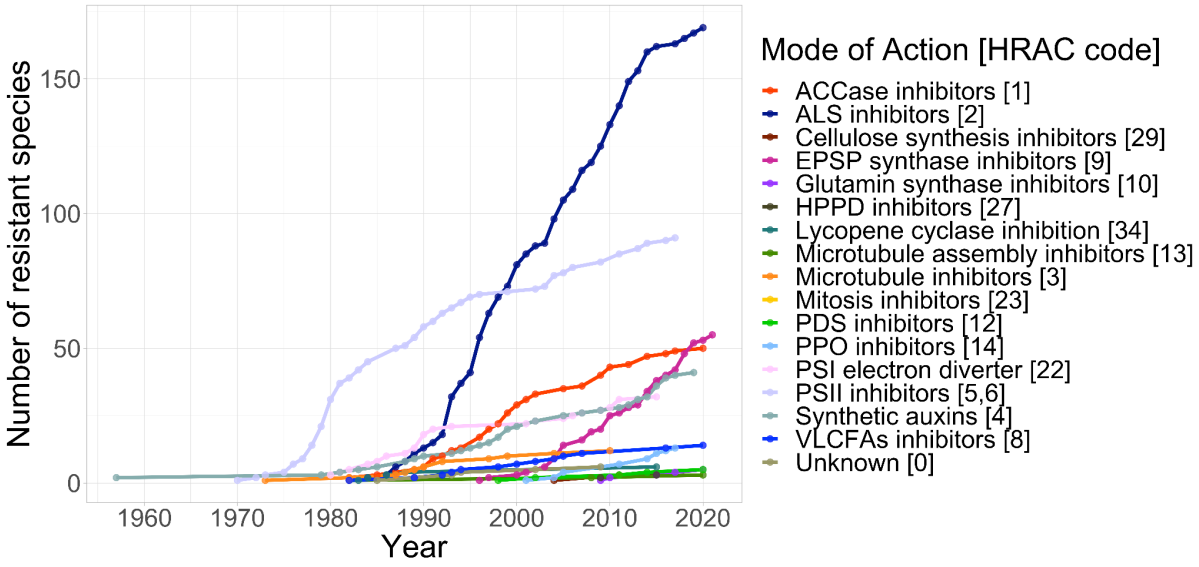


Figure 1: Timeline of the accumulation of herbicide resistant weed species per mode of action (MoA). **Credit:** I. Heap, www.weedscience.org, data downloaded: 05.02.2022 (Heap 2022), adapted by S. Kersten.

The example of **glyphosate** illustrates how quickly widespread resistance can develop as a result of massive herbicide usage: since glyphosate is effective on almost all weed species, it has a wide application range in all kinds of agronomic crops. With the introduction of genetically modified organisms (GMO) crops in 1996 that can survive glyphosate treatments, widespread field applications of glyphosate began. By 2014, herbicide tolerant crops were grown for maize, soybean, and cotton at around 90% of the cultivated area in the United States and sprayed accordingly (United States Department of Agriculture 2020). Until then, the widespread glyphosate application resulted in more than 40 resistant weed species, some of them particularly problematic like *Amaranthus palmeri*, *Amaranthus tuberculatus*, *Ambrosia* spp., *Kochia scoparia*, *Digitaria insularis*, *Sorghum halepense* and *Lolium rigidum* (Gould et al. 2018; Heap 2022). As a consequence, new GMO crops tolerant to synthetic auxins such as herbicides 2,4-D and dicamba were engineered, whereupon resistance to these agents also quickly further increased (Behrens et al. 2007; Busi et al. 2018).

Photosystem II (PS II) inhibitors, another early popular MoA, include triazines, triazinones, triazolinones, uracils, pyridazinones and phenyl-carbamates. The best known of these is

triazine, which was used extensively in maize cultivation in the USA from 1960 to 1990. Large-scale applications led to increasing resistance problems since 1970. Particularly problematic are the genera *Amaranthus, Chenopodium* and *Solanum* sp. in the maize growing areas across the US and Europe (Heap 2014b).

Acetolactate synthase (ALS) inhibitors comprise the MoA group with the largest number of herbicides (sulfonylureas (SUs), imidazolinones (IMIs), triazolopyrimidines, pyrimidinyl-thio-benzoates), and the first resistance cases were reported already four years after market introduction in the 1980s (Heap 2014b). Nevertheless, since their introduction they have been widely preferred due to their low application rates, low mammalian toxicity, broad spectrum of weed control and wide application windows (Tranel and Wright 2002). The invention of 'leaf-active' safeners in the 1990s, further expanded the application range of both ALS and Acetyl-CoA carboxylase (ACCase) herbicides, as they could achieve selectivity in cereal crops when used in combination (reviewed in Kraehmer et al. 2014). As a consequence, an exceptionally high number of species (169) developed resistance to ALS inhibitors, which makes them the most prominent example for convergent evolution of herbicide resistance (Baucom 2016).

Acetyl-CoA carboxylase (ACCase) inhibitors such as aryloxyphenoxy-propionates (FOPs), phenylpyrazolines (DENs), and cyclohexanediones (DIMs) are among the most effective herbicides against grass weeds and thus have been widely used predominantly in cereal crops (Délye et al. 2005). ACCase inhibitors were introduced in 1974, and until 2014 a total of 43 grass species had acquired resistance (Heap 2014a).

The largest number of resistant weed species by far belongs to the Poaceae family, the monocotyledonous grass weeds (Figure 2a) (Heap 2022). Historically, grass weeds in the agricultural landscape have adapted and synchronized their life cycles in parallel with the major crops wheat, rice and corn from the same plant family (Kraehmer 2019). Hence, the most problematic weed species are annual outcrossing and very prevalent in the field flora (Heap 2014b). They have large populations with high genetic diversity from which they can adapt very rapidly to strong herbicide selection pressures. Due to limited availability of MoAs, mostly ACCase and ALS inhibitors are used as postemergence herbicides in cereal crops to control grass weeds. As a consequence of this, an overwhelmingly high number of species have developed resistance to these two MoAs (Figure 2b) (Heap 2022). In particular, resistant grass weeds like *Avena fatua, Lolium rigidum, Echinochloa crus-galli* and

Alopecurus myosuroides are widespread and have the greatest economic impact on agricultural crops (Heap 2014b).

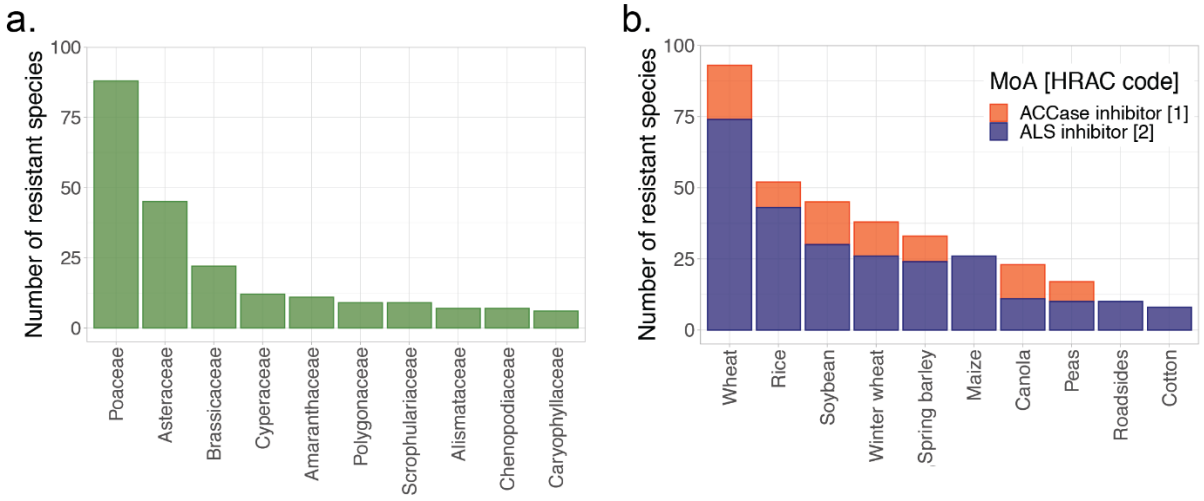


Figure 2: Number of herbicide resistant species grouped by plant families and crops. **a.** The ten weed families with the highest number of resistant species for all modes of action (MoA). **b.** The ten field crops with the highest emergence of resistant weeds to the two MoA groups of ACCase inhibitors [1] and ALS inhibitors [2]. **Credit:** I. Heap, www.weedscience.org, data downloaded: 05.02.2022 (Heap 2022), adapted by the author.

To date, resistant weed populations have been reported against almost every MoA available (Gould et al. 2018). Resistance is also exacerbated by the fact that access to MoAs is limited due to stricter regulations and nowadays the cost of discovering new pesticides is eight times what it used to be 50 years ago (Duke 2012).

1.3 Molecular basis of herbicide resistance

Evolutionary rescue prevents extinction of populations in a changing environment based on adaptive, beneficial alleles that increase in number and maintain the natural population (Orr and Unckless 2014). This has led to different types and mechanisms of herbicide adaptation, which can be grouped into two broad categories of resistance: target-site resistance (TSR) and non-target-site resistance (NTSR). TSR prevents the herbicide molecules from binding to the target protein by altering the structure of the binding site. Alternatively, TSR can also result from overexpression of the target-site gene, generally caused by gene amplification (Gaines et al. 2010). On the other hand, NTSR involves various metabolic degradation processes. These are enhanced metabolism, reduced herbicide uptake and/or translocation and increased herbicide sequestration (Délye 2013; Heap 2014a).

TSRs typically arise from single mutations in the herbicide target gene and often evolve rapidly as a response to the repeated use of the same MoA. These large-effect mutations usually result in a “specialist” type of resistance, mostly to only one herbicide mode of action (Comont et al. 2020). NTSR originates from the upregulation of a number of metabolic and stress response pathways and leads to a "generalist" resistance phenotype. NTSR arises predominantly from heterogeneous selection environments such as application of herbicide mixtures and involves multiple genes (Comont et al. 2020). Both TSR and NTSR can occur as cross resistance or multiple resistance. Cross resistance means a single mechanism leads to resistance to multiple herbicides, whereas multiple resistance refers to several resistance mechanisms combined in a single plant. This evolves primarily through repeated selection with different MoAs or through gene and pollen flow. It is also possible for multiple resistance mechanisms to coexist in a population (Heap 2014b).

Target-site resistance:

Glyphosate inhibits 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), a key enzyme in the shikimic acid pathway, responsible for the biosynthesis of the aromatic amino acids phenylalanine and tyrosine (Steinrücken and Amrhein 1980). There are two known TSR mechanisms by which glyphosate resistance can be conferred. The first one is through changes in the codon positions Gly 101, Thr 102 or Pro 106 in the *EPSPS* gene, with the latter being the most widespread due to the full preservation of normal EPSPS activity (Healy-Fried et al. 2007). However, a significant fitness cost for the double mutation Thr-102-Ile/ Pro-106-Ser was reported in the weed species *E. indica* (Han et al. 2017). The second mechanism is *EPSPS* gene amplification (Gaines et al. 2010). The first appearance of resistance through *EPSPS* gene amplification was found in *A. palmeri*, with the gene duplicated in a range from 4 to over 100 copies across the chromosomes in the genome, and corresponding enhancement in mRNA expression and protein level (Gaines et al. 2010). Interestingly, *A. tuberculatus* shows a similar level of resistance, but only 4-10 duplicated *EPSPS* copies, which rather than spread throughout the genome, occur in tandem, most likely originating from unequal recombination mediated by repetitive DNA (Dillon et al. 2016). Since then, additional species with *EPSPS* gene duplication have been reported, with the trend being that higher *EPSPS* copy number leads to higher resistance (reviewed in Table 2, Gaines et al. 2020). Unlike non-synonymous mutations that confer resistance, higher copy numbers in *A. palmeri* were not associated with fitness costs, which may be the reason for their widespread emergence (Vila-Aiub et al. 2014).

ALS is the first enzyme in the biosynthetic pathway for the branched-chain amino acids valine, leucine and isoleucine, and it is therefore targeted by several herbicide classes (reviewed in Zhou et al. 2007). Herbicide resistance emerged rapidly, because the *ALS* site of action can accumulate many different point mutations without losing its functionality (Heap 2022). This is because the ALS herbicide molecules do not mimic the substrate for the enzyme, but block the opening of the channel to the catalytic domain, thus providing a more general mechanism leading in most cases to multiple resistances to several classes of ALS inhibitors (McCourt et al. 2006). Changes in the following eight amino acids lead to resistance: Ala 122, Pro 197, Ala 205, Asp 376, Arg 377, Trp 574, Ser 653, and Gly 654, with mutations at Pro 197 being the most widespread, followed by Trp 574 (Tranel and Wright 2002; Délye and K Boucansaud 2007). Although two studies have reported a negative fitness effect on the Pro-197-His and Trp-574-Leu mutations in *Lactuca serriola* and *Amaranthus powellii* (Tardif et al. 2006; Alcocer-Ruthling et al. 1992), a more recent study found no significant fitness effects on plant growth for Pro-197-Ser and Trp-574-Leu mutations in *Lolium rigidum* (Yu et al. 2010).

ACCase inhibitors block the first step in fatty acid synthesis by inhibiting the catalytic activity of the enzyme acetyl coenzyme carboxylase (Walker et al. 1988). They act specifically on grasses because they target the homomeric plastidic ACCase, which is almost exclusively found in monocots and encoded in the nuclear genome (Incledon and Hall 1997). ACCase inhibitors acquire target-site resistance for the same reason as ALS inhibitors. There are a number of point mutations that lead to amino acid changes and thus prevent binding of the herbicide in the target protein. The known *ACCase* codon positions are: Ile 1781, Trp 1999, Trp 2027, Ile 2041, Asp 2078, Cys 2088, Gly 2096 (reviewed in Kaundun 2014). Depending on the mutation, changes confer resistance to one or several of the three different classes of ACCase inhibitor herbicides, Ile-1781-Leu and Asp-2078-Gly being resistant to all three classes (Beckie and Tardif 2012). Furthermore, fitness costs also differ between the TSR mutations. While Ile-1781-Leu and Ile-2041-Asn show no effect on fitness, plants carrying the Asp-2078-Gly allele have lower plant height, vegetative dry biomass, and seed set. Similarly, plants with the Trp-2027-Cys allele have a lower seed production (Menchari et al. 2008; Du et al. 2019; Vila-Aiub et al. 2015).

PS II inhibitors target the maternally inherited chloroplastic *psbA* gene to impede photosynthesis. By binding to the D1 protein within the PS II complex, the herbicide compounds compete with plastoquinone at the plastoquinone binding site. This inhibits the PS II electron transport and as a consequence nicotinamide adenine dinucleotide phosphate

hydrogen (NADPH) and adenosine triphosphate (ATP) synthesis in the chloroplasts (Gronwald 1997), which in turn compromises CO_2 fixation and causes oxidative stress. Triazine herbicides have been mostly used in maize since 1950. Resistance is primarily caused by a single amino acid substitution in the *psbA* gene, Ser-26-Gly, that has independently evolved on a global scale. It prevents triazine binding, but allows plastoquinone binding. However, it comes with deleterious fitness effects on CO_2 assimilation and plant development (Ireland et al. 1988; Ort et al. 1983). In addition, there are other five less relevant TSR mutations for non-triazine binding herbicides in the D1 protein that also cause resistance (reviewed in Powles and Yu 2010).

Synthetic auxins mimic the activity of the plant hormone auxin (indole-3-acetic acid, IAA) and thus affect cell growth and plant development. Different resistance mechanisms target four major protein classes essential in the auxin pathway: auxin transporters (PIN, ABCB, AUX/LAX), transcriptional repressors (Aux/IAAs), auxin response factors (ARFs), and the Skp1-Cullin-F-box TIR1/AFB E3 ubiquitin ligase complex ($SCF^{TIR1/AFB}$) (reviewed in Todd et al. 2020). TSR mechanisms are highly specific to different auxin herbicide classes, but still require more research investigation, including sequencing information on the involved genes.

In conclusion it is evident that TSR for the various MoAs repetitively and independently evolve at the same positions even across species. This phenomenon of widespread convergent evolution in herbicide resistant field populations quite likely originates from genomic constraints, thus favoring certain mutational target positions (Baucom 2016).

Non-target-site resistance:

NTSR is also widespread, found in many species and causing major resistance problems due to cross-resistance. Much less is known about NTSR than TSR in weed species and it is therefore still a research field that needs considerable investigation (Délye et al. 2011). There are four main gene superfamilies that are suspected to play a key role: genes encoding cytochrome P450 monooxygenases, glutathione S-transferases, ATP-binding cassettes (ABC) transporters, and glycosyltransferases.

Cytochrome P450 monooxygenases are bound to the endoplasmic reticulum and mostly located at the membrane. They represent one of the largest protein families with the highest diversity in plants. They are involved in the synthesis of hormones, fatty acid derivatives, defense compounds and metabolism of endogenous substances (Schuler and

Werck-Reichhart 2003). The inactivation of herbicide molecules occurs usually through hydroxylation or dealkylation. They have been found to inhibit several classes of herbicides and preferentially cause cross-resistance (Barrett 1995).

Glutathione S-transferases catalyze the conjunction of glutathione with the herbicide molecules, which are then sequestered in the vacuoles or excreted via the root tips. Expression of glutathione transferases is induced primarily in the cytosol as a response to plant injuries and oxidative stress through herbicides (reviewed in Dixon et al. 2002). Their resistance function was first demonstrated in triazine herbicides and later applied to genetically engineered herbicide resistant crops (Milligan et al. 2001).

ABC transporters are a large family of proteins responsible for cross-membrane transport. They are involved in reduced translocation, excretion, or sequestration of compounds (Schulz and Kolukisaoglu 2006) and their overexpression has been linked to herbicide resistance (Windsor et al. 2003; Jo et al. 2004).

Glycosyltransferases are commonly transcriptionally activated as a common response to plant injury and oxidative stresses such as herbicides. They cause glycosylation of molecules and detoxify a variety of plant metabolites, phytotoxins and xenobiotics (Bowles et al. 2005). Their expression can be also induced by certain chemical agents. As with cytochrome P450 monooxygenases and glutathione S-transferases, this property was used for the development of safeners, which are chemicals that protect crops from herbicide damage (Hatzios and Burgos 2004).

A large number of studies has focused on differential expression analysis to find potential candidate genes for NTSR in a variety of plant and weed species (Iwakami et al. 2014; Dücker et al. 2019; Wright et al. 2018; Nakka et al. 2017; Lu et al. 2015; Cagnac et al. 2004). Unfortunately, there are only few genomic studies in natural populations to identify adaptive alleles that confer resistance (Kreiner et al. 2021; Van Etten et al. 2020). Kreiner et al. (2021) conducted a genome-wide association study (GWAS) on resistance to glyphosate in *Amaranthus tuberculatus* in which, in addition to the *EPSPS* gene, they identified several regulatory genes involved in major metabolic detoxification pathways. In another recent study, Cytochrome P450 monooxygenases, glycosyltransferases and ABC transporters have been found in a genome-wide scan in glyphosate resistant populations of *Ipomoea purpurea* (Van Etten et al. 2020). Notably, while one genomic region showed parallel evolution across populations, other regions displayed divergent signals (Van Etten et al. 2020). This highlights

the importance of conducting genomic studies to detect the presence of these adaptive alleles at a species-specific level in natural populations.

1.4 A major threat to crop production: Biology and herbicide resistance of *A. myosuroides*

Alopecurus myosuroides Huds. (common name: blackgrass) is an annual, diploid grass distributed over the temperate climate zone and found primarily on the European continent (Figure 3) (Naylor 1972; Global Biodiversity Information Facility 2021). It grows tufted and slender to a height of 80 cm. The base of the leaf sheath is usually purple-red in color. Leaves are glabrous, lanceolate and approximately 3-17 cm long and 2-8 mm broad with a 5 mm long ligule. The spike-like inflorescence is 4-12 cm long and 3-6 mm wide and turns from green to brown as it matures, which is where the name blackgrass comes from. The spikelets are 4-7 mm long, single-flowered, subsessile, and divided below the glumes, with a geniculate awn (Riches 2008; Naylor 1972).

Due to rapid development of herbicide resistance, it has become the most problematic weed species in cereal crops in many parts of Europe, causing significant yield and economic losses (Moss 2017). A recent study estimated the annual costs of herbicide resistance in *A. myosuroides* in the United Kingdom at £400 million in lost gross profit (Varah et al. 2019).

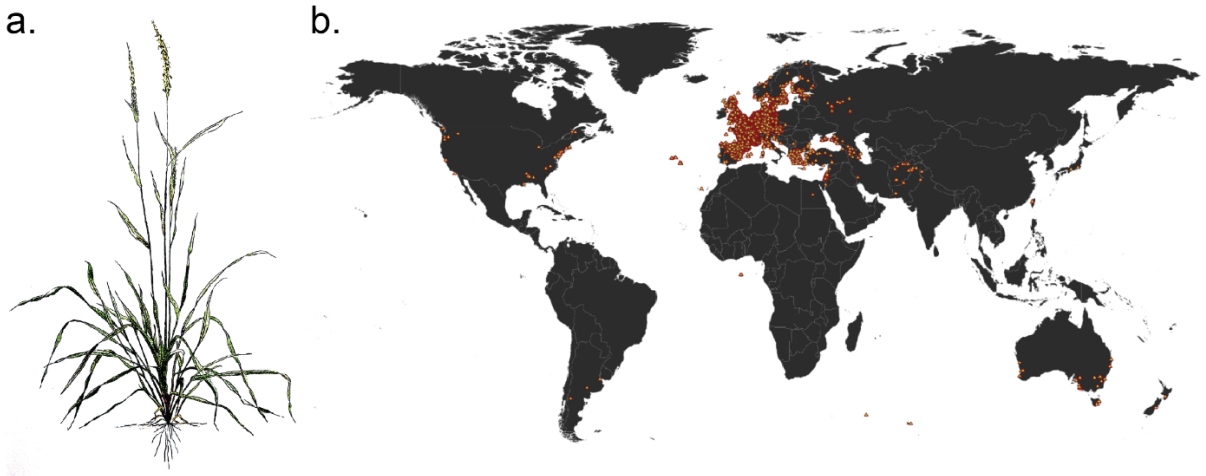


Figure 3: Morphology and geographical distribution of *Alopecurus myosuroides*. **a.** Drawing on watercolor paper with pencils and ink. **b.** Global distribution. Credit: GBIF, www.gbif.org, data downloaded: 10.03.2020 (Global Biodiversity Information Facility 2021), Adapted by the author.

Seeds of *A. myosuroides* generally display primary dormancy, with fresh seed germination varying from 38% to 70% (Colbach et al. 2002). While fresh seeds show a higher

germination rate under light, this effect disappears with increasing seed age of more than 350 days (Colbach et al. 2002). Germination occurs in two cohorts, a main germination in fall and a secondary emergence in spring. The spring cohorts show faster development and early flowering. Therefore, it is suspected that there are two distinct biotypes, which differ in germination and flowering requirements (Wallgren and Avholm 1987; Chauvel et al. 2002). The amount of seeds is usually around 50-100 viable seeds per tiller, and up to 100 tillers (> 5,000 seeds) per plant under greenhouse conditions (Chauvel et al. 2005). More than 900 tillers per m^2 have been observed in the field (Moss 1983). Flowering can be promoted by vernalization during the pre-germination phase (Chauvel et al. 2002). *Alopecurus myosuroides* is an outcrosser with a partial self-incompatibility resulting in a low seed viability after enforced self-fertilization (Chauvel 1991). In the field, *A. myosuroides* synchronizes with the crops, but completes the life cycle slightly before (Moss 1983). It is also affected by high plant densities and suppressed by the crop at high nitrogen levels (Chauvel et al. 2005; Naylor 1972). The dispersal of pollen and seeds is spatially very limited: more than 70% of pollen and seed dispersal takes place within a meter (Colbach and Sache 2001). However, Busi *et al.* (2008) found that in another grass weed species, *Lolium rigidum*, pollen dispersal, and therefore resistant plants, can reach 3,000 m, but it is uncertain whether this would also be applicable to *A. myosuroides*. Long-distance seed dispersal is suspected to occur primarily through agricultural machinery (Wallinga et al. 2002). Seed viability in soil declines 74% per year and only a small proportion survives five years in soil (Moss 1985).

The first herbicide resistance in *A. myosuroides* was detected in Israel in 1979 against methabenzthiazuron, a PS II inhibitor (Heap 2022). Shortly after, in 1982, resistance to chlortoluron and isoproturon, which also belong to the PS II inhibitors, was reported in the United Kingdom (Moss and Cussans 1985). In the following decades, it became evident that *A. myosuroides* evolves resistance rapidly, including multiple resistance to several MoAs (C. Délye et al. 2011; Petit et al. 2010). Until today, blackgrass has acquired resistance to five different MoA groups, including various active ingredients and cross resistances (Heap 2022).

Due to high reliance on herbicides in the past and limited number of available MoAs, resistance adaptation in *A. myosuroides* is especially problematic in winter cereals (Moss 2017). In particular, ALS- and ACCase-inhibiting herbicides have widely lost their efficacy. A study in Germany found an increase in TSR mutations in *ACCase* from 5.0% in 2004 to 54.3% in 2011, and a similar trend was quantified in *ALS* from 0.8% in 2007 to 13.9% in

2012, without mentioning the large fraction of NTSR (Rosenhauer et al. 2013). Délye et al. (2007) conducted a resistant survey in northeastern France in 2000. They harvested seeds from 243 fields of winter wheat with suspected herbicide resistance. Of 22,300 seedlings analyzed, 99.2% were resistant, of which 58.6% contained TSR alleles while the remaining resistance was attributed to NTSR. Several follow-up studies, including a European resistance survey of the *ACCase* gene further confirm high overall levels of resistance with different resistance mechanisms involved (Délye et al. 2010; Petit et al. 2010; Petit et al. 2010; Délye et al. 2008). In the United Kingdom, the resistance situation in blackgrass is probably the most severe. A nation-wide study of 138 blackgrass populations in 2014 revealed that 79% of the populations acquired multiple resistance to several MoAs of herbicides (Hicks et al. 2018).

Target-site resistance to ALS and ACCase inhibitors can arise through seven different mutations in the codon region of each gene (Délye et al. 2005; Xu et al. 2014; Tranel and Wright 2002; Délye and K Boucansaud 2007, see also previous section). Depending on the mutation, variable cross-resistances to specific ACCase inhibitor groups are found. Ile-1781-Leu and Asp-2078-Gly confer resistance to all three groups: FOPs, DIMs and DENs. Plants with the Trp-2027-Cys and Ile-2041-Asn mutations survive treatments with FOPs and DENs, while Gly-2096 confers resistance exclusively to FOPs (Délye 2005; Délye et al. 2008; Petit et al. 2010). The degree of cross-resistance of TSRs is thought to be one of the factors that determines the frequency at which they are found (Gaines et al. 2020; Powles and Yu 2010). Another factor is their respective fitness effect. Therefore, the most frequently occurring mutation, Ile-1781, leads to resistance to all three ACCase inhibitor groups and appears to have no deleterious fitness effect (Menchari et al. 2008; Délye et al. 2013). In ALS, the most common resistance mutation in blackgrass is Pro-197, followed by Trp-574, comparable to other weed species (Heap 2022; Marshall and Moss 2008). While Trp-574 confers resistance to ALS inhibitor groups SUs and IMIs, Pro-197 leads to resistance only to SUs (Gaines et al. 2020; Marshall and Moss 2008). However, as described earlier, there are at least eleven possible nucleotide substitutions that lead to amino acid changes that confer resistance through the Pro-197, hence these mutations arise more frequently (summarized in Powles and Yu 2010).

In addition to TSR, NTSR in *A. myosuroides* is widespread and often confers cross resistance to other MoAs (Délye 2013). There are several candidate gene families identified in *A. myosuroides* of being involved in NTSR, such as glutathione-S-transferases (Cummins et al. 2013; Brazier et al. 2002; Dücker et al. 2019; Cummins et al. 1999; Franco-Ortega et

al. 2021), Cytochrome P450 monooxygenases (Hall et al. 1997; Franco-Ortega et al. 2021; Hyde et al. 1996; Letouzé and Gasquez 2003), UDP-glycosyltransferases (Franco-Ortega et al. 2021) and ABC transporters (Franco-Ortega et al. 2021). There is also compelling evidence that the genetic architecture of NTSR is complex, involving several genes, and that it can vary between and even within populations (Petit et al. 2010; Franco-Ortega et al. 2021; Cai et al. 2021). The detailed genetic architecture is unknown and given the variable, but nearly always present, fraction of resistance not explained by known TSR mutations in every major survey of agricultural fields, NTSR should be a main focus of research in the coming years (Rosenhauer et al. 2013; Comont et al. 2020; Délye et al. 2010; Délye et al. 2007).

1.5 Molecular population genetics in herbicide resistance

Modern molecular population genetics is the study of the dynamics of mutations and their variation in allele frequencies, generally across the entire genome. These patterns are captured by markers that describe genetic variation in individuals and populations. Although genetic markers were first used to determine ABO blood groups as early as 1900 (Landsteiner 1900), the beginnings of molecular population genetics can be traced back to the studies of Harris, Lewontin, and Hubby in 1966, who were the first to use allozymes to describe genetic variation (Harris 1966; Lewontin and Hubby 1966). However it took until 1982 for the first studies that described variation on a nucleotide basis to appear and provide us the necessary tools to make inferences about evolutionary forces as well as the structure and demographic histories of populations (Kreitman 1983; Aquadro and Greenberg 1983; Langley et al. 1982). The advances in sequencing technologies, especially next-generation-sequencing (NGS) over the past two decades offered us high-throughput methods to analyze genome-wide variation in diverse species (1000 Genomes Project Consortium 2012; 1001 Genomes Consortium 2016). In particular, studies on non-model species have become feasible and affordable due to these improvements in sequencing technologies and protocols. Therefore, it is not surprising that there is now a greater focus on population genomic studies in the field of weed science to decipher the evolutionary mechanisms that drive the spread of resistance (Küpper et al. 2018; Martin et al. 2020; Gaines et al. 2021; Ravet et al. 2021; Evans et al. 2018; Kreiner et al. 2019; Kreiner et al. 2022). One sequencing strategy is whole-genome shotgun (WGS) sequencing, which provides comprehensive analysis possibilities for herbicide adaptation, such as selective sweep detection (Kreiner et al. 2019). Unfortunately, WGS sequencing is associated with relatively high sequencing costs. Therefore, reduced representation methods, such as restriction site-associated DNA sequencing (RADseq) and exome capture, offer the

possibility of obtaining a subset of genome-wide markers at a reasonable cost and can be performed reference free (Andrews et al. 2016).

Several recent population studies in weeds that have made use of RADseq markers concluded that weeds in agricultural fields tend to have high genetic diversity, multiple parallel adaptation events, low genetic differentiation between populations and the possibility of long distance gene flow (Küpper et al. 2018; Martin et al. 2020; Gaines et al. 2021). Similar results were obtained in two other studies using simple sequence repeats (SSR) in *Bassia scoparia* and exome capture in *Panicum virgatum* (Ravet et al. 2021; Evans et al. 2018). Although they represent only a fraction of the genome, reduced representation methods are still useful for reflecting genetic diversity and providing valuable ecological and evolutionary insights into the species of interest.

To investigate the genetic diversity and the population structure in *A. myosuroides*, I adapted two reduced representation methods: A double-digest RADseq (ddRADseq) protocol using the two restriction enzymes Mph1103I and EcoRI (see Chapter 1, Lang et al. 2020, protocol: https://github.com/SonjaKersten/Laboratory-protocols/tree/master/ddRAD_protocol) and an exome capture with custom baits (*unpublished*, see Addendum in Chapter 1). Furthermore, I generated a high quality Pacific Biosciences (PacBio) reference genome from a single plant of a sensitive German population (Kersten et al. 2021). First, I conducted a local collection of 21 field populations in a variety of summer and winter annual crops in southern Germany (*unpublished*, Tuebingen_RADseq_collection.xlsx). Nine farmers were represented with two to four fields, including two organic farmers (SA and FR). Field histories were documented and 10-20 plants were sampled evenly distributed across each of the fields (in total 330 samples). I generated ddRADseq libraries and multiplexed all samples by making use of a dual barcoding system and sequenced them on three HiSeq Illumina lanes in paired-end mode and with a read length of 150 bp. After adapter trimming (Martin 2011), alignment to the reference genome (Li 2013), variant calling (Van der Auwera et al. 2013) and filtering (Puritz et al. 2014; Danecek et al. 2011), I obtained 146,139 single nucleotide polymorphisms (SNPs) and a mean coverage of 24.3x, with 98% of samples having coverage greater than 10x for further analysis.

Although I found only low genetic differentiation (fixation index [F_{ST}] range: 0.019 to 0.080), the populations clearly clustered by geographic location, with the most Eastern fields, BM and FR, each forming a separate cluster (Figure 4a). Altitude may also contribute to the differentiation since the fields of farmers BM and FR were located at a higher altitude (753

m.a.s.l and 733 m.a.s.l., respectively) than the fields of the farmers LG, KN, HA, HE, SC, SA and BU (356 - 545 m.a.s.l). It is worth noting that farmer FR, whose fields represent the most differentiated populations according to PC1, has been an organic Demeter farmer since 1954.

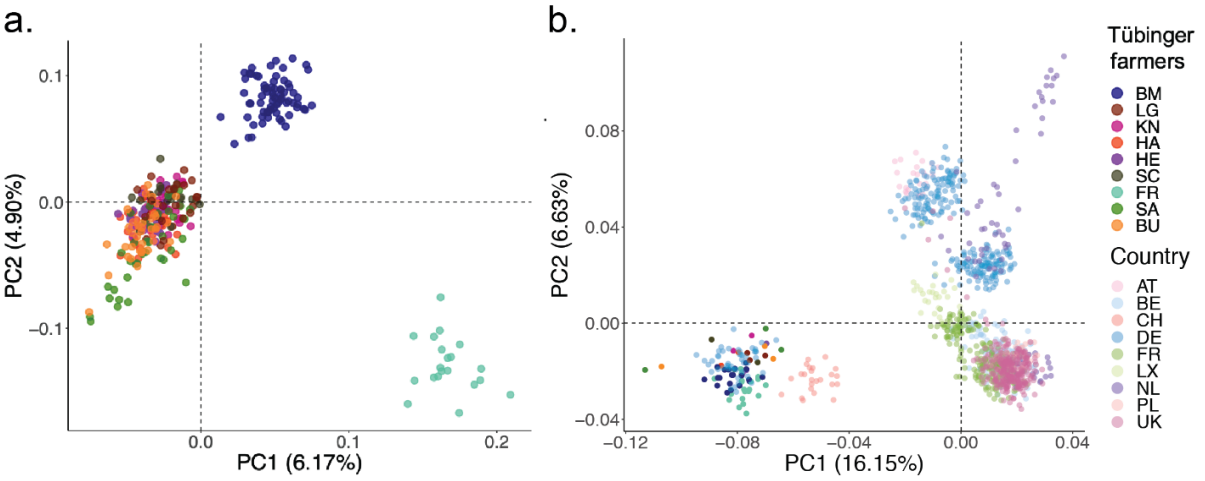


Figure 4: Principal component analysis (PCA) of *Alopecurus myosuroides* field populations, showing the first two principal components (explained variance in parenthesis). **a.** Local collection from southern Germany with nine farmers and different crops. **b.** 20 representative samples from each of the three clusters in (a.), embedded in a Europe-wide *A. myosuroides* collection. The southern German samples (Tübingen) can be found within the bottom-left German cluster of the European set.

Other local or nation-wide studies in *A. myosuroides* based on allozymes and AFLP markers also report high genetic variation and low levels of differentiation (Menchari et al. 2007; Chauvel and Gasquez 1994; Cavan et al. 1998). In contrast to these previous studies and a more recent study in the United Kingdom, which also used RADseq markers (Dixon et al. 2020), I found a clear geographic population structure in our local collection (Figure 4a).

The promising results at a very local scale encouraged us to analyze population structure in an European collection with the same method (see Chapter 2, Kersten et al. 2021). For a joint analysis of both data sets, I combined 20 representative samples from each cluster of the South Germany data set with the complete European data set of 1,122 samples from nine countries. Thereby, the local South Germany samples fell into one of the three major German clusters of the European data set (Figure 4b).

1.6 Objectives

Alopecurus myosuroides is a non-model organism with a relatively large genome size of 3.6 Gb. Without a reference genome at hand, the first goal of this study was to adapt a ddRADseq protocol to the species and characterize the population structure in our own local field collection from southern Germany. Then, I further applied it to the European dataset, in which I also generated PacBio long-read amplicons of the two target-site genes *ACCase* and *ALS* from all samples. I then simulated whether the target-site resistance resulted from existing genetic variation or from *de novo* mutations. In the last part of this thesis, I adapted the long-read amplicon workflow to pools to develop a cost-effective and accurate screening method for the TSR locus *ACCase* that can be easily adapted to any gene and species of interest.

The specific goals of this study were:

- Adapt a ddRADseq protocol to *A. myosuroides* and describe the population structure on a local scale.
- Determine the genetic diversity and population structure in an European collection and generation of a high quality PacBio reference genome for an accurate alignment of short read data.
- Perform a bulked-segregant analysis for NTSR against ACCase inhibitors.
- Characterize TSRs in European populations using PacBio long read amplicons for the *ALS* and *ACCase* loci.
- Identify the evolutionary origin of TSR resistances: from standing genetic variation or from *de novo* mutations.
- Adapt the precise long read workflow of amplicon technology to pools to make it more cost effective and high throughput.

2. Chapter 1

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

Received: 2 October 2019 | Revised: 6 March 2020 | Accepted: 30 March 2020

DOI: 10.1111/1755-0998.13168

SPECIAL FEATURE: GENOMICS OF NATURAL HISTORY COLLECTIONS FOR UNDERSTANDING EVOLUTION IN THE WILD

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

Funding information
This work was supported by the German Research Foundation (DFG; project 324876998 of SPP1374) and by the Presidential Innovation Fund of the Max Planck Society.

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

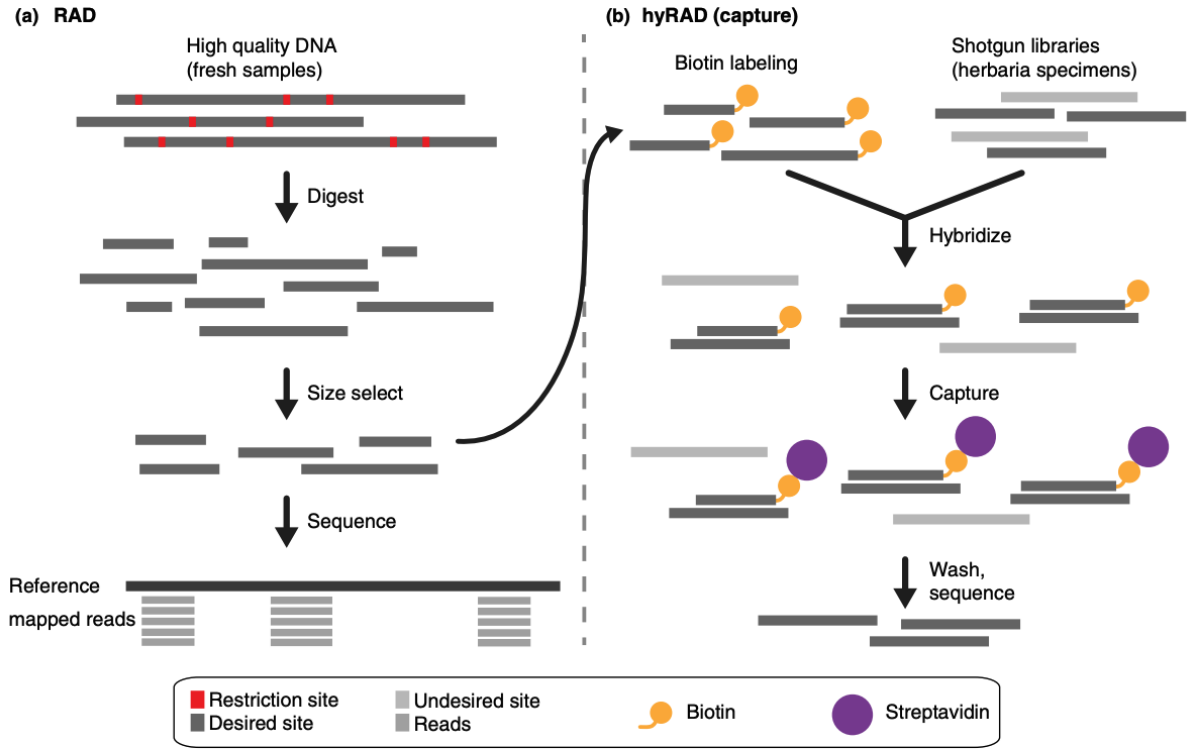


FIGURE 1 RAD and hyRAD method overview. (a) RAD: restriction enzyme(s) (one for RAD, two for double-digest RAD, ddRAD) cut the DNA. Prior to sequencing, the fraction of the genome that will be part of the RAD library is reduced using size-selection, reducing the complexity of the library (reduced-representation 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 [Colour figure can be viewed at wileyonlinelibrary.com]

Fisher, FD0274) and *NsiI*/*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

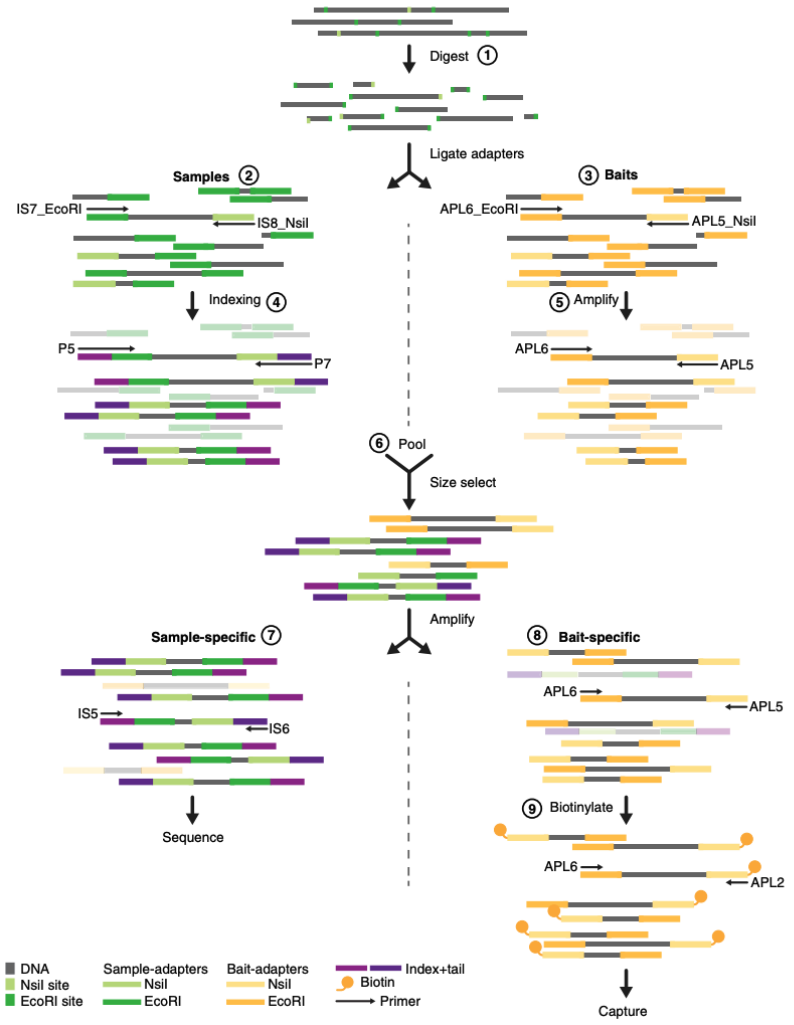


FIGURE 2 Method overview for ddRAD and parallel bait preparation. (1) Fresh high-quality DNA samples are digested with two restriction enzymes (*Eco*RI and *Nsi*I/Mph1103I). Depending on whether the samples will be (2) sequenced or (3) become baits for hybridization capture, different sets of double-stranded 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 [Colour figure can be viewed at wileyonlinelibrary.com]

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

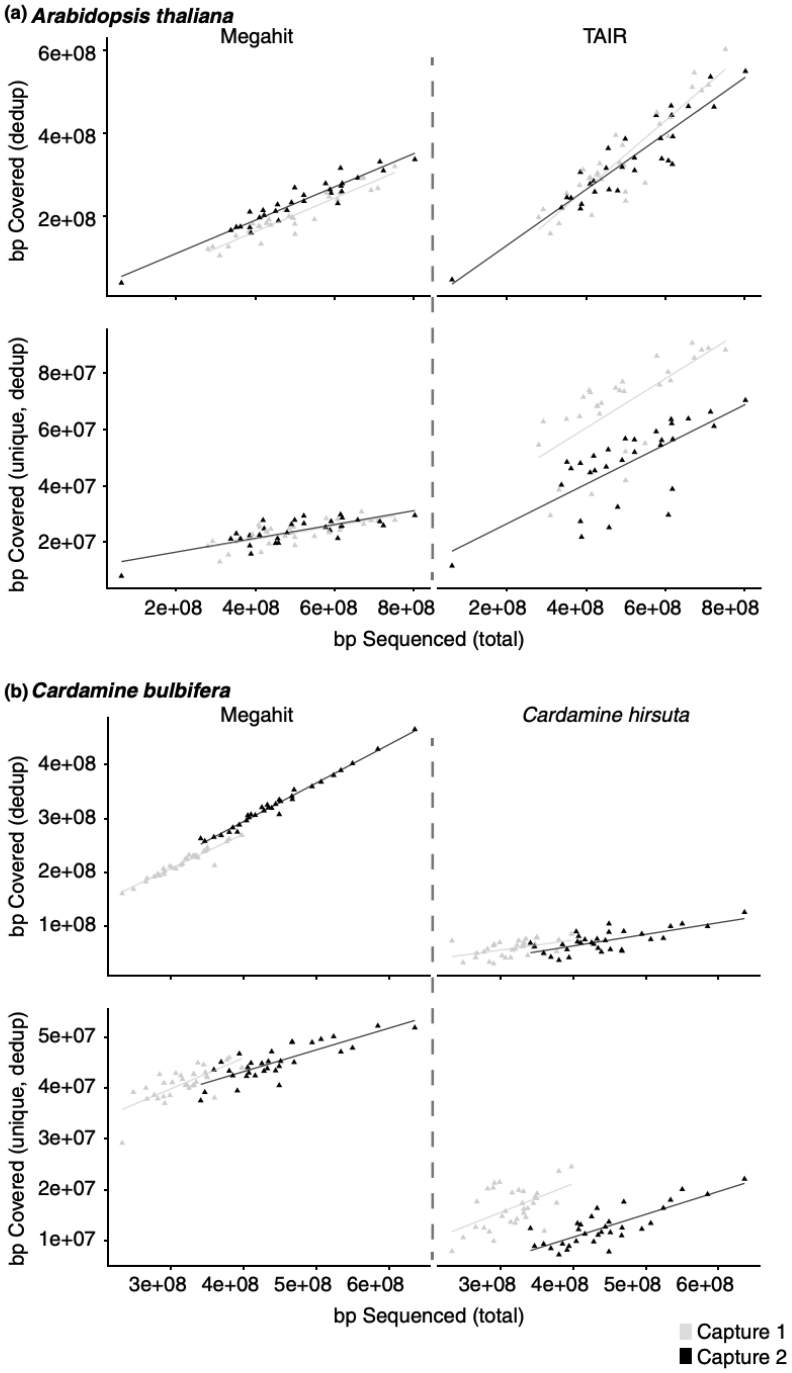


FIGURE 3 Efficiency of a single versus two subsequent rounds of hyRAD captures. Comparison of total and unique base pairs covered per base pair sequenced, mapping aDNA sequences against either a published whole-genome reference or the ddRAD-based MEGAHIT assembly. (a) *Arabidopsis thaliana* samples, 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

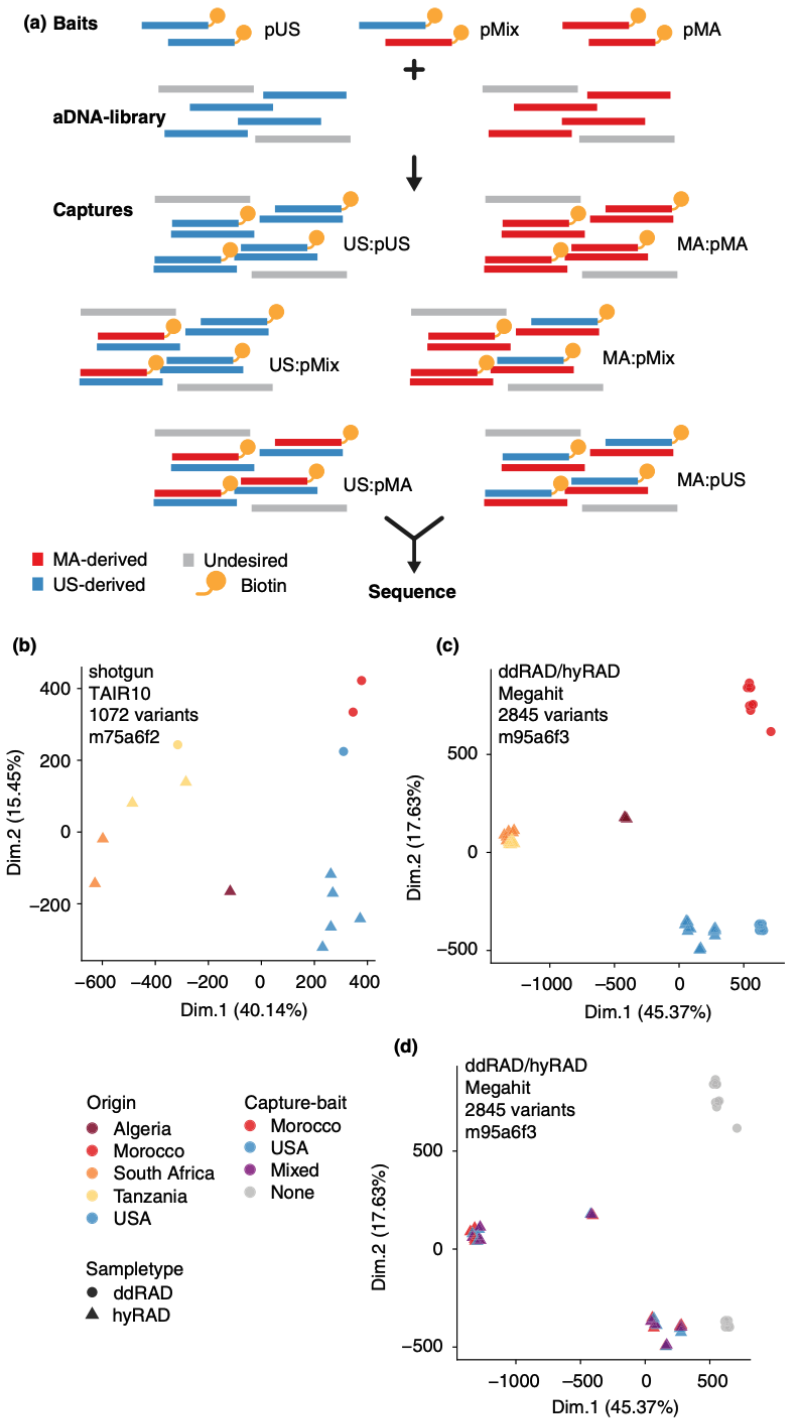


FIGURE 4 *Arabidopsis thaliana* pilot capture. (a) Pilot design with biotinylated baits made from fresh tissue of US (blue) and Moroccan plants (MA, red) that are used to capture ancient DNA libraries from the same geographical locations. For each library, three captures were performed, with either the 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) [Colour figure can be viewed at wileyonlinelibrary.com]

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

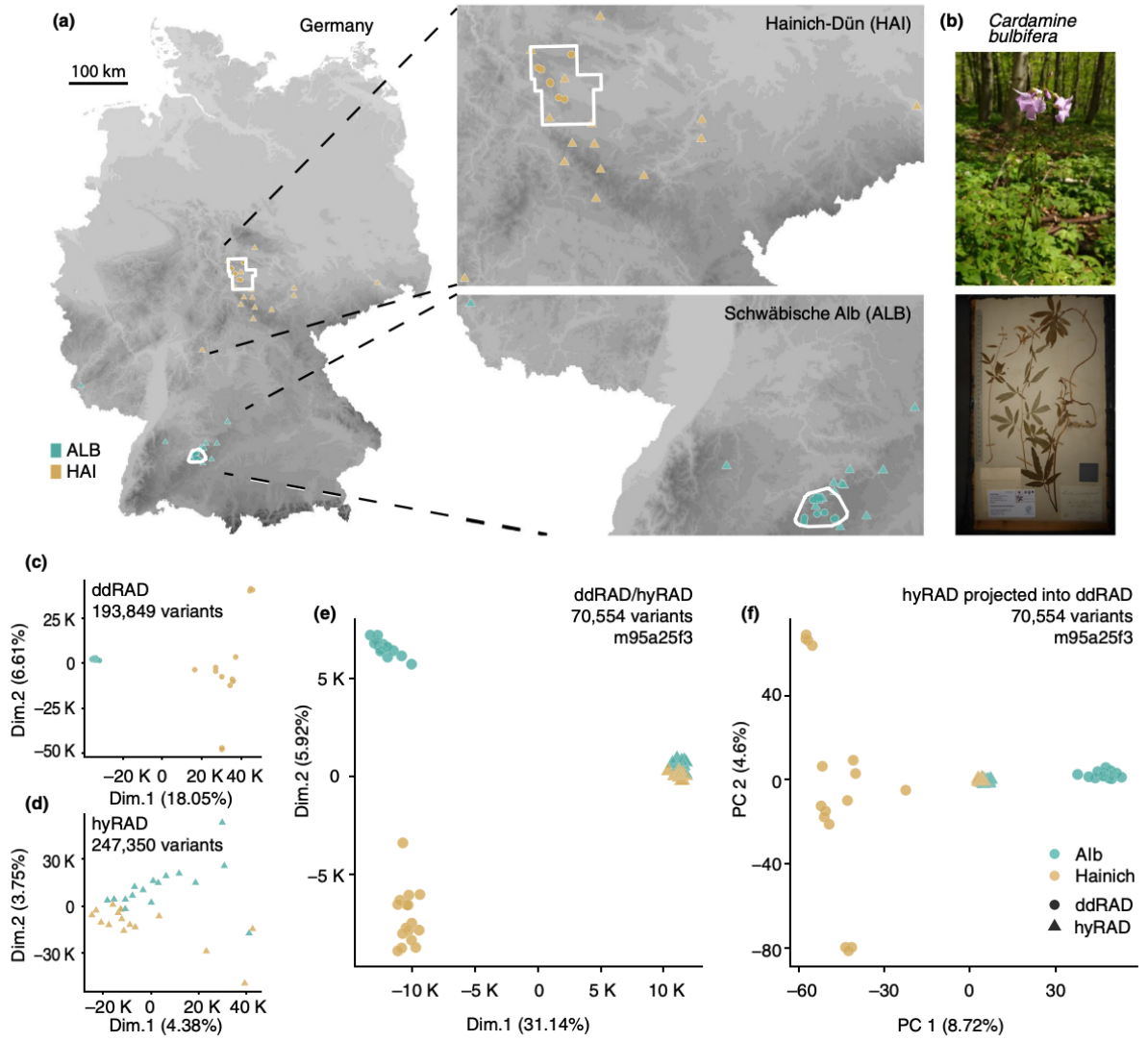


FIGURE 5 ddRAD and hyRAD with the nonmodel species *Cardamine bulbifera*. (a) Overview and zoomed maps of Germany showing the geographical origin of samples; circles represent contemporary samples, triangles historical samples, turquoise colour for samples from Schwäbische Alb, beige from Hainich, and 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) [Colour figure can be viewed at wileyonlinelibrary.com]

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 ($1.75–7.43 \times 10^6$), 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.

ACKNOWLEDGEMENTS

We thank the Tübingen botanical garden for fresh plant samples, Angela Hancock for sharing seeds of the Moroccan *A. thaliana* accessions (Elh-2, Arb-0), Gautam Shirsekar for sharing unpublished short-read sequencing data for H2081, and Guido Brandt for advice on operating the Illumina platform using different sequencing primers simultaneously. We are grateful to Cornelia Krause, Anette Rosenbauer and Jochen Müller from the herbaria in Tübingen, Stuttgart and Jena, respectively, for their introduction to and help in the herbaria, and the kind permission to sample specimens. We thank Fernando Rabanal for help with initial bioinformatic processing of sequencing information, our collaborators on the DFG-financed project Oliver Bossdorf, Franziska Willems and J. F. Scheepens for discussion, and The AGE group and Moises Exposito-Alonso for discussion and input. We thank Detlef Weigel (MPI Tübingen) and Dominique Bergmann (Stanford University) for supporting P. L. M. Lang during the final stages of the project. We thank the managers of the three Exploratories, Kirsten Reichel-Jung, Iris Steitz, and Sandra Weithmann (Alb) and Katrin Lorenzen and Juliane Vogt (Hainich) and all former managers for their work in maintaining the plot and project infrastructure; Christiane Fischer and Jule Mangels for giving support through the central office, Michael Owonibi and Andreas Ostrowski for managing the central database, and Markus Fischer, Eduard Linsenmair, Dominik Hessenmöller, Daniel Prati, Ingo Schöning, François Buscot, Ernst-Detlef Schulze, Wolfgang W. Weisser and the late Elisabeth Kalko for their role in setting up the Biodiversity Exploratories project. The work has been (partly) funded by the DFG Priority Program 1374 "Infrastructure-Biodiversity-Exploratories" (324876998). Fieldwork permits were issued by the responsible state environmental offices of Baden-Württemberg and Thüringen.

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.

How to cite this article: Lang PLM, Weiß CL, Kersten S, et al. Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Mol Ecol Resour*. 2020;20:1228–1247. https://doi.org/10.1111/1755-0998.13168

2.2 Supplementary

Supporting Information for:

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

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

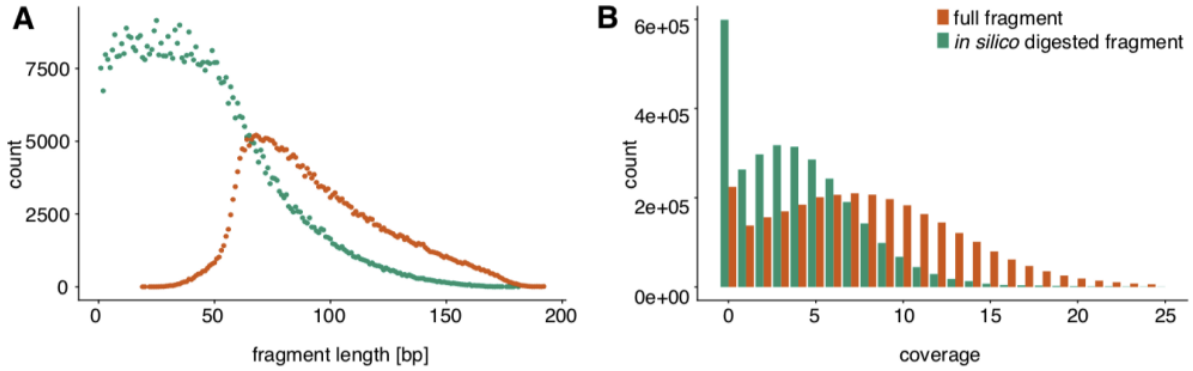
[1]Research Group for Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany
[2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany
[3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[4]Centre for Life's Origins and Evolution, Department of Genetics, Evolution, and Environment, University College London, London, WC1H 0AG, UK
[5]Department of Biology, Stanford University, Stanford, CA, USA 94305
[6]Department of Genetics, Stanford University, Stanford, CA, USA 94305

Table of Contents

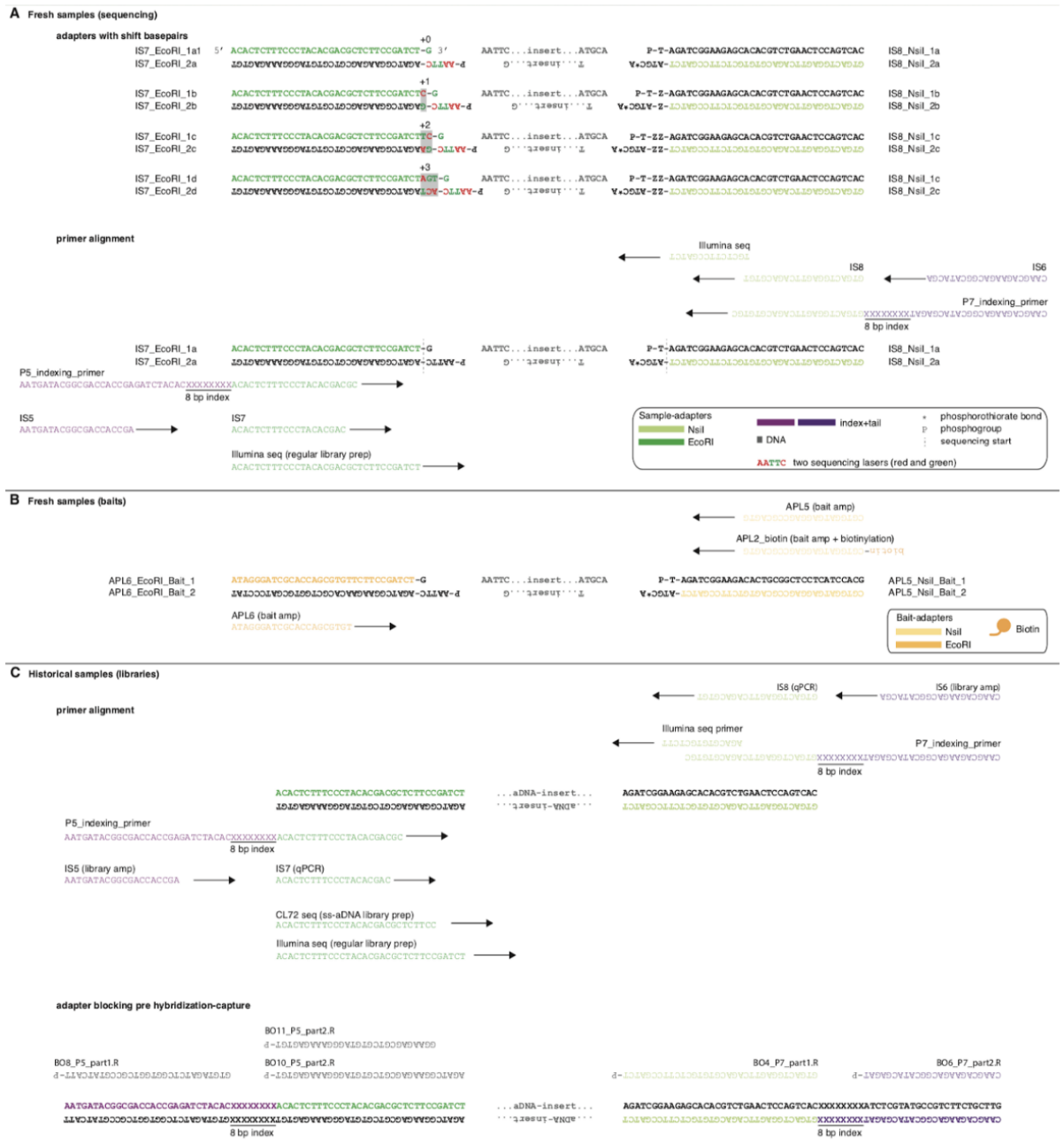
Supplemental Tables

Supplemental Figures

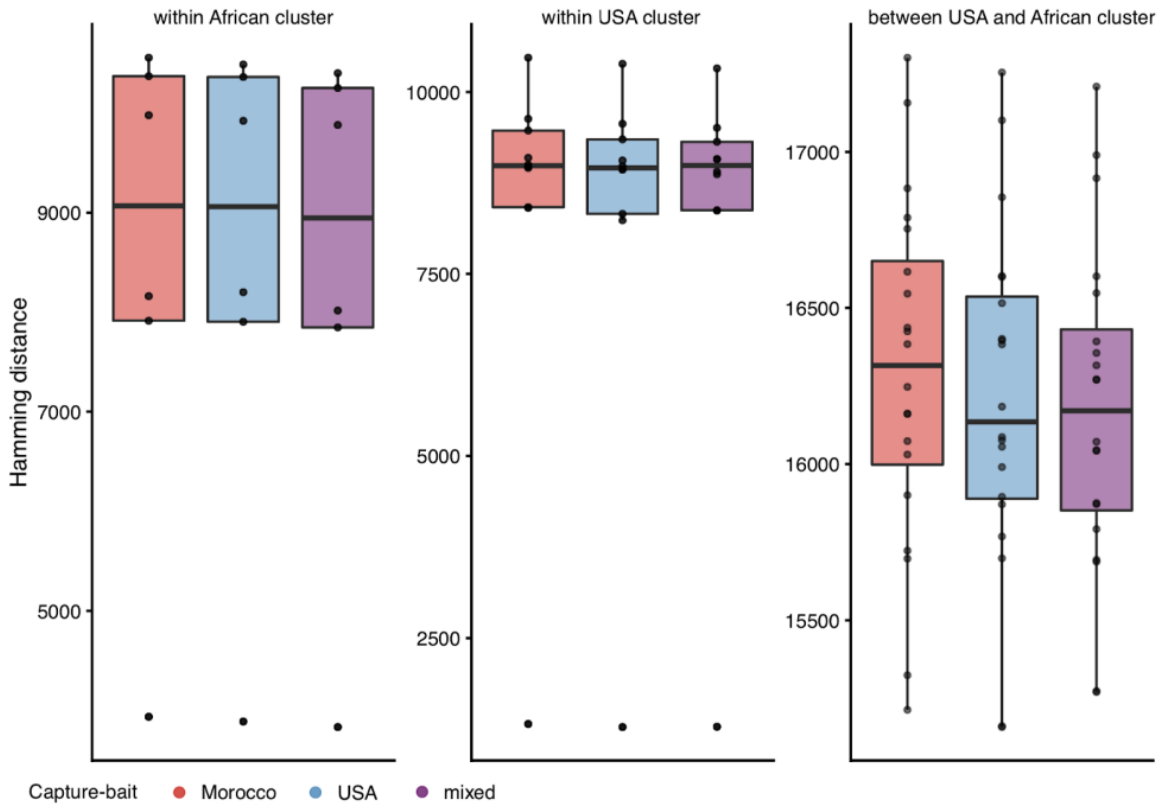
*** Supplemental Tables 3 and 5, and a high-resolution version of Figure S2 are available online at https://github.com/patlang/phenotech**



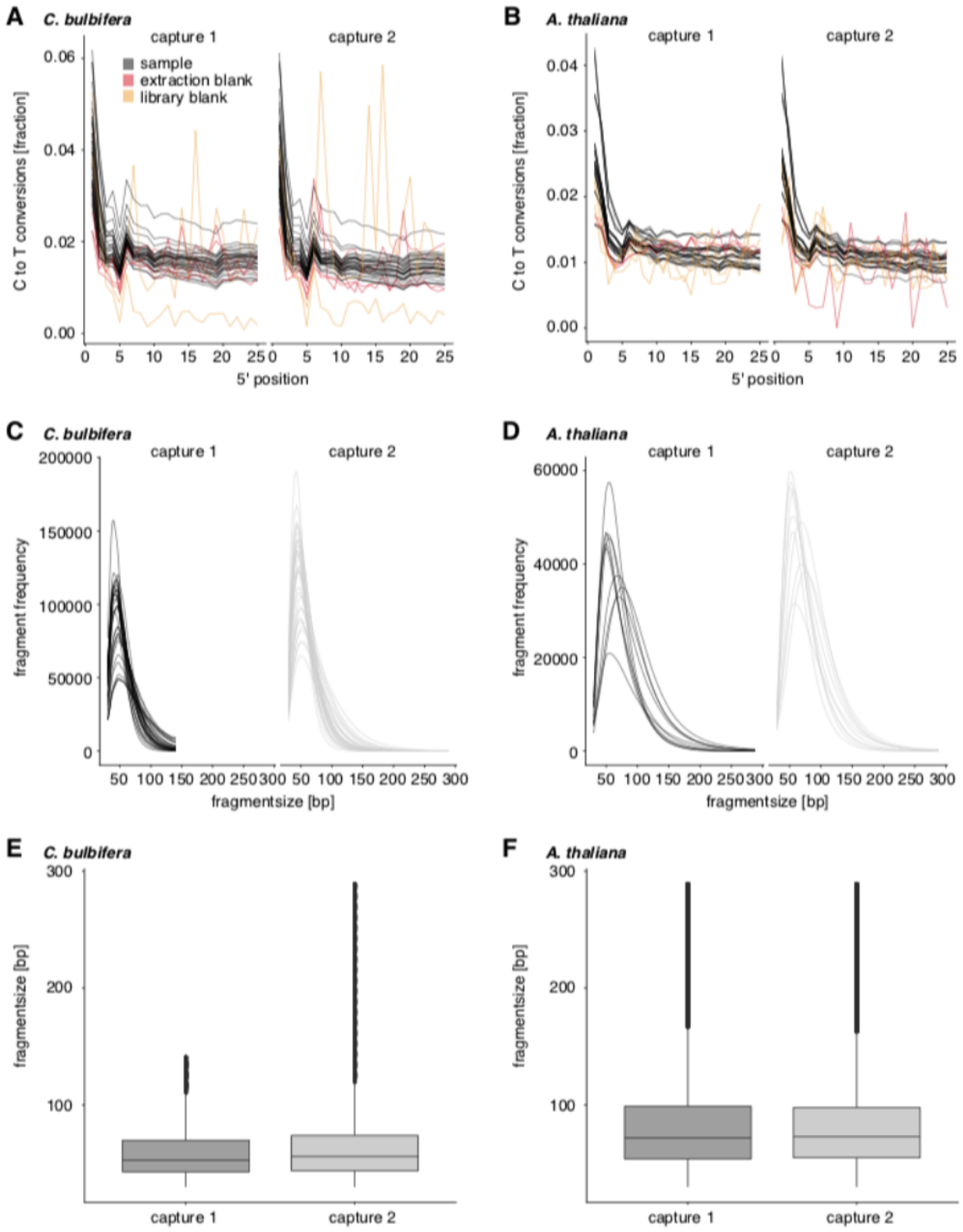
Supplemental Figure 1. In silico digest of historical sequencing information. a) Fragment size distribution of all historical fragments of an *Arabidopsis thaliana* herbarium library with a KpnI restriction site, before and after in silico digestion at these sites. **b)** Coverage histogram of all regions 50 bp up- and downstream of KpnI-site exact matches in the TAIR10 reference genome after mapping either the entire historical library (orange), or exclusively *in silico* digested fragments containing the KpnI-site (green), to TAIR10. Mapping was performed using bwa mem with default parameters, and coverage was assessed using bedtools coverage.



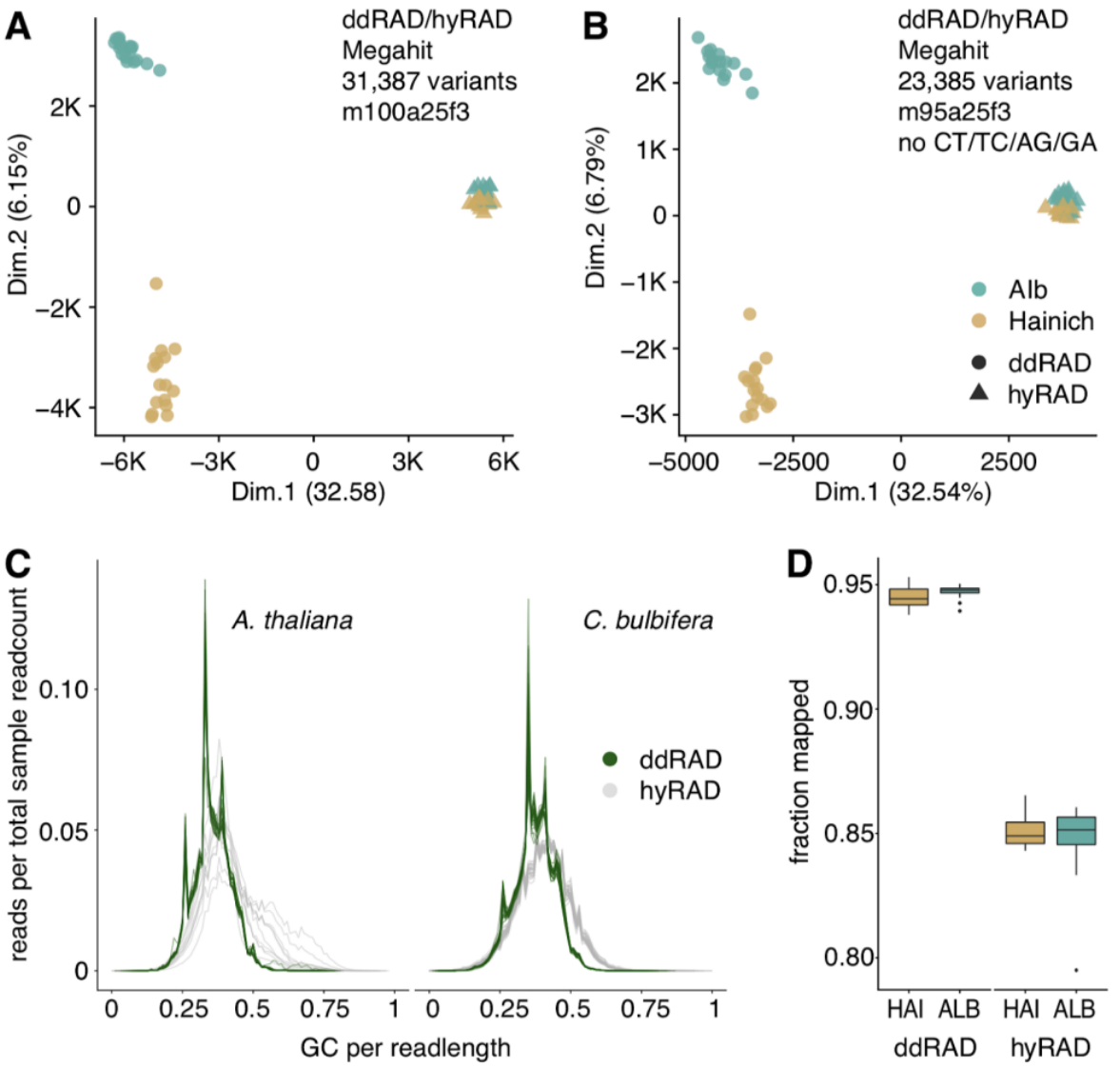
Supplemental Figure 2. Primer overview. a) and **b)** Fresh samples, **a)** intended for sequencing, with specific adapters including shift bases and primers for subsequent indexing, amplifications and sequencing. **b)** Samples intended to become baits, and their specific adapters (different from the adapters used for fresh libraries) and biotin primer. **c)** Historical libraries with specific adapters and primers for indexing, amplification and sequencing. Blocking oligos to prevent re-hybridization of double-stranded historical libraries pre-hybridization capture.



Supplemental Figure 3. Comparison of bait efficiency in *Arabidopsis thaliana*. Pairwise genetic distances between samples captured with the same bait-type (pUS, pMA or pMix) were measured within and between the two main clusters of either African or North American (US) samples. Distances per bait-type are displayed in box and whiskers plots, color-coded and from left to right for pMA (red), pUS (blue) and pMix (purple), separating (from left to right) distances within the African cluster, the North American cluster, and distances between both clusters.



Supplemental Figure 4. Damage patterns of hyRAD samples. Fraction of C-to-T conversion (i.e. deamination) observed in the 5' end of sequenced fragments in the first and second round of capture, in **a)** *Cardamine bulbifera* and **b)** *Arabidopsis thaliana*. Sizes of sequenced (deduplicated) fragments per capture round, evidencing the different sequencing strategies employed (i.e. 70 bp versus 150 bp in capture round 1 and 2, respectively), in **c** and **e)** *C. bulbifera* and **d** and **f)** *A. thaliana*.



Supplemental Figure 5. *Cardamine bulbifera* historical and modern separation investigation. MDS using **a)** only sites with full information, **b)** no sites with the allele pairs C/T or A/G, which might be caused by age-related deamination instead of reflecting true genetic variation. **c)** GC content of historical and modern *C. bulbifera* and *Arabidopsis thaliana* samples, calculated from sequences after merging of read pairs. **d)** Fraction of reads (fresh ddRAD samples, of total; historical hyRAD samples, of deduplicated) that map to the Megahit assembly.

2.3 Addendum, exome capture with custom baits in *A. myosuroides*

Captures are reduced representation methods highly targeted to the regions of interest and can be used to address a variety of evolutionary and ecological questions (Jones and Good 2016). A widely used method is exome capture, which aims to enrich the coding sequence of the genome (Ng et al. 2009). It usually requires probe design and costly synthesis. To design the probes, one also needs a reference genome. To overcome these limitations, Puritz and Lotterhos (2018) developed a protocol in which custom complementary DNA (cDNA) probes are directly prepared from RNA samples and can be applied to genomic DNA Illumina libraries. This allows the study of local adaptation based on expressed alleles in any organism and saves the cost of probe synthesis. To reduce the presence of higher expressed transcripts in the custom probes, a normalization step with duplex-specific nuclease (DSN) is included. The protocol was benchmarked in oyster (*Crassostrea virginica*) (Puritz and Lotterhos 2018).

I attempted to transfer the method to an *A. myosuroides* collection in a single field (440 samples; Exome_capture_collection.xlsx). In total, there were four consecutive collection time points each before and after herbicide treatment. In spring 2018, the ALS inhibitor Atlantis WG® (29.2 g kg^{-1} of mesosulfuron and 5.6 g kg^{-1} of iodosulfuron, HRAC group 2) was applied in wheat, and in fall 2018, microtubule assembly inhibitor Kerb™ Flo (400 l^{-1} Propyzamid, HRAC group 3) was sprayed in oilseed rape.

Methods

DNA libraries were generated with a custom Nextera protocol using a purified *Tn5* transposase as described in Picelli et al. (2014). For cDNA probe generation a large number of expressed genes is required, therefore I treated each of the 60 *A. myosuroides* populations with the herbicides Atlantis WG®, Kerb™ Flo and Axial® 50 (50 g l^{-1} of pinoxaden + 12.5 g l^{-1} Cloquintocet-mexyl, HRAC group: 1) at recommended field rates, as well as no treatment, to maximize the pools of transcripts that will be used as probes. Leaf material was taken after 24 and 72 hours, as the highest response expression is usually observed after these times (Gardin et al. 2015). RNA from different timepoints was extracted following the protocol from Yaffe et al. (2012) and pooled equally into one large pool. To prepare the probes, I followed the steps in the EecSeq protocol of Puritz and Lotterhos 2018 (https://github.com/jpuritz/EecSeq, Puritz and Lotterhos 2018) using the NEBNext Ultra II library prep kit (NEB #E7760) and the Illumina DSN Normalization protocol (DSN, Evrogen #EA003 , protocol: #15014673). It is advisable that the generated cDNA probes cannot bind

to the Illumina flow cell, even when they are removed by a treatment with restriction enzymes. Therefore, I made a modification to the adapters and corresponding biotinylated amplification primers and adapted already established primer sequences named APL1 and APL6 (Fu et al. 2013). The final hybridization of the probes with the DNA libraries was performed according to the capture protocol of the thesis Chapter 2.1 (Lang et al. 2020). The captured samples were sequenced on an Illumina HiSeq3000 sequencer in one lane with paired end mode and 150 bp read length.

To calculate capture efficiency, I aligned the reads to our *A. myosuroides* reference genome with bwa-mem v0.7.17-r1194-dirty (Li 2013). Samtools depth v1.9 (Li et al. 2009) was used to estimate whether the gene fraction was overrepresented compared to the rest of the genome. A similar analysis was also performed for the annotated TEs. If the method would have been efficient, the proportion of reads belonging to protein-coding genes should have been overrepresented.

Results

For the reference genome of *A. myosuroides* in our own study, I obtained 50,029 annotated genes comprising a total of 139 Mb (3.9%) of the genome (Kersten et al. 2021). Contrary to expectations, the annotated exomes were underrepresented in the custom capture (median: 3.2%, mean: 3.2%) (Figure 5a). That is, the experimental mean values were lower than the expected genome-wide average coverage of the exome. To investigate which other sequences in the genome could have been responsible for the lack of enrichment, I focused on transposable elements (TEs), since they are also expressed and might have contributed unproportionally to our custom cDNA probes. This turned out to be a plausible explanation, as in our *A. myosuroides* genome 83% are annotated as TEs, but in the capture TEs are slightly overrepresented with a median of 86.5% and a mean of 86.4% (Figure 5b).

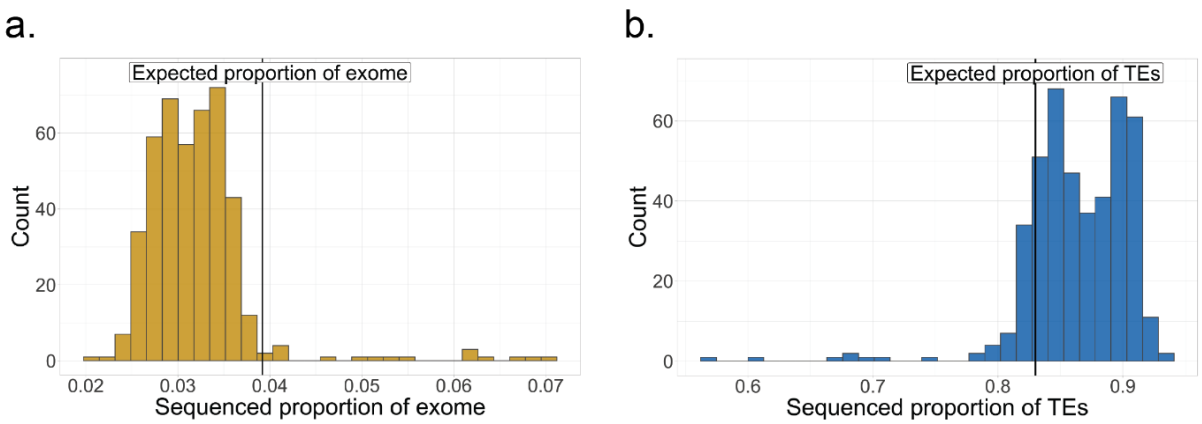


Figure 5: Capture efficiency evaluation of the custom capture in *Alopecurus myosuroides*. **a.** Sequenced proportion of the genome-wide exome. Black line shows the expected proportion of genes. **b.** Sequenced proportion of transposable elements (TEs). Black line indicates the expected proportion of TEs.

Discussion

In plants, transposable elements are a major drive of genome expansion, and can represent from as low as 3% to as high as 85% of plant genomes (reviewed in Lee and Kim 2014). The Puritz exome capture protocol (Puritz and Lotterhos 2018) was tested in the oyster *Crassostrea virginica*, which has a genome size of 685 Mb (NCBI assembly Bioproject PRJNA376014), nearly six times smaller than *A. myosuroides*. Notably, the percentage of TEs in *C. virginica* is only 48% compared to 83% in *A. myosuroides*, which is at the upper end of the distribution of TE percentages in plant genomes. The much larger fraction and absolute number of TEs in *A. myosuroides* may not be suitable for the generation of experimental exome probes. Therefore, I advise caution when applying custom probes to plant species.

3. Chapter 2

3.1 Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass

PNAS RESEARCH ARTICLE | AGRICULTURAL SCIENCES OPEN ACCESS

Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass

Sonja Kersten[a,b], Jiyang Chang[c,d], Christian D. Huber[e], Yoav Voichek[f], Christa Lanz[b], Timo Hagmaier[b], Patricia Lang[b,1], Ulrich Lutz[b], Insa Hirschberg[g], Jens Lerchl[h], Aimone Porri[h], Yves Van de Peer[c,d,i,j], Karl Schmid[a], Detlef Weigel[b,2], and Fernando A. Rabanal[b]

Contributed by Detlef Weigel; received April 19, 2022; accepted March 10, 2023; reviewed by Todd A. Gaines and Pleuni S. Pennings

Repeated herbicide applications in agricultural fields exert strong selection on weeds such as blackgrass (*Alopecurus myosuroides*), which is a major threat for temperate climate cereal crops. This inadvertent selection pressure provides an opportunity for investigating the underlying genetic mechanisms and evolutionary processes of rapid adaptation, which can occur both through mutations in the direct targets of herbicides and through changes in other, often metabolic, pathways, known as non-target-site resistance. How much target-site resistance (TSR) relies on de novo mutations vs. standing variation is important for developing strategies to manage herbicide resistance. We first generated a chromosome-level reference genome for *A. myosuroides* for population genomic studies of herbicide resistance and genome-wide diversity across Europe in this species. Next, through empirical data in the form of highly accurate long-read amplicons of alleles encoding acetyl-CoA carboxylase (ACCase) and acetolactate synthase (ALS) variants, we showed that most populations with resistance due to TSR mutations—23 out of 27 and six out of nine populations for *ACCase* and *ALS*, respectively—contained at least two TSR haplotypes, indicating that soft sweeps are the norm. Finally, through forward-in-time simulations, we inferred that TSR is likely to mainly result from standing genetic variation, with only a minor role for de novo mutations.

Alopecurus myosuroides | herbicide resistance | rapid adaptation | blackgrass

Significance

Because herbicides are designed to kill weeds, spontaneous mutants that are resistant to herbicide application have an enormous selective advantage and will often come to quickly dominate weed populations. While this is a nuisance for farmers, it provides opportunities for investigating in detail how organisms rapidly respond to strong selection, especially what role newly arising mutations play vs. mutations that are already present in a population. We first assembled a reference genome for blackgrass, the most economically damaging herbicide-resistant weed in Europe, and then combined analyses of known herbicide-resistant loci with forward-in-time simulations to show that target-site resistance mutations likely often predate the application of herbicides.

The agricultural use of herbicides has inadvertently selected for many herbicide-resistant grass weeds over the past several decades. Among these, blackgrass (*Alopecurus myosuroides*) has become the most economically damaging herbicide-resistant weed in Europe (1, 2). In England alone, the annual cost of resistance was estimated to be £0.4 billion (€ 0.47 billion) in lost gross profit (3).

We distinguish two resistance mechanisms. First, there is target-site resistance (TSR), which is caused by coding sequence mutations in or amplification of the genes encoding the proteins targeted by herbicides (4–9). Second, there is non-TSR (NTSR), which is associated with enhanced metabolic processes such as herbicide detoxification or sequestration (6, 10). To better understand how either type of resistance arises and may potentially come to dominate *A. myosuroides* populations, we need to learn more about the population structure and genetic diversity of the species across Europe. Previous regional studies have only found weak, if any, population structure, suggesting a very rapid and recent spread of the species (11, 12).

Two important drivers of the modes of evolution of herbicide resistance are the genetic architecture of the trait and the types of mutations that can give rise to it. TSR is conferred by mutations in single genes, with only a very small number of coding sequence changes allowing for herbicide resistance without eliminating the activity of the targeted protein. As in many other weeds, TSR in *A. myosuroides* has increased rapidly (13, 14), and as a consequence, herbicides that inhibit the action of acetolactate synthase (ALS, also known as acetohydroxyacid synthase) and acetyl-CoA carboxylase (ACCase) have widely lost their efficacy as weed control agents. In contrast, several different gene families, encoding detoxifying enzymes and transporters such as cytochrome P450 monooxygenases, glutathione S-transferases, ATP-binding cassette, and MFS-type transporters as well as glycosyltransferases, have been found to contribute to NTSR (reviewed in ref. 15). NTSR now accounts for a substantial proportion of resistance in agricultural fields and is becoming a major focus of herbicide resistance research (13).

Finally, the rapid speed with which herbicide resistance spreads in individual weed species raises the question of whether this is primarily due to repeated selection for rare de novo mutations, or more commonly arising from standing genetic variation, with herbicide-resistant alleles segregating in the population already before the widespread adoption of herbicide application. An optimal framework for distinguishing between these hypotheses is provided by forward-in-time genetic simulations (16).

Competing interest statement: J.L. and A.P. are employees of BASF, which manufactures and sells herbicides. D.W. holds equity in Computomics, which advises breeders. D.W. advises KWS SE, a plant breeder and seed producer. All other authors declare no competing or financial interests.

[1]Present address: Department of Biology, Stanford University, Stanford, CA 94305.

[2]To whom correspondence may be addressed. Email: weigel@tue.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2206808120/-/DCSupplemental.

Published April 12, 2023.

To enable a better understanding of herbicide resistance evolution in *A. myosuroides*, we have generated a high-quality reference genome with PacBio long reads. Genotyping with double-digest restriction-site associated DNA (ddRAD) sequencing markers in 47 European field populations revealed considerable geographical population structure along with high effective population sizes. To characterize TSR haplotype diversity at the field level, we generated PacBio long-read amplicons for the known TSR genes *ACCase* and *ALS* and compared our empirical data with the results from probabilistic models of adaptation via selective sweeps and forward simulations. We infer that standing genetic variation is the most likely mechanism behind the TSR mutations of independent origin, with only a minor role for de novo mutations.

Results

Genome Assembly and Annotation. For genome sequencing and assembly, we selected a single plant from an herbicide-sensitive population (Appels Wilde Samen GmbH, Darmstadt) from Germany and ascertained that it did not carry known TSR mutations at the nuclear *ACCase* and *ALS* or the chloroplast *psbA* loci (*Methods*). Previous genome size estimates of *A. myosuroides* based on Feulgen photometry ranged from 4.2 Gb (17) to 4.7 Gb (18). To estimate the genome size of the selected individual more accurately, we performed flow cytometry using rye (*Secale cereale*) as a reference standard (19), which yielded an estimated haploid genome size of 3.56 Gb (Fig. 1*A*). Next, we generated ~90× genome coverage of PacBio continuous long reads (CLRs), ~44× genome coverage of Illumina PCR-free short reads, and ~66× genome coverage of Hi-C chromatin contact data. We de novo assembled the genome with *FALCON-Unzip* (20), deduplicated primary contigs with *purge_dups* (21), and scaffolded contigs with *HiRise* (22) (Table 1). The size of the final assembly was 3.53 Gb and consisted of seven super-scaffolds (Fig. 1*B* and *SI Appendix*, Fig. S1*A*), in agreement with the known karyotype of the species with seven chromosomes (18).

Given that in plants, transposable elements are a major driver of genome size, it is not surprising that repetitive sequences account for 85.2% of the *A. myosuroides* genome, with 63.8% classified as long

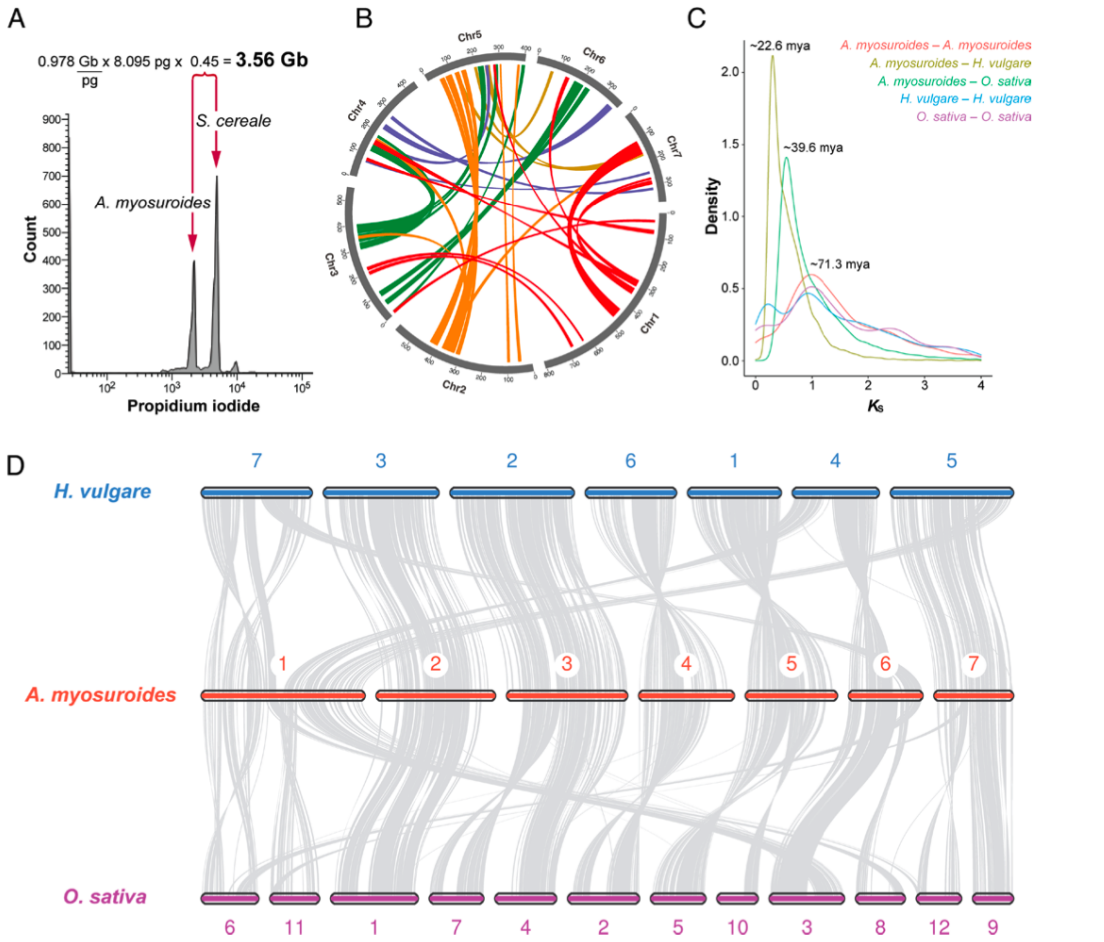


Fig. 1. Reference genome of an *A. myosuroides* individual from a German herbicide-sensitive population. (*A*) Histogram of relative DNA content from flow cytometry of propidium iodide-stained nuclei of *A. myosuroides* and the reference standard *Secale cereale* cv. Daňkovské (diploid genome size = 16.9 pg). (*B*) Circos plot of the *A. myosuroides* genome, with colored lines connecting anchor pairs (genes in collinear regions) with synonymous substitution rates (K_S) > 0.5. Numbers represent megabases. (*C*) K_S distributions for paralogs within the *A. myosuroides, Hordeum vulgare* (23), and *Oryza sativa* (24) genomes and for ortholog pairs shared by the three species. Divergence time, expressed as Mya, was estimated based on 7.0×10^{-9} as the substitution rate in grasses (25). (*D*) Syntenic relationships between the chromosomes of *A. myosuroides* and *H. vulgare* (*Top*) and *O. sativa* (*Bottom*).

Table 1. Genome assembly metrics

Descriptor	Value
Total assembly size	5,218,837,661 bp
Number of contigs	6,215
Contig N50	1,783,999 bp
Largest contig	11,174,859 bp
Chromosome-level assembly size	3,529,081,863 bp
Chromosome N50	554,019,051 bp
Largest chromosome	807,086,175 bp
Number of protein-coding genes	50,029
Mean gene length	2,789 bp
BUSCO score	C:94.6% [S:82.0%, D:12.6%], F:0.9%, M:4.5%
TE content	Class I [LTR: 63.8%, non-LTR: 0.1%] Class II [TIR: 10.9%, Helitron: 8.2%] Other repeated regions: 2.15%

Contig metrics are shown before deduplication. Benchmarking Universal Single-Copy Orthologs (*BUSCO*) (26) scores were obtained with the 'embryophyta_odb10' gene set (n = 1,614). Complete (C), single copy (S), duplicated (D), fragmented (F) and missing (M) genes are indicated. Transposable element (TE) content was determined with the *Extensive* de novo *TE Annotator* (EDTA) (27).

terminal repeats retrotransposons (Table 1). We annotated 50,029 protein-coding genes based on a combination of both RNA-seq and PacBio Iso-Seq long transcripts from five different tissues (leaves, roots, whole inflorescences, anthers, and pollen), *ab initio* prediction, and protein homology. Transcriptome data supported 87.5% of the annotated genes, and 95% of all genes could be assigned functions with *InterProScan* (28). Of the 1,614 near-universal single-copy orthologs predicted by *BUSCO* (26), 94.6% were found as complete genes (82.0% as single copy, 12.6% as duplicated genes). On average, protein-coding genes in *A. myosuroides* are 2,789 bp long and contain 3.71 exons (Table 1). Potential shortcomings of our assembly are discussed in the *Methods* section.

Chromosome level synteny with other grasses was high, particularly with the more closely related *Hordeum vulgare* genome (23), for which the distribution of the number of synonymous substitutions per synonymous site (K_S) for orthologous gene pairs indicated a divergence time of ~22.6 Mya (Fig. 1*C*). Chromosomes 2, 3, 4, 5, and 7 in *A. myosuroides* have a near 1:1 relationship with chromosomes 3, 2, 6, 1, and 5 in *H. vulgare*, respectively (Fig. 1*D*). An exception is chromosome 1 in *A. myosuroides* (807 Mb), which contains sequences that are syntenic with chromosomes 4, 5, and 7 in *H. vulgare*.

Population Structure. Our new reference genome allowed us to easily assess the distribution of genome-wide diversity across Europe. To this end, we performed ddRAD-Seq in 1,123 individuals. These represented 44 populations, each with 22 to 24 plants, across nine European countries, and came from farmers with suspected herbicide resistance in their fields. For comparison, we included three herbicide-sensitive reference populations (Fig. 2). We defined 109,924 single nucleotide polymorphisms (SNPs) with an average sequencing depth of 22.6× (*SI Appendix*, Fig. S2*A*). A clear phylogeny per country was not discernible from the maximum likelihood (ML) tree (Fig. 2*A*), but Treemix captured the geographic distribution at the country scale without significant migration events, as indicated by the F3 statistic (*SI Appendix*, Fig. S3).

Overall genetic differentiation between populations was low (F_{ST} range: 0.01 to 0.05, n = 47; Fig. 2*C*), consistent with other studies of *A. myosuroides* (11, 12) and other wild grasses such as *Panicum virgatum* (29). The relatedness of individuals within populations was high (F_{IS} = 0.1; range 0.06 to 0.12). In the admixture analysis, we could identify between 7 and 9 ancestry groups (Fig. 2*D* and *SI Appendix*, Fig. S4*C*) that were consistent with the clusters formed in a principal component analysis (PCA; Fig. 2 *B* and *D* and *SI Appendix*, Fig. S4*A*). Individuals from Belgium (BE), the United Kingdom (UK), Luxemburg (LX), and France (FR) were genetically very similar and clustered together. A population from the Netherlands (NL), NL11330, was most differentiated from all others with F_{ST}–values up to 0.05 (Fig. 2*C*). Germany (DE) was divided into three subclusters, one having common ancestry with individuals from Switzerland (CH), one with Austria (AT), and the third cluster being highly admixed. The populations from Poland (PL) shared common ancestry with the Netherland population NL01664. In summary, there is a clear geographical population structure across Europe, although the data at this point do not allow us to infer colonization and migration histories.

The mean observed SNP heterozygosity was 0.11, with no significant difference between populations that were under herbicide selection and those that were not (*SI Appendix*, Fig. S2*B*). We further estimated Watterson's theta θ_W on the 1.1% sequenced fraction of our genome. The θ_W (mean = 0.0047) estimates are within the range of other outcrossing plant species (30). With these θ_W estimates and the mutation rate of 3.0×10^{-8} from maize (31), we determined effective population sizes ranging from 30,366 to 41,941 individuals (*SI Appendix*, Fig. S2*C*). Among countries for which we had more than six populations, Germany had significantly smaller (*P* value range: 0.01 to 0.03) effective population sizes than France, Belgium, and the United Kingdom (*SI Appendix*, Fig. S2*D*), but the causes for this difference remain unknown. Given that we have estimated effective population sizes in *A. myosuroides* mostly from populations under selection that have already experienced a decrease in population sizes, it is very likely that we rather underestimate the long-term effective population sizes of our *A. myosuroides* field populations (32). Messer and Petrov (2013) noted that temporal fluctuations in population size can strongly influence estimates of effective population size, especially in recent bottlenecks, as would be the case with adaptation processes to herbicides (33). Adaptation to strong selection pressure is a rapid process, and the probability of adaptive mutations arising is higher for larger population sizes (34, 35).

Haplotype Networks of Herbicide Target Genes ALS and ACCase. Much of the work on the molecular mechanisms underlying herbicide resistance has focused on mutations in the genes that encode the enzymes inhibited by herbicides. Two prominent herbicide targets are the genes encoding ALS and ACCase, both of which can be inhibited by a range of structurally diverse chemicals (36, 37). At both loci, mutations at multiple conserved codons are known to confer inhibitor resistance (4–7). To understand the diversity not only of specific mutations but also of entire haplotypes on which these mutations arose, we aimed to characterize the *ALS* and *ACCase* loci, including the extended linked sequences that surround them. We amplified ~13.2 kb for *ACCase* and ~3.6 kb for *ALS* by long-range PCR and analyzed complete amplicons with PacBio Circular Consensus Sequencing (CCS) for all individuals in our European collection. We applied very stringent criteria to call haplotype sequences in our dataset—requiring high accuracy (>99% or q20) and a minimal CCS read depth per sample of 25×. This enabled us to characterize entire haplotypes for 1,046 individuals for *ACCase* and 842 individuals for *ALS*. We were able to recover two haplotypes for the vast majority of our samples that passed quality control filters, 84.9% for *ACCase* and 59.8% for

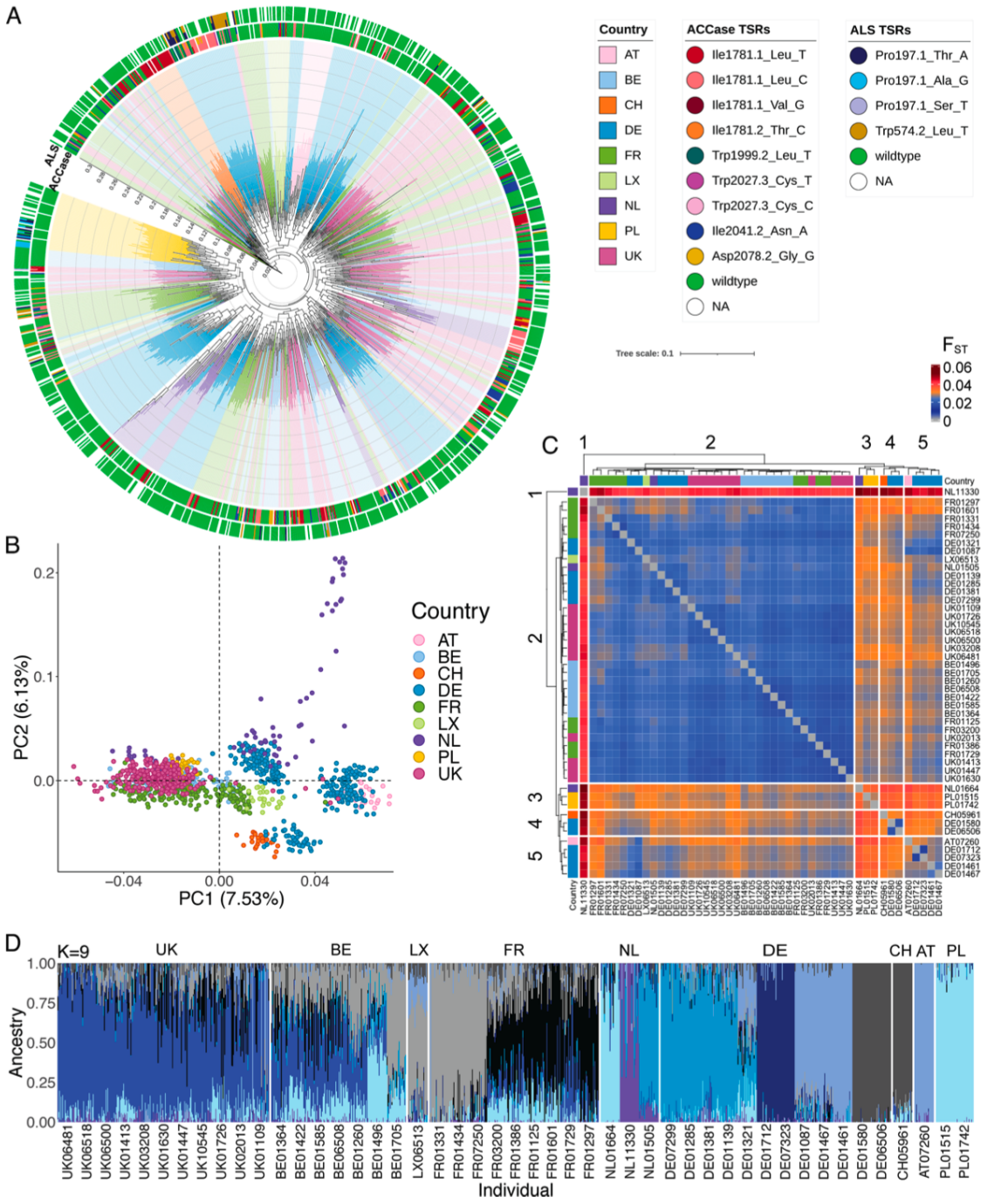


Fig. 2. Population structure analysis of 47 European *A. myosuroides* populations with 109,924 genome-wide ddRAD-seq markers. (*A*) Maximum-likelihood tree. Branch ends are marked in country colors. TSR mutations for *ALS* and *ACCase* in each individual are indicated in the outer and inner rings, respectively. (*B*) Principal component analysis (PCA) showing the first two eigenvectors, with explained genetic variance in parentheses. Colors reflect country-specific origin of the populations. (*C*) Heatmap of fixation index (F_{ST}) values in contrast between the different populations, ranging from close to 0 (blue) to 0.06 (dark red). Populations are clustered by similarity of F_{ST} patterns, and colors of the branch tips indicate countries of origin of the populations. (*D*) Admixture proportions with ancestry groups of K = 9. Admixture proportions with ancestry groups of K = 7 and K = 8 can be found in *SI Appendix*, Fig. S4. Each bar corresponds to one individual, grouped by country [Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK)].

ALS. We assume that the remaining individuals are homozygous for the same haplotype.

Some TSR mutations were less common than others. For example, Trp2027Cys and Asp2078Gly in *ACCase* were underrepresented, consistent with these mutations reducing fitness in the absence of herbicide selection in other species (38–41). The most common mutation was Ile1781Leu, consistent with pleiotropic effects of this substitution that increase fitness also in the absence of selection (38, 39) (Fig. 3 *A*–*C* and Dataset S2). Finally, TSR mutations typically act in a dominant fashion, and the majority of TSR mutations (71.4%) at *ACCase* in our data occurred as heterozygotes.

PCA of *ACCase* haplotypes distinguishes three major groups (*SI Appendix*, Fig. S5). Although each group included representatives from all countries (including those countries for which we only analyzed single populations), and although alleles without obvious TSR mutations were present in all major groups, it is difficult to judge with current information whether the major groups of haplotypes arose before the species colonized Europe or whether they reflect relatively recent migration events. TSR substitutions Ile1781Leu, Ile1781Thr, Trp2027Cys, and Ile2041Asn were the most wide-spread and found in all European groups, while Asp2078Gly and Ile1781Val were found in two and Trp1999Leu in one group. This pattern of the same TSR mutation arising independently in separate geographic locations across Europe (Fig. 2*A*) extends previous observations made at local or country scales using small samples of short amplicons that included only a limited number of variable sites for haplotype detection (42–44).

To better characterize *ACCase* haplotype diversity, we inferred haplotype trees and networks at the level of single fields (Fig. 3 and Dataset S2). We observe haplotype networks of varying complexity (Fig. 3 *A*–*C*), likely reflecting the selection pressure to which each population was subjected. If the allele frequency of a single mutation—on a single haplotype—increases rapidly in a population, this is called a "hard sweep". If, on the other hand, there are several different haplotypes in a population that confer resistance—whether they all carry the same beneficial mutation or different ones—and increase in frequency at the same time, this is referred to as a "soft sweep" (32). In our collection, only four out of the 27 populations with recorded TSRs contained a single TSR haplotype—and in these four cases, the TSR haplotypes were found at low frequency, with fewer than 10% of sequences having the corresponding TSR mutation. In principle, this pattern may well reflect an early state of a hard sweep, but that the other 23 populations contain at least two haplotypes with TSR mutations indicates that soft sweeps are the norm (*SI Appendix*, Table S1 and Fig. S6). In 14 of these populations, we found different haplotypes with the same TSR mutation resulting from multiple independent mutation events, as opposed to the same TSR mutation being transferred to other haplotypes by recombination (Fig. 3*D* and Dataset S2). This observation confirms in an unbiased manner inferences from earlier explorative studies (42–45). We found seven instances in which two or three different TSR mutations had arisen in a single field, from the same haplotype (Fig. 3 *B*, *C*, and *E* and Dataset S2). The maximum number of independent (nonrecombinant) *ACCase* TSR haplotypes within a field population was 10 (*SI Appendix*, Table S1).

Having observed multiple TSR haplotypes in the same field, we also asked the converse question, whether the same haplotypes could be found across populations. The *ACCase* sequences from our collection of 1,046 individuals could be clustered into 250 nonredundant haplotypes, of which a quarter (n = 62) carried one of the known TSR substitutions (Dataset S1). A third of these (n = 20) were shared across multiple populations, which was essentially the same as wild-type haplotypes shared by multiple populations (55 out of 188; chi-squared test *P* value = 0.7736). Furthermore, the most common TSR haplotypes had evolved from the most common wild-type haplotypes. For instance, three of the eight most abundant TSR haplotypes were only one mutation away from the single most abundant wild-type haplotype, which made up 12.6% of all individual haplotypes (n = 264). We note that identical TSR haplotypes do not have to have a single origin, given that parallel herbicide resistance evolution is common, and identical TSR mutations can occur not only on different *ACCase* haplotypes in *A. myosuroides* but also in other genes associated with resistance to herbicides of different species (46).

In the complete assembly of the *A. myosuroides* genome, we discovered at least two copies of the *ALS* gene in chromosome 1 (*Methods*). These copies have full-length open reading frames without introns, and high-quality Iso-Seq reads (>99.9% or q30) span full-length transcripts (*SI Appendix*, Fig. S7). The copy most similar to the GenBank sequence AJ437300.2 (47) was designated as *ALS1*, and we selectively amplified *ALS1* with primers that should not target the other *ALS* loci (or locus). The existence of multiple *ALS* copies in *A. myosuroides* may have confounded previous studies, which relied on primers in the coding region to genotype *ALS* TSR mutations. This strategy was used in a pyrosequencing assay (48) commonly used for this type of study, which, differently from our work, did not lead to the identification of homozygous or trans-heterozygous Pro197Thr genotypes (49, 50).

ALS1 haplotypes fell into three major Europe-wide groups (*SI Appendix*, Fig. S8). In our collection, TSR mutations for this gene were only present in Germany, France, United Kingdom, and Poland. TSR mutations Pro197Thr and Trp574Leu were found in two of these groups, Pro197Ala and Pro197Ser only in one, and no obvious TSR mutation was found in the third group. Similar to *ACCase*, although less often, two or more TSR haplotypes of independent origin could be detected within single fields, in six of the nine populations with recorded TSRs (*SI Appendix*, Table S1 and Dataset S3).

Simulations of Standing Genetic Variation vs. de novo Mutations. Strong selection pressure exerted by herbicides leads to very rapid adaptation, but a major question is whether herbicide resistance evolves predominantly from standing genetic variation that was present already before the onset of herbicide selection or from de novo mutations that arose after herbicide selection began. In other words, are the typical population dynamics in terms of effective population size and drift compatible with a reservoir of TSR mutations available before exposure to herbicides and are spontaneous TSR mutations sufficiently frequent for rapid resistance evolution?

To answer this question, we first used equations from Hermisson and Pennings (34) to derive expectations for the probability of adaptation (i.e., evolution of herbicide resistance via TSR mutations) and the likelihood that this adaptation is due to standing genetic variation. First, we calculated the probability of adaptation, assuming a mutation rate of 3.0×10^{-8} (31), a mutational target size of seven nucleotides, corresponding to the TSR mutations investigated here, and onset of herbicide selection 30 generations ago. As mentioned above, estimates of N_e based on genetic diversity integrate over a long period of time and past bottlenecks will reduce it, leading to estimates that are lower than the actual N_e before the bottlenecks (33). Therefore, we considered both $N_e = 42{,}000$, which is the highest estimate from our field populations, and $N_e = 84{,}000$.

With $N_e = 42{,}000$, we observed that the probability of adaptation strongly depends on the beneficial selection coefficient of the TSR mutation during the herbicide selection phase (s_{ben}). For strong positive selection ($s_{ben} \geq 1$), which we expect for herbicide

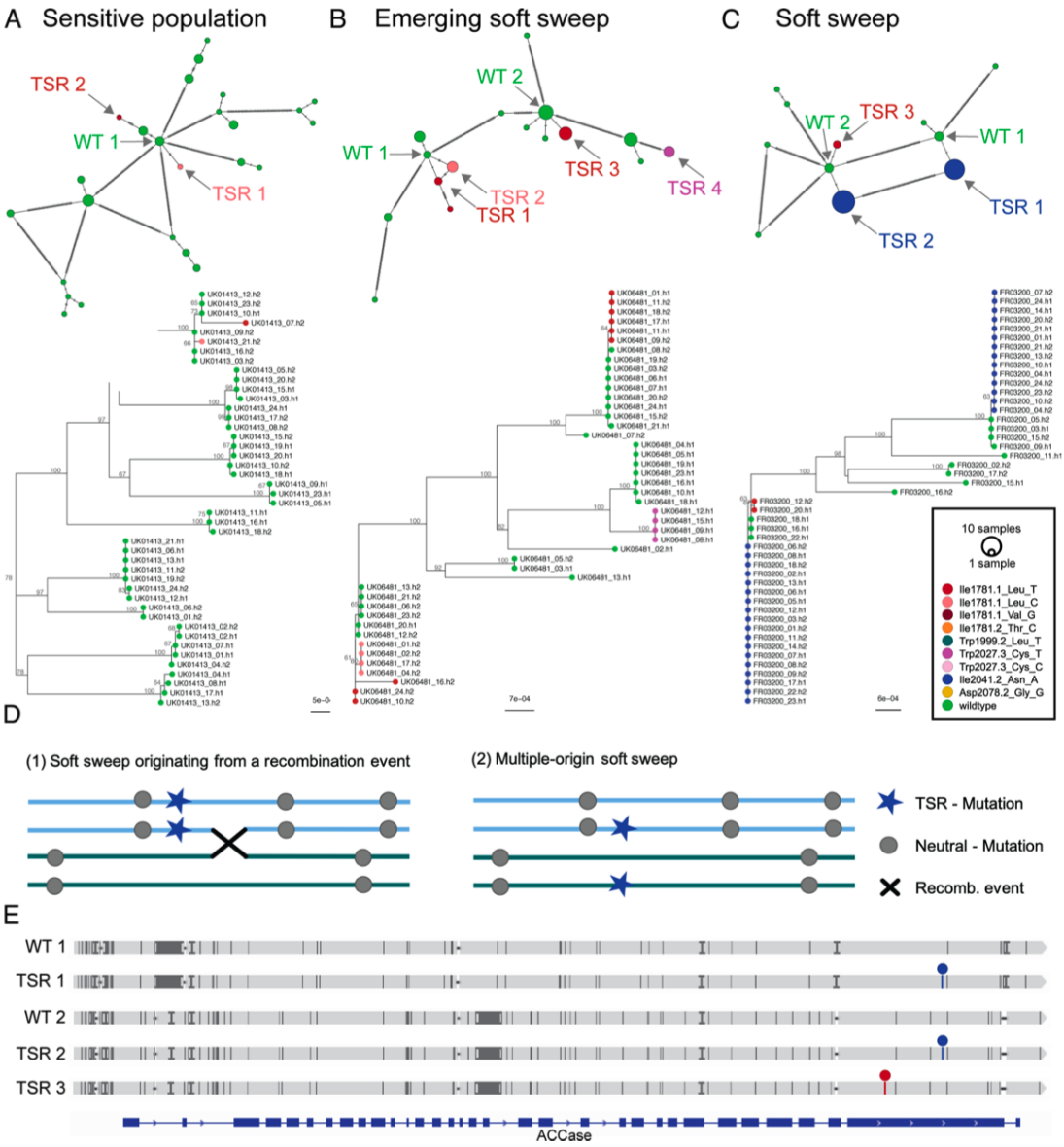


Fig. 3. Haplotype analysis of the complete *ACCase* gene (13.2 kb). (*A*) Network and maximum likelihood (ML)-tree of 44 haplotypes from the sensitive reference population UK01413 (HerbiSeed standard), which has not been under herbicide selection. The color code in all networks and trees (*A–C*) indicates different target-site resistance (TSR) mutations, with haplotypes that have wild-type sequence at known TSR positions in green. Likely wild-type haplotypes of origin for TSR mutations are indicated (WT). (*B*) Network and ML-tree of 44 haplotypes from the British population UK06481, which shows a selection pattern characteristic of an emerging soft sweep for TSR mutations. (*C*) Network and ML-tree of 46 haplotypes from the French population FR03200, with a predominant soft sweep pattern for the TSR mutation Ile2041.2_Asn_A. (*D*) Schematic representation of alternative origins of soft sweep patterns: recombination vs. independent mutation events. (*E*) In the FR03200 population, two distinct wild-type haplotypes (WT 1 and WT 2) have independently sustained the same TSR mutation, giving rise to haplotypes TSR 1 and 2. In addition, wild-type haplotype WT 2 has given rise to a second TSR haplotype (TSR 3). Positions of TSR Ile1781.1_Leu_T and TSR Ile2041.2_Asn_A mutations are marked with red and blue circles, respectively.

application, the probability of adaptation is high (> 50%). Only for weakly beneficial mutations ($s_{ben} \leq 0.01$) it decreases below 20% (Fig. 4 *A*, *Left*). With $N_e = 84{,}000$, the probability of adaptation increases to up to 80% due to both higher levels of standing genetic variation and a larger rate of de novo mutations (Fig. 4 *A*, *Right*). The deleterious selection coefficient of TSR mutations before the onset of herbicide selection has only a minor influence on the probability of adaptation.

Next, we asked whether adaptation to herbicide selection pressure predominantly occurs via standing genetic variation or de novo mutation (34). We observed that fixation from standing genetic variation is more probable (>50%) for neutral or almost neutral mutations ($s_{del} < 1e^{-4}$) and has a probability larger than 40% even for deleterious mutations ($s_{del} = 1e^{-3}$) and the smaller population size ($N_e = 42{,}000$) (Fig. 4 *B*, *Left*). The probability of adaptation from standing genetic variation generally increases with smaller s_{ben}

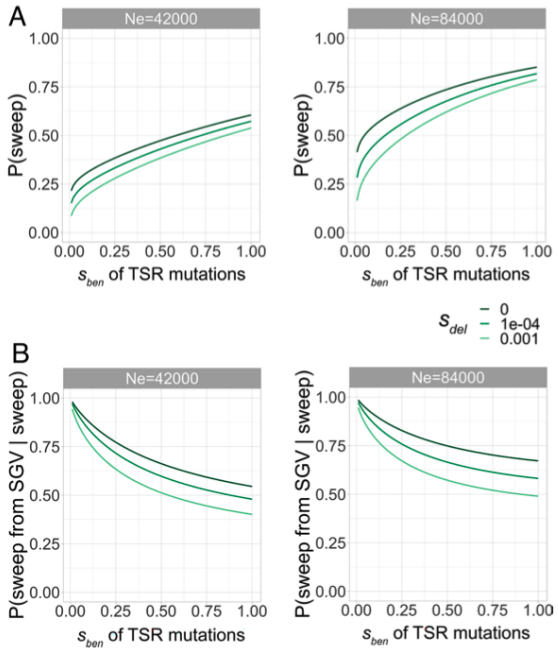


Fig. 4. Simulations of different scenarios for adaptation. Equations from Hermisson and Pennings (34) were used to derive (*A*) expectations for the probability of adaptation via a selective sweep from beneficial target-site resistance (TSR) mutations in general and (*B*) from standing genetic variation in particular. Probabilities of sweeps with different effective population sizes (N_e) are estimated as a function of the strength of selection.

or larger N_e (N_e = 84,000; Fig. 4 *B*, *Right*) because of the decreasing fixation probability of de novo mutations and the increasing levels of standing genetic variation, respectively (34). These results suggest that herbicide resistance should occur predominantly, although not exclusively, via standing genetic variation.

The remarkable diversity of TSR haplotypes of independent origin observed in individual fields (Fig. 3) prompted us to examine the speed of adaptation and the expected level of TSR diversity in more detail through forward-in-time simulations with the software SLiM (16). We defined two possible scenarios in which TSR mutations arise: one in which resistant alleles were already present in the population before the start of herbicide selection (standing genetic variation) and one in which they emerged only after selection pressure was imposed (de novo mutation) (*SI Appendix*, Fig. S9). Our model assumes that individuals with at least one TSR mutation have a 20 times higher chance of surviving the herbicide treatment than individuals without any TSR mutation. We applied a gamma distribution of deleterious mutations in exons (51) (Fig. 5), but a model with exclusively neutral mutations gave similar results (*SI Appendix*, Fig. S10). We ran one thousand simulations for two different N_e values, analyzing the changes in allele frequencies of TSR mutations, and the number of independent TSR mutations per population.

With N_e = 42,000, TSR mutations were more often attributable in the simulations to standing genetic variation (25.5%) than solely de novo mutations (4.2%). Also, we found more independent TSR mutations per population for the standing genetic variation scenario, at least two mutations in 41 simulation runs, with a maximum of four mutations in one run (Fig. 5*A*). Under the de novo scenario, there were at most two mutations, which were found in a single run (Fig. 5*B*). Since we had observed in most of our populations with TSR, namely in 23 out of 27, at least two independent TSR mutations, the simulations tend to agree mostly with standing genetic variation as a main driver of herbicide resistance evolution in our system. However, even under this scenario, the simulations fell short with respect to the extent of TSR due to soft sweeps observed in our populations.

Highly effective population sizes favor rapid adaptation processes because advantageous alleles are more likely to be immediately available and at higher frequencies (32, 34). With N_e = 84,000, 40.3% simulation runs had TSR mutations due to standing genetic variation and 5.7% had de novo mutations (Fig. 5 *C* and *D*), favoring the former scenario even more than with N_e = 42,000. Also, we found up to five independent TSR haplotypes due to standing genetic variation (Fig. 5*C*). In the single simulation run—under the model that considered deleterious mutations in exons—that led to five independent TSR mutations, these were already present at the onset of selection.

Even with a larger N_e, our simulations seem to underestimate both the ratio of TSR soft sweeps over hard sweeps, as well as the maximum number of independent TSR haplotypes in a given population, indicating that our N_e estimates are conservative. Actual N_e might be even higher since estimates based on RAD-seq data tend to underestimate genetic diversity (53, 54), and empirical census population sizes are higher than our estimated N_e (55). Very high effective population sizes would be consistent with reports from farmers of heavy infestations in fields due to difficulties in managing resistance. Large effective population sizes likely reflect large census population sizes, thus maintaining genetic variation under herbicide selection and providing a large genetic pool for accumulation of resistance mutations. Factors that promote high census population sizes in specific years include climatic variables that lead to poor weed control conditions, seed dormancy, reduced tillage efficiency, large seed banks, and crop rotations with high amounts of winter cereals (1, 55, 56).

Simulations have shown that in the time it takes for a particular allele to become fixed in a population starting from standing genetic variation, the same mutation can arise de novo (34). In the case of de novo mutations, our simulations reveal that there is a considerable risk that they are directly lost again through drift since they are on average initially much rarer than mutations that are part of standing genetic variation (Fig. 5 *A* and *C* and *SI Appendix*, Fig. S10 *A* and *C*), added to the possibility that some TSR mutations are slightly beneficial even in the absence of herbicide application (38, 39). And although we cannot exclude de novo mutations as a source of TSR alleles, they are characterized by a slow initial phase of adaptation (after 10 to 15 generations under selection) in our simulations, thus cannot compete with preexisting mutations from standing genetic variation (Fig. 5 *B* and *D* and *SI Appendix*, Fig. S10 *B* and *D*). Therefore, the standing genetic variation scenario, with the presence of multiple alleles, as is typical for soft sweeps, is closer to what we observed in our experimental data. Furthermore, to estimate how many TSR alleles per generation are present as standing genetic variation in a field, we ran 100 simulations under neutrality. This revealed the emergence and loss due to random genetic drift in field populations before the start of herbicide selection by farmers. We could detect up to four TSR alleles at the same time (*SI Appendix*, Fig. S11).

Previous studies documented rapid adaptation of grass weeds to herbicide applications within a few generations, sometimes as quickly as three or four generations (57, 58), which is in agreement with anecdotal reports from farmers. The degree of herbicide resistance in a field is also closely correlated with the frequency of application (59). This would be consistent with

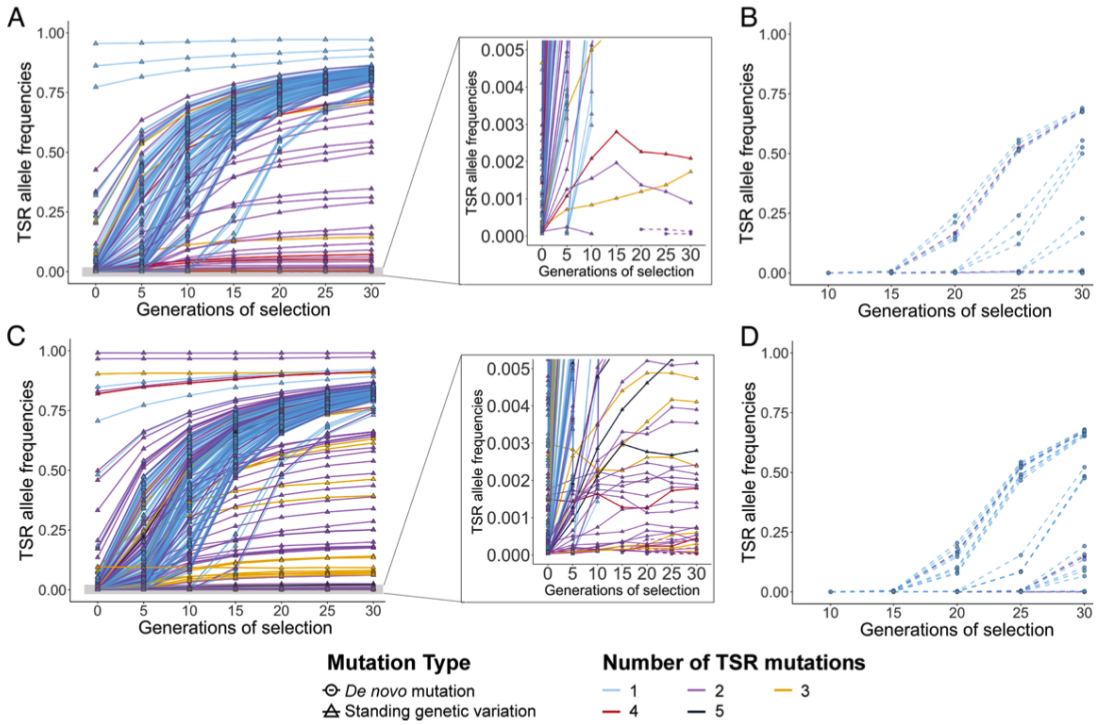


Fig. 5. Simulations of expected allele frequencies for TSR alleles arising from standing genetic variation or de novo mutation. The distribution of mutations was generated with a generic gene model that has the same number of exons and introns, and ratio between coding and noncoding sequences as the *ACCase* gene. Mutations in introns and noncoding regions were considered to be neutral, while exons had a ratio of 0.25/0.75 (neutral/deleterious) mutations according to Messer and Petrov (52), with selection coefficients (s) for deleterious mutations drawn from a gamma distribution with E[s] = −0.000154 and a shape parameter of 0.245 (51). Five hundred of one thousand simulation runs are shown for an effective population size (N_e) of (*A* and *B*) 42,000 individuals and (*C* and *D*) 84,000 individuals. Continuous lines represent mutations originating from standing genetic variation; de novo TSR mutations are shown with dashed lines. Colors indicate the total number of TSR mutations per population. (*A* and *C*) Standing genetic variation scenario, with TSR mutations preexisting in the populations before herbicide selection. Shown is the increase in TSR allele frequencies under herbicide selection of up to 30 generations, with one herbicide application per generation. The right panel shows a truncated *Y* axis at 0.005 TSR allele frequencies. Notice that some TSR de novo mutations have also arisen in runs that had preexisting TSR alleles. (*B* and *D*) De novo mutation scenario. Any TSR mutation that might have arisen before the start of selection has been lost again, so that no TSR mutations are present at generation 0 of selection.

TSR mutations having been present already at low frequency before herbicides came into use, as shown through herbicide treatment of naive populations of *Lolium rigidum* (60) or the analysis of herbarium samples of *A. myosuroides* collected before the advent of modern herbicides (61). In the case of *L. rigidum*, the frequency of sulfometuron-methyl resistance in previously untreated populations was around 10^{-4} (60) while among 685 *A. myosuroides* herbarium specimens, one individual collected nearly hundred years before the introduction of herbicides carried the *ACCase* Ile-1781-Leu mutation (61). In fact, we found TSR mutations in two of our three sensitive reference populations by deep amplicon sequencing (HerbiSeed standard, WHBM72 greenhouse standard APR/HA from September 2014), although it is unknown when these populations were collected with respect to the relevant herbicides coming into broad use. On the other hand, in a study that aimed to empirically determine the de novo mutation rate of TSRs in a natural population of grain amaranth (*Amaranthus hypochondriacus*), not a single spontaneous resistant genotype was found among 70 million screened plants (62). This would give 1.4×10^{-8} as an approximate upper bound of spontaneous mutations conferring resistance to a specific herbicide, which is in the range of spontaneous mutation rates that have been empirically measured in plants for single sites (63, 64).

Discussion

Plants have evolved a remarkable number of mechanisms to protect themselves against damage and extinction from changing environmental conditions, including ones due to human activity. In particular, the outsized selection imposed by repeated application of herbicides had led to extraordinarily rapid evolutionary adaptation in many weed species.

While several TSR mutations incur fitness penalties in the absence of herbicide applications (40, 41), some do not, and there is even at least one report of a TSR mutation being favorable independently of herbicide application (38, 39). The extent of fitness costs before herbicide application began in a population will in turn affect whether TSR mutations can accumulate in a population that is not under herbicide selection. Our study shows that standing genetic variation is primarily responsible for the observed level of per-field diversity of TSR haplotypes of independent origin that is associated with the rapid evolution of herbicide resistance in *A. myosuroides* populations. This suggests that the TSR mutations that are the focus of our investigation have limited fitness costs in the absence of herbicide treatment.

Another factor that likely influences the speed of TSR resistance is the presence and abundance of NTSR alleles, which in turn will be affected by the herbicide regime in that particular

population. For example, herbicide mixtures promote unspecific resistance through NTSR due to enhanced metabolism of herbicides with diverse modes of action (65). In our collection, there is a substantial fraction of individuals resistant to either ACCase or ALS inhibitors that cannot be attributed to known TSR mutations (*SI Appendix*, Fig. S12). The ratio of TSR to NTSR varies greatly, with some populations having only one or the other, and other populations having both (*SI Appendix*, Fig. S13). In many cases, evolutionary adaptation in response to a change in the environment occurs via soft selective sweeps, as this allows a greater proportion of ancestral genetic diversity to be maintained (66). In our case, there was a wide range in the fraction of TSR individuals per population but no correlation with genome-wide nucleotide diversity (π) (Pearson's r: 0.26, *P*-value: 0.075) (*SI Appendix*, Fig. S14), consistent with previous analyses using AFLP markers (11). The preservation of genetic diversity is particularly important in agricultural fields with highly variable conditions in terms of crop rotation, pest management, and other field management measures and can be crucial for weed populations to thrive under a range of different environmental conditions. Resistance apparently evolves in parallel in numerous fields with weed populations, occurring through different mechanisms that nature has at its disposal—TSR at a particular locus being one of them—depending on which resistance pathways have preexisting mutations.

Highly accurate long-reads have enabled us to resolve entire haplotypes of TSR genes and thus to ascertain their independent origin. To determine the contribution of gene flow, the reconstruction of larger haplotypes extending dozens or hundreds of kilobases will be necessary. Haplotype information can be inferred from whole-genome shotgun sequencing data by ancestral recombination graphs (67) or by targeted long-read sequencing to reassemble a larger genomic region. The high-quality genome assembly we have disclosed here provides a foundation for such future analyses.

In conclusion, our examination of different scenarios for adaptation to herbicides indicates that with the diversity of resistance mechanisms available, a large fraction of *A. myosuroides* populations is likely to have the genetic prerequisites not only for rapid evolution of resistance to currently used herbicide modes of action but also to potential new future modes of action.

Materials and Methods

For detailed experimental and analytical procedures, please see *SI Appendix, Supporting Text*.

Reference Genome. A single plant from an herbicide-sensitive population (Appels Wilde Samen GmbH, Darmstadt) from Germany was sequenced with CLR in a PacBio Sequel I system. FALCON-Unzip toolkit was used for initial assembly (20), and contigs were subjected to deduplication with purge_dups v1.0.0 (21). Hi-C library reads were used as input data for HiRise for chromosome-level scaffolding (22). To aid gene annotation, both Illumina RNA-seq and PacBio Iso-seq data from five tissues (anthers, whole inflorescences, leaves, pollen, and roots) from the same individual were generated.

Population Studies. For the population structure analysis of 47 European *A. myosuroides* populations, the ddRAD libraries were prepared according to a published method for fresh samples (68) and sequenced in an Illumina NovaSeq 6000 system on a S2 FlowCell in paired-end mode and with a read length of 150 bp to an average coverage of 22.6x read depth. Variants were called with GATK v4.1.3.0 (69), and SNPs were filtered following the recommendations of the RAD-Seq variant-calling pipeline 'dDocent' (70). The ML phylogenetic tree that shows the genetic relationship between the samples of our European dataset was inferred with RAXML-NG v0.9.0 (71) and visualized with the interactive Tree Of Life online tool (72) (Fig. 2*A*). The identification of ancestry groups was performed with ADMIXTURE (73) (Fig. 2*D*). Effective population sizes were calculated after the formula $N_e = \theta_W/4^*\mu$ for a diploid organism. Watterson thetas θ_W were estimated with ANGSD v0.930 (74) exclusively from the ddRAD-sequenced portion of the *A. myosuroides* assembly. The mutation rate $\mu = 3.0 \times 10^{-8}$ was adopted from *Zea mays* (31).

Amplicon Analysis. *ALS* and *ACCase* long-range amplicons were generated with the barcoded primers listed in Dataset S1, sequenced in a PacBio Sequel I system, and converted to haplotypes with the tool PacBio Amplicon Analysis (https://github.com/PacificBiosciences/pbAA). Multiple alignments of all haplotypes per population were performed with MAFFT v7.407 (75), trees were inferred with RAXML-NG v0.9.0 (71), and minimum spanning networks were visualized with POPART v.1.7 (76) (Fig. 3 and Datasets S2 and S3).

Simulations. To model the general probability of adaptation through a sweep and then specifically from standing genetic variation (Fig. 4), we used equations 8, 11, 14, 18, and 20 from Hermisson and Pennings (34). For forward-in-time simulations (Fig. 5 and *SI Appendix*, Fig. S10), we used the software SLiM v3.4 (16) with the *ACCase* locus (12,250 bp) as a template, a burn-in period of $10 \times N_e$ generations, and 30 generations of selection (*SI Appendix*, Fig. S9). Both the mutation rate (3.0×10^{-8}) (31) and genome-wide average recombination rate (7.4×10^{-9}) (77) were adopted from *Z. mays*. We set the population size to 42,000 individuals, which is the highest possible N_e from the populations characterized with RAD-Seq data. Since diversity estimates of N_e integrate over a long period of time and past bottlenecks will reduce it, leading to estimates that are lower than the actual N_e before the bottlenecks (33), we additionally simulated the doubled effective population size of 84,000 individuals.

Data, Materials, and Software Availability. Raw data including PacBio CLR and Iso-seq reads, Illumina PCR-free, Hi-C, and RNA-seq reads can be accessed in the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under project accession number PRJEB49257 (78), assembly accession CASDCE010000000 (79). Raw ddRAD-seq data for the population study, and PacBio CCS q20 reads can be downloaded from the ENA project accession number PRJEB49288 (80). Annotation files for the genome assembly, the SNP matrix for the ddRAD-seq experiment, and the fasta files with the haplotypes of *ACCase* and *ALS* can be found at https://doi.org/10.5281/zenodo.7634530 (81). Scripts and experimental protocols to reproduce the analyses in this study are deposited in the GitHub repository of this study (https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022) (82).

ACKNOWLEDGMENTS. We thank Andreas Landes (BASF SE, Germany) for the European blackgrass accessions, Daniel Hewitt and John Cussans (National Institute of Agricultural Botany, United Kingdom) for the sensitive reference seeds of the Broadbalk long-term experiment (Rothamsted, 2013), Johannes Herrmann for helpful discussions on population structure and haplotype networks, Derek Lundberg for advice on amplicon sequencing, Frank Chan for facilitating the use of the flow cytometer, Marek Kučka for providing purified *Tn5* transposase, Dovetail Genomics for Hi-C library preparation and genome scaffolding, and Rudi Antonise (KeyGene, Netherlands) for the genotyping-by-sequencing service of herbicide-sensitive populations that allowed selection of the reference population. S.K. was supported by a stipend from the Landesgraduiertenförderung (State Graduate Scholarship, LGFG) of the State of Baden-Württemberg. F.A.R. was supported by a Human Frontiers Science Program Long-Term Fellowship (LT000819/2018-L). The majority of funding was provided by BASF and the Max Planck Society.

Author affiliations: [a]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany; [b]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany; [c]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium; [d]Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium; [e]Department of Biology, The Eberly College of Science, Penn State University, State College, PA 16801; [f]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Bio Center, 1030 Vienna, Austria; [g]Friedrich Miescher Laboratory 72076 Tübingen, Germany; [h]Agricultural Research Station, BASF SE, 67117 Limburgerhof, Germany; [i]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa; and [j]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

Preprint Servers: bioRxiv: 10.1101/2021.12.14.472587.

Author contributions: S.K., D.W., and F.A.R. designed research; S.K., C.L., T.H., P.L., U.L., I.H., and F.A.R. performed research; J.L. and A.P. contributed new reagents/analytic tools; S.K., J.C., C.D.H., Y.V., and F.A.R. analyzed data; Y.V.d.P., K.S., D.W. and F.A.R. supervised research; and S.K., D.W., and F.A.R. wrote the paper.

Reviewers: T.A.G., Colorado State University; and P.S.P., San Francisco State University.

1. S. R. Moss, S. A. M. Perryman, L. V. Tatnell, Managing herbicide-resistant blackgrass (Alopecurus Myosuroides): Theory and practice. *Weed Technol.* **21**, 300–309 (2007).
2. M. Rosenhauer, B. Jaser, F. G. Felsenstein, J. Petersen, Development of target-site resistance (TSR) in Alopecurus myosuroides in Germany between 2004 and 2012. *J. Plant Dis. Prot.* **120**, 179–187 (2013).
3. A. Varah *et al.*, The costs of human-induced evolution in an agricultural system. *Nat. Sustainability* **3**, 63–71 (2019).
4. C. Délye, X.-Q. Zhang, S. Michel, A. Matéjicek, S. B. Powles, Molecular bases for sensitivity to acetyl-coenzyme A carboxylase inhibitors in black-grass. *Plant Physiol.* **137**, 794–806 (2005).
5. H. Xu *et al.*, Mutations at codon position 1999 of acetyl-CoA carboxylase confer resistance to ACCase-inhibiting herbicides in Japanese foxtail (Alopecurus japonicus). *Pest Manag. Sci.* **70**, 1894–1901 (2014).
6. P. J. Tranel, T. R. Wright, Resistance of weeds to ALS-inhibiting herbicides: What have we learned? *Weed Sci.* **50**, 700–712 (2002).
7. C. Délye, K. K. Boucansaud, A molecular assay for the proactive detection of target site-based resistance to herbicides inhibiting acetolactate synthase in Alopecurus myosuroides. *Eur. Weed Res. Soc. Weed Res.* **48**, 97–101 (2007).
8. D. M. Shah *et al.*, Engineering herbicide tolerance in transgenic plants. *Science* **233**, 478–481 (1986).
9. T. A. Gaines *et al.*, Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1029–1034 (2010).
10. C. Délye, Unravelling the genetic bases of non-target-site-based resistance (NTSR) to herbicides: A major challenge for weed science in the forthcoming decade. *Pest Manag. Sci.* **69**, 176–187 (2013).
11. Y. Menchari, C. Délye, V. Le Corre, Genetic variation and population structure in black-grass (Alopecurus myosuroides Huds.), a successful, herbicide-resistant, annual grass weed of winter cereal fields. *Mol. Ecol.* **16**, 3161–3172 (2007).
12. A. Dixon, D. Comont, G. T. Slavov, P. Neve, Population genomics of selectively neutral genetic structure and herbicide resistance in UK populations of Alopecurus myosuroides. *Pest Manag. Sci.* **77**, 1520–1529 (2020).
13. C. Délye, J. Gardin, K. Boucansaud, B. Chauvel, C. Petit, Non-target-site-based resistance should be the centre of attention for herbicide resistance research: Alopecurus myosuroides as an illustration. *Weed Res.* **51**, 433–437 (2011).
14. C. Petit, B. Duhieu, K. Boucansaud, C. Délye, Complex genetic control of non-target-site-based resistance to herbicides inhibiting acetyl-coenzyme A carboxylase and acetolactate-synthase in Alopecurus myosuroides Huds. *Plant Sci.* **178**, 501–509 (2010).
15. S. L. Martin *et al.*, Population genomic approaches for weed science. *Plants* **8**, 354 (2019).
16. B. C. Haller, P. W. Messer, SLiM 3: Forward genetic simulations beyond the wright-fisher model. *Mol. Biol. Evol.* **36**, 632–637 (2019).
17. J. Pellicer, I. J. Leitch, The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* **226**, 301–305 (2019).
18. V. K. Sieber, B. G. Murray, The cytology of the genus Alopecurus (Gramineae). *Bot. J. Linn. Soc.* **79**, 343–355 (1979).
19. J. Doležel, J. Greilhuber, S. Lucretti, A. Meister, Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Ann. Bot.* **82**, 17–26 (1998).
20. C.-S. Chin *et al.*, Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
21. D. Guan *et al.*, Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
22. N. H. Putnam *et al.*, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
23. International Barley Genome Sequencing Consortium *et al.*, A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711–716 (2012).
24. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
25. B. S. Gaut, B. R. Morton, B. C. McCaig, M. T. Clegg, Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 10274–10279 (1996).
26. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
27. S. Ou *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
28. E. Quevillon *et al.*, InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
29. J. Evans *et al.*, Extensive genetic diversity is present within North American switchgrass germplasm. *Plant Genome* **11** (2018).
30. E. M. Leffler *et al.*, Revisiting an old riddle: What determines genetic diversity levels within species? *PLoS Biol.* **10**, e1001388 (2012).
31. N. Yang *et al.*, Contributions of Zea mays subspecies mexicana haplotypes to modern maize. *Nat. Commun.* **8**, 1874 (2017).
32. J. Hermisson, P. S. Pennings, Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**, 700–716 (2017).
33. P. W. Messer, D. A. Petrov, Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).
34. J. Hermisson, P. S. Pennings, Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352 (2005).
35. B. Charlesworth, Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
36. Q. Yu, S. B. Powles, Resistance to AHAS inhibitor herbicides: Current understanding. *Pest Manag. Sci.* **70**, 1340–1350 (2014).
37. H. K. Takano, R. F. L. Ovejero, G. G. Belchior, G. P. L. Maymone, F. E. Dayan, ACCase-inhibiting herbicides: Mechanism of action, resistance evolution and stewardship. *Sci. Agric.* **78**, e20180387 (2020).
38. T. Wang, J. C. Picard, X. Tian, H. Darmency, A herbicide-resistant ACCase 1781 Setaria mutant shows higher fitness than wild type. *Heredity* **105**, 394–400 (2010).
39. C. Délye, Y. Menchari, S. Michel, E. Cadet, V. Le Corre, A new insight into arable weed adaptive evolution: Mutations endowing herbicide resistance also affect germination dynamics and seedling emergence. *Ann. Bot.* **111**, 681–691 (2013).
40. H. Darmency, Y. Menchari, V. Le Corre, C. Délye, Fitness cost due to herbicide resistance may trigger genetic background evolution. *Evolution* **69**, 271–278 (2015).
41. L. Du *et al.*, Fitness costs associated with acetyl-coenzyme A carboxylase mutations endowing herbicide resistance in American sloughgrass (Beckmannia syzigachne Steud.). *Ecol. Evol.* **9**, 2220–2230 (2019).
42. C. Délye, C. Straub, A. Matéjicek, S. Michel, Multiple origins for black-grass (Alopecurus myosuroides Huds) target-site-based resistance to herbicides inhibiting acetyl-CoA carboxylase. *Pest Manag. Sci.* **60**, 35–41 (2004).
43. C. Délye, C. Straub, S. Michel, V. Le Corre, Nucleotide variability at the acetyl coenzyme A carboxylase gene and the signature of herbicide selection in the grass weed Alopecurus myosuroides (Huds.). *Mol. Biol. Evol.* **21**, 884–892 (2004).
44. Y. Menchari *et al.*, Weed response to herbicides: Regional-scale distribution of herbicide resistance alleles in the grass weed Alopecurus myosuroides. *New Phytol.* **171**, 861–873 (2006).
45. J. Herrmann, S. T. Hess M, H. Strek, O. Richter, R. Beffa, Spatial and temporal development of ACCase and ALS resistant Black-grass (Alopecurus myosuroides Huds.) populations in neighboring fields in Germany. *Julius-Kühn-Archiv.* **443**, 273–279 (2014).
46. R. S. Baucom, The remarkable repeated evolution of herbicide resistance. *Am. J. Bot.* **103**, 181–183 (2016).
47. C. Délye, K. Boucansaud, A molecular assay for the proactive detection of target site-based resistance to herbicides inhibiting acetolactate synthase in Alopecurus myosuroides. *Weed Res.* **48**, 97–101 (2008).
48. R. Beffa *et al.*, Weed resistance diagnostic technologies to detect herbicide resistance in cereal-growing areas: A review. *Julius-Kühn-Archiv.* **434**, 75–80 (2012).
49. R. Marshall, S. R. Moss, Characterisation and molecular basis of ALS inhibitor resistance in the grass weed Alopecurus myosuroides. *Weed Res.* **48**, 439–447 (2008).
50. C. M. Knight, "*Investigating the evolution of herbicide resistance in UK populations of Alopecurus myosuroides*," University of Warwick, Coventry, United Kingdom. (2015).
51. C. D. Huber, A. Durvasula, A. M. Hancock, K. E. Lohmueller, Gene expression drives the evolution of dominance. *Nat. Commun.* **9**, 2750 (2018).
52. P. W. Messer, D. A. Petrov, Frequent adaptation and the McDonald-Kreitman test. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8615–8620 (2013).
53. B. Arnold, R. B. Corbett-Detig, D. Hartl, K. Bomblies, RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22**, 3179–3190 (2013).
54. M. Cariou, L. Duret, S. Charlat, How and how much does rad-seq bias genetic diversity estimates? *BMC Evol. Biol.* **16**, 240 (2016).
55. B. Chauvel, J.-P. Guillemin, N. Colbach, Evolution of a herbicide-resistant population of Alopecurus myosuroides Huds. in a long-term cropping system experiment. *Crop Prot.* **28**, 343–349 (2009).
56. B. Chauvel, J. P. Guillemin, N. Colbach, J. Gasquez, Evaluation of cropping systems for management of herbicide-resistant populations of blackgrass (Alopecurus myosuroides Huds.). *Crop Prot.* **20**, 127–137 (2001).
57. J. Petersen, M. Dresbach-Runkel, J. Wagner, A method to determine the pollen-mediated spread of target-site resistance to acetyl-coenzyme A carboxylase inhibitors in black grass (Alopecurus myosuroides Huds.). *J. Plant Dis. Prot.* **117**, 122–128 (2010).
58. P. Neve, S. Powles, High survival frequencies at low herbicide use rates in populations of Lolium rigidum result in rapid evolution of herbicide resistance. *Heredity* **95**, 485–492 (2005).
59. H. L. Hicks *et al.*, The factors driving evolved herbicide resistance at a national scale. *Nat. Ecol. Evol.* **2**, 529–536 (2018).
60. C. Preston, S. B. Powles, Evolution of herbicide resistance in weeds: Initial frequency of target site-based resistance to acetolactate synthase-inhibiting herbicides in Lolium rigidum. *Heredity* **88**, 8–13 (2002).
61. C. Délye, C. Deulvot, B. Chauvel, DNA analysis of herbarium Specimens of the grass weed Alopecurus myosuroides reveals herbicide resistance pre-dated herbicides. *PLoS One* **8**, e75117 (2013).
62. F. A. Casale, D. A. Giacomini, P. J. Tranel, Empirical investigation of mutation rate for herbicide resistance. *Weed Sci.* **67**, 361–368 (2019).
63. M.-L. Weng *et al.*, Fine-grained analysis of spontaneous mutation spectrum and frequency in Arabidopsis thaliana. *Genetics* **211**, 703–714 (2019).
64. M. Exposito-Alonso *et al.*, The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).
65. D. Comont *et al.*, Evolution of generalist resistance to herbicide mixtures reveals a trade-off in resistance management. *Nat. Commun.* **11**, 3086 (2020).
66. B. A. Wilson, P. S. Pennings, D. A. Petrov, Soft selective sweeps in evolutionary rescue. *Genetics* **205**, 1573–1586 (2017).
67. J. M. Kreiner *et al.*, Repeated origins, widespread gene flow, and allelic interactions of target-site herbicide resistance mutations. *Elife* **11**, e70242 (2022).
68. P. L. M. Lang *et al.*, Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Mol. Ecol. Resour.* **20**, 1228–1247 (2020).
69. G. A. Van der Auwera *et al.*, From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
70. J. B. Puritz, C. M. Hollenbeck, J. R. Gold, dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**, e431 (2014).
71. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
72. I. Letunic, P. Bork, Interactive Tree Of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
73. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
74. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
75. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
76. J. W. Leigh, D. Bryant, popart: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015).

77. E. Bauer *et al.*, Intraspecific variation of recombination rate in maize. *Genome Biol.* **14**, R103 (2013).
78. F. A. Rabanal, S. Kersten, D. Weigel, BioProject ID PRJEB49257: Chromosome-level assembly of an Alopecurus myosuroides individual from the German herbicide-sensitive population DE01087 using PacBio sequencing. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB49257. Deposited 30 April 2022.
79. F. A. Rabanal, S. Kersten, D. Weigel, Accession ID CASDCE010000000.1: *Alopecurus myosuroides* genome assembly. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/CASDCE010000000.1. Deposited 18 February 2023.
80. F. A. Rabanal, S. Kersten, D. Weigel, BioProject ID PRJEB49288: Herbicide resistance evolution in blackgrass. *European Nucleotide Archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB49288. Deposited 30 April 2022.
81. S. Kersten, C. Jiyang, Y. Van de Peer, D. Weigel, F. A. Rabanal, Additional Files for "Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass". *Zenodo*. https://doi.org/10.5281/zenodo.7634530. Deposited 13 February 2023.
82. S. Kersten. GitHub repository for "Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass". *GitHub*. https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022. Deposited 17 December 2022.

3.2 Supplementary

PNAS

Supporting Information for

Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass

Sonja Kersten[1,2], Jiyang Chang[3,4], Christian D. Huber[5], Yoav Voichek[6], Christa Lanz[2], Timo Hagmaier[2], Patricia Lang[2†], Ulrich Lutz[2], Insa Hirschberg[7], Jens Lerchl[8], Aimone Porri[8], Yves Van de Peer[3,4,9,10], Karl Schmid[1], Detlef Weigel[2], and Fernando A. Rabanal[2]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany.
[2]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany.
[3]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium.
[4]Center for Plant Systems Biology, Vlaams Instituut voor Biotechnologie, 9052 Ghent, Belgium.
[5]Department of Biology, The Eberly College of Science, Penn State University, State College, PA 16801, USA.
[6]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna BioCenter, 1030 Vienna, Austria.
[7]Friedrich Miescher Laboratory, 72076 Tübingen, Germany.
[8]Agricultural Research Station, BASF SE, 67117 Limburgerhof, Germany.
[9]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa.
[10]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, 210095, China.
[†]current address: Department of Biology, Stanford University, Stanford, CA 94305, USA.

Detlef Weigel
Email: weigel@weigelworld.org

This PDF file includes:

Supporting text
Figures S1 to S14
Tables S1 and S2
Legends for Datasets S1 to S3
SI References

Other supporting materials for this manuscript include the following:

Datasets S1 to S3

Supporting Information Text

Reference genome sequencing, assembly and annotation

Plant selection and flow cytometry

A single plant from a sensitive German reference population provided by BASF was selected. All required tissues for all described reference related sequencing methods were collected from the same plant. We confirmed the absence of known TSR mutations on the *ACCase*, *ALS* and *psbA* loci using Illumina amplicon sequencing. PCR products of the three target genes (Dataset S1) were pooled, and sequencing libraries were generated with a purified *Tn5* transposase as described in a previous study (1). The library was spiked into an Illumina HiSeq 3000 lane. The resulting reads were checked for known TSR mutations causing herbicide resistance (2, 3).

Leaf tissue from both the selected *A. myosuroides* plant and the reference standard *Secale cereale* cv. Daňkovské (4) were simultaneously chopped with a razor blade in 250 µl of nuclei extraction buffer (CyStain PI Absolute P kit; P/N 05-5022). After the addition of 1 ml of staining solution (including 6 µl of propidium iodide (PI) and 3 µl of RNase from the same kit) the suspension was filtered through a 30 µm filter (CellTrics®; P/N 04-0042-2316). Five replicates of these samples were stored in darkness for 4 h at 4°C prior to flow cytometry analysis. PI-area was detected with a BD FACSMelody™ Cell Sorter (BD Biosciences) equipped with a yellow-green laser (561 nm) and 613/18BP filtering. A total of 25,000 events were recorded per replicate, and the ratio of the mean PI-area values of each target sample and reference standard 2C peaks was used to estimate DNA content according to ref. (5) (mean = 3.53 Gb; s.d. = 0.0052 Gb; n = 5).

Whole-genome PacBio sequencing

Prior to high-molecular weight (HMW) extraction the reference plant was kept for 48 hours in the dark to reduce the starch accumulation. We harvested ca. 30 g of young leaf material and ground it in liquid nitrogen. Nuclei isolation was performed according to a published protocol (6) with the following modifications: we used 16 reactions, each with 1 g input material in a 20 ml nuclear isolation buffer. The filtered cellular homogenate was

centrifuged at 3500 x g, followed by 3x washes in nuclear isolation buffer. The isolated plant cell nuclei were resuspended in 60 µl Proteinase K (#19131, Qiagen). For HMW-DNA recovery, the Nanobind Plant Nuclei Big DNA Kit (SKU NB-900-801-01, Circulomics) was used. In total, we obtained approximately 80 µg of HMW-DNA, which was subjected to needle shearing once (FINE-JECT® 26Gx1'' 0.45x25mm, LOT 14-13651). A 75-kb template library was prepared with the SMRTbell® Express Template Preparation Kit 2.0, and size-selected with the BluePippin system (SageScience) with 15-kb cutoff and a 0.75% agarose, 1-50kb cassette (BLF7510, Biozym) according to the manufacturer's instructions (P/N 101-693-800-01, Pacific Biosciences, California, USA). The library was sequenced on a Sequel I system (Pacific Biosciences) using the Binding Kit 3.0. and MagBead loading. In total, we sequenced 18 SMRT cells of 10 hours and 8 SMRT cells of 20 hours movie time.

Illumina PCR-free library sequencing

The genomic DNA was fragmented to 350 bp size using a Covaris S2 Focused Ultrasonicator (Covaris) with the following settings: duty cycle 10%, intensity 5, 200 cycles and 45s treatment time. The library prep was performed according to the manufacturer's instructions for the NxSeq® AmpFREE Low DNA Library Kit from Lucigen® (Cat No. 14000-2) with the addition of a large-cutoff bead-cleanup (0.6 : 1, bead:library ratio) after the adapter ligation, followed by the recommended standard bead-cleanup at the final purification step. The library was quantified with the Qubit Fluorometer (Invitrogen) and quality checked on a Bioanalyzer High Sensitivity Chip on an Agilent Bioanalyzer 2100 (Kit #5067-4626, Agilent Technologies). The library was sequenced on two lanes of an Illumina HiSeq 3000 system in paired-end mode and with a read length of 150bp.

Short-read RNA-seq

RNA was extracted from five tissues (leaves, whole inflorescences, anthers, pollen, roots) following a published protocol (7). Remaining DNA was removed with DNaseI (#EN0521, Thermo Scientific) following manufacturer's recommendations. The quality was checked with an RNA 6000 Nano Chip on an Agilent Bioanalyzer 2100 (Kit

#5067-1511, Agilent Technologies). All RNA integrity number (RIN) scores were above 5.4.

For the library preparation, the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina in combination with the Poly(A) mRNA Magnetic Isolation Module (#E7760, #E7490, NEB) was used. The heat fragmentation was performed for a duration of 9 min resulting in final library sizes of around 545 bp. All 5 libraries were equally pooled and sequenced on one lane of an Illumina HiSeq 3000 system in paired-end mode and with a read length of 150 bp.

Long-read PacBio RNA Iso-seq

We extracted RNA from the same five tissue samples as for the short-read sequencing. To ensure a high RNA quality for long-read sequencing, we used a published protocol (8), which is a CTAB based method for high-quality total RNA applications from different plant tissues. The remaining DNA was removed with the TURBO DNA-free Kit (Invitrogen), designed for optimal preservation of RNA during the DNase treatment. The quality check on the Agilent Bioanalyzer 2100 (Agilent Technologies) with an RNA Nano 6000 Chip resulted in RIN scores higher than 7.6 for all tissues.

The IsoSeq libraries were prepared following the PacBio protocol for 'Iso-Seq™ Express Template Preparation for Sequel and Sequel II Systems' (P/N 101-763-800 Version 02; October 2019, Pacific Biosciences, California, USA). The cDNA was amplified in 12 cycles and purified using the 'standard' workflow for samples primarily composed of transcripts centered ~2 kb.

Hi-C library preparation

Hi-C libraries were prepared in a similar manner as described (9). Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp

mean fragment size and sequencing libraries were generated using NEBNextUltra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X.

Genome assembly

Genome assembly was done with the FALCON and FALCON-Unzip toolkit (10) distributed with the 'PacBio Assembly Tool Suite' (falcon-kit 1.3.0; pypeflow 2.2.0; https://github.com/PacificBiosciences/pb-assembly). For the pre-assembly step, in which CLR subreads are aligned to each other for error correction, we opted for auto-calculating our own seed read length ('length_cutoff = -1') with 'genome_size = 3530000000' and 'seed_coverage = 40' . Details of the FALCON assembly parameters used in this study are provided in the dedicated GitHub for this study (https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022). Primary contigs were subjected to deduplication with purge_dups v1.0.0 (11) using cutoffs (5, 36, 60, 72, 120, 216). For scaffolding, deduplicated primary contigs and Hi-C library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (12). Dovetail Hi-C library sequences were aligned to the draft input assembly using bwa (13). The separations of Dovetail Hi-C pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold.

Genome annotation

Transposable elements were annotated with the tool Extensive *de-novo* TE Annotator (EDTA) v1.9.7 (14). The protein-coding gene annotation pipeline involved merging three independent approaches: RNA-aided annotation, *ab initio* prediction and protein homology search. The first approach is based on both RNA-seq and Iso-seq data from five tissues, anthers, whole inflorescences, leaves, pollen and roots. Pre-processing of Iso-seq data was carried out with PacBio® tools (https://github.com/PacificBiosciences/pbbioconda) that included in a first step the

generation of Circular Consensus Sequencing (CCS) reads (minimum predicted accuracy 0.99 or q20) with ccs v5.0.0 and demultiplexing with lima v2.0.0. In a second step, poly-A trimming and concatemer removal were done at the sample level (i.e., separately for each tissue) while clustering was carried out for all tissues combined with functions from isoseq3 v3.4.0. Unique isoforms had a mean length of 2,210 bp.

Iso-seq clusters were aligned to the *A. myosuroides* genome using GMAP v2017-11-15 using default parameters(15), whereas RNA-seq datasets were first mapped to the *A. myosuroides* genome using Hisat2 (16) and subsequently assembled into transcripts by StringTie2 (17). All transcripts from Iso-seq and RNA-seq were combined using Cuffcompare (18). Transdecoder v5.0.2 (https://github.com/TransDecoder) was then used to find potential open reading frames (ORFs) and to predict protein sequences. To further maximize sensitivity for capturing ORFs that may have functional significance, BLASTP(19) (v2.6.0+, arguments -max_target_seqs 1 -evalue 1e-5) was used to compare potential protein sequences with the Uniprot database (20). In the second approach, *ab initio* prediction was performed by BRAKER2 (21) using a model trained with RNA-seq data from *A. myosuroides*. For the third approach, consisting of homology prediction, the protein sequences from five closely related species (*Brachypodium distachyon*, *Oryza sativa*, *Setaria italica*, *Sorghum bicolor* and *Hordeum vulgare*) that belong to the same family were used as query sequences to search the reference genome using TBLASTN (e < 1e-5). These databases were downloaded from Plaza v4.5 (22) (https://bioinformatics.psb.ugent.be/plaza/). Regions mapped by these query sequences were subjected to Exonerate (23) to generate putative transcripts.

Finally, EvidenceModeler v1.1.1 (24) was used to integrate all of the above sources of evidence, and the Benchmarking Universal Single-Copy Orthologs (BUSCO; v4.0.4; embryophyta_odb10) gene set to assess the quality of annotation results (25). Putative gene functions were identified using InterProScan (26) with different databases, including PFAM, Gene3D, PANTHER, CDD, SUPERFAMILY, ProSite, GO. Meanwhile, functional annotation of these predicted genes was obtained by aligning the protein sequences of these genes against the sequences in public protein databases and the UniProt database using BLASTP (e-value $<1 \times 10^{-5}$).

Comparative genomics

Analyses related to synonymous substitution rates (K_S) were performed using the wgd package (27). First, the paranome (entire collection of duplicated genes) was obtained with 'wgd mcl' using all-against-all BLASTP and MCL clustering. Then, the K_S distribution of *A. myosuroides* was calculated using 'wgd ksd' with default settings, MAFFT V7.453 (28) for multiple sequence alignment, and codeml from PAML package v4.4c (29) for maximum likelihood estimation of pairwise synonymous distances. Anchors or anchor pairs (duplicates located in collinear or syntenic regions of the genome) were obtained using i-ADHoRe (30), employing the default settings in 'wgd syn'.

Plant genomes typically contain both whole-genome and segmental duplications. We therefore investigated collinear regions indicative of recent duplications. When we analyzed the divergence of closely related paralogs present in these regions based on synonymous substitution rates (K_S), we noticed two main peaks, one at K_S ~0.16 and another one at K_S ~1.2 (Figure S1B). The K_S of the first peak is unusually low and would normally indicate very recent duplicates. To explore the nature of the gene pairs with low K_S, we extracted all gene pairs in these regions with $K_S \leq 0.5$ and asked how they are distributed in the genome. Collinear blocks containing these pairs are generally very close and always within the same chromosome (Figure S1C), while pairs with $K_S > 0.5$ are located in different chromosomes (Figure 1B). One explanation would be that these blocks are the products of recent duplication events, although there is not much evidence for large-scale local duplications in plant genomes. Alternatively, they could be an artifact of the assembly process, as in highly heterozygous genomes, different alleles can be assembled independently into different contigs. If these duplicates are not properly purged, which is particularly difficult if alleles are very dissimilar, then during scaffolding they are placed close to each other on the same chromosome. With the data at hand, it is difficult to distinguish between these two possibilities, but based on the close paralogs being almost always present close to each other, we favor the second explanation. The second peak (K_S ~1.2), mostly representing paralogs in different chromosomes (Figure 1B, Figure S1B), coincides with a known whole-genome duplication (WGD) event common in all grasses (31, 32) that occurred ~70 million years ago (mya). The list of anchors and their K_S values is available in Dataset S1.

MCscan JCVI (33) was used to do the analysis of syntenic relationships and depth ratio by providing the coding DNA sequences (CDS) and annotation file in gff3 format. TBtools was used to visualize the results via a Circos plot (34).

Population studies

Sample collection and DNA extraction

Seeds from 44 *A. myosuroides* populations from nine European countries were provided by BASF. The seeds were collected from farmers with suspected herbicide resistance in their fields against ACCase – and/or ALS-inhibiting herbicides. In addition, we included three sensitive reference populations (HerbiSeed standard, Broadbalk long-term experiment Rothamsted 2013, WHBM72 greenhouse standard APR/HA from September 2014).

The seeds of all 47 populations were sown in vermiculite substrate and stratified in a 4°C climatic chamber for one week, and subsequently placed in the greenhouse at 23°C / 8 h daytime, 18°C / 16 h nighttime regime. After one week in the greenhouse, one plant per pot was transferred to standard substrate (Pikiererde Typ CL P, Cat. No. EN12580, Einheitserde) for a total of 27 plants per population. We aimed to collect 8-weeks-old leaf tissue from 24 individuals per population, but due to insufficient germination in two populations, we were unable to collect material from two individuals and therefore finally obtained 1,126 samples for further processing. 300 mg of plant material was collected into a 2 ml screw cap tube filled with 4-5 porcelain beads and ground with a FastPrep tissue disruptor (MP Biomedicals). For DNA extraction, we used a lysis buffer consisting of 100 mM Tris (pH 8.0), 50 mM EDTA (pH 8.0), 500 mM NaCl, 1,3% SDS and 0.01 mg/ml RNase A. The DNA was precipitated with 5M potassium acetate, followed by two bead-cleanups for DNA purification. For a detailed hands-on protocol, see https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022.

Phenotyping

For phenotyping, 27 plants per population described in the previous section were divided into two treatment and two control groups, following a specific tray design to minimize

spatial growth effects. Treatment 1: Atlantis WG® (Bayer Crop Science) + Synergist Atlantis WG® (10 plants per population). Control 1: Only Synergist Atlantis WG® (three plants per population). Treatment 2: Axial® 50 (Syngenta) + Synergist Hasten (10 plants per population). Control 2: Only Synergist Hasten (four plants per population). All plants were sprayed 11 weeks after transplanting. Herbicides and synergists were applied with a lab sprayer (Schachtner), nozzle Teejet 8001 EVS and an air pressure of between 200-225 kPa. The sprayer was calibrated for a field application rate of 400 l/ha in four rounds of three independent replicates each (M = 396.4, SD = 7.53). Axial® 50 (50 g/l of pinoxaden + 12.5 g/l Cloquintocet-mexyl) was applied in combination with the synergist Hasten (716 g/l rapeseed oil ethyl and methyl esters, 179 g/l nonionic surfactants, ADAMA Deutschland GmbH). Atlantis WG® (29.2 g/kg of mesosulfuron and 5.6 g/kg of iodosulfuron) was used with the provided synergist (276,5 g/l sodium salt, fatty alcohol ether sulfate, Bayer Crop Science). Control plants were sprayed only with the synergists. Axial® 50 was applied at the recommended field rate of 1.2 l/ha , Atlantis WG® at 800 g/ha and both synergists at 1 l/ha. After four weeks all plants were scored according to the scheme in (Figure S12A), where the score D1 represents completely dead plants and the score A6 represents plants without any growth reductions compared to the control plants of the respective population.

ddRAD library preparation and sequencing

The ddRAD libraries were prepared according to a published method for fresh samples (35). 200 ng input DNA per sample were digested with the two restriction enzymes EcoRI (#FD0274, Thermo Fisher Scientific) and Mph1103I (FD0734, Thermo Fisher Scientific), followed by double-stranded custom-adapter ligation. The custom-adapters contain different numbers of additional nucleotides to shift the sequencing of the restriction enzyme sites and prevent the sequencer from causing an error due to unique signaling. After the restriction enzyme digestion step and the adapter ligation, large cutoff bead-cleanups (0.6:1, bead:library ratio) with homemade magnetic beads (Sera-Mag SpeedBeads™, #65152105050450, GE Healthcare Life Sciences) in PEG/NaCl buffer (36) were used to clean the samples from the buffers and remove large fragments above ~600 bp length. We used a dual-indexing PCR to be able to multiplex up to six 96-well plates of samples. Thus, two pools of libraries were sufficient for all our samples. Since it is challenging to determine exact library concentrations, our strategy

consisted of pooling all samples to the best of our abilities with the concentrations at hand and spike them into an Illumina HiSeq 3000 lane for about 5% of the total coverage. Afterwards, the library concentrations were re-calculated from the read coverage output and re-pooled accordingly to achieve a more even coverage. Size selection was performed using a BluePippin system (SageScience) with a 1.5% agarose cassette, 250bp-1.5kb (#BDF1510, Biozym) for a size range of 300–500 bp. The library pools were quantified with the Qubit Fluorometer (Invitrogen) and quality checked on a Bioanalyzer High Sensitivity Chip on an Agilent Bioanalyzer 2100 (Kit #5067-4626, Agilent Technologies). A detailed hands-on protocol can be found here: https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022.

First, each library pool was sequenced in-house on an Illumina HiSeq 3000 lane in paired-end mode and 150 bp read length to assess the performance and quality. Afterwards, both pools were submitted to CeGaT GmbH, Tübingen, and sequenced with an Illumina NovaSeq 6000 system on a S2 FlowCell with XP Lane Loading in paired-end mode and with a read length of 150 bp. Total data output was 1.4 Tb, representing an average coverage of 22.6x read depth.

Alignment, SNP calling and SNP filtering

Demultiplexed raw reads were first trimmed for the base-shifts of the custom adapters in the 5' and 3' fragment ends. Afterwards, all remaining adapter sequences and low-quality bases were removed and only reads with a minimum read length of 75 bp were kept using cutadapt v2.4 (37). The read quality was checked before and after trimming with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Paired-end reads were first merged using Flash v1.2.11 (38), then the extended and the unmerged reads were independently aligned to the reference genome using bwa-mem v0.7.17-r1194-dirty (13). We used samtools v1.9 (39) to sort and index the bam-files and to finally combine the bam files of the extended and unmerged aligned reads per sample.

Variant calling was performed with the HaplotypeCaller function of GATK v4.1.3.0 (40). For joint genotyping, we broke the reference at N-stretches and generated an interval list with Picard's v2.2.1 function 'ScatterIntervalsByNs'

(http://broadinstitute.github.io/picard/). Next, we generated a genomic database by using GATK v4.1.3.0 'GenomicsDBImport', followed by joint genotyping with 'GenotypeGVCFs'. A first missing data filter (--max-missing 0.3) was applied with VCFtools v0.1.15 (41) to the VCF outputs of all intervals to reduce the number of unusable variants. Afterwards all interval VCFs were merged with Picard v2.2.1 'MergeVcfs'. The combined VCF was filtered following the recommendations of the RAD-Seq variant-calling pipeline 'dDocent' (42). First, basic filters were applied with VCFtools v0.1.15 (--max-missing 0.5 --mac 3 --minQ 30 --minDP 3 --max-meanDP 35), followed by advanced filter options for RAD-Seq data with 'vcffilter' (ABHet > 0.25 & ABHet < 0.75 | ABHet < 0.01 & QD > 5 & MQ > 40 & MQRankSum > (0 - 5) & MQRankSum < 5 & ExcessHet < 30 & BaseQRankSum > (0 - 5) & BaseQRankSum < 5) (https://github.com/vcflib/vcflib). We also filtered individuals with missing data more than 0.5, which removed four individuals from our dataset, and we ended up with a total of 1,122 individuals. Lastly, we used a population specific variant filter, which allowed for 30% missing data, but every variant had to be called in at least 10 populations. Our final VCF for further analysis contained 109,924 informative SNPs.

Phylogeny and population genetics statistics

A maximum likelihood (ML) phylogenetic tree was inferred with RAXML-NG v0.9.0 (43) to display the genetic relationship between the samples of our European dataset. We inferred a single ML-tree without bootstrapping using the model GTR+G+ASC_LEWIS of nucleotide evolution with ascertainment bias correction since we inferred it on RAD-seq data. The annotation of the tree for the known TSR mutations was done based on the *ALS* and *ACCase* amplicons described below. For visualization, we used the interactive Tree Of Life (iTOL) online tool (44).

To assess the population structure of our European collection we ran a principal component analysis (PCA) with the R-package SNPrelate (45) on 101,114 biallelic informative SNPs. To perform the admixture analysis on shared ancestry, we first pruned the dataset with PLINK v1.90b4.1 (46) for only biallelic SNPs. Admixture v1.3.0 (47) was run for up to 10 k groups, using a 10-fold cross-validation procedure to infer the right amount of k groups. TreeMix v1-13 (48) was run on the VCF filtered with PLINK as previously described in the admixture analysis. The transformation into the right input file

format was done with STACKS v1.48 (--treemix) (49). The tree was rooted with the most divergent outgroup population NL11330 (-root NL11330) and inferred in windows of 50 SNPs (-k 50) with 5 bootstrap replicates (-bootstrap 5). Since the treemix F3 statistic did not show significant migration, no migration events were added to the tree. FSTs were calculated with STACKS v1.48 (--fstats) (49) and visualized with the R package ComplexHeatmap 2.0.0 (50).

Since we only covered about 1.1% of the entire genome with our ddRAD-Seq reads, we calculated the Watterson thetas θ_W and effective population sizes exclusively from the sequenced portion of our genome. Therefore, we used ANGSD v0.930 (51) on our previously generated bam-files and applied some basic filters (-uniqueOnly 1 -remove_bads 1 -only_proper_pairs 0 -trim 0 -C 50 -baq 1 -minMapQ 20), followed by calculation of the site-frequency spectra (SFS) (-doCounts 1 -GL 1 -doSaf 1) and Watterson's theta estimator θ_W in sliding windows of 50,000 bp with a step size of 10,000 bp. The effective population size was calculated after the formula $N_e = \theta_W / 4*\mu$ for a diploid organism. The mutation rate μ for the calculation was taken from the *Zea mays* literature (52) as a genome-wide average of 3.0×10^{-8}.

VCFtools v0.1.15 (41) was used to calculate the coverage (--depth) of the SNP markers and the observed homozygosity O(HOM) (--het). Using the number of sites N_SITES, the proportion of observed heterozygous sites can be calculated according to the formula (N_SITES - O(HOM)) / N_SITES.

ALS and *ACCase* amplicon analysis

ALS and *ACCase* PacBio amplicon sequencing

To generate *ALS* and *ACCase* amplicons for long-read PacBio sequencing we used the same DNA from the European collection described in a preceding section. Before PCR amplification DNA was normalized to 10 ng/µl. Then, 30 ng (*ALS*) and 50 ng (*ACCase*) total input DNA was used for the PCR Master Mix reaction (1 µl P5 indexing primer (5 µM), 1 µl P7 indexing primer (5 µM), 4 µl of 5x Prime STAR buffer, 1.6 µl dNTPs, 0.4 µl Prime STAR polymerase (Takara, R050B), filled up to 20 µl with water). The indexing PCR program for *ALS* was a 2-step PCR with 10 seconds of denaturation at 98°C and 210 seconds of annealing and extension at 68°C for 28 cycles, followed by a final

extension for 10 min at 72°C. For *ACCase*, the annealing and extension step was elongated to 660 seconds. Amplicons were then pooled equally per gene and bead cleaned. In the case of the 13.2 kb amplicon from *ACCase*, we added a BluePippin (SageScience) size selection to remove any remaining fragments below 10 kb. PacBio libraries were created according to the following PacBio amplicon protocol (part number 101-791-800 version 02 (April 2020)) and SMRT cells were loaded on a PacBio Sequel I system with Binding Kit and Internal Ctrl Kit 3.0 (part number 101-461-600 version 10; October 2019). An extended hands-on protocol can be found at https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022.

PacBio amplicon analysis

Most steps were carried out with tools developed by PacBio (https://github.com/PacificBiosciences/pbbioconda). First, CCS reads were generated with ccs v6.0.0 (minimum predicted accuracy 0.99 or q20). Then, demultiplexing was carried out with lima v1.11.0 (with parameters '--ccs --different --peek-guess --guess 80 --min-ref-span 0.875 --min-scoring-regions 2 --min-length 13000 --max-input-length 14000' for *ACCase* while for *ALS* similar parameters were used except for '--min-length 3200 --max-input-length 4200'). Next, pbaa cluster (v1.0.0) was run with default parameters followed by a series of amplicon-specific filtering steps.

For *ACCase*, we required a minimum of 25 CCS reads per sample, and only "passed clusters" were further considered for analysis. Samples with either 0 or more than 2 clusters were discarded. In samples in which a single cluster was identified (i.e., homozygous individuals for this locus), both haplotypes were assigned the same cluster sequence. In samples in which two different clusters (haplotypes) were identified, the difference between their respective frequencies had to be ≤ 0.50, otherwise the sample was discarded.

In the case of *ALS*, PCR amplification had been uneven, as our primers preferentially amplified certain haplotypes in individuals heterozygous for this locus. We presumed this was due to various structural variations downstream of the gene between major haplotypes (Figure S7). Therefore, to be able to analyze haplotype diversity of this locus, we employed less strict filtering steps than for *ACCase*. For *ALS*, a minimum of 25 CCS

reads per sample were required, while both 'passed clusters' and originally 'failed clusters' (mostly due to low frequency) were re-evaluated. First, only samples with cluster diversity ≤ 0.40 and cluster quality ≥ 0.7 were kept. In samples in which a single cluster was identified (i.e., homozygous individuals for this locus), its frequency had to be ≥ 0.98 to then assign the same cluster sequence to both haplotypes. In samples in which two different clusters (haplotypes) were identified, the difference between their respective frequencies was allowed to be ≤ 0.85, otherwise the sample was discarded. In the few samples in which three or more different clusters (haplotypes) were identified, the sum of the frequencies of the two main clusters had to be ≥ 0.96, and their difference ≤ 0.85 to be considered for downstream analyses.

Haplotype networks, haplotype trees and haplotype PCA

To annotate the clusters generated with pbaa with TSR metadata information, the single cluster fasta files representing two alleles per individual were first converted to fastq files using 'Fasta_to_fastq' (https://github.com/ekg/fasta-to-fastq). The resulting fastq files were aligned to the *ACCase* reference using minimap2 v2.15-r913-dirty (53), followed by sorting and indexing of the output bam files with samtools v1.9 (39). Read groups were assigned with the Picard function 'AddOrReplaceReadGroups' (RGID=$SAMPLE RGLB=ccs RGPL=pacbio RGPU=unit1 RGSM=$SAMPLE) (http://broadinstitute.github.io/picard/), followed by variant calling using GATK v4.1.3.0 (40) with functions 'HaplotypeCaller' (-R $REF --min-pruning 0 -ERC GVCF) and 'GenotypeGVCFs' with default settings. Variant annotation in the resulting VCF was performed with SnpEff v4.3t (54). The VCF was loaded in R to extract the TSR information and annotate the haplotype networks, trees and PCA with custom R scripts.

For the multiple alignments per population, we first combined all respective individual fasta files of the pbaa clusters into a single fasta file and then aligned them using MAFFT v7.407 (--thread 20 --threadtb 10 --threadit 10 --reorder --maxiterate 1000 --retree 1 --genafpair) (28). We used PGDSpider v2.1.1.5 (55) to transfer the multiple alignment fasta file into a Nexus-formatted file. Minimum spanning networks were inferred and visualized with POPART v.1.7 (56). Per population haplotype trees were inferred with RAXML-NG v0.9.0 (43) from the multiple sequence alignment files. 'Tree search' was performed with 20 distinct starting trees and bootstrapping analysis with the model

GTR+G and 10,000 bootstrap replicates. Tree visualization was done in R with ggtree v1.16.6 (57). The packages treeio v1.8.2 (58) and tibble v3.0.4 (https://github.com/tidyverse/tibble/) were used to add the TSR metadata information to the tree object. The branch length and node support values were extracted from Felsenstein's bootstrap proportions (FBP) output files. The haplotype PCAs were performed using the R package SNPrelate (45) on the previously generated VCFs for ALS and ACCase and visualized using ggplot2 (59).

Identification of *ALS* copies

Using the *ALS* GenBank sequence of *A. myosuroides* AJ437300.2 (60) as a query, BLASTN v2.2.29+ (61) retrieved three hits in chromosome 1 of our assembly. These loci corresponded to three gene models annotated as the largest subunit of ALS: model.Chr1.12329 (identity = 1921/1923 bp; 99.8%; hereafter *ALS1*), model.Chr1.11275 (identity = 1820/1915 bp; 95.0%; hereafter *ALS2*) and model.Chr1.11288 (identity = 1818/1915; 94.9%; hereafter *ALS3*).

To better characterize the relationship between these putative copies of the *ALS* gene, we analyzed synonymous substitution rates (K_S) and Iso-seq full-transcripts. K_S values between paralogs *ALS1-ALS2* and paralogs *ALS1-ALS3* were 0.153 and 0.165, respectively, while between paralogs *ALS2-ALS3* was 0.028. Although all K_S values between these paralogs were below 0.5, they are not present in our list of anchor pairs from the comparative genomics analysis (Dataset S1) for not being located among the collinear regions identified by i-ADHoRe (30).

For the analysis of Iso-seq data, we first generated very high-quality reads, with a minimum predicted accuracy 0.999 or q30, per tissue up until the poly-A trimming and concatemer removal step with isoseq3 v3.4.0 as described before for genome annotation. Next, we combined the q30 Iso-seq transcripts from all tissues and extracted only those that matched the following internal *ALS* sequences conserved among the three loci: 'CGCGCTACCTGCCCGCCTC', 'GTCTCCGCGCTCGCCGATGCT, 'GTCCAAGATTGTGCACAT' and 'GAGTGAAGTCCGTGCAGCAATC'. We obtained 343 Iso-seq q30 full-length transcripts, and it is worth mentioning that different internal *ALS* sequences yield near identical numbers of transcripts. Since Iso-seq q30 reads have

heterogeneous lengths, we used cutadapt v2.4 (37) to trim all reads at the 5' and 3' borders (-a CTTATTAATCA -g CCACAGCCGTCGC) of the CDS to make them all the same length. Finally, clustering with pbaa v1.0.0 (--min-read-qv 30) resulted in only three clusters with 143 reads corresponding to *ALS1*, 100 reads to *ALS2* and 100 reads to *ALS3*. Representative full-length Iso-seq reads with average read quality of q93 from each cluster were used for Figure S7. Therefore, all *ALS* gene models can produce full-length transcripts. Taking together K_S values and Iso-seq data, we could only conclude that *ALS1* is clearly distinct from *ALS2* and *ALS3*, but we could not distinguish whether *ALS2* and *ALS3* are two distinct loci or two alleles of the same locus.

Model simulations

Using equations 8, 11, 14, 18 and 20 from Hermisson and Pennings (62), we first modeled the general probability of adaptation through a sweep and then specifically from standing genetic variation. We set the population size to 42,000 individuals, which is the highest possible N_e from the populations characterized with RAD-Seq data. Since diversity estimates of N_e integrate over a long period of time and past bottlenecks will reduce it, leading to estimates that are lower than the actual N_e before the bottlenecks (63), we additionally simulated the doubled effective population size of 84,000 individuals. As maize is a diploid grass with a similar genome size to *A. myosuroides*, we adopted the mutation rate 3.0×10^{-8} (52). Both target site resistance genes in our study contain seven well described SNP positions that cause resistance (2, 3, 64, 65), therefore we set the mutational target size to seven. Before selection, we assumed three different selection coefficients for those mutations: 0, 1e-04, 0.001. Under selection, those TSR positions were beneficial in a range from 0 to 1 (Figure 4A,B, x-axes). The number of generations of selection was set to 30.

Standing genetic variation model vs. *de novo* model

Forward simulations were executed on a computing cluster with SLiM v3.4 (66) using SLiMGui v3.4 for model development. We used the *ACCase* locus (12,250 bp) as a template for all our simulations. Since we sequenced 585 bp upstream and 364 bp downstream of the gene, we defined the length of our simulated genomic element as 13,199 bp with TSR mutations at the following positions: 11052 (Ile1781), 11706

(Trp1999), 11790 (Trp2027), 11832 (Ile2041), 11943 (Asp2078), 11973 (Cys2088), 11997 (Gly2096). We further defined three genomic element types: exon, intron and non-coding region. For introns and non-coding regions, all mutations were considered to be neutral. In exons, a ratio of 0.25/0.75 (neutral/deleterious) mutations was used according to Messer and Petrov (67), with selection coefficients (s) for deleterious mutations drawn from a gamma distribution with E[s] = -0.000154 and a shape parameter of 0.245 (68). Since *A. myosuroides* is an annual grass, all models were built as Wright-Fisher models with non-overlapping generations and standard Wright-Fisher model assumptions (http://benhaller.com/slim/SLiM_Manual.pdf, p.35/36). As described above, we set the population size to 42,000 and 84,000 individuals. Both the mutation rate (3.0×10^{-8}) (52) and genome-wide average recombination rate (7.4×10^{-9}) (69) were adopted from maize. We implemented a burn-in period of $10 \times N_e$ generations to generate the initial genetic diversity and, since this is a computationally intensive process, we scaled our models down by a factor of 5.

We ran the model in one thousand independent runs per population size (42,000 and 84,000 individuals), and with (Figure 5) or without exons and introns, in which case all mutations were considered to be neutral (Figure S10), until generation $10 \times N_e$. After this generation, we applied herbicide selection for which mutations at the specified TSR positions became highly beneficial and dominant, with a selection coefficient s_i of 1.0 and a dominance coefficient h_i of 1.0 (fitness model for TSR individuals, homozygous: $1 + s_i = 1 + 1 = 2$, and heterozygous: $1 + h_i * s_i = 1 + 1 * 1 = 2$) (Figure S9). In practice, an herbicide is usually applied in the field once or twice each year. Since in *A. myosuroides* one generation time corresponds to about one year, we simulated one selection event per generation. Foster *et al.* 1993 (70)) specifically reported an ACCase inhibiting herbicide efficiency rate of 95-97%. However, it is likely that some *A. myosuroides* plants without TSR mutations will later emerge and thus escape the lethal effect of herbicide treatment contributing to the genetic diversity in the field. Therefore, in our simulations, we assume a remaining fitness of 10% for individuals that do not carry a TSR mutation to account for plants that escaped herbicide treatment or germinated at a later time point (fitness model for individuals without TSR, homozygous: $1 + s_i = 1 + (-0.9) = 0.1$). The selection pressure was applied at the end of every generation for a total of 30 generations. Only survivor individuals could reproduce and contribute to the next generation.

Generation 10 x N_e was a checkpoint for the presence of TSR mutations. If at least one individual in the total population was carrying at least one of the TSR mutations in a heterozygous state, the run belonged to the standing genetic variation scenario. If the first TSR mutation emerged only after herbicide selection, the run belonged to the *de novo* scenario. TSR allele frequencies and proportion of resistant individuals for 7 different time points (before selection, 5, 10, 15, 20, 25 and 30 generations after start of selection) were written to a log file and plotted with ggplot2 (59).

TSR occurrence

Furthermore, we examined how often a TSR mutation occurs on our simulated *ACCase* locus and how long it remains in the population before either being lost due to genetic drift or increase in frequency toward fixation under neutral conditions. This allows us to quantify how often resistance mutations are present as standing genetic variation in a field population before herbicide selection starts. To this end, we used a modified version of the model described in the previous section, this time without preexisting mutations, and ran it for 1,000 generations under neutrality. The other general parameters stayed the same as described above: 42,000 and 84,000 individuals, maize mutation rate 3.0 x 10^{-8} [52], maize recombination rate 7.4 x 10^{-9} [69]. Mutations were modeled using the described intron/exon gene model for the *ACCase* locus. After each generation, we output the number of TSR mutations at the predetermined TSR positions in the population. We performed 100 independent simulation runs per N_e. Detailed scripts for all simulations can be found at https://github.com/SonjaKersten/Herbicide_resistance_evolution_in_blackgrass_2022.

Data manipulation and plotting

The visualization of our data was done with R version v3.6.1 (71) and RStudio v1.1.453 (http://www.rstudio.com). All R packages and versions used for general data manipulation and visualization can be found in Table S2.

Supporting Information Figures

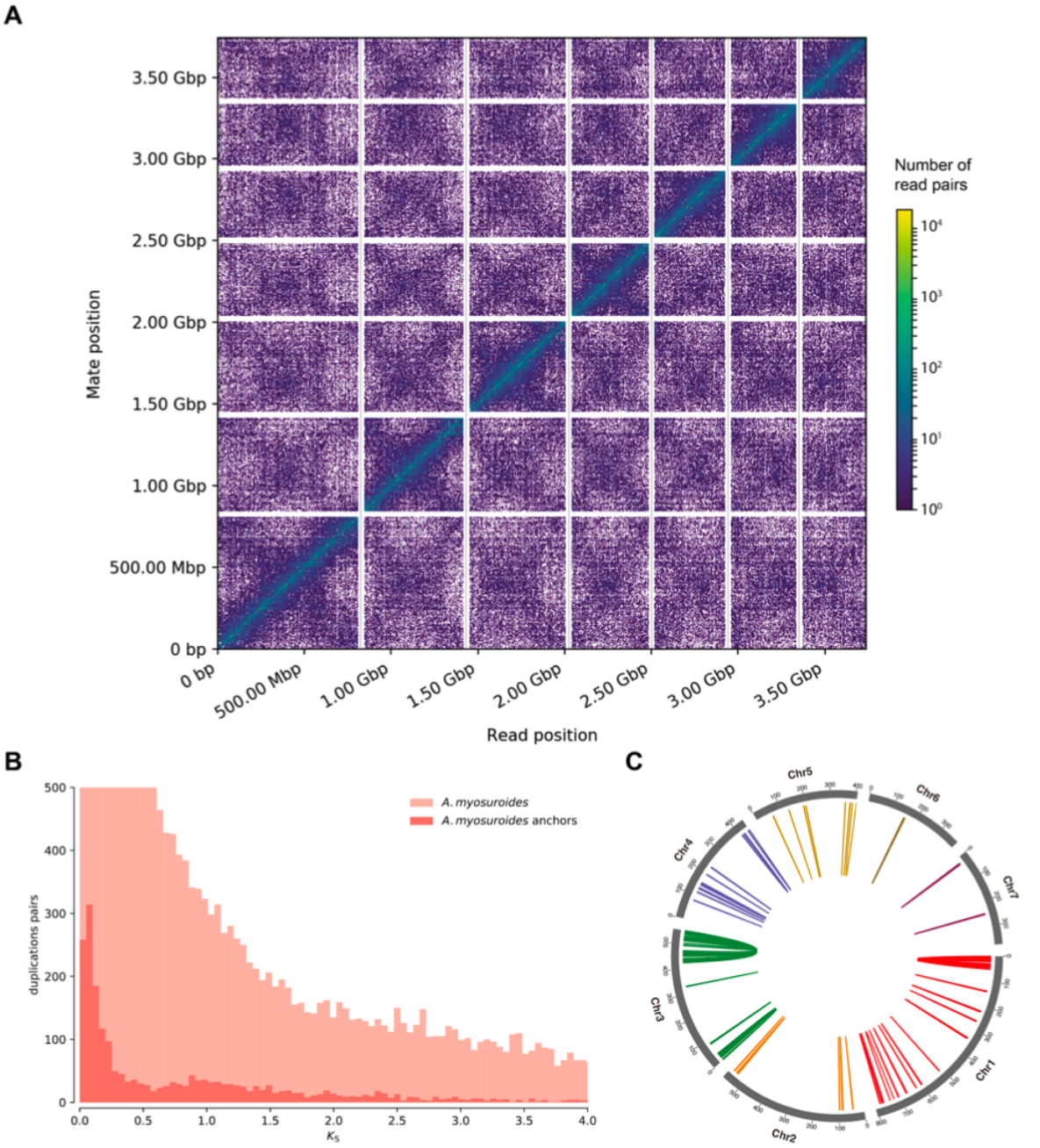


Figure S1. Genome scaffolding and analysis of anchors. **A,** Link density plot of the mapping positions of the first (x-axis) and second read (y-read) in the read pair, grouped into bins. The color of each square indicates the number of read pairs in that bin. Scaffolds < 1 Mb are excluded. **B,** K_S distributions for all paralogs within the *A. myosuroides* (light color) and for the paralogs retained in collinear regions, also known as anchors (dark color). **C,** Circos plot of the *A. myosuroides* genome, with colored lines connecting anchor pairs (genes in the collinear regions) with K_S < 0.5 (Dataset S1).

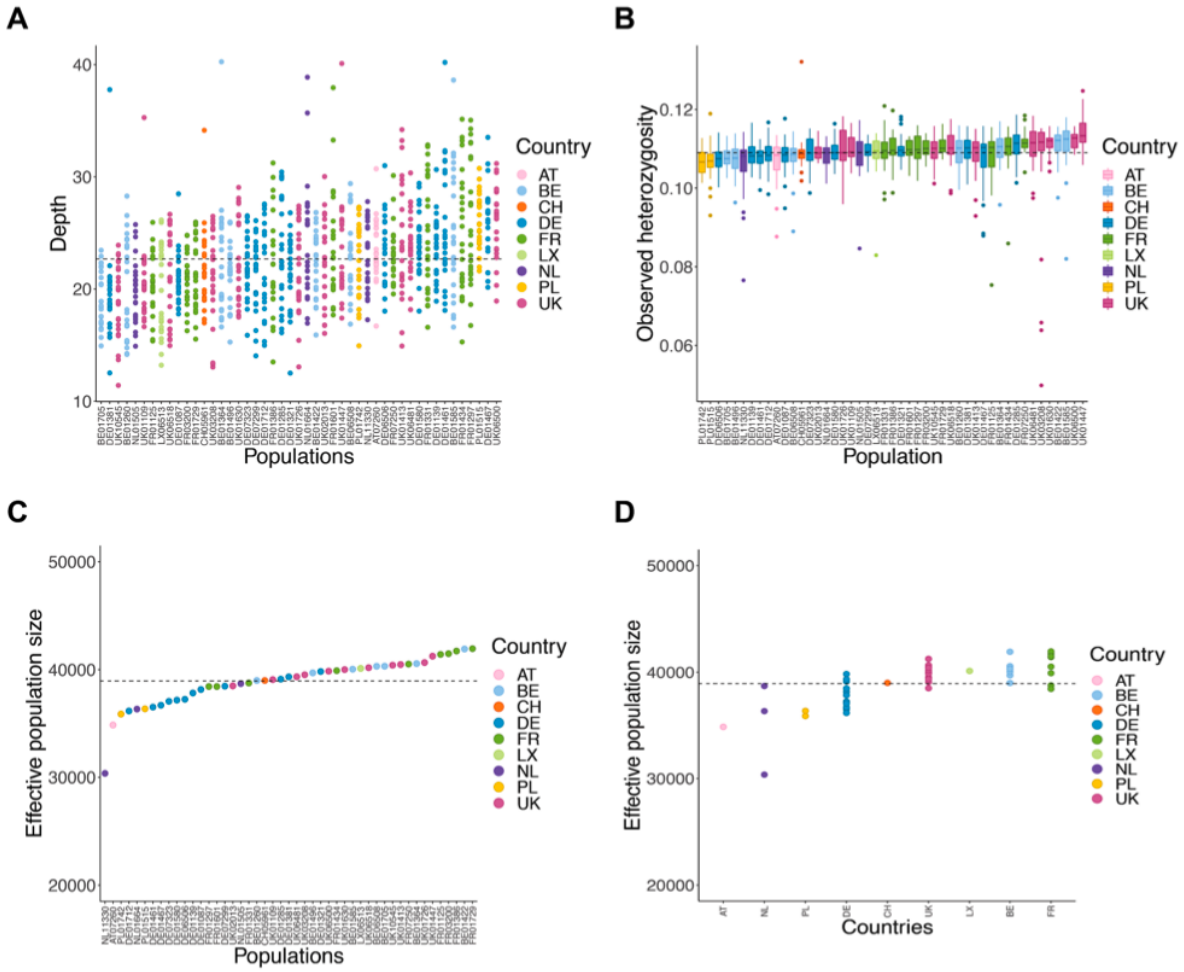


Figure S2. Basic statistics of the ddRAD-Seq dataset and diversity metrics. Colors reflect country-specific origin of the populations. **A,** Sequencing depth. **B,** Observed SNP heterozygosity. **C,** Effective population sizes. Mean= 38,912 individuals (dashed line) **D,** Effective population sizes ordered by countries. Tukey's HSD test showed a significant difference between the mean effective population size of DE and UK ($p<0.03$), DE and BE ($p<0.03$), DE and FR ($p<0.01$). Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK).

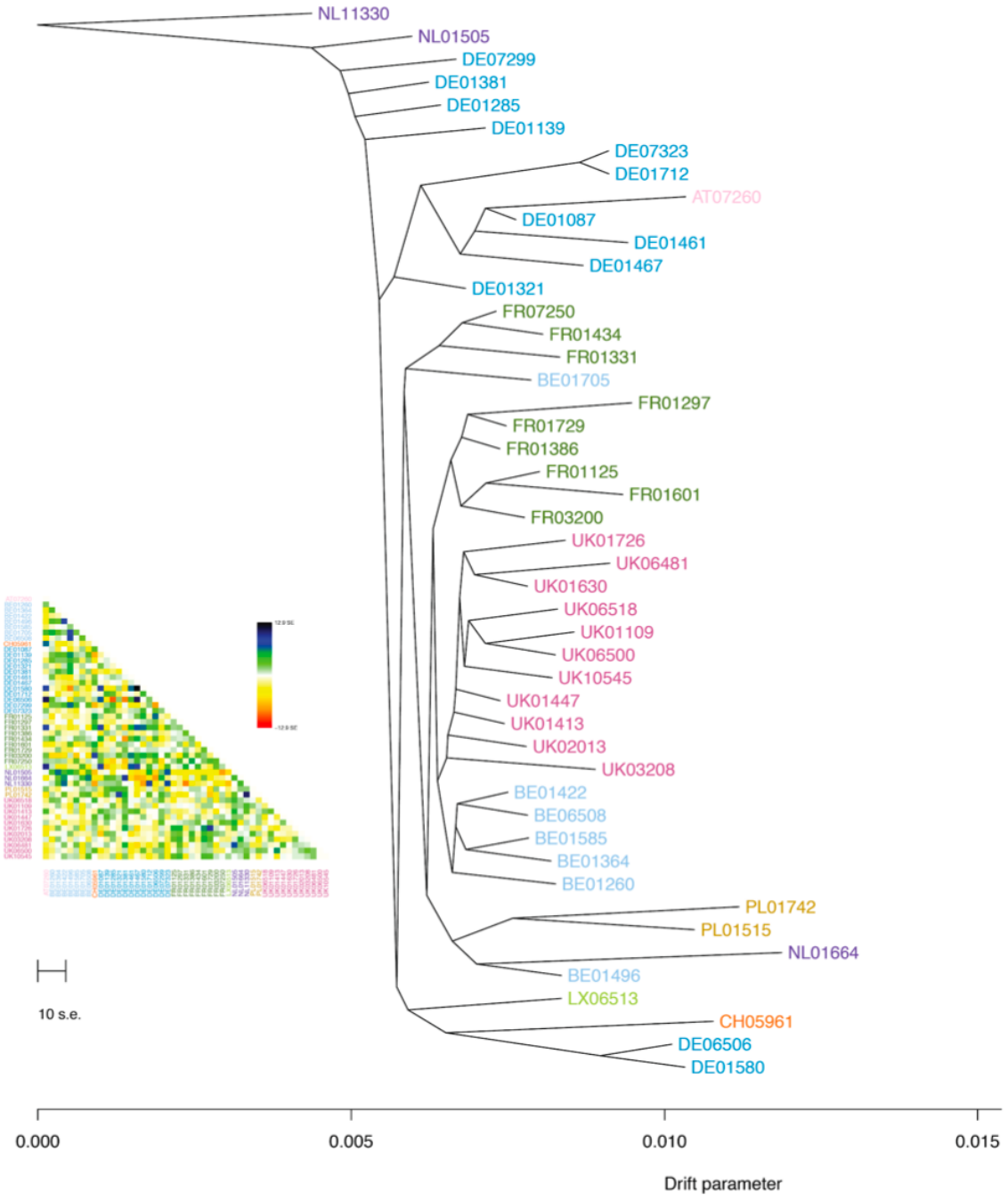


Figure S3. Treemix plot of the relationship of populations with residuals. Colors reflect the country-specific origin of the populations. Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK).

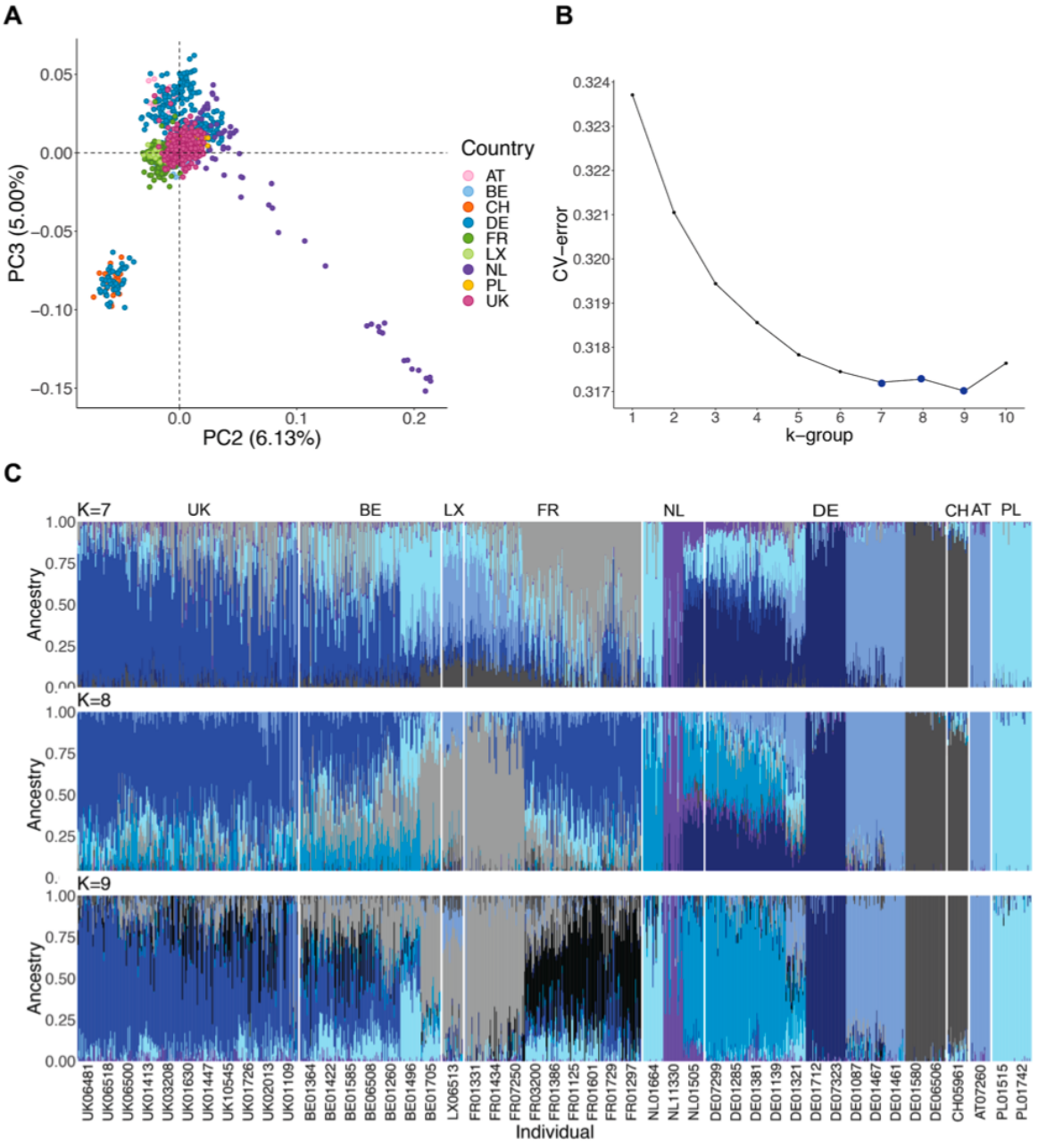


Figure S4. Population structure analysis. **A,** Second and third eigenvectors of the principal component analysis (PCA). The genetic variance of the second and third component is shown in brackets. Colors reflect country-specific origin of the populations. **B,** Cross validation error as a function of K of the admixture analysis. **C**, Admixture proportions with ancestry groups of K=9, K=7 and K=8. Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK).

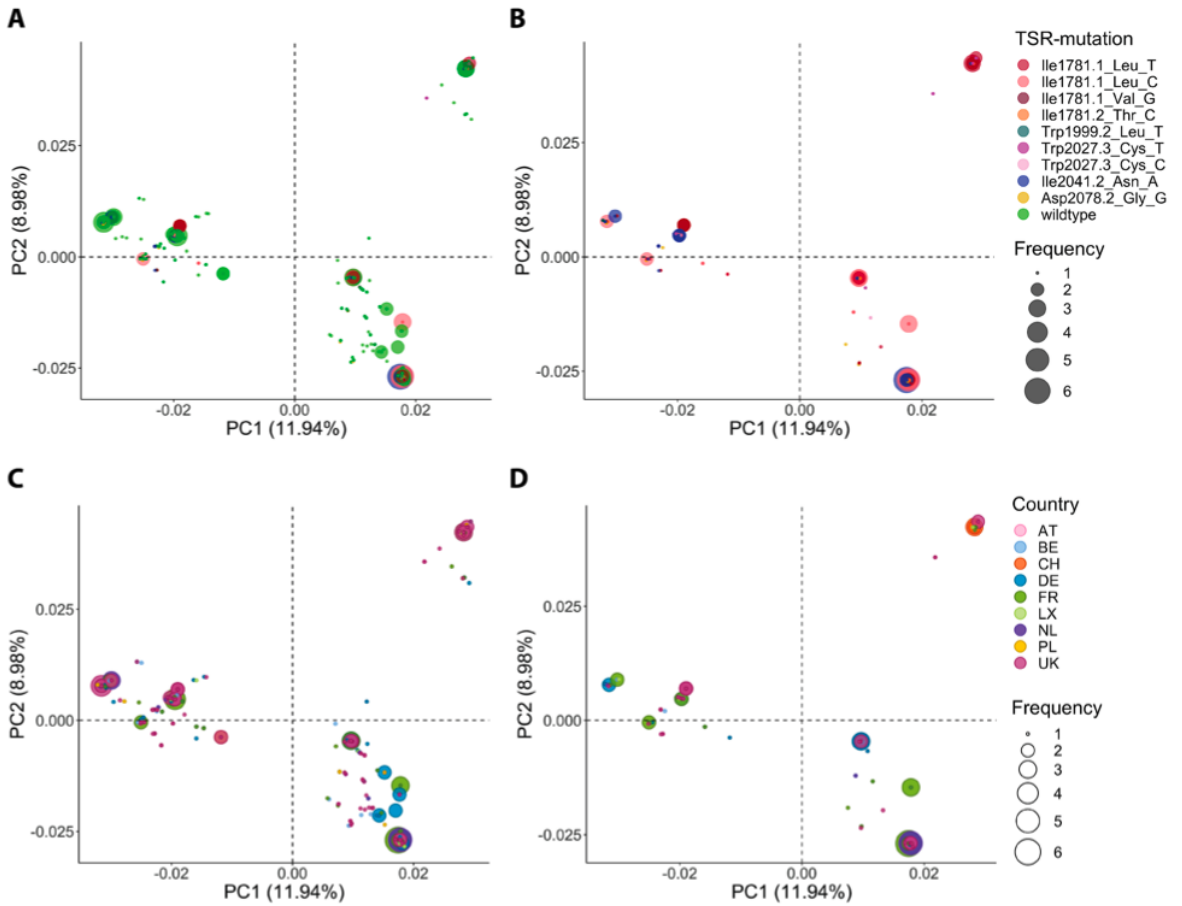


Figure S5. ACCase haplotype principal component analysis (PCA). Eigenvectors of the first two components are shown. **A,** Target-site-resistance (TSR) annotation of all existing haplotypes including wildtype haplotypes. **B,** Only TSR haplotypes. **C,** Country-specific coloring of all existing haplotypes. **D,** Country-specific coloring of exclusively TSR haplotypes. The values in brackets show the explained variance. Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK).

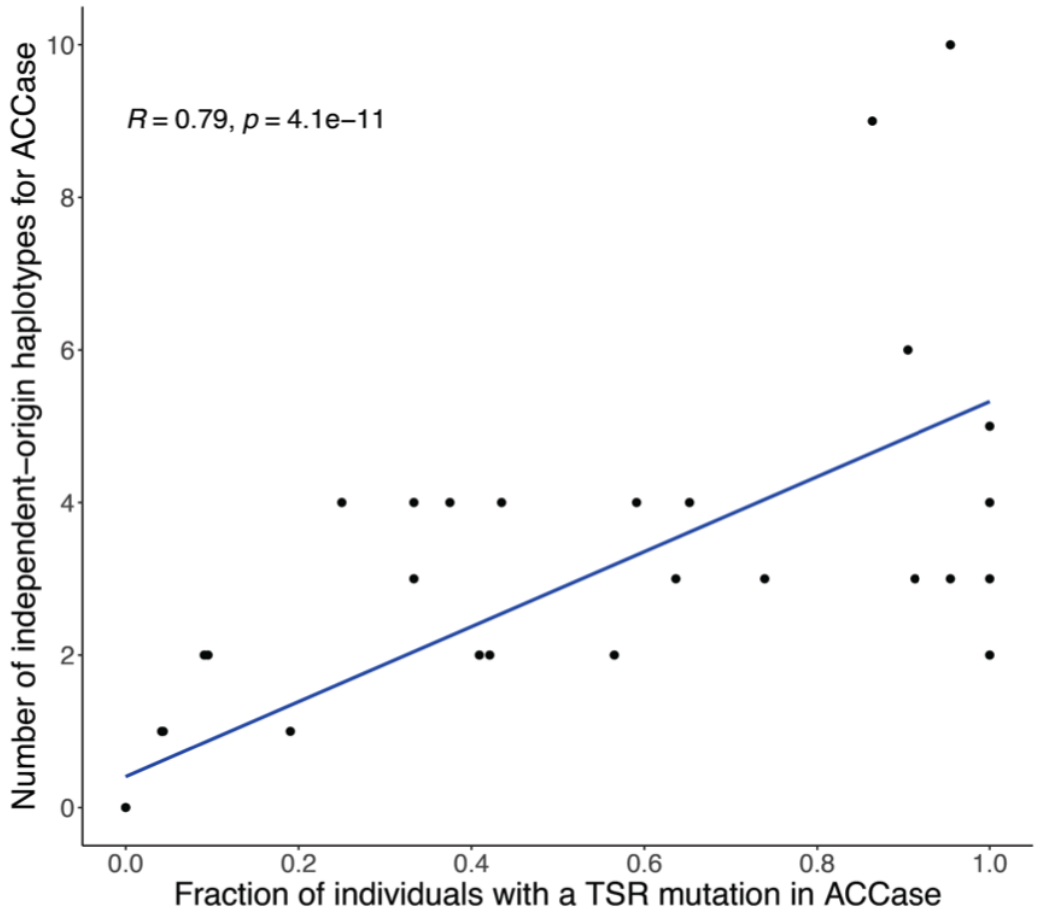


Figure S6. Correlation between the fraction of individuals with TSR mutations and the number of TSR haplotypes for the *ACCase* gene. Every dot represents a population.

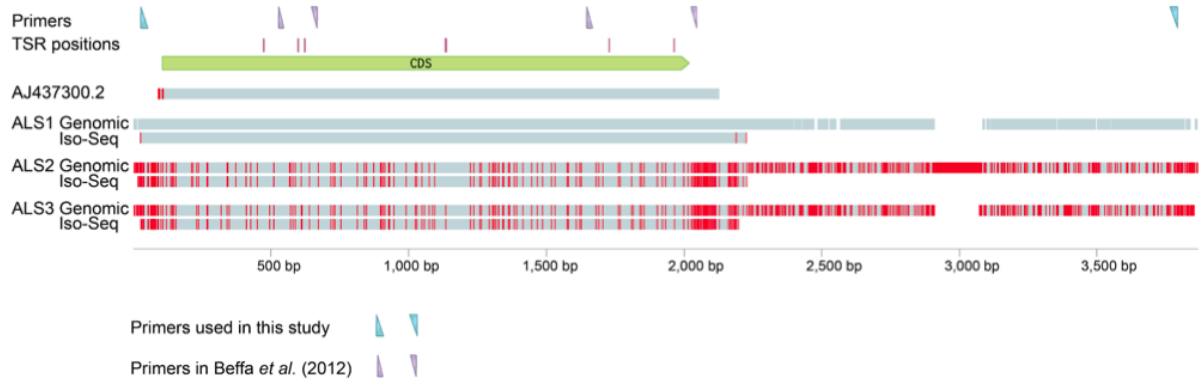


Figure S7. *ALS* copies in *A. myosuroides* genome. Multiple alignment performed with Clustal Omega (72) between the widely studied *ALS* GenBank entry of *A. myosuroides* AJ437300.2 (60), three genomic loci encoding *ALS* genes, and three representative Iso-Seq reads (each with an average read quality of q93) corresponding to each of the three Iso-Seq clusters determined by pbaa (https://github.com/PacificBiosciences/pbAA) with data from all five tissues combined. Indicated are also the positions of the seven known TSR mutations in *ALS*, the primers used in this study to selectively amplify *ALS1*, and the two pairs of primers commonly used to genotype TSRs Pro197 and Ala205 (first pair), and Trp574 and Ser653 (second pair) (73).

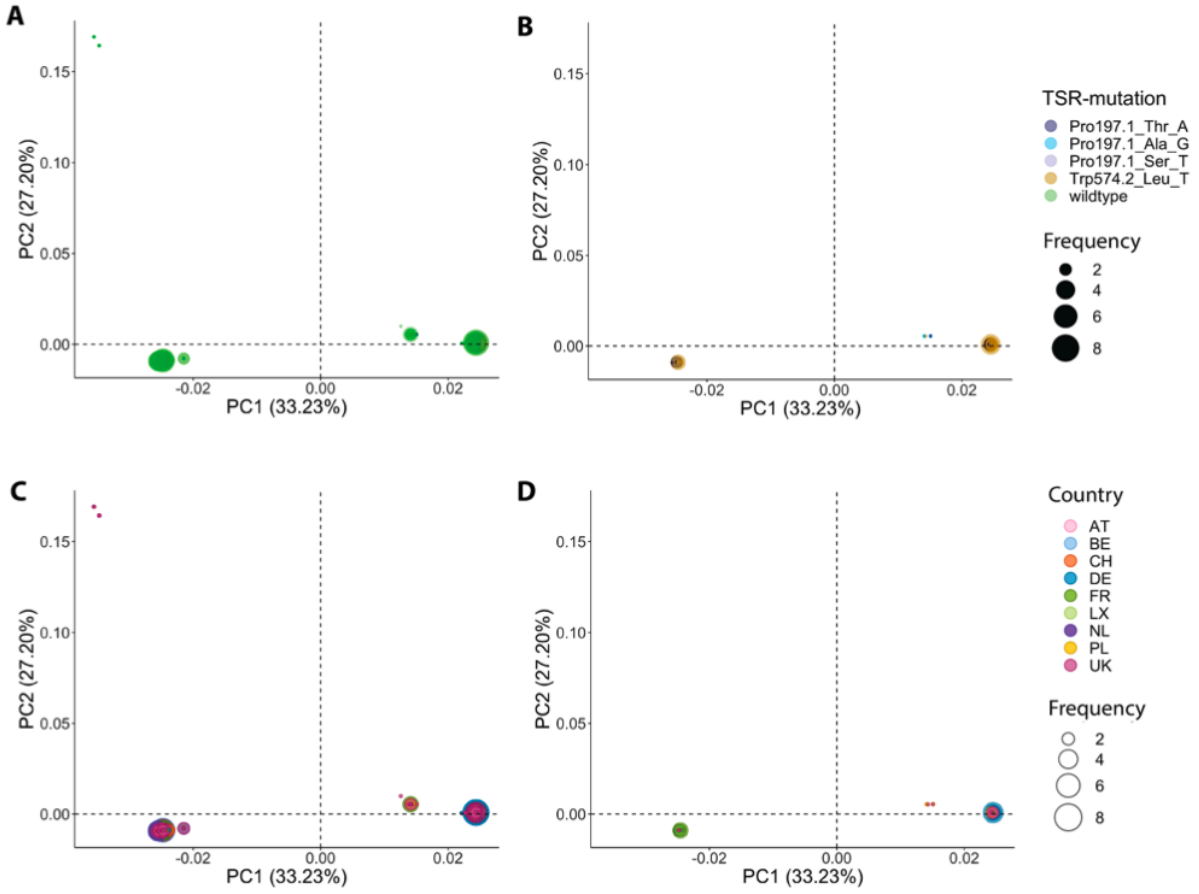


Figure S8. *ALS* haplotype principal component analysis (PCA). Eigenvectors of the first two components are shown. **A,** Target-site-resistance (TSR) annotation of all existing haplotypes including wildtype haplotypes. **B,** Only TSR haplotypes. **C,** Country-specific coloring of all existing haplotypes. **D,** Country-specific coloring of exclusively TSR haplotypes. The values in brackets show the explained variance. Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK).

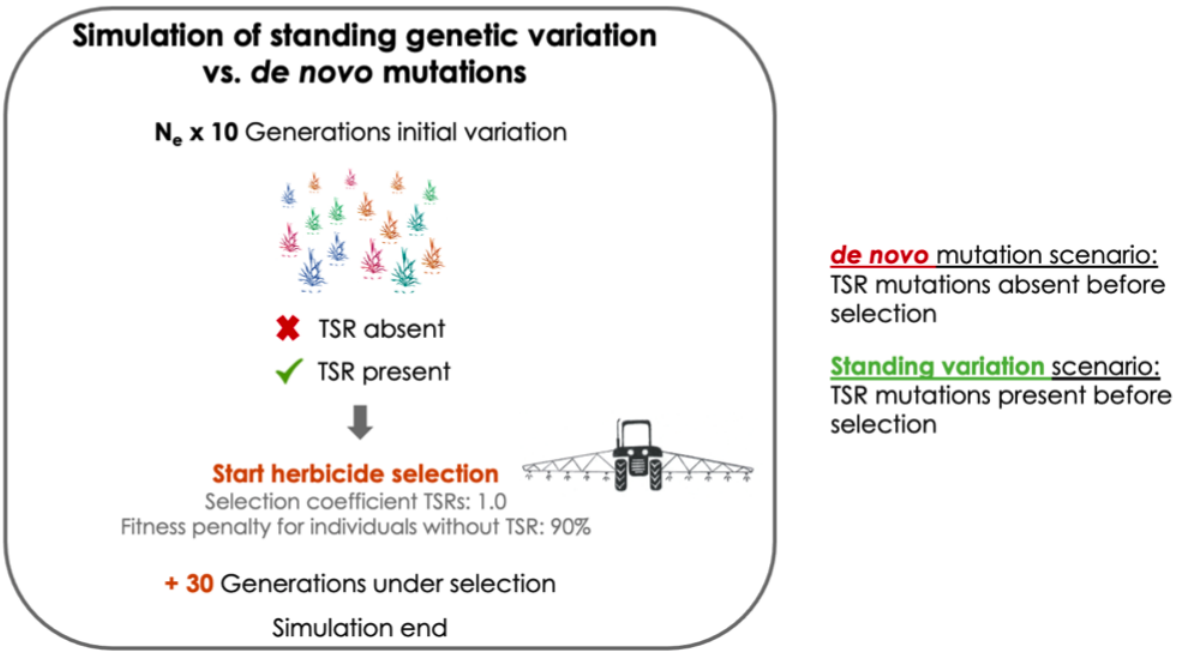


Figure S9. Simulation of herbicide resistance evolution. Visualization of the SLiM simulation model and the two scenarios for the origin of TSR mutation: TSR mutations emerging from standing genetic variation if they were present before the start of herbicide selection, or from *de novo* mutation if they appeared after herbicide selection. The model was run using the intron/exon structure of the *ACCase* locus as a template and without it. For the model with introns/exons, mutations in introns and non-coding regions were considered to be neutral, while exons had a ratio of 0.25/0.75 (neutral/deleterious) mutations according to Messer and Petrov (67), with selection coefficients (s) for deleterious mutations drawn from a gamma distribution with E[s] = -0.000154 and a shape parameter of 0.245 (68). For the latter, all mutations were considered to be neutral. Furthermore, we simulated two N_e values: 42,000 and 84,000 individuals.

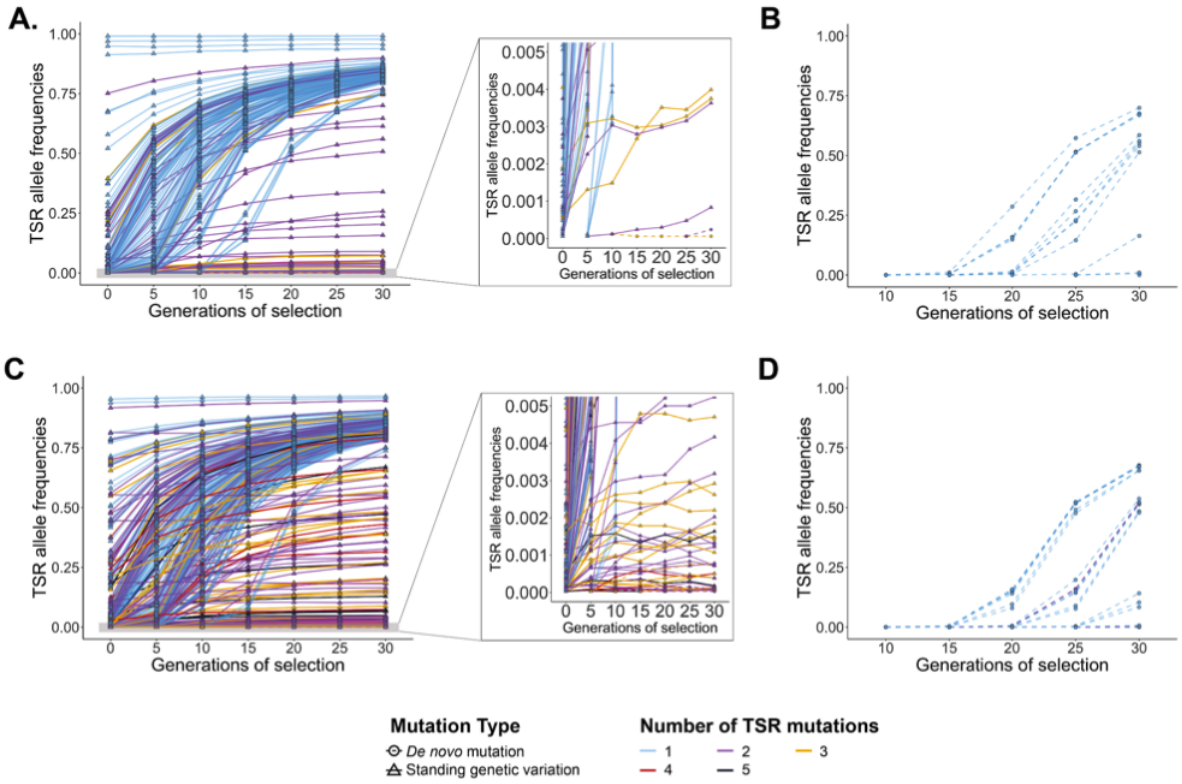


Figure S10. Simulations of expected allele frequencies for TSR alleles arising from standing genetic variation or *de novo* mutation (without intron/exon structure). All mutations were considered to be neutral before the start of selection. Five hundred of one thousand simulation runs are shown for an effective population size (N_e) of (**A, B**) 42,000 individuals and (**C, D**) 84,000 individuals. Continuous lines represent mutations originating from standing genetic variation; *de novo* TSR mutations are shown with dashed lines. Colors indicate the total number of TSR mutations per population. **A, C,** Standing genetic variation scenario, with TSR mutations pre-existing in the populations before herbicide selection. Shown is the increase in TSR allele frequencies under herbicide selection of up to 30 generations, with one herbicide application per generation. The right panel shows a truncated y-axis at 0.005 TSR allele frequencies. **B, D,** *De novo* mutation scenario. Any TSR mutation that might have arisen before the start of selection has been lost again, so that no TSR mutations are present at generation 0 of selection.

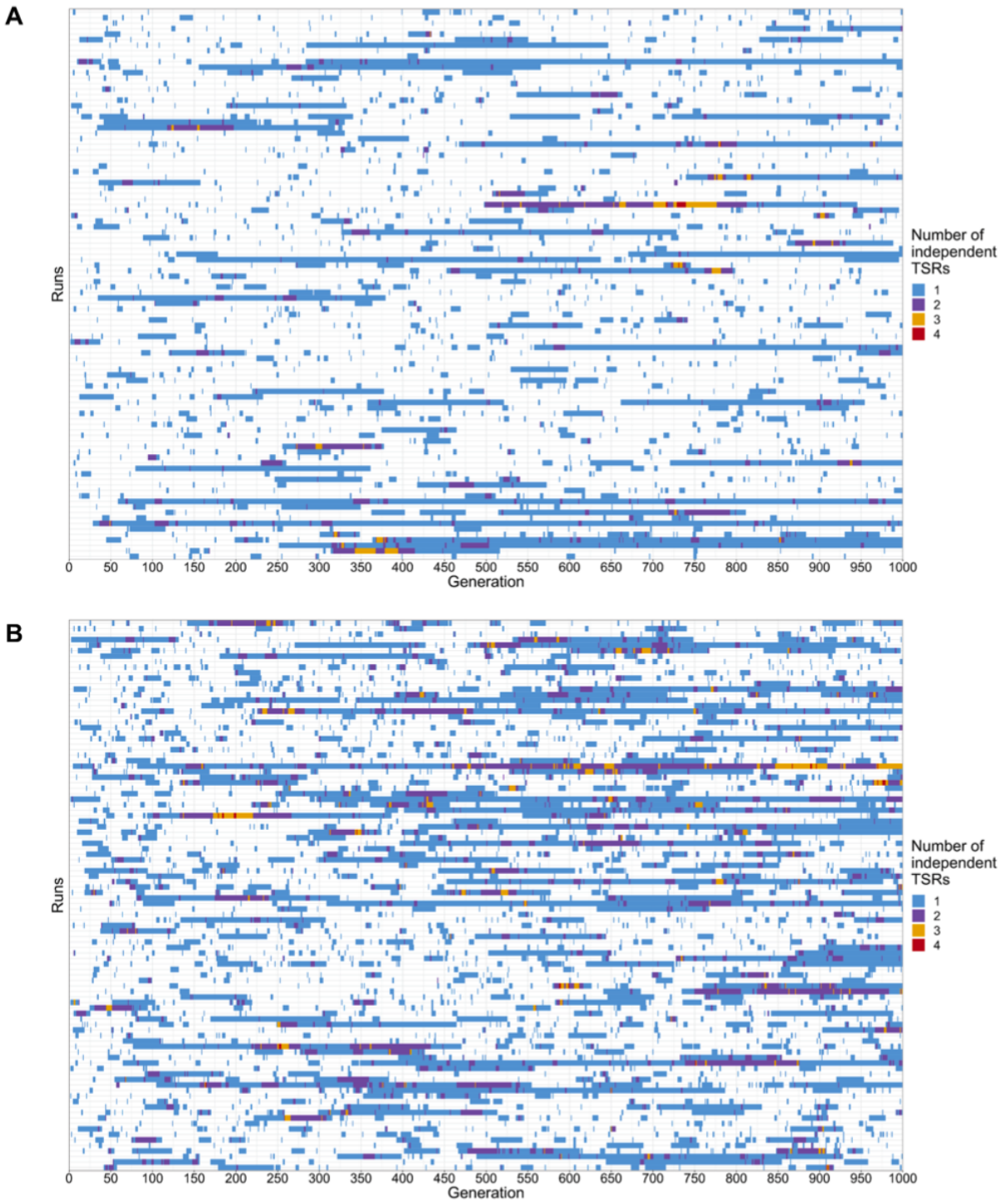


Figure S11. Simulations of TSR abundance under neutrality. SLiM simulations representing 100 independent simulation runs of population evolution for (**A**) 42,000 individuals and (**B**) 84,000 individuals over 1,000 generations under neutrality. The occurrence and loss of TSRs due to genetic drift can be observed. Colors indicate the number of TSRs present in each generation in a given run.

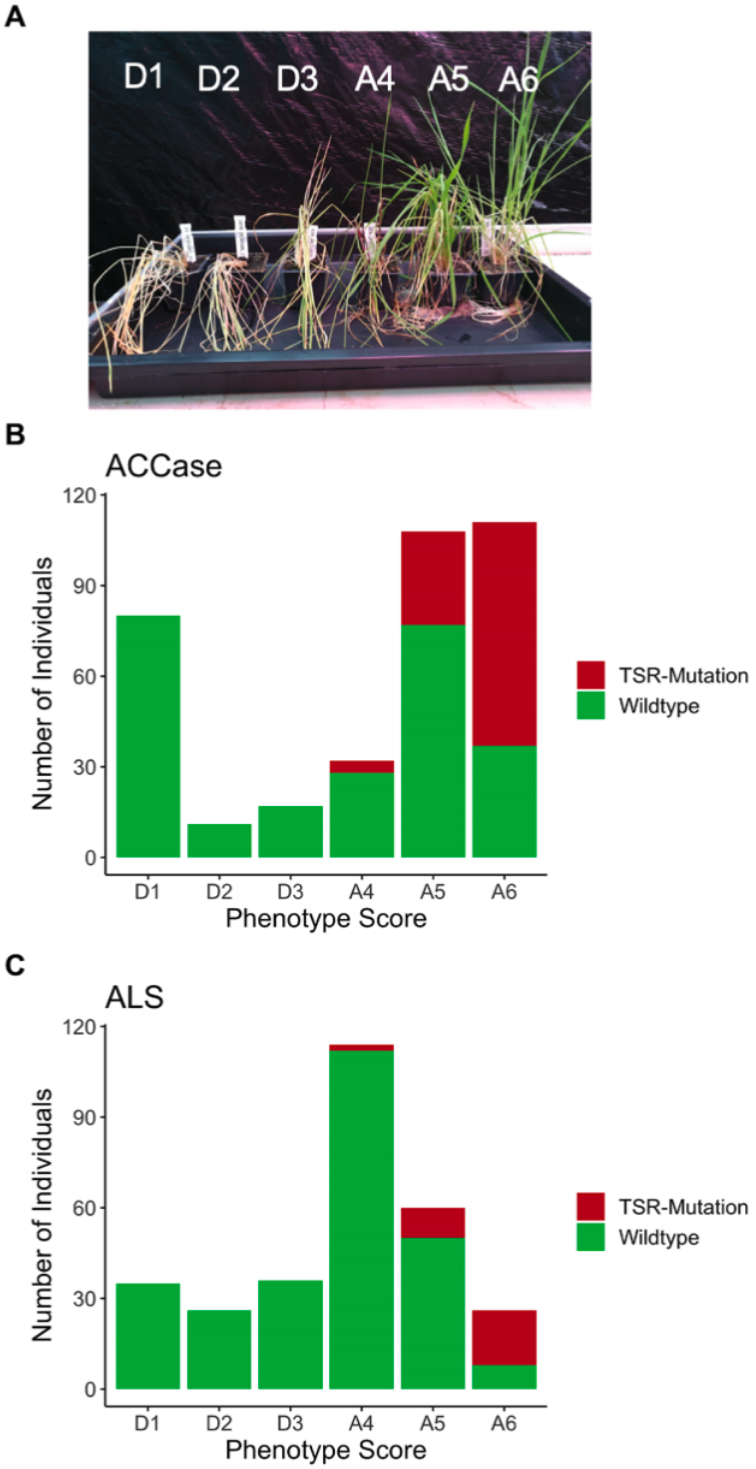


Figure S12. Phenotyping and genotyping of TSR mutations in single individuals. **A,** Phenotype scoring scheme, where the score D1 represents completely dead plants (no green material visible) and the score A6 represents plants without any growth reductions compared to the control plants of the respective population. **B,** Distribution of phenotype scores after treatment with ACCase inhibitor Axial® 50 (pinoxaden + cloquintocet-mexyl). In red, the number of individuals that carry a TSR mutation. In green, wildtype individuals. **C,** Distribution of phenotype scores after treatment with ALS inhibitor Atlantis WG® (mesosulfuron + iodosulfuron).

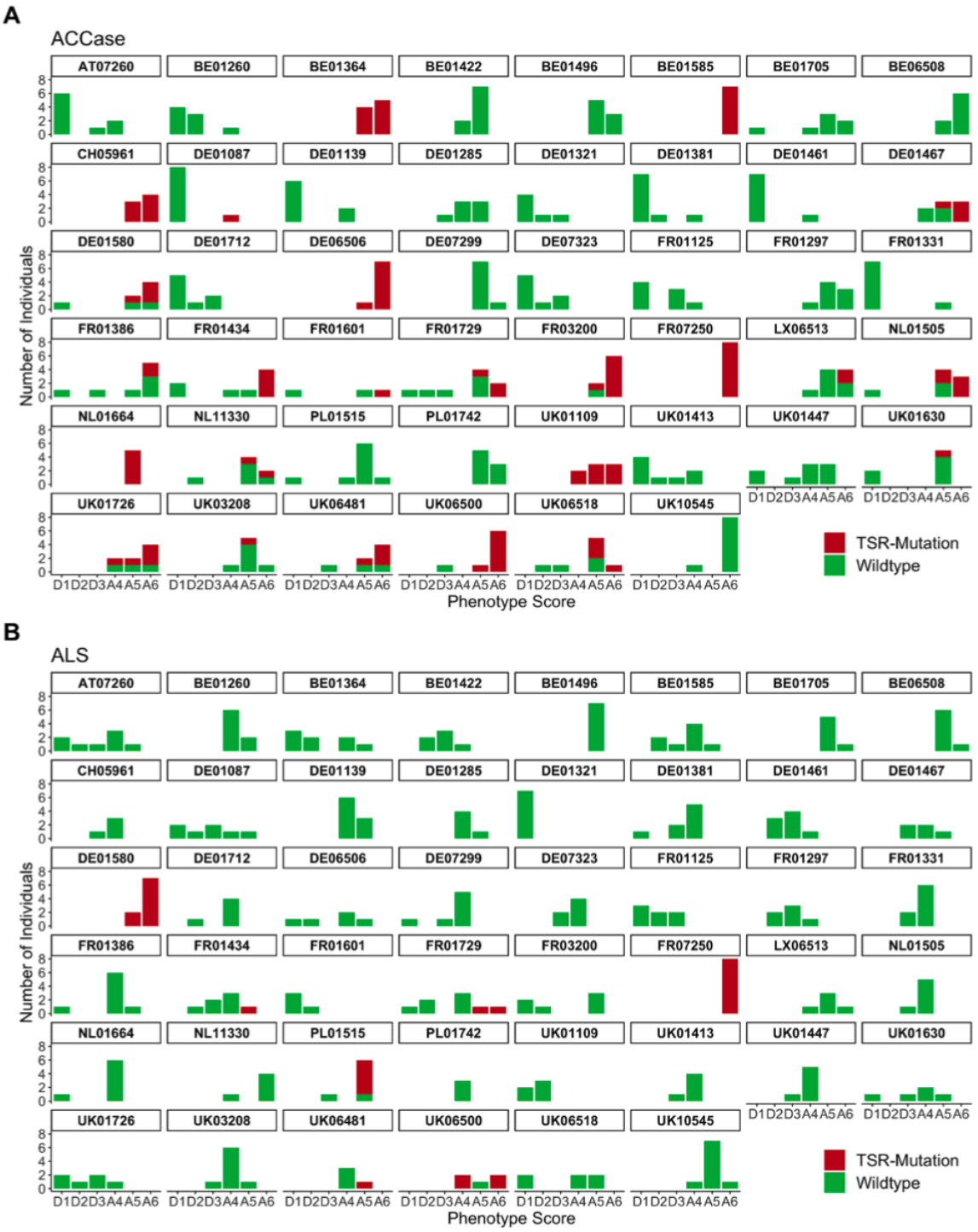


Figure S13. Phenotyping and genotyping of TSR mutations in single individuals per population. Phenotype scoring according to the scheme in Figure S12A, where the score D1 represents completely dead plants (no green material visible) and the score A6 represents plants without any growth reductions compared to the control plants of the respective population. **A,** Distribution of phenotype scores per population after treatment with ACCase inhibitor Axial® 50 (pinoxaden + cloquintocet-mexyl). In red, the number of individuals that carry a TSR mutation. In green, wildtype individuals. **B,** Distribution of phenotype scores after treatment with ALS inhibitor Atlantis WG® (mesosulfuron + iodosulfuron).

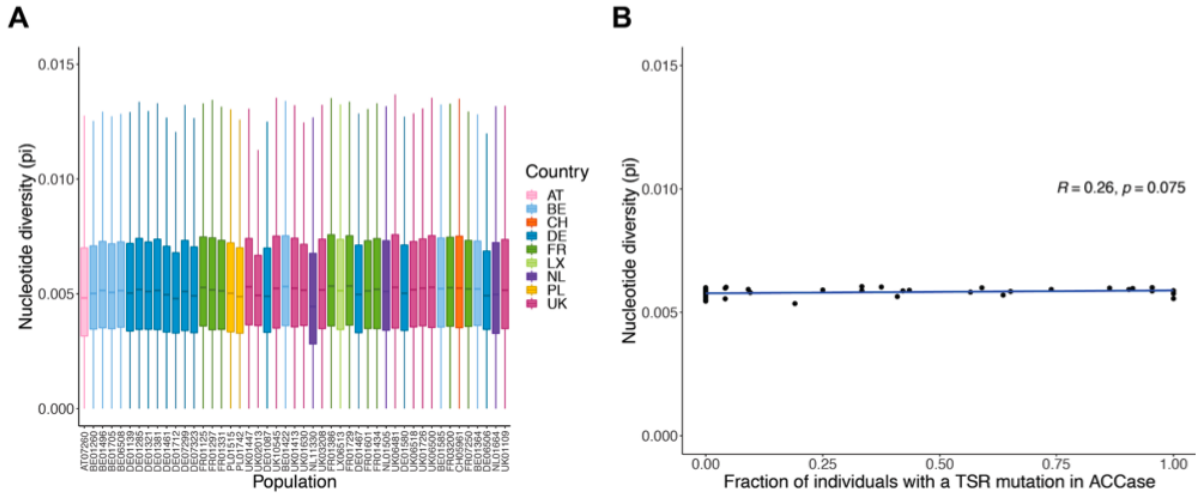


Figure S14. Relationship between nucleotide diversity and fraction of individuals with a TSR mutation in *ACCase* per population. **A,** Nucleotide diversity (pi) per population estimated from ddRAD-Seq data. Populations are sorted, in increasing order, according to the fraction of individuals with a TSR mutation in the *ACCase* gene. Colors reflect the country-specific origin of the populations. Austria (AT), Belgium (BE), Switzerland (CH), Germany (DE), France (FR), Luxembourg (LX), Netherlands (NL), Poland (PL), United Kingdom (UK). **B,** Correlation between the fraction of individuals with TSR mutations in the *ACCase* gene and nucleotide diversity (pi) per population. Every dot represents a population.

Supporting Information Tables

Table S1. TSR and TSR haplotype number per population.

Population	*ACCase*			*ALS*		
	TSR number	TSR haplotype number	Min. indep. haplotype origins	TSR number	TSR haplotypes	Min. indep. haplotype origins
AT07260	0	0	0	0	0	0
BE01260	0	0	0	0	0	0
BE01364	4	5	5	0	0	0
BE01422	1	1	1	0	0	0
BE01496	0	0	0	0	0	0
BE01585	5	7	6	0	0	0
BE01705	0	0	0	0	0	0
BE06508	0	0	0	0	0	0
CH05961	3	4	3	0	0	0
DE01087	1	1	1	0	0	0
DE01139	0	0	0	0	0	0
DE01285	0	0	0	0	0	0
DE01321	0	0	0	0	0	0
DE01381	0	0	0	0	0	0
DE01461	0	0	0	0	0	0
DE01467	2	2	2	0	0	0
DE01580	3	3	3	2	2	2
DE01712	0	0	0	0	0	0
DE06506	3	4	4	0	0	0
DE07299	0	0	0	0	0	0
DE07323	0	0	0	0	0	0
FR01125	0	0	0	0	0	0

FR01297	0	0	0	0	0	0
FR01331	0	0	0	0	0	0
FR01386	3	4	4	1	1	1
FR01434	3	4	4	2	2	2
FR01601	2	2	2	0	0	0
FR01729	3	4	4	1	2	2
FR03200	2	3	3	0	0	0
FR07250	6	10	10	2	3	3
LX06513	2	3	3	0	0	0
NL01505	2	2	2	0	0	0
NL01664	2	2	2	0	0	0
NL11330	1	1	1	0	0	0
PL01515	0	0	0	2	2	2
PL01742	0	0	0	0	0	0
UK01109	2	4	3	0	0	0
UK01413	2	2	2	0	0	0
UK01447	0	0	0	0	0	0
UK01630	2	2	2	0	0	0
UK01726	3	3	3	0	0	0
UK02013	0	0	0	0	0	0
UK03208	2	4	4	0	0	0
UK06481	3	5	4	1	1	1
UK06500	4	9	9	2	3	3
UK06518	3	5	4	0	0	0
UK10545	1	1	1	1	1	1

Table S2. R-packages used for data manipulation and visualization.

Package name and version	Reference
ComplexHeatmap 2.0.0	Gu, 2016 (50)
dplyr 1.0.2	Wickham, 2020 (74)
gdsfmt 1.20.0	Zheng, 2012 (45)
GetoptLong 1.0.4	Gu, 2020 (https://github.com/jokergoo/GetoptLong)
ggplot 3.3.2	Wickham, 2016 (59)
ggpubr 0.4.0	Kassambara, 2020 (https://github.com/kassambara/ggpubr/)
ggtree 1.16.6	Yu, 2017 (57)
ggthemes 4.2.0	Arnold, 2021 (https://github.com/jrnold/ggthemes/)
gtable 0.3.0	Wickham, 2019 (https://github.com/r-lib/gtable)
haplotypes 1.1.2	Aktas, 2020 (https://cran.r-project.org/web/packages/haplotypes/haplotypes.pdf)
patchwork 1.1.0	Pedersen, 2020 (https://github.com/thomasp85/patchwork)
plotly 4.9.2.1	Sievert, 2019 (75)
plyr 1.8.6	Wickham, 2011 (76)
qqman 0.1.4	Turner, 2021 (https://github.com/stephenturner/qqman)
SNPRelate 1.18.1	Zheng, 2012 (45)
stats 3.6.1	The R core team (71)
tibble 3.1.6	Müller,2021 (https://github.com/tidyverse/tibble/)
tidyr 1.1.2	Wickham, 2020 (https://github.com/tidyverse/tidyr/)
tidyverse 1.3.0	Wickham, 2019 (77)
treeio 1.18.1	Wang, 2020 (58)
vcfR 1.12.0	Knaus, 2017 (78)

Legends for Datasets

Dataset S1. Sheet1, List of paralogs retained in collinear regions (anchors), their K_S values, and whether they are part of Figure S1C. **Sheet2,** List of 250 non-redundant *ACCase* haplotypes. **Sheet3,** List of primers used in this study.

Dataset S2. *ACCase* networks and trees for 47 European populations. Haplotype network and maximum likelihood (ML)-tree per population. The color code in all networks and trees shows target-site resistances (TSRs) and wildtype haplotypes in green.

Dataset S3. *ALS* networks and trees for 47 European populations. Haplotype network and maximum likelihood (ML)-tree per population. The color code in all networks and trees shows target-site resistances (TSRs) and wildtype haplotypes in green.

Supporting Information References

1. S. Picelli, *et al.*, Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

2. C. Délye, X.-Q. Zhang, S. Michel, A. Matéjicek, S. B. Powles, Molecular bases for sensitivity to acetyl-coenzyme A carboxylase inhibitors in black-grass. *Plant Physiol.* **137**, 794–806 (2005).

3. H. Xu, *et al.*, Mutations at codon position 1999 of acetyl-CoA carboxylase confer resistance to ACCase-inhibiting herbicides in Japanese foxtail (Alopecurus japonicus). *Pest Manag. Sci.* **70**, 1894–1901 (2014).

4. J. Doležel, J. Greilhuber, S. Lucretti, A. Meister, Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Annals of Botany* **82**, 17–26 (1998).

5. J. Dolezel, J. Bartos, Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.* **95**, 99–110 (2005).

6. R. Workman, *et al.*, High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing v1 (protocols.io.4vbgw2n) https:/doi.org/10.17504/protocols.io.4vbgw2n.

7. H. Yaffe, *et al.*, LogSpin: a simple, economical and fast method for RNA isolation from infected or healthy plants and other eukaryotic tissues. *BMC Res. Notes* **5**, 45 (2012).

8. A. Acosta-Maspons, I. González-Lemes, A. A. Covarrubias, Improved protocol for isolation of high-quality total RNA from different organs of Phaseolus vulgaris L. *Biotechniques* **66**, 96–98 (2019).

9. E. Lieberman-Aiden, *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

10. C.-S. Chin, *et al.*, Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).

11. D. Guan, *et al.*, Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

12. N. H. Putnam, *et al.*, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).

13. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

14. S. Ou, *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

15. T. D. Wu, C. K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

16. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

17. S. Kovaka, *et al.*, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

18. C. Trapnell, *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

19. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

20. UniProt Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

21. T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, M. Borodovsky, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108 (2021).

22. M. Van Bel, *et al.*, PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* **46**, D1190–D1196 (2018).

23. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

24. B. J. Haas, *et al.*, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

25. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

26. P. Jones, *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

27. A. Zwaenepoel, Y. Van de Peer, wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155 (2019).

28. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

29. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

30. S. Proost, *et al.*, i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).

31. A. H. Paterson, J. E. Bowers, B. A. Chapman, Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9903–9908 (2004).

32. International Brachypodium Initiative, Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**, 763–768 (2010).

33. H. Tang, *et al.*, Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).

34. C. Chen, *et al.*, TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* **13**, 1194–1202 (2020).

35. P. L. M. Lang, *et al.*, Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Mol. Ecol. Resour.* **20**, 1228–1247 (2020).

36. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).

37. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

38. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

39. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

40. G. A. Van der Auwera, *et al.*, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).

41. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

42. J. B. Puritz, C. M. Hollenbeck, J. R. Gold, dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**, e431 (2014).

43. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

44. I. Letunic, P. Bork, Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–8 (2011).

45. X. Zheng, *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

46. C. C. Chang, *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

47. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).

48. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

49. N. C. Rochette, A. G. Rivera-Colón, J. M. Catchen, Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.* **28**, 4737–4754 (2019).

50. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

51. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

52. N. Yang, *et al.*, Contributions of Zea mays subspecies mexicana haplotypes to modern maize. *Nat. Commun.* **8**, 1874 (2017).

53. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

54. P. Cingolani, *et al.*, A program for annotating and predicting the effects of single nucleotide

72. F. Sievers, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

73. R. Beffa, *et al.*, Weed resistance diagnostic technologies to detect herbicide resistance in cereal-growing areas. A review. *Julius-Kühn-Archiv* **434**, 75–80 (2012).

74. H. Wickham, R. François, L. Henry, K. Müller, A Grammar of Data Manipulation [R package dplyr version 1.0.2] (2020) (November 17, 2021).

75. C. Sievert, Interactive web-based data visualization with R, plotly, and shiny (2020).

76. H. Wickham, The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40** (2011).

77. H. Wickham, *et al.*, Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

78. B. J. Knaus, N. J. Grünwald, vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).

4. Chapter 3

4.1 Deep haplotype analyses of target-site resistance locus ACCase in blackgrass enabled by pool-based amplicon sequencing

Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2023), pp. 1–14 doi: 10.1111/pbi.14033

Deep haplotype analyses of target-site resistance locus *ACCase* in blackgrass enabled by pool-based amplicon sequencing

Sonja Kersten[1,2], Fernando A. Rabanal[2,*], Johannes Herrmann[3], Martin Hess[3], Zev N. Kronenberg[4], Karl Schmid[1] and Detlef Weigel[2,*]

[1]*Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany*
[2]*Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen, Germany*
[3]*Agris42 GmbH, Stuttgart, Germany*
[4]*Pacific Biosciences, Menlo Park, California, USA*

Received 27 June 2022;
revised 27 December 2022;
accepted 6 February 2023.
**Correspondence* (Tel +49 7071 601 1417; email fernando.rabanal@tue.mpg.de (F. A. R.) and Tel +49 7071 601 1410; fax +49 7071 601 1412; email weigel@tue.mpg.de (D. W.))

Keywords: Amplicon sequencing, HiFi long reads, *pbaa*, *Alopecurus myosuroides*, herbicide resistance, *ACCase*.

Summary

Rapid adaptation of weeds to herbicide applications in agriculture through resistance development is a widespread phenomenon. In particular, the grass *Alopecurus myosuroides* is an extremely problematic weed in cereal crops with the potential to manifest resistance in only a few generations. Target-site resistances (TSRs), with their strong phenotypic response, play an important role in this rapid adaptive response. Recently, using PacBio's long-read amplicon sequencing technology in hundreds of individuals, we were able to decipher the genomic context in which TSR mutations occur. However, sequencing individual amplicons are costly and time-consuming, thus impractical to implement for other resistance loci or applications. Alternatively, pool-based approaches overcome these limitations and provide reliable allele frequencies, although at the expense of not preserving haplotype information. In this proof-of-concept study, we sequenced with PacBio High Fidelity (HiFi) reads long-range amplicons (13.2 kb), encompassing the entire *ACCase* gene in pools of over 100 individuals, and resolved them into haplotypes using the clustering algorithm PacBio amplicon analysis (*pbaa*), a new application for pools in plants and other organisms. From these amplicon pools, we were able to recover most haplotypes from previously sequenced individuals of the same population. In addition, we analysed new pools from a Germany-wide collection of *A. myosuroides* populations and found that TSR mutations originating from soft sweeps of independent origin were common. Forward-in-time simulations indicate that TSR haplotypes will persist for decades even at relatively low frequencies and without selection, highlighting the importance of accurate measurement of TSR haplotype prevalence for weed management.

Introduction

Since the introduction of herbicides in agriculture in the 1940s, numerous plant species have evolved resistance to these chemicals. The two main mechanisms that led to rapid adaptation are non-target-site resistance (NTSR) and target-site resistance (TSR). NTSR refers to processes that degrade or physically prevent the active ingredient from reaching its target, such as enhanced metabolization, decreased absorption or translocation and sequestration (Devine and Shukla, 2000; Heap, 2014b). NTSR typically involves multiple genes (Cai *et al.*, 2022; Franco-Ortega *et al.*, 2021; Kreiner *et al.*, 2021; Van Etten *et al.*, 2020), resistance is often quantitative and several candidate gene families contribute to it, including cytochromes P450 monooxygenases, glycosyltransferases or glutathione S-transferases (reviewed in Gaines *et al.*, 2020). TSR has more qualitative effects, it is usually characterized by resistance to high levels of the herbicide, and it can often be traced back to large-effect gene mutations that change individual amino acids in herbicide target enzymes. More rarely, TSR is associated with overexpression of the target enzyme (Devine and Shukla, 2000).

The first TSR mutation was discovered in the *psbA* gene (Golden and Haselkorn, 1985). The *psbA* product, chlorophyll-binding protein D1, normally binds plastoquinone and serves as an essential component of photosystem II (PS II). The herbicide triazine competes with plastoquinone at the plastoquinone-binding site of protein D1, thus inhibiting PS II electron transport (reviewed in Gronwald (1997)). The amino acid substitution Ser-264-Gly prevents triazine binding, while still allowing plastoquinone binding (Gronwald, 1997). However, it comes with deleterious effects on CO_2 assimilation and plant development (Ireland *et al.*, 1988; Ort *et al.*, 1983). In the following decades, TSR mutations were identified in other genes including the genes for L-tubulin (Anthony *et al.*, 1998; Chu *et al.*, 2018; Délye et al., 2004a; Hashim *et al.*, 2012; Yamamoto *et al.*, 1998), acetolactate synthase (*ALS*) (Délye and Boucansaud, 2007; Tranel and Wright, 2002), acetyl-CoA carboxylase (*ACCase*) (reviewed in Kaundun (2014)) and 5-enolpyruvylshikimate-3-phosphate

Please cite this article as: Kersten, S., Rabanal, F A., Herrmann, J., Hess, M., Kronenberg, Z N., Schmid, K. and Weigel, D. (2023) Deep haplotype analyses of target-site resistance locus *ACCase* in blackgrass enabled by pool-based amplicon sequencing. *Plant Biotechnol J.*, https://doi.org/10.1111/pbi.14033.

synthase (*EPSPS*) (reviewed in Sammons and Gaines (2014)). In some cases, only a single amino acid substitution has been found to confer herbicide resistance, while in other genes, including *ALS* and *ACCase*, mutations at several residues can lead to herbicide resistance.

Many weeds have evolved independent resistances to multiple herbicides. Among them are European populations of the grassy weed *Alopecurus myosuroides*, where herbicide resistance results in significant yield losses for farmers (Rosenhauer *et al.*, 2013; Varah *et al.*, 2019). In fact, widespread resistance to ACCase inhibitors in *A. myosuroides* has greatly limited the ability of farmers to effectively control this problematic weed (Délye *et al.*, 2010; Heap, 2014a; Hess *et al.*, 2022; Rosenhauer *et al.*, 2013). Aryloxyphenoxy-propionates (FOPs), phenylpyrazolines (DENs) and cyclohexanediones (DIMs) all block the first step in fatty acid synthesis by inhibiting ACCase catalytic activity (Walker *et al.*, 1988). These herbicides act specifically on grasses because they target the homomeric plastidic ACCase, which is almost exclusively found in monocots and which is encoded in the nuclear genome (Incledon and Hall, 1997). All seven known sites at which TSR mutations occur are located in the penultimate exon, which encodes the C-terminal domain: Ile1781, Trp1999, Trp2027, Ile2041, Asp2078, Cys2088 and Gly2096. Depending on the mutation, amino acid substitutions confer resistance to one or several of the three different classes of ACCase inhibitor herbicides, Ile1781Leu and Asp2078Gly being resistant to all three classes (Beckie and Tardif, 2012). In *A. myosuroides*, plants with the Trp2027Cys and Ile2041Asn mutations survive treatments with FOPs and DENs, while Gly2096 confers resistance exclusively to FOPs (Délye, 2005; Délye *et al.*, 2008; Petit *et al.*, 2010). The degree of cross-resistance provided by TSRs is thought to be one of the factors that determine the frequency at which they are found (Gaines *et al.*, 2020; Powles and Yu, 2010). Another important factor comes from the effect each mutation has on herbicide-independent plant fitness. For instance, mutations at the most frequently affected site, Ile1781, appear to have no deleterious fitness effect (Délye *et al.*, 2013b; Menchari *et al.*, 2008). On the other hand, plants carrying the Asp2078Gly allele are shorter, have less vegetative dry biomass and set fewer seeds. Similarly, plants with the Trp2027Cys allele have lower seed production (Du *et al.*, 2019; Menchari *et al.*, 2008; Vila-Aiub *et al.*, 2015). The frequencies of *ACCase* TSR mutations have been investigated in several studies (Délye *et al.*, 2010; Délye *et al.*, 2004b; Menchari *et al.*, 2006; Rosenhauer *et al.*, 2013), but usually without considering the genomic context of the complete *ACCase* gene. Complete haplotype information is important in several respects, including establishing the number of times with which a specific mutation has occurred independently, and whether TSR mutations occur preferentially on specific haplotype backgrounds (Kersten *et al.*, 2023; Kreiner *et al.*, 2022).

Pool sequencing with Illumina short reads offers a cost- and time-saving option for the analysis of many individuals by combining barcoded DNA from multiple samples before sequencing (Ferretti *et al.*, 2013; Schlötterer *et al.*, 2014). This has included pooled amplicon sequencing approaches for resistance diagnosis assays in multiple species (Délye *et al.*, 2020, 2015; Schlipalius *et al.*, 2019). Unfortunately, due to the limited read lengths (from 50 to 300 bases in paired-end mode), to preserve haplotype information of an entire gene, variant calls have to be phased based on known patterns of linkage disequilibrium, with phasing accuracy depending on the co-occurrence of variants within paired-end reads. Long-read amplicons offer many advantages to solve the above-mentioned limitations. The most widely used third-generation long-read sequencing technologies are from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Unfortunately, in their native form, both suffer from limited per-base accuracy (below 90%), which until recently made reliable variant calling or haplotype determination in ONT and PacBio long reads difficult (Korlach, 2013). With the introduction of the PacBio circular consensus sequencing strategy to generate High Fidelity (HiFi) reads (Wenger *et al.*, 2019), random sequencing errors can now be corrected, and average per-base accuracy above 99% (q20) can now be routinely achieved (Travers *et al.*, 2010; Wenger *et al.*, 2019). In combination with the new clustering software PacBio amplicon analysis (*pbaa*) from Pacbio (Kronenberg *et al.*, 2021), this offers completely new perspectives for the application of amplicon sequencing in diagnostics.

In this study, we describe a high-throughput PacBio amplicon workflow for pooled samples that can be easily adapted to any gene of interest, independently of the organism. We demonstrate its feasibility with the TSR gene *ACCase*, for which we amplified a ~14 kb long fragment, which includes 585 bp upstream of the CDS, the 32 exons and 31 introns of the gene (12.5 kb) and 364 bp downstream. We provide a detailed hands-on laboratory protocol to amplify and long-read sequence loci such as *ACCase* as well as analysis recommendations for using the software *pbaa* in pools with up to 200 samples. We applied this workflow to German field populations of *A. myosuroides*. With the exception of a few low-frequency haplotypes, we were able to recover all individual haplotypes in the pools we tested. Furthermore, we found TSRs resulting from soft sweeps in almost all populations. Using SLiM simulations, we demonstrate that these TSR mutations may persist in field populations for decades to centuries, depending on their starting allele frequencies, even when selection is no longer applied. Therefore, it is strongly advised not only to base weed management strategies solely on herbicide applications but also to integrate mechanical weed management and crop rotation, to keep the incidence of weeds in the field continuously low with a combination of chemical and non-chemical measures.

Results

Workflow to sequence and analyse long-read amplicon pools

In a recent study, we sequenced PacBio long-read amplicons of the TSR locus *ACCase* in individuals of 47 European *A. myosuroides* populations (Kersten *et al.*, 2023). We discovered a recurrent pattern within field populations of different haplotypes with the same TSR mutation resulting from independent mutation events, as opposed to the same TSR mutation being transferred to other haplotypes by recombination. Characterizing the TSR diversity of entire haplotypes to this level of resolution was enabled by two main factors: sequencing of single individuals with HiFi reads and the clustering of these reads to reconstruct both haplotypes in each individual with the *pbaa* tool (Kronenberg *et al.*, 2021). However, performing independent DNA extractions and generating long-range amplicons with dual barcodes per individual proved to be both time-consuming and costly. To mitigate these limitations in future studies, we evaluated whether haplotype-level resolution can be achieved by sequencing per-field pools of large numbers of individuals. This is of interest for the further characterization of the origin and evolutionary tempo of herbicide resistance.

For benchmarking purposes, we selected nine populations from our previous study (Kersten *et al*., 2023) and compared the *ACCase* haplotypes determined from 22 to 24 independently sequenced individuals to *ACCase* haplotypes inferred from pools of 200 individuals. Each population was sown separately in the greenhouse. Then, we used a paper-size template to harvest similar amounts of 4-week-old leaf tissue from each plant and pooled them per population prior to DNA extraction (Figure 1a). Next, from 50 ng of DNA (on average ~65 diploid genome copies per individual in the pool; see Experimental procedures), we amplified a 13.2 kb long-range PCR fragment that encompasses the entire *ACCase* coding sequence including introns, plus 585 bp upstream and 364 bp downstream sequences. We used direct dual-indexing per pool, which later allowed multiplexing of all pools on a single SMRT cell. We paid special attention to combine similar amounts of PCR amplicons from all pools, by determining amplicon concentrations with a Qubit fluorometer and an additional gel electrophoresis for cross-validation before combining the pools (Figure 1a). A PacBio amplicon library was then created, size-selected using a BluePippin system and sequenced on the Sequel II system (Figures 1b, S1).

The current practice for *de novo* assembly studies, structural variant calling and amplicon analyses is to start from q20 HiFi reads, that is reads with an accuracy of at least 99% (Travers *et al*., 2010; Wenger *et al*., 2019). Since we were working with pools consisting of hundreds of individuals, our downstream analysis relied heavily on the precision of each individual read. Therefore, we increased the quality of the input HiFi reads to q30 (≥99.9% accuracy; Figure 1c). To compare haplotype frequencies between populations, we normalized all HiFi reads to the pool with the lowest number of reads, 16000 reads, corresponding to an average read depth of 40 for each amplicon represented in the sample (200 diploid individuals). Since the most common error types of HiFi reads are indels in homopolymer contexts (Travers *et al*., 2010; Wenger *et al*., 2019), we applied further filters including 'minimum cluster-read-count 20' (half the expected depth per single haplotype) and 'minimum-cluster-frequency 0.00125', which referred to the fraction of reads to support a true cluster in our data sets (Figure 1d).

Individuals versus pools – a *pbaa* cluster quality assessment

pbaa has been exclusively tested either on single individuals of diploid or polyploid species or on up to six HLA genes of the same individual (Kronenberg *et al*., 2021). After read-to-read alignment, for each focal read, *pbaa* sorts the alignments in decreasing identity and retains only the top 'n' alignments, which we call a pile. The frequency of each haplotype in the pool affects the parameter choices for *pbaa*. For a perfectly balanced pool, where every haplotype has the same number of reads, the pile size should match the expected haplotype read count. Therefore, to reduce spurious cluster formation, we adjusted the pile size to be about a quarter larger than the expected haplotype read count. The pile is used for error correcting each focal read. If the pile size is set too high, the pile will contain many cross-haplotype alignments and the haplotype-specific variant in the focal read will be corrected away. Similarly, the minimum variant frequency within a pile can affect which variants are masked out. Assuming the pile contains a high fraction of within haplotype alignments, a variant frequency cut-off of 0.4 performs well across a range of parameters. We then compared the resulting haplotypes in pools to haplotypes inferred from individuals of the same populations (Figure 2, Table 1).

The Belgium population BE01585 had a high diversity of TSR haplotypes of independent origin, a classical sign of soft sweeps due to herbicide selection pressure. Among the original 24 diploid individuals, we identified a total of 15 unique haplotypes, seven of which are haplotypes with TSR mutations (Table 1). In the pool data set, we successfully recovered 13 of the original 15 unique haplotypes identified in individuals, including all TSR haplotypes (Figure 2a). The two missing haplotypes in the pool were haplotypes found only in a single individual. Moreover, since the pool contained a larger number of individuals, we could identify 19 additional rare haplotypes, three of which were also TSR haplotypes. Notably, *pbaa* applied to pools was able to correctly resolve haplotypes that differed by a single mutation (Figure 2b,c).

The ability to recover a haplotype in the pools was influenced by its prevalence in the population. All haplotypes that were present in at least four out of 24 individuals were found in the corresponding pool, as were more than 85% of haplotypes present in two or three individuals (Figure 2d). Haplotypes found in only one out of 24 individuals were recovered in 71% of cases. This is most likely a reflection of the experimental design, in which the pools and 24 individuals were drawn from the same seed lots, which contained thousands of seeds, but the 24 individuals were not a subset of the pools of hundreds of individuals. Nevertheless, there was a high correlation ($R = 0.85$, $P < 2.2$ e-16) between haplotype frequency in individuals and in pools (Figure 2e). *pbaa* missed a few TSR haplotypes in the pools compared to the 24 individuals, but in all but one case, the analysis recovered additional TSR haplotypes in the pools. As one would expect from the deeper sampling, the number of haplotypes detected in the pools always exceeded the number of haplotypes found among the 24 individuals, from 15% to over twofold (Table 1). Thus, not only did the pools provide valuable, detailed information on the haplotype composition of field populations, but with the identification of up to 12 additional TSR haplotypes, they provided information of importance for resistance monitoring and herbicide use management (Hawkins *et al*., 2018; Powles and Yu, 2010). In addition, the collection of plant pools can constitute a valuable resource for the implementation of standardized epidemiological diagnostic methods, essential for monitoring future resistances (Comont and Neve, 2021).

Haplotype clustering reveals the evolutionary context

We employed our pool approach to survey TSR haplotype diversity in a German-wide contemporary collection of agricultural fields, for which seeds had been harvested in the year 2019. We selected 64 *A. myosuroides* populations collected in fields of winter annual crops: 49 populations that showed widespread resistance to the *ACCase*-inhibiting herbicides Axial® (active ingredients 50 g/L of pinoxaden and 12.5 g/L cloquintocet-mexyl), 13 populations with an incidence of Axial® resistance below 10% and two organic fields without a recent history of herbicide application (Data S1). Seventeen farms were represented with multiple fields. We used the same workflow as described above (Figure 1), but using pools of 150 individuals. To make results comparable across populations, the resulting HiFi reads were normalized to 5300 reads per pool, corresponding to an average read depth of 17.6 per chromosome (150 diploid individuals). The reads were filtered for 'minimum cluster-read-

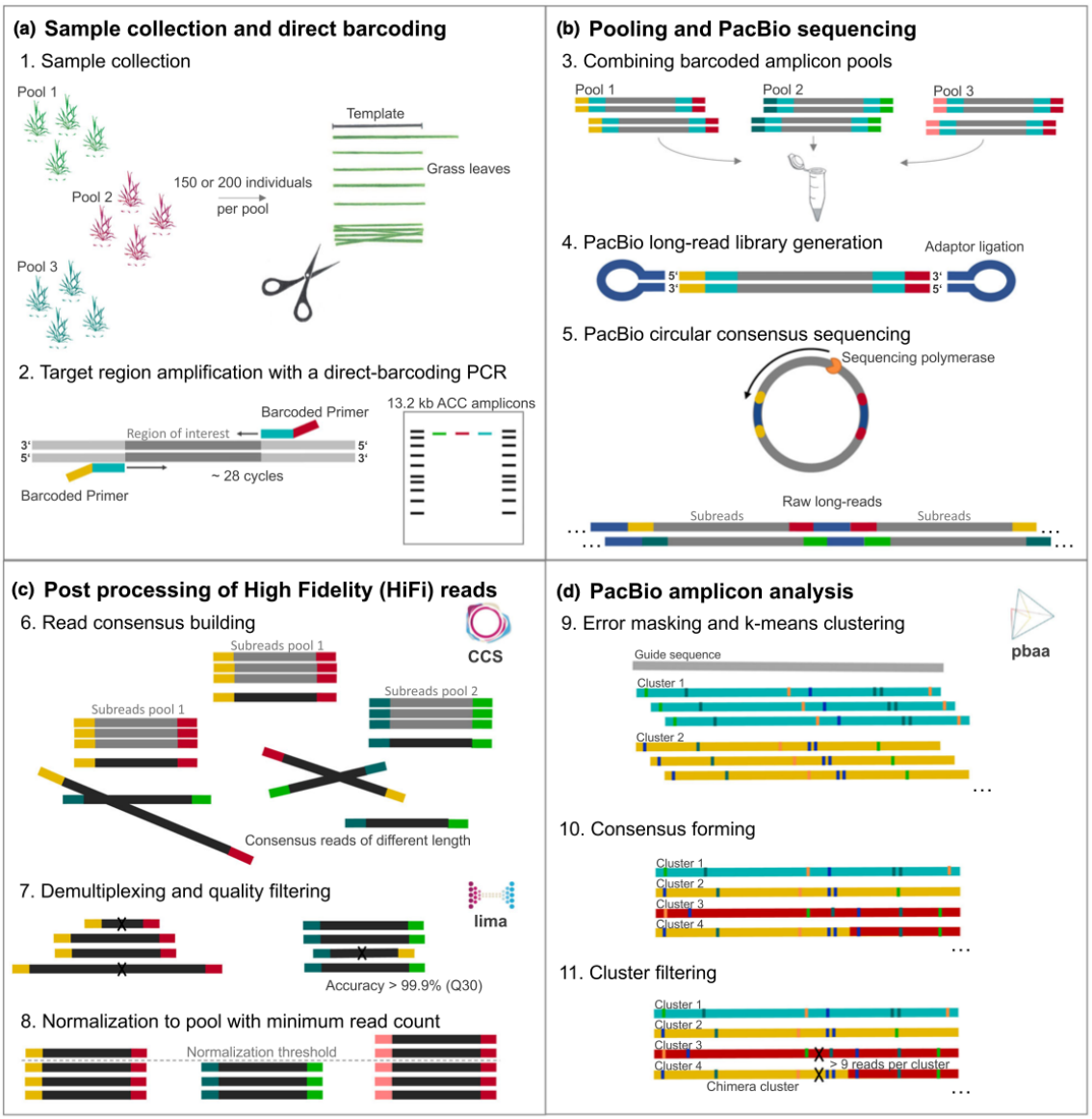


Figure 1 Workflow to generate and analyse long-read amplicons in pools. (a) Leaf material is collected using a size template to ensure equal sample representation in each pool. Long-range amplicon products are obtained by PCR with direct barcoding to individually tag each pool. Products are visualized by gel electrophoresis for quality control and validation of amplicon concentration measurements. (b) All population pools are combined in equal amounts in a single tube. A PacBio library is generated and sequenced in circular consensus mode on a Sequel II system. (c) Computational processing includes read-consensus building, demultiplexing and filtering of raw reads. (d) *pbaa* clustering is used for variant detection and filtering (Kronenberg *et al.*, 2021). The output is *fasta* files listing all haplotypes per population pool and including meta information on read coverage of each haplotype.

count 9' and 'minimum-cluster-frequency 0.0017', which led to an average number of 25 clusters per population (range 15–35).

Conventional single-nucleotide polymorphism (SNP) calling approaches have typically been used for variant calling and analysis of pooled data (Schlötterer *et al.*, 2014), but they generally ignore the underlying genomic context. Based solely on allele frequencies, we can estimate the abundance of each TSR mutation (Figure 3a), but we do not know in which context they emerged (Figure 3b). To further assess the accuracy of the pooled clustering approach, we compared the TSR haplotype frequencies with allele frequencies from a conventional SNP calling approach (Figure 3a,b). Pearson correlation coefficients were highly significant and ranged from 0.85 to 1 for all six TSR mutations, with only a few low-frequency *pbaa* clusters not captured (Figures 3c, S2).

The overall most common TSR mutation is Ile1781Leu, which has been reported in previous studies to increase the fitness of individuals in the absence of herbicide selection and therefore

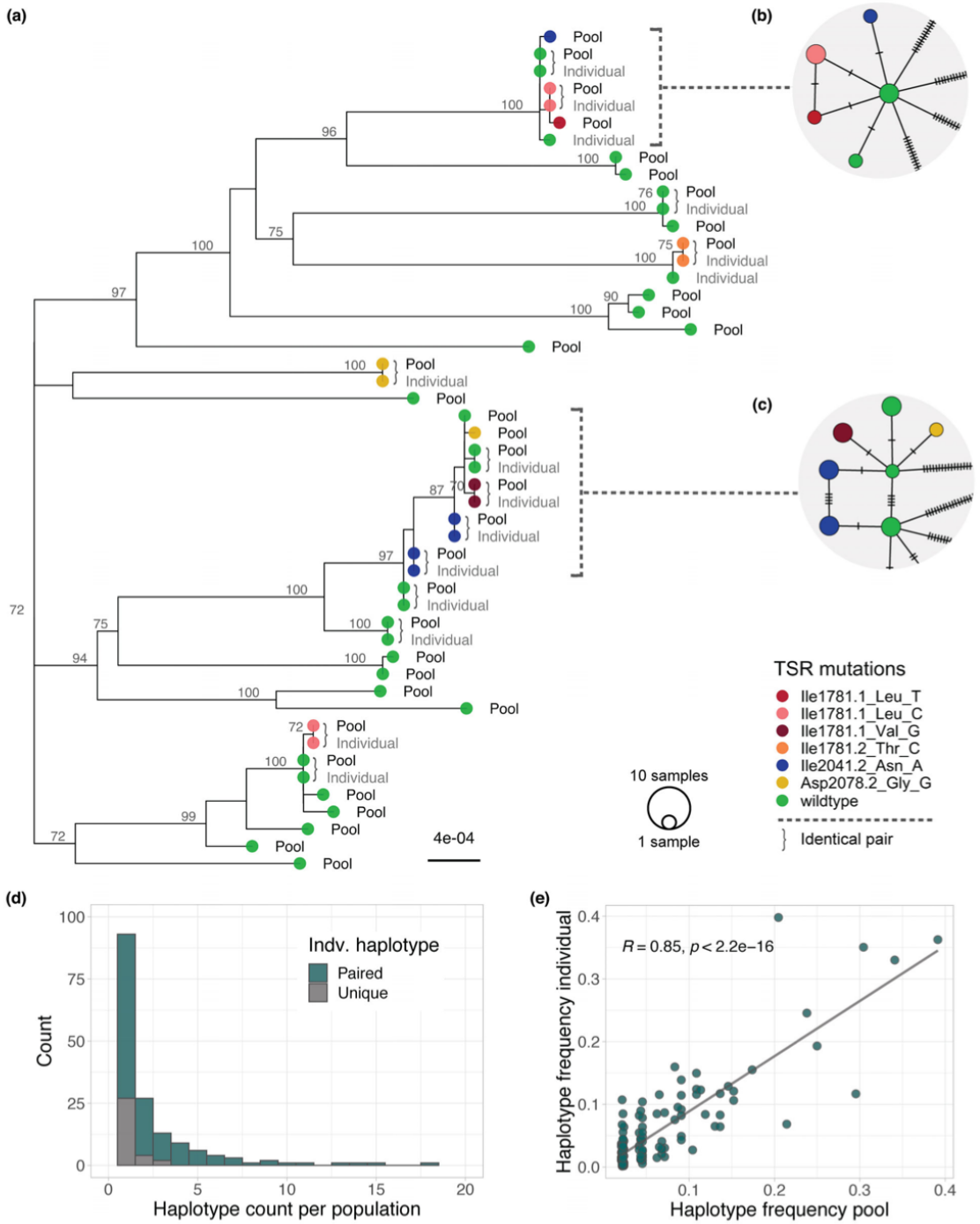


Figure 2 Unique *ACCase* haplotypes identified by *pbaa* in individuals compared to pools for the same population. (a) Maximum-likelihood tree of haplotypes identified in the pool data set (200 samples), and haplotypes inferred in the individual data set (24 samples). Samples were collected in an agricultural field in Belgium (BE01585). For simplicity, in the individual data set, only unique haplotypes are shown. Tree labels indicate the data set of origin (Pool vs. Individual). Coloured tree tips show target-site-resistance (TSR) mutations. Curly brackets mark identical haplotype pairs found in both the individual and the pool data set from the same population. (b, c) Haplotype network representing the corresponding clade in the tree. *pbaa* can successfully recover haplotypes that differ only in one mutation (tick bar). (d) Haplotype counts per population in the individual data set. The number of haplotypes that could have been successfully identified in the pool data set is marked in green. Only a fraction of the low abundant ones could not be recovered (grey). (e) Correlation of haplotype frequencies in the pool data set versus the individual data set.

Table 1 Individual haplotype recovery in pools

Population	Number of haplotypes: pool	Number of haplotypes: individuals	Number of correct pairs	Number of correct TSR pairs	Unidentified TSRs in pools	Additional TSRs in pools
BE01585	34	15	13 (15)	7 (7)	0	3
DE01467	26	16	14 (16)	2 (2)	0	2
DE01580	31	18	14 (18)	3 (3)	0	3
FR01434	41	21	17 (21)	3 (4)	1	7
FR01729	35	24	19 (24)	3 (4)	1	6
FR03200	24	11	8 (11)	3 (3)	0	3
FR07250	39	18	14 (18)	7 (10)	3	12
NL01505	26	22	19 (22)	2 (2)	0	0
UK06481	34	19	13 (19)	4 (5)	1	6

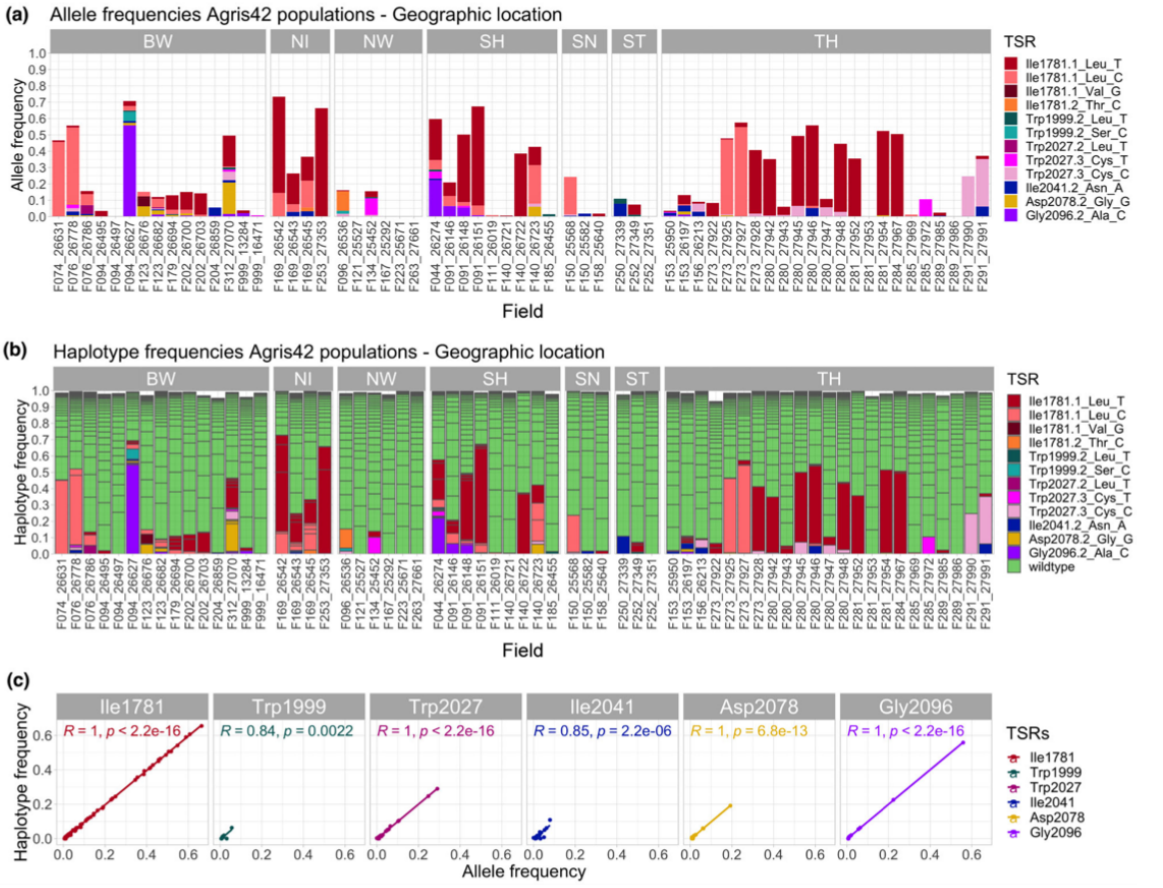


Figure 3 Comparison between conventional single-nucleotide polymorphism (SNP) mapping and *pbaa* haplotype clustering. (a) TSR allele frequencies obtained by SNP mapping. Colours indicate different TSR mutations. (b) Haplotype frequencies were inferred using *pbaa* (Kronenberg *et al*., 2021). Colours refer to TSR and wild-type haplotypes. (c) Correlation between allele frequencies and haplotype frequencies summarized per TSR amino acid position. Correlation coefficients and *P* values are shown separately in each TSR panel. BW, Baden-Württemberg; NI, Lower Saxony; NW, North Rhine-Westphalia; SH, Schleswig-Holstein; SN, Saxony; ST, Saxony-Anhalt; TH, Thuringia.

may have already existed in favourable conditions prior to herbicide application (Délye *et al*., 2013b; Du *et al*., 2019; Wang *et al*., 2010). While the number of populations per state was too small to make definitive statements about regional variation, the state with the smallest fields and farms, Baden-Württemberg, had the most diverse set of TSR mutations and haplotypes. However, states also vary in their history of herbicide use and thus are not that easily comparable. Very few TSR mutations were observed in

field populations of North Rhine-Westphalia, Saxony and Saxony-Anhalt, whereas the field populations in Thuringia seemed to be mainly dominated by single TSR haplotypes.

We refer to a hard sweep when a single haplotype dominates in a population. If, on the other hand, multiple adaptive haplotypes in a population increase in frequency at the same time, this is called a soft sweep (Hermisson and Pennings, 2017). In 38 of 55 German *A. myosuroides* populations containing TSR mutations, we can observe the latter phenomenon, confirming our previous results from European populations where herbicide adaptation occurred predominantly via soft sweeps through TSR mutations of independent origin (Kersten *et al.*, 2023). We also find a significant proportion of NTSR for the *ACCase*-inhibiting herbicide Axial® in this German data set, as the biotests reveal significantly more resistance than the TSR frequencies can explain (Figure S3a,c). However, the phenotypic resistance to Focus Ultra correlates significantly with the frequency of TSR mutations Ile1781Leu and Asp2078Gly, as reported before (Powles and Yu, 2010) (Figure S3b,d).

Organically farmed fields show TSRs of independent origin

Among the nine phenotypically sensitive populations included in the study, there were two organically farmed fields that have not been treated with herbicides going back as far as at least 1980, which predates the introduction of *ACCase* inhibitors to the market. In these fields, we found TSR haplotypes at low frequencies, from 0.3% to 2.0% (Figure 4), in agreement with our previous inferences that standing genetic variation is the most likely evolutionary mechanism behind herbicide selection (Kersten *et al.*, 2023). This is considerably higher than in a phenotyping-based study of the grass *Lolium rigidum*, the frequency of resistant individuals to ALS inhibitors in untreated populations ranged from 0.001% to 0.012% (Preston and Powles, 2002). The observation of TSR mutations in organic fields without a history of herbicide use is in agreement with the *ACCase* TSR mutation Ile1781Leu having been detected in one out of 685 (0.146%, or 0.073% at the haplotype level) *A. myosuroides* herbarium specimens collected about a 100 years ago (Délye *et al.*, 2013a). Under herbicide selection, strong resistance can develop within a few generations in such populations. This is due to the fact that mutations present as standing genetic variation have raised to certain frequencies and could already more easily establish in the populations (Hermisson and Pennings, 2005). This is further facilitated by high census population sizes, which can rapidly emerge in years with insufficient weed control and therefore provide a large genetic resource for resistance mutations (Menchari *et al.*, 2007).

Besides standing genetic variation, another potential source for TSR mutations in these organic fields could be gene flow and seed dispersal by wind, pollen flow, agricultural machinery or wildlife (Colbach and Sache, 2001; Somerville *et al.*, 2019). However, in the case of the organically farmed fields, we find not only the TSR haplotypes but also a corresponding wild-type haplotype, which differs by only one mutation in the entire 13.2 kb amplicon. This is true in all three cases, making it likely that the TSR mutations arose from these wild-type alleles independently in the fields – noticeably, for each field, there are also three pairs of wild-type haplotypes that differ by only a single mutation from each other. Moreover, we find the wild-type haplotypes in two out of the three cases with higher frequency than the corresponding TSR haplotypes (Figure 4), further suggesting that gene flow as a source is not very likely. Instead, the abundant plants with the matching wild-type haplotypes in these fields are the most likely source for different TSR mutations of independent origin.

TSRs will likely remain in fields for many decades even without selection

Adaptation to a new environment is often constrained due to pleiotropic fitness effects in previous conditions (reviewed in Purrington (2000)). TSR mutations in weed populations represent this special case in agricultural fields, where they become highly beneficial under herbicide application and rise in frequency. However, in the absence of herbicide selection, they have predominantly neutral or even detrimental effects (Du *et al.*, 2019; Menchari *et al.*, 2008; Tardif *et al.*, 2006; Vila-Aiub *et al.*, 2015). At least one *ACCase* mutation, Ile1781Leu, is known to be beneficial under neutral conditions (Délye *et al.*, 2013b; Wang *et al.*, 2010). On the other hand, there have been reports of fitness effects in several TSR mutations in the absence of selection, although quite often the differences do not persist when assessed in realistic field conditions or in competition with other plants (Du *et al.*, 2019). Unfortunately, the fitness proxies used – for example, biomass, netto assimilation rate, relative growth rate, plant height, leaf area ratio, seed production, or *ACCase*-specific activity – are difficult to compare and it is difficult to translate these observations into uniform estimates of selection coefficients (Anthimidou *et al.*, 2020; Sabet Zangeneh *et al.*, 2016; Vila-Aiub *et al.*, 2009, 2015, Yu *et al.*, 2007).

Because herbicide resistance has become such a serious problem in recent decades, it is important to learn whether the foregoing herbicide application for certain intervals is sufficient to remove a given TSR mutation from a field population via genetic drift. To tackle this question, we generated forward-in-time simulations with the software SLiM (Haller and Messer, 2019). While most studies focus on a few individuals of many populations (Délye *et al.*, 2004c; Menchari *et al.*, 2006), the depth of our pools allows us to assess more realistic haplotype frequencies of TSRs from our empirical data set (Figure 3). We used high (0.7; Figure 5a,e), intermediate (0.4; Figure 5b,f) and low (0.1 and 0.05; Figure 5c,d,g,h) initial TSR frequencies for our simulations, considering that many TSRs are usually present in the heterozygous state. The simulated selection coefficients ranged from 0 (no detrimental effect in the absence of herbicide selection) to 0.4 (40% fitness cost in the absence of herbicide selection). Within this range, we included the reported selection coefficient estimates for TSR mutations Trp2027Cys and Asp2078Gly, for which under realistic field scenarios, seed production was significantly reduced by 20% and 30%, respectively (Du *et al.*, 2019). Other parameters, such as effective population size, mutation and recombination rate, were obtained from the literature (Bauer *et al.*, 2013; Kersten *et al.*, 2023; Yang *et al.*, 2017). For each mutation and initial allele frequency, we simulated two different dominance coefficients derived from fitness experiments in *A. myosuroides* (Menchari *et al.*, 2008), an intermediate, codominant coefficient of 0.5 (Figure 5a–d), and a recessive coefficient of 0.25 (Figure 5e–h). We conducted 400 independent SLiM simulation runs per parameter combination and estimated the average number of generations for a TSR to be removed from the population by genetic drift.

The SLiM simulations indicated that under the best-case scenario, with a low initial allele frequency (0.05), a strong deleterious selection coefficient (–0.4), and codominance (0.5), it would take, on average, 36 generations (the average number of

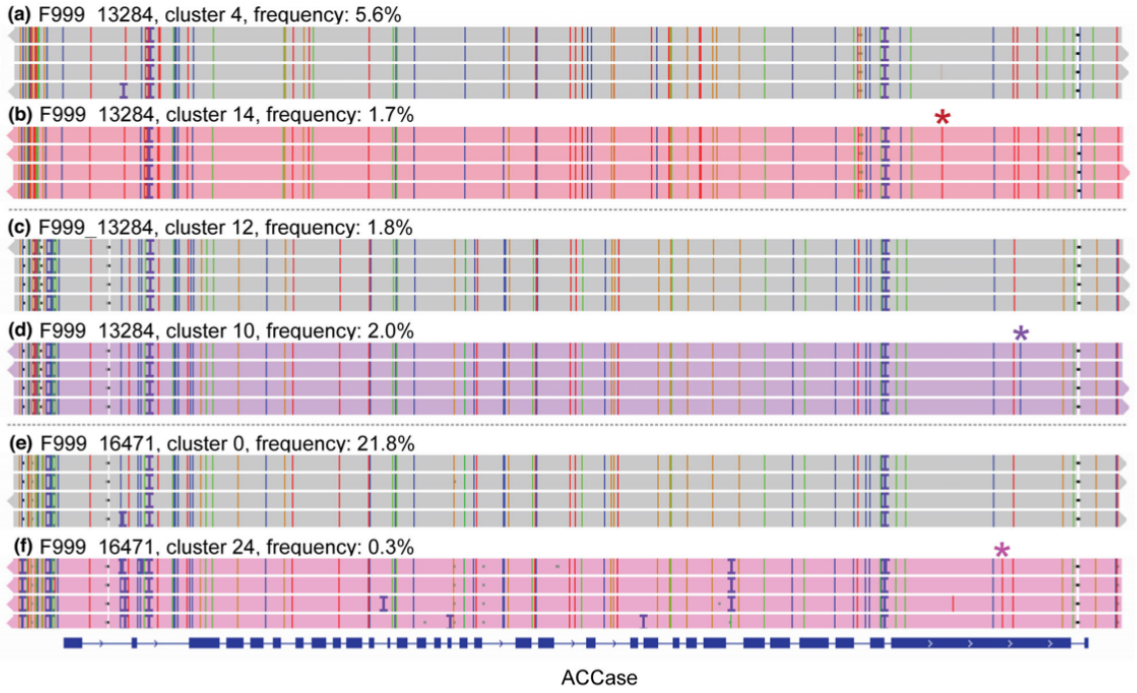


Figure 4 TSR haplotypes and the corresponding wild-type haplotypes from which they arose in organic fields. (a, c, e) Wild-type haplotypes, (b) a haplotype with the TSR mutation Ile1781.1Leu_T, (d) a haplotype with the TSR mutation Gly2096.2Ala_C, (f) a haplotype with the TSR mutation Trp2027.3Cys_T.

generations at which all simulations reach 0 allele frequency) until the TSR mutation is lost (Figure 5d). Unfortunately, farmers often recognize a resistance problem in their fields only once the TSR mutations have already risen to high frequency. Furthermore, most TSR mutations do not seem to reach such a strong deleterious fitness effect (Du *et al.*, 2019; Menchari *et al.*, 2008). For example, under a milder selection coefficient of −0.1 (still below what has been reported for Trp2027Cys; Du *et al.*, 2019), codominant mutations would persist in a non-treated field, on average, for up to 204 generations (when the initial allele frequency = 0.7), and up to 330 generations when more recessive. Since these numbers of generations are mostly beyond the lifespan of a farmer, not to mention the economical loss incurred by a field being fallow for decades, additional measures need to be taken to manage and prevent herbicide resistance.

Discussion

Pool sequencing of amplicons with PacBio HiFi reads is a cost-effective method for sequencing thousands of samples while preserving haplotype resolution. The *pbaa* clustering software eliminates the need for read alignment against a reference and phasing. Instead, HiFi sequences are clustered directly, preserving the full information contained and reducing reference bias. This opens up new avenues for the discovery of unknown structural variants and genetic diversity. Furthermore, the complete amplicon workflow can be easily established as a high-throughput method for almost any gene of interest in any organism. A notable exception for any pool-based approach in a single locus would be the monitoring of recent gene amplification, such as the alternative resistance mechanism in response to glyphosate discovered for the *EPSPS* gene (Gaines *et al.*, 2010).

Importantly, in cases, where the genes of interest are shorter than *ACCase*, long amplicons that include intergenic sequence up- and downstream of the gene are likely to provide even higher resolution of alleles, as variation in intergenic sequences is usually higher than in more constrained genic regions. Although *pbaa* was able to resolve haplotypes that differ by a single variant since the technology has difficulties with homopolymers, it would be prudent to mask these, at least beyond a certain length (e.g. >6 bp).

A valid concern is whether PacBio HiFi technology is appropriate for applications that require the fast return of sequencing data. In a commercial setting, a large number of samples collected in a short-time frame will help to quickly fill an entire SMRT cell. Alternatively, one could use another long-read technology, such as Oxford Nanopore Technologies. Although single-read accuracy has been limiting for Oxford Nanopore data, the latest developments with duplex read promise to reach q30 (https://github.com/nanoporetech/duplex-tools), as used here for PacBio HiFi-based amplicons.

Based on the two population studies we conducted in *A. myosuroides*, Europe-wide and in Germany, we can conclude that herbicide resistance arises independently in different field populations. This puts farmers and consultants in charge to investigate their fields carefully and obtain the status quo of their fields in terms of the resistance situation because once a resistance mechanism is established in a field population, it is highly unlikely to be lost over the course of a human lifetime,

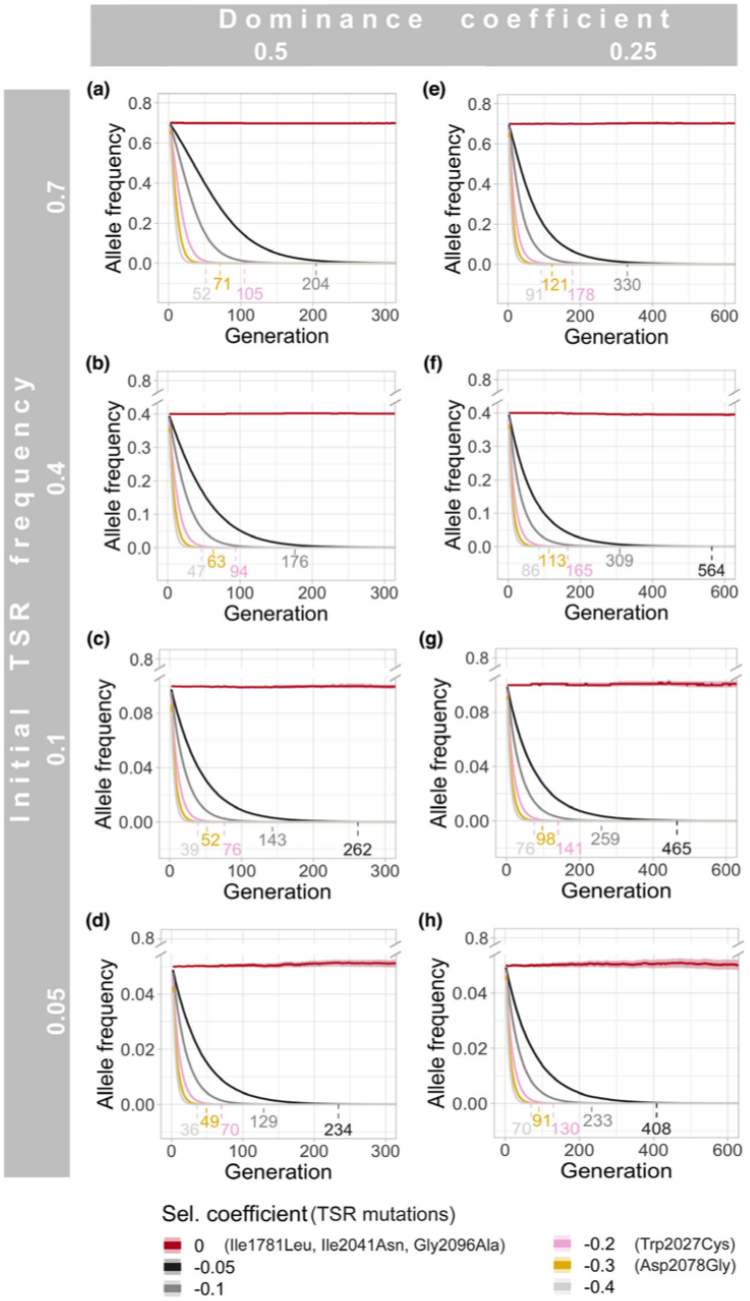


Figure 5 Simulations of the number of generations in which TSR alleles remain in *A. myosuroides* field populations in the absence of selection, assuming different selection coefficients, as estimated from fitness experiments (Du *et al.*, 2019; Menchari *et al.*, 2008). While homozygous individuals suffer the full consequences of deleterious TSR mutations, we simulated two different dominance coefficients for heterozygous allele states: an intermediate codominance of 0.5 (a–d) and a more recessive coefficient of 0.25 (d–h). The coloured numbers above the x-axis indicate the average number of generations at which the mutations shown at the bottom are lost in the different scenarios. Means and 0.95 confidence intervals per parameter combination are shown.

even after herbicide application is stopped. The variation in resistance across the fields sampled in the current study supports the assertion that weed management strategies should focus on the field level, requiring accurate and up-to-date information on the prevalence of herbicide resistance in a given field. A recent survey in Germany found that while only 20% of agricultural

fields suffered from high levels of infestation with *A. myosuroides*, resistance to the *ACCase* inhibitor pinoxaden could be detected in 80% of samples (Hess *et al*., 2022). This indicates that successful resistance management requires precautionary control of the census population size of the weed. Management strategies should therefore focus not only on chemical but also non-chemical measures, such as delayed seeding, moldboard ploughing and crop rotation (Lutman *et al*., 2013; Moss *et al*., 2007).

Experimental procedures

European sample collection

The European collection was provided by BASF. Amplicon sequencing data of 22–24 single individuals from 47 populations has been described (Kersten *et al*., 2023). For this study, we selected nine of those populations containing TSR mutations and resowed and sequenced pools of 200 individuals to assess the potential of *pbaa* clustering in pools versus individuals.

German sample collection and phenotyping

In the course of a Germany-wide herbicide resistance assessment (2019), a collection of *A. myosuroides* seeds on 1369 agricultural fields was conducted (Hess *et al*., 2022) All samples came from fields sown with winter wheat or triticale in the year of sampling and were screened in a biotest prior to sequencing. Seeds were sown in sandy-loam substrate and treated at BBCH 12/13. Two *ACCase* inhibitors were used for the screening, Axial® (50 g/L of pinoxaden and 12.5 g/L cloquintocet-mexyl) and Focus® Ultra (100 g/L cycloxydim). Herbicide application was done with 200 L water in a Research Track Sprayer Generation III using a Teejet-8002-EVS-Nozzle and field rates of 1.2 L/ha for Axial® and 2.5 L/ha for Focus® Ultra. A visual assessment of the efficacy was done 21 days after treatment. All plants were screened in two replicates together with well-characterized standard populations. Sixty-four samples were later chosen based on the number of seeds available to conduct further tests, the suitability to form regional clusters, and variations in the degree of efficacy of the tested herbicides. Besides two samples from organic farms, all other samples were collected from conventional farms.

Growth conditions, harvesting and DNA extraction

All seeds were sown in a standard substrate (Pikiererde Typ CL P, Cat.No EN12580; Einheitserde) and stratified at 4 °C in a climate chamber. Then they were transferred to the greenhouse at 22 °C with 16 h daylight. For the pilot experiment, we harvested 200 individuals per pool in the European collection. For the German data set, 150 individuals were collected from each population. To ensure equal representation of all individuals per pool, grass leaves were cut using a 2.5 cm size template (ca. 10 mg leaf material per plant). Care was also taken to ensure that the leaves were of similar width. All pool samples were collected in 50 mL Falcon tubes filled with 4–5 metal beads and ground with a FastPrep-24™ 5G tissue disruptor using the CoolBigPrep™ 2 × 50 mL-Adapter filled with dry ice (Prod. No machine: 15260488, Prod. No adapter: 11471525; MP Biomedicals, Irvine, CA).

DNA purification was performed as detailed in our online hands-on DNA extraction protocol in GitHub. Briefly, 300 mg of plant leaf powder per pool was incubated for 60 min at 60 °C in 800 µL of lysis buffer (100 mM Tris pH 8, 50 mM EDTA pH 8, 500 mM NaCl, 1.3% SDS and 0.01 mg/mL RNase A) in a 2 mL screw cap tube. After centrifugation at 12 000 *g* for 1 min, 200 µL of the supernatant were transferred to a fresh 2 mL Eppendorf Safe-Lock tube (Prod. No. 0030120094). To precipitate proteins, 65 µL of 5 M potassium acetate was added to each sample. After vigorous vortexing and a short spin, samples were incubated for 15 min at −20 °C. Following a centrifugation step at 12 000 *g* for 1 min, 200 µL of supernatant was transferred to a fresh 2 mL Eppendorf tube. A first cleanup was performed for 5 min by adding 300 µL of 0.4% solution of SeraMag™ SpeedBead Carboxylate-Modified [E3] Magnetic Particles (Prod. No. 65152105050450; GE Healthcare). After placing the tube on a magnet, the supernatant was discarded and beads were washed twice with 80% ethanol while keeping the tube on the magnet. Elution was performed with 50 µL of water. A second cleanup was performed with 50 µL of 0.4% solution of SeraMag™ beads, and ethanol washes and elution were done as before.

ACCase amplicon generation and PacBio sequencing

To generate the *ACCase* amplicons, we used a direct dual barcoding approach with target-specific primers (24 forward and 16 reverse) that had the barcode sequences in their 5′ ends (Data S1). We conducted four independent PCR reactions (using different primer pairs) for each pool. For each PCR, we use 50 ng of DNA as a template. The number of template copies in 50 ng of input DNA of a genome estimated to be 3.56 Gbp (Kersten *et al*., 2023) was estimated to be 13012 according to the following equation:

$$\text{Number of copies} = \frac{\text{Amount input DNA (ng)} \times 6.022 \times 10^{23}\ \text{(molecules/mole)}}{\text{Length of haploid genome (bp)} \times 1 \times 10^{9}\ \text{(ng/g)} \times 650\ \text{(g/mole of bp)}}$$

That is, 32.5 and 43.4 copies per haploid genome per PCR reaction for pools of 400 and 300 diploid individuals, respectively.

The 13.2-kb-long target region was amplified using a master mix reaction with 1 µL Forward indexing primer (5 µM), 1 µL Reverse indexing primer (5 µM), 4 µL 5× Prime STAR buffer, 1.6 µL dNTPs (2.5 mM each), 0.4 µL Prime STAR GXL polymerase (1.25 U/µL) (R050B; Takara Bio Inc., Shiga, Japan), filled up to 20 µL with water in a two-step PCR reaction with 28 cycles (denaturation: 98 °C, 10 s; annealing: 68 °C, 11 min; final extension: 72 °C, 10 min; hold: 4 °C). For a quality check, 5 µL of each amplicon pool was visualized on a 0.8% agarose gel, and the concentration was determined with a Qubit™ system. Then, all amplicons were pooled equally into a large pool, bead-cleaned and size-selected using a BluePippin system (Sage Science, Beverly, MA) with High-Pass Plus 0.75% agarose cassettes, 15 kb (342BPLUS03; Biozym Scientific GmbH, Germany). Only fragments larger than 10 kb were retained (Figure S1). The correct fragment size selection was verified with a Femto Pulse system (Agilent, Santa Clara, CA). The PacBio library was created following protocol no. 101-791-800 version 01 (June 2019) with the SMRTbell Express Template Prep Kit 2.0 (part number 100-938-900). Sequel® II loading was performed according to manufacturer specifications with Sequel® II Binding Kit 2.0 and Int Ctrl 1.0 (part number 101-842-900). A detailed hands-on amplicon protocol can be found in Github.

Generation and demultiplexing of q30 HiFi reads

Pre-processing steps were carried out with PacBio tools (https://github.com/PacificBiosciences/pbbioconda). This included the generation of circular consensus sequences (ccs) with ccs v6.0.0

with a minimum predicted accuracy of 0.999 (q30), demultiplexing of pools with lima v1.11.0 with parameter settings '--ccs --different --peek-guess --guess 80 --split-bam-named --min-ref-span 0.875 --min-scoring-regions 2', and conversion of the resulting bam to fastq format with bam2fastq v1.3.0.

Pbaa clustering

Prior to the *pbaa* clustering, we concatenated the q30 HiFi reads of all PCR reactions corresponding to each pool and normalized the number of reads in each pool to the population with the lowest read counts of each data set by random sampling with fastqtools v0.8.3 (fastq-sample -n<read_number>) (https://github.com/dcjones/fastq-tools) and indexed each pool with samtools faidx v1.9 (Li *et al*., 2009). In the European collection, we used 16 000 reads per pool and in the German collection 5300 reads. Furthermore, one population in the German data set did not have enough reads; therefore, it was excluded from further analyses.

The provided guide sequence for the reference-aided clustering approach covered the complete *ACCase* gene sequence and originated from a sensitive plant of a northern Germany reference population (WHBM72 greenhouse standard APR/HA from Sep. 2014) (sequence provided in the GitHub repository for this project).

In the European data set, *pbaa* v1.0.0 (commit 691333c) clustering was performed with --min-read-qv 30 --max-alignments-per-read 16 000 --max-reads-per-guide 16 000 --pile-size 50 --min-var-frequency 0.4 --min-cluster-read-count 20 --min-cluster-frequency 0.00125. In the German data set, we used the following adjusted parameters: --max-alignments-per-read 5300 --max-reads-per-guide 5300 --pile-size 25 --min-var-frequency 0.4 --min-cluster-read-count 9 --min-cluster-frequency 0.0017 (https://github.com/PacificBiosciences/pbAA). Finally, to extract the consensus sequences generated in the clustering step, including meta information of each haplotype, and re-orient them – when necessary – in the forward orientation, we used a homemade script, which can be found in the dedicated GitHub for this study.

Pbaa validation in the European data set

All clusters inferred by *pbaa* in the pools and all unique haplotypes from the individuals were combined into a joint fasta file per population. MAFFT v7.407 was used for the multiple alignments (–-thread 20 --threadtb 10 --threadit 10 --reorder --maxiterate 1000 --retree 1 --genafpair) (Katoh and Standley, 2013) and PGDSpider v2.1.1.5 to transfer the multiple alignment fasta file into a nexus formatted file (Lischer and Excoffier, 2012). The maximum-likelihood (ML) tree was generated with RAXML-NG v0.9.0 using the GTR + G model and 1000 bootstraps (Kozlov *et al*., 2019). The minimum spanning network was inferred and visualized with POPART v.1.7 (Leigh and Bryant, 2015). The TSR information for the colouring of the haplotype tree and network was retrieved from a classical alignment of the *pbaa* clusters to the *ACCase* reference gene (see Section 'Annotation of TSR mutations'). The resulting VCF was loaded and manipulated in R to annotate the ML tree and minimum spanning network. Used R packages can be found in Table S1.

Based on the multiple alignments per population, haplotypes in the pool and individual datasets were counted with the R package 'haplotypes' (https://cran.r-project.org/web/packages/haplotypes/haplotypes.pdf) and summarized in Table 1. Haplotype frequencies were calculated with homemade R scripts and the correlations of individual and pool haplotype frequencies were calculated and visualized using the packages 'ggpubr' (https://github.com/kassambara/ggpubr/) and 'ggplot2' (Wickham, 2016).

Comparison of conventional SNP mapping with *pbaa* clustering in the German data set

For the conventional alignment and SNP calling, the reads of each pool were aligned to the *ACCase* reference sequence with pbmm2 (https://github.com/PacificBiosciences/pbmm2). All resulting bam files were merged, sorted and indexed with samtools v1.9 (Li *et al*., 2009). SNP calling was performed with freebayes v1.3.2 (freebayes -f $REF --min-mapping-quality 20 --min-alternate-fraction 0.005 --pooled-continuous --report-monomorphic) (Garrison and Marth, 2012). All single VCF files were compressed, indexed and merged using tabix v0.2.5 (Li, 2011). Before extracting allele depth (AD) and total depth (DP) information for SNPs at TSR positions to compare REF and ALT counts, the multi-allelic positions were split into multiple rows of biallelic calls with bcftools v1.9-15-g7afcbc9 (bcftools norm -m -any -Oz) (Danecek and McCarthy, 2017), followed by converting the variants into a table using the VariantsToTable function from GATK 4.1.3.0 (Van der Auwera *et al*., 2013). The table was loaded and manipulated in R version 3.6.1 (Team, 2018), and allele frequencies were plotted with the R package 'ggplot2' (Wickham, 2016).

Annotation of TSR mutations

To annotate the clusters generated with the *pbaa* clustering approach with the TSR information, the single cluster fasta files were transferred to *fastq* files in which all bases were assigned quality 'I', with Fasta_to_fastq (https://github.com/ekg/fasta-to-fastq). Afterwards, the *fastq* files containing a single read representing the corresponding cluster were aligned to the *ACCase* reference with minimap2 v2.15-r913-dirty (Li, 2018). The resulting bam file was sorted and indexed with samtools v1.9 (Li *et al*., 2009) and the read groups were adjusted with Picard's v2.2.1 function AddOrReplaceReadGroups (RGID = $SAMPLE RGLB = ccs RGPL = pacbio RGPU = unit1 RGSM = $SAMPLE) (http://broadinstitute.github.io/picard). Variant calling, which results in the VCF file, was performed with the HaplotypeCaller from GATK 4.1.3.0 (--R $REF --min-pruning 0 -ERC GVCF), followed by GenotypeGVCFs with standard settings (Van der Auwera *et al*., 2013). Variants in the resulting VCF file were annotated with SnpEff v4.3t (Cingolani *et al*., 2012).

Organic fields with TSR mutations of independent origin

The TSR information of the organic fields was extracted from the previously described haplotype clustering in the German data set. The clusters in the BAM files were coloured with *pbaa* v.1.0.0 bampaint and visualized in the Integrative Genomics Viewer IGV_2.11.9 (Robinson *et al*., 2011).

SliM simulations

We performed forward simulations with SLiM v3.4 (Haller and Messer, 2019) under Wright-Fisher model assumptions to determine the number of generations that TSR mutations persist in agricultural fields without being under herbicide selection. A population size of 42 000 individuals was assumed, following calculations from the previous publication (Kersten *et al*., 2023). Similarly, we adopted the mutation rate of 3.0×10^{-8} (Yang *et al*., 2017) and genome-wide average recombination rate of 7.4×10^{-9} (Bauer *et al*., 2013) from maize, a diploid grass with a comparable genome size. We modelled a range of selection coefficients (s_i) from 0 to –0.4, covering values based on

literature that compared seed production of wild-type and mutant American sloughgrass (*Beckmannia syzigachne* Steud.) plants during competition with wheat plants under field conditions (s_i = −0.2 for Trp2027Cys, and s_i = −0.3 for Asp2078Gly) (Du *et al.*, 2019). We used two dominance coefficients (h_i) 0.5 and 0.25 for the TSR mutations as reported in *A. myosuroides* (Menchari *et al.*, 2008). The fitness model for individuals carrying a homozygous TSR mutation is $1 + s_i$, and for a heterozygous one is $1 + h_i * s_i$. Initial haplotype frequencies were extracted from our empirical pool data and set to 0.05, 0.1, 0.4 and 0.7. We performed 400 independent SLiM runs per parameter combination and calculated the mean values and the 0.95 confidence intervals in R with the package 'rcompanion' (https://rcompanion.org/handbook/). Visualization was done with 'ggplot2' (Wickham, 2016).

Acknowledgements

We thank Andreas Landes and Jens Lerchl (BASF SE) for providing the Europe-wide populations of *A. myosuroides*, Angela Kuttler and Jakob Keck for help with the sowing and leaf material sampling in the greenhouse and Christa Lanz (MPI for Biology Tübingen) for assistance with the PacBio amplicon library preparation and the Sequel II loading. Open Access funding enabled and organized by Projekt DEAL.

Funding

S.K. was supported by a stipend from the Landesgraduiertenförderung (LGFG) of the State of Baden-Württemberg. F.A.R. was supported by a Human Frontiers Science Program (HFSP) Long-Term Fellowship (LT000819/2018-L). The majority of funding was provided by the Max Planck Society.

Conflict of interest

J.H. is the founder and M.H. is the owner of Agris42, a company providing herbicide resistance testing services and weed management consultation to farmers. D.W. holds equity and S.K. is an employee of Computomics, which advises breeders. Z.N.K. is an employee and shareholder of Pacific Biosciences, a company developing single molecule-sequencing technologies. Other authors declare no competing or financial interest.

Author contributions

Conceptualization: F.A.R.; Investigation: S.K. with support from F.A.R., J.H. and M.H.; Software: Z.N.K.; Formal Analysis: S.K.; Resources: J.H. and M.H.; Writing – Original Draft: S.K.; Writing – Review and Editing Preparation: S.K., F.A.R. and D.W.; Visualization: S.K.; Supervision: F.A.R., K.S. and D.W.; Funding Acquisition: D.W.

Data availability statement

PacBio HiFi q30 reads for each pool have been deposited in the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under project accession number PRJEB[illegible] Experimental protocols, SLiM simulations and custom scripts to reproduce the analyses in this study are deposited on GitHub (https://doi.org/10.5281/zenodo.7646820).

References

Anthimidou, E., Ntoanidou, S., Madesis, P. and Eleftherohorinos, I. (2020) Mechanisms of Lolium rigidum multiple resistance to ALS- and ACCase-inhibiting herbicides and their impact on plant fitness. *Pestic. Biochem. Physiol.* **164**, 65–72.

Anthony, R.G., Waldin, T.R., Ray, J.A., Bright, S.W. and Hussey, P.J. (1998) Herbicide resistance caused by spontaneous mutation of the cytoskeletal protein tubulin. *Nature* **393**, 260–263.

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L. *et al.* (2013) Intraspecific variation of recombination rate in maize. *Genome Biol.* **14**, R103.

Beckie, H.J. and Tardif, F.J. (2012) Herbicide cross resistance in weeds. *Crop Prot.* **35**, 15–28.

Cai, L., Comont, D., MacGregor, D., Lowe, C., Beffa, R., Neve, P. and Saski, C. (2022) The blackgrass genome reveals patterns of non-parallel evolution of polygenic herbicide resistance. *New Phytol.* **237**, 1891–1907

Chu, Z., Chen, J., Nyporko, A., Han, H., Yu, Q. and Powles, S. (2018) Novel α-tubulin mutations conferring resistance to dinitroaniline herbicides in Lolium rigidum. *Front. Plant Sci.* **9**, 97.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92.

Colbach, N. and Sache, I. (2001) Blackgrass (*Alopecurus myosuroides* Huds.) seed dispersal from a single plant and its consequences on weed infestation. *Ecol. Model.* **139**, 201–219.

Comont, D. and Neve, P. (2021) Adopting epidemiological approaches for herbicide resistance monitoring and management. *Weed Res.* **56**, 1–13.

Danecek, P. and McCarthy, S.A. (2017) BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039.

Délye, C. (2005) Weed resistance to acetyl coenzyme A carboxylase inhibitors: an update. *Wees* **53**, 728–746.

Délye, C. and Boucansaud, K. (2007) A molecular assay for the proactive detection of target site-based resistance to herbicides inhibiting acetolactate synthase in *Alopecurus myosuroides*. *Eur. Weed Res. Soc. Weed Res.* **48**, 97–101.

Délye, C., Menchari, Y., Michel, S. and Darmency, H. (2004a) Molecular bases for sensitivity to tubulin-binding herbicides in green foxtail. *Plant Physiol.* **136**, 3920–3932.

Délye, C., Straub, C., Matéjicek, A. and Michel, S. (2004b) Multiple origins for black-grass (*Alopecurus myosuroides* Huds) target-site-based resistance to herbicides inhibiting acetyl-CoA carboxylase. *Pest Manag. Sci.* **60**, 35–41.

Délye, C., Straub, C., Michel, S. and Le Corre, V. (2004c) Nucleotide variability at the acetyl coenzyme A carboxylase gene and the signature of herbicide selection in the grass weed *Alopecurus myosuroides* (Huds.). *Mol. Biol. Evol.* **21**, 884–892.

Délye, C., Matéjicek, A. and Michel, S. (2008) Cross-resistance patterns to ACCase-inhibiting herbicides conferred by mutant ACCase isoforms in *Alopecurus myosuroides* Huds. (black-grass), re-examined at the recommended herbicide field rate. *Pest Manag. Sci.* **64**, 1179–1186.

Délye, C., Michel, S., Bérard, A., Chauvel, B., Brunel, D., Guillemin, J.-P. *et al.* (2010) Geographical variation in resistance to acetyl-coenzyme A carboxylase-inhibiting herbicides across the range of the arable weed *Alopecurus myosuroides* (black-grass). *New Phytol.* **186**, 1005–1017.

Délye, C., Deulvot, C. and Chauvel, B. (2013a) DNA analysis of herbarium Specimens of the grass weed *Alopecurus myosuroides* reveals herbicide resistance pre-dated herbicides. *PLoS One* **8**, e75117.

Délye, C., Menchari, Y., Michel, S., Cadet, E. and Le Corre, V. (2013b) A new insight into arable weed adaptive evolution: mutations endowing herbicide resistance also affect germination dynamics and seedling emergence. *Ann. Bot.* **111**, 681–691.

[illegible] https://www.ebi.ac.uk/ena/browser/home R., Gautier, V., Poncet, C. and Michel, S. (2015) Using next-generation sequencing to detect mutations endowing resistance to pesticides: application to acetolactate-synthase (ALS)-based resistance in barnyard grass, a polyploid grass weed. *Pest Manag. Sci.* **71**, 675–685.

Délye, C., Michel, S., Pernin, F., Gautier, V., Gislard, M., Poncet, C. and Le Corre, V. (2020) Harnessing the power of next-generation sequencing technologies to the purpose of high-throughput pesticide resistance diagnosis. *Pest Manag. Sci.* **76**, 543–552.

Devine, M.D. and Shukla, A. (2000) Altered target sites as a mechanism of herbicide resistance. *Crop Prot.* **19**, 881–889.

Du, L., Qu, M., Jiang, X., Li, X., Ju, Q., Lu, X. and Wang, J. (2019) Fitness costs associated with acetyl-coenzyme A carboxylase mutations endowing herbicide resistance in American sloughgrass (*Beckmannia syzigachne* Steud.). *Ecol. Evol.* **9**, 2220–2230.

Ferretti, L., Ramos-Onsins, S.E. and Pérez-Enciso, M. (2013) Population genomics from pool sequencing. *Mol. Ecol.* **22**, 5561–5576.

Franco-Ortega, S., Goldberg-Cavalleri, A., Walker, A., Brazier-Hicks, M., Onkokesung, N. and Edwards, R. (2021) Non-target site herbicide resistance is conferred by two distinct mechanisms in black-grass (*Alopecurus myosuroides*). *Front. Plant Sci.* **12**, 636652.

Gaines, T.A., Zhang, W., Wang, D., Bukun, B., Chisholm, S.T., Shaner, D.L. *et al*. (2010) Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. *Proc. Natl. Acad. Sci.* **107**, 1029–1034.

Gaines, T.A., Duke, S.O., Morran, S., Rigon, C.A.G., Tranel, P.J., Küpper, A. and Dayan, F.E. (2020) Mechanisms of evolved herbicide resistance. *J. Biol. Chem.* **295**, 10307–10330.

Garrison, E. and Marth, G. (2012) *Haplotype-based variant detection from short-read sequencing*. arXiv [q-bio.GN].

Golden, S.S. and Haselkorn, R. (1985) Mutation to herbicide resistance maps within the psbA gene of Anacystis nidulans R2. *Science* **229**, 1104–1107.

Gronwald, J.W. (1997) Resistance to PS II inhibitor herbicides. In *Weed and Crop Resistance to Herbicides*(De Prado, R., Jorŕn, J. and García-Torres, L., eds), pp. 53–59. Dordrecht: Springer Netherlands.

Haller, B.C. and Messer, P.W. (2019) Slim 3: Forward genetic simulations beyond the Wright-Fisher Model. *Mol. Biol. Evol.* **36**, 632–637.

Hashim, S., Jan, A., Sunohara, Y., Hachinohe, M., Ohdan, H. and Matsumoto, H. (2012) Mutation of alpha-tubulin genes in trifluralin-resistant water foxtail (*Alopecurus aequalis*). *Pest Manag. Sci.* **68**, 422–429.

Hawkins, N.J., Bass, C., Dixon, A. and Neve, P. (2018) The evolutionary origins of pesticide resistance. *Biol. Rev. Camb. Philos. Soc.* **94**, 135–155.

Heap, I. (2014a) Global perspective of herbicide-resistant weeds. *Pest Manag. Sci.* **70**, 1306–1315.

Heap, I. (2014b) Herbicide resistant weeds. In *Integrated Pest Management: Pesticide Problems*, Vol. **3** (Pimentel, D. and Peshin, R., eds), pp. 281–301. Dordrecht: Springer Netherlands.

Hermisson, J. and Pennings, P.S. (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352.

Hermisson, J. and Pennings, P.S. (2017) Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**, 700–716.

Hess, M., Herrmann, J., Bollmann, N. and Wagner, J. (2022) Results of a weed monitoring 2019-2021: Development of infestation and resistance situation. *Jul.-Kühn-Archiv.* **468**, 199–205.

Incledon, B.J. and Hall, J.C. (1997) Acetyl-coenzyme A carboxylase: Quaternary structure and inhibition by graminicidal herbicides. *Pestic. Biochem. Physiol.* **57**, 255–271.

Ireland, C.R., Telfer, A., Covello, P.S., Baker, N.R. and Barber, J. (1988) Studies on the limitations to photosynthesis in leaves of the atrazine-resistant mutant of *Senecio vulgaris* L. *Planta* **173**, 459–467.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

Kaundun, S.S. (2014) Resistance to acetyl-CoA carboxylase-inhibiting herbicides. *Pest Manag. Sci.* **70**, 1405–1417.

Kersten, S., Chang, J., Huber, C.D., Voichek, Y., Lanz, C., Hagmaier, T. *et al*. (2023) Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass. *Proc. Natl. Acad. Sci.* https://doi.org/10.1073/pnas.2206808120

Knaus, B.J. and Grünwald, N.J. (2017) vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53.

Korlach, J. (2013) *Understanding Accuracy in SMRT Sequencing*.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B. and Stamatakis, A. (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455.

Kreiner, J.M., Tranel, P.J., Weigel, D., Stinchcombe, J.R. and Wright, S.I. (2021) The genetic architecture and population genomic signatures of glyphosate resistance in *Amaranthus tuberculatus*. *Mol. Ecol.* **30**, 5373–5389.

Kreiner, J.M., Sandler, G., Stern, A.J., Tranel, P.J., Weigel, D., Stinchcombe, J.R. and Wright, S.I. (2022) Repeated origins, widespread gene flow, and allelic interactions of target-site herbicide resistance mutations. *eLife* **11**, e70242.

Kronenberg, Z., Töpfer, A. and Harting, J. (2021) *pbaa: PacBio Amplicon Analysis [version 1.0.0]*. GitHub. Available online: https://github.com/PacificBiosciences/pbAA

Leigh, J.W. and Bryant, D. (2015) popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116.

Li, H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Lischer, H.E.L. and Excoffier, L. (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299.

Lutman, P.J.W., Moss, S.R., Cook, S. and Welham, S.J. (2013) A review of the effects of crop agronomy on the management of*Alopecurus myosuroides*. *Weed Res.* **53**, 299–313.

Menchari, Y., Camilleri, C., Michel, S., Brunel, D., Dessaint, F., Le Corre, V. and Délye, C. (2006) Weed response to herbicides: regional-scale distribution of herbicide resistance alleles in the grass weed *Alopecurus myosuroides*. *New Phytol.* **171**, 861–873.

Menchari, Y., Délye, C. and Le Corre, V. (2007) Genetic variation and population structure in black-grass (*Alopecurus myosuroides* Huds.), a successful, herbicide-resistant, annual grass weed of winter cereal fields. *Mol. Ecol.* **16**, 3161–3172.

Menchari, Y., Chauvel, B., Darmency, H. and Délye, C. (2008) Fitness costs associated with three mutant acetyl-coenzyme A carboxylase alleles endowing herbicide resistance in black-grass *Alopecurus myosuroides*: Fitness cost in ACCase-resistant black-grass. *J. Appl. Ecol.* **45**, 939–947.

Moss, S.R., Perryman, S.A.M. and Tatnell, L.V. (2007) Managing herbicide-resistant blackgrass (*Alopecurus myosuroides*): theory and practice. *Weed Technol.* **21**, 300–309.

Ort, D.R., Ahrens, W.H., Martin, B. and Stoller, E.W. (1983) Comparison of photosynthetic performance in triazine-resistant and susceptible biotypes of *Amaranthus hybridus*. *Plant Physiol.* **72**, 925–930.

Petit, C., Bay, G., Pernin, F. and Délye, C. (2010) Prevalence of cross- or multiple resistance to the acetyl-coenzyme A carboxylase inhibitors fenoxaprop, clodinafop and pinoxaden in black-grass (*Alopecurus myosuroides* Huds.) in France. *Pest Manag. Sci.* **66**, 168–177.

Powles, S.B. and Yu, Q. (2010) Evolution in action: plants resistant to herbicides. *Annu. Rev. Plant Biol.* **61**, 317–347.

Preston, C. and Powles, S.B. (2002) Evolution of herbicide resistance in weeds: initial frequency of target site-based resistance to acetolactate synthase-inhibiting herbicides in Lolium rigidum. *Heredity* **88**, 8–13.

Purrington, C.B. (2000) Costs of resistance. *Curr. Opin. Plant Biol.* **3**, 305–308.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26.

Rosenhauer, M., Jaser, B., Felsenstein, F.G. and Petersen, J. (2013) Development of target-site resistance (TSR) in *Alopecurus myosuroides* in Germany between 2004 and 2012. *J. Plant Dis. Prot.* **120**, 179–187.

Sabet Zangeneh, H., Mohammaddust Chamanabad, H.R., Zand, E., Asghari, A., Alamisaeid, K., Travlos, I.S. and Alebrahim, M.T. (2016) Study of fitness cost in three rigid ryegrass populations susceptible and resistant to acetyl-CoA carboxylase inhibiting herbicides. *Front. Ecol. Evol.* **4**.

Sammons, R.D. and Gaines, T.A. (2014) Glyphosate resistance: state of knowledge. *Pest Manag. Sci.* **70**, 1367–1377.

Schlipalius, D.I., Tuck, A.G., Pavic, H., Daglish, G.J., Nayak, M.K. and Ebert, P.R. (2019) A high-throughput system used to determine frequency and distribution of phosphine resistance across large geographical regions. *Pest Manag. Sci.* **75**, 1091–1098.

Schlötterer, C., Tobler, R., Kofler, R. and Nolte, V. (2014) Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**, 749–763.

Somerville, G.J., Melander, B., Kudsk, P. and Mathiassen, S.K. (2019) Modelling annual grass weed seed dispersal in winter wheat, when influenced by hedges and directional wind. *Ecol. Model.* **410**, 108729.

Tardif, F.J., Rajcan, I. and Costea, M. (2006) A mutation in the herbicide target site acetohydroxyacid synthase produces morphological and structural alterations and reduces fitness in *Amaranthus powellii*. *New Phytol.* **169**, 251–264.

Team, R.C. (2018) *R: a language and environment for statistical computing computer program, version 3.5. 0*. Vienna, Austria: R Foundation for Statistical Computing.

Tranel, P.J. and Wright, T.R. (2002) Resistance of weeds to ALS-inhibiting herbicides: what have we learned? *Weed Sci.* **50**, 700–712.

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S. and Turner, S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A. *et al*. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33.

Van Etten, M., Lee, K.M., Chang, S.-M. and Baucom, R.S. (2020) Parallel and nonparallel genomic responses contribute to herbicide resistance in Ipomoea purpurea, a common agricultural weed. *PLoS Genet.* **16**, e1008593.

Varah, A., Ahodo, K., Coutts, S.R., Hicks, H.L., Comont, D., Crook, L. *et al*. (2019) The costs of human-induced evolution in an agricultural system. *Nat. Sustain.* **3**, 63–71.

Vila-Aiub, M.M., Neve, P. and Powles, S.B. (2009) Fitness costs associated with evolved herbicide resistance alleles in plants. *New Phytol.* **184**, 751–767.

Vila-Aiub, M.M., Yu, Q., Han, H. and Powles, S.B. (2015) Effect of herbicide resistance endowing Ile-1781-Leu and Asp-2078-Gly ACCase gene mutations on ACCase kinetics and growth traits in Lolium rigidum. *J. Exp. Bot.* **66**, 4711–4718.

Walker, K.A., Ridley, S.M., Lewis, T. and Harwood, J.L. (1988) Fluazifop, a grass-selective herbicide which inhibits acetyl-CoA carboxylase in sensitive plant species. *Biochem. J.* **254**, 307–310.

Wang, T., Picard, J.C., Tian, X. and Darmency, H. (2010) A herbicide-resistant ACCase 1781 Setaria mutant shows higher fitness than wild type. *Heredity* **105**, 394–400.

Wang, L.-G., Lam, T.T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T. *et al*. (2020) Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T. *et al*. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162.

Wickham, H. (2007) Reshaping Data with the reshape Package. *J. Stat. Softw.* **21**, 1–20.

Wickham, H. (2011) The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R. *et al*. (2019) Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686.

Wickham, H., François, R., Henry, L., and Müller, K. (2020) *dplyr: A Grammar of Data Manipulation [R package dplyr version 1.0.2]*. Available online: https://dplyr.tidyverse.org

Yamamoto, E., Zeng, L. and Baird, W.V. (1998) Alpha-tubulin missense mutations correlate with antimicrotubule drug resistance in *Eleusine indica*. *Plant Cell* **10**, 297–308.

Yang, N., Xu, X.-W., Wang, R.-R., Peng, W.-L., Cai, L., Song, J.-M. *et al*. (2017) Contributions of Zea mays subspecies mexicana haplotypes to modern maize. *Nat. Commun.* **8**, 1874.

Yu, Q., Collavo, A., Zheng, M.-Q., Owen, M., Sattin, M. and Powles, S.B. (2007) Diversity of acetyl-coenzyme A carboxylase mutations in resistant Lolium populations: evaluation using clethodim. *Plant Physiol.* **145**, 547–558.

Yu, G., Smith, D.K., Zhu, H., Guan, Y. and Lam, T.T. (2017) Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1 Phenotyping of German populations, target-specific primers with barcode sequences attached, and barcodes to sample correspondence.
Figure S1 Insert size distribution of the PacBio amplicon library.
Figure S2 Correlation between allele frequencies and haplotype frequencies for TSR amino acid positions Trp1999, Ile2041 and Asp2078.
Figure S3 Correlations between TSR haplotype frequencies and phenotyping with ACCase inhibitors.
Table S1 R-packages used for data manipulation and visualization.

4.2 Supplementary

Supporting Information for

Deep haplotype analyses of target-site resistance locus *ACCase* in blackgrass enabled by pool-based amplicon sequencing

Sonja Kersten[1,2], Fernando A. Rabanal[2,*], Johannes Herrmann[3], Martin Hess[3], Zev N. Kronenberg[4], Karl Schmid[1], Detlef Weigel[2,*]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany.
[2]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen, Germany.
[3]Agris42 GmbH, Stuttgart, Germany.
[4]Pacific Biosciences, Menlo Park, CA, USA.

Authors for correspondence:
Fernando A. Rabanal: fernando.rabanal@tue.mpg.de
Detlef Weigel: weigel@weigelworld.org

This PDF file includes:
Figures S1 to S3
Table S1

Other supporting materials for this manuscript include the following:
Data S1
GitHub repository (https://doi.org/10.5281/zenodo.7646820)

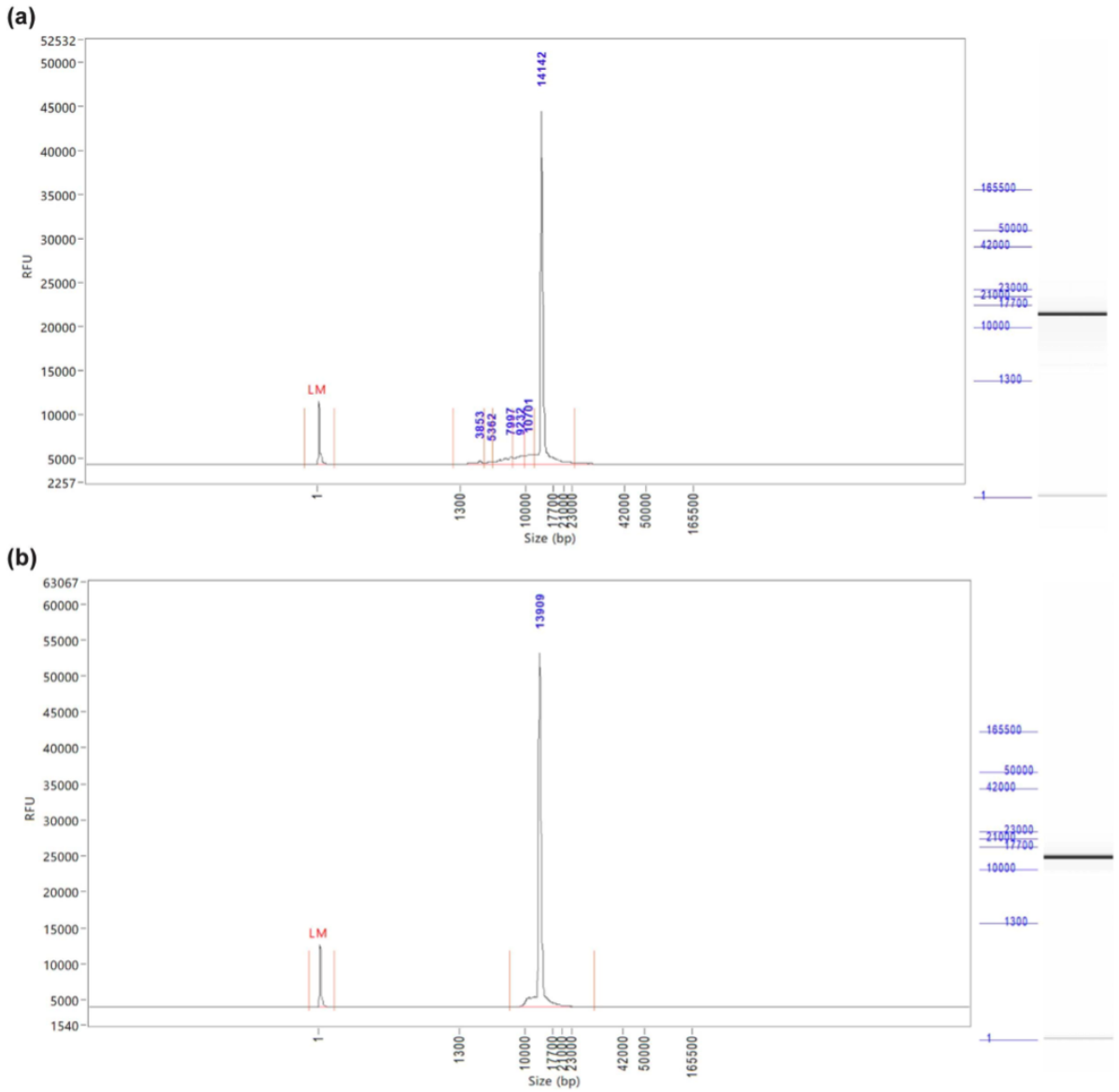


Figure S1. Insert size distribution of the PacBio amplicon library. **(a)** Before and **(b)** after size-selection on the BluePippin instrument as measured on a Femto Pulse System. Only fragments larger than 10 kb were retained.

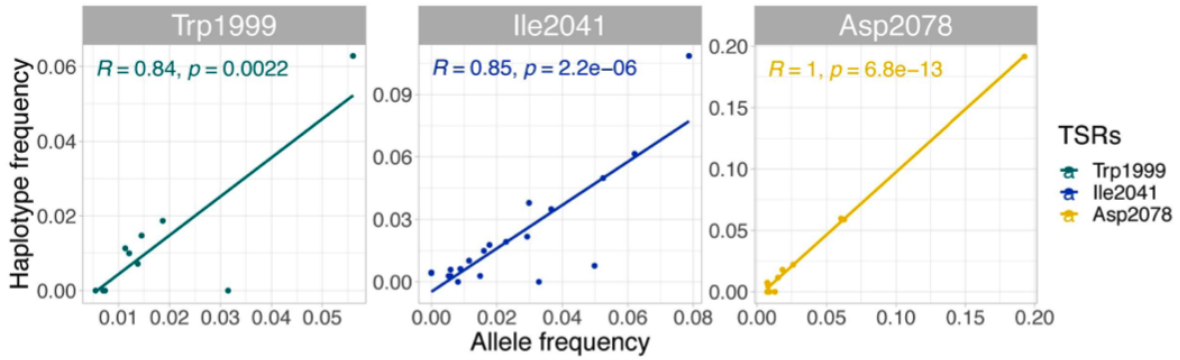


Figure S2. Correlation between allele frequencies and haplotype frequencies for TSR amino acid positions Trp1999, Ile2041 and Asp2078. Correlation coefficients and p-values are shown separately in each TSR panel.

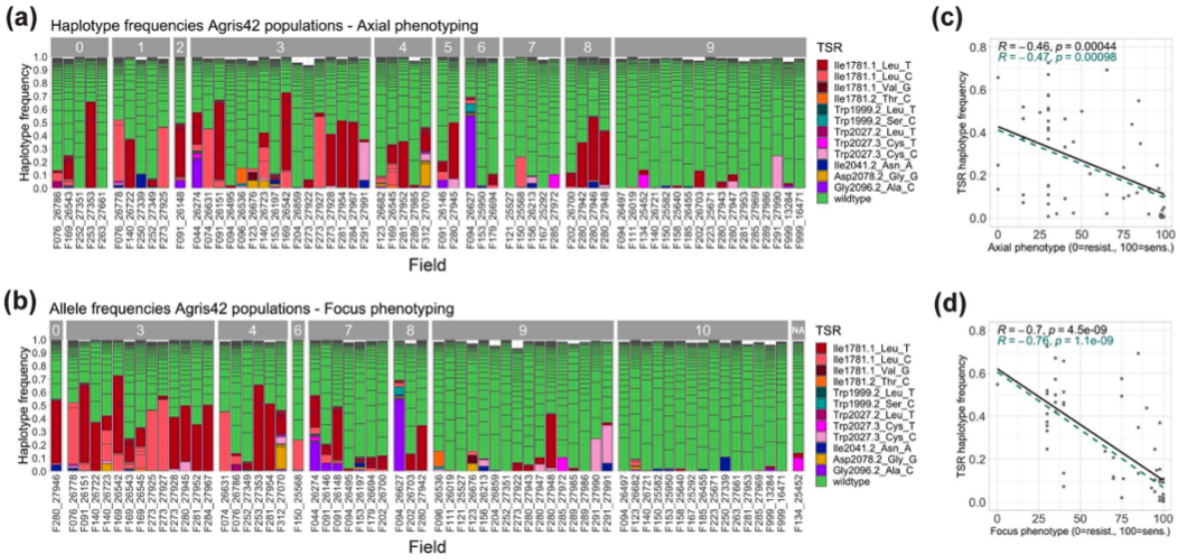


Figure S3. Correlations between TSR haplotype frequencies and phenotyping with ACCase inhibitors. Haplotype frequencies were inferred using *pbaa*. Colors refer to TSR and wild-type haplotypes. **a.** Bins represent the remaining efficiencies of the herbicide Axial® (Bin 0: 0 to 10% efficiency, which means 90–100% survivor plants; Bin 9: 90 to 99% efficiency). Correlation coefficients and p-values for TSR haplotypes and their respective phenotypes are shown in the panel on the right. **b.** Bins represent remaining efficiencies of the herbicide Focus Ultra (Bin 0: 0 to 10% efficiency; Bin 9: 90 to 99% efficiency; Bin 10: 100% efficiency with 0% survivor plants). **c,d.** Correlation coefficients and p-values for TSR haplotypes and their respective phenotypes are shown in the corresponding panels on the right. Black line shows the correlation of all TSR mutations, the green line only the TSR mutations with reported resistance to the respective herbicides (summarized in Table 3 of Powles and Yu, 2010).

Table S1. R-packages used for data manipulation and visualization.

Package name and version	Reference
dplyr 1.0.2	(Wickham *et al.*, 2020)
ggplot 3.3.2	(Wickham, 2016)
ggpubr 0.4.0	Kassambara, 2020 (ref (https://github.com/kassambara/ggpubr/)
ggtree 1.16.6	(Yu *et al.*, 2017)
haplotypes 1.1.2	Aktas, 2020 (ref https://cran.r-project.org/web/packages/haplotypes/haplotypes.pdf)
plyr 1.8.6	(Wickham, 2011)
rcompanion 2.4.1	Mangiafico, 2016 (https://rcompanion.org/handbook)
reshape 0.8.8	(Wickham, 2007)
tidyr 1.1.2	Wickham, 2020 (ref (https://github.com/tidyverse/tidyr/)
tidyverse 1.3.0	(Wickham *et al.*, 2019)
treeio 1.18.1	(Wang *et al.*, 2020)
vcfR 1.12.0	(Knaus and Grünwald, 2017)

5. General Discussion

5.1 Target-site resistances arise from soft selective sweeps

Resistance adaptation to herbicide applications can leave characteristic footprints of selection in the genome called selective sweeps. There are two types of sweeps, depending on the availability of beneficial alleles. *Hard sweeps* are the increase in frequency of a beneficial allele that arose in a single ancestor. *Soft sweeps* result from the parallel rise of several beneficial alleles that arose in multiple ancestors in a population (Hermisson and Pennings 2017). If a beneficial allele emerges after the onset of selection pressure, it is more likely to manifest as a hard sweep (Messer and Petrov 2013). Soft sweeps constitute the predominant type of adaptation in many species and originate from the supply of genetic variation either from standing genetic variation or from recurrent *de novo* mutations. They can arise either from independent origins of multiple alleles or from recombination that adds the beneficial mutation to different genetic backgrounds (Hawkins et al. 2018). Conditions that favor soft sweeps are large population size, large mutational targets, and the waiting time for a new adaptive allele being shorter than for an existing beneficial allele to sweep through the population (Messer and Petrov 2013). In general, soft sweeps are associated with less loss of genetic variation as several different haplotypes simultaneously rise in frequency and are present in intermediate frequencies (Hermisson and Pennings 2017).

In my study of *A. myosuroides* (Chapter 3), I analyzed haplotypes for the two TSR loci *ACCase* and *ALS* determined by highly accurate circular consensus (CCS) long-read amplicons encompassing the complete genes (~13.2 kb for *ACCase* and ~3.6 kb for *ALS)*. My haplotype networks and trees showed different complexity, quite likely reflecting the different selection pressures through herbicides to which they have been exposed. Overall, I find high haplotype diversity in my populations, even though they have been under selection pressure and have developed resistance. Furthermore, I showed that TSR mutations evolved independently from multiple different origins, characteristic of soft sweeps. This is consistent with the theory that large mutational target sites such as *ACCase* and *ALS*, each with at least seven known amino acid positions that when mutated confer resistance, can adapt very quickly through soft sweeps, thus maintaining genetic diversity.

5.2 Resistance adaptation from standing genetic variation versus *de novo* mutations

Regardless of the nature of a sweep, one of the key questions about the emergence of herbicide resistance is its evolutionary origin, specifically, whether alleles that confer resistance arise from standing genetic variation or *de novo* mutations. In line with convention, I define standing genetic variation as mutations that were present in field populations before selection began, whereas *de novo* mutations occurred more recently, when selection pressure already existed. The factors that influence the type and speed of resistance adaptation in natural weed populations are effective population size, mutation rate, selection coefficients of resistance mutations, mutational target site, strength of selection, weed species, and gene flow (Kreiner et al. 2018; Heap 2014a). Depending on the biological, genetic, and ecological characteristics of the weed populations, combinations of these factors are more likely to lead to adaptation through standing genetic variation or *de novo* mutations (Hermisson and Pennings 2005).

Strong selection pressure and high positive selection coefficients support the establishment of *de novo* mutations, since they face a high risk of being immediately lost through genetic drift (Hawkins et al. 2018; Hermisson and Pennings 2005). There is a waiting time for *de novo* mutations, which is associated with a delay in resistance adaptation compared to preexisting variation (Orr and Unckless 2014). Therefore, *de novo* mutations have a greater impact on genetic diversity because the waiting period reduces the minimum population size more severely, so populations that adapt through *de novo* mutations often suffer a greater loss of genetic diversity than when there is standing genetic variation to select on (Orr and Unckless 2014). In general, *de novo* mutations are more likely when the mutation rate of the species is high and there has been a massive recent population expansion. This is the case, for example, in the insect species *Drosophila melanogaster* due to seasonal fluctuations that led to repeated and independent insecticide resistance even within the same continent (Karasov et al. 2010). In weeds, this scenario is more rare, likely due to the persistent seed banks in the soil that contribute to the genetic variation of a field population. These seed banks often contain beneficial adaptive alleles that predate herbicide selection, thus the probability of adaptation from a new *de novo* mutation is much lower. However, Kreiner et al. (2022) found evidence in *A. tuberculatus* of several parallel resistance mutations arising from a recent species expansion and originating from *de novo* mutations.

Adaptation from standing genetic variation is likely for species with large effective population sizes, because neutral and even deleterious mutations exist most of the time as permanent genetic adaptive variation under mutation-selection-drift balance (Messer and Petrov 2013). Polymorphisms from standing genetic variation can be present in a single gene or as a polygenic trait (Pritchard and Di Rienzo 2010). The speed of adaptation is faster from standing genetic variation, because the beneficial allele is already present in several copies (Hermisson and Pennings 2005). Furthermore, there is a high probability for soft sweeps, because there may already be several different adaptive alleles that rise in parallel in frequency under favorable conditions (Hermisson and Pennings 2005). Since mutations from standing genetic variation had more time to establish in a population, rarer changes can also emerge and more complex traits involving multiple mutations (Matuszewski et al. 2015). Under low mutation rates, adaptation from standing genetic variation is favored. High mutation rates, on the other hand, favor both that a new allele arises *de novo*, but also that it is already present as standing genetic variation (Hermisson and Pennings 2005). In the latter case, the fitness cost of the mutation and the strength of selection have a major influence on the type of adaptation (Messer and Petrov 2013).

Most problematic weed species with herbicide resistance, such as *Amaranthus* spp., *Lolium rigidum*, *Conyza canadensis* and *Avena fatua*, have large census population sizes and a high genetic diversity, which corresponds to large effective population sizes (Heap 2014; reviewed in Kreiner et al. 2018). This suggests that the predominant type of herbicide resistance adaptation in weeds comes from standing genetic variation – as long as the fitness costs for resistance mutations are not very high. This can be the case for both large-effect TSR mutations, as well as for NTSR, which arises from several small effect genes. Further evidence for pre-existing adaptive variation has been provided by several studies in which herbicide resistances were found prior to the widespread introduction of herbicides (Délye et al. 2013; Preston and Powles 2002; Baucom and Mauricio 2010).

Since many different factors are involved in the evolution of herbicide resistance, simulation of populations and selection is an important approach for understanding the interaction of these factors and predicting resistance development, especially since empirical data are often limited in size and completeness and do not include as many factor combinations, simulations are often compared to empirical data to confirm conclusions. This is because large-scale and long term experiments are expensive or unfeasible. Simulations can deal with large population sizes and model also aspects that are difficult to measure such as rare alleles and seed banks in soil (Renton et al. 2014). At the same time, the limits of modeling

need to be known to interpret the results with appropriate care. The most important aspect is that the initial parameters assumptions such as mutation rate, recombination rate and fitness coefficients are essential for modeling and may be difficult to determine, especially in non-model organisms.

In my population study in *A. myosuroides*, I found high TSR diversity at the *ACCase* and *ALS* loci (Chapter 2). To infer the evolutionary origin of these resistance mutations, I generated simulation models based on the *ACCase* gene locus and its corresponding TSRs. I then compared the TSR diversity of my empirical data from 47 populations across Europe with the simulated data to determine whether the observed TSR complexity can be explained by standing genetic variation or recent *de novo* mutations. For the initial modeling parameters, I used my empirical data and genome-wide ddRADseq markers to determine the effective population size. I inferred a maximum effective population size of 42,000 individuals, which is likely an underestimation given farmer reports of field infestations and the uncertainties inherent from deriving past effective population sizes for fluctuating populations under herbicide selection (Messer and Petrov 2013). I therefore ran my simulations in addition with twice the effective population size. Mutation rate and recombination rate was adapted from maize (Yang et al. 2017; Bauer et al. 2013). TSR mutations were considered to be dominant and highly beneficial under herbicide selection pressure. I built my models using the forward simulation software SLiM, which is based on individuals and can account for extinction of single individuals without adaptive TSR mutation under herbicide selection.

In my simulations, I found that the observed TSR diversity can be better explained by standing genetic variation in *A. myosuroides*. I nevertheless note that while in my empirical data, adaptation seems to have occurred mainly through soft sweeps, the majority of simulated *de novo* mutations manifested as hard sweeps, which suggests that some of the initial parameters may need further adjustment (see below). Given the rapid rate of resistance development in agricultural fields, resistance development through standing genetic variation is more likely because the mutations are already established in the population. Although I would not exclude *de novo* mutations as a source of adaptive variation in *A. myosuroides*, I conclude that most TSR mutations pre-exist as standing genetic variation.

To better adapt the simulations to weed populations that have low mobility, future simulation models could be further optimized to take into account extended parameters like spatial

distribution, soil seed banks, bottle neck and off-year scenarios. On the practical side, detailed simulations of climatic conditions, genetics, environment and farming practices can help to predict resistance development and decide on future farming management strategies. In addition, more comprehensive sequencing data such as WGS can provide new analysis possibilities to trace population histories using coalescent simulations and to track migration events. WGS sequencing, in combination with detailed metadata about location and farming practices, can be also used to investigate the extent of parallel resistance evolution as well as long and short distance gene flow. Unfortunately, one of main limitations of this study was the lack of metadata such as geographic location, herbicide treatments, crop cycle, and so on. Also, due to the large genome size of *A. myosuroides* (3.6 Gb), WGS sequencing is very expensive for a high number of individuals, therefore I chose ddRADseq as a reduced representation method. This comes with limitations like not being able to apply advanced statistics for gene flow inference or ancestral recombination graphs that are based on linkage information and phased haplotypes (Speidel et al. 2019; Kreiner et al. 2022).

5.3 Resistance management strategies for *A. myosuroides* and future perspectives

Weed management is essential for agricultural productivity and, together with genetic and biological factors, has a major influence on the rate of resistance development. There are three main trends in agricultural practices that have favored the emergence of *A. myosuroides* plants in agricultural fields, contributing to a rapid spread of resistance in recent decades: the prevalence of winter cereals in crop rotations, earlier sowing in fall and minimum tillage instead of plowing (Moss 2017).

Nowadays, in the context of integrated weed management (IWM) strategies, a combination of chemical and non-chemical weed control measures is recommended to reduce the reliance on herbicides (Shaner and Beckie 2014). Non chemical measures include crop rotations, increased cultivation of spring crops, fallow years, delayed autumn drilling, rotational plowing, competitive cultivars, higher sowing rates and minimizing the risk of seed dispersal (Lutman et al. 2013; Moss 2017). Regarding herbicide applications, a correct timing is crucial, as well as the alternation of the MoAs. Patch spraying and application of herbicide mixtures can further reduce the risk for resistance to emerge, though not entirely prevent it (Beckie 2006). Notably, in a recent study, it was found that mixtures can also lead to a more ‘generalist’ type of resistance, meaning that several MoAs lose their efficacy (Comont et al. 2020). Resistance monitoring can therefore make a major contribution to the

early detection and prevention of resistance (Comont and Neve 2021). Although IWM strategies have long been promoted, implementation in practice has been rather hesitant and has only occurred in recent years out of increasing necessity (Shaner and Beckie 2014). This is primarily due to the measures being usually more complex, time-consuming, less effective, and in some cases even more expensive (Hurley and Frisvold 2016).

There is only a limited number of possible herbicide target sites and it is unclear whether more MoAs even exist, therefore other future directions need to be taken (Duke 2012). Natural phytotoxins have been investigated, but have not proven suitable, either because they have insufficient physicochemical properties for the application in natural environments or the uptake and translocation in plants (reviewed in Dayan and Duke 2014). Nevertheless, they could potentially provide clues to new MoAs. An exciting new direction is the study of natural by-products of microorganisms or plant extracts, but so far the development has not yet reached market maturity (Westwood et al. 2018).

Biological weed control based on arthropods and parasites has been shown to be a functional alternative approach in some species, though never became a widespread application in agricultural fields, and rather has served to control invasive weeds in natural ecosystems (species list: Winston 2014). However, in the case of microbes, some agents could be registered as herbicides, commonly referred to as ‘bioherbicides’ (reviewed in Westwood et al. 2018). Unfortunately, they tend to be unreliable due to variations in performance in different environments and under different climatic conditions, but there is much potential left to be discovered in the coming decades (Westwood et al. 2018).

Another new control method under research is RNA interference to silence key genes of weeds. These RNA sequences can be highly specific, but there are currently major technical issues, such as the large-scale production of RNAs and the efficient uptake and transportation into the plant (Green 2014). Other approaches exist that involve genetic engineering of the plant or weed, such as the "gene drive" technology. These modifications are now possible through the widespread use of Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9, but the implementation is still uncertain due to ethical concerns and the legal hurdles for release of organisms generated by CRISPR/Cas9 in many countries (Confédération paysanne and Others 2018; Barrett et al. 2019).

Robotics and machine learning are new technological approaches for weed control and are currently experiencing an enormous boom. The limiting factor at the moment is the sensor

technology, which still requires improvements in visual weed detection. Promising approaches are hyperspectral imaging and machine learning algorithms to classify plant images (reviewed in Su 2020).

The recent trend of collecting “big data” in life sciences can be also applied to the collection of parameters regarding weed infestations and resistances. This provides valuable information for future research as well as for the development of advising software and monitoring apps (Westwood et al. 2018).

5.4 The shift in technology: Targeted long-read sequencing

Long reads are on their way to become the state of the art also in population resequencing. Not only do they represent the state of the art for *de novo* reference genome assembly, but they also enable new findings based on a range of other applications such as long-read RNA (IsoSeq) and long-read amplicon sequencing (reviewed in Pollard et al. 2018).

Many groundbreaking discoveries of the past decades were made using Illumina short-read data (Sudmant et al. 2015; Sudmant et al. 2010; Becker et al. 2011; Schmitz et al. 2011; 1001 Genomes Consortium. 2016; Cagan et al. 2022). Although Illumina data can be generated at high throughput, provide cost-effectiveness and high accuracy (above 99.9%), the read length (up to 300 bp) limits the resolution of complex and repetitive genomic regions. Due to the limitation of aligning short reads to a single reference genomes, not all structural variation such as indels and deletions affecting more than 50 bp, can be detected (Logsdon et al. 2020). However, long-read sequencing has revolutionized the field by enabling the characterization of large structural variants and the analysis of complete haplotypes instead of single SNPs (Bayer et al. 2020).

The two current major sequencing technologies for long-reads are Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). ONT sequencing can be achieved from a pocket-size device and is based on linear DNA molecules that pass through nanopores that are embedded in a polymer membrane (Jain et al. 2016). A flow cell contains hundreds to thousands of these pores surrounded by an electrolytic solution that produces an electric current inside the nanopores. Motor proteins bound to the adapters guide the molecules through the pores in a controlled way. Since different bases cause specific changes in the electric current of the pore, these can be detected and decoded (Deamer et al. 2016). While ONT has the potential to produce reads longer than 1 Mb, it comes with lower per base accuracies of 87-98% (Logsdon et al. 2020). The single-molecule real-time

(SMRT) sequencing technology of PacBio uses zero-mode waveguides for massive parallelisation. The DNA molecule is circularized by ligation of hairpin adapters. Each of the eight million zero-mode waveguides in the current version of the platform contains a single active DNA polymerase that processes around the molecule and incorporates new fluorescently labeled nucleotides. A laser excites the signal, which is then detected by a camera, and the fluorophore is cleaved off again. This procedure is repeated thousands of times as long as the polymerase is active (Eid et al. 2009). Two different run modes are available on the PacBio Sequel instruments: Continuous long reads (CLR) and Circular consensus reads (CCS), also known as HiFi reads. CLRs maximize insert size (> 35 kb) at the cost of read accuracy (87-92%) (Logsdon et al. 2020). CCS uses circular semi-long reads (15 - 20 kb) through which the polymerase circulates several times. The more often the polymerase passes, the better the random sequencing errors can be corrected by the generation of the consensus from the subreads, thus increasing read accuracy to more than 99% (Wenger et al. 2019).

In my study (Chapter 2), I showed that with PacBio long-read amplicon sequencing it is possible to retrieve complete haplotypes of *ACCase* (13.2 kb) and *ALS* (3.5 kb) gene sequences in CCS mode in individuals. Additionally, this technology overcame the limitations of the also highly accurate Sanger sequencing employed in previous studies (Délye et al. 2004a; Délye et al. 2004b; Menchari et al. 2006), where individual overlapping sequencing products had to be aligned to each other and, in the case of heterozygosity, the individual amplified fragments had to be cloned into separate plasmids to be properly phased. To infer haplotypes, I used the new clustering software pacbio amplicon analysis (pbaa) for a reference-free clustering approach of the q20 CCS reads (Kronenberg et al. 2021). The accurate haplotype inference achieved in individuals motivated us to extend the advantages of this method to pools of 150 - 200 individuals (Chapter 3). In the context of field monitoring for herbicide resistance, this allows for a more cost-effective, high-throughput analytical method for thousands of samples. I compared haplotypes retrieved from individuals to a pool of 200 individuals from the same population. Except for a few low-frequency haplotypes, I recovered all individual haplotypes in the pools. Next, I applied the method to a Germany-wide dataset and compared a classical SNP calling approach to the new haplotype clustering performed by pbaa. In addition to simple allele frequencies as with classical SNP calling, I was able to obtain TSR mutations and wild types with their full haplotype context in the clustering approach. Notably, I could retrieve TSR and wildtype haplotypes that differed only in a single mutation.

Long-read sequencing technologies continuously experience rapid improvements, and targeted sequencing of relevant genetic regions is of major interest, particularly for clinical applications (Neveling et al. 2019). Through long-reads, new regions have become accessible and distinct haplotypes can be phased. However, most amplicon protocols require long-range PCRs, which can introduce amplification biases (Thompson et al. 2002). This limitation has now been overcome by a new amplification free application of PacBio: target enrichment with CRISPR/Cas9. In addition, it provides the opportunity of sequencing information in the vicinity of the region of interest beyond the limitation of a PCR product with predefined primers, which is ideal for constructing larger haplotypes. After generating a normal PacBio library from native genomic DNA, Cas9 is recruited by a complementary, short guide RNA to the targeted sequences of interest among the population of circularized SMRTbell molecules, where it causes its characteristic double-strand sequence breaks. Then, new hairpin adapters are ligated to the molecules that have been cleaved, and SMRTbell molecules that contain the region of interest are enriched on magnetic beads (Tsai et al. 2015). A similar Cas9-based enrichment strategy has also been developed by Oxford Nanopore (Oxford Nanopore Technologies 2020).

Although at the time of my study base accuracy of ONT reads was lower than that of CCS reads (Logsdon et al. 2020), Oxford Nanopore technology innovations are advancing rapidly as well. Since the accuracy highly depends on the base calling algorithm, new methods were developed that focus on single read base calling at the molecule sequencing step. In addition, new nanopores more sensitive to homopolymers have been developed, such a way that variant calls for structural variants reached 96% accuracy, with simple SNP calls being comparable to short-read data (Oxford Nanopore Technologies 2020). Compared to PacBio, Oxford Nanopore is more versatile under remote field conditions because of its size, cost and implementation of fast and simple protocols and bioinformatic pipelines (Pomerantz et al. 2022). With the innovation of ‘Adaptive sampling’ and the continuous improvements of the software controlled enrichment, it is possible to enrich genetic regions of interest in real time during the sequencing process. To do this, template sequences of the region of interest must be provided. During the processing of the first few hundred bases of a molecule, a decision is made whether the molecule matches the template sequence and sequencing continues or whether the molecule is ejected from the pore (Payne et al. 2021; Oxford Nanopore team 2018). With this strategy, successful targeted enrichment in a microbial mock community has been already demonstrated (Martin et al. 2022).

6. Concluding Remarks

The impressive adaptability of plants to different environments and selection pressures is a fascinating research field. In the present thesis, I investigated the evolutionary and genetic mechanisms leading to rapid resistance adaptation in the weed species *A. myosuroides* through repeated herbicide applications in agricultural fields. I found that herbicide resistance is largely explained by standing genetic variation, implying that there is a reservoir of adaptive alleles in weed populations that provide high potential to also overcome future MoAs. Using forward simulations, I showed that even without continued selection, TSR mutations persist in field populations for decades to centuries. These findings should be taken into consideration in weed management strategies to avoid exacerbating the resistance situation in the coming decades. The best strategy to prevent herbicide resistance in agricultural fields is to keep the size of the weed population low, which also includes the seed bank in the soil. Given the strict regulations for certain MoAs and the uncertainty of whether new MoAs will be available in the coming decades, there is an urgent need to find alternative ways to address this challenge. Fortunately, there are promising future approaches to address the problem with new technologies in sensors and robotics, and molecular weed control methods such as targeted RNA. These approaches, in combination with timely resistance monitoring through innovative long read sequencing and analysis of the genomic contexts in which resistance emerges, are crucial in the battle against resistance evolution in weeds.

7. Acknowledgments

I would like to express my gratitude to my supervisors Karl Schmid and Detlef Weigel for letting me work on this interesting topic, and for their constant and productive advice. In particular, I thank Karl for welcoming me in his lab and providing me with an inspiring work environment. I am grateful to Detlef for initiating the research on herbicide resistance in *Alopecurus myosuroides*, letting me be part of it and making the generation of the diverse population data possible. Thanks to Kelly Swarts for being part of my committee and her excellent suggestions.

Sincere thanks to my industry collaborators from BASF, Jens Lerchl, Aimone Porri and Andreas Landes, for sharing with us the European collection of *A. myosuroides* seeds, and for facilitating funding and providing valuable project ideas. Thanks to Zev Kronenberg from Pacific Biosciences for developing the clustering software Pacbio amplicon analysis (pbaa), essential for core analyses in my research, and for his patient advice on parameter optimization. Thanks to our academic collaborator from the Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium, and Christian Huber from the Department of Biology, Penn State University, USA for their contribution to one of my publications. I have greatly appreciated the fruitful discussions and it has been a pleasure to work with all of you.

Some colleagues have been particularly supportive and inspiring during my PhD. First of all, Talia Karasov and Gautam Shirsekar, who encouraged me to do my PhD. Derek Lundberg, the ultimate expert on lab techniques, gave me very valuable advice on primer design and PCR conditions. Max Collenberg and Sergio Latorre provided me a smooth start in bioinformatics analysis by supporting me with programming hints and example codes. I would also like to give a special mention to Christa Lanz from the Max Planck Genome Center. Thanks to her, we had fantastic sequencing results, and I learned a lot about next and third generation sequencing. I will make sure to always maintain her high lab standards.

Warm thanks to Fernando for very valuable project ideas and teaching me most things I know about bioinformatics. His professional working style was very inspiring for me and I also sincerely appreciate his constructive comments on this thesis. A very special thanks to Gloria, who took care of my little daughter Helena to provide me time to write the thesis. Thanks to my parents, my sister and brother and the rest of my family for their constant support and motivation and for always being there whenever I needed them.

Last but not least, I would like to thank the Max Planck Society for funding the research and the Landesgraduiertenförderung in Hohenheim for providing me a PhD scholarship.

8. Bibliography

1000 Genomes Project Consortium (2012) ‘An integrated map of genetic variation from 1,092 human genomes’, *Nature*, 491(7422), pp. 56–65.

1001 Genomes Consortium. (2016) ‘1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*’, *Cell*, 166(2), pp. 481–491.

Alcocer-Ruthling, M., Thill, D.C. and Carol Mallory-Smith (1992) ‘Monitoring the Occurrence of Sulfonylurea-Resistant Prickly Lettuce (*Lactuca serriola*)’, *Weed technology: a journal of the Weed Science Society of America*, 6(2), pp. 437–440.

Alexander, H.K. *et al.* (2014) ‘Evolutionary rescue: linking theory for conservation and medicine’, *Evolutionary applications*, 7(10), pp. 1161–1179.

Andrews, K.R. *et al.* (2016) ‘Harnessing the power of RADseq for ecological and evolutionary genomics’, *Nature reviews. Genetics*, 17(2), pp. 81–92.

Aquadro, C.F. and Greenberg, B.D. (1983) ‘Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals’, *Genetics*, 103(2), pp. 287–312.

Barrett, L.G. *et al.* (2019) ‘Gene drives in plants: opportunities and challenges for weed control and engineered resilience’, *Proceedings. Biological sciences / The Royal Society*, 286(1911), p. 20191515.

Barrett, M. (1995) ‘Metabolism of herbicides by cytochrome P450 in corn’, *Drug metabolism and drug interactions*, 12(3-4), pp. 299–315.

Baucom, R.S. (2016) ‘The remarkable repeated evolution of herbicide resistance’, *American journal of botany*, 103(2), pp. 181–183.

Baucom, R.S. and Mauricio, R. (2010) ‘Defence against the herbicide RoundUp® predates its widespread use’, *Evolutionary ecology research*, 12(1), pp. 131–141.

Bauer, E. *et al.* (2013) ‘Intraspecific variation of recombination rate in maize’, *Genome biology*, 14(9), p. R103.

Bayer, P.E. *et al.* (2020) ‘Plant pan-genomes are the new reference’, *Nature plants*, 6(8), pp. 914–920.

Becker, C. *et al.* (2011) ‘Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome’, *Nature*, 480(7376), pp. 245–249.

Beckie, H.J. (2006) ‘Herbicide-Resistant Weeds: Management Tactics and Practices1’, *Weed Technology*, 20(3), pp. 793–814.

Beckie, H.J. and Tardif, F.J. (2012) ‘Herbicide cross resistance in weeds’, *Crop protection* , 35, pp. 15–28.

Behrens, M.R. *et al.* (2007) ‘Dicamba resistance: enlarging and preserving biotechnology-based weed management strategies’, *Science*, 316(5828), pp. 1185–1188.

Bowles, D. *et al.* (2005) ‘Glycosyltransferases: managers of small molecules’, *Current opinion in plant biology*, 8(3), pp. 254–263.

Brazier, M., Cole, D.J. and Edwards, R. (2002) 'O-Glucosyltransferase activities toward phenolic natural products and xenobiotics in wheat and herbicide-resistant and herbicide-susceptible black-grass (*Alopecurus myosuroides*)', *Phytochemistry*, 59(2), pp. 149–156.

Busi, R. *et al.* (2008) 'Long distance pollen-mediated flow of herbicide resistance genes in *Lolium rigidum*', *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 117(8), pp. 1281–1290.

Busi, R. *et al.* (2018) 'Weed resistance to synthetic auxin herbicides', *Pest management science*, 74(10), pp. 2265–2276.

Cagan, A. *et al.* (2022) 'Somatic mutation rates scale with lifespan across mammals', *Nature*, 604(7906), pp. 517–524.

Cagnac, O. *et al.* (2004) 'AtOPT6 transports glutathione derivatives and is induced by primisulfuron', *Plant physiology*, 135(3), pp. 1378–1387.

Cai, L. *et al.* (2021) 'The blackgrass genome reveals patterns of divergent evolution of non-target site resistance to herbicides', *bioRxiv*. doi:10.1101/2021.12.14.472569.

Campe, R. *et al.* (2018) 'A new herbicidal site of action: Cinmethylin binds to acyl-ACP thioesterase and inhibits plant fatty acid biosynthesis', *Pesticide biochemistry and physiology*, 148, pp. 116–125.

Cavan, Biss and Moss (1998) 'Localized origins of herbicide resistance in *Alopecurus myosuroides*', *Weed research*, 38(3), pp. 239–245.

Chauvel, B. (1991) Polymorphisme génétique et sélection de la résistance aux urées substituées chez *Alopecurus myosuroides Huds*. Paris 11. Available at: https://www.theses.fr/1991PA112174.

Chauvel, B. *et al.* (2002) 'Effect of vernalization on the development and growth of *Alopecurus myosuroides*', *Weed research*, 42(2), pp. 166–175.

Chauvel, B. and Gasquez, J. (1994) 'Relationships between genetic polymorphism and herbicide resistance within *Alopecurus*', *Heredity*, 72, pp. 336–344.

Chauvel, B., Guillemin, J.P. and Letouze, A. (2005) 'Effect of intra-specific competition on development and growth of *Alopecurus myosuroides* Hudson', *European journal of agronomy: the journal of the European Society for Agronomy*, 22(3), pp. 301–308.

Clarke, J., Moss, S. and Orson, J. (2000) 'The future for grass weed management in the UK', *Pesticide Outlook*, 11(2), pp. 59–63.

Colbach, N. *et al.* (2002) 'Effect of environmental conditions on *Alopecurus myosuroides* germination. I. Effect of temperature and light', *Weed research*, 42(3), pp. 210–221.

Colbach, N. and Sache, I. (2001) 'Blackgrass (*Alopecurus myosuroides* Huds.) seed dispersal from a single plant and its consequences on weed infestation', *Ecological modelling*, 139(2), pp. 201–219.

Comont, D. *et al.* (2020) 'Evolution of generalist resistance to herbicide mixtures reveals a trade-off in resistance management', *Nature communications*, 11(1), p. 3086.

Comont, D. and Neve, P. (2021) 'Adopting epidemiological approaches for herbicide resistance monitoring and management', *Weed research*. Available at:

https://onlinelibrary.wiley.com/doi/abs/10.1111/wre.12420.

Confédération paysanne and Others (Judgment of 25 July 2018). C-528/16, EU:C:2018:583.

Cummins, I. *et al.* (2013) 'Key role for a glutathione transferase in multiple-herbicide resistance in grass weeds', *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), pp. 5812–5817.

Cummins, I., Cole, D.J. and Edwards, R. (1999) 'A role for glutathione transferases functioning as glutathione peroxidases in resistance to multiple herbicides in black-grass', *The Plant journal: for cell and molecular biology*, 18(3), pp. 285–292.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics* , 27(15), pp. 2156–2158.

Darwin, C. (1859) *The origin of species by means of natural selection*. John Murray.

Dayan, F.E. and Duke, S.O. (2014) 'Natural compounds as next-generation herbicides', *Plant physiology*, 166(3), pp. 1090–1105.

Deamer, D., Akeson, M. and Branton, D. (2016) 'Three decades of nanopore sequencing', *Nature biotechnology*, 34(5), pp. 518–524.

Délye, C., Straub, C., Matéjicek, A., *et al.* (2004) 'Multiple origins for black-grass (*Alopecurus myosuroides* Huds) target-site-based resistance to herbicides inhibiting acetyl-CoA carboxylase', *Pest management science*, 60(1), pp. 35–41.

Délye, C., Straub, C., Michel, S., *et al.* (2004) 'Nucleotide variability at the acetyl coenzyme A carboxylase gene and the signature of herbicide selection in the grass weed *Alopecurus myosuroides* (Huds.)', *Molecular biology and evolution*, 21(5), pp. 884–892.

Délye, C. *et al.* (2005) 'Molecular bases for sensitivity to acetyl-coenzyme A carboxylase inhibitors in black-grass', *Plant physiology*, 137(3), pp. 794–806.

Délye, C. (2005) 'Weed resistance to acetyl coenzyme A carboxylase inhibitors: an update', *Weed Science*, 53(5), pp. 728–746.

Délye, C. *et al.* (2007) 'Status of black grass (*Alopecurus myosuroides*) resistance to acetyl-coenzyme A carboxylase inhibitors in France', *Weed research*, 47(2), pp. 95–105.

Délye, C. *et al.* (2010) 'Geographical variation in resistance to acetyl-coenzyme A carboxylase-inhibiting herbicides across the range of the arable weed *Alopecurus myosuroides* (black-grass)', *The New phytologist*, 186(4), pp. 1005–1017.

Délye, C. *et al.* (2011) 'Non-target-site-based resistance should be the center of attention for herbicide resistance research: *Alopecurus myosuroides* as an illustration', *Weed research*, 51(5), pp. 433–437.

Délye, C. *et al.* (2013) 'A new insight into arable weed adaptive evolution: mutations endowing herbicide resistance also affect germination dynamics and seedling emergence', *Annals of botany*, 111(4), pp. 681–691.

Délye, C. (2013) 'Unravelling the genetic bases of non-target-site-based resistance (NTSR) to herbicides: a major challenge for weed science in the forthcoming decade', *Pest management science*, 69(2), pp. 176–187.

Délye, C., Deulvot, C. and Chauvel, B. (2013) 'DNA analysis of herbarium Specimens of the

grass weed *Alopecurus myosuroides* reveals herbicide resistance pre-dated herbicides', *PloS one*, 8(10), p. e75117.

Délye, C. and K Boucansaud, K. (2007) 'A molecular assay for the proactive detection of target site-based resistance to herbicides inhibiting acetolactate synthase in *Alopecurus myosuroides*', *European Weed Research Society Weed Research*, 48, pp. 97–101.

Délye, C., Matéjicek, A. and Michel, S. (2008) 'Cross-resistance patterns to ACCase-inhibiting herbicides conferred by mutant ACCase isoforms in *Alopecurus myosuroides* Huds. (black-grass), re-examined at the recommended herbicide field rate', *Pest management science*, 64(11), pp. 1179–1186.

Dillon, A. *et al.* (2016) 'Physical Mapping of Amplified Copies of the 5-Enolpyruvylshikimate-3-Phosphate Synthase Gene in Glyphosate-Resistant *Amaranthus tuberculatus*', *Plant physiology*, 173(2), pp. 1226–1234.

Dixon, A. *et al.* (2020) 'Population genomics of selectively neutral genetic structure and herbicide resistance in UK populations of *Alopecurus myosuroides*', *Pest management science* [Preprint]. doi:10.1002/ps.6174.

Dixon, D.P., Lapthorn, A. and Edwards, R. (2002) 'Plant glutathione transferases', *Genome biology*, 3(3), p. REVIEWS3004.

Dobzhansky, T. (1937) *Genetics and the origin of species*. New York, NY: Columbia University Press (Columbia University Biological Series (volume 11)).

Dücker, R., Zöllner, P., Parcharidou, E., *et al.* (2019) 'Enhanced metabolism causes reduced flufenacet sensitivity in black-grass (*Alopecurus myosuroides* Huds.) field populations', *Pest management science*, 75(11), pp. 2996–3004.

Dücker, R., Zöllner, P., Lümmen, P., *et al.* (2019) 'Glutathione transferase plays a major role in flufenacet resistance of ryegrass (*Lolium* spp.) field populations', *Pest management science*, 75(11), pp. 3084–3092.

Duke, S.O. (2012) 'Why have no new herbicide modes of action appeared in recent years?', *Pest management science*, 68(4), pp. 505–512.

Du, L. *et al.* (2019) 'Fitness costs associated with acetyl-coenzyme A carboxylase mutations endowing herbicide resistance in American sloughgrass (*Beckmannia syzigachne* Steud.)', *Ecology and evolution*, 9(4), pp. 2220–2230.

Eid, J. *et al.* (2009) 'Real-time DNA sequencing from single polymerase molecules', *Science*, 323(5910), pp. 133–138.

Evans, J. *et al.* (2018) 'Extensive Genetic Diversity is Present within North American Switchgrass Germplasm', *The plant genome*, 11(1). doi:10.3835/plantgenome2017.06.0055.

Fisher, R.A. (1930) *The genetical theory of natural selection*. Oxford, Clarendon Press, p. 302.

Food and Agriculture Organization (2019) The state of food security and nutrition in the World 2019 The state of food security and nutrition in the World 2019: safeguarding against economic slowdowns and downturns. Rome, Italy: *Food & Agriculture Organization of the United Nations (FAO)*.

Franco-Ortega, S. *et al.* (2021) 'Non-target Site Herbicide Resistance Is Conferred by Two Distinct Mechanisms in Black-Grass (Alopecurus myosuroides)', *Frontiers in plant science*,

12, p. 636652.

Gaines, T.A. *et al.* (2010) ‘Gene amplification confers glyphosate resistance in *Amaranthus palmeri*’, *Proceedings of the National Academy of Sciences*, 107(3), pp. 1029–1034.

Gaines, T.A. *et al.* (2020) ‘Mechanisms of evolved herbicide resistance’, *The Journal of biological chemistry*, 295(30), pp. 10307–10330.

Gaines, T.A. *et al.* (2021) ‘Investigating the origins and evolution of a glyphosate-resistant weed invasion in South America’, *Molecular ecology*, 30(21), pp. 5360–5372.

Gianessi, L.P. (2013) ‘The increasing importance of herbicides in worldwide crop production’, *Pest management science*, 69(10), pp. 1099–1105.

Global Biodiversity Information Facility (2021) *Alopecurus myosuroides Huds. in GBIF Secretariat*, *GBIF | Global Biodiversity Information Facility*. doi:10.15468/39omei.

Gould, F., Brown, Z.S. and Kuzma, J. (2018) ‘Wicked evolution: Can we address the sociobiological dilemma of pesticide resistance?’, *Science*, 360(6390), pp. 728–732.

Green, J.M. (2014) ‘Current state of herbicides in herbicide-resistant crops’, *Pest management science*, 70(9), pp. 1351–1357.

Gressel, J. (2011) ‘Global advances in weed management’, *The Journal of agricultural science*, 149(S1), pp. 47–53.

Gronwald, J.W. (1997) ‘Resistance to PS II Inhibitor Herbicides’, in De Prado, R., Jorrín, J., and García-Torres, L. (eds) *Weed and Crop Resistance to Herbicides*. Dordrecht: Springer Netherlands, pp. 53–59.

Haldane, J.B.S. (1932) ‘The Time of Action of Genes, and Its Bearing on some Evolutionary Problems’, *The American naturalist*, 66(702), pp. 5–24.

Hall, L.M., Moss, S.R. and Powles, S.B. (1997) ‘Mechanisms of Resistance to Aryloxyphenoxypropionate Herbicides in Two Resistant Biotypes ofAlopecurus myosuroides(blackgrass): Herbicide Metabolism as a Cross-Resistance Mechanism’, *Pesticide biochemistry and physiology*, 57(2), pp. 87–98.

Han, H. *et al.* (2017) ‘A double EPSPS gene mutation endowing glyphosate resistance shows a remarkably high resistance cost’, *Plant, cell & environment*, 40(12), pp. 3031–3042.

Harris, H. (1966) ‘Enzyme polymorphisms in man’, *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society* , 164(995), pp. 298–310.

Hatzios, K.K. and Burgos, N. (2004) ‘Metabolism-based herbicide resistance: regulation by safeners’, *Weed Science*, 52(3), pp. 454–467.

Hawkins, N.J. *et al.* (2018) ‘The evolutionary origins of pesticide resistance’, *Biological reviews of the Cambridge Philosophical Society* [Preprint]. doi:10.1111/brv.12440.

Healy-Fried, M.L. *et al.* (2007) ‘Structural basis of glyphosate tolerance resulting from mutations of Pro101 in Escherichia coli 5-enolpyruvylshikimate-3-phosphate synthase’, *The Journal of biological chemistry*, 282(45), pp. 32949–32955.

Heap, I. (2014a) ‘Global perspective of herbicide-resistant weeds’, *Pest management science*, 70(9), pp. 1306–1315.

Heap, I. (2014b) ‘Herbicide Resistant Weeds’, in Pimentel, D. and Peshin, R. (eds) *Integrated Pest Management: Pesticide Problems, Vol.3*. Dordrecht: Springer Netherlands, pp. 281–301.

Heap, I. (2022) ‘International herbicide-resistant weed database’. Available at: www.weedscience.org (Accessed: 2 May 2022).

Herbicide Resistance Action Committee (2022) Global herbicide classification lookup. Available at: https://hracglobal.com/tools/classification-lookup/.

Hermisson, J. and Pennings, P.S. (2005) ‘Soft sweeps: molecular population genetics of adaptation from standing genetic variation’, *Genetics*, 169(4), pp. 2335–2352.

Hermisson, J. and Pennings, P.S. (2017) ‘Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation’, *Methods in ecology and evolution / British Ecological Society*, 8(6), pp. 700–716.

Hicks, H.L. *et al.* (2018) ‘The factors driving evolved herbicide resistance at a national scale’, *Nature Ecology & Evolution* [Preprint]. doi:10.1038/s41559-018-0470-1.

Hilton, H.W. (1957) ‘Herbicide tolerant strains of weeds, Hawaiian Sugar Plant’, *Assoc. Annu. Rep.* [Preprint], (69).

Hurley, T.M. and Frisvold, G. (2016) ‘Economic Barriers to Herbicide-Resistance Management’, *Weed Science*, 64(sp1), pp. 585–594.

Huxley and Julian (1943) *Evolution : the modern synthesis*. New York: Harper & Brothers.

Hyde, R.J., Hallahan, D.L. and Bowyer, J.R. (1996) ‘Chlorotoluron metabolism in leaves of resistant and susceptible biotypes of the grass weed *Alopecurus myosuroides*’, *Pesticide science*, 47(2), pp. 185–190.

Incledon, B.J. and Hall, J.C. (1997) ‘Acetyl-coenzyme A carboxylase: Quaternary structure and inhibition by graminicidal herbicides’, *Pesticide biochemistry and physiology*, 57(3), pp. 255–271.

Ireland, C.R. *et al.* (1988) ‘Studies on the limitations to photosynthesis in leaves of the atrazine-resistant mutant of *Senecio vulgaris* L’, *Planta*, 173(4), pp. 459–467.

Iwakami, S. *et al.* (2014) ‘Cytochrome P450 CYP81A12 and CYP81A21 Are Associated with Resistance to Two Acetolactate Synthase Inhibitors in *Echinochloa phyllopogon*’, *Plant physiology*, 165(2), pp. 618–629.

Jain, M. *et al.* (2016) ‘The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community’, *Genome biology*, 17(1), p. 239.

Jo, J. *et al.* (2004) ‘Paraquat resistance of transgenic tobacco plants over-expressing the Ochrobactrum anthropi pqrA gene’, *Biotechnology letters*, 26(18), pp. 1391–1396.

Karasov, T., Messer, P.W. and Petrov, D.A. (2010) ‘Evidence that adaptation in *Drosophila* is not limited by mutation at single sites’, *PLoS genetics*, 6(6), p. e1000924.

Kaundun, S.S. (2014) ‘Resistance to acetyl-CoA carboxylase-inhibiting herbicides’, *Pest management science*, 70(9), pp. 1405–1417.

Kersten, S. *et al.* (2021) ‘Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass’, *bioRxiv*. doi:10.1101/2021.12.14.472587.

Kraehmer, H. *et al.* (2014) ‘Herbicides as weed control agents: state of the art: I. Weed control research and safener technology: the path to modern agriculture’, *Plant physiology*, 166(3), pp. 1119–1131.

Kraehmer, H. (2019) ‘Why have grasses become so successful?’, in *Grasses*. Chichester, UK: John Wiley & Sons, Ltd, pp. 549–554.

Kreiner, J.M. *et al.* (2019) ‘Multiple modes of convergent adaptation in the spread of glyphosate-resistant *Amaranthus tuberculatus*’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(42), pp. 21076–21084.

Kreiner, J.M. *et al.* (2021) ‘The genetic architecture and population genomic signatures of glyphosate resistance in *Amaranthus tuberculatus*’, *Molecular ecology*, 30(21), pp. 5373–5389.

Kreiner, J.M. *et al.* (2022) ‘Repeated origins, widespread gene flow, and allelic interactions of target-site herbicide resistance mutations’, *eLife*, 11. doi:10.7554/eLife.70242.

Kreiner, J.M., Stinchcombe, J.R. and Wright, S.I. (2018) ‘Population Genomics of Herbicide Resistance: Adaptation via Evolutionary Rescue’, *Annual review of plant biology*, 69, pp. 611–635.

Kreitman, M. (1983) ‘Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*’, *Nature*, 304(5925), pp. 412–417.

Kronenberg, Z., Töpfer, A. and Harting, J. (2021) ‘pbaa: PacBio Amplicon Analysis’, *GitHub* [Preprint]. GitHub. Available at: https://github.com/PacificBiosciences/pbAA.

Küpper, A. *et al.* (2018) ‘Population Genetic Structure in Glyphosate-Resistant and -Susceptible Palmer Amaranth (*Amaranthus palmeri*) Populations Using Genotyping-by-sequencing (GBS)’, *Frontiers in plant science*, 9, p. 29.

Landsteiner, K. (1900) ‘Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe’, *Centralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten*, 27, pp. 357–362.

Langley, C.H., Montgomery, E. and Quattlebaum, W.F. (1982) ‘Restriction map variation in the Adh region of *Drosophila*’, *Proceedings of the National Academy of Sciences of the United States of America*, 79(18), pp. 5631–5635.

Lang, P.L.M. *et al.* (2020) ‘Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA’, *Molecular ecology resources*, 20(5), pp. 1228–1247.

Letouzé, A. and Gasquez, J. (2003) ‘Enhanced activity of several herbicide-degrading enzymes: a suggested mechanism responsible for multiple resistance in blackgrass (*Alopecurus myosuroides* Huds.)’, *Agronomie*, 23(7), pp. 601–608.

Lewontin, R.C. and Hubby, J.L. (1966) ‘A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*’, *Genetics*, 54(2), pp. 595–609.

Li, H. (2013) ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’, *arXiv [q-bio.GN]* [Preprint]. Available at: http://arxiv.org/abs/1303.3997.

Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) ‘Long-read human genome sequencing and its applications’, *Nature reviews. Genetics*, 21(10), pp. 597–614.

Lucas, J.A., Hawkins, N.J. and Fraaije, B.A. (2015) ‘Chapter Two - The Evolution of Fungicide Resistance’, in Sariaslani, S. and Gadd, G.M. (eds) *Advances in Applied Microbiology*. Academic Press, pp. 29–92.

Lutman, P.J.W. *et al.* (2013) ‘A review of the effects of crop agronomy on the management ofAlopecurus myosuroides’, *Weed research*, 53(5), pp. 299–313.

Lu, Y.C., Zhang, S. and Yang, H. (2015) ‘Acceleration of the herbicide isoproturon degradation in wheat by glycosyltransferases and salicylic acid’, *Journal of hazardous materials*, 283, pp. 806–814.

Marshall, R. and Moss, S.R. (2008) ‘Characterisation and molecular basis of ALS inhibitor resistance in the grass weed *Alopecurus myosuroides*’, *Weed research*, 48(5), pp. 439–447.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), pp. 10–12.

Martin, S. *et al.* (2022) ‘Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples’, *Genome biology*, 23(1), p. 11.

Martin, S.L. *et al.* (2020) ‘High gene flow maintains genetic diversity following selection for high EPSPS copy number in the weed kochia (Amaranthaceae)’, *Scientific reports*, 10(1), p. 18864.

Matuszewski, S., Hermisson, J. and Kopp, M. (2015) ‘Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum’, *Genetics*, 200(4), pp. 1255–1274.

Mayr, E. (1942) *Systematics and the origin of species from the viewpoint of a zoologist*. Columbia University Press.

McCourt, J.A. *et al.* (2006) ‘Herbicide-binding sites revealed in the structure of plant acetohydroxyacid synthase’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(3), pp. 569–573.

Melander, A.L. (1914) ‘Can insects become resistant to sprays’, *Journal of economic entomology*, 7(2), pp. 167–173.

Menchari, Y. *et al.* (March, 13 2008) ‘Fitness costs associated with three mutant acetyl-coenzyme A carboxylase alleles endowing herbicide resistance in black-grass *Alopecurus myosuroides*: Fitness cost in ACCase-resistant black-grass’, *The Journal of applied ecology*, 45(3), pp. 939–947.

Menchari, Y. *et al.* (2006) ‘Weed response to herbicides: regional-scale distribution of herbicide resistance alleles in the grass weed *Alopecurus myosuroides*’, *The New phytologist*, 171(4), pp. 861–873.

Menchari, Y., Délye, C. and Le Corre, V. (2007) ‘Genetic variation and population structure in black-grass (*Alopecurus myosuroides* Huds.), a successful, herbicide-resistant, annual grass weed of winter cereal fields’, *Molecular ecology*, 16(15), pp. 3161–3172.

Mendel, G. (1866) *Versuche über Pflanzenhybriden - Experiments in Plant Hybridisation*. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, 1865 (Abhandlungen: 3–47).

Messer, P.W. and Petrov, D.A. (2013a) ‘Frequent adaptation and the McDonald-Kreitman test’, *Proceedings of the National Academy of Sciences of the United States of America*,

110(21), pp. 8615–8620.

Messer, P.W. and Petrov, D.A. (2013b) ‘Population genomics of rapid adaptation by soft selective sweeps’, *Trends in ecology & evolution*, 28(11), pp. 659–669.

Milligan, A.S. *et al.* (2001) ‘The expression of a maize glutathione S-transferase gene in transgenic wheat confers herbicide tolerance, both in planta and in vitro’, *Molecular breeding: new strategies in plant improvement*, 7(4), pp. 301–315.

Moss, S. (2017) ‘Black-grass (*Alopecurus myosuroides*): Why has this Weed become such a Problem in Western Europe and what are the Solutions?’, *Outlooks on Pest Management*, 28(5), pp. 207–212.

Moss, S.R. (1983) ‘The production and shedding of *Alopecurus myosuroides* Huds. seeds in winter cereals crops’, *Weed research*, 23(1), pp. 45–51.

Moss, S.R. (1985) ‘The survival of *Alopecurus myosuroides* Huds. seeds in soil’, *Weed research*, 25(3), pp. 201–211.

Moss, S.R. and Cussans, G.W. (1985) ‘Variability in the susceptibility of *Alopecurus myosuroides* (black-grass) to chlortoluron and isoproturon’, *Aspects of applied biology / Association of Applied Biologists* [Preprint]. Available at: https://agris.fao.org/agris-search/search.do?recordID=US201301450340.

Mota-Sanchez, D. and Wise, J.C. (2022) ‘The Arthropod Pesticide Resistance Database’. Michigan State University. Available at: http://www.pesticideresistance.org (Accessed: 2022).

Nakka, S. *et al.* (2017) ‘Rapid detoxification via glutathione S-transferase (GST) conjugation confers a high level of atrazine resistance in Palmer amaranth (*Amaranthus palmeri*)’, *Pest management science*, 73(11), pp. 2236–2243.

Naylor, R.E.L. (1972) ‘*Alopecurus Myosuroides* Huds. (A. Agrestis L.)’, *The Journal of ecology*, 60(2), pp. 611–622.

Neveling, K. *et al.* (2019) ‘The Value of Long Read Amplicon Sequencing for Clinical Applications’. Available at: https://www.pacb.com/wp-content/uploads/Aro-AGBTPH-2019-The-value-of-long-read-amplicon-sequencing-for-clinical-applications.pdf.

Oerke, E.-C. (2006) ‘Crop losses to pests’, *The Journal of agricultural science*, 144(1), pp. 31–43.

Orr, H.A. and Unckless, R.L. (2014) ‘The population genetics of evolutionary rescue’, *PLoS genetics*, 10(8), p. e1004551.

Ort, D.R. *et al.* (1983) ‘Comparison of Photosynthetic Performance in Triazine-Resistant and Susceptible Biotypes of *Amaranthus hybridus*’, *Plant physiology*, 72(4), pp. 925–930.

Oxford Nanopore team (2018) ‘Read Until: Research tool for adaptive sampling’, *GitHub* [Preprint]. GitHub. Available at: https://github.com/nanoporetech/read_until_api.

Oxford Nanopore Technologies (2020) *At NCM, announcements include single-read accuracy of 99.1% on new chemistry and sequencing a record 10 Tb in a single PromethION run*, *Nonopore Tech update*. Available at: https://nanoporetech.com/about-us/news/ncm-announcements-include-single-read-accuracy-991-new-chemistry-and-sequencing#:~:text=During%20an%20update%20at%20the,advances%20in%20its%20sequencing%20technology.

Palumbi, S.R. (2001) ‘Humans as the world’s greatest evolutionary force’, *Science*, 293(5536), pp. 1786–1790.

Payne, A. *et al.* (2021) ‘Readfish enables targeted nanopore sequencing of gigabase-sized genomes’, *Nature biotechnology*, 39(4), pp. 442–450.

Petit, C., Duhieu, B., *et al.* (2010) ‘Complex genetic control of non-target-site-based resistance to herbicides inhibiting acetyl-coenzyme A carboxylase and acetolactate-synthase in *Alopecurus myosuroides* Huds’, *Plant science: an international journal of experimental plant biology*, 178(6), pp. 501–509.

Petit, C., Bay, G., *et al.* (2010) ‘Prevalence of cross- or multiple resistance to the acetyl-coenzyme A carboxylase inhibitors fenoxaprop, clodinafop and pinoxaden in black-grass (*Alopecurus myosuroides* Huds.) in France’, *Pest management science*, 66(2), pp. 168–177.

Pollard, M.O. *et al.* (2018) ‘Long reads: their purpose and place’, *Human molecular genetics*, 27(R2), pp. R234–R241.

Pomerantz, A. *et al.* (2022) ‘Rapid in situ identification of biological specimens via DNA amplicon sequencing using miniaturized laboratory equipment’, *Nature protocols* [Preprint]. doi:10.1038/s41596-022-00682-x.

Powles, S.B. and Yu, Q. (2010) ‘Evolution in action: plants resistant to herbicides’, *Annual review of plant biology*, 61, pp. 317–347.

Preston, C. and Powles, S.B. (2002) ‘Evolution of herbicide resistance in weeds: initial frequency of target site-based resistance to acetolactate synthase-inhibiting herbicides in Lolium rigidum’, *Heredity*, 88(1), pp. 8–13.

Pritchard, J.K. and Di Rienzo, A. (2010) ‘Adaptation - not by sweeps alone’, *Nature reviews. Genetics*, 11(10), pp. 665–667.

Puritz, J.B., Hollenbeck, C.M. and Gold, J.R. (2014) ‘dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms’, *PeerJ*, 2, p. e431.

Ravet, K. *et al.* (2021) ‘Genomic-based epidemiology reveals independent origins and gene flow of glyphosate resistance in *Bassia scoparia* populations across North America’, *Molecular ecology*, 30(21), pp. 5343–5359.

Renton, M. *et al.* (2014) ‘Herbicide resistance modelling: past, present and future’, *Pest management science*, 70(9), pp. 1394–1404.

REX Consortium (2013) ‘Heterogeneity of selection and the evolution of resistance’, *Trends in ecology & evolution*, 28(2), pp. 110–118.

Riches, C. (2008) *Alopecurus myosuroides* (black grass), CABI. Invasive Species Compendium. *Wallingford, UK*. Available at: https://www.cabi.org/isc/datasheet/4360#CB10FF29-C5BE-43FD-97AC-BE23D2D95C67.

Rosenhauer, M. *et al.* (2013) ‘Development of target-site resistance (TSR) in *Alopecurus myosuroides* in Germany between 2004 and 2012’, *Journal of plant diseases and protection: scientific journal of the German Phytomedical Society* , 120(4), pp. 179–187.

Schmitz, R.J. *et al.* (2011) ‘Transgenerational epigenetic instability is a source of novel methylation variants’, *Science*, 334(6054), pp. 369–373.

Schuler, M.A. and Werck-Reichhart, D. (2003) ‘Functional genomics of P450s’, *Annual review of plant biology*, 54, pp. 629–667.

Schulz, B. and Kolukisaoglu, H.U. (2006) ‘Genomics of plant ABC transporters: the alphabet of photosynthetic life forms or just holes in membranes?’, *FEBS letters*, 580(4), pp. 1010–1016.

Shaner, D.L. and Beckie, H.J. (2014) ‘The future for weed control and technology’, *Pest management science*, 70(9), pp. 1329–1339.

Speidel, L. *et al.* (2019) ‘A method for genome-wide genealogy estimation for thousands of samples’, *Nature genetics*, 51(9), pp. 1321–1329.

Steinrücken, H.C. and Amrhein, N. (1980) ‘The herbicide glyphosate is a potent inhibitor of 5-enolpyruvylshikimic acid-3-phosphate synthase’, *Biochemical and biophysical research communications*, 94(4), pp. 1207–1212.

Sudmant, P.H. *et al.* (2010) ‘Diversity of human copy number variation and multicopy genes’, *Science*, 330(6004), pp. 641–646.

Sudmant, P.H. *et al.* (2015) ‘An integrated map of structural variation in 2,504 human genomes’, *Nature*, 526(7571), pp. 75–81.

Su, W.-H. (2020) ‘Advanced Machine Learning in Point Spectroscopy, RGB- and Hyperspectral-Imaging for Automatic Discriminations of Crops and Weeds: A Review’, *Smart Cities*, 3(3), pp. 767–792.

Switzer, C.M. (1957) ‘The existence of 2,4-D resistant strains of wild carrot’, *Proc. N.E.W.C.C.*, 11, pp. 315–318.

Tardif, F.J., Rajcan, I. and Costea, M. (2006) ‘A mutation in the herbicide target site acetohydroxyacid synthase produces morphological and structural alterations and reduces fitness in *Amaranthus powellii*’, *The New phytologist*, 169(2), pp. 251–264.

Thompson, J.R., Marcelino, L.A. and Polz, M.F. (2002) ‘Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by “reconditioning PCR”’, *Nucleic acids research*, 30(9), pp. 2083–2088.

Todd, O.E. *et al.* (2020) ‘Synthetic auxin herbicides: finding the lock and key to weed resistance’, *Plant science: an international journal of experimental plant biology*, 300, p. 110631.

Tranel, P.J. and Wright, T.R. (2002) ‘Resistance of weeds to ALS-inhibiting herbicides: what have we learned?’, *Weed Science*, 50(6), pp. 700–712.

Tsai, S.Q. *et al.* (2015) ‘GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases’, *Nature biotechnology*, 33(2), pp. 187–197.

United States Department of Agriculture (USDA (July,17,2020) *Recent trends in GE adoption in ‘Adoption of Genetically Engineered Crops in the U.S (1996-2020).’* Available at: https://www.ers.usda.gov/data-products/adoption-of-genetically-engineered-crops-in-the-us/recent-trends-in-ge-adoption/.

Van der Auwera, G.A. *et al.* (2013) ‘From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline’, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 43, pp. 11.10.1–11.10.33.

Van Etten, M. *et al.* (2020) 'Parallel and nonparallel genomic responses contribute to herbicide resistance in *Ipomoea purpurea*, a common agricultural weed', *PLoS genetics*, 16(2), p. e1008593.

Varah, A. *et al.* (2019) 'The costs of human-induced evolution in an agricultural system', *Nature Sustainability* [Preprint]. doi:10.1038/s41893-019-0450-8.

Vila-Aiub, M.M. *et al.* (2014) 'No fitness cost of glyphosate resistance endowed by massive EPSPS gene amplification in Amaranthus palmeri', *Planta*, 239(4), pp. 793–801.

Vila-Aiub, M.M. *et al.* (2015) 'Effect of herbicide resistance endowing Ile-1781-Leu and Asp-2078-Gly ACCase gene mutations on ACCase kinetics and growth traits in *Lolium rigidum*', *Journal of experimental botany*, 66(15), pp. 4711–4718.

Walker, K.A. *et al.* (1988) 'Fluazifop, a grass-selective herbicide which inhibits acetyl-CoA carboxylase in sensitive plant species', *Biochemical Journal*, 254(1), pp. 307–310.

Wallgren, B. and Avholm, K. (1987) 'Dormancy and germination of *Apera spica-venti* L. and *Alopecurus myosuroides* Huds. seeds', *Swedish Journal of Agricultural Research*, 8(1), pp. 11–15.

Wallinga, J., Kropff, M.J. and Rew, L.J. (2002) 'Patterns of spread of annual weeds', *Basic and applied ecology*, 3(1), pp. 31–38.

Wenger, A.M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature biotechnology*, 37(10), pp. 1155–1162.

Westwood, J.H. *et al.* (2018) 'Weed Management in 2050: Perspectives on the Future of Weed Science', *Weed Science*, 66(3), pp. 275–285.

Windsor, B., Roux, S.J. and Lloyd, A. (2003) 'Multiherbicide tolerance conferred by AtPgp1 and apyrase overexpression in *Arabidopsis thaliana*', *Nature biotechnology*, 21(4), pp. 428–433.

Winston RL, Schwarzländer M, Hinz HL, Day MD, Cock MJW, Julien MH (2014) *Biological Control of Weeds: A World Catalogue of Agents and Their Target Weeds*. USDA Forest Service, Forest Health Technology Enterprise Team, Morgantown, West Virginia.

Wright, A.A. *et al.* (2018) 'Multiple Herbicide–Resistant Junglerice (*Echinochloa colona*): Identification of Genes Potentially Involved in Resistance through Differential Gene Expression Analysis', *Weed Science*, 66(3), pp. 347–354.

Wright, S. (1931) 'Evolution in Mendelian Populations', *Genetics*, 16(2), pp. 97–159.

Xu, H. *et al.* (2014) 'Mutations at codon position 1999 of acetyl-CoA carboxylase confer resistance to ACCase-inhibiting herbicides in Japanese foxtail (*Alopecurus japonicus*)', *Pest management science*, 70(12), pp. 1894–1901.

Yang, N. *et al.* (2017) 'Contributions of *Zea mays* subspecies mexicana haplotypes to modern maize', *Nature communications*, 8(1), p. 1874.

Yu, Q. *et al.* (2010) 'AHAS herbicide resistance endowing mutations: effect on AHAS functionality and plant growth', *Journal of experimental botany*, 61(14), pp. 3925–3934.

Zhou, Q. *et al.* (2007) 'Action mechanisms of acetolactate synthase-inhibiting herbicides', *Pesticide biochemistry and physiology*, 89(2), pp. 89–96.

Supplementary material available exclusively in electronic format

The following folders and associated files are located on the enclosed CD under 'Supplementary_exclusively_electronic' or at the corresponding websites of each of the peer-reviewed publications (Chapter 1: https://doi.org/10.1111/1755-0998.13168; Chapter 2: https://doi.org/10.1073/pnas.2206808120; Chapter 3: https://doi.org/10.1111/pbi.14033).

Introduction

- Exome_capture_collection.xlsx
- Tuebingen_RADseq_collection.xlsx

Chapter 1

- men13168-sup-0002-tables1-s2-s4.xlsx
- men13168-sup-0003-tables3.xlsx
- men13168-sup-0004-tables5.xlsx

Chapter 2

- pnas.2206808120.sd01.xlsx
- pnas.2206808120.sd02.pdf
- pnas.2206808120.sd03.pdf

Chapter 3

- pbi14033-sup-0001-data_s1.xlsx