DOI: 10.1038/s41559-019-0807-4. https://www.nature.com/articles/s41559-019-0807-4

Final published version available via SharedIt read only link: https://rdcu.be/borZY

Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*

Patrick Monnahan[1], Filip Kolář[2,3,4], Pierre Baduel[1], Christian Sailer[1], Jordan Koch[1], Robert Horvath[5], Benjamin Laenen[5], Roswitha Schmickl[2,4], Pirita Paajanen[1], Gabriela Šrámková[2], Magdalena Bohutínská[2,4], Brian Arnold[6], Caroline M. Weisman[7], Karol Marhold[2,8], Tanja Slotte[5], Kirsten Bomblies[1], and Levi Yant[1, 9, *]

1. Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK
2. Department of Botany, Faculty of Science, Charles University, Benátská 2, 128 01 Prague, Czech Republic
3. Department of Botany, University of Innsbruck, Sternwartestraße 15, A-6020 Innsbruck, Austria
4. Institute of Botany, The Czech Academy of Sciences, Zámek 1, 252 43 Průhonice, Czech Republic
5. Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden
6. Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115 USA
7. Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA, 02138 USA
8. Plant Science and Biodiversity Centre, Slovak Academy of Sciences, Dúbravská cesta 9, SK-845 23 Bratislava, Slovak Republic
9. School of Life Sciences and Future Food Beacon, University of Nottingham, Nottingham, UK

Patrick Monnahan, Filip Kolář, and Pierre Baduel contributed equally.

***Author for correspondence:** levi.yant@nottingham.ac.uk; Tel: 0749 025 3006

Date of submission: 2 August 2018
Word count (Introduction, Results & Discussion): 3,455
Keywords: polyploidy; selection; introgression; population genetics; evolution
Running title: Population genomic consequences of autopolyploidy

Abstract

Ploidy-variable species allow direct inference of the effects of chromosome copy number on fundamental evolutionary processes. While an abundance of theoretical work suggests polyploidy should leave distinct population genomic signatures, empirical data remains sparse. We sequenced ~300 individuals from 39 populations of *Arabidopsis arenosa*, a naturally diploid-autotetraploid species. We find the impacts of polyploidy on population genomic processes are subtle yet pervasive, including reduced efficiency of purifying selection, differences in linked selection, and rampant gene flow from diploids. Initial masking of deleterious mutations, faster rates of nucleotide substitution, and interploidy introgression likely conspire to shape the evolutionary potential of polyploids.

Introduction

Whole-genome duplications (WGD) have occurred throughout the tree of life [1, 2] and are associated with biological phenomena of great socio-economic importance such as crop domestication [3] and carcinogenesis [4]. The direct effects of WGD or polyploidy can be far-reaching, ranging from cellular [5] through organism-level phenotypes [6], up to population and ecosystem-level processes [7-9].

Population genetic theory predicts substantive effects of ploidy on both neutral and selective processes [10-16]. With higher ploidy, neutral diversity is expected to rise, while the rate of population differentiation due to genetic drift will slow [17]. Given the additional chromosomal partners available during recombination, linkage disequilibrium should decrease and haplotype diversity correspondingly increase. Polyploidy also has unique effects on migration. Although polyploidization is traditionally viewed as a means of instant speciation [18-20], the ploidy barrier may be permeable, particularly from diploids to polyploids [21, 22]. Additionally, polyploids may lack reproductive incompatibilities found in diploid progenitors [23]. Upon secondary contact, interploidy introgression could then further enrich polyploid diversity.

The effect of ploidy on selective processes can be primarily attributed to differential manifestation of allelic dominance. Added masking of deleterious alleles should elevate equilibrium frequencies at mutation-selection balance, potentially increasing genetic load [24]. Similarly, beneficial alleles are not observed as readily in polyploids, which slows the fixation of individual alleles [25]. These disadvantages can be mitigated by polyploids' propensity to receive, maintain, and generate genetic variation. By itself, the increased rate at which beneficial mutations are introduced in polyploids can be sufficient for a faster overall rate of adaptation [16, 26]. Additionally, introgression is increasingly being recognized as an important source of adaptive variation [27, 28]. Though genomic evidence in a ploidy-variable system is lacking, a greater tendency for polyploids to accept variation from locally adapted populations of the same or different ploidy (or even species), may facilitate adaptation and expansion of polyploid lineages.

Lack of population genomic data is particularly pronounced for autopolyploids, which arise from within-species WGD [29]. In contrast to the better studied allopolyploids, where effects of polyploidy are confounded with subgenome divergence, autopolyploids allow direct investigations of the role of polyploidy *per se*. Using a new model for autopolyploidy, *Arabidopsis arenosa* [30], we generated the most comprehensive range-wide genomic dataset to date of a natural tetrasomic autotetraploid [31, 32] (182 individuals / 24 populations) and its diploid sister lineages (105 / 15; Figure 1a). The tetraploids, whose ecological niche largely overlaps with the genetically divergent diploids, trace to a single origin (~30 kya) [31] and subsequently spread across much of Europe, occasionally coming into secondary contact with diploids [31, 33].

We focus on three main questions concerning the genomic impact of selection and migration in this system: First, we investigate if purifying selection is relaxed in autotetraploids as predicted from the increased masking of deleterious alleles. Second, given the inherent effects of polyploidy on processes governing diversity and recombination, we ask whether signals of linked selection markedly differ in one ploidy versus the other. Lastly, we focus on two independent contact zones to assess the impact of interploidy gene flow on polyploid evolution. Overall, our empirical analyses provide insights into the complexity of autopolyploid evolution, supporting some but not all theoretical predictions. Altered selective processes and introgressions shape the genomic landscape of tetraploids, and perhaps their evolutionary potential as well.

Results

High diversity and population differentiation in natural A. arenosa

As previously reported [34], the diploid populations form five divergent, geographically-separated groups (Fig. 1b, c): the *Baltic* lineage, the highly distinct *Pannonian* and *Dinaric* lineages ($\overline{F_{ST}}$ = 0.31 and 0.34, respectively), and the less differentiated Southern Carpathian (*S. Carp.*), and Western Carpathian (*W. Carp.*) lineages ($\overline{F_{ST}}$ = 0.25, with evidence of past, Table S4, and recent hybridization, e.g. HNI Fig 1b). The tetraploids comprise four lineages: *S. Carp., W. Carp, C. Europe* (Alps and western Central Europe), and the *Ruderal* lineage. The latter group is the most widespread yet ecologically distinct, occupying man-made sites (e.g. railways) from southern Germany to Sweden (Fig. 1a). Ploidy is explicitly indicated as a suffix (*2x* or *4x*) hereafter. We find lower differentiation among tetraploid populations than diploid ones (Table 1, S3, Fig. 1c, d), in line with the greater age of diploids and the neutral expectation that, all else equal, the rate of drift is halved relative to diploids [17]).

Arabidopsis arenosa is an obligate outcrosser, and all populations exhibit high genome-wide diversity ($\bar{\theta}_{\pi}$ = 0.015, Table 1), an order of magnitude higher than the predominantly self-fertilizing *A. thaliana* [35]. All else equal, polyploidy is expected to increase diversity ($8N_e\mu$ in tetraploids versus

$4Ne\mu$ in diploids). Although tetraploid populations exhibit slightly higher Watterson's θ_W at non-synonymous sites (zero-fold degenerate sites, 0-dg), we observe no significant increase of θ_π or θ_W in tetraploid populations at putatively neutral sites (four-fold degenerate sites, 4-dg). These results were robust to exclusion of tetraploid populations with evidence of interploidy admixture (DRA, LAC, TZI, KOW, STE, TBG; Table S2). Such an impact of genome duplication on θ_W (in contrast to θ_π) is consistent with a recent origin of tetraploids, as θ_W is more sensitive to accumulation of rare variants.

Ploidy effects on purifying selection

Though we find only mild differences between ploidies in θ_π or θ_W , we observe a highly significant difference for the ratio of 0-dg θ_W to 4-dg θ_W (Wilcoxon rank-sum test, W=54, $p = 0.001$; Table 1, S1), which is consistent with expectations of relaxed purifying selection in tetraploids [24]. To further explore this hypothesis, we assessed how gene-level diversity varied with gene expression, a proxy for selective constraint (e.g. [36-38]). We confirmed that highly expressed genes exhibit reduced nonsynonymous diversity in both ploidies (multiple linear model, MLM; $p < 0.0001$, F-test of expression effect on θ_W at 0-dg sites and the 0-dg/4-dg ratio of θ_W; Fig. 2a, 2b and Table S5). However, the 0-dg/4-dg θ_W ratio was generally higher in tetraploids ($p < 0.0001$) due to elevated nonsynonymous diversity (α_p coefficient in MLM; $p < 0.0001$, Table 1, Fig. 2b, S5). These results were robust across data subsets and upon including various cofactors (e.g. population sizes; Tables S5 & S6). This confirms that, beyond the increased mutational input resulting from doubled genome copies in tetraploids, there is an additional increase of non-synonymous diversity (thus increasing the 0-dg/4-dg diversity ratio), which likely reflects an overall relaxation of purifying selection.

Such relaxation could be due to either a reduction in the *strength* of selection or simply because selection is less efficient in tetraploids. If mutations are purely recessive, the homozygotes bearing the deleterious phenotype are much less frequent in tetraploids (q^2 versus q^4, assuming random mating

[32]), making purifying selection inefficient relative to diploids even if the fitness cost of the mutant homozygote (i.e. selection *strength*) is equivalent across ploidies. To distinguish between these two explanations, we evaluated the distribution of fitness effects (DFE) across both ploidies, finding no apparent differences in the strength of purifying selection in diploid vs. tetraploid populations (Fig. 2d). From this analysis, it seems purifying selection is not weaker *per se*, but rather less efficient at reducing allele frequencies because deleterious mutations are better masked in autotetraploids.

However, we note two important assumptions in DFE estimation methods [39] that complicate interploidy comparisons. First, a diploid model of allele frequencies at mutation-selection-drift balance is assumed. Since frequencies are expected to be higher in autotetraploids [15], this model would be biased towards inferring weaker selection than necessary to explain polyploid data. Second, deleterious mutations are assumed to be additive. If deleterious mutations are recessive, equilibrium frequencies can be orders of magnitude greater in tetraploids, further amplifying the first bias. If purifying selection were truly weaker in tetraploids, these biases would make this more apparent; instead, we find no evidence for ploidy differences in the DFE (Fig. 2d, S7 and Table S7 and S14).

In the long run, these selective effects (along with increased mutational input) are expected to result in higher genetic load for tetraploids (under partial recessivity; load should be equivalent at equilibrium for complete recessivity) [24]. To obtain a crude estimate of genetic load in each population, we counted homozygous genotypes per-individual for *derived*, nonsynonymous alleles. Under complete recessivity, the estimated load is currently lower in tetraploids than in diploids (Wilcoxon rank-sum test of population means, W = 264, $p < 0.0001$ and W = 195, $p < 0.0001$ with/without interploidy-admixed tetraploid populations, respectively; Fig. 2c). However, the relatively young tetraploid lineages may not have reached equilibrium [31], which could take hundreds of thousands of generations [16]. Furthermore, the actual load may be substantially higher in tetraploids if deleterious mutations are at least partially recessive [24]. Unfortunately, current methodologies do not allow for relaxation of the assumption of complete recessivity for tetraploids.

Ploidy effects on positive and linked selection

If dominance when in single-copy is comparable across ploidy, such that *Aa* and *Aaaa* genotypes are equivalent, the greater mutational opportunity in tetraploids should ultimately lead to higher rates of adaptation [16]. Using DFE-alpha analysis [38], we estimated the proportion of nonsynonymous (0-dg) sites fixed by positive selection in each population. Using either α or ω_{α}, this proportion was significantly higher in tetraploid populations (W = 14 or 6, respectively; $p < 0.0001$ for both; Fig. 2e, S8 and Table S7 and S14), possibly indicating increased rates of adaptive substitution. This does not simply reflect admixture (below), as the difference remained significant when we removed the six tetraploid populations admixed with diploids (Table S7). However, similar to the preceding section, multiple non-selective processes can lead to mis-estimation of parameters within DFE-alpha and similar methods [40-42] (discussed in Supplementary Text 3).

Although DFE-alpha suggests a higher proportion of adaptive substitutions in tetraploids, the fixation of particular mutations is generally expected to take longer [25], with implications for the degree that linked selection reduces diversity during selective sweeps. Using the average squared genotypic correlation between SNPs, we approximated linkage disequilibrium (LD) (Fig. 3a), finding an overall reduction in tetraploids (50% lower mean correlations at 1kb distance in tetraploids). We then assessed the impact on linked selection by comparing across genomic windows excess nonsynonymous divergence ($E_{NS} = d_N - d_S$) and 4-dg site diversity (Fig. 3b). Regardless of ploidy, E_{NS} and 4-dg θ_{π} were consistently negatively correlated (Table S8), suggesting that divergent selection had reduced diversity at linked, neutral sites. The parabola shape ($p < 0.001$ for quadratic term; Table S8) indicates that diversity is also reduced for E_{NS}<<0 regions (i.e. those under background selection). The reductive effect of E_{NS} on neutral diversity was significantly stronger in gene-dense regions (upper

20%, Fig. 3d; interaction of E_{NS} and gene-density: $p < 0.001$ in Table S8) than in gene poor regions (lower 20%, Fig. 3c). Within gene-dense regions, we observed higher neutral diversity in tetraploids in particular for negative E_{NS} values (Fig. 3d, p=0.002 for 3-way interaction between E_{NS}, gene-density, and ploidy, Table S8). This difference in slope suggests background selection is less effective at reducing diversity in tetraploids, while selective sweeps reduce diversity similarly across ploidies. No such differences were observed in low gene-density regions (Fig. 3c), where linked selection will be less pronounced.

While slower fixation times in tetraploids would dampen a signature of linked selection, two factors could effectively counter this effect: 1.) the evolution of reduced per-base recombination in tetraploids (to avoid deleterious multivalents forming during meiosis [58]) and 2.) systematic differences across ploidies in the age of selective sweeps (due to the comparatively recent tetraploid formation). Such reduced recombination is not evident, genome-wide, in tetraploids. In fact, our LD approximation is generally lower in tetraploids, reflecting a higher population recombination rate ($\rho = 8N_er$ in tetraploids and $\rho = 4N_er$ in diploids; Fig. S22) and/or the more recent population expansion [39]. Unfortunately, the lack of genetic maps and of a workable phasing algorithm prevents inclusion of the recombination landscape in our regression modelling approach. Furthermore, estimation of the age and strength of selection is not currently possible on a genomic scale. Understanding the interplay between fixation times, recombination landscapes, and natural history will be the focus of future investigations.

Single origin of tetraploids and interploidy introgression

Although previous work supported a single tetraploid origin in the W. Carpathians [31], local tetraploids clustered genetically with locally co-occurring diploids in two parallel cases (Southern Carpathians and Baltic coast; Fig. 1a,c, S3, S4B). This might suggest multiple tetraploid origins followed by widespread gene flow among tetraploids, as these two tetraploid lineages still share a sizeable portion of polymorphisms with the widespread tetraploid lineages (*W. Carp.-4x*, *C. Europe-4x*, Fig. 1b). However, we find multiple lines of evidence supporting a single tetraploid origin followed instead by interploidy gene flow from locally co-occurring diploids (see Supplementary Text 1 for further discussion). First, coalescent simulations (*fastsimcoal2*) consistently favour scenarios with a single tetraploid origin (~20k – 31k generations ago) followed by admixture (Fig. 4a,b, S9, S10; Table S9). Second, frequencies of alleles diagnostic of the putative diploid ancestor of all tetraploids (*W. Carp.-2x* lineage) are elevated and positively correlated across all tetraploid populations (Fig. 4c, S11). Finally, alleles of several key meiosis genes are shared among all tetraploids, yet consistently divergent from diploids (Fig. 4d, S12).

Interploidy gene flow could be mediated either by viable triploids (virtually absent in natural *A. arenosa* [32]) or by a one-step production of tetraploid hybrids via merger of unreduced gametes of a diploid (2n) with a normal (reduced) gamete of a tetraploid (also 2n) [52]. In the Southern Carpathian contact zone 37% of the *S. Carp.-4x* possessed the regionally-specific plastid haplotypes typical for the *S. Carp.-2x*, suggesting gene flow between ploidies sometimes involves diploid mothers. *Ruderal-4x* populations, on the other hand, only shared plastid haplotypes with other tetraploid groups (*W. Carp.-4x* and *C. Europe-4x*); in this contact zone, either selection favours tetraploid plastids, or gene flow primarily involves male gametes from diploids. In addition, multiple tetraploid (but no diploid) populations showed elevated frequencies of nuclear (Fig. S13) and occasionally plastid (Fig. S14) markers otherwise private to *Arabidopsis lyrata* – a partially sympatric species known to hybridize with *A. arenosa* at the tetraploid but not diploid level [23, 43].

The maintenance of tetraploid alleles at key meiosis genes in the face of introgression from diploids implies that some genomic regions are more or less resistant to interploidy admixture. To identify such regions in each contact zone, we first evaluated the weights of topologies supporting tetraploid monophyly (TM) vs. local admixture (LA) in windows across the genome (Fig. 5, S15). Generally, the tendency was for no single topology to dominate a particular window, yet occasionally, we observed windows where the vast majority of weight was given to either the TM or LA topology. Within the latter regions, we then looked for the specific pattern of: 1) reduced genetic divergence of the focal tetraploid lineage to sympatric diploids versus non-sympatric diploids (as expected with local admixture) and 2) elevated genetic divergence of the focal tetraploid to all other tetraploids. Lastly, we looked for signatures of positive selection using Fay and Wu's H. In each contact zone, we identified multiple regions with such three-fold evidence (Fig. 5; Fig. S16). Within the 1% outliers for both LA topology weight and H, we found a number of gene coding loci (Table S10) with some indication of functional enrichment (see Supplementary Text 2). Conversely, windows with high weight given to the TM topology (Topology 1) often exhibited elevated divergence to *all* diploids and non-elevated divergence to tetraploids. Additionally, these windows often included meiotic genes previously identified as exhibiting the strongest signatures of tetraploid-specific selection in a subset of *A. arenosa* populations [58]. Together, this is consistent with a strong tetraploid resistance to diploid introgression in these regions, suggestive of their *ongoing* role in the maintenance of stable autopolyploid chromosome segregation.

Discussion

Using the largest population resequencing dataset to date of a ploidy-variable plant species, we observe pervasive differences in how forces governing genome evolution shape genetic diversity and divergence in nature. In diploid and autotetraploid *A. arenosa*, we find subtly distinct signatures of linked and purifying selection. Additionally, multiple sources of evidence indicate substantial introgression from diploids to tetraploids. We discuss these results in terms of the inherent effects of genome doubling and the implications for the evolutionary potential of polyploid lineages.

The effects of genome doubling on selective processes, and consequently the patterns of genomic variation they leave behind, are multifarious and sometimes counter-acting, making it difficult to observe and distinguish individual causes. The challenge is heightened by the lack of methodologies generalized for higher ploidy and the potential for demographic events associated with the creation, establishment, and expansion of nascent tetraploids to obfuscate the genomic signals of selection. Yet, the fundamental impact of genome doubling on dominance relationships and mutational and recombinatorial opportunity are clearly reflected in our analyses of linked and purifying selection.

For positive selection, the increased masking of a beneficial mutation's effect in tetraploid populations is likely not sufficient to slow adaptation. The higher estimated proportion of nonsynonymous polymorphisms fixed by positive selection in tetraploids (Fig. 2e) supports the notion that increased mutational input is sufficient to overcome any hindrance to adaptation posed by the reduced efficiency of selection [16, 26, 44]. Furthermore, increased fixation times (via increased masking) and mutational and recombinatorial opportunity in tetraploids promote retention of haplotype (Fig 3a) and nucleotide diversity (Fig 3d) following selection. *A. arenosa* tetraploids expanded well beyond their ancestor diploid's range, including postglacial and man-made habitats [34]. Increased mutational input and retention of diversity may aid polyploids in adapting to the fluctuating or otherwise challenging environments that are often associated with polyploids [7, 45, 46].

With purifying selection, nonsynonymous polymorphism is governed simultaneously by selection against deleterious alleles and their recurrent introduction via mutation; genome doubling favouring increased polymorphism in both cases. Importantly, the former may result solely from the added masking of recessive deleterious mutations in heterozygous genotypes; the *strength* of selection need not differ. In this context, increased diversity in tetraploids is detrimental, leading to higher genetic load at equilibrium [24]. Our estimate of genetic load (assuming recessivity) is currently lower for tetraploids, even though nonsynonymous diversity is higher for genes under purifying selection. In addition to reasons discussed above (see *Results*), double reduction (a unique phenomenon in autopolyploids where the resolution of multivalents occasionally causes sister chromatids to segregate into the same gamete) may also play a role, by increasing homozygosity and allowing more efficient purging of deleterious alleles, although this would only affect more distal chromosome regions [47]. Furthermore, the actual load in tetraploids could be much higher if deleterious alleles are at least partially recessive [24], as has been demonstrated previously in a natural plant system [48]. Currently, no comparable demonstration exists for autotetraploids, although ploidy-variable species, such as *A. arenosa*, provide a compelling system in which this could be addressed.

Despite an increased recognition of adaptive introgression [49], introgression between ploidy cytotypes could be maladaptive [50, 51]. Here, the most salient example lies in meiotic genes, which have been shown to exhibit the strongest signatures of selection in tetraploids (presumably to promote proper segregation of additional chromosomes to gametes [58]). The introduction of diploid-like meiotic alleles into a tetraploid population would increase the frequency of multivalent formation, thus decreasing fitness. In line with this, meiotic genes consistently show the strongest signatures of introgression resistance in tetraploids: elevated divergence between ploidies, reduced diversity within tetraploids, and tetraploid monophyly in both diploid-tetraploid contact zones (Figs. 4d, 5). On the other hand, we found coding regions with diploid-like derived alleles that have swept to higher frequencies in co-occurring tetraploids (Fig. 5), implying that interploidy introgression can be adaptive in tetraploids. In fact, the most widespread tetraploid lineage (*Ruderal-4x*), which evolved a different, weedy, life strategy [52], colonizing man-made habitats across central and northern Europe [53], is the only lineage with traces of introgression from both a distinct diploid *A. arenosa* lineage (*Baltic-2x*), [54] as well as another species – *A. lyrata* (Figs. 1, S13). Overall, this points to the ability of tetraploids to accumulate diversity from various lineages, while retaining essential tetraploid- or locality-specific adaptations.

Much work remains to understand the drivers of successful establishment and spread of newly formed polyploid lineages. Relative to ecological explanations [55, 56], population genomic processes have not been thoroughly assessed in natural populations despite being invoked [7, 16]. Our results provide empirical insight, generally supporting pervasive yet subtle effects of ploidy on certain neutral and selective processes. Despite slightly increased nonsynonymous diversity, tetraploids may still be benefiting from the masking of potentially deleterious recessive mutations, and also exhibit consistently higher frequencies of adaptive nonsynonymous substitutions. Finally, multiple events of strong introgression into tetraploids may provide additional substrate for local adaptation. This supports the view of polyploids as diverse and adaptable evolutionary amalgamates from multiple distinct ancestral lineages [57].

Online Methods

Plant Material and Library Preparation

In addition to eight previously sequenced populations, [32, 58, 59] we collected 31 new populations throughout the distribution range of *A. arenosa* (see Table S11 and Fig. S17) and its closest relative, *A. croatica*. We aimed to cover each main evolutionary lineage distinguished by previous RADseq studies [31, 34] by multiple populations, and also representatively cover the ploidy level (15 diploid, 24 tetraploid populations), altitudinal, (range 1 – 2,240 m a.s.l.) and edaphic variation (17 calcareous, 21 siliceous, 1 serpentine substrate).

We extracted DNA from silica-dried leaf tissue according to a CTAB protocol [60] with the following modifications: 75 – 100 mg of dry leaf tissue were ground in 2 mL tubes (Retsch swing mill), 200 units of RNase A per extraction were added to the isolation buffer, and the DNA pellets were washed twice with 70% ethanol. DNA was resuspended in 50 µL TE-buffer for storage, and small fragments were removed using Agencourt AMPure XP beads (Beckman Coulter, Massachusetts, USA) following the manufacturer’s instructions with 0.4x DNA:beads ratio.

We quantified the extracted gDNA using the dsDNA HS assay (Q32854) from ThermoFisher Scientific (Life Technologies Ltd. Paisley, UK) with their Qubit 2.0 or 3.0 (Q33216). We prepared Illumina (Illumina United, Fulbourn, UK) Nextera XT (FC-131-1024) and TruSeq PCR-free (FC-121-3003) sequencing libraries for 350 bp insert length of genomic DNA, as well as Nextera sequencing libraries (FC-121-1030). For PCR free libraries we used 300 to 500 ng DNA as input instead of the recommended 1 µg. We quantified the NGS libraries using Qubit as described above.

Sequencing and Variant Calling

We multiplexed libraries based on Qubit concentration and ran those pools on an initial quantification lane. According to the yields for each sample, we increased loading of the same multiplex-mix on several lanes to achieve a minimum of 10× coverage, based on the number of raw reads. Samples that had less than our target coverage were remixed and run on another lane (top-up lane). We sequenced 125 bp pair end reads on Illumina's HiSeq 2500 platform for all sequencing runs.

Our data processing pipeline involved three main parts: 1) Preparing the raw sequencing data, 2) Mapping and re-aligning the sequencing data and 3) Variant discovery (GATK *v.3.5*, following GATK Best Practices). All steps and parameters are summarised in File S2. To prepare the raw sequencing data for mapping we concatenated the fastq.gz files from the different sequencing lanes, followed by trimming off the adapter sequence from reads that had inserts shorter than 250 bp, using cutadapt *v.1.9* [61]. We mapped the reads to a North American *Arabidopsis lyrata* reference genome [62] using bwa [63]. At this stage, we added *A. arenosa* sequencing data from previous studies [32, 58, 59]. For Nextera (PCR-based) libraries, we removed duplicated reads using 'MarkDuplicates' from picard-tools 1.134 [64] followed by 'AddOrReplaceReadGroups' to add read groups and indices to the bam files. We then used GATK *v.3.5* 'RealignerTargetCreator' and 'IndelRealigner' [65] to re-align the reads around indels. Prior to variant discovery, we excluded individuals that had less than 40% of bases < 4× coverage (assessed via GATK 'DepthOfCoverage' with the restriction to a minimum base quality of 25 and a minimum mapping quality of 25). Our final dataset for analysis contained 287 *A. arenosa* and four *A. croatica* individuals from 39 populations (see File S1 for population details and File S3 for a summary of processing quality assessments).

We called variants for the 291 bam files (287 *A. arenosa* and four *A. croatica*) using 'HaplotypeCaller' and 'GenotypeGVCFs' (GATK *v.3.5*). For each bam file, 'HaplotypeCaller' was run in parallel for each scaffold with ploidy specified accordingly and retaining all sites (variant and non-variant). We combined the single-sample GVCF output from HaplotypeCaller to multisample GVCFs and then ran 'GenotypeGVCFs' to jointly genotype these GVCFs, which greatly aids in distinguishing rare variants from sequencing errors. Using GATK's 'SelectVariants', we first excluded all indel and mixed sites and restricted the remaining variant sites to biallelic. Second, we removed sites that failed GATK Best Practices quality recommendations (QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, HaplotypeScore < 13.0). Third, we masked genes that showed excess heterozygosity (fixed heterozygous in at least five SNPs in two or more diploid populations) in the dataset, i.e. potential paralogues mapped on top of each other. At the same step, we masked sites that had excess read depth that we defined as 1.6× the second mode (with the first mode being low coverage sites indicative of mismapping) of the read depth distribution (DP > 6400).

Polarization and Variant Classification

We repolarized a subset of sites using a collection of genotyped individuals across closely related diploid *Arabidopsis* species thus avoiding polarization against a single individual (the reference genome, N. American *A. lyrata*). We used two individuals from each of the following diploid *Arabidopsis* species (genotyped in the same way as our *A. arenosa* samples): European *A. lyrata*, *A. croatica*, and *A. halleri*. For a site, we considered only species with complete genotypes and only considered a site with at least two species represented. We required the alternative allele frequency to be > 0.5 in each species, if all species were represented at a site. However, if only two species were represented, we doubly weighted allele frequency for the species by preferring species with expected higher genetic variation of its European populations (i.e. with decreasing priority for *A. halleri* > *A. lyrata* > *A. croatica*) and required mean allele frequency to be > 0.5. In total, this identified ~145,000 sites for repolarization. We classified sites as 4-fold (4-dg) or 0-fold (0-dg) degenerate based on their position in the *A. lyrata* gene model annotation Araly1_GeneModels_FilteredModels6.gff. 0-dg sites are those where any mutation is expected to result in an amino acid change, and 4-dg are the opposite (same amino acid regardless of mutation).

Population Structure

We inferred relationships among the 39 *A. arenosa* and one *A. croatica* populations (the full dataset, as well as each separate ploidy) based on putatively neutral 4-fold degenerate SNPs. Synonymous sites are not necessarily free of constraints, e.g. due to potential codon usage bias, but are nevertheless the closest to effectively neutral of any site class in the genome [66]. After quality filtering our demographic analysis is based on a genome-wide dataset consisting of 1,350,328 four-fold degenerate SNPs, allowing for a maximum of 10% missing alleles per site (1.2% missing data). Firstly, we calculated principal component analysis (PCA) using *glPCA* function in *adegenet* [67] replacing the missing values (1.2% in total) by average allele frequency for that locus. Next, we calculated Nei's [68] distances among all individuals in *StAMPP* [69] and displayed it using the neighbour network algorithm in *SplitsTree* [70]. Third, we selected the 553 (503 for the diploid-only dataset) most parsimony-informative genes based on the following criteria: 1.) for each accession, we excluded genes with ≥ 10% missing data, and 2.) we excluded genes with ≥ 10% missing accessions. We constructed a maximum likelihood tree from each gene using *RAxML v.8* [77] with model GTRCAT and 100 (rapid) bootstrap replicates [77]. In each gene alignment for *RAxML*, accessions were represented by the consensus sequence, with different alleles represented as ambiguous sites in the consensus sequence. Ambiguous sites are treated by *RAxML* as invariant sites, hence, the standard nucleotide substitution model needed to be utilized; the ascertainment bias correction model that is usually used for SNP matrices is not appropriate in such case. The resulting gene trees were summarized under the multispecies coalescent using *Astral v.4.10.10* [78]; bootstrapping was performed with 100 replicates each.

We further determined grouping of the populations using three clustering approaches: model-based Bayesian clustering using *fastStructure v.1.0* [71] and STRUCTURE *v.2.3.2,* [72] and a non-parametric k-means clustering using *adegenet* [67]. The analyses were performed separately for (i) the entire data set of *A. arenosa* (*A. croatica* excluded; 9,543 SNPs after random thinning over windows of 50 kb to reduce effect of linkage and removing singletons, 2.4% of missing data), (ii) diploids only (12,655 SNPs, 4.1% missing data) and (iii) tetraploids only (9,596 SNPs, 2.3% missing data). In *fastStructure*, five replicate runs for K (number of groups) ranging from 1 to 10 were carried out under default settings. We selected the optimal K value based on the similarity coefficient (~1 for optimal K [73]) across replicates (Fig. S18). As *fastStructure* does not handle polyploid genotypes, we randomly subsampled two alleles per each tetraploid locus (following [74]) using a custom script. To check for the effect of such subsampling, we also ran the original STRUCTURE program, which handles mixed-ploidy datasets, for optimal K values according to *fastStructure*. We ran the admixture model with uncorrelated allele frequencies using a burn-in of 100,000 iterations followed by 1,000,000 additional iterations. Finally, we ran k-means clustering using 1000 random starts and selected the partition with the lowest Bayesian information criterion (BIC) value. Population groupings were consistent across algorithms (Fig. 1b, c, S1 – S5), although some methods identified finer sub-structure within the *S. Carp.-2x* and *C. Europe-4x* lineages (Fig. S5).

We used Treemix *v.1.3* to infer migration events and relationships between the 39 *A. arenosa* populations using one *A. croatica* population as outgroup. We used the 4-dg sites to build a tree without any migration events and used this tree as basis for migration models to make comparisons easier (option '-g'). We modelled zero to eight migrations and graphically assessed the residuals after each additional migration modelled, using the R-scripts supplied with the Treemix package. If specific population pairs had high residuals, we modelled an additional migration event. We continued until the residuals were small and evenly spread across population pairs and/or until an additional migration event involved the outgroup (we consider this admixture unlikely due to very local occurrence and spatial isolation of the *A. croatica*).

To quantify differentiation among populations, we calculated genome-wide F_{ST} and Rho coefficients (similarly as in the window-based analyses described below) and performed analysis of molecular variance (AMOVA) based on the Nei's distances using the *amova* function in the *pegas* R package [75]. We tested for isolation by distance relationships through comparison of matrices of geographic and genetic (Nei's among-population) distances among the populations using *mantel.randtest* function in *ade4* R package [76]. For each tetraploid population, we calculated the frequency of alleles diagnostic to each diploid lineage. The allele was defined as diagnostic if it exhibited minimum frequency 0.3 (to avoid including sequencing errors as diagnostic alleles) in that diploid lineage and was absent in any other diploid lineage (except for the putatively admixed Baltic diploids, Table S13). For all populations we also calculated frequency of *A. lyrata*-like alleles, i.e. reference alleles that were otherwise rare in the complete *A. arenosa* dataset (a rarity cut-off of 6.8%, i.e., equivalent to two tetraploid populations of 8 individuals). As these alleles were nearly absent in *A. arenosa* diploid populations, i.e. the ancestors of tetraploids, we assume they more likely represent hybridisation from *A. lyrata* than ancestral variation shared among both species.

Finally, we inferred phylogenetic relationships among plastomes of our samples and previously published plastomes of other *Arabidopsis* species [74]. We mapped the reads to a custom *A. arenosa* plastome assembly constructed using org.ASM (http://pythonhosted.org/ORG.asm/) and performed variant calling and filtration as described above, with the exception of setting ploidy = 1 in GATK *HaplotypeCaller* and retaining SNPs and invariant sites with depth > 4 in at least 90% of the individuals. We aligned all sequences using *Mafft* [77] and reconstructed relationships using maximum likelihood in *RAxML* using GTR model with Gamma distribution of rate variation.

Demographic analysis

We compared various demographic models and estimated parameters using the coalescent simulation software *fastsimcoal2 v.25* [78]. The models differed in topology and presence/absence of migration (admixture) events (Figs 4, S9, and S10), and each model was fit to a multi-dimensional site frequency spectrum calculated from the observed four-fold degenerate SNP data. Our primary interest in these analyses lie in confirming whether or not the additional populations that we sampled supported the single origin of tetraploids previously determined in [31]. Specifically, we focused on populations in the two diploid/tetraploid contact zones (Southern Carpathian and Baltic-Ruderal contact zones).

We attempted to discriminate between single versus independent origins using population quartets involving representatives from both putative parental diploid lineages (*S. Carp-2x* and *W. Carp.-2x* for *S. Carp.-4x*; *Baltic-2x* and *W. Carp.-2x* for *Ruderal-4x*; i.e. the genetically closest two in the descriptive distance-based and clustering analyses, Fig. 1 and 4), the *W. Carp.-4x* that is genetically closest to the putative ancestor of the widespread tetraploids [31] and the focal tetraploid (Fig. S9 and S10). In order to maintain a realistic number of scenarios while permuting the parameters (11 models for each population quartet), we modelled both uni- and bi-directional admixture within the same ploidy level, but only unidirectional interploidy admixture – from diploids to tetraploids. This decision reflects no signs of admixture of the diploids in clustering analyses (in contrast to the highly admixed tetraploids, Fig 1B) and virtual absence of triploids in nature [33], i.e. the only possible mediators of gene flow in the tetraploid-to-diploid direction [56]. In addition, we tested for the potentially admixed origin of the Baltic diploids (*Baltic-2x*) [34] using population trios involving representatives of each diploid lineage (*W. Carp.-2x* and *S. Carp.-2x*) as well as the focal *Baltic-2x* population (Fig S19 and Table S4).

For each scenario and population trio/quartet, we performed 50 independent *fastsimcoal* runs to overcome local maxima in the likelihood surface (see File S7 for example template file). In order to minimize the population-specific effects, we ran the analyses for different iterations of well-covered populations falling within the particular lineage, leading to 12 different population quartets ("natural replicates”) for each scenario testing the origin of the *S. Carp.-4x* and *Ruderal-4x* and four trios in the *Baltic-2x* scenarios. We then extracted the best likelihood partition for each *fastsimcoal* run, calculated Akaike information criterion (AIC) and summarized the AIC values across the 50 independent fastsimcoal runs over the scenarios tested within each population trio/quartet. The scenario with consistently lowest AIC values within particular population trio/quartet was preferred (Figs. S9 and S10). In order to calculate confidence intervals for the demographic parameters (Table S9), we sampled with replacement from the 4-dg SNPs to create 100 bootstrapped datasets and performed additional *fastsimcoal2* analyses under the preferred scenario with these 100 distinct datasets. For these analyses we also included representative (best covered) populations from the putatively non-admixed *C. Europe-4x* lineage. Finally, we used the mutation rate of 4.3×10^{-8} estimated by [31] to calibrate coalescent simulations and obtain absolute values of population sizes and divergence times.

In addition, we used PSMC 0.6.4 [79] to infer changes in effective population size (N_e) through time using information from whole-genome sequences of the *A. arenosa* diploids. We plotted 75 samples out of the 93 sequenced diploids, i.e. excluding samples with too low a coverage (below 12×) and too much missing data. Coverage and missing data might have large effects on the PSMC estimates [80]; therefore, our results should be interpreted only in conjunction with other analysis methods. We ran PSMC with parameters: psmc -N25 -t15 -r5 -p "4+25*2+4+6" and then plotted the past changes in N_e assuming a mutation rate of $3.7x10^{-8}$ substitutions per site per generation and generation time of two years.

Window-based metric calculation

In order to facilitate comparisons of windows across populations or population contrasts, we chose to calculate population genetic metrics in windows defined by a given number of base pairs. We repeated all calculations for two window sizes, 10kb and 50kb. We used the 50kb windows for characterizing broad, genome or chromosome-level patterns, whereas the former was used for finer, gene-level analyses. For 50kb windows, patterns of LD decay suggest a minimal degree of non-independence among windows relative to the genome background (Fig. 3a).

For each of the 36 populations with at least five individuals, we excluded all individuals with < 8× average coverage, except for populations SZI, KZL, and SNO as excluding individuals from these populations would drop them below required minimum of 5 individuals. After excluding these individuals, we excluded sites if the number of missing individuals was greater than 10%, on a population-specific basis. We calculated per-site nucleotide diversity (θ_π and Watterson's Theta θ_W; divided by the total number of sites with sufficient coverage) and Tajima's D following [81]. To equalise the expected variance of these metrics, thereby facilitating cross-population comparisons, we randomly chose 5 individuals with sufficient coverage from each population, doing so independently at each site. Differences in the diversity statistics among populations of different ploidy were tested using non-parametric Wilcoxon rank-sum test (wilcox.test in R package *stats*), taking populations as replicates.

We calculated the following divergence metrics for each possible pairwise population comparison using our custom scripts available at https://github.com/pmonnahan/ScanTools: F_{ST} [82], ρ [17], d_{XY} [83], and the number and proportion of fixed differences. The multi-locus implementation of F_{ST} and ρ was translated from the software SPAGeDi [84].

Topology weighting and detection of local introgression

We quantified the relative support for alternative phylogenetic relationships among populations using the topology weighting approach implemented in Twisst [85]. We used only 4-fold degenerate sites and used only individuals with > 8x coverage. Using bcftools, we converted the VCF files to a simplified tabular genotype file containing only the relevant individuals. We filtered this file using the filterGenotypes.py script that accompanies the Twisst software. At a site, we required genotype calls for at least 200 out of the 254 high coverage individuals (i.e. allowing ~20% missing data). We used only biallelic sites and required that the minor allele be present in at least 2 individuals. We then ran phyml_sliding_windows.py using 100 SNP windows (-w 100 and –M 20), which fits an ML phylogenetic tree for each window. Ideally, Twisst should be run on phased data; however, we were unable to find a workable phasing software that could handle diploids and tetraploids despite multiple attempts. Instead, we used the phasing algorithm internal to Twisst, which forms haplotypes by maximizing pairwise LD in each window.

We then ran Twisst for a number of scenarios, specifying individual population or groups of populations (lineages) as taxa. Twisst implements an iterative sub-sampling algorithm based on the phyML results to determine the support or weight of each possible taxon topology within each window. We requested the program calculate the complete weightings (completely searching sample space) if possible and used an approximate method, where sampling ceases after a given threshold of confidence is reached, when necessary. We allowed for 2000 sampling iterations before opting for the backup method. After this limit, we used the "Wilson" method at the 5% level, which will enforce sampling until the binomial 95% confidence interval is less than 5% of the weight value.

We used a combination of information from Twisst as well as divergence metrics to diagnose regions of both excessively strong and weak interploidy introgression in the two highly admixed *S. Carp.-4x* and *Ruderal-4x* lineages. First, introgressed regions should show an elevated weight for topologies wherein the proximal diploid/tetraploid pair are placed sister to one another (Topology 3 in Fig. 5). Second, when comparing the focal tetraploid to other tetraploid populations, an introgressed region should show elevated divergence while at the same time exhibiting reduced divergence to the focal diploid population. Conversely, introgression-resistant regions should show elevated Topology 1 and a combination of low divergence from tetraploids and elevated divergence from all diploids. We looked for evidence of selection on introgressed regions by overlapping window outliers for Topology 3 and Fay and Wu's H (in 10kb windows) in the focal tetraploid (99th percentile for both metrics).

Gene expression analysis of purifying selection

We evaluated patterns of diversity at the gene level using gene expression levels as a proxy for selective pressure based on evidence that higher-expressed genes generally show stronger signs of purifying selection in both plants and animals [36, 86-88]. To obtain gene-wise estimates of diversity, we performed a separate mapping process (again, using *A. lyrata* as the reference genome) using a subset of the total *A. arenosa* dataset that covers all major diploid and tetraploid lineages (9 tetraploid and 9 diploid populations, comprising 74 and 70 individuals, respectively, listed in Table S12). We retained sites with read depth of 4 or higher for at least 5 individuals across each population (9 – 14 million sites per population, Table S12). Sites were downsampled to 5 individuals independently at each site to homogenize chromosome depth across sites.

First, we extracted RNA from leaves of 3-week old individuals with three biological replicates for each of three diploid populations (HNI, RZA, SNO) to complete our previous dataset [54] of seven tetraploid populations (TBG, BGS, STE, KAS, CA2, HOC, SWA) using the RNeasy Plant Mini Kit (Qiagen). We synthesized single-strand cDNA from 500ng of total RNA using VN-anchored poly-T(23) primers with MuLV Reverse Transcriptase (Enzymatics) according to the manufacturer's recommendations. We made RNAseq libraries using the TruSeq RNA Sample Prep Kit v2 (Illumina) and sequenced libraries on an Illumina HiSeq 2000 with 50bp single-end reads. We sequenced between 9.8 and 18.8 million reads (avg 13.6 million). We aligned reads to the *A. lyrata* genome using TopHat2 [89] and re-aligned unmapped reads using Stampy [90]. We acquired read counts for each of the 32,670 genes using HTseq-count [91] with *A. lyrata* gene models. We normalized for sequencing depth using DEseq2 in R, [92] and further analyses were performed in MATLAB (MathWorks).

Analysis of differential expression between diploid and tetraploid expression patterns were performed using a one-way analysis of variance (ANOVA), and *p*-values were corrected for false discovery rate [93]. To avoid low-expression genes, we filtered for genes presenting a least one sample with normalized counts above 25, and computed the log-ratio of the average population expression in tetraploid populations against the average expression in diploids (positive when the expression of a gene is higher in tetraploid and negative when it is higher in diploids).

We obtained 6,504 genes with statistically significant differential expression ($p < 0.05$) between diploids and tetraploids (33% of 19,319 genes), but only 321 of these presented fold-change above 1.78x (5% two-tail threshold, Fig. S20A) and 214 above 2x. Overall, the average mean expression across populations is very strongly correlated between ploidies (slope = 1.02, $R^2 = 0.93$, Fig. S20B). To estimate mutational patterns we limited ourselves to the set of 18,998 genes non-differentially expressed (NDE) between ploidies.

We then filtered genes exhibiting a dependence of diversity metrics on the number of sites, specifically the genes that showed a correlation of number of sites with diversity (indicating potential mis-mapping of reads; Fig. S21). This effect of 4-dg θ_π and θ_W was strong for genes with fewer than 20 sites or more than a 100 using a locally weighted linear regression (LOWESS) for genes with a minimum of 5 sites of each fold (0-dg and 4-dg). Between these two boundaries, the number of sites only has a weak effect on 4-dg diversity. We observed a similar pattern in terms of 0-dg diversity with loci with fewer than 30 or more than 400 0-dg sites (Fig. S21C&D). After exclusion of loci outside of these bounds (for both 4-dg and 0-dg) from any downstream analysis we were able to cover around 45% of all NDE genes.

We then visualized the correlation of diversity of each gene with the average gene expression within the ploidy of the population with a locally weighted linear regression (LOWESS). For genes with expression levels above a certain expression threshold (50), nonsynonymous diversity (0-dg θ_π

and θ_W) showed a clear negative correlation with expression (proxy for strength of purifying selection) for both ploidies (Fig. 2a, Fig. S6: bold lines). Notably, this trend seems to break for very high expression (>2250 i.e. top 0.35%), possibly due to the low coverage of this expression range (67 genes). After removal of genes outside of these thresholds, we obtained 5,900 NDE genes per population to be used for multiple linear model (MLM) fitting.

We evaluated the effect of gene expression on 0-dg/4-dg diversity ratio and on 0-dg diversity for each population by modelling them (y) as a function of the ploidy of the population (p) with coefficient α_p, the average gene expression measured in ploidy p (E_p) with coefficient β, and an interaction term γ_p as follows:

$$(y) \sim 1+\alpha_p + \beta*\log(E_p) + \gamma_p*\log(E_p)$$

To estimate N_g, we first estimated effective population sizes using synonymous diversity as an estimator of θ, the estimated mutation rate (μ) of $4.3\text{x}10^{-8}$ for *A. arenosa* [31] and their theoretical relationship given by $\theta = 4\mu N_e$ in diploids and $\theta = 8\mu N_e$ in tetraploids. This gave an estimate of effective population sizes around 240,000 individuals for diploids and around 130,000 for tetraploids. In terms of number of haploid genomes, this difference in effective census sizes is more than compensated by tetrasomy (~480,000 in tetraploids vs ~520,000 in diploids).

The second MLM equation for evaluating the impact of population size (N_g) on 0-dg diversity or on 0-dg/4-dg diversity ratio was established using stepwise regression, evaluating the addition or removal of each term based on the *p*-value for an *F*-test of the change in the sum of squared error. The final formula for 0-dg diversity was:

$$(0\text{-dg } \theta_W) \sim 1+\alpha_p + \beta*\log(E_p) + \delta*N_g + \gamma_N*\log(E_p) + \varepsilon_p*N_g$$

where the interaction term with log expression γ is now dependent on N_g, δ represents the fixed effect of N_g, with an additional interaction term ε_p dependent on ploidy (p). The final formula for the 0-dg/4-

dg diversity ratio was:

$$(0\text{-dg } \theta_W) \sim 1+\alpha_p + \beta*\log(E_p) + \delta*N_g + \gamma_p*\log(E_p)$$

where the interaction term with log expression γ is now dependent on ploidy (p) only. The MLM estimates are presented in Table S5 and S6, and the estimated effects for values of the predictor chosen to show large responses are plotted in Fig 2b: log Expression: 3.9124 to 7.7098; Ng: 366058 (low) to 488976 (med) to 611894 (high).

In addition, we calculated recessive load as a number of sites with derived allele in homozygote state per each individual with at least 5 million SNPs called (240 individuals in total) and tested for difference among population means of diploid and tetraploid populations using Wilcoxon rank-sum test.

Distribution of fitness effects

Using the allele frequency spectra (AFS) for 4-dg and 0-dg sites (separately) for each of the 36 populations with ≥ 5 individuals screened, we estimated the distribution of fitness effects (DFE) [39], the proportion of adaptive substitutions relative to the total number of nonsynonymous substitutions (α) [40], and the proportion of adaptive substitutions relative to neutral divergence (ω_a; [94]; DFE-alpha v2.16; http://www.homepages.ed.ac.uk/pkeightl/software.html). This method implements a maximum-likelihood-based procedure to jointly estimate the parameters of a gamma-distributed DFE and a simple stepwise population size change model from site frequency spectrum data. Divergence was obtained from the polarized unfolded AFS and used to estimate α and ω_a, while correcting for the effect of slightly deleterious mutations using the estimated DFE. For all parameters estimated, we obtained 95% confidence intervals by analyses of 200 bootstrapped data sets. For each population, we fit two demographic models (constant population size and stepwise population size change), selected the best-fit model using a likelihood ratio test (LRT), and then estimated the parameters of the DFE, α and ω_a

under this model. The DFE is estimated using a gamma distribution with a shape parameter (β) and a scale parameter that represents the strength of purifying selection. As the strength of selection is dependent of the effective population size N_e, the result of DFE are often summarized by binning the distribution in 3 bins of $-N_e*s$. A $-N_es$ of 0-1 represents nearly neutral sites, 1-10 mildly deleterious mutations, and > 10 highly deleterious mutations.

For all populations, the stepwise population size change model was preferred. We ran DFE-alpha using both unfolded and folded site-frequency spectra. As the results were very consistent using the folded or unfolded allele frequency spectrum, we chose to focus on estimates based on the folded spectra, which should be more robust. We tested whether diploids and tetraploids differed with respect to the proportion of new nonsynonymous mutations in each bin, using Wilcoxon rank-sum tests.

Linked selection analysis and calculation of genotypic associations (linkage disequilibrium)

We inspected the relationship between the excess nonsynonymous divergence (d_{XY}) relative to synonymous divergence, as a proxy for divergent selection, and synonymous diversity (θ_π) in 50kb windows [95]. Both nonsynonymous and synonymous divergence was calculated for each population in each window as the average divergence at (non)synonymous sites for all pairwise contrasts between the focal population and all other populations in the dataset. We natural-log transformed these values and standardized them to be on the same scale. Then, we simply took the difference between nonsynonymous and synonymous scaled, transformed divergence values in each window. We refer to this difference as E_{NS}.

We also square root-transformed θ_π for normality purposes, removed windows with fewer than 20 SNPs, and removed populations with fewer than 2,000 non-missing windows, retaining a total of 27 populations (10 diploid and 17 tetraploid, listed in Supplementary File S1) and an average of 2,660 windows per population (~60% of genome). A negative relationship with θ_π is interpreted as evidence of a reductive effect of selection on linked, neutral diversity (i.e. linked selection). More specifically, we were interested to see if this relationship was dependent on ploidy level.

We used a multiple regression approach to infer this relationship and its dependence on ploidy level. We also included information on gene density (the proportion of bases in the window occupied by genic sequences according to the *A. lyrata* annotation) and proportion of missing data in each window. When calculating missingness in each window, we considered all biallelic sites and simply averaged the proportion of missing data across all 287 individuals in the study at each site within the window. Given the strong negative relationship between gene density and missingness, we combined gene density and missingness into a single, compound variable

$$\text{GDM} = \textit{gene density} * (1 - \textit{missingness})$$

where high values indicate windows with high gene density and low missingness and low values indicate the opposite. We fit a mixed model, using restricted maximum likelihood ('lmer' function) via the R package *lme4* [96], with E_{NS} and GDM as continuous variables, ploidy as fixed categorical variable, and populations as a random categorical variable. We determined significance using Wald Chi-square tests in the *car* package. We used the default distribution family (i.e. normal distribution) to model the residual variance. We also included a quadratic effect of E_{NS} to investigate the possibility of a nonlinear relationship with neutral diversity. Our initial model included all possible interactions, and we selected our final model by eliminating non-significant higher order interaction terms. The results were not qualitatively different following removal of tetraploid populations admixed by non-sister diploids (*S. Carp -4x*: DRA, LAC and TZI, and *Ruderal-4x:* KOW, STE and TBG). Similarly, results did not change upon removing the apparent outlier associated with the maximum observed 4-fold diversity.

To calculate genotypic correlations (a proxy for linkage disequilibrium), we recoded genotypes at all sites on chromosome 2 to represent the number of alternative alleles (0 - 2 for diploids and 0 - 4 for tetraploids). We calculated r^2 for pairs of loci and is simply the square of the correlation coefficient. An r value of 1.0 (and thus an r^2 of 1.0) means that genotypes are perfectly correlated for a particular pair of loci. However, because we do not have phase information, this r^2 value is not equivalent to the r^2 often reported when discussing LD. Therefore, we do not technically measure LD, but rather a related measure of genotypic associations. To visualize LD decay (Fig. 4a), we averaged r^2 value for all pairs of loci that fall in bins of a given distance apart, only considering populations with 8 or greater individuals. For populations with >8 individuals, we downsampled to include only the 8 highest coverage individuals. We performed the r^2 calculation for each population separately to avoid confounding effects of population differentiation.

To observe the impacts of various factors in our data on our LD approximation, we first simulated unlinked data and varied the number of sites and individuals as well as ploidy. At each site, we randomly drew allele frequencies from a uniform distribution, and then drew genotypes from the binomial distribution with p equal to the drawn allele frequencies and n of 2 or 4, depending on ploidy. The average r^2 value for each data set indicates that the number of individuals is the primary determinant of the expected r^2 value for unlinked sites, with the other factors exhibiting a negligible effect. We also simulated neutral linked data (100 replicate data sets; 1.5 Mb sequences; recombination rate = 1×10^{-8}, mutation rate = 1×10^{-8} ; $N_e = 100{,}000$) using msprime [97]. From each replicate, we created diploid and tetraploid genotype data by grouping the simulated haplotypes into sets of 2 or 4, respectively, to create 10 individuals of each. We then calculated r^2 for each replicate simulation. We simulated data for multiple parameter sets, focusing on the effects of the mutation rate, population size, recombination rate, and ploidy on r^2. We observed a slight downward bias for tetraploid data generally, but this effect was negligible compared to the effects of recombination rate and population size. As expected, the mutation rate did not affect r^2 as this measure should be proportional to the population-scaled recombination rate (a function of the per-base recombination rate and population size). If we double diversity by doubling the mutation rate, we find no observable effect on r^2 (Fig. S22).

Code Availability

Custom scripts used to generate genome scan metrics are available at https://github.com/pmonnahan/ScanTools. Other analysis scripts are available at https://github.com/pmonnahan/ArenosaPloidy.

Data Availability

Sequence data that support the findings of this study have been deposited in the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) with the primary accession code PRJNA484107 (available at http://www.ncbi.nlm.nih.gov/bioproject/484107) and PRJNA472485 for RNAseq data.

Acknowledgements

The authors thank Eliška Záveská, Magdalena Lučanová and Stanislav Španiel for help with fieldwork and John Brookfield and Simon Martin for helpful comments on versions of the manuscript. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme Projects of Large Research, Development, and Innovations Infrastructures, and by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2017/7-174. LY acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 679056) and the UK Biological and Biotechnology Research Council (BBSRC) via grant BB/P013511/1 to the John Innes Centre. KB acknowledges European Research Council Consolidator grant: CoG EVO-MEIO 681946 US National Science Foundation: IOS-1146465. Additional support was provided by Czech Science Foundation (project 16-10809S to KM, and 17-20357Y to FK), Charles University (project Primus/SCI/35 to FK), and a SNSF Early Postdoc Mobility fellowship (P2ZHP3_158773 to CS).

Author Contributions

LY, KB, FK, PB and PM conceived the study. PM, FK, PB, BL, CS, JK, RH, RS and PP performed analyses with input from LY, KB, RH, and TS. CS, PB, GF, MB and CW performed laboratory experiments. PM, FK and PB wrote the manuscript with primary input from KB, LY, BA, CS and TS. All authors edited and approved of the final manuscript.

Competing Interests statement

The authors declare no competing interests.

Materials & Correspondence

Correspondence and material requests should be addressed to Levi Yant at levi.yant@nottingham.ac.uk

References:

1. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences. 2009;106(33):13875-9. doi: 10.1073/pnas.0811575106.

2. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nature Reviews Genetics. 2017;18:411. doi: 10.1038/nrg.2017.26.

3. Salman-Minkov A, Sabath N, Mayrose I. Whole-genome duplication as a key factor in crop domestication. Nature plants. 2016;2:16115.

4. Storchova Z, Pellman D. From polyploidy to aneuploidy, genome instability and cancer. Nature reviews Molecular cell biology. 2004;5(1):45-54.

5. Yant L, Bomblies K. Genome management and mismanagement—cell-level opportunities and challenges of whole-genome duplication. Genes & development. 2015;29(23):2405-19.

6. Levin DA. The role of chromosomal change in plant evolution: Oxford University Press; 2002.

7. Parisod C, Holderegger R, Brochmann C. Evolutionary consequences of autopolyploidy. New phytologist. 2010;186(1):5-17.

8. te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubešová M, et al. The more the better? The role of polyploidy in facilitating plant invasions. Annals of Botany. 2011;109(1):19-45.

9. Segraves KA. The effects of genome duplications in a community context. New Phytologist. 2017.

10. Haldane JBS. The causes of evolution: Princeton University Press; 1932.

11. Wright S. The distribution of gene frequencies in populations of polyploids. Proceedings of the National Academy of Sciences. 1938;24(9):372-7.

12. Fisher R. The theoretical consequences of polyploid inheritance for the mid style form of Lythrum salicaria. Annals of Human Genetics. 1941;11(1):31-8.

13. Stebbins GL. Chromosomal evolution in higher plants. Chromosomal evolution in higher plants. 1971.

14. Haldane JB. Theoretical genetics of autopolyploids. Journal of Genetics. 1930;22(3):359-72.

15. Bever JD, Felber F. The theoretical population genetics of autopolyploidy. Oxford surveys in evolutionary biology. 1992;8:185-.

16. Otto SP, Whitton J. Polyploid Incidence and Evolution. Annual Review of Genetics. 2000;34(1):401-37. doi: 10.1146/annurev.genet.34.1.401. PubMed PMID: 11092833.

17. Ronfort J, Jenczewski E, Bataillon T, Rousset F. Analysis of population structure in autotetraploid species. Genetics. 1998;150(2):921-30.

18. Grant V. Plant speciation: New York: Columbia University Press xii, 563p.-illus., maps, chrom. nos.. En 2nd edition. Maps, Chromosome numbers. General (KR, 198300748); 1981.

19. Coyne JA, Orr HA. Speciation. Sunderland, MA. Sinauer Associates, Inc; 2004.

20. Mallet J. Hybrid speciation. Nature. 2007;446(7133):279.

21. Slotte T, Huang H, Lascoux M, Ceplitis A. Polyploid speciation did not confer instant reproductive isolation in Capsella (Brassicaceae). Molecular Biology and Evolution. 2008;25(7):1472-81.

22. Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, Nichols RA, et al. Unidirectional diploid–tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. Molecular ecology. 2016;25(11):2413-26.

23. Lafon-Placette C, Johannessen IM, Hornslien KS, Ali MF, Bjerkan KN, Bramsiepe J, et al. Endosperm-based hybridization barriers explain the pattern of gene flow between Arabidopsis lyrata and Arabidopsis arenosa in Central Europe. Proceedings of the National Academy of Sciences. 2017:201615123.

24. Ronfort J. The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. Genetics Research. 1999;74(1):31-42.

25. Hill R. Selection in autotetraploids. TAG Theoretical and Applied Genetics. 1971;41(4):181-6.

26. Selmecki AM, Maruvka YE, Richmond PA, Guillet M, Shoresh N, Sorenson AL, et al. Polyploidy can drive rapid adaptation in yeast. Nature. 2015;519(7543):349-52.

27. Schmickl R, Marburger S, Bray S, Yant L, Henderson I. Hybrids and horizontal transfer: introgression allows adaptive allele discovery. Journal of Experimental Botany. 2017.

28. Arnold ML, Kunte K. Adaptive Genetic Exchange: A Tangled History of Admixture and Evolutionary Innovation. Trends in ecology & evolution. 2017;32(8):601-11.

29. Bomblies K, Madlung A. Polyploidy in the Arabidopsis genus. Chromosome Research. 2014;22(2):117-34. doi: 10.1007/s10577-014-9416-x.

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

31. Arnold B, Kim S-T, Bomblies K. Single Geographic Origin of a Widespread Autotetraploid Arabidopsis arenosa Lineage Followed by Interploidy Admixture. Molecular Biology and Evolution. 2015;32(6):1382-95. doi: 10.1093/molbev/msv089.

32. Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K. Genetic adaptation associated with genome-doubling in autotetraploid Arabidopsis arenosa. PLoS genetics. 2012;8(12):e1003093.

33. Kolář F, Lučanová M, Záveská E, Fuxová G, Mandáková T, Španiel S, et al. Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the Arabidopsis arenosa group (Brassicaceae). Biological Journal of the Linnean Society. 2016;119(3):673-88.

34. Kolář F, Fuxová G, Záveská E, Nagano AJ, Hyklová L, Lučanová M, et al. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model Arabidopsis arenosa. Molecular ecology. 2016;25(16):3929-49.

35. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell2016. p. 481-91.

36. Ingvarsson PK. Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Molecular biology and evolution. 2007;24(3):836-44.

37. Wright SI, Yau CK, Looseley M, Meyers BC. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Molecular biology and evolution. 2004;21(9):1719-26.

38. Popescu CE, Borza T, Bielawski JP, Lee RW. Evolutionary rates and expression level in Chlamydomonas. Genetics. 2006;172(3):1567-76.

39. Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 2007;177(4):2251-61.

40. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009;26(9):2097-108.

41. Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. Overestimation of the adaptive substitution rate in fluctuating populations. Biology Letters. 2018;14(5).

42. Venkat A, Hahn MW, Thornton JW. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat Ecol Evol. 2018;2(8):1280-8. doi: papers3://publication/doi/10.1038/s41559-018-0584-5.

43. Schmickl R, Koch MA. Arabidopsis hybrid speciation processes. Proceedings of the National Academy of Sciences. 2011;108(34):14192-7.

44. Gerstein AC, Otto SP. Ploidy and the causes of genomic evolution. Journal of Heredity. 2009;100(5):571-81.

45. Favarger C. Cytogeography and biosystematics. Plant biosystematics. 1984:453-76.

46. Brochmann C, Brysting A, Alsos I, Borgen L, Grundt H, Scheen A-C, et al. Polyploidy in arctic plants. Biological journal of the Linnean society. 2004;82(4):521-36.

47. Butruille DV, Boiteux LS. Selection–mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. Proceedings of the National Academy of Sciences. 2000;97(12):6608-13. doi: 10.1073/pnas.100101097.

48. Willis JH. Inbreeding Load, Average Dominance and the Mutation Rate for Mildly Deleterious Alleles in &lt;em&gt;Mimulus guttatus&lt;/em&gt. Genetics. 1999;153(4):1885.

49. Schmickl R, Marburger S, Bray S, Yant L. Hybrids and horizontal transfer: introgression allows adaptive allele discovery. Journal of experimental botany. 2017;68(20):5453-70.

50. Lowe WH, Muhlfeld CC, Allendorf FW. Spatial sorting promotes the spread of maladaptive hybridization. Trends in Ecology & Evolution. 2015;30(8):456-62. doi: https://doi.org/10.1016/j.tree.2015.05.008.

51. Yukilevich R. ASYMMETRICAL PATTERNS OF SPECIATION UNIQUELY SUPPORT

REINFORCEMENT IN DROSOPHILA. Evolution. 2012;66(5):1430-46. doi: 10.1111/j.1558-5646.2011.01534.x.

52. Baduel P, Arnold B, Weisman CM, Hunter B, Bomblies K. Habitat-associated life history and stress-tolerance variation in Arabidopsis arenosa. Plant physiology. 2016;171(1):437-51.

53. Hylander N. Cardaminopsis suecica (Fr.) Hiit., a northern amphidiploid species. Bulletin du Jardin botanique de l'Etat, Bruxelles/Bulletin van den Rijksplantentuin, Brussel. 1957:591-604.

54. Baduel P, Hunter B, Yeola S, Bomblies K. Genetic basis and evolution of rapid cycling in railway populations of tetraploid Arabidopsis arenosa. PLOS Genetics. 2018;14(7):e1007510. doi: 10.1371/journal.pgen.1007510.

55. Husband BC, Sabara HA. Reproductive isolation between autotetraploids and their diploid progenitors in fireweed, Chamerion angustifolium (Onagraceae). New Phytologist. 2004;161(3):703-13.

56. Kolář F, Čertner M, Suda J, Schönswetter P, Husband BC. Mixed-Ploidy Species: Progress and Opportunities in Polyploid Research. Trends in Plant Science. 2017.

57. Soltis DE, Soltis PS. Polyploidy: recurrent formation and genome evolution. Trends in Ecology & Evolution. 1999;14(9):348-52.

58. Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD, Franklin FCH, et al. Meiotic adaptation to genome duplication in Arabidopsis arenosa. Current biology. 2013;23(21):2151-6.

59. Arnold BJ, Lahner B, DaCosta JM, Weisman CM, Hollister JD, Salt DE, et al. Borrowed alleles and convergence in serpentine adaptation. Proceedings of the National Academy of Sciences. 2016;113(29):8320-5.

60. Doyle JJ. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11-5.

61. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):pp. 10-2.

62. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature genetics. 2011;43(5):476-81.

63. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

64. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011;43(5):491-8.

65. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.

66. Wright SI, Lauga B, Charlesworth D. Rates and patterns of molecular evolution in inbred and outbred Arabidopsis. Molecular Biology and Evolution. 2002;19(9):1407-20.

67. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403-5.

68. Nei M. Genetic distance between populations. The American Naturalist. 1972;106(949):283-92.

69. Pembleton LW, Cogan NO, Forster JW. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Molecular ecology resources. 2013;13(5):946-52.

70. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics (Oxford, England). 1998;14(1):68-73.

71. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573-89.

72. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus

genotype data. Genetics. 2000;155(2):945-59.

73. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. The pattern of polymorphism in Arabidopsis thaliana. PLoS biology. 2005;3(7):e196.

74. Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, et al. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nature genetics. 2016;48(9):1077-82.

75. Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. 2010;26(3):419-20.

76. Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. Journal of statistical software. 2007;22(4):1-20.

77. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.

78. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS genetics. 2013;9(10):e1003905.

79. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475(7357):493.

80. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Molecular ecology. 2016;25(5):1058-72.

81. Zeng K, Fu Y-X, Shi S, Wu C-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 1996;174(3):1431-9.

82. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. evolution. 1984;38(6):1358-70.

83. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular ecology. 2014;23(13):3133-57.

84. Hardy OJ, Vekemans X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Resources. 2002;2(4):618-20.

85. Martin SH, Van Belleghem SM. Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. Genetics. 2017. doi: 10.1534/genetics.116.194720.

86. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Molecular biology and evolution. 2000;17(1):68-070.

87. Rocha EP, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. Molecular biology and evolution. 2004;21(1):108-16.

88. Slotte T, Bataillon T, Hansen TT, St. Onge K, Wright SI, Schierup MH. Genomic Determinants of Protein Evolution and Polymorphism in Arabidopsis. Genome Biology and Evolution. 2011;3:1210-9. doi: 10.1093/gbe/evr094.

89. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36.

90. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome research. 2011;21(6):936-9.

91. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.

92. Love M, Anders S, Huber W. Differential analysis of count data–the DESeq2 package. Genome Biology. 2014;15:550.

93. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological).

1995:289-300.

94. Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, et al. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Molecular biology and evolution. 2010;27(8):1822-32.

95. Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, et al. Natural selection and genetic diversity in the butterfly Heliconius melpomene. Genetics. 2016;203(1):525-41.

96. Bates D, Martin M, Ben B, Walker S. lme4: Linear mixed effects models using Eigen and S4.(R package v. 1.0–6). See http://CRAN. R-project. org/package= lme4; 2014.

97. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology. 2016;12(5):e1004842. doi: 10.1371/journal.pcbi.1004842.

Figure Legends

Fig. 1 | Geographic distribution and range-wide genetic variation of *Arabidopsis arenosa*. a, Distribution of the 39 *A. arenosa* populations (red labels - diploids, blue - tetraploids) with average proportions of cluster membership inferred by FastStructure (panel b). Color shades highlight highly admixed tetraploid populations (*Ruderal* and *S. Carpathian-4x*) together with the diploid sources of admixture. **b**, Posterior probabilities of cluster membership of the 287 *A. arenosa* individuals as inferred by FastStructure under K=6. **c**, Neighbor-joining tree of Nei's genetic distances between all individuals and the outgroup *Arabidopsis croatica*. Individuals from admixed populations are highlighted correspondingly. Inset: distribution of pairwise genetic divergence of populations (ρ) within each ploidy. **d**, Principal component analysis of all but the two most divergent diploid (*Pannonian* and *Dinaric*) *A. arenosa* lineages (shades correspond to admixed populations).

Fig. 2 | Effects of ploidy on purifying selection, genetic load, and the distribution of fitness effects (DFE). a, Genic nonsynonymous (0-dg) diversity versus average gene expression (log-scale) for each population and each ploidy (resp. faint and bold LOWESS curves). The two outlier populations, *2x*-SNO and *4x*-SCH, are indicated. **b**, Standardized effects with confidence intervals in multiple linear model of haploid effective population size (N_g), ploidy, and levels of expression on nonsynonymous (0-

dg) θ_W (upper panel) and on 0-dg/4-dg θ_W ratio (lower panel). The interaction terms of N_g with ploidy and with expression are represented for 0-dg θ_W. **c**, Recessive load in tetraploid individuals estimated as number of homozygous 0-dg derived alleles. **d**, DFE by ploidy and binned by strength of purifying selection. **e**, Proportion of adaptive substitution (α) and proportion of adaptive substitution relative to neutral (ω_α) by ploidy. Errors bars represent 95% confidence interval based on 200 bootstrap replicates.

Fig. 3 | Ploidy effects on linkage disequilibrium and the strength of linked selection. a, Decay of genotypic correlations (proxy for LD) within each population and averaged for each ploidy (heavy lines) as a function of distance between sites. **b**, Curvilinear relationship between excess nonsynonymous (0-dg) divergence (E_{NS}) on neutral diversity (4-dg θ_π,). **c-d**, Linear relationship between excess 0-dg divergence on neutral diversity (4-dg θ_π) for gene-poor (<20th gene-density percentile) and gene-dense regions (>90th gene-density percentile), respectively.

Fig. 4 | Evidence for single origin of tetraploids. a-b, Single origin of the *S. Carp.-4x*, and *Ruderal-4x* tetraploids, respectively, followed by local admixture from their geographically proximal diploids inferred as most likely scenario (large) by fastsimcoal2 coalescent simulations vs. competing scenarios (small schemes). Range of median maximum-likelihood estimates of divergence times in generations across different population quartets are indicated. **c**, Allele frequencies in each tetraploid lineage of alleles diagnostic to particular diploid *A. arenosa* lineages. Significant differences within each category of diploid alleles, as identified by Tukey's honestly significant difference (HSD) post hoc test, are designated by distinct letters. **d**, Topology weights (TM, tetraploid monophyly; LA, local-admixture; ILS, incomplete lineage sorting) in set of 6 meiosis-related genes compared with genome-wide average (WG).

Fig. 5 | Signals of interploidy introgression and loci resisting the gene flow. Topology weightings for the three diagnostic topologies relating *S. Carp.-2x*, *S. Carp.-4x*, *W. Carp-4x* and the outgroup, *Dinaric-2x* across arms of scaffolds 8 (left) and 4 (right). Zoomed-in panels from top to bottom: topology weighting, average divergence (ρ) of *S. Carp.-4x* to all other tetraploids (black line) and to diploid lineages (colored lines), and Fay and Wu's H. Left zoom-in: example of locus locally introgressed from diploids and under positive selection (dominant LA topology, low divergence with local diploids specifically, and deeply negative Fay and Wu's H). Right zoom-in: resistance to local introgression of key meiotic locus, *ASY3*, with narrower peaks consistently with more ancient origin of the region.

Table 1 Measures of within-population diversity and among-population divergence in diploid and tetraploid *A. arenosa*

	Divergence			Diversity[1]										
	Rho / F_{st}[2]	AMOVA[3]	rM[4]	pairwise diversity (θ_π)			Watterson's θ (θ_W)			Tajima's D			π_{NS}/π_S	θ_{NS}/θ_S
Sites	4-dg	4-dg	4-dg	all	4-dg	NS (0-dg)	all	4-dg	NS (0-dg)	all	4-dg	NS (0-dg)	-	-
Diploids	0.30 /	71	0.14	0.016	0.022	0.0054	0.015	0.022	0.005	0.03	0.16	-0.09	0.242	0.255
(14 pops)	0.29		n.s.	(0.003)	(0.003)	(0.0007)	(0.004)	(0.003)	(0.0009)	(0.21)	(0.18)	(0.23)	(0.017)	(0.017)
Tetraploids	0.20 /	48	0.55	0.015	0.023	0.0055	0.016	0.023	0.006	-0.23	0.00	-0.41	0.237	0.263
(22 pops)	0.11		***	(0.004)	(0.006)	(0.0013)	(0.004)	(0.005)	(0.0013)	(0.29)	(0.27)	(0.28)	(0.007)	(0.007)
Difference[5]	-	-	-	n.s.	n.s.	n.s.	n.s.	n.s.	.	**	.	***	n.s.	***

Populations with < 5 individuals were excluded; for populations with > 5 individuals, sites were randomly downsampled to five to facilitate comparison across populations.

[1] values averaged across populations within the ploidy, standard deviation is in parentheses

[2] values averaged over pairwise comparisons of populations belonging to that ploidy

[3] % of explained variance among populations (compared to variance within populations) in Analysis of Molecular Variance (AMOVA)

[4] Isolation by distance tested by Mantel test; the rM for diploid populations became 0.23* when spatially distant but genetically proximal Baltic populations were excluded

[5] Wilcoxon rank-sum test; n.s. non significant, $p \leq 0.07$ * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

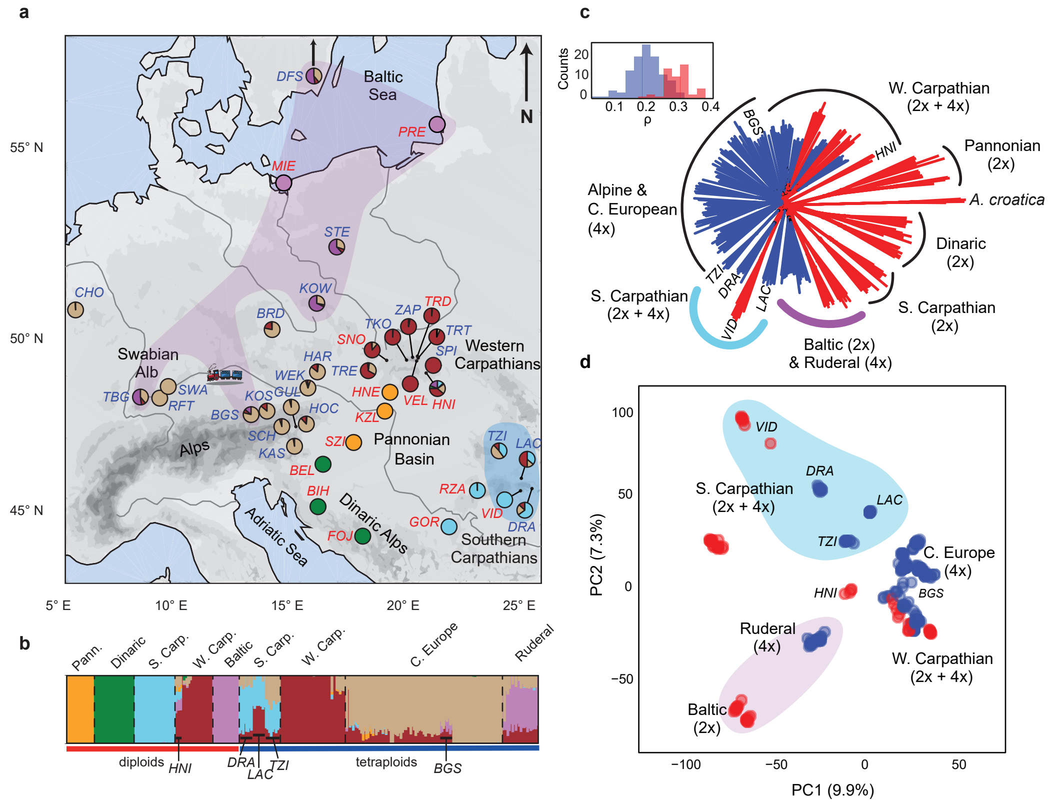

Fig. 1 | Geographic distribution and range-wide genetic variation of *Arabidopsis arenosa*. a, Distribution of the 39 *A. arenosa* populations (red labels - diploids, blue - tetraploids) with average proportions of cluster membership inferred by FastStructure (panel b). Color shades highlight highly admixed tetraploid populations (*Ruderal* and *S. Carpathian-4x*) together with the diploid sources of admixture. **b**, Posterior probabilities of cluster membership of the 287 *A. arenosa* individuals as inferred by FastStructure under K=6. **c**, Neighbor-joining tree of Nei's genetic distances between all individuals and the outgroup *Arabidopsis croatica*. Individuals from admixed populations are highlighted correspondingly. Inset: distribution of pairwise genetic divergence of populations (ρ) within each ploidy. **d**, Principal component analysis of all but the two most divergent diploid (*Pannonian* and *Dinaric*) *A. arenosa* lineages (shades correspond to admixed populations).

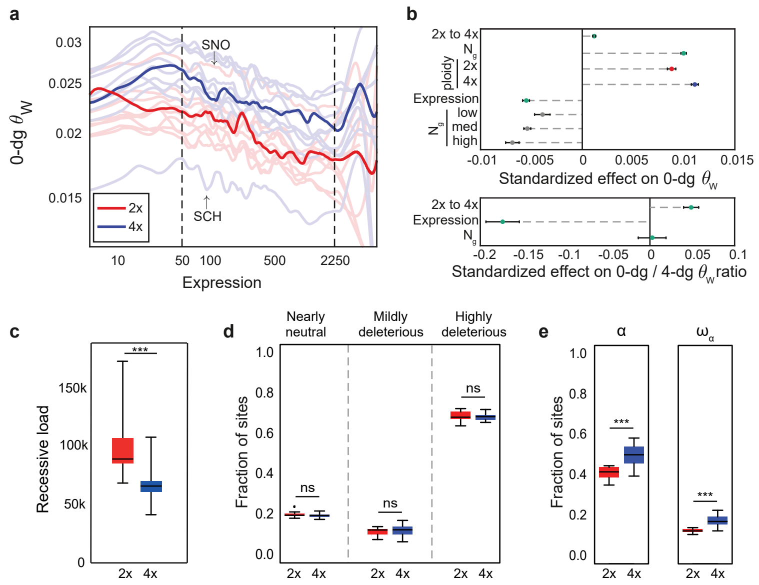

Fig. 2. | Effects of ploidy on purifying selection, genetic load, and the distribution of fitness effects (DFE). a, Genic nonsynonymous (0-dg) diversity versus average gene expression (log-scale) for each population and each ploidy (resp. faint and bold LOWESS curves). The two outlier populations, 2x-SNO and 4x-SCH, are indicated. **b**, Standardized effects with confidence intervals in multiple linear model of haploid effective population size (N_g), ploidy, and levels of expression on nonsynonymous (0-dg) θ_W (upper panel) and on 0-dg/4-dg θ_W ratio (lower panel). The interaction terms of N_g with ploidy and with expression are represented for 0-dg θ_W. **c**, Recessive load in tetraploid individuals estimated as number of homozygous 0-dg derived alleles. **d**, DFE by ploidy and binned by strength of purifying selection. **e**, Proportion of adaptive substitution (α) and proportion of adaptive substitution relative to neutral (ω_α) by ploidy. Errors bars represent 95% confidence interval based on 200 bootstrap replicates.

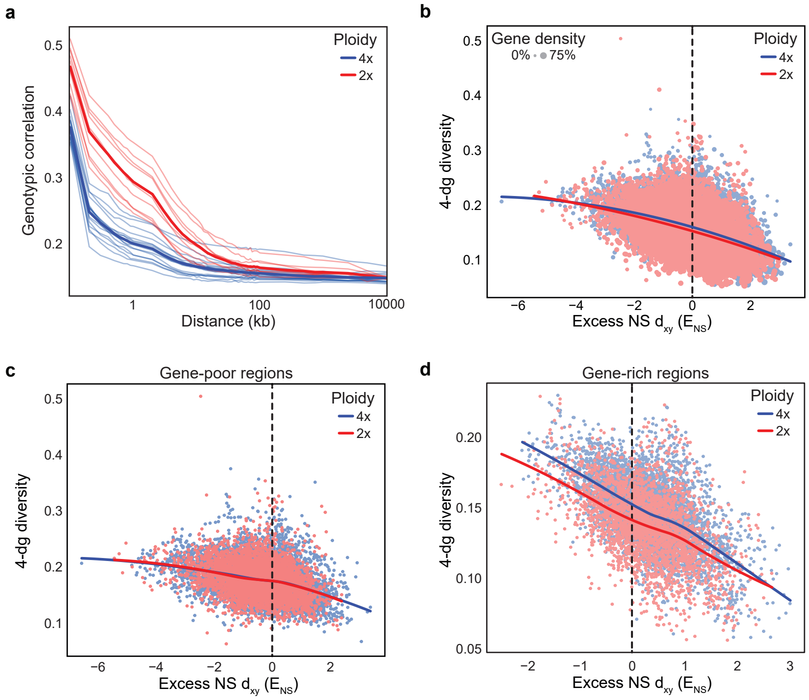

Fig. 3 | Ploidy effects on linkage disequilibrium and the strength of linked selection. a, Decay of genotypic correlations (LD estimator) within each population and averaged for each ploidy (heavy lines) as a function of distance between sites. **b,** Curvilinear relationship between excess nonsynonymous (0-dg) divergence (E_{NS}) on neutral diversity (4-dg θ_π). **c-d,** Linear relationship between excess 0-dg divergence on neutral diversity (4-dg θ_π) for gene-poor (<20th gene-density percentile) and gene-dense regions (>90th gene-density percentile), respectively.

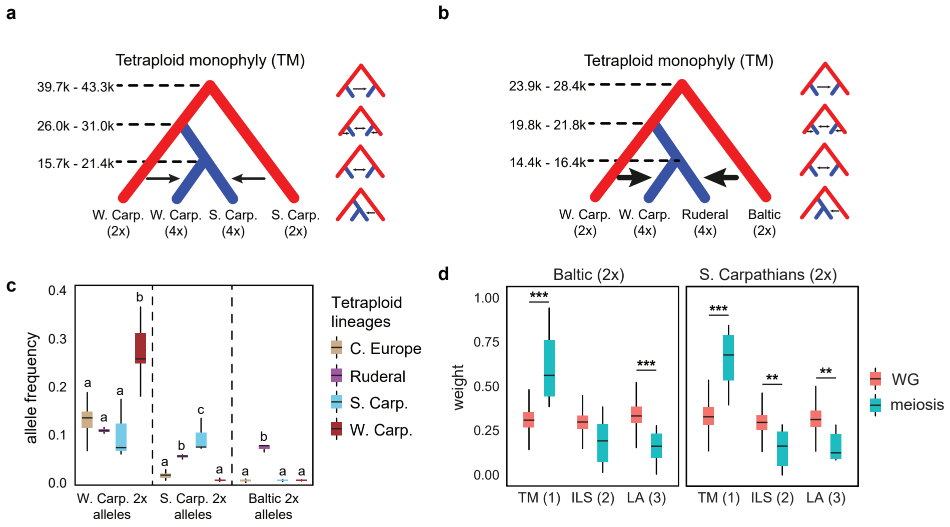

Fig. 4 | Evidence for single origin of tetraploids. a-b , Single origin of the *S. Carp.-4x*, and *Ruderal-4x* tetraploids, respectively, followed by local admixture from their geographically proximal diploids inferred as most likely scenario (large) by fastsimcoal2 coalescent simulations vs. competing scenarios (small schemes). Range of median maximum-likelihood estimates of divergence times in generations across different population quartets are indicated. **c**, Allele frequencies in each tetraploid lineage of alleles diagnostic to particular diploid *A. arenosa* lineages. Significant differences within each category of diploid alleles, as identified by Tukey's honestly significant difference (HSD) post hoc test, are designated by distinct letters.**d**, Topology weights (TM, tetraploid monophyly; LA, local-admixture; ILS, incomplete lineage sorting) in set of 6 meiosis-related genes compared with genome-wide average (WG).

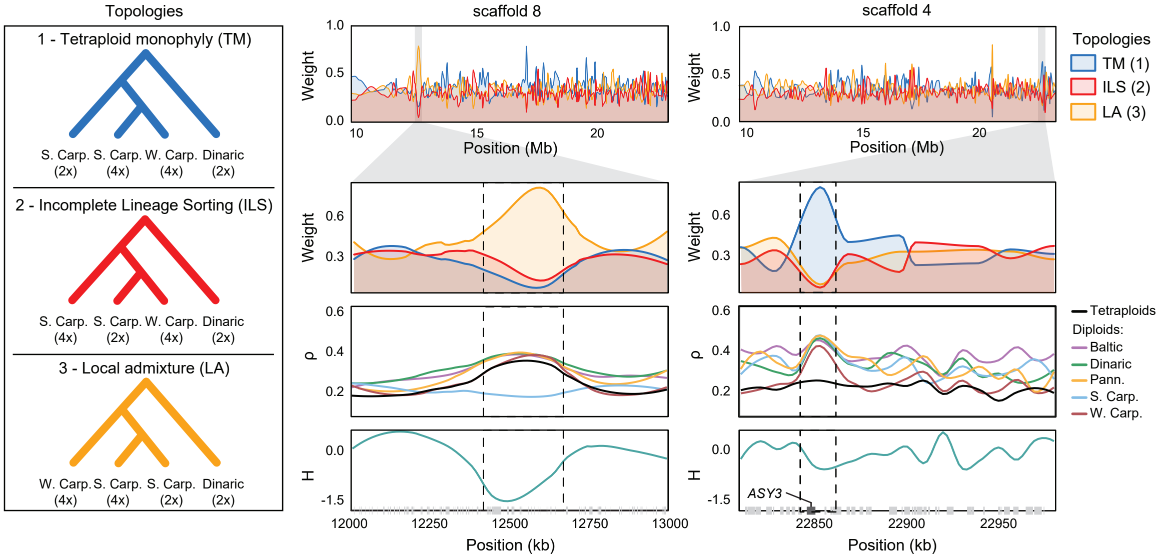

Fig. 5 | Signals of interploidy introgression and barrier loci. Topology weightings for the three diagnostic topologies relating *S. Carp.-2x, S. Carp.-4x, W. Carp-4x* and the outgroup, *Dinaric-2x* across arms of scaffolds 8 (left) and 4 (right). Zoomed-in panels from top to bottom: topology weighting, average divergence (ρ) of *S. Carp.-4x* to all other tetraploids (black line) and to diploid lineages (colored lines), and Fay and Wu's H. Left zoom-in: example of locus locally introgressed from diploids and under positive selection (dominant LA topology, low divergence with local diploids specifically, and deeply negative Fay and Wu's H). Right zoom-in: resistance to local introgression of key meiotic locus, *ASY3*, with narrower peaks consistently with more ancient origin of the region