Population genomic analysis of diploid - autopolyploid species

Magdalena Bohutinská[1,2], Jakub Vlček[1], Patrick Monnahan[3], Filip Kolář[1,2]

[1]Department of Botany, Faculty of Science, Charles University, Prague, Czech Republic
[2]Institute of Botany of the Czech Academy of Sciences, Průhonice, Czech Republic
[3]Department of Pediatrics, University of Minnesota, Minneapolis, United States of America

correspondence: filip.kolar@natur.cuni.cz

running head: Analysis of mixed-ploidy SNP data

Abstract

This chapter outlines an empirical analysis of genome-wide single nucleotide polymorphism (SNP) variation and its underlying drivers among multiple natural populations within a diploid-autopolyploid species. The aim is to reconstruct genetic structure among natural populations of varying ploidy and infer footprints of selection in these populations, framed around specific questions that are typically encountered when analysing a mixed-ploidy data set, i.e. addressing the relevance of natural whole-genome duplication for speciation and adaptation. We briefly review the options for analysis of polyploid population genomic data involving variant calling, population structure, demographic history inference and selection scanning approaches. Further, we provide suggestions for methods and associated software, possible caveats and examples of their application to mixed-ploidy and autopolyploid data sets.

Keywords

Arabidopsis, autopolyploidy, genetic variation, population differentiation, selection scans

Introduction

In this chapter, we guide the reader through an analysis of genome-wide SNP variation in natural populations within a diploid-autopolyploid species. Fundamental assumptions of this chapter are (i) *a priori* knowledge of ploidy of all individuals (e.g., identified by flow cytometry or karyology), (ii) polysomic inheritance of the polyploid populations (involving random pairing among all homologous chromosomes) and thus autopolyploid origin and (iii) availability of a suitable reference genome either for the focal species or for a closely related species. For specific analyses tailored towards allopolyploidy and mixed-inheritance systems, we refer to the chapters by Scott et al. and Roux et al. in this volume. Instead of a complete review of available models, population genetic approaches and methodologies that are well discussed elsewhere [1, 2], we rather directly guide the reader through a set of procedures applicable to autopolyploid data. This chapter is thus primarily directed to empirical evolutionary biologists facing the challenge of analysing genome-wide DNA polymorphism data in autotetraploid species (although most methods are directly extendable to higher ploidy levels). Although our primary focus is on wild populations, the analyses presented here are applicable to other autopolyploid systems such as autopolyploid crops (e.g., potato, alfalfa, watermelon, pak-choi), crop-wild species complexes (e.g., interploidy crop-to-wild admixture) and evolving autopolyploid cancer cells.

Specifically, we provide an example set of analyses applicable on large (tested for up to 600 resequenced genomes) genome-wide SNP data sets of multiple diploid and autotetraploid populations. When appropriate, we refer to the results of such analyses published elsewhere [3–6]. Throughout the chapter, we illustrate our examples primarily using SCANTOOLS, a toolset specifically designed to work with autopolyploid and mixed-ploidy data. It contains a set of python3 scripts for window-based and single variant-based analyses of population differentiation, taking as an input a set of diploid and/or autopolyploid VCF (variant call format) files.

In each section, we will illustrate the application of the discussed analysis on a smaller example data set of two diploid and two autotetraploid populations of *Arabidopsis arenosa* (L.) Lawalrée (Brassicaceae). We also prepared a tutorial which can guide you through the practical aspects of analyses of the example data set, available at https://github.com/mbohutinska/PolyplChapter. *A. arenosa* is an obligate outcrosser [7] widely distributed across much of Europe, and its established diploid (2x) and autotetraploid (4x) cytotypes naturally meet at several contact zones across its range [8, 9]. The evolutionary history of the species and its intraspecific lineages are well-known, encompassing five distinct diploid lineages, one of which gave rise to a

tetraploid cytotype approximately 11,000–30,000 generations ago [3, 10, 11]. From this area of origin, the tetraploid lineage spread and came into secondary contact with the earlier diverging diploid lineages in three regions, leaving distinct traces of interploidy (2x to 4x) gene flow in tetraploid populations [3]. Moreover, the species, predominantly dwelling in foothill rocky habitats, exhibits remarkable expansion to distinct, highly-selective habitats such as toxic soils, railway tracks or alpine areas and those habitats have been colonized mainly by tetraploids [7]. Notably, alpine habitats have been independently colonized by both diploid and tetraploid populations in several areas, forming a distinct alpine ecotype bearing footprints of environmental adaptation [4, 12] (Fig. 1).

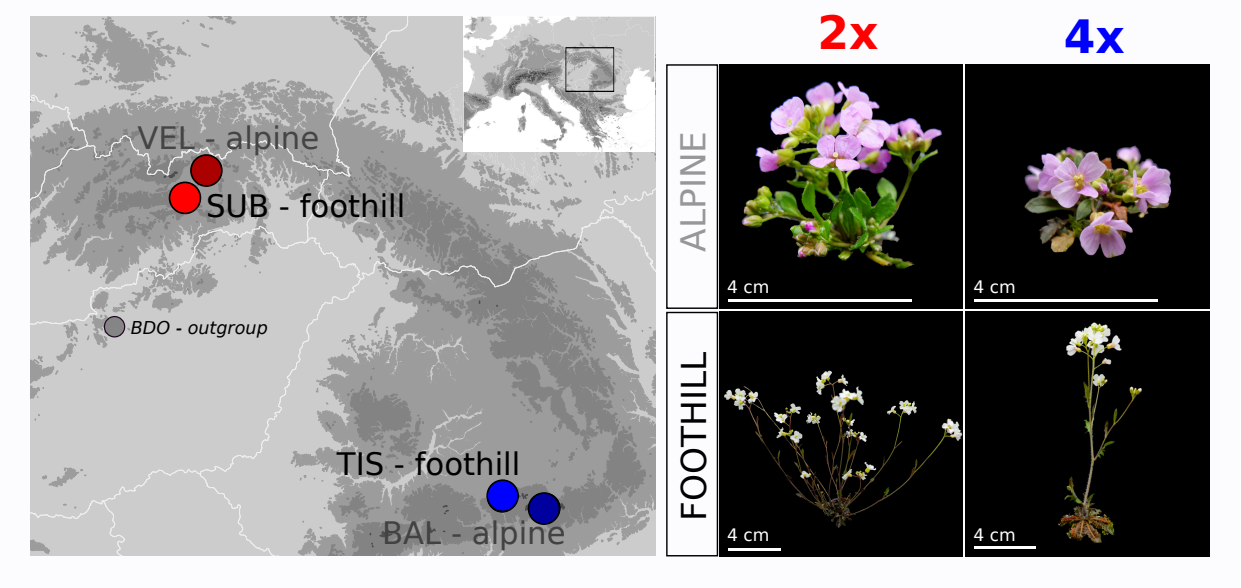

Fig. 1 Distribution of populations used in the example data set and typical phenotypes of alpine and foothill populations of *Arabidopsis arenosa*, following a cultivation over two generations in a common garden (diploid = red and tetraploid = blue; dark shade = alpine, light shade = foothill). Note that while there are no striking phenotypic differences between ploidies (a frequent phenomenon in diploid-autopolyploid systems) there are conspicuous heritable differences between foothill and alpine ecotypes within each ploidy.

Variant (SNP) calling and filtering

Regardless of ploidy, accurate and reliable variant calling is the foundational step for successful population genomic analysis. In our example, we will describe a relatively standardized way using one of the most popular variant calling tools, The Genome Analysis Toolkit (GATK); some alternative approaches are shortly summarized in the next section. Our genotyping is based on typical short-read sequence data (150 bp paired-end Illumina reads), and the pipeline consists of three main steps: mapping to a reference genome, variant calling and variant filtering. These steps are nearly identical for both diploids and tetraploids; the few deviations are explained

below. Mapping to the reference genome is unbiased by ploidy level because all four chromosomes are homologous (unlike allopolyploids where “composite” references of both diploid subgenomes are typically used). Individual physical reads are assigned to a matching position on the reference genome based on nucleotide sequence, e.g., using the program BWA [13]. Afterwards, we used GATK that allows joint variant calling and genotyping of diploid and polyploid individuals [14]. In this two-step process, genotypes are first inferred separately for each individual based on the number of reads supporting each allele and sequencing error rate using a designated likelihood model (*HaplotypeCaller* GATK tool) and then refined for the entire set of samples (callset) (*GenotypeGVCFs* tool). There are no ploidy-specific considerations for the *GenotypeGVCFs* step in which a mixture of diploids and tetraploids is analyzed together. On the other hand, the --ploidy flag in the initial per-individual *HaplotypeCaller* tool must be set correctly for each individual in order to produce the appropriate genotype calls.

Another important ploidy-specific consideration emerges during the process of filtering the VCF file produced by GATK. For an equivalent level of coverage, polyploid genotype calls tend to be less reliable than diploid genotypes, especially when the depth of coverage is low as further discussed in the chapters by Scott et al. and Meirmans in this volume. In diploids, sequencing errors (as evidenced by uneven coverage of an allele) are more easily detected as there are only three genotypes of a bi-allelic locus (*aa, Aa, AA*, with alleles *a* and *A*). In tetraploids, however, there are five possible genotypes (*aaaa, aaaA, aaAA, aAAA, AAAA*) and especially heterozygotes with uneven allele dosage (*aaaA, aAAA*) can be mistakenly assigned as errors. The potential genotyping bias can be limited by removing sites with low coverage (e.g., removing sites supported by less than *N* reads per individual). Additional filters on potentially repetitive or paralogous regions can be applied, based on criteria such as excessive heterozygosity and coverage, similarly as in diploids (e.g., [3]). However, we caution the use of relative thresholds of genotype quality (GQ annotation in the VCF file) for filtering mixed-ploidy data set. The difference between the best and second-best genotype likelihoods, and thus the value of GQ, are typically lower for polyploid genotypes (distinction between 5 genotype classes) as compared to diploids (distinction between 3 genotype classes) so that default filtering strategies for diploids based on relative GQ thresholds do not readily generalize to higher ploidy levels. Filtering can be performed in any tool capable of handling VCF files with polyploid genotypes, e.g., GATK (*VariantFiltration* and *SelectVariant* tools) or BCFTOOLS.

Allele frequency estimation and inference of genetic diversity

Inference of allele frequencies lies at the core of population genetic analysis as it represents the very first step in calculation of population diversity and differentiation among populations. Optimally, complete genotypes are reconstructed for each sampled individual from molecular markers (as in our GATK example explained above), allowing a straightforward calculation of observed allele frequencies. Population genomic analysis of polyploids or mixed-ploidy systems, however, has been plagued by several challenges not present in diploids [1, 2]. Firstly, the information on allele dosage —the number of copies of each allele present within an individual — may be challenging to infer (see for instance the chapter by Meirmans in this volume). As a consequence, the resulting incomplete genotypes are handled as missing data, in some cases leading to biases. To overcome them, specific methods allowing for genotype imputation or allele frequency estimation from genotype likelihoods can be applied (see the next paragraph). Second, certain regions of the autopolyploid genome are more susceptible to double reduction, which leads to nonstandard segregation of alleles into gametes and increased homozygosity [15, 16]. Finally, incomplete tetrasomy (or heterosomy, see Roux et al., this volume) may arise either by polyploid formation from partially divergent genomes of distinct lineages or by ongoing homolog divergence of initially “pure” autopolyploids. In such a case, preferential pairing of hom(e)ologous chromosomes may lead to segregation distortion in particular loci [1], which is manifested e.g., by increased frequency of intermediate heterozygotes (*aaAA*).

Alternative approaches in allele frequency estimation of autopolyploids are thus emerging and represent novel promising avenues in polyploid research. One direction is leveraging genotype likelihoods directly for the inference of allele frequencies and other population genomic parameters. The power of such an approach is that it by-passes the need to reconstruct complete genotypes that may be challenging in autopolyploids. With simulations, Blischak et al. [17] demonstrated that by using genotype likelihoods, one can overcome genotype uncertainty resulting from both low-coverage sequencing data, allelic dosage uncertainty and non-standard inheritance patterns. Such an approach for estimating population genetic parameters from genotype likelihoods is implemented in the programs EBG [17] and ENTROPY [18]. Another approach is modelling the particular biases associated with polyploid data such as increased variation in sequencing error, additional variation among samples beyond a simple binomial model (overdispersion) and systematic allelic bias (e.g., in UPDOG [19] and FITPOLY [20]).

Genetic diversity is usually the first indicator inferred from allele frequency data informing about past and present evolutionary processes experienced by the population. Theory predicts that, at equilibrium, the amount of nucleotide diversity in an autotetraploid population is likely to be

higher than in a similarly sized diploid population [2, 21]. This is due to the higher number of chromosome copies, which may accumulate a higher number of mutations under the same mutation rate. Owing to more chromosomes, autopolyploid populations also experience less severe effects of genetic drift than diploid populations of the same size [22]. The basic measure of genomic variation within a population, both genome-wide and locally, is nucleotide diversity, i.e. the proportion of pairwise differences among two randomly drawn chromosomes from a population (designated as π or Θ_π) [23]. π is usually calculated for a given category of variants (e.g., non-synonymous SNPs) and for a proper estimation of the proportions, invariant positions shall be also called and filtered in a similar way as SNPs. Given full-dosage genotypes reconstructed or allele frequencies estimated directly using the designated models explained above, this is a straightforward measure that has been implemented in multiple tools, including SCANTOOLS, GENODIVE (see Meirmans, this volume) and ADEGENET R package (Table 1).

A complementary way of estimating diversity is leveraging the information on allele frequencies within individuals, estimating observed (H_O) and expected heterozygosity (either indicated as H_E or H_S). While such quantification is very straightforward in diploids, it is more challenging in autopolyploids due to the presence of several classes of heterozygous individuals (e.g., *aaaA, aaAA, aAAA* at an autotetraploid locus with two alleles, *a* and *A*). A proper analysis is discussed elsewhere [2] and is implemented e.g., in GENODIVE [24]. As heterozygosity estimation is based on the reliable reconstruction of complete genotypes, it is particularly challenging in polyploids, especially in genome-wide studies where per-individual genome-wide coverage is usually achieved at a cost of lower sequencing depth.

Utilizing metrics of genetic diversity to reconstruct evolutionary history is greatly aided by information on past demographic events each population has experienced. Tajima's *D* [25] calculated on a genome-wide subset of putatively selectively neutral sites serves as a useful summary statistic (Table 1), indicating the presence of past population contractions (positive genome-wide *D* values) or expansions (negative values), barring confounding effects due to selection at linked loci. In *A. arenosa,* Monnahan et al. [3] discovered on average lower Tajima's *D* values in autotetraploid populations as compared to their diploid counterparts suggesting tetraploid expansion, in correspondence with their recent origin and large geographical area occupied. As Tajima's *D* values are strongly influenced by frequencies of rare alleles, proper inference of allele frequency and robust calling and filtering even of the rarest variants (incl. singletons) is crucial. If the ability to detect and genotype rare alleles differs across ploidies, this will introduce systematic bias in *D* across ploidies. A novel method for calculating Tajima's *D* in

autopolyploid populations from high-throughput data, accounting for individual genotype uncertainty, has been recently proposed [26]

Analysis of population genetic structure

Detecting population structure is the first step in a thorough understanding of evolutionary forces shaping diversity of natural populations, like natural selection and hybridization. It may also provide valuable insights into quantitative genetic inquiries in cultivars and laboratory strains. Specifically, population structure defines a 'neutral background', against which non-neutral processes may be highlighted and modelled accordingly. Additionally, for ploidy-variable species, determining structure within and among ploidy lineages can inform the process of polyploid origin (single vs. multiple diploid lineages involved in a polyploid's origin), spatio-temporal context of genome duplication (time, place and number of genome doubling events) and inter-ploidy gene flow (direction and potential asymmetry in the intensity of migration).

While inference of population genetic structure has a long tradition in population genetics and later also genomics of diploid organisms [27], the range of available tools is still very limited in polyploids and best practices need to be established. Here, we highlight several options provided by recent developments allowing for analysis of medium-sized data sets (up to ~100k SNPs, typically included in one VCF file comprising all scaffolds and all individuals) using a combination of ordination analyses, Bayesian clustering and network-reconstruction algorithms which are already implemented in various programs (e.g., ADEGENET and STAMPP R packages, STRUCTURE, TREEMIX and GENODIVE). Where appropriate, we also highlight possible extensions towards custom-based tools designed to handle massive parallel analyses of large genome-wide sets (>> 100 k, up to 10-20 M SNPs over hundreds of individuals; e.g., SCANTOOLS). Instead of a complete overview of available population genetic methods (reviewed elsewhere; [1, 2]), we present a set of approaches applicable on a large genome-wide SNP data set based on complete genotypes (full dosage) and moderate (~3 % - 10 %) levels of missing data in the diploid-autotetraploid species *A. arenosa*. Such a data set can be the expected outcome of a short-read resequencing project in a non-model species with a draft reference genome available (see e.g., *Cardamine amara [28], Cochlearia officinalis [29]* for recent applications to non-model species). Similarly to analyses of diploid population genomic data, we start by calculating basic population differentiation statistics, followed by an exploratory analysis of the complete data set of all individuals. After outlining these descriptive approaches, we shall turn to model-based approaches for testing specific evolutionary hypotheses using coalescent theory.

Population differentiation

Proper and unbiased estimation of population differentiation and divergence is crucial for understanding the relationships between studied populations, their evolutionary history and migration dynamics. They are also invaluable for designing further analyses focused on specific processes affecting genetic diversity such as gene flow and selection. The most widely used metrics to estimate genetic differentiation is the fixation coefficient, F_{ST}, summarizing genetic differentiation between populations inversely scaled by the diversity within populations [30]. If autopolyploids have on average higher genetic diversity than diploids, F_{ST} values between tetraploid populations should on average be lower than between diploids [2]. This may prove a problem if the study aims at direct comparison of the differentiation between pairs of diploid vs. pairs of tetraploid populations. For these purposes, one may use an alternative metric, *Rho* (ρ), which is designed to be comparable between ploidy levels [22]. It expresses the average differentiation between individuals from compared populations and should be the statistic of choice for studies of population structure in polyploids since its expected value is independent of ploidy, frequency of polysomic inheritance, rate of double reduction and mating system [2]. On the other hand, *Rho* is less suitable for comparisons across ploidy levels in a single data set as the degree of divergence between ploidy levels quantified by *Rho* typically gets inflated [2] (see Table 2). Thus, presenting both F_{ST} and *Rho* is advisable for a general overview about the differentiation in the entire mixed-ploidy data set. There are multiple programs designed to simultaneously calculate F_{ST} and *Rho*, e.g., GENODIVE (smaller data sets, also including incomplete dosage information) and SCANTOOLS (larger data sets, complete dosage) and additional options are listed elsewhere [1, 2].

Finally, analysis of molecular variance (AMOVA) offers a formalized quantification and test of hierarchical differentiation between individuals within a population, between populations and optionally also groups of populations or lineages. While originally developed for diploids, its extensions to autopolyploids with both complete and incomplete dosage have been recently developed and implemented in GENODIVE [31] and POLYGENE [32].

Principal component analysis

Unconstrained ordination method such as principal component analysis (PCA) provides perhaps the most frequently used visualization of genome-wide differentiation among individuals. The advantages are fast analysis of large multivariate data sets (consisting of thousands of SNPs) and straightforward visualization of the major orthogonal trends in variation. As input, a set of allele frequencies within individuals (0, 0.25, 0.5, 0.75 or 1 for a non-reference allele of a bi-

allelic locus of an autotetraploid) or populations of all loci are taken. Such analyses are very straightforward when full genotypes are resolved and missing data (typically replaced by average allele frequencies for that locus prior to the analysis) are not pervasive. However, large amounts of missing data (the presence of individuals with > 10-20 % missing genotypes) greatly hampers resolution of these analyses and such problematic individuals, more typically tetraploids in case of low coverage and/or incomplete dosage inference, tend to be dragged towards the centre of the PCA plot. Note that even without missing data, tetraploids may be positioned more centrally than diploids due to the fewer occurrences of the extreme genotype classes, i.e. the full homozygotes. This results in less 0 and 1 frequencies, and therefore a lower variance in allele frequencies in tetraploids compared to diploids [2]. Such an effect might be hard to separate from real demographic effects causing lower differentiation among tetraploids such as recent and potentially recurrent origin, slower divergence and intense gene flow. Additionally, as PCA is designed to capture major orthogonal trends in the variation, it is particularly sensitive to sampling (e.g., few highly diverged samples in an otherwise homogeneous data set may strongly affect the outcome). It is thus advisable to present multiple PC axes to fully capture the population structure and to complement PCA with additional approaches, listed below.

Clustering approaches

A vital complement to simple plotting of individual relatedness is the statistical inference of each individual's ancestry to an *a priori* defined number of groups. *K*-means clustering, which is not based on an explicit statistical model, offers the simplest (and quickest) approach directly applicable to any ploidy level. However, this approach is unable to reconstruct mixed ancestry of individuals and is associated with ploidy-related biases when dosage information is incomplete [33].

More elaborate, model-based approaches have been developed, allowing for inference of mixed ancestry from a hypothetical ancestral population and thus inference of potential hybrids and/or individuals exhibiting shared polymorphism. STRUCTURE [34] has been the most popular tool in this respect for the last 20 years. However, it has been primarily developed for the analysis of Sanger-sequence, microsatellite and restriction-fragment data sets, which significantly limits its application on massive parallel sequencing data. Thus, several alternative clustering methods based on similar principles have been developed for large population genomic data sets of diploids, e.g., ADMIXTURE [35] or FASTSTRUCTURE [36]. While the original STRUCTURE program also allowed for analysis of polyploid data sets (and with certain encoding drawbacks also

mixed-ploidy [33]), until very recently there was no faster alternative for large sequencing data in polyploids. Thus, the few available studies dealing with diploid-autopolyploid population structure so far relied on workarounds such as random subsampling of two alleles for analysis using programs designed to analyze diploids, e.g., ADMIXTURE or FASTSTRUCTURE [3, 37]. Encouragingly, a recent simulation-based study demonstrated that such approaches lead to only minor biases in analysis of co-dominant data with fully resolved genotypes (such as genome-wide SNPs) [33]. Recently, Shastry et al. [18] developed a clustering software specifically designed to infer population structure and ancestry in large genome-wide mixed-ploidy data sets. ENTROPY represents the very first program allowing such analysis in populations and species of varying ploidy using both high- and low-depth sequencing data, using genotype likelihoods as a direct input. If established in the community, ENTROPY may become a standard tool for postulating evolutionary hypotheses on polyploid origin and population structure.

Tree and network-reconstruction algorithms

Standard phylogenetic tools, such as Maximum Likelihood reconstruction of concatenated sequences or reconstruction under multispecies coalescent, are also generally applicable for diploid-autopolyploid systems as they typically operate with per-individual consensus sequences. However, the effect of collapsing polyploid genotypes to consensus remains to be evaluated. Applicability of phylogenetic methods is limited in systems of recent divergence, such as many polyploid complexes, where much of the variation is still segregating, ancestral polymorphism is frequent and gene flow ubiquitous. For example, while likelihood-based tree reconstruction approaches provided robust phylogeny of diploid lineages in *A. arenosa*, they failed to reconstruct relationships among the recently expanding autotetraploid lineages [3].

Thus, alternative approaches based on allele frequencies may be more informative. Perhaps the simplest (and quickest) method is calculation of genetic distances from genome-wide allele frequency data followed by their visualisation in a network, e.g., using Neighbor joining networks. As the conversion of genome-wide polymorphisms into a single distance index means dramatic simplification, selection of a suitable distance index is the crucial step in this analysis. There are multiple options for polyploid or mixed-ploidy data sets, varying in the fidelity of representation of within vs. inter-ploidy relationships and potential biases associated with missing data including incomplete genotypes; discussed elsewhere [1, 2].

Allele frequency covariance graphs, implemented in TREEMIX [38], represent an exciting alternative, which is now frequently used in large genomic data sets in diploids and has recently

also been applied to polyploid species [3, 28, 39, 40]. Unlike standard phylogenetic tools assuming bifurcation events, TREEMIX also allows for visualization of additional residual variation, unexplained by the backbone tree, as migration edges, which could be interpreted as additional pulses of gene flow. However, certain interpretation issues may arise when the levels of gene flow are high enough to dominate the allele frequency covariance patterns, and complex scenarios involving multiple polyploid origins combined with recurrent inter-ploidy introgression may occur (such as in *Neobatrachus* frogs [40]). Thus, TREEMIX serves as a hypothesis-generating method, not a formal test. To infer a realistic number of migration events, TREEMIX can be run in replicates assuming different numbers of migration events, and a partition when these models reach likelihood saturation can be identified (see [40]).

Interpretation and example application in *A. arenosa*

In *Arabidopsis arenosa*, applying the processes outlined above allowed us to unravel the complex evolutionary history of this diploid-autotetraploid complex [3, 11]. First, we revealed high overall nucleotide diversity, in line with large outcrossing populations typical for that species, which have likely survived in multiple glacial refugia. Surprisingly, nucleotide diversity of diploid and tetraploid populations was nearly the same. On the other hand, across the range, the differentiation was higher among diploid populations, as quantified by both F_{ST} and *Rho*, corresponding with relatively long divergence among diploid lineages and a recent tetraploid origin. Such a finding was further corroborated by ordination analysis (PCA), maximum likelihood phylogenies and Bayesian clustering analyses that demonstrated five divergent diploid lineages, one of which was closely related with a single yet still very diverse autotetraploid cluster. On top of that, further genetic sub-structuring within the tetraploid cluster including the tendency for clustering of spatially close diploid and tetraploid populations suggested additional locally-restricted interploidy gene flow. This hypothesis was further addressed by coalescent simulations (see the next section).

A similar trend is recapitulated also in our simplified example data set presented below (populations corresponding to Fig. 1). The populations clustered primarily according to region (and thus ploidy) as indicated by separation along PC1 (Fig. 2A), distance-based Neighbor joining tree (Fig. 2B), TREEMIX graph (Fig. 2C, D) and clustering in ENTROPY (Fig. 2E). Differentiation appeared to be higher between diploid populations than between tetraploids as identified by the degree of separation along the second PC axis (PC2; Fig. 2A) and the higher F_{ST} values (Table 2). However, the opposite is found when *Rho* values are compared, which are more appropriate for such comparisons, showing that the PC2 differentiation results are biased

by ploidy itself, in line with simulation studies [2]. There was only slightly higher diversity in tetraploid populations (Table 1).

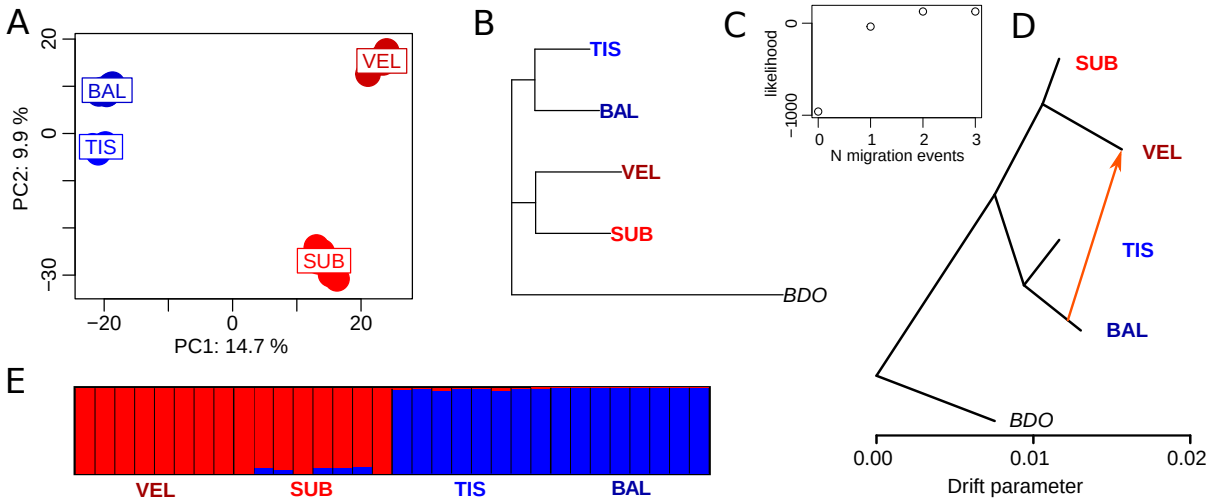

Fig. 2 Relationship between diploid and autotetraploid populations of *Arabidopsis arenosa* from two regions and of two elevational ecotypes reconstructed using ~200 k putatively neutral fourfold degenerate SNPs. (A) Principal component analysis of the individuals, coloured according to ploidy (corresponding with the region) and alpine/foothill ecotype (red - diploid, blue - tetraploid; foothill light, alpine dark). (B) Neighbor joining tree based on Nei's genetic distances between populations calculated from allele frequencies. (C, D) Allele frequency covariance graph of populations, inferred by TREEMIX, and the likelihood of models differing by number of migration events. (C) One migration event has been selected based on the highest improvement in likelihood values. (D) Distances along the x-axis show the drift parameter estimates, corresponding to the number of generations separating the two populations and their effective population sizes ($t/2N_e$). The arrow indicates the most likely migration event. The outgroup (BDO) is a diploid population from the Pannonian Basin, previously inferred as the earliest-diverging lineage of the entire species. (E) Posterior probability of assignment of each sequenced individual to the two groups as inferred by ENTROPY based on ~7 k unlinked SNPs (subsampled to a minimum distance of 1 kb to avoid linked sites).

<u>Best practices (example programs in brackets, summarized in Table 3)</u>

- Based on genome annotation information, extract putatively selectively neutral sites such as intergenic (if data quality permits) and/or fourfold-degenerate genic sites (GATK or BCFTOOLS) *[input: .vcf file]*
- Load the VCF file into R (using VCFR) and calculate PCA (ADEGENET) and inspect for outlier individuals (erroneous species, contaminants, hybrids/allopolyploids?) and biases stemming from low-coverage individuals with excessive amounts of missing data, positioned towards the centre of the plot *[input: .vcf file]*
- Follow-up with quality checks: calculate genetic distances among individuals (STAMPP), plot them in a Neighbor joining network (SPLITSTREE) and check for additional spurious

clustering in the network, e.g., individuals not clustering with other members of their populations (contaminants, labelling errors?) or connected with multiple alternative splits of similar weight (hybrids / allopolyploids?). Remove such individuals, clearly document the criteria used for such a decision and check the PCA and Neighbor joining network once again. *[input: .vcf for R packages and distance matrix .dst file for SPLITSTREE]*

- For the clean data set, calculate diversity statistics and genome-wide Tajima's *D* for each population (SCANTOOLS or GENODIVE), investigate potential sources of diversity variation among populations (ploidy, spatial distribution, etc.) to postulate hypotheses about past population size changes *[input: SCANTOOLS - .vcf file + population key, GENODIVE - special format or several population genetic formats, e.g., structure-formatted .str file]*
- Calculate F_{ST} and *Rho* differentiation among populations (SCANTOOLS or GENODIVE); in comparisons focused on relative differentiation among 2x-2x vs. 4x-4x population pairs use *Rho*. Optionally, the matrix of inter-population distances may also be plotted as a network using SPLITSTREE.
- Run clustering analysis (STRUCTURE, ENTROPY, and/or *K*-means clustering in ADEGENET or GENODIVE) following general recommendations for mixed-ploidy data [33] and infer major genetic groups / lineages and potential admixed individuals / populations *[input: structure-formatted .str file]*
- Infer the relationships among individuals (e.g., Neighbor-Joining networks in SPLITSTREE, PCA in ADEGENET) and populations (e.g., TREEMIX) in a network-like manner *[input: .vcf file]*
- Integrate the information from population differentiation, PCA, clustering and network-based analyses, and formulate hypotheses on population relationships, the extent of intra- and inter-ploidy gene flow and number and geographic locations of autopolyploid origins.

Inference of population demographic history

Coalescent simulations followed by a model-selection approach are becoming a standard tool for selecting among competing scenarios of population demographic history as well as for the inference of important demographic parameters such as past and present effective population size (*Ne*), divergence time and past gene flow (migration)(see also Roux et al., this volume). Comparing coalescent-based simulations of population genetic data for different models of demographic history allows statistical testing of hypotheses formulated *a priori* (e.g., based on

ploidy differentiation, spatial distribution, ecological preference or addressing controversial scenarios resulting from the exploratory approaches). For autotetraploids with tetrasomic inheritance, Arnold et al. [41] demonstrated that at a locus without recombination, the ancestral genetic process in a large population may be approximated using Kingman's standard coalescent, with a coalescent effective population size 4*N*. Thus, coalescent simulation programs developed for diploids can be adapted to study population history in autotetraploids by interpreting the timescale in units of 4*N* generations.

There is an expanding set of tools for coalescent simulations available, here we will focus on FASTSIMCOAL2, a versatile tool that has been used most frequently in the few so far available inquiries of autotetraploid and mixed-ploidy systems [3, 10, 12, 39, 42]. A first step when analyzing empirical data is retrieving allele frequency spectra (AFS, also referred to as site frequency spectra, SFS) that could be directly calculated from existing allele frequency data stored in a VCF file (e.g., using SCANTOOLS). A specialized approach focused on estimation of joint AFS in autotetraploids, additionally taking into consideration potential deviations from random mating, has been recently developed yet not implemented in software [26].

Coalescent simulations are a demanding multi-step process that is strongly conditional on *a priori* knowledge on the system studied; not only the results of population structure reconstruction, but also overall biology of the species, mode of polyploid origin (auto- vs. allo-polyploidy), local geological history, etc. Thus, it requires careful construction of realistic yet not overly complicated models, as runtimes increase drastically as the number of populations and thus possible population iterations and parameters to be estimated increase. Briefly, expected AFS are simulated under a given set of parameters of each modelled scenario and their composite likelihood is computed. An efficient optimization algorithm is used to find optimal parameter values from which simulations best match the observed AFS, thus maximising the likelihood (for more details see dedicated studies [43]). Finally, likelihood values inferred for different scenarios are compared, e.g., using the Akaike Information Criterion (AIC). Alternatively, Approximate Bayesian Computation (ABC) methods maybe used to perform simulation-based inference and model selection in a Bayesian framework (see Roux et al., this volume).

For mixed-ploidy systems, a scenario of single vs. multiple autopolyploid origin can be compared in a computationally efficient way via a population-quartet design, if the source diploid populations are sampled [3]. Two general topologies can be contrasted (i) sister position of populations of each ploidy (scenario 1 in Fig. 3) and (ii) sister position of a diploid and tetraploid population paired by, e.g., sympatry or similar morphology (scenario 2 in Fig. 3). Such

topologies may or may not be accompanied by additional inter- or intra-ploidy gene flow (e.g., within regions) and population size changes (e.g., initial bottleneck followed by population growth in expanding polyploid). If more populations per each ploidy/lineage are sampled, it is advisable to run the analysis repeatedly, iterating different (meaningful) population quartets to leverage such natural replicates while keeping the models manageably simple. For example, when comparing scenarios of repeated polyploid origin vs. single origin followed by regional interploidy gene flow, Monnahan et al. [3] used such quartets and for each scenario iterated different populations of each ploidy and lineage. They found that regardless of population combination, the scenario of single origin was preferred, suggesting selection of the winning scenario was not driven by specific properties of one particular population. Additional to modelling polyploid origin, hypotheses on the origin of ecologically distinct populations, population divergence and historical population size changes can be addressed within each ploidy or in mixed-ploidy systems in a similar way to diploid systems [4, 5, 12] (Fig. 3).

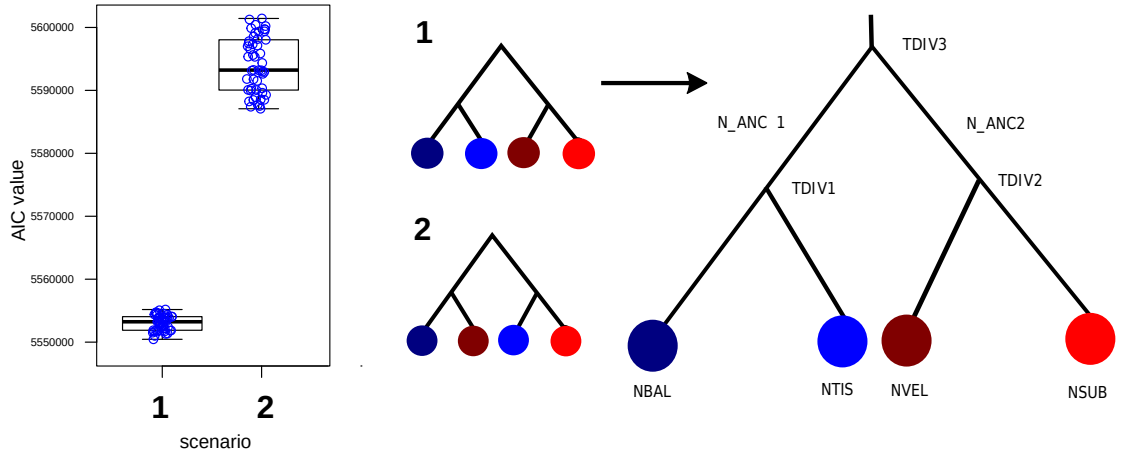

Fig. 3 Comparison of evolutionary scenarios simulated in coalescent framework in FASTIMCOAL. Although population structure analyses (Fig. 2) strongly indicate regional (ploidy) clustering, formal rejection of an alternative scenario (single origin of alpine ecotype followed by independent genome duplication within each region) can be done by a model comparison of distinct scenarios. We simulated a four-population model assuming either regional = ploidy (scenario 1) or ecotype (scenario 2) sister relationships. Alpine (dark shade) vs. foothill (light shade) ecotype, diploid (red) and tetraploid (blue) cytotype. Boxplot summarizes AIC values over 50 FASTSIMCOAL optimization runs for each scenario, demonstrating that the scenario of parallel origin of the alpine ecotype consistently reaches lower AIC and thus is preferred. Scenario 1 is further depicted with all parameters used in the simulations, namely effective population sizes (codes starting with 'N') and divergence times ('TDIV').

Best practices

- Create multidimensional joint allele frequency spectra (AFS, SFS) from the well-filtered VCF file of putatively selectively neutral sites of a particular population set (module

.generateFSC2input in SCANTOOLS). Filtering should be done carefully in order to estimate proper frequencies of all classes of mutations, including the rarest categories (e.g., do not apply minor allele frequency filters, do not remove singletons, etc.). Missing data can be handled by downsampling to a smaller number of individuals for which complete genotype information is available. *[input: .vcf file]*

- Carefully consider the scenarios to be compared and encode them into designated fastsimcoal files (*.est* and *.tpl* files). Construct as simple models as possible, in order to avoid exploring unrealistic broad parameter space and/or comparing too many scenarios. To achieve generality, consider running a similar set of scenarios over multiple combinations of natural populations (natural replicates). Set realistic, sufficiently broad ranges for initial parameter values.
- Specifically to polyploidy, it is good to consider specific possibilities such as limited/asymmetrical gene flow between ploidy levels, independent polyploid origins, large and increasing population sizes for expanding polyploid lineages or, in contrast, a bottleneck if the recent polyploid lineage had not yet restored diversity by recurrent origins or gene flow.
- Run the simulations following recommendations in the FASTSIMCOAL2 manual (e.g., at least 50 replicates per scenario and population combination), collect the likelihoods, calculate AIC and perform model comparison for each particular population combination. It is also worth exploring the distribution of estimated parameter values to check whether the replicates are converging to similar (and realistic) optimal values under each scenario. Likely, various combinations of initial parameter settings and scenarios have to be explored prior to reaching final conclusions. *[input: empirical joint allele frequency spectra in .obs files, parameter .est and .tpl files]*
- For the selected scenario, bootstrap the SNP matrix used for creating the empirical AFS (bootstrapping option is available within the module *.generateFSC2input* in SCANTOOLS), re-run the analysis and calculate confidence intervals for your parameters of interest.

Population genomic inference of selection

Identification of genomic regions differentiated by positive selection is a way to gain insights into the genomic basis of adaptive evolution. Specifically to polyploid systems, such analyses in mixed-ploidy species may be particularly relevant for understanding the genomic basis of adaptation towards organismal and cellular-level challenges imposed by whole-genome

duplication [6, 44], significance of (adaptive) inter-ploidy introgression [45, 46] as well as mechanisms of selection and (local) adaptation in polyploid populations [4, 42].

Numerous statistical methods have been developed to detect genomic regions bearing footprints of positive selection in diploid data sets (called selection scans hereafter). Applications of these methods range from the study of recent speciation (scanning for microevolutionary changes between subdivided populations) to selection scans for macroevolutionary changes between species [47]. While the broad diversity of selection scan methods covers various timescales of divergence between compared populations and species (Fig. 4), they mostly assume (and were tested for) diploid outcrossing populations without sharp demography changes [48]. This makes it difficult to directly infer their applicability for analysis of polyploid data sets and simulation-based validations or tests are generally lacking. Nevertheless, here we present several complementary approaches which we have applied for analysing the footprints of selection in genome resequencing data sets representing both mixed-ploidy and purely autopolyploid populations [4, 5] (Table 4).

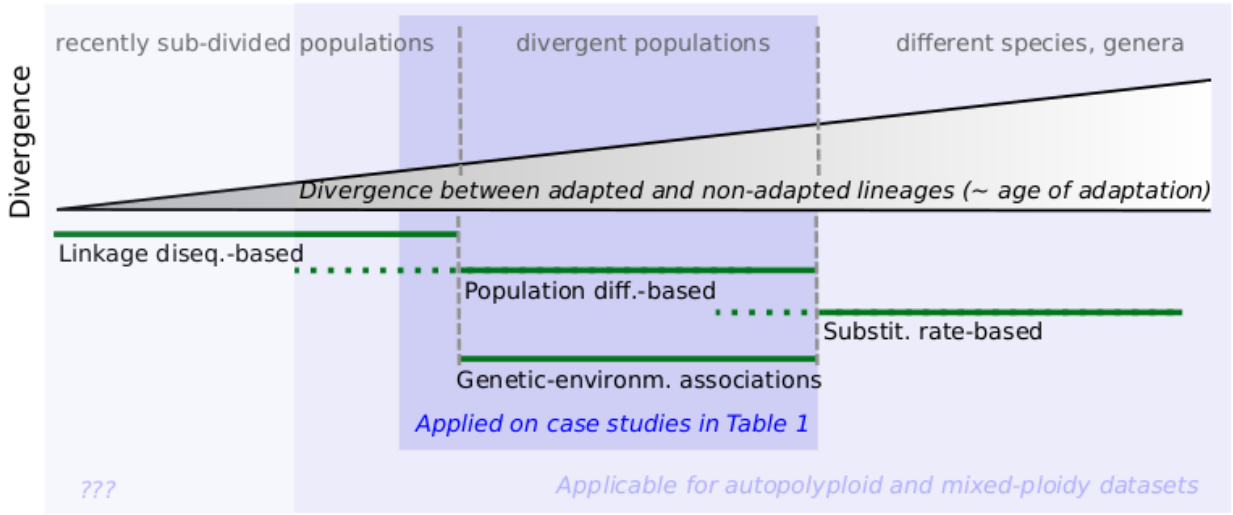

Fig. 4 Overview of selection scan methods to detect signatures of positive selection and their applicability for autopolyploid and mixed-ploidy data sets. The question marks denote so far unexplored approaches of haplotype-based selection scans.

Very roughly, selection scans can be divided into four main categories: (1) substitution rate-based selection scans, suitable for between-species or deeper divergences, (2) population differentiation-based scans, primarily identifying signals of rapid allele frequency changes (hard sweeps) between populations, (3) linkage disequilibrium-based scans, searching for recent modest allele frequency changes (soft sweeps) within populations and (4) environmental association analyses, which attempt to link environmental variables with associated genetic

variation, providing more direct evidence for the drivers of local adaptation. For more details see dedicated reviews [47, 49–51].

In mixed-ploidy and autopolyploid data sets, we applied three of these approaches, each suitable for different divergence between the compared units: substitution rate-based and population differentiation-based scans and environmental association analyses (Table 4, Fig. 4). All three approaches are based on genotype/allele frequencies, whose estimation is feasible in autopolyploids (conditioned on detection of full genotypes, reliable imputation of partially resolved genotypes and/or likelihood-based allele frequency estimation; for details see the third section of this chapter). The use of linkage disequilibrium-based methods is perhaps the most challenging as it requires proper inference of genomic linkage information from short read data which is very difficult in autopolyploids [52]. Although this was recently achieved in highly heterozygous autotetraploid potato [53], we will introduce here only the three approaches based on genotype/allele frequencies. On the bright side, there is empirical evidence of positively selected standing alleles (swept by soft sweeps, [54]) identified in autotetraploid *A. arenosa* using allele frequency-based methods [5, 6] suggesting that soft sweeps can be at least partially captured by the population-differentiation approaches.

Selection scan approaches are typically based on a specific design, comparing sets of populations from contrasting environments (called here ‘background’ and focal ‘adapted’ populations). Such design may for example involve the contrast of a plant population growing on toxic soil (‘adapted’), with its close relative occupying non-toxic soil (‘background’ - note that even a background population is expected to be adapted to its local environment). Specifically to mixed-ploidy and autopolyploid data sets, there are three possible selection scans designs (Table 4). First, one may wish to scan for selection between diploids and tetraploids (contrast 2x-4x), for example to ask which genes are involved in adaptation to whole-genome duplication. Such a comparison in *A. arenosa* revealed a set of genes involved in meiosis likely underlying the adaptation to polyploid state [44]. Second, the study design may include multiple adapted and background population pairs, each pair comprising populations of the same ploidy but the pairs differing in ploidy (i.e multiple 2x-2x and 4x-4x pairs). This setup tests for the presence of parallel adaptation in both ploidies. Examples include adaptation to harsh alpine environments, repeatedly occurring in diploid *Arabidopsis halleri* as well as diploid and autotetraploid *A. arenosa* [4]. Third, the data set might contrast various autopolyploid populations only, e.g., following adaptive divergence of established autopolyploid lineages. An example is adaptation of an autotetraploid *A. arenosa* to toxic serpentine soils *[5]*. While in the first example, polyploidization represents the selection agent of interest, in the second and third case it acts

rather as a confounding factor. Nevertheless, the selection scan methods listed below are applicable to all three of these designs. Some of the candidate genes which we identified using these approaches have also been followed further experimentally. This follow-up research provided multiple hints that the adapted alleles code a new function in comparison to the backgound alleles [5, 55] (Table 4).

Each of the selection scan approaches is implemented in various methods and programs. First, for simple population differentiation scans, any software which enables calculation of F_{ST}, D_{XY}, *Rho* or related estimators of population differentiation and divergence from allele frequency data in autopolyploids should be applicable (we used SCANTOOLS). Calculation of population differentiation for each window/SNP should be then followed by identification of the most outlier regions, either by dedicated programs (e.g. SCANTOOLS) or by custom scripts. Differentiation scans are intuitive and straightforward, and they may provide a convenient selection detection tool. However, their application requires good understanding of the evolutionary history of studied populations (see Roux et al., this volume) and correct setting of contrasts of interest (i.e., carefully selecting pairs of 'adapted' and phylogenetically close 'background' populations). Further, one may wish to apply scans for differentiation that are model-based, explicitly accounting for the covariance structure among the population allele frequencies reflecting common history of the populations under study. This may improve the robustness of differentiation scans by considering population structure while estimating differentiation (see dedicated papers i.e. [57, 58]). With this motivation, we applied BAYPASS [57], which requires reference and alternative allele counts as an input, enabling us to include autopolyploid data, and identified set of candidate loci potentially involved in alpine adaptation. By overlapping such candidate lists across multiple foothill-alpine population constrasts we also identified a subset of such candidates which probably evolved in parallel [4] (see Fig. 5 for an example). Finally, to approximate a likely functional consequence of candidate regions, one might combine population differentiation measures with estimation of the impact of corresponding amino acid substitutions on the protein function. This approach is implemented in the method FINEMAV [59], originally developed for diploid human data, but has been modified also for non-model autopolyploid data sets [6].

Second, for tests for selection based on the distribution of substitutions between populations, like the McDonald-Kreitman test [60], one requires knowledge of the number of synonymous and nonsynonymous substitutions and polymorphisms. Apart from identifying specific candidate genes for positive selection, this type of procedure can also establish the genome-wide adaptive proportion of genetic variation, which may inform the debate about possible higher adaptability

of autopolyploids. However, estimates of adaptive evolution from the McDonald-Kreitman test may be biased downwards by the presence of slightly deleterious mutations [61]. Thus, filtering out low-frequency polymorphisms and other workarounds were suggested by simulations [61] and applied for diploids [62, 63]. Although we miss similar simulation-based recommendations for autopolyploids, their capacity to harbour higher recessive variability than diploids [21] may make them even more prone to these biases, calling for applying similar recommendations as those developed for diploids.

Third, to identify alleles associated with population-specific covariates (e.g., environmental variables, phenotypic traits) in diploids and tetraploids, we used BAYPASS [57] and latent factor mixed models (LFMM; [64]). As mentioned above, BAYPASS is a Bayesian method to identify genomic regions subjected to adaptive divergence across populations. On top of this framework, it also involves models enabling the assessment of association of allele frequency differences with population-specific information about environmental conditions or phenotype [57]. LFMM uses similar logic, testing the association between genetic markers and variables of interest, yet here using linear mixed models with a discrete number of ancestral population groups modelled as latent factors. LFMM is, however, much faster and computationally less demanding. Both methods, BAYPASS and LFMM, require population allele counts as an input and their application in mixed-ploidy and autopolyploid data sets is thus straightforward.

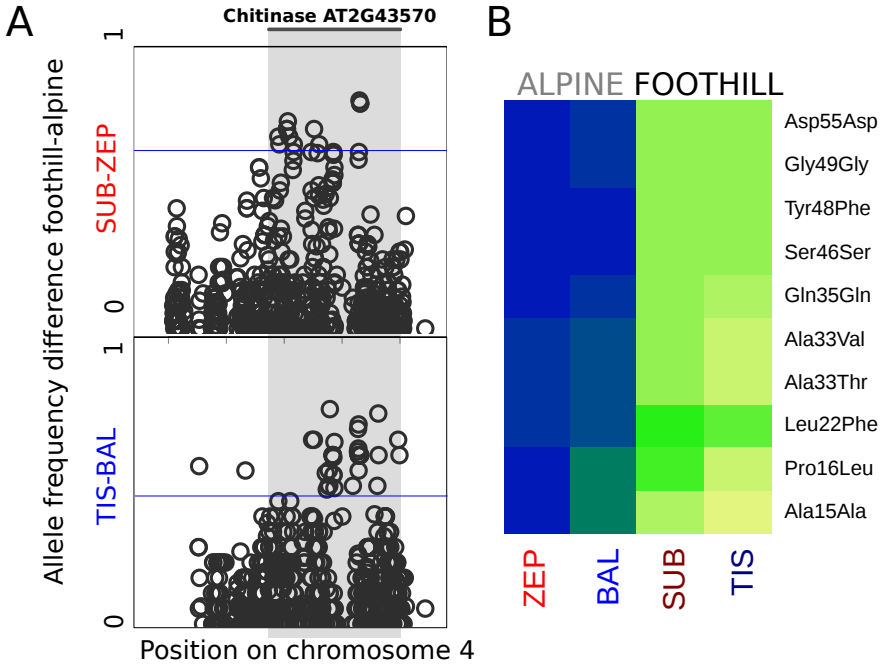

Fig. 5 Example of a candidate locus for alpine adaptation identified by overlapping simple divergence scanning approaches while leveraging detailed knowledge of the evolutionary history of the system. (A)

Pattern of allele frequency differentiation between foothill and alpine population in a selected locus that has been revealed as a top candidate in both independent population contrasts (mountain ranges) investigated. (B) Heatmap showing allelle frequency difference of this candidate in the complete data set, the frequency of derived allele scales with the depth of blue.

Assumptions and limitations

To account for possible biases connected to the analysis of autopolyploid data, one should first understand the population genomic consequences of autopolyploidy to the pattern of allele frequency changes left by positive selection [21]. In short, simulations showed that peaks of genetic differentiation formed by the action of positive selection are narrower in autopolyploids, likely as a result of elevated recombination rates and longer times to fixation. This may lead to better localisation of the specific gene or region targeted by positive selection. On the other hand, differentiation tends to be more subtle in autopolyploids, especially when the selected allele is (partially) dominant. That leads to overall slower allele frequency shifts under positive selection. For a comprehensive overview of the effect of autopolyploidy on the patterns of selection sweeps see [21]; here we focus on possible practical consequences of these effects.

First, all approaches introduced above strongly depend on the assumption that analysed autopolyploids exhibit random segregation of homologous chromosomes (general assumption of the entire chapter). Deviations from this rule may lead to segregation distortions and bias our estimation of genetic differentiation [2, 65]. Second, if the divergence between background and adapted population is low and/or adaptation proceeds via recessive alleles, the peak of differentiation may be too low to be detected under stringent outlier/significance thresholds [21]. This again highlights the need for good understanding of the evolutionary history of the studied system. Still, the overlap of candidate genes discovered using a higher number of complementary selection scanning approaches with lower thresholds (see e.g., [4, 44]) may partially help to overcome these issues, at the cost of allowing potentially more false positives in the discovery process. Third, while the better localisation of selection sweeps in autopolyploids in principle provides an advantage [21], more sweeps might remain undetected when using reduced representation sequencing techniques like RADseq as the linkage is weaker in autopolyploids. Finally, although genetic differentiation estimators may be biased when comparing values inferred for population pairs of different ploidy (e.g., F_{ST} in case of 2x-2x vs. 4x-4x comparisons [2]), they are still helpful in outlier selection scans based on relative comparison to a genome-wide distribution for that particular pair, like, e.g., 5 % of the top-differentiated regions. While keeping in mind drawbacks of using relative thresholds (e.g., allowing an unknown amount of false positives; [49]), we recommend against the usage of

absolute value of genetic differentiation measures as outlier threshold over multiple population pairs differing by ploidy, as the absolute values may be affected by ploidy.

In conclusion, selection scans may provide interesting insights into the genomics of adaptive evolution in both diploids and autopolyploids [7]. However, we stress that similarly to the diploid case, and likely even more, selection scans should serve as hypothesis-generating tools to stimulate further research, not an approach to confidently identify the adaptive genes. Asserting an adaptive role of specific SNPs will require more rigorous functional or *in silico* validation of candidate selected alleles, i.e. modelling the impact of candidate allele on a protein structure or introduction of adaptive allele into a common background and testing for its effect in a common garden experiment.

Best practices

- Carefully design population sampling based on a good knowledge of the evolutionary and ecological background of the study system (known ploidy and population structure, number of polyploidization events - see the section 'Analysis of population genetic structure') and additional information on the adaptive events of interest (information on selective factor in natural populations, e.g., abiotic and biotic parameters; optimally complemented by transplant or common garden experiments determining heritable basis of the adaptive traits and overall fitness response).
- Consider the divergence between adapted and background populations (section 'Inference of population demographic history') and up-to-date knowledge about available selection scans methods to design the optimal selection scanning procedure. Ideally, combine multiple complementary selection scan approaches:
    - substitution-based methods, calculating the number of substitutions, i.e. fixed non-reference alleles and polymorphisms for each category and applying the McDonald-Kreitman test. Use SnpEff [66] to classify synonymous and nonsynonymous variants and then use either custom scripts or a dedicated model-based web tool (Table 3). Be aware that this test is not applicable to recent selection events, in which both adaptive and background alleles are still segregating within populations. [*input: .vcf file (SnpEff)]*
    - differentiation-based methods (F_{ST}, D_{XY}, *Rho* - ScanTools, BayPass, FineMAV + SnpEff) [*input: .vcf file (ScanTools, SnpEff), allele count tables (BayPass)*]
    - environmental association analyses (BayPass, LFMM) [*input: allele count tables*]

- Visualise the differentiation metrics at the candidate loci, use visual feedback to fine-tune your selection scan design, yet anticipate that visual manifestations of selection are influenced by ploidy; ScanTools, scripts. Consider using different method settings and thresholds, explore how changing thresholds alter the results while avoiding cherry picking.
- Collapse the resulting candidate regions (sites/windows) to genes, interpret them in a functional context (GO enrichment, KEGG pathway analysis, STRING protein interaction networks, functional descriptions from databases like Uniprot).
- Consider follow-up predictive or functional validation of the top candidate loci.

Acknowledgements

We are grateful to Arthur Zwaenepoel, Patrick Meirmans, Josselin Clo, Nélida Padilla García and Polina Novikova for very useful comments to an earlier version of this chapter. We thank Veronika Konečná, Nélida Padilla García and Gabriela Šrámková for help with running particular analyses of the example data set and Doubravka Požárová for photos of alpine plants. This work was supported by the Czech Science Foundation (project 20-22783S to FK) and the long-term research development project No. RVO 67985939 of the Czech Academy of Sciences. Access to computing and storage facilities has been provided by the National Grid Infrastructure MetaCentrum provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042).

References

1. Dufresne F, Stift M, Vergilino R, Mable BK (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol Ecol 23:40–69. https://doi.org/10.1111/mec.12581

2. Meirmans PG, Liu S, van Tienderen PH (2018) The Analysis of Polyploid Genetic Data. J Hered 109:283–296. https://doi.org/10.1093/jhered/esy006

3. Monnahan P, Kolář F, Baduel P, et al (2019) Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. Nature Ecology & Evolution 3:457. https://doi.org/10.1038/s41559-019-0807-4

4. Bohutínská M, Vlček J, Yair S, et al (2021) Genomic basis of parallel adaptation varies with divergence in Arabidopsis and its relatives. PNAS 118:. https://doi.org/10.1073/pnas.2022713118

5. Konečná V, Bray S, Vlček J, et al (2021) Parallel adaptation in autopolyploid Arabidopsis arenosa is dominated by repeated recruitment of shared alleles. bioRxiv 2021.01.15.426785. https://doi.org/10.1101/2021.01.15.426785

6. Bohutínská M, Handrick V, Yant L, et al (2021) De Novo Mutation and Rapid Protein (Co-)evolution during Meiotic Adaptation in Arabidopsis arenosa. Molecular Biology and Evolution 38:1980–1994. https://doi.org/10.1093/molbev/msab001

7. Yant L, Bomblies K (2017) Genomic studies of adaptive evolution in outcrossing Arabidopsis species. Current Opinion in Plant Biology 36:9–14. https://doi.org/10.1016/j.pbi.2016.11.018

8. Kolář F, Lučanová M, Záveská E, et al (2016) Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the Arabidopsis arenosa group (Brassicaceae). Biol J Linn Soc Lond 119:673–688. https://doi.org/10.1111/bij.12479

9. Morgan EJ, Čertner M, Lučanová M, et al (2020) Niche similarity in diploid-autotetraploid contact zones of Arabidopsis arenosa across spatial scales. American Journal of Botany 107:1375–1388. https://doi.org/10.1002/ajb2.1534

10. Arnold B, Kim S-T, Bomblies K (2015) Single geographic origin of a widespread autotetraploid Arabidopsis arenosa lineage followed by interploidy admixture. Molecular biology and evolution 32:1382–1395

11. Kolář F, Fuxová G, Záveská E, et al (2016) Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model Arabidopsis arenosa. Molecular Ecology 25:3929–3949. https://doi.org/10.1111/mec.13721

12. Knotek A, Konečná V, Wos G, et al (2020) Parallel Alpine Differentiation in Arabidopsis arenosa. Front Plant Sci 11:. https://doi.org/10.3389/fpls.2020.561526

13. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

14. McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110

15. Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. New Phytol 186:5–17. https://doi.org/10.1111/j.1469-8137.2009.03142.x

16. Butruille DV, Boiteux LS (2000) Selection–mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. PNAS 97:6608–6613

17. Blischak PD, Kubatko LS, Wolfe AD (2018) SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics 34:407–415. https://doi.org/10.1093/bioinformatics/btx587

18. Shastry V, Adams PE, Lindtke D, et al (2021) Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy. Molecular Ecology Resources 21:1434–1451. https://doi.org/10.1111/1755-0998.13330

19. Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping Polyploids from Messy Sequencing Data. Genetics 210:789–807. https://doi.org/10.1534/genetics.118.301468

20. Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12:172. https://doi.org/10.1186/1471-2105-12-172

21. Monnahan P, Brandvain Y (2020) The effect of autopolyploidy on population genetic signals of hard sweeps. Biology Letters 16:20190796. https://doi.org/10.1098/rsbl.2019.0796

22. Ronfort J, Jenczewski E, Bataillon T, Rousset F (1998) Analysis of Population Structure in Autotetraploid Species. Genetics 150:921–930. https://doi.org/10.1093/genetics/150.2.921

23. Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. PNAS 76:5269–5273

24. Meirmans PG genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. Molecular Ecology Resources n/a: https://doi.org/10.1111/1755-0998.13145

25. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

26. Ferretti L, Ribeca P, Ramos-Onsins SE (2018) The Site Frequency/Dosage Spectrum of Autopolyploid Populations. Front Genet 9:. https://doi.org/10.3389/fgene.2018.00480

27. Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nat Rev Genet 16:727–740. https://doi.org/10.1038/nrg4005

28. Bohutínská M, Alston M, Monnahan P, et al (2021) Novelty and convergence in adaptation to whole genome duplication. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msab096

29. Bray SM, Wolf EM, Zhou M, et al (2020) Convergence and novelty in adaptation to whole genome duplication in three independent polyploids. bioRxiv 2020.03.31.017939. https://doi.org/10.1101/2020.03.31.017939

30. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589

31. Meirmans PG, Liu S (2018) Analysis of Molecular Variance (AMOVA) for Autopolyploids. Front Ecol Evol 6:. https://doi.org/10.3389/fevo.2018.00066

32. Huang K, Wang T, Dunn DW, et al (2021) A generalized framework for AMOVA with multiple hierarchies and ploidies. Integrative Zoology 16:33–52. https://doi.org/10.1111/1749-4877.12460

33. Stift M, Kolář F, Meirmans PG (2019) Structure is more robust than other clustering methods in simulated mixed-ploidy populations. Heredity 123:429–441. https://doi.org/10.1038/s41437-019-0247-6

34. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

35. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664. https://doi.org/10.1101/gr.094052.109

36. Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics 197:573–589. https://doi.org/10.1534/genetics.114.164350

37. Novikova PY, Hohmann N, Nizhynska V, et al (2016) Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nature Genetics 48:1077–1082. https://doi.org/10.1038/ng.3617

38. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8:e1002967. https://doi.org/10.1371/journal.pgen.1002967

39. Wos G, Mořkovská J, Bohutínská M, et al (2019) Role of ploidy in colonization of alpine habitats in natural populations of Arabidopsis arenosa. Ann Bot 124:255–268. https://doi.org/10.1093/aob/mcz070

40. Novikova PY, Brennan IG, Booker W, et al (2020) Polyploidy breaks speciation barriers in Australian burrowing frogs Neobatrachus. PLOS Genetics 16:e1008769. https://doi.org/10.1371/journal.pgen.1008769

41. Arnold B, Bomblies K, Wakeley J (2012) Extending Coalescent Theory to Autotetraploids.

Genetics 192:195–204. https://doi.org/10.1534/genetics.112.140582

42. Arnold BJ, Lahner B, DaCosta JM, et al (2016) Borrowed alleles and convergence in serpentine adaptation. PNAS 113:8320–8325. https://doi.org/10.1073/pnas.1600405113

43. Excoffier L, Dupanloup I, Huerta-Sánchez E, et al (2013) Robust demographic inference from genomic and SNP data. PLoS Genetics 9:e1003905

44. Yant L, Hollister JD, Wright KM, et al (2013) Meiotic adaptation to genome duplication in Arabidopsis arenosa. Current Biology 23:2151–2156. https://doi.org/10.1016/j.cub.2013.08.059

45. Marburger S, Monnahan P, Seear PJ, et al (2019) Interspecific introgression mediates adaptation to whole genome duplication. Nat Commun 10:1–11. https://doi.org/10.1038/s41467-019-13159-5

46. Schmickl R, Yant L Adaptive introgression: how polyploidy reshapes gene flow landscapes. New Phytologist n/a: https://doi.org/10.1111/nph.17204

47. Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting Natural Selection in Genomic Data. Annual Review of Genetics 47:97–120. https://doi.org/10.1146/annurev-genet-111212-133526

48. Stephan W (2019) Selective Sweeps. Genetics 211:5–13. https://doi.org/10.1534/genetics.118.301319

49. Hoban S, Kelley JL, Lotterhos KE, et al (2016) Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. The American Naturalist 188:379–397. https://doi.org/10.1086/688018

50. Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences 365:185–205. https://doi.org/10.1098/rstb.2009.0219

51. Booker TR, Jackson BC, Keightley PD (2017) Detecting positive selection in the genome. BMC Biology 15:98. https://doi.org/10.1186/s12915-017-0434-y

52. Motazedi E, Finkers R, Maliepaard C, de Ridder D (2018) Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. Briefings in Bioinformatics 19:387–403. https://doi.org/10.1093/bib/bbw126

53. Siragusa E, Haiminen N, Finkers R, et al (2019) Haplotype assembly of autotetraploid potato using integer linear programing. Bioinformatics 35:3279–3286. https://doi.org/10.1093/bioinformatics/btz060

54. Hermisson J, Pennings PS (2005) Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics 169:2335–2352. https://doi.org/10.1534/genetics.104.036947

55. Morgan C, Zhang H, Henry CE, et al (2020) Derived alleles of two axis proteins affect meiotic

traits in autotetraploid Arabidopsis arenosa. PNAS. https://doi.org/10.1073/pnas.1919459117

56. Lee KM, Coop G (2017) Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data. Genetics 207:1591–1619. https://doi.org/10.1534/genetics.117.300417

57. Gautier M (2015) Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics 201:1555–1579. https://doi.org/10.1534/genetics.115.181453

58. Cheng JY, Racimo F, Nielsen R (2019) Ohana: detecting selection in multiple populations by modelling ancestral admixture components. bioRxiv 546408. https://doi.org/10.1101/546408

59. Szpak M, Mezzavilla M, Ayub Q, et al (2018) FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biology 19:5. https://doi.org/10.1186/s13059-017-1380-2

60. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654. https://doi.org/10.1038/351652a0

61. Charlesworth J, Eyre-Walker A (2008) The McDonald–Kreitman Test and Slightly Deleterious Mutations. Molecular Biology and Evolution 25:1007–1015. https://doi.org/10.1093/molbev/msn005

62. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227–1234

63. Zhang L, Li W-H (2005) Human SNPs Reveal No Evidence of Frequent Positive Selection. Molecular Biology and Evolution 22:2504–2507. https://doi.org/10.1093/molbev/msi240

64. Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution 30:1687–1699. https://doi.org/10.1093/molbev/mst063

65. Meirmans PG, Van Tienderen PH (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity 110:131–137. https://doi.org/10.1038/hdy.2012.80

66. Cingolani P, Platts A, Wang LL, et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92. https://doi.org/10.4161/fly.19695

Table 1 Nucleotide diversity (π or Θ_π) and Tajima's D calculated over ~200k putatively neutral four-fold degenerate sites genome-wide in the four *A. arenosa* example populations (subsampled to seven individuals per population at each site). Population codes correspond to Fig. 1.

population	ploidy	type	nucleotide diversity	Tajima's *D*
VEL	2x	alpine	0.049	0.10
SUB	2x	foothill	0.055	-0.01
TIS	4x	foothill	0.059	0.09
BAL	4x	alpine	0.054	0.44

Table 2 Population differentiation metrics calculated over ~200k four-fold degenerate sites genome-wide in a pairwise manner between the four *A. arenosa* example populations.

Pop. pair	ploidy contrast	*Rho*	F_{ST}	Fixed differences[1]
SUB-VEL	2x-2x	0.17	0.12	63
SUB-BAL	2x-4x	0.25	0.15	82
SUB-TIS	2x-4x	0.23	0.13	14
VEL-BAL	2x-4x	0.26	0.15	33
VEL-TIS	2x-4x	0.26	0.15	25
TIS-BAL	4x-4x	0.22	0.09	3

[1]Total number of fixed differences between the corresponding populations.

Table 3 Reference to programs noted in the 'Best practices' examples

Software/method	Purpose	Link
GATK	variant calling & genotyping	https://gatk.broadinstitute.org/hc/en-us
GATK-wrapper	variant calling & genotyping pipeline for parallelized analysis on a cluster	https://github.com/vlkofly/Fastq-to-vcf
EBG	polyploid genotyping	https://github.com/pblischak/polyploid-genotyping
BCFTOOLS	filtering and manipulation of VCF files	http://samtools.github.io/bcftools/bcftools.html
SCANTOOLS	population genomic analysis using large genomic data sets	https://github.com/mbohutinska/ScanTools_ProtEvol
GENODIVE	population genetic analysis of diploids and polyploids; data conversion	https://www.bentleydrummer.nl/software/software/GenoDive.html
ENTROPY	inferring ancestry proportions of individuals to genetic groups, tailored to mixed-ploidy systems and large SNP data	https://anaconda.org/bioconda/popgen-entropy
STRUCTURE	inferring ancestry proportions of individuals to genetic groups	https://web.stanford.edu/group/pritchardlab/structure.html
STRUCTURE-input	data conversion VCF file -> Structure format	https://github.com/MarekSlenker/snipStrup
TREEMIX	inferring population relationships	https://bitbucket.org/nygcresearch/treemix/wiki/Home
FASTSIMCOAL2	coalescent simulations	http://cmpg.unibe.ch/software/fastsimcoal2/
McDonald-Kreitman test	selection test	http://benhaller.com/messerlab/asymptoticMK.html

Table 4 Three different selection scans designs, varying in the combination of diploid and polyploid populations and divergence between them.

Selective factor	**Ploidy contrast**	**Divergence**	**Focus**	**Methods**	**Validation**	**Ref**
Whole-genome duplication	2x-4x	~20,000	protein evolution	divergence scan on amino acid substitutions, FineMAV	functional assays of two candidates	[6, 28, 44]
Alpine environment	2x-2x, 4x-4x	~10,000	gene evolution	BayPass, divergence scan	modelling (signal confirmed in 151 out of 190 candidates)	[4]
Serpentine soil	4x-4x	~ 5,000	gene evolution	environmental association analysis, divergence scan	modelling (signal condirmed in 19 out of 61 candidates)	[5, 42]

Note: Ploidy is given for the focal contrast(s). Divergence between adapted and background populations is shown, in generations. Distinguishing among Modes of Convergent adaptation (DMC, [56]) has been used to formally test selection signal against a neutral scenario for each candidate.