

How predictable is genome evolution? Insights from convergent adaptations across Brassicaceae

Je evoluce genomu předvídatelná?
O čem vypovídají opakované adaptace v čeledi brukvovité



PhD thesis
Prague, 2021

Magdalena Bohutínská

Department of Botany
Charles University, Faculty of Science
Supervised by Filip Kolář, Ph.D.



Hereby I declare that I made this thesis independently, using the mentioned references. I have not submitted or presented any part of this thesis for any other degree or diploma.

Prohlašuji, že jsem práci vypracovala samostatně a uvedla jsem všechny použité informační zdroje a literaturu. Žádná část práce nebyla využita k získání jiného vysokoškolského titulu.

Magdalena Bohutínská, Prague, 2021

Author contribution statement

I declare that I have contributed to all papers included in the thesis. My contributions to particular papers are as follows:

CS1: Bohutínská, M., Vlček, J., Yair, S., Leanen, B., Konečná, V., Fracassetti, M., Slotte, T., & Kolář, F. (2021). Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives. Proceedings of the National Academy of Sciences of the United States of America. doi.org/10.1073/pnas.2022713118

Field sampling, data analysis and interpretation, manuscript writing; total contribution 80%

CS2: Wos, G., Bohutínská, M., Nosková, J., Mandáková, T., & Kolář, F. (2021). Parallelism in gene expression between foothill and alpine ecotypes in *Arabidopsis arenosa*. The Plant Journal, tpj.15105. https://doi.org/10.1111/tpj.15105

Population genomic data analysis and interpretation, manuscript editing; total contribution 30%

CS3: Marburger, S., Monnahan, P., Seear, P. J., Martin, S. H., Koch, J., Paajanen, P., Bohutínská, M., Higgins, J., Schmickl, R., & Yant, L. (2019). Interspecific introgression mediates adaptation to whole genome duplication. Nature Communications. 10(1). doi.org/10.1038/s41467-019-13159-5

Analysis of adaptive introgression at a locus level, manuscript editing; total contribution 10%

CS4: Bohutínská, M., Handrick, V., Yant, L., Schmickl, R., Kolář, F., Bomblied, K., & Paajanen, P. (2021). De-novo mutation and rapid protein (co-)evolution during meiotic adaptation in *Arabidopsis arenosa*. Molecular Biology and Evolution. doi.org/10.1093/molbev/msab001

Study design, population genomic data analysis and interpretation, manuscript writing; total contribution 60%

CS5: Bohutínská, M., Alston, M., Monnahan, P., Mandáková, T., Bray, S., Paajanen, P., Kolář, F., & Yant, L. (2021). Novelty and convergence in adaptation to whole genome duplication. Molecular Biology and Evolution. doi.org/10.1093/molbev/msab096

Field sampling and plant cultivation, population genomic data analysis and interpretation, manuscript writing; total contribution 70%

CS6: Bray, S. M., Wolf, E. M., Zhou, M., Busoms, S., Bohutínská, M., Paajanen, P., Monnahan, P., Koch, J., Fisher, S., Koch, M., & Yant, L. (2020). Convergence and novelty in adaptation to whole genome duplication in three independent polyploids. Preprint at BioRxiv. doi.org/10.1101/2020.03.31.017939 (manuscript)

Study and analysis design on quantifying parallelism, manuscript editing; total contribution 10%

Table of Contents

Abstract.....	1
Abstrakt.....	2
Acknowledgements.....	3
Part A – General chapters.....	4
A1: Introduction.....	4
(Convergent) adaptation, the deterministic aspect of evolution.....	4
Divergence matters? The role of different sources of adaptive variation.....	5
Beyond gene reuse: the effect of function-level convergence.....	6
A2: Aims and model systems.....	8
Central hypotheses:.....	8
Objective 1 – the presence of genomic convergence:.....	9
Objective 2 – varying extent of genomic convergence at shallow divergences:.....	9
Objective 3 – varying extent of genomic convergence at deeper divergences:.....	9
A3: Methods.....	11
A4: Key results of my studies.....	12
1. Shallow divergence levels (within <i>Arabidopsis</i>).....	12
2. Deeper divergence levels (beyond <i>Arabidopsis</i>).....	13
A5: Conclusions: Divergence matters!.....	17
References.....	20
Part B – Case studies.....	27
Case study 1.....	28
Case study 2.....	39
Case study 3.....	54
Case study 4.....	66
Case study 5.....	82
Case study 6.....	120

Abstract

Adaptation, the process of propagation of beneficial mutations, enables populations and species to face changing environmental conditions. Cases of convergent (considered synonym to 'parallel' here) adaptation highlight natural selection's capacity to shape biological diversity, and provide natural replicates to investigate the extent of predictability in the genetic basis of adaptation. Recently, a wealth of genomic studies has identified widespread genomic convergence. However, the evidence has taken many forms, from responses in the same functions but different loci (function-level convergence) down to the precision of repeated adaptation via the same genes (gene reuse), raising a question if such variation can be explained by some unifying force/mechanism. It has been speculated that patterns of genomic convergence differ among studies because the scale of divergence differs from case to case. Yet, this observation has not been tested on a unified model system across a divergence continuum and so underlying factors remain unknown.

In my PhD project I conducted an empirical investigation on how and why patterns of genomic convergence change with increasing divergence. To do so, I studied the genomic basis of convergent adaptation to outer (alpine habitats) and inner (whole genome duplication) environmental challenges. I focused on convergently adapting lineages across plant model family Brassicaceae, spanning ~0.01 – 25 million years divergence. Leveraging such naturally replicated system, I aimed to test if the level of gene reuse in convergent adaptation decreases with increasing divergence between the compared units, if this reflects the availability of pre-existing variation and genetic constraints such as pleiotropy and what is the role of function-level convergence.

Using whole genome resequencing and statistical analysis, complemented with experiments, I identified convergent footprints of selection shared across natural populations and quantified the extent of genomic convergence. Among the case studies forming my PhD project, the degree of gene reuse in convergent adaptation strongly depended on genetic divergence between the compared lineages – while I found substantial gene reuse between closely related populations, shared genetic underpinnings of adaptation were rare above the genus level. At such deeper divergences, the lack of gene reuse was compensated by significant function-level convergence. Finally, at shallow divergence levels, decreasing gene reuse reflected decreasing probability of allele reuse, i.e. repeated recruitment of the same standing or introgressed variation by positive selection. This provided a first mechanistic explanation for the observed divergence scale-dependency of genetic convergence.

In summary, I showed that the gene reuse in convergent adaptation scales with divergence, reflecting different population-level processes determining the availability of adaptive alleles at a within-species level. Further, adaptation via different loci involved in the same pathway become the dominant source of repeatability once the divergence is high and allele sharing is limiting. Generally, the results of my PhD thesis bring a novel empirical contribution to the ongoing lively discussion about the drivers of convergent adaptation and the (un)predictability of evolution. Consequently, they may inform a variety of conservation and medicinal applications that rely on evolutionary predictability and may be of interest to geneticists leveraging natural replicates of convergence in studies of adaptation.

Abstrakt

Adaptace, proces šíření výhodných alel, umožňuje populacím a druhům čelit nepříznivým podmínkám prostředí. Příklady konvergentní adaptace (zde použita jako synonymum k paralelní) ukazují důležitost přírodního výběru pro formování biologické diversity a umožňují zjišťovat, do jaké míry je genetická podstata evoluce předvídatelná. Mnoho nedávných genomických studií odhalilo, že je konvergentní evoluce na úrovni genomu rozšířená. Jenže konkrétní podoba se velmi liší, od odpovědi pomocí stejných molekulárních drah ale různých míst v genomu (funkční konvergence) po opakovanou adaptaci pomocí stejných genů (opětovné použití genů). To vede k otázce, jestli takto variabilní systém umožňuje jakékoliv predikce ohledně mechanismů zodpovědných za konvergentní adaptaci. Spekuluje se, že se způsoby konvergentní adaptace liší podle příbuznosti zkoumaných linií. Jenže tato možná závislost nikdy nebyla testována na jednotném modelovém systému napříč škálou různých příbuzností, a tak se nemohla potvrdit ani ona, ani její možné příčiny.

V předkládané PhD práci jsem empiricky zjišťovala jak a proč se genomická konvergence liší s klesající příbuzností. Jako příklad jsem použila konvergentní adaptaci k vnějším (alpínské prostředí) a vnitřním (celogenomová duplikace) podmínkám. Ty jsem studovala v rostlinné modelové čeledi brukvovité (Brassicaceae) a zahrnovaly spektrum příbuzností mezi 0,01 – 25 miliony let. Díky těmto modelovým systémům jsem testovala, jestli se míra opětovného využití genů v konvergentní adaptaci snižuje se snižující se příbuzností a jestli to odráží dostupnost sdílených alel nebo genetická omezení daná pleiotropií genů. Nakonec jsem se ptala, jakou roli v opakované adaptaci hraje funkční konvergence.

Za použití sekvenování genomu a statistických analýz, doplněných experimenty, jsem identifikovala konvergentní změny související s adaptací a vyčíslila tak míru genomické konvergence. Mezi studiemi, které byly součástí mého PhD projektu, míra opětovného využití genů v konvergentní adaptaci úzce souvisela s genetickou příbuzností srovnávaných linií. Identifikovala jsem vysokou míru opětovného využití genů mezi blízce příbuznými populacemi a naopak velmi nízkou míru nad úrovní rodu. U takto vzdáleně příbuzných konvergentních linií potom byla nízká míra genové konvergence kompenzována zvýšenou mírou konvergence na úrovni funkčních molekulárních drah. Nakonec jsem zjistila, že u blízce příbuzných linií je opětovné využití genů pravděpodobnější díky jejich schopnosti sdílet společné alely – buď výměnou genovým tokem nebo zděděné od předků. Tato zjištění představují první objasnění mechanismů stojících za snižující se mírou opětovného využití genů s narůstající evoluční vzdáleností mezi liniemi.

Celkově jsem ukázala, že míra opětovného použití genů v konvergentní adaptaci souvisí s příbuzností, což je zapříčiněno lepší dostupností společných alel mezi populacemi uvnitř druhu. Opakovaná adaptace pomocí rozdílných genů účinkujících ve stejné molekulární dráze naopak dominuje, když je příbuznost mezi konvergentními liniemi (a tedy možnost sdílet výhodné alely) limitně nízká. Obecně tyto výsledky nabízejí nové empirické poznatky k debatě o opakovatelnosti adaptace a z ní vyplývající (ne)předvídatelnosti evoluce. To může v důsledku přispět k navrhování programů ochrany přírody nebo nalézt využití v oborech medicíny pracující s předvídatelností evoluce.

Acknowledgements

I am very grateful to my supervisor Filip Kolář for giving me the opportunity to work on this project, for his great advices on the project design, data analysis and on interpretation, for helping me develop fruitful collaborations, for teaching me patiently how to write a manuscript and for providing me so much space for my scientific development. I am further grateful to other senior scientists which I could collaborate with during this project, mainly to Levi Yant, Tanja Slotte, Pirita Paajanen and Kirsten Bomblies. I learnt a lot by working with them! Finally, I thank all my colleagues and friends who are co-authors on articles included in the thesis – without them, these would never happen.

Finishing a PhD project is challenging by itself and having two babies during the process requires an incredible amount of support and understanding from supervisor, colleagues and family. That is why I am very grateful to Filip, who was not only an excellent supervisor, but also a great friend, always inventing ways to keep me involved. Great thanks for allowing me to join field trips, excursions and retreats with babies, for letting his office be destroyed by them and for supporting my (and sometimes babies') active participation in the team. Also, big thanks to the whole Plant ecological genomics team for their friendliness and patience! I wish that the incredibly supportive attitude which I could experience during my PhD once becomes a standard for every parent in science. My great thanks belongs to my beloved husband Daniel, for his constant help, love and interest in what I am working on. Thanks for joining our field trips, retreats and abroad stays and taking care of our kids anytime possible. My thanks and love goes also to our sons Daneček and Dominik. Finally, I am very grateful to my parents, for sharing their love for nature and science with me and to the long list of babysitters (including many family members and friends) for allowing me to experience the precious peaceful and focused moments without my beloved kids.

Part A – General chapters

A1: Introduction

(Convergent) adaptation, the deterministic aspect of evolution

Adaptive evolution, the propagation of fitter alleles through the action of positive selection gives rise to innovation in nature, the process which is particularly important under a changing environment (Lande & Shannon, 1996). Indeed, an adaptively evolving fraction of the genome is detected in almost all eukaryotic organisms – simulations and empirical evidence showed that as much as 40 % of genetic variation is likely to be targeted by positive selection (Booker, Jackson, & Keightley, 2017; Messer & Petrov, 2013). This provides evidence for the deterministic aspect of evolution and contributes to the neutralist-selectionist debate about what fraction of genetic variation evolves under deterministic positive selection and what under stochastic genetic drift (Duret, 2008).

Convergent adaptation is the repeated evolution of similar traits by the same or different genes (gene reuse vs. function-level convergence) leading to a fitness advantage in multiple independent lineages (Arendt & Reznick, 2008). For three reasons, it provides a useful framework to study adaptive evolution. First, if a gene and consequently a phenotypic trait emerges multiple times as a response to a certain environmental trigger, it provides strong evidence that it is a generally needed response to that trigger (Blount, Lenski, & Losos, 2018). Second, convergence naturally provides much needed replicates to determine the predictability of adaptive evolution (Gould, 1989), specifically, how likely a certain trait evolves by gene reuse (Agrawal & Stinchcombe, 2009). Finally, the replicates may be used to identify genetic factors which are likely underlying this repeatability. Such repeatability may then provide a framework to estimate a likely level of evolutionary predictability across natural cases of adaptation (Stern & Orgogozo, 2009; Yeaman, Gerstein, Hodgins, & Whitlock, 2018). Thus, our attempts to understand adaptive evolution naturally start with a question: how frequently does adaptation repeat itself? Particularly, what is the fraction of the genome reused in adaptation?

However, the answer is not straightforward: individual case studies of convergent adaptation demonstrate large variation. It ranges from absence of any gene reuse (Zou & Zhang, 2015), similarity in functional pathways but not genes (Birkeland et al., 2020; Cooper et al., 2014), reuse of a limited number of genes (Foote et al., 2015; Takuno et al., 2015) to abundant convergence at both gene and functional levels (Lim, Witt, Graham, & Avalos, 2019; Manceau, Domingues, Linnen, Rosenblum, & Hoekstra, 2010). For example, high-altitude adaptation of songbirds in Taiwan repeatedly uses the same single nucleotide polymorphisms (Lai et al., 2019) whereas repeated adaptation to arctic environments in three different species of Brassicaceae is mediated by different genes but comparable molecular pathways (Birkeland et al., 2020). This opens a pressing question: can we predict an evolutionary process which leads to such variable outcomes?

Divergence matters? The role of different sources of adaptive variation.

The divergence between convergently evolving lineages may represent a unifying factor which underlies the variability in the convergently evolving fraction of genome, and, consequently, informs about the predictability of adaptive evolution (Blount et al., 2018). This is an intuitive idea, which has been also supported by some (yet so far indirect) observations. Specifically, phenotype-oriented meta-analyses suggest that both phenotypic convergence (Ord & Summers, 2015) and gene reuse underlying particular phenotypic traits (Conte, Arnegard, Peichel, & Schluter, 2012) decrease with increasing time to the common ancestor. Moreover, my brief review of published genomic studies suggests that genome-wide gene reuse in convergent adaptation tends to scale with divergence (Bohutínská, Vlček, et al., 2021). Thus, there are some hints that divergence provides a significant factor determining gene reuse in adaptation. Having said that, the evidence is scattered in between unrelated reviews and meta analyses, without a dedicated empirical system in which such a relationship could be tested.

An additional question, testable with such empirical inquiry, surrounds the mechanisms underlying gene reuse in adaptation and thus governing its divergence-dependency. There are two types of these mechanisms discussed in the literature, yet again without unified support across divergence scales (Blount et al., 2018; Stern & Orgogozo, 2009; Yeaman et al., 2018).

First mechanism, which may underlie predictability of convergent adaptation between closely related lineages, is allele reuse. It refers to the repeated sweep of the same haplotype that is shared among populations or species either via gene flow or from ancestral (standing) variation (Barrett & Schluter, 2008). For example, the ample gene reuse in repeated high-altitude adaptation of songbirds in Taiwan was possible thanks to their access to shared pool of alleles. These alleles, proven to be beneficial in high-altitude environment, were already present in the common ancestor of the high-altitude lineages, showing the importance of standing genetic variation in mediating adaptation (Lai et al., 2019). While allele reuse has been documented in studies of convergence among closely related lineages (Alves et al., 2019; Haenel, Roesti, Moser, MacColl, & Berner, 2019; Jones et al., 2012; Lai et al., 2019; Oziolor et al., 2019), the alternative scenario, convergent adaptation from independent de-novo mutations targeting the same locus dominates empirical inquiries comparing distantly related taxa (Martin & Orgogozo, 2013; Yeaman et al., 2018). Similarly, some studies report a decreasing probability of hemiplasy (apparent convergence resulting from incomplete lineage sorting) with divergence in phylogeny-based studies (Goldstein, Pollard, Shah, & Pollock, 2015; Mendes, Hahn, & Hahn, 2016). This suggests that the degree of allele reuse may be the primary factor underlying the hypothesized divergence-dependency of gene reuse in convergent adaptation, at least at shallow divergence levels (up to sister species) where incomplete lineage sorting and gene flow are still frequent. Once again, however, such question has not been systematically addressed in a suitable model system varying in divergence.

Second, at deeper evolutionary timescales, allele sharing is essentially ruled out by impermeable reproductive barriers and completed lineage sorting (Hudson & Coyne, 2002). Still, there are reports of convergent adaptation by gene reuse among different genera, families or even kingdoms (Martin & Orgogozo, 2013), indicating that different mechanisms underlie the predictability of gene reuse at deeper divergences. Theory suggests that when

multiple novel beneficial alleles originate within one lineage, positive selection may preferentially act only on a subset of them (Stern, 2010). That is because the adaptive potential of some alleles is constrained by their pleiotropic side-effects, in which individual variants affect the expression of more than one trait, some of which may be then maladaptive (Fisher, 1930; Stern, 2000). Consequently, adaptation is expected to be mediated by a subset of low-pleiotropy genes (Connallon & Hall, 2018). This reduction in the number of possibly adaptive genes can result in increased convergence by repeated selection of the same optimally pleiotropic genes (Stern, 2010). This was shown for example in genes of the anthocyanin pathway, mediating change in the plant petal colour, and in consequence the adaptation to pollinator preferences. While a change in multiple genes of the pathway results in petal colour shift, in nature such change is repeatedly mediated via the most downstream gene of the pathway, which does not have negative side-effects on other functions like UV-protection or pest resistance (Kopp, 2009). Gene functions and the structure of gene regulatory networks together determine the level of pleiotropy. They also change through time, leading to varying pleiotropic constraints between distantly related species (Conte et al., 2012). This variation is increasing with divergence between the convergently adapting species, resulting in decreased probability of gene reuse in adaptation (Martin & Orgogozo, 2013). Thus, the second candidate mechanism for the divergence-dependency of adaptation, is the diversification of pleiotropic constraints.

In summary, theoretical and scattered empirical evidence suggests that the repeatability of adaptation is divergence-dependent. Two mechanisms may underlie this relationship, allele reuse among populations and closely relates species and pleiotropic constraints diversification at deeper divergences. Yet, the limited focus of individual studies of convergent adaptation on a single level of divergence does not allow a unified comparison across divergence scales. Thus, the hypothesis that gene reuse in convergent adaptation scales with divergence has not yet been systematically tested genome-wide and across sufficiently broad divergence scale and the underlying evolutionary mechanisms remain poorly understood.

Beyond gene reuse: the effect of function-level convergence.

Adaptive evolution may also repeat itself at different levels than by selection targeting the same locus (gene reuse). Changes in different (often regulatory) genes may affect expression of the same gene or different genes yet still located in the same functional pathway, leading to repeated acquisition of the same adaptive phenotype (Elmer & Meyer, 2011; Manceau et al., 2010). Such function-level convergence is frequently reported even when gene reuse is absent, perhaps due to the high divergence between the two convergently evolving lineages (Birkeland et al., 2020; Cooper et al., 2014). For example, some vertebrates evolve darker coat color by mutations in the gene *MC1R*, whereas others achieve the very same adaptation by selecting mutation in a different gene, *Agouti*, acting in the same molecular pathway (Kingsley, Manceau, Wiley, & Hoekstra, 2009; Manceau et al., 2010).

Because the probability of adaptation by gene reuse is likely to increase with availability of shared alleles and under higher pleiotropic constraints (see above), one may expect that much of the convergent adaptation among closely related lineages will rely on gene reuse,

not function-level convergence. Under such scenario, function-level convergence should be high once lineages repeatedly adapt via *de-novo* alleles and genes in the adapting molecular pathway have similar level of pleiotropy. In contrast, if the mutational target size causing certain adaptation is high (i.e. high number of sites may mutate to achieve the adaptive phenotype), like in case of loss of function mutations in regulatory genes (Hoekstra & Coyne, 2007; Johanson et al., 2000; Kopp, 2009), the function-level convergence might be prevalent and unrelated to divergence (Yeaman et al., 2018). From this reasoning, we may draw two scenarios of function-level convergence (in its strict definition including only convergence by *different* genes from the same pathway). First, it is prevalent at all divergence levels because multiple independent mutations, targeting different genes, may cause the needed phenotype. Second, it is increasing with decreasing possibility to reuse the same alleles and with lowering pleiotropic constraints, which may lead to its positive divergence-dependency. Yet, available literature does not provide empirical system to support either of these scenarios, or to come with alternative one, leaving space for further investigation. Thus, it is of interest to investigate how the importance of function-level convergence, both in absolute and relative terms, varies with divergence between lineages encompassing convergent adaptation. Further, this brings a question whether function-level convergence may compensate for the rare gene reuse at deeper divergences.

A2: Aims and model systems

The overall aim of my PhD project was to understand genetic mechanisms governing repeatability in adaptive evolution of a genome. To do so, I used a multidisciplinary approach leveraging naturally replicated extreme-adapted plant populations. I used two model selection pressures: adaptation to whole genome duplication (intrinsic change) and adaptation to alpine environment (external environmental change). They provide conveniently strong selection pressures, represented by well-defined set of conditions (Bomblies, 2020; Körner, 2003) and occurring repeatedly across plant species and populations.

I, together with my co-authors, sampled and whole genome resequenced sets of ancestral non-adapted (diploid / foothill) and derived adapted (tetraploid / alpine) populations. Using a combination of experiments, population genome scans for positive selection and statistical modelling, I identified candidate genes associated with adaptation to each factor. Then, I quantified the genome-wide extent of gene reuse and function-level convergence across the repeated instances of each case of adaptation. Finally, I inquired about the source of the candidate adaptive variants using population genomic modelling of different parallel evolution scenarios (Lee & Coop, 2017).

Overall, I analyzed replicated instances of adaptation to a whole genome duplication and to a challenging alpine environment, spanning a range of divergence from populations within plant model *Arabidopsis* to tribes within the plant family Brassicaceae. I tested the following three hypotheses (Fig. 1):

Central hypotheses:

- Gene reuse in convergent adaptation negatively scales with evolutionary divergence between repeatedly adapting populations, species and genera.
- The decreasing allele reuse at short divergences and increasing diversification of pleiotropic constraints at deeper divergences jointly drive this relationship.
- The function-level convergence compensates for limited gene reuse at deeper divergences.

I performed an empirical assessment of gene and function-level parallelism in convergent adaptation across species of the model plant family Brassicaceae, spanning ~0.01 – 25 million years of divergence (Fig. 1, (Hohmann, Wolf, Lysak, & Koch, 2015; Novikova et al., 2016)). I quantified the contribution of allele reuse to evolutionary repeatability, covering various cases of adaptation to whole genome duplication and alpine environment along this divergence scale. Leveraging a uniquely broad set of genetic tools and resources that were developed for the leading plant model *Arabidopsis thaliana*, I addressed specific questions on the effect of allele sharing and pleiotropy on the predictability of gene reuse in adaptation and on the relative importance of function-level convergence. I divided the project into three parts based on the divergence, addressing following objectives:

Objective 1 – the presence of genomic convergence:

Does selection repeatedly target the same genomic regions or functional categories during repeated adaptation towards similar factors (CS1, CS2, CS3, CS5, CS6)?

Objective 2 – varying extent of genomic convergence at shallow divergences:

Does the extent of genomic convergence decrease with increasing genetic divergence between lineages? To what extent is this relationship explained by repeated recruitment of the same adaptive alleles (CS1, CS3)?

Objective 3 – varying extent of genomic convergence at deeper divergences:

What is the role of pleiotropic constraint in genome-wide convergence (CS4, CS5, CS6)? Is the putative decrease of genomic convergence with increasing divergence between lineages compensated by function-level convergence (CS5, SC6)?

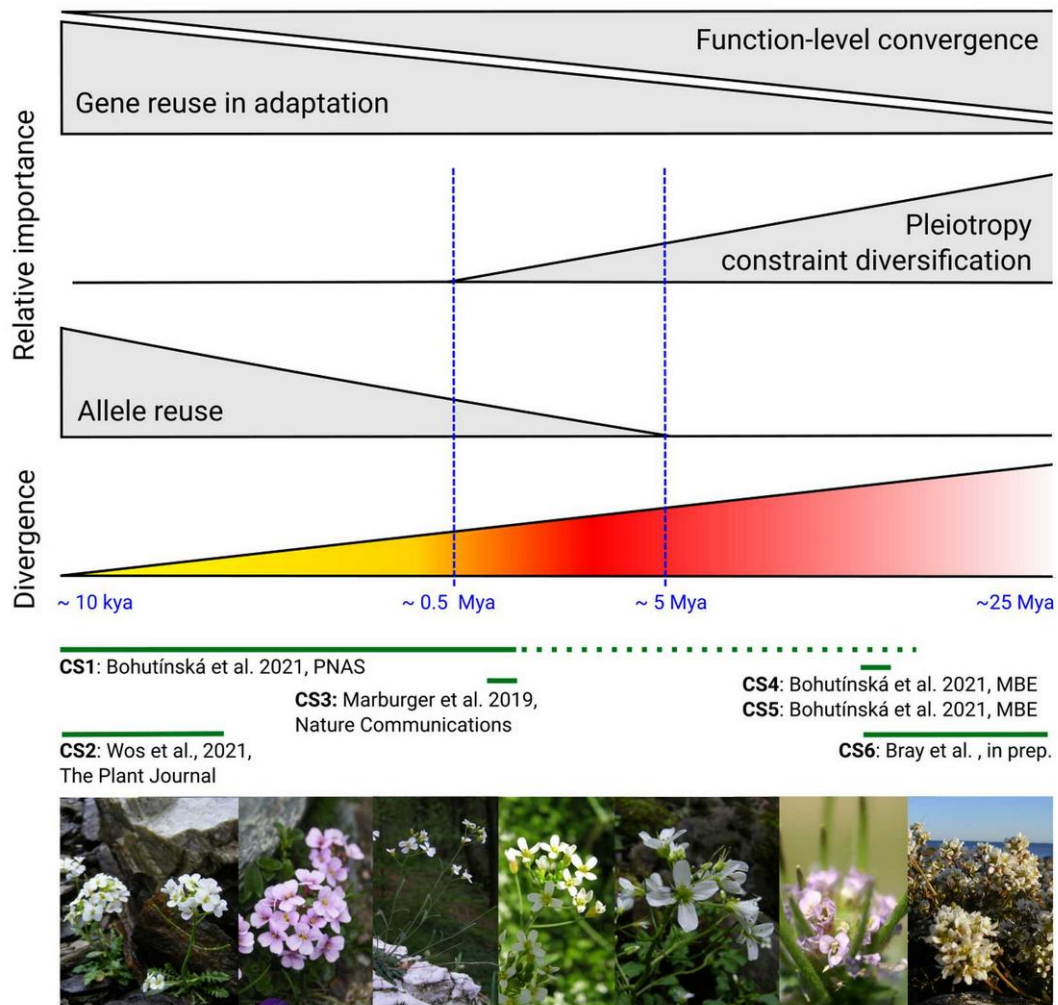


Fig. 1: Hypotheses regarding relationships between genomic basis of convergent adaptation and divergence and the case studies used to address these hypotheses. I propose that genetically closer lineages adapt to a similar challenge more frequently by gene reuse, sampling suitable variants from the shared pool (standing or introgressed allele reuse), which makes their adaptive evolution more predictable. Color ramp symbolizes increasing divergence between the lineages (~0.01 – 25 Mya across my studies). Horizontal green lines show the divergence covered by each of the case studies (numbered following the Author contribution statement). Pictured are species studied throughout the thesis. The divergence scale is not linear in order to aid visualization.

A3: Methods

The principal approach of my PhD project lies in the investigation of naturally replicated cases of adaptation in wild plant populations. I combined systematic field sampling of multiple ancestral non-adapted (diploid / foothill) and derived adapted (tetraploid / alpine) populations, followed by whole genome resequencing of these populations, scans for genes likely targeted by positive selection between them and further interpretation of the identified candidate genes in an evolutionary-history as well as functional context via subsequent statistical analysis, modelling and experiments.

Each of the case studies forming my PhD project started with whole genome resequencing, followed by sequence mapping to the corresponding reference genome and variant calling (following procedure first introduced in (Monnahan et al., 2019)). Using the variant single nucleotide polymorphism (SNP) data, I identified genomic regions showing signatures of positive selection associated with whole genome duplication or alpine colonization. To gather reliable candidates I used a conservative approach combining various selection scans approaches, reviewed in (Oleksyk, Smith, & O'Brien, 2010; Vitti, Grossman, & Sabeti, 2013; Weigand & Leese, 2018). Next, I annotated the candidate genes into corresponding molecular pathways or functions, using Gene ontology enrichment analysis (Alexa & Rahnenführer, 2018) and predictions of protein-protein interactions (Szkarczyk et al., 2015). By overlapping candidate genes and functions among repeated instances of adaptation, I quantified the genome-wide extent of gene reuse and function-level convergence (CS1-CS6). For the whole procedure, I used python3 ScanTools_ProtEvol pipeline, a toolset which I developed by extending the approach of Monnahan et al., 2019 in order to analyze genome differentiation across multiple populations and used follow up R scripts to further interpret and functionally annotate the candidate genes (all available at my GitHub account <https://github.com/mbohutinska>).

To better understand the evolutionary background and functional interpretation of the convergent adaptation cases, I further applied modelling and experiments. First, I used modelling combined with likelihood-based model selection (DMC method, (Lee & Coop, 2017)) and contrasting gene-tree topology weighting (Twisst, (Marburger et al., 2019)) to identify genes in which selection repeatedly acted on alleles shared among populations and species, pointing towards gene reuse (CS1, CS3). Then I used RNASeq to inquire about transcription changes associated with alpine adaptation as a proxy for function-level convergence (CS2). Finally, I used cytology analyses to better understand the phenotypic repeatability in adaptation to whole genome duplication (CS3, CS5).

A4: Key results of my studies

1. Shallow divergence levels (within *Arabidopsis*)

Alpine adaptation

Populations of alpine plants provide suitable systems for addressing mechanisms of convergent adaptation in a genomic context. This is especially true in the species of well-researched *Arabidopsis* genus. The species thrives mostly in low to mid-elevations (up to ~1,000 m a.s.l.) of Central and Eastern Europe, but occasionally occurs in treeless alpine habitats (> 2,000 m a.s.l.) (Knotek et al., 2020). Alpine environments are good models to inquire about convergent adaptation: they pose a spectrum of challenges to plant life and occur as islands in the landscape, potentially triggering directional selection (Körner, 2003). The challenges include freezing and fluctuating temperatures, strong winds, increased UV radiation or a short summer season. Such pressures may result in emergence of distinct alpine morphotypes including contracted rosette plants, dense cushions, large flowers and big roots (Körner, 2003; Christian Rellstab et al., 2020). Indeed, we showed that alpine *A. arenosa* and *A. halleri* constitutively exhibit a distinct morphotype characterized by lower stature, less-lobed and thicker leaves, larger flowers and wider siliques (Bohutínská et al., 2021, CS1; Knotek et al., 2020; Šrámková-Fuxová et al., 2017).

Following up these results, I and co-authors leveraged seven natural replicates of the adaptation to stressful alpine environments in two outcrossing *Arabidopsis* species spanning ~0.6 million years of divergence (Bohutínská et al., 2021, CS1). We analyzed whole genome sequences of 174 individuals from seven *Arabidopsis* lineages and found that the degree of gene reuse in this convergent adaptation strongly depends on genetic divergence between lineages. A designated model-based approach further revealed that the probability of allele reuse (repeated recruitment of the same standing or introgressed variation by positive selection) was the major driver of this pattern. The novelty of this approach lied in the systematic empirical analysis of genome-wide convergence and its underlying causes over a wide range of divergence. Such a unified comparison made it possible to identify divergence as a significant factor shaping the magnitude of genomic parallelism. This highlighted the importance of considering the demographic history of populations, and the consequent availability of standing variation, when interpreting the outcomes of convergent evolution (Bohutínská et al., 2021, CS1).

In a complementary study (Wos, Bohutínská, Nosková, Mandáková, & Kolář, 2021, CS2), we inquired about the presence of function-level convergence across the subset of four out of the seven alpine adaptation cases from the previous study (Bohutínská et al., 2021, CS1), spanning divergence of ~10 – 30 thousand years (Arnold, Kim, & Bomblies, 2015). We used convergent gene expression changes as a proxy for function-level convergence. That is because similar expression shifts of the same genes are hypothesized to lead to the same impact on plant functioning but are often caused by genetic changes in different regulatory genes (Manceau et al., 2010). Thus, although the expression-level convergence still suggests a functional repeatability of adaptation, it is often not driven by gene reuse. Thus,

by comparing leaf transcriptomes of four distinct foothill–alpine population pairs across four treatments, we asked about the functional consequences of possible gene expression convergence in alpine adaptation. We found significant convergence in gene expression at the level of individual loci with an over-representation of genes involved in biotic stress response. In addition, we demonstrated a shared differential response of the originally foothill versus alpine populations to environmental variation across mountain regions. However, the overlap between these parallel expression candidates and our previously identified parallel genomic candidates was very limited, suggesting the relationship between genomic and regulatory convergence is not straightforward as it may be expected by complexity of regulatory networks in plants (Jacobs et al., 2020; Sobel & Streisfeld, 2013). In summary, these results suggest frequent evolutionary repeatability in gene expression changes associated with the colonization of a challenging environment. Such functional repeatability combines constitutive expression differences and plastic interaction with the surrounding environment (Wos, Bohutínská, Nosková, Mandáková, & Kolář, 2021, CS2).

Adaptation to whole genome duplication

Whole genome duplication (WGD), is a massive genomic mutation which is also traumatic event for the cell (Badauel, Bray, Vallejo-Marin, Kolář, & Yant, 2018). Core processes, from meiosis to cell cycle regulation, ion homeostasis and transcription, have to adapt to WGD. Usually this is too much to manage, leading to extinction. From time to time however, a young polyploid lineage adapt to stabilise, potentially possessing benefits (Hollister et al., 2012). In the next study of convergence, we investigated a case where repeated WGDs occurred in two sister *Arabidopsis* species, leading to successful adaptation (Marburger et al., 2019, CS3).

The analysis rested upon the whole genome resequencing of 92 diploid and tetraploid individuals of the two *Arabidopsis* outcrossers, *A. arenosa* and *A. lyrata*, diverging ~ 0.6 million of years ago (Novikova et al., 2016). We identified regions with signals of selective sweeps in tetraploids of both species. Majority of these regions overlapped between the two species, suggesting high degree of gene reuse in adaptation. The reused genes mostly encoded a set of physically and functionally interacting proteins governing meiosis crossover number and distribution. We further demonstrated that the species exchanged the reused alleles via interspecific gene flow, which enhanced rapid and efficient adaptation. Overall, sharing of these alleles allowed the sum of these species to become better than their constituent parts, allowing the repeatedly originated autotetraploids to survive the unstable post-WGD phase and to escape an extinction. This further highlighted allele reuse as a crucial mechanism promoting repeated recruitment of the same allele in adaptation of recently diverged sister species, which are still able to access their shared pool of genetic variation (Marburger et al., 2019, CS3).

2. Deeper divergence levels (beyond *Arabidopsis*)

Alpine adaptation

Alpine adaptation in plants is widespread and broad diversity of Angiosperms can thrive in higher altitudes above a treeline (Körner, 2003). Many cases of alpine adaptation have been

described in the model plant family Brassicaceae (Christian Rellstab et al., 2020; T. Zhang et al., 2019). Plants of this family are often characterised by small genomes and their relations to model plant *Arabidopsis thaliana* make them good models for genomic studies of adaptation. Indeed, Brassicaceae literature involves six genome-wide studies of alpine adaptation (Günther, Lampei, Barilar, & Schmid, 2016; Hämälä & Savolainen, 2019; Kubota et al., 2015; C. Rellstab et al., 2017; J. Zhang et al., 2016; T. Zhang et al., 2019), including five species diverging 0.5 – 18 millions of years ago (Hohmann et al., 2015; Novikova et al., 2016). Thus, we made use of the available candidate gene lists from these studies, complemented them with our parallel candidate gene lists of *Arabidopsis arenosa* and *A. halleri* and tested whether the relationship between the degree of gene reuse and divergence persists at deeper phylogenetic scales (Bohutínská et al., 2021, CS1). We were able to identify significant gene reuse among different *Arabidopsis* species and function-level convergence among the Brassicaceae genera. However, the degree of gene reuse was significantly higher for comparisons within a genus (*Arabidopsis*) than between genera while such a trend was absent for convergent functions. That suggests that there are limits to gene reuse at above genus-level divergences. However, the limited degree of gene reuse did not allow to test for the contribution of pleiotropic constraints to genomic convergence in this model system. Taken together, these results suggest that there are likely similar functions associated with alpine adaptation among different lineages, species and even genera from distinct tribes of Brassicaceae. Yet, the probability of reusing the same genes within these functions decreases with increasing divergence among the lineages, thus reducing the chance to identify gene reuse among diverged lineages (Bohutínská et al., 2021, CS1).

Adaptation to whole genome duplication

The instant meiotic and physiological consequences of WGD necessitate the concerted adjustment of a wide range of core functions (Bomblies, 2020), but nevertheless natural outcomes of WGD have repeatedly survived across kingdoms (Bomblies, Higgins, & Yant, 2015). Given this repeated adaptation despite obvious challenges, we asked: how do distant and thus genetically independent lineages survive WGD? Are the solutions to this well-defined selection pressure constrained to a limited set of genes, suggesting gene reuse due to the pleiotropy constraints? To answer this, we investigated a repeated adaptation to WGD between three distant Brassicaceae species, *Arabidopsis arenosa* (Bohutínská, Handrick, et al., 2021, CS4), *Cardamine amara* (Bohutínská, Alston, et al., 2021, CS5) and *Cochlearia* species complex (Bray et al., 2020, CS6), separated by 25 million years (Hohmann et al., 2015).

First, we studied the adaptation to whole genome duplication in *A. arenosa*, which has become an increasingly important model for understanding the causes and consequences of the adaptive evolution of meiosis (Bohutínská, Handrick, et al., 2021, CS4). We performed sets of genome scans for tetraploid-specific selection in *A. arenosa* and identified a number of candidate genes for the evolution of tetraploid stability. We further found that the magnitude of molecular adaptation to tetraploidy, both in terms of number of loci and the extent of SNP changes, much exceeded our expectations based on the adaptive evolution of diploid lineages. In tetraploids, the number of positively selected amino acid changes, and the extent of selection on conserved amino acids and amino acids with predicted functional changes stand out both qualitatively and quantitatively. We also found that it was unlikely

that most of the tetraploid-specific alleles were selected from standing variation, suggesting they mostly accumulated by evolution and/or co-evolution of novel alleles in the tetraploid lineage. These findings had fundamental implications for understanding how essential conserved processes can respond to sudden selection pressures.

Building on the list of candidate genes for the evolution of tetraploid stability in *A. arenosa* (and the same adaptive alleles introgressed with *A. lyrata*, see Results of CS3), we next asked if these solutions have been repeated in a divergent yet still genomically comparable Brassicaceae species *C. amara* (Bohutínská, Alston, et al., 2021, CS5). Our analysis rested upon the population genomic scans for selection and cytological phenotyping in the diverse wild outcrosser *C. amara*, which experienced a WGD event fully independent of *A. arenosa* and *A. lyrata* (~ 17 million years of divergence (Hohmann et al., 2015)). Importantly, *C. amara* is a perennial herb harbouring high level of genetic diversity (similar to both *A. arenosa* and *A. lyrata*) and shares a similar evolutionary history, with a likely single geographic origin, followed by autotetraploid expansion associated with glacial oscillations (Zozomová-Lihová et al., 2015). We sampled 100 *C. amara* individuals in a replicated resequencing scheme, generated a novel reference genome, and cytologically assessed both cytotypes to understand meiotic behavior before and after WGD. We localized signals of selection to gene-sized peaks in each independently formed tetraploid, comparing *C. amara* with *A. arenosa*.

We discovered that the specific genes required to repeatedly adapt to WGD were remarkably flexible. Our results pointed to a minimally constrained, highly polygenic basis for the distributed control of meiosis, DNA repair, and cell cycle following WGD. While this observation supported our hypothesis about decreasing gene reuse with increasing divergence, the limited gene overlap did not allow us to test for the effect of pleiotropic constraints on gene reuse.

Despite the minimal gene reuse, we found a strong support for function-level convergence. Both species adapted by the same functional classes (as determined by gene ontology enrichment) and by the same protein-protein interaction networks. Based on this, we concluded that there are multiple solutions to WGD-associated challenges, allowing diverse species to establish as autopolyploids. We further showed that gene reuse was limited at the divergence levels where species were no longer able to share potentially adaptive alleles. This was in contrast to our expectations, suggesting that pleiotropic constraints likely not largely affected adaptation to whole genome duplication. Finally, we propose that function-level convergence, i.e. adaptation by different genes leading to the same function, may be the dominant mechanism of convergent adaptation between distantly related species, reflecting the decreasing relative importance of gene reuse (Bohutínská, Alston, et al., 2021, CS5).

In the final study, we extended the divergence scale up to ~ 25 million years by including diploid-autotetraploid *Cochlearia* system from Great Britain (Bray et al, 2020). Like *C. amara* and *A. arenosa*, the British diploid and tetraploid populations are also characterised by high genetic diversity and evolutionary history involving likely single origin of autotetraploids and their following spread (Bray et al., 2020, CS6; Gill, 2008). We detected genes and processes under selection following WGD in the *Cochlearia* species complex by performing a scan for selective sweeps following WGD by resequencing two diploid and six tetraploid populations of British *Cochlearia*. We then contrasted our results with two independent WGDs in

Arabidopsis arenosa and *Cardamine amara*. Similarly to previous study, we found that the specific genes recruited to respond to WGD were highly flexible among the species, suggesting minimal gene reuse. However, we again found that WGD required the evolution of similar convergently adapting functional processes in all three cases, further supporting the observation that function-level convergence compensates for limited gene reuse (Bray et al., 2020, CS6).

A5: Conclusions: Divergence matters!

In my PhD project, I analyzed genome-wide variation over multiple instances of naturally replicated extreme-adapted populations and species. To address my objectives more broadly, I studied cases of genomic adaptation towards two distinct selective agents: alpine environment (CS1, CS2) and whole genome duplication (CS3 – CS6). I worked with convergence across a broad divergence scale, starting at shallow divergences among populations of the same species, extending up to deeper divergences among different tribes of the family Brassicaceae (Hohmann et al., 2015).

First, I asked for the presence of genomic convergence – if positive selection repeatedly targets the same genomic regions or functional categories in cases of repeated adaptation. Using a combination of population genome scans for positive selection and statistical modelling, I identified significant genomic convergence at the level of gene reuse and repeated functional pathways (CS1 – CS6). I further empirically demonstrated that the extent of gene reuse decreases with increasing divergence between compared lineages. In contrast, function-level convergence has been present across the whole divergence scale studied and, unlike gene reuse, its magnitude in absolute terms did not scale with divergence (CS1, CS2, CS5, CS6).

Second, I inquired about mechanisms underlying varying extent of genomic convergence at shallow divergence levels. I showed that the negative relationship between gene reuse and divergence was largely explained by the decreasing role of allele reuse among related *Arabidopsis* lineages, which share ancestral or introgressed alleles (CS1, CS3). That could possibly reflect either genetic (weak hybridization barriers, widespread ancestral polymorphism between closely related lineages (Hudson & Coyne, 2002) or ecological reasons (lower niche differentiation and geographical proximity (Bradburd & Ralph, 2019; Graham, Storch, & Machac, 2018)). Further, I identified a significant function-level convergence, the reuse of different genes from the same pathway, as a mechanism further increasing repeatability among closely related lineages (CS2).

Third, I investigated possible factors explaining the varying extent of genomic convergence at deeper divergences, to test for the effect of pleiotropic constraint diversification on gene reuse. In contrast to high gene reuse among closely related populations and species, I identified critically low levels of gene reuse among less related species (CS4 – CS6). This offers two hypotheses for the role of pleiotropy constraints diversification on shaping genomic convergence: (i) it has a significant impact on the limited gene reuse and high degree of function-level convergence by directing selection towards different low pleiotropy genes in the distant Brassicaceae species, (ii) it has only minor role and gene reuse at higher divergences has other causes or is purely stochastic. These hypotheses could be further followed by understanding of function-level convergence among distantly related lineages in the context of the selected gene regulatory networks. Finally, the low gene reuse in convergent adaptation among distantly related species of the family Brassicaceae was compensated by significant function-level convergence, in which species repeatedly adapted by different genes involved in the same functional pathways (CS5, CS6). That corresponds to the reports of frequent function-level convergence in other genomic studies of convergent adaptation among diverged species (Birkeland et al., 2020; Cooper et al., 2014; Whiting et

al., 2021).

Altogether, these findings provide empirical support for predictions on genetic convergence (Conte et al., 2012; Ord & Summers, 2015), and unravel general mechanisms that may help explain ample variability in the extent of genomic convergence in adaptation that was reported, yet remained unexplained, in many case studies (Blount et al., 2018; Hibbins, Gibson, & Hahn, 2020; Martin & Orgogozo, 2013; Morales et al., 2019). The decreasing role of allele reuse with divergence agrees with theoretical expectations that the evolutionary potential of a population depends on the availability of preexisting (standing or introgressed) genetic variation (Barrett & Schluter, 2008; Ralph & Coop, 2015; Thompson, Osmond, & Schluter, 2019) and that the extent of shared polymorphism decreases with increasing differentiation between diverging lineages (Charlesworth, Charlesworth, & Barton, 2003; Hudson & Coyne, 2002). The overall low gene reuse at higher divergence levels highlights function-level convergence as an important factor underlying repeatability of adaptive evolution (Manceau et al., 2010; Whiting et al., 2021). In general, my PhD project brings an empirical contribution to the understanding of the drivers of repeated convergent adaptation and the following (un)predictability of adaptive evolution. It further demonstrates the importance of a quantitative understanding of divergence for the assessment of predictability of adaptive evolution (Blount et al., 2018) and brings support to the emerging view of the ubiquitous influence of divergence scale on different evolutionary and ecological mechanisms (Graham et al., 2018).

Based on my results, I suggest that further studies shall focus on genome-wide convergence among distantly related lineages in the context of regulatory networks to better understand the role of pleiotropic constraint diversification in shaping the magnitude of gene reuse. Further, it is of interest to test if the role of varying pleiotropic constraints is diminished within a species due to frequent sharing of beneficial alleles. Finally, further quantitative and functional genetic research shall bring experimental evidence for fitness advantage of the reused alleles and for their possible pleiotropic consequences.

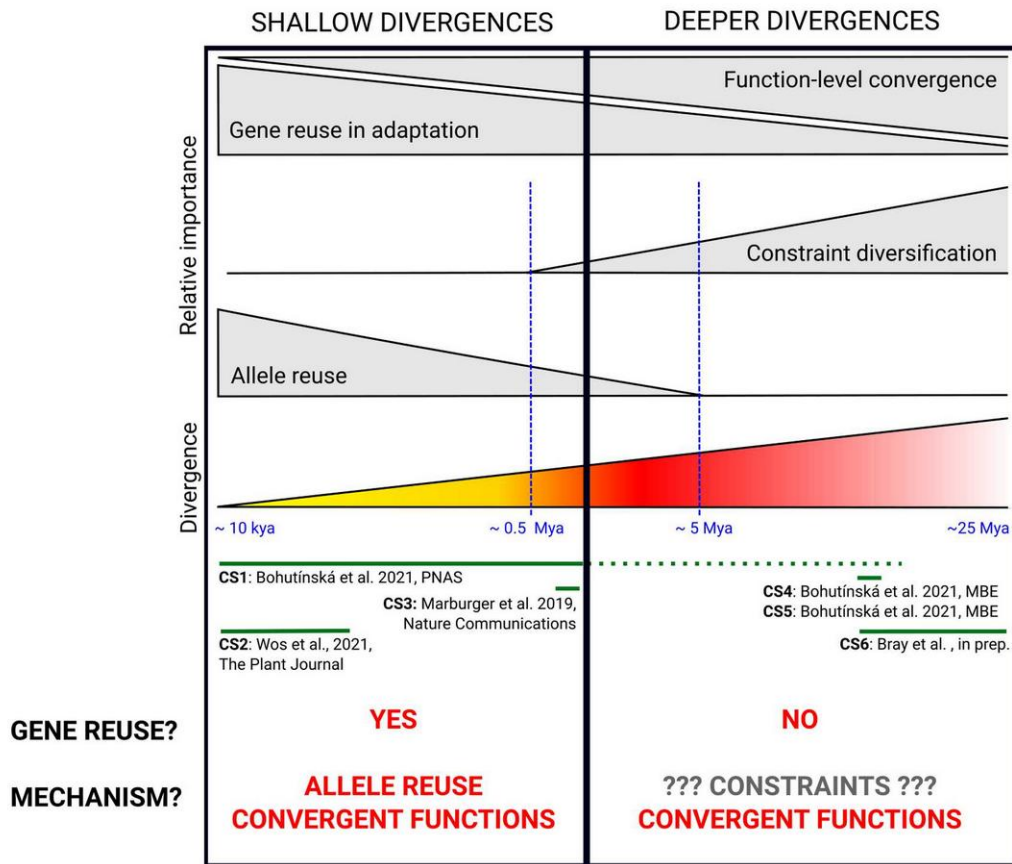


Fig. 2: Schematic summary of the principal findings of my PhD project on the relationships between the genomic basis of convergent adaptation and divergence.

References

- Agrawal, A. F., & Stinchcombe, J. R. (2009). How much do genetic covariances alter the rate of adaptation? *Proceedings of the Royal Society B: Biological Sciences*, 276(1659), 1183–1191. <https://doi.org/10.1098/rspb.2008.1671>
- Alexa, A., & Rahnenführer, J. (2018). *Gene set enrichment analysis with topGO*. Retrieved from <http://www.mpi-sb.mpg.de/~alexa>
- Alves, J. M., Carneiro, M., Cheng, J. Y., Matos, A. L. de, Rahman, M. M., Loog, L., ... Jiggins, F. M. (2019). Parallel adaptation of rabbit populations to myxoma virus. *Science*, 363(6433), 1319–1326. <https://doi.org/10.1126/SCIENCE.AAU7285>
- Arendt, J., & Reznick, D. (2008). Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology and Evolution*, 23(1), 26–32. <https://doi.org/10.1016/j.tree.2007.09.011>
- Arnold, B., Kim, S.-T., & Bomblies, K. (2015). Single Geographic Origin of a Widespread Autotetraploid *Arabidopsis arenosa* Lineage Followed by Interploidy Admixture. *Molecular Biology and Evolution*, 32(6), 1382–1395. <https://doi.org/10.1093/molbev/msv089>
- Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F., & Yant, L. (2018, August 20). The “Polyploid Hop”: Shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Frontiers in Ecology and Evolution*. Frontiers Media S.A. <https://doi.org/10.3389/fevo.2018.00117>
- Barrett, R. D. H., & Schluter, D. (2008, January 1). Adaptation from standing genetic variation. *Trends in Ecology and Evolution*. Elsevier Current Trends. <https://doi.org/10.1016/j.tree.2007.09.008>
- Birkeland, S., Lovisa, A., Gustafsson, S., Brysting, A. K., Brochmann, C., & Nowak, M. D. (2020). Multiple genetic trajectories to extreme abiotic stress adaptation in Arctic Brassicaceae. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msaa068/5804990>
- Blount, Z. D., Lenski, R. E., & Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life’s tape. *Science*, 362(655). <https://doi.org/10.1126/SCIENCE.AAM5979>
- Bohutínská, M., Alston, M., Monnahan, P., Mandáková, T., Bray, S., Paajanen, P., ... Yant, L. (2021). Novelty and convergence in adaptation to whole genome duplication. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msab096>
- Bohutínská, M., Handrick, V., Yant, L., Schmickl, R., Kolář, F., Bomblies, K., & Paajanen, P. (2021). *De-novo* mutation and rapid protein (co-)evolution during meiotic adaptation in *Arabidopsis arenosa*. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msab001>

- Bohutínská, M., Vlček, J., Yair, S., Leanen, B., Konečná, V., Fracassetti, M., ... Kolář, F. (2021). Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives. *Proceedings of the National Academy of Sciences of the United States of America*. <https://doi.org/10.1073/pnas.2022713118>
- Bomblies, K. (2020). When everything changes at once: finding a new normal after genome duplication. *Proceedings of the Royal Society B: Biological Sciences*, 287(1939), 20202154. <https://doi.org/10.1098/rspb.2020.2154>
- Bomblies, K., Higgins, J. D., & Yant, L. (2015). Meiosis evolves: adaptation to external and internal environments. *New Phytologist*, 208(2), 306–323. <https://doi.org/10.1111/nph.13499>
- Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017, October 30). Detecting positive selection in the genome. *BMC Biology*. BioMed Central Ltd. <https://doi.org/10.1186/s12915-017-0434-y>
- Bradburd, G. S., & Ralph, P. L. (2019). Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics*, 50, 427–449. <https://doi.org/10.1146/annurev-ecolsys-110316>
- Bray, S. M., Wolf, E. M., Zhou, M., Busoms, S., Bohutinska, M., Paajanen, P., ... Yant, L. (2020). Convergence and novelty in adaptation to whole genome duplication in three independent polyploids. *BioRxiv*, 2020.03.31.017939. <https://doi.org/10.1101/2020.03.31.017939>
- Charlesworth, B., Charlesworth, D., & Barton, N. H. (2003). The Effects of Genetic and Geographic Structure on Neutral Variation. *Annual Review of Ecology, Evolution, and Systematics*, 34(1), 99–125. <https://doi.org/10.1146/annurev.ecolsys.34.011802.132359>
- Connallon, T., & Hall, M. D. (2018). Genetic constraints on adaptation: a theoretical primer for the genomics era. *Annals of the New York Academy of Sciences*, 1422(1), 65–87. <https://doi.org/10.1111/nyas.13536>
- Conte, G. L., Arnegard, M. E., Peichel, C. L., & Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proc. R. Soc. B*, 279, 5039–5047. <https://doi.org/10.1098/rspb.2012.2146>
- Cooper, K. L., Sears, K. E., Uygur, A., Maier, J., Baczkowski, K.-S., Brosnahan, M., ... Tabin, C. J. (2014). Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature*, 511(7507), 41–45. <https://doi.org/10.1038/nature13496>
- Duret, L. (2008). Neutral Theory: The Null Hypothesis of Molecular Evolution | Learn Science at Scitable. *Nature Education*, 1(1). Retrieved from <https://www.nature.com/scitable/topicpage/neutral-theory-the-null-hypothesis-of-molecular-839/>
- Elmer, K. R., & Meyer, A. (2011, June 1). Adaptation in the age of ecological genomics: Insights from parallelism and convergence. *Trends in Ecology and Evolution*. Elsevier Current Trends. <https://doi.org/10.1016/j.tree.2011.02.008>

- Fisher, R. A. (1930). *The genetical theory of natural selection. The genetical theory of natural selection*. Clarendon Press. <https://doi.org/10.5962/bhl.title.27468>
- Foote, A. D., Liu, Y., Thomas, G. W. C., Vinař, T., Alföldi, J., Deng, J., ... Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature GeNetics*, 47. <https://doi.org/10.1038/ng.3198>
- Gill, E. (2008). Conservation genetics of the species complex *Cochlearia officinalis* L. s.l. in Britain.
- Goldstein, R. A., Pollard, S. T., Shah, S. D., & Pollock, D. D. (2015). Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Molecular Biology and Evolution*, 32(6), 1373–1381. <https://doi.org/10.1093/molbev/msv041>
- Gould, S. J. (1989). *Wonderful life : the Burgess Shale and the nature of history*. Norton.
- Graham, C. H., Storch, D., & Machac, A. (2018). Phylogenetic scale in ecology and evolution. *Global Ecology and Biogeography*, 27(2), 175–187. <https://doi.org/10.1111/geb.12686>
- Günther, T., Lampei, C., Barilar, I., & Schmid, K. J. (2016). Genomic and phenotypic differentiation of *Arabidopsis thaliana* along altitudinal gradients in the North Italian Alps. *Molecular Ecology*, 25(15), 3574–3592. <https://doi.org/10.1111/mec.13705>
- Haenel, Q., Roesti, M., Moser, D., MacColl, A. D. C., & Berner, D. (2019). Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evolution Letters*, 3(1), 28–42. <https://doi.org/10.1002/evl3.99>
- Hämälä, T., & Savolainen, O. (2019). Genomic Patterns of Local Adaptation under Gene Flow in *Arabidopsis lyrata*. *Molecular Biology and Evolution*, 32(11), 2557–2571. <https://doi.org/10.1093/molbev/msz149>
- Hibbins, M. S., Gibson, M. J. S., & Hahn, M. W. (2020). Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. *BioRxiv*, 2020.04.15.043752. <https://doi.org/10.1101/2020.04.15.043752>
- Hoekstra, H. E., & Coyne, J. A. (2007). THE LOCUS OF EVOLUTION: EVO DEVO AND THE GENETICS OF ADAPTATION. *Evolution*, 61(5), 995–1016. <https://doi.org/10.1111/j.1558-5646.2007.00105.x>
- Hohmann, N., Wolf, E. M., Lysak, M. A., & Koch, M. A. (2015). A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. *The Plant Cell*, 27(10), 2770–2784. <https://doi.org/10.1105/tpc.15.00482>
- Hollister, J. D., Arnold, B. J., Svedin, E., Xue, K. S., Dilkes, B. P., & Bomblies, K. (2012). Genetic Adaptation Associated with Genome-Doubling in Autotetraploid *Arabidopsis arenosa*. *PLoS Genetics*, 8(12), e1003093. <https://doi.org/10.1371/journal.pgen.1003093>
- Hudson, R. R., & Coyne, J. A. (2002). Mathematical consequences of the genealogical species concept. *Evolution*, 56(8), 1557–1565. <https://doi.org/10.1111/j.0014->

- Jacobs, A., Carruthers, M., Yurchenko, A., Gordeeva, N. V., Alekseyev, S. S., Hooker, O., ... Elmer, K. R. (2020). Parallelism in eco-morphology and gene expression despite variable evolutionary and genomic backgrounds in a Holarctic fish. *PLOS Genetics*, *16*(4), e1008658. <https://doi.org/10.1371/journal.pgen.1008658>
- Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., & Dean, C. (2000). Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science*, *290*(5490), 344–347. <https://doi.org/10.1126/science.290.5490.344>
- Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., ... Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. <https://doi.org/10.1038/nature10944>
- Kingsley, E. P., Manceau, M., Wiley, C. D., & Hoekstra, H. E. (2009). Melanism in *Peromyscus* is caused by independent mutations in Agouti. *PLoS ONE*, *4*(7), 6435. <https://doi.org/10.1371/journal.pone.0006435>
- Knotek, A., Konečná, V., Wos, G., Požárová, D., Šrámková, G., Bohutínská, M., ... Kolář, F. (2020). Parallel Alpine Differentiation in *Arabidopsis arenosa*. *Frontiers in Plant Science*, *11*, 1949. <https://doi.org/10.3389/fpls.2020.561526>
- Kopp, A. (2009). Metamodels and phylogenetic replication: a systematic approach to the evolution of developmental pathways. *Evolution*, *63*(11), 2771–2789. <https://doi.org/10.1111/j.1558-5646.2009.00761.x>
- Körner, C. (2003). *Alpine Plant Life*. Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-98018-3>
- Kubota, S., Iwasaki, T., Hanada, K., Nagano, A. J., Fujiyama, A., Toyoda, A., ... Morinaga, S. I. (2015). A Genome Scan for Genes Underlying Microgeographic-Scale Local Adaptation in a Wild *Arabidopsis* Species. *PLoS Genetics*, *11*(7), e1005361. <https://doi.org/10.1371/journal.pgen.1005361>
- Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., ... Li, S.-H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(6), 2152–2157. <https://doi.org/10.1073/pnas.1813597116>
- Lande, R., & Shannon, S. (1996). *The role of genetic variation in adaptation and population persistence in a changing environment*. *BRIEF COMMUNICATIONS Evolution* (Vol. 50).
- Lee, K. M., & Coop, G. (2017). Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data. *Genetics*, *207*(4), 1591–1619. <https://doi.org/10.1534/GENETICS.117.300417>
- Lim, M. C. W., Witt, C. C., Graham, C. H., & Avalos, L. M. D. (2019). Parallel Molecular Evolution in Pathways, Genes, and Sites in High-Elevation Hummingbirds Revealed by Comparative Transcriptomics. *Genome Biol. Evol.*, *11*(6), 1573–1585.

<https://doi.org/10.5061/dryad.v961mb4>

- Manceau, M., Domingues, V. S., Linnen, C. R., Rosenblum, E. B., & Hoekstra, H. E. (2010). Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 365(1552), 2439–2450. <https://doi.org/10.1098/rstb.2010.0104>
- Marburger, S., Monnahan, P., Seear, P. J., Martin, S. H., Koch, J., Paajanen, P., ... Yant, L. (2019). Interspecific introgression mediates adaptation to whole genome duplication. *Nature Communications*, 10(1). <https://doi.org/10.1038/s41467-019-13159-5>
- Martin, A., & Orgogozo, V. (2013). The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution*, 67(5), 1235–1250. <https://doi.org/10.1111/evo.12081>
- Mendes, F., Hahn, Y., & Hahn, M. W. (2016). Gene Tree Discordance Can Generate Patterns of Diminishing Convergence over Time. *Molecular Biology and Evolution*, 3299–3307. <https://doi.org/10.1093/molbev/msw197>
- Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*. Elsevier Ltd. <https://doi.org/10.1016/j.tree.2013.08.003>
- Monnahan, P., Kolář, F., Baduel, P., Sailer, C., Koch, J., Horvath, R., ... Yant, L. (2019). Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nature Ecology and Evolution*, 3(3). <https://doi.org/10.1038/s41559-019-0807-4>
- Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2019). Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Science Advances*, 5(12). <https://doi.org/10.1126/sciadv.aav9963>
- Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., ... Nordborg, M. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, 48(9), 1077–1082. <https://doi.org/10.1038/ng.3617>
- Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537), 185–205. <https://doi.org/10.1098/rstb.2009.0219>
- Ord, T. J., & Summers, T. C. (2015). Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evolutionary Biology*, 15(1), 137. <https://doi.org/10.1186/s12862-015-0424-z>
- Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., Guberman VerPloeg, S., Bruns, P. C., ... Matson, C. W. (2019). Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science (New York, N.Y.)*, 364(6439), 455–457. <https://doi.org/10.1126/science.aav4155>

- Ralph, P. L., & Coop, G. (2015). The Role of Standing Variation in Geographic Convergent Adaptation. *The American Naturalist*, 186(S1), S5-23. <https://doi.org/10.1086/682948>
- Rellstab, C., Fischer, M. C., Zoller, S., Graf, R., Tedder, A., Shimizu, K. K., ... Gugerli, F. (2017). Local adaptation (mostly) remains local: Reassessing environmental associations of climate-related candidate SNPs in *Arabidopsis halleri*. *Heredity*, 118(2), 193–201. <https://doi.org/10.1038/hdy.2016.82>
- Rellstab, Christian, Zoller, S., Sailer, C., Tedder, A., Gugerli, F., Shimizu, K. K., ... Fischer, M. C. (2020). Genomic signatures of convergent adaptation to Alpine environments in three Brassicaceae species. *Molecular Ecology*, mec.15648. <https://doi.org/10.1111/mec.15648>
- Sobel, J. M., & Streisfeld, M. A. (2013). Flower color as a model system for studies of plant evo-devo. *Frontiers in Plant Science*, 4, 321. <https://doi.org/10.3389/fpls.2013.00321>
- Šrámková-Fuxová, G., Závěská, E., Kolář, F., Lučanová, M., Španiel, S., & Marhold, K. (2017). Range-wide genetic structure of *Arabidopsis halleri* (Brassicaceae): glacial persistence in multiple refugia and origin of the Northern Hemisphere disjunction. *Botanical Journal of the Linnean Society*, 185(3), 321–342. <https://doi.org/10.1093/botlinnean/box064>
- Stern, D. (2010). *Evolution, Development, and the Predictable Genome*. Retrieved from <https://www.nhbs.com/evolution-development-and-the-predictable-genome-book>
- Stern, D. L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution*. Society for the Study of Evolution. <https://doi.org/10.1111/j.0014-3820.2000.tb00544.x>
- Stern, David L., & Orgogozo, V. (2009, February 6). Is genetic evolution predictable? *Science*. Cambridge Univ. Press. <https://doi.org/10.1126/science.1158997>
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... Von Mering, C. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1), D447–D452. <https://doi.org/10.1093/nar/gku1003>
- Takuno, S., Ralph, P., Swarts, K., Elshire, R. J., Glaubitz, J. C., Buckler, E. S., ... Ross-Ibarra, J. (2015). Independent Molecular Basis of Convergent Highland Adaptation in Maize. *Genetics*, 200(4), 1297–1312. <https://doi.org/10.1534/genetics.115.178327>
- Thompson, K. A., Osmond, M. M., & Schluter, D. (2019). Parallel genetic evolution and speciation from standing variation. *Evolution Letters*, 3(2), 129–141. <https://doi.org/10.1002/evl3.106>
- Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, 47, 97–120. <https://doi.org/10.1146/annurev-genet-111212-133526>
- Weigand, H., & Leese, F. (2018). Detecting signatures of positive selection in non-model species using genomic data. *Zoological Journal of the Linnean Society*, 184(2), 528–

583. <https://doi.org/10.1093/zoolinnean/zly007>

- Whiting, J. R., Paris, J. R., van der Zee, M. J., Parsons, P. J., Weigel, D., & Fraser, B. A. (2021). Drainage-structuring of ancestral variation and a common functional pathway shape limited genomic convergence in natural high- and low-predation guppies. *PLOS Genetics*, *17*(5), e1009566. <https://doi.org/10.1371/journal.pgen.1009566>
- Wos, G., Bohutínská, M., Nosková, J., Mandáková, T., & Kolář, F. (2021). Parallelism in gene expression between foothill and alpine ecotypes in *Arabidopsis arenosa*. *The Plant Journal*, tpj.15105. <https://doi.org/10.1111/tpj.15105>
- Yeaman, S., Gerstein, A. C., Hodgins, K. A., & Whitlock, M. C. (2018). Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genetics*, *14*(10), e1007717. <https://doi.org/10.1371/journal.pgen.1007717>
- Zhang, J., Tian, Y., Yan, L., Zhang, G., Wang, X., Zeng, Y., ... Sheng, J. (2016). Genome of Plant Maca (*Lepidium meyenii*) Illuminates Genomic Basis for High-Altitude Adaptation in the Central Andes. *Molecular Plant*, *9*(7), 1066–1077. <https://doi.org/10.1016/j.molp.2016.04.016>
- Zhang, T., Qiao, Q., Novikova, P. Y., Wang, Q., Yue, J., Guan, Y., ... Qiong, L. (2019). Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude. *Proceedings of the National Academy of Sciences*, *116*(14), 7137–7146. <https://doi.org/10.1073/PNAS.1817580116>
- Zou, Z., & Zhang, J. (2015). No Genome-Wide Protein Sequence Convergence for Echolocation. *Molecular Biology and Evolution*, *32*(5), 1237–1241. <https://doi.org/10.1093/molbev/msv014>
- Zozomová-Lihová, J., Malánová-Krásná, I., Vít, P., Urfus, T., Senko, D., Svitok, M., ... Marhold, K. (2015). Cytotype distribution patterns, ecological differentiation, and genetic structure in a diploid-tetraploid contact zone of *Cardamine amara*. *American Journal of Botany*, *102*(8), 1380–1395. <https://doi.org/10.3732/ajb.1500052>

Part B – Case studies

- Case study 1: Bohutínská, M., Vlček, J., Yair, S., Leanen, B., Konečná, V., Fracassetti, M., Slotte, T., & Kolář, F. (2021). Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives. *Proceedings of the National Academy of Sciences of the United States of America*. doi.org/10.1073/pnas.2022713118
- Case study 2: Wos, G., Bohutínská, M., Nosková, J., Mandáková, T., & Kolář, F. (2021). Parallelism in gene expression between foothill and alpine ecotypes in *Arabidopsis arenosa*. *The Plant Journal*, tpj.15105. <https://doi.org/10.1111/tpj.15105>
- Case study 3: Marburger, S., Monnahan, P., Seear, P. J., Martin, S. H., Koch, J., Paajanen, P., Bohutínská, M., Higgins, J., Schmickl, R., & Yant, L. (2019). Interspecific introgression mediates adaptation to whole genome duplication. *Nature Communications*, 10(1). doi.org/10.1038/s41467-019-13159-5
- Case study 4: Bohutínská, M., Handrick, V., Yant, L., Schmickl, R., Kolář, F., Bomblies, K., & Paajanen, P. (2021). De-novo mutation and rapid protein (co-)evolution during meiotic adaptation in *Arabidopsis arenosa*. *Molecular Biology and Evolution*. doi.org/10.1093/molbev/msab001
- Case study 5: Bohutínská, M., Alston, M., Monnahan, P., Mandáková, T., Bray, S., Paajanen, P., Kolář, F., & Yant, L. (2021). Novelty and convergence in adaptation to whole genome duplication. *Molecular Biology and Evolution*. doi.org/10.1093/molbev/msab096
- Case study 6: Bray, S. M., Wolf, E. M., Zhou, M., Busoms, S., Bohutínská, M., Paajanen, P., Monnahan, P., Koch, J., Fisher, S., Koch, M., & Yant, L. (2020). Convergence and novelty in adaptation to whole genome duplication in three independent polyploids. *BioRxiv*. doi.org/10.1101/2020.03.31.017939 (manuscript)

Case study 1.

Genomic basis of parallel adaptation varies with divergence in
Arabidopsis and its relatives





Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives

Magdalena Bohutínská^{a,b,1}, Jakub Vlček^{a,c,d}, Sivan Yair^e, Benjamin Laenen^f, Veronika Konečná^{a,b}, Marco Fracassetti^f, Tanja Slotte^f, and Filip Kolář^{a,b,1}

^aDepartment of Botany, Faculty of Science, Charles University, 128 01 Prague, Czech Republic; ^bInstitute of Botany, Czech Academy of Sciences, 252 43 Průhonice, Czech Republic; ^cBiology Centre, Czech Academy of Sciences, 370 05 České Budějovice, Czech Republic; ^dDepartment of Zoology, Faculty of Science, University of South Bohemia, 370 05 České Budějovice, Czech Republic; ^eCenter for Population Biology, University of California, Davis, CA 95616; and ^fDepartment of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden

Edited by Joy M. Bergelson, The University of Chicago, Chicago, IL, and approved March 30, 2021 (received for review October 30, 2020)

Parallel adaptation provides valuable insight into the predictability of evolutionary change through replicated natural experiments. A steadily increasing number of studies have demonstrated genomic parallelism, yet the magnitude of this parallelism varies depending on whether populations, species, or genera are compared. This led us to hypothesize that the magnitude of genomic parallelism scales with genetic divergence between lineages, but whether this is the case and the underlying evolutionary processes remain unknown. Here, we resequenced seven parallel lineages of two *Arabidopsis* species, which repeatedly adapted to challenging alpine environments. By combining genome-wide divergence scans with model-based approaches, we detected a suite of 151 genes that show parallel signatures of positive selection associated with alpine colonization, involved in response to cold, high radiation, short season, herbivores, and pathogens. We complemented these parallel candidates with published gene lists from five additional alpine Brassicaceae and tested our hypothesis on a broad scale spanning ~0.02 to 18 My of divergence. Indeed, we found quantitatively variable genomic parallelism whose extent significantly decreased with increasing divergence between the compared lineages. We further modeled parallel evolution over the *Arabidopsis* candidate genes and showed that a decreasing probability of repeated selection on the same standing or introgressed alleles drives the observed pattern of divergence-dependent parallelism. We therefore conclude that genetic divergence between populations, species, and genera, affecting the pool of shared variants, is an important factor in the predictability of genome evolution.

parallelism | evolution | genomics | alpine adaptation | *Arabidopsis*

Evolution is driven by a complex interplay of deterministic and stochastic forces whose relative importance is a matter of debate (1). Being largely a historical process, we have limited ability to experimentally test for the predictability of evolution in its full complexity (i.e., in natural environments) (2). Distinct lineages that independently adapted to similar conditions by similar phenotype (termed “parallel,” considered synonymous to “convergent” here) can provide invaluable insights into the issue (3, 4). An improved understanding of the probability of parallel evolution in nature may inform on constraints on evolutionary change and provide insights relevant for predicting the evolution of pathogens (5–7), pests (8, 9), or species in human-polluted environments (10, 11). Although the past few decades have seen an increasing body of work supporting the parallel emergence of traits by the same genes and even alleles, we know surprisingly little about what makes parallel evolution more likely and, by extension, what factors underlie evolutionary predictability (1, 12).

A wealth of literature describes the probability of “genetic” parallelism, showing why certain genes are involved in parallel adaptation more often than others (13). There is theoretical and empirical evidence for the effect of pleiotropic constraints, availability of beneficial mutations or position in the regulatory network all having an impact on the degree of parallelism at the

level of a single locus (3, 13–18). In contrast, we know little about causes underlying “genomic” parallelism (i.e., what fraction of the genome is reused in adaptation and why). Individual case studies demonstrate large variation in genomic parallelism, ranging from absence of any parallelism (19), similarity in functional pathways but not genes (20, 21), and reuse of a limited number of genes (22–24) to abundant parallelism at both gene and functional levels (25, 26). Yet, there is little consensus about what determines variation in the degree of gene reuse (fraction of genes that repeatedly emerge as selection candidates) across investigated systems (1).

Divergence (the term used here to consistently describe both intra- and interspecific genetic differentiation) between the compared instances of parallelism appears as a potential driver of the variation in gene reuse (14, 27, 28). Phenotype-oriented meta-analyses suggest that both phenotypic convergence (28) and genetic parallelism underlying phenotypic traits (14) decrease with increasing time to the common ancestor. Although a similar targeted multiscale comparison is lacking at the genomic level, our brief review of published studies (29 cases, [Dataset S1](#)) suggests that also gene reuse tends to scale with divergence (Fig. L4 and [SI Appendix, Fig. S1](#)). Moreover, allele reuse (repeated sweep of the same haplotype that is shared among populations either via gene flow or from standing genetic variation) frequently underlies

Significance

Repeated evolution tends to be more predictable. The impressive spectrum of recent reports on genomic parallelism, however, revealed that the fraction of the genome that evolves in parallel varies greatly, possibly reflecting different evolutionary scales investigated. Here, we demonstrate divergence-dependent parallelism using a comprehensive genome-wide dataset comprising 12 cases of parallel alpine adaptation and identify decreasing probability of adaptive re-use of genetic variation as the major underlying cause. This finding empirically demonstrates that evolutionary predictability is scale dependent and suggests that availability of preexisting variation drives parallelism within and among populations and species. Altogether, our results inform the ongoing discussion about the (un)predictability of evolution, relevant for applications in pest control, nature conservation, or the evolution of pathogen resistance.

Author contributions: M.B. and F.K. designed research; M.B. performed research; M.B., B.L., V.K., M.F., and T.S. contributed new reagents/analytic tools; M.B., J.V., S.Y., and V.K. analyzed data; and M.B. and F.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

¹To whom correspondence may be addressed. Email: magdalena.holcova@natur.cuni.cz or filip.kolar@natur.cuni.cz.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2022713118/-DCSupplemental>.

Published May 17, 2021.

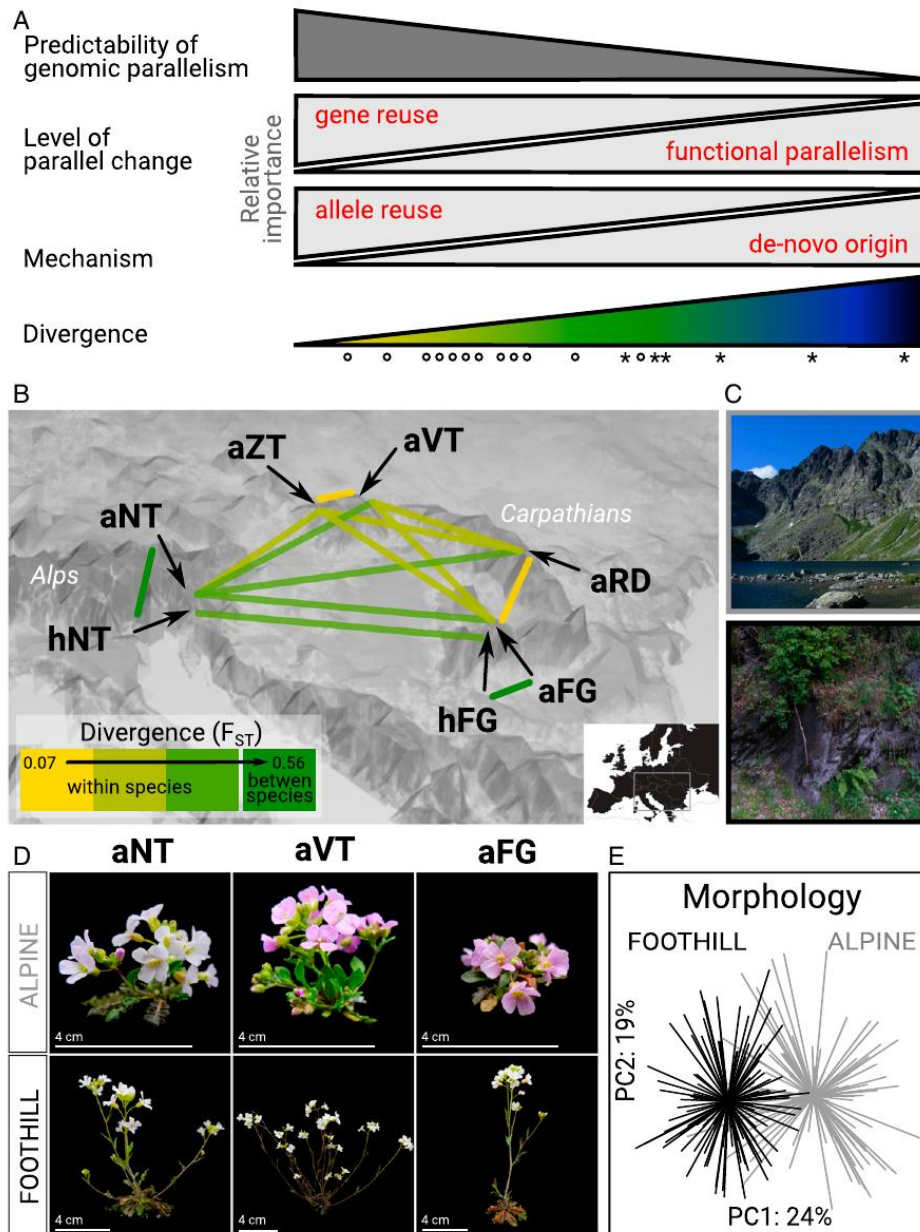


Fig. 1. Hypotheses regarding relationships between genomic parallelism and divergence and the *Arabidopsis* system used to address these hypotheses. (A) Based on our literature review, we propose that genetically closer lineages adapt to a similar challenge more frequently by gene reuse, sampling suitable variants from the shared pool (allele reuse), which makes their adaptive evolution more predictable. Color ramp symbolizes rising divergence between the lineages (~0.02 to 18 Mya in this study); the symbols denote different divergence levels tested here using resequenced genomes of 22 *Arabidopsis* populations (circles) and meta-analysis of candidates in Brassicaceae (asterisks). (B) Spatial arrangement of lineages of varying divergence (neutral F_{ST} ; bins only aid visualization; all tests were performed on a continuous scale) encompassing parallel alpine colonization within the two *Arabidopsis* outcrossers from central Europe: *A. arenosa* (diploid: aVT; autotetraploid: aNT, aZT, aRD, and aFG) and *A. halleri* (diploid: hNT and hFG). Note that only two of the ten between-species pairs (dark green) are shown to aid visibility. The color scale corresponds to the left part of the color ramp used in A. (C) Photos of representative alpine and foothill habitat. (D) Representative phenotypes of originally foothill and alpine populations grown in common garden demonstrating phenotypic convergence. Scale bar corresponds to 4 cm. (E) Morphological differentiation among 223 *A. arenosa* individuals originating from foothill (black) and alpine (gray) populations from four regions after two generations in a common garden. Principal component analysis was run using 16 morphological traits taken from ref. 45.

parallel adaptation between closely related lineages (29–32), while parallelism from independent de novo mutations at the same locus dominates between distantly related taxa (13). Similarly, previous studies reported a decreasing probability of hemiplasy (apparent convergence resulting from gene tree discordance) with divergence in phylogeny-based studies (33, 34). This suggests that the degree of allele reuse may be the primary factor underlying the hypothesized divergence-dependency of parallel genome evolution, possibly reflecting either weak hybridization barriers, widespread ancestral polymorphism between closely related lineages (35), or ecological reasons (lower niche differentiation and geographical proximity) (36, 37). However, the generally restricted focus of individual studies of genomic parallelism on a single level of divergence does not lend itself to a unified comparison across divergence scales. Although different ages of compared lineages affect a variety of evolutionary–ecological processes such as diversification rates, community structure, or niche conservatism (37), the hypothesis that genomic parallelism scales with divergence has not yet been systematically tested, and the underlying evolutionary processes remain poorly understood.

Here, we aimed to test this hypothesis and investigate whether allele reuse is a major factor underlying the relationship. We analyzed replicated instances of adaptation to a challenging alpine environment, spanning a range of divergence from populations to tribes within the plant family Brassicaceae (38–43) (Fig. 1A). First, we took advantage of a unique naturally replicated setup in the plant model genus *Arabidopsis* that was so far neglected from a genomic perspective (Fig. 1B). Two predominantly foothill-dwelling *Arabidopsis* outcrossers (*A. arenosa*, *A. halleri*) exhibit scattered, morphologically distinct alpine occurrences at rocky outcrops above the timberline (Fig. 1C). These alpine forms are separated from the widespread foothill population by a distribution gap spanning at least 500 m of elevation. Previous genetic and phenotypic investigations and follow-up analyses presented here showed that the scattered alpine forms of both species represent independent alpine colonization in each mountain range, followed by parallel phenotypic differentiation (Fig. 1D and E) (44–46). Thus, we sequenced genomes from seven alpine and adjacent foothill population pairs, covering all European lineages encompassing the alpine ecotype. We discovered a suite of 151 genes from multiple functional pathways relevant to alpine stress that were repeatedly differentiated between foothill and alpine populations. This points toward a polygenic, multifactorial basis of parallel alpine adaptation.

We took advantage of this set of well-defined parallel selection candidates and tested whether the degree of gene reuse decreases with increasing divergence between the compared lineages (Fig. 1A). By extending our analysis to five additional alpine Brassicaceae species, we further tested whether there are limits to gene reuse above the species level. Finally, we inquired about possible underlying evolutionary processes by estimating the extent of allele reuse using a designated modeling approach. Overall, our empirical analysis provides a perspective to the ongoing discussion about the variability in the reported magnitude of parallel genome evolution and identifies allele reuse as an important evolutionary process shaping the extent of genomic parallelism between populations, species, and genera.

Results

Parallel Alpine Colonization by Distinct Lineages of *Arabidopsis*. We retrieved whole-genome sequences from 11 alpine and 11 nearby foothill populations (174 individuals in total, seven to eight per population) covering all seven mountain regions with known occurrence of *A. arenosa* or *A. halleri* alpine forms (a set of populations from one mountain region is further referred to as a “lineage”; Fig. 1B and SI Appendix, Fig. S2 and Tables S1 and S2). Within each species, population structure analyses based on genome-wide fourfold degenerate (4d) synonymous single nucleotide polymorphisms

(SNPs) demonstrated clear grouping according to lineage but not alpine environment, suggesting parallel alpine colonization of each mountain region by a distinct genetic lineage (SI Appendix, Figs. S3 and S4). This was in line with separation histories between diploid populations of *A. halleri* estimated in Relate (SI Appendix, Fig. S5) and previous coalescent simulations on broader population sampling of *A. arenosa* (45). The only exception was the two spatially closest lineages of *A. arenosa* (aVT and aZT) for which alpine populations clustered together, keeping the corresponding foothill populations paraphyletic. Due to considerable pre- (spatial segregation) and postzygotic (ploidy difference) barriers between the alpine populations from these two lineages (47), we left aZT and aVT as separate units in the following analyses for the sake of clarity (exclusion of this pair of lineages did not lead to qualitatively different results; SI Appendix, Text S1).

We observed a gradient of neutral differentiation among the seven lineages, quantified as average pairwise 4d- F_{ST} between foothill populations from each lineage, ranging from 0.07 to 0.56 (SI Appendix, Table S3). To control for potential effects of linked selection on our divergence estimates, we also calculated F_{ST} differentiation using noncoding sites that are distant from selectively constrained sites (Materials and Methods). These F_{ST} values strongly correlated with 4d- F_{ST} (Pearson’s $r = 0.93$, P value < 0.001). Further, 4d- F_{ST} values correlated with absolute neutral divergence (4d- D_{XY} , Pearson’s $r = 0.89$, $P < 0.0001$), and we further refer to them consistently as “divergence.” All populations showed high levels of 4d-nucleotide diversity (mean = 0.023, SD = 0.005), as expected for strict outcrossers, and no remarkable deviation from neutrality [the range of 4d-Tajima’s D was -0.16 to 0.6 , well within the neutrality interval -2 to 2 proposed by Tajima (48); SI Appendix, Table S4]. We found no signs of severe demographic change that would be associated with alpine colonization (similar 4d-nucleotide diversity and 4d-Tajima’s D of alpine and foothill populations; Wilcoxon rank test, $P = 0.70$ and 0.92 , respectively; $n = 22$). Coalescent-based demographic inference further supported a no-bottleneck model even for the outlier population with the highest 4d-Tajima’s D value (population LAC of aFG lineage, SI Appendix, Fig. S6).

Genomic Basis of Parallel Alpine Adaptation. Leveraging whole-genome resequencing data of the seven natural replicates, we identified a set of genes showing signatures of parallel directional selection associated with alpine colonization. We used a conservative approach taking the intersection of F_{ST} -based divergence scans designed to control for potential confounding signal of local selection within each ecotype (Materials and Methods) and candidate detection under a Bayesian framework that accounts for neutral processes (BayPass) and identified from 100 to 716 gene candidates in the seven lineages. Of these, we identified 196 gene candidates that were shared between at least two lineages and further tested whether they are consistent with parallel adaptation using neutral simulations in the Distinguishing Modes of Convergence (DMC) maximum composite likelihood framework (49) (Materials and Methods). Out of the 196 shared gene candidates, we identified 151 genes showing significantly higher support for the parallel alpine selection model as compared to a neutral model assuming no selection in DMC (further referred to as “parallel gene candidates”). This set of genes contains an enrichment of differentiated nonsynonymous SNPs (SI Appendix, Table S5), and we did not find any evidence that this was explained by weaker selective constraint compared to the rest of the genome (approximated by ratio of their nonsynonymous-to-synonymous diversity; SI Appendix, Table S6 and Fig. S7 and Text S2). Further, F_{ST} values calculated for the 5% outlier windows do not correlate with recombination rate (Pearson’s $r = 0.037$, $P = 0.67$), and such genes do not tend to cluster in regions of low recombination rates (SI Appendix, Figs. S8 and S9).

Bohutinská et al.

Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives

PNAS | 3 of 10

https://doi.org/10.1073/pnas.2022713118

Functional annotations of the parallel gene candidates using The Arabidopsis Information Resource (TAIR) database and associated publications (Dataset S2), protein-protein interaction database STRING (SI Appendix, Fig. S10), and gene ontology (GO) enrichment analysis (Dataset S3) suggest a complex polygenic basis of alpine adaptation, involving multiple major functional categories, well-matching expectations for a response to a multifactorial environmental stress (SI Appendix, Text S3). Six of the physiological adaptations to alpine environment, encompassing both abiotic and biotic stress, stand out (broadly following ref. 50), both in terms of number of associated parallel candidate genes and functional pathways (Fig. 2). We further discuss these putative alpine adaptations and their functional implications in (SI Appendix, Text S3).

Ubiquitous Gene and Function-Level Parallelism and Their Relationship with Divergence. Using the set of parallel gene candidates identified in *Arabidopsis* lineages, we quantified the degree of parallelism at the level of genes and gene functions (biological processes). We overlapped the seven lineage-specific candidate gene lists across all 21 pairwise combinations of the lineages and identified significant parallelism (nonrandom number of overlapping genes, $P < 0.05$, Fisher's exact test, Fig. 3A) among 15 (71%) lineage pairs (SI Appendix, Table S7). Notably, the overlaps were significant for

10 out of 11 pairwise comparisons among the lineages within a species but only in five out of 10 pairwise comparisons across species (Dataset S4). We then annotated the functions of gene candidates using "biological process" GO terms in each lineage, extracted only significantly enriched functions, and again overlapped them across the seven lineages. Of these, we found significant overlaps ($P < 0.05$, Fisher's exact test) among 17 (81%) lineage pairs, and the degree of overlap was similar within and across species (82 and 80%, respectively, Fig. 3B and SI Appendix, Table S7 and Dataset S5).

Then, we quantified the degree of parallelism for each pair of *Arabidopsis* lineages as the proportion of overlapping gene and function candidates out of all candidates identified for these two lineages. The degree of parallelism was significantly higher at the function level (mean proportion of parallel genes and functions across all pairwise comparisons = 0.045 and 0.063, respectively, $D = 437.14$, degrees of freedom [df] = 1, $P < 0.0001$, generalized linear model [GLM] with binomial errors). Importantly, the degree of parallelism at the gene level (i.e., gene reuse) significantly decreased with increasing divergence between the lineages (negative relationship between Jaccard's similarity in candidate gene identity among pairs of lineages and 4d-Fst; Mantel $rM = -0.71$, $P = 0.001$, 999 permutations, Fig. 3C). In contrast, the degree of parallelism by function did not correlate with divergence ($rM = 0.06$, $P = 0.6$, 999 permutations, $n = 21$, Fig. 3D).

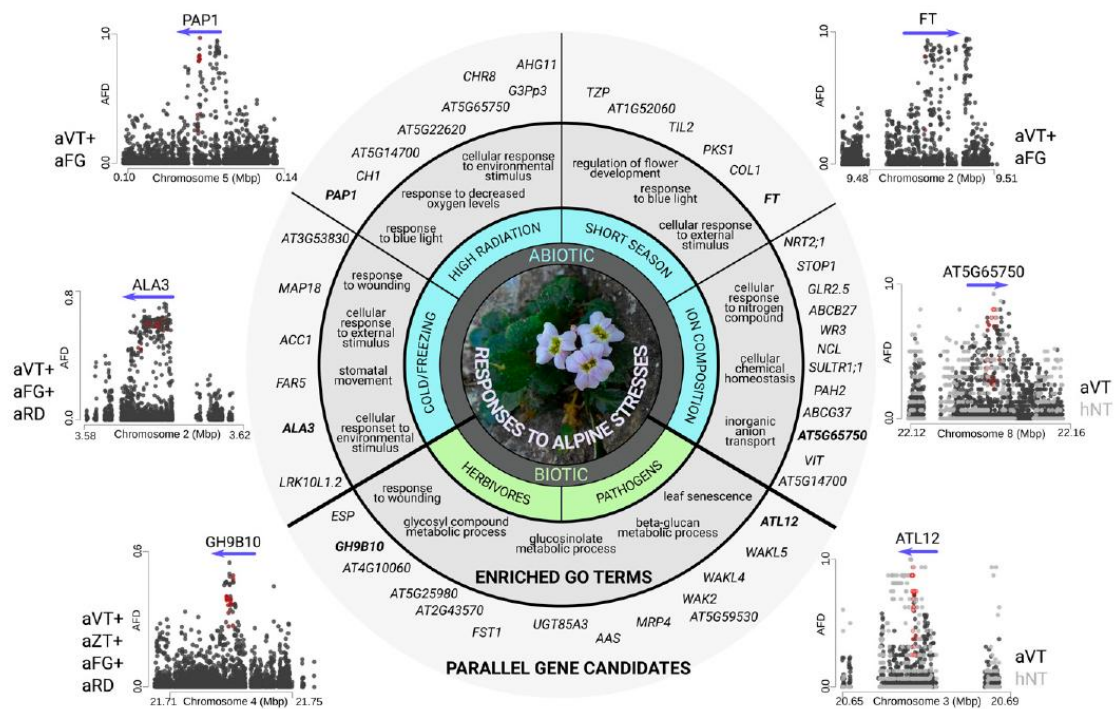


Fig. 2. Physiological responses to alpine stresses in *A. arenosa* and *A. halleri*, identified based on functional annotation of parallel gene candidates (circle) and signatures of parallel directional selection at the corresponding loci (surrounding dotplots). The circle scheme is based on the annotated list of 151 parallel gene candidates (Dataset S2) and corresponding enriched GO terms within the biological process category (Dataset S3). For purposes of functional interpretation and visualization, we also classified the enriched GO terms in the context of major alpine stressors following ref. 50, and list a subset of corresponding 47 well-annotated parallel gene candidates in the outer circle. For the complete list of all genes, refer to Dataset S2, and for more details on functional interpretations, refer to SI Appendix, Text S3. Dotplots show allele frequency difference (AFD) at SNPs between foothill and alpine populations summed over all lineages showing a parallel differentiation in a given gene (blue arrow). The lineage names are listed on the sides. Loci with two independently differentiated haplotypes likely representing de novo mutations (AT5G65750 and ATL12) are represented by peaks of black and gray dots, corresponding with the two parallel lineages. Red circles highlight nonsynonymous variants.

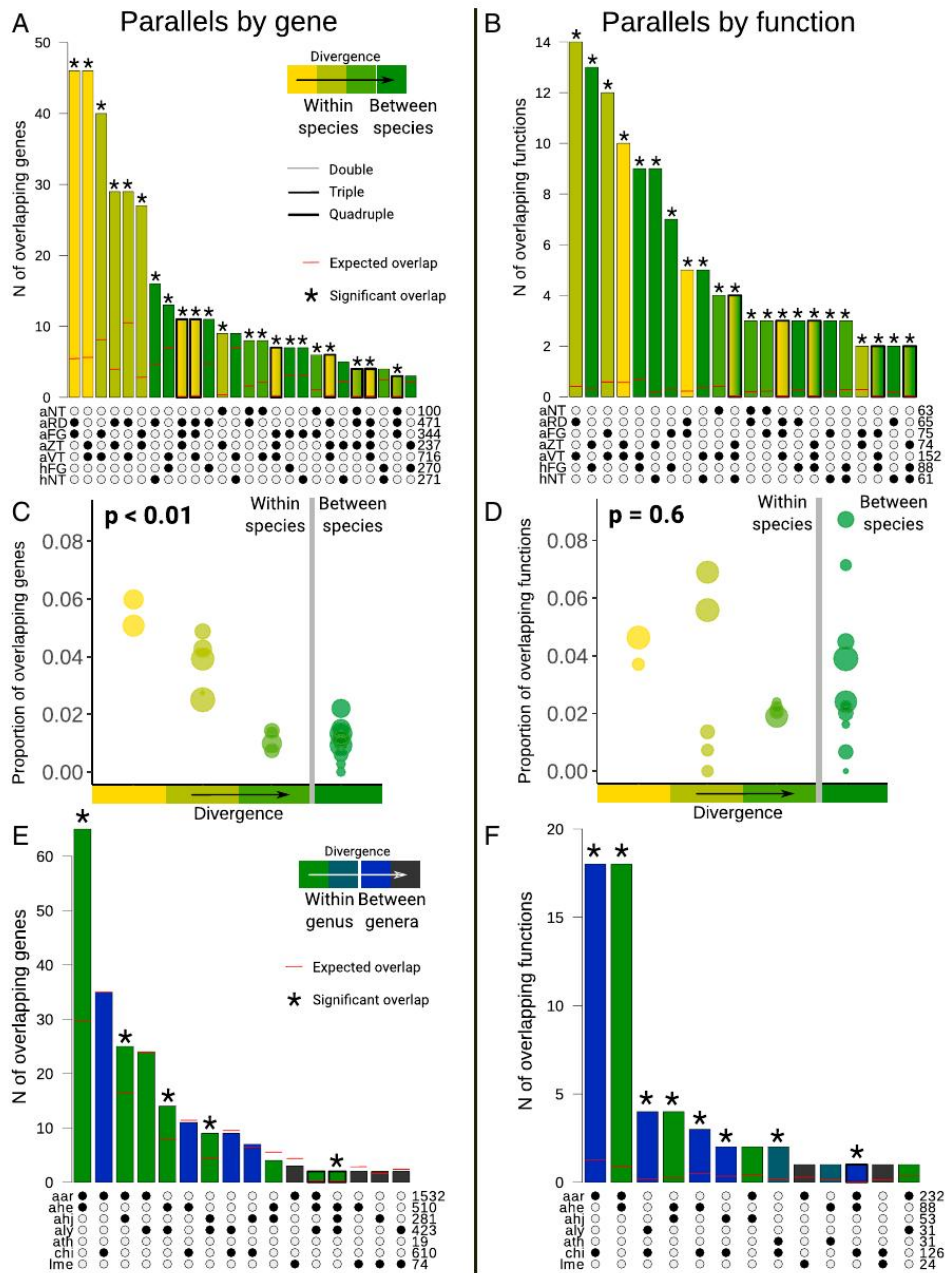


Fig. 3. Variation in gene and function-level parallelism and their relationship with divergence in *A. arenosa* and *A. halleri* (A–D) and across species from Brassicaceae family (E and F). (A and B) Number of overlapping candidate genes (A) and functions (B; enriched GO terms) for alpine adaptation colored by increasing divergence between the compared lineages. Only overlaps of >2 genes and >1 function are shown (for a complete overview, refer to [Datasets S4–S7](#)). Numbers in the bottom-right corner of each panel show the total number of candidates in each lineage. Categories indicated by an asterisk exhibited higher than random overlap of the candidates ($P < 0.05$, Fisher's exact test). For lineage codes, see Fig. 1B. Categories with overlap over more than two lineages are framed in bold and filled by a gradient. (C and D) Proportions of parallel genes (C; gene reuse) and functions (D) among all candidates identified within each pair of lineages (dot) binned into categories of increasing divergence (bins correspond to Fig. 1B and only aid visualization; size of the dot corresponds to the number of parallel items). Significance of the association was inferred by Mantel test over continuous divergence scale. (E and F) Same as A and B but for species from Brassicaceae family, spanning higher divergence levels. Codes: aar: our data on *A. arenosa*; ahe: our data on *A. halleri* combined with *A. halleri* candidates from Swiss Alps (39); ahj: *Arabidopsis halleri* subsp. *gemmifera* from Japan (38); aly: *A. lyrata* from Northern Europe (40); ath: *A. thaliana* from Alps (43); chi: *Crulichimalaya himalaica* (42); and lme: *Lepidium meyenii* (41).

Bohutinská et al.
Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives

PNAS | 5 of 10
<https://doi.org/10.1073/pnas.2022131118>

We further tested whether the relationship between the degree of parallelism and divergence persists at deeper phylogenetic scales by complementing our data with candidate gene lists from six genome-wide studies of alpine adaptation from the Brassicaceae family (38–43) [involving five species diverging 0.5 to 18 Mya (51, 52), *SI Appendix, Supplementary Methods and Tables S8 and S9*]. While we still found significant parallelism both at the level of candidate genes and functions (Fig. 3 *E* and *F* and *Datasets S6 and S7*), their relationship with divergence was nonsignificant (Mantel $r_M = -0.52/-0.22$, for genes/functions respectively, $P = 0.08/0.23$, 999 permutations, $n = 21$). However, the degree of gene reuse was significantly higher for comparisons within a genus (*Arabidopsis*) than between genera ($D = 15.37$, $df = 1$, $P < 0.001$, GLM with binomial errors) while such a trend was absent for parallel function candidates ($D = 0.38$, $df = 1$, $P = 0.54$), suggesting that there are limits to gene reuse at above-genus-level divergences. Taken together, these results suggest that there are likely similar functions associated with alpine adaptation among different lineages, species, and even genera from distinct tribes of Brassicaceae. However, the probability of reusing the same genes within these functions decreases with increasing divergence among the lineages, thus reducing the chance to identify parallel genome evolution.

Probability of Allele Reuse Underlies the Divergence Dependency of Gene Reuse. Repeated evolution of the same gene in different lineages could either reflect repeated recruitment of the same allele from a shared pool of variants (“allele reuse”) or adaptation via alleles representing independent mutations in each lineage (“de novo origin”) (49). To ask whether varying prevalence of these two evolutionary processes could explain the observed divergence-dependency of gene reuse, we quantified the contribution of allele reuse versus de novo origin to the gene reuse in each pair of *A. arenosa* and *A. halleri* lineages and tested whether it scales with divergence.

For each of the 151 parallel gene candidates, we inferred the most likely source of its candidate variant(s) by using a designated likelihood-based modeling approach that investigates patterns of shared hitchhiking from allele frequency covariance at positions surrounding the selected site [DMC method (49)]. We contrasted three models of gene reuse, involving 1) selected allele acquired via gene flow, 2) sourced from ancestral standing variation (both 1 and 2 representing allele reuse), and 3) de novo origin of the selected allele. In line with our expectations, the degree of allele reuse decreased with divergence ($D = 34.28$, $df = 16$, $P < 0.001$, GLM with binomial errors; Fig. 4A). In contrast, the proportion of variants sampled from standing variation remained relatively high even at the deepest interspecific comparison (43%; Fig. 4A and *SI Appendix, Fig. S11*). The absolute number of de novo-originated variants was low across all divergence levels investigated (*Dataset S8*). This corresponds to predictions about a substantial amount of shared variation between related species with high genetic diversity (35) and frequent adaptive transspecific polymorphism in *Arabidopsis* (10, 53–55). Absence of interspecific parallelism sourced from gene flow was in line with the lack of genome-wide signal of recent migration between *A. arenosa* and *A. halleri* inferred by coalescent simulations (*SI Appendix, Fig. S12*).

Importantly, allele reuse covered a dominant fraction of the variation in gene reuse that was explained by divergence (Fig. 4B), suggesting allele reuse is the major factor contributing to the observed divergence-dependency of gene reuse. We also observed a strong correlation between divergence and the maximum composite likelihood estimate of the amount of time the allele was standing in the populations between their divergence and the onset of selection (Pearson’s $r = 0.83$, $P < 0.0001$, Fig. 4C). This suggests that the onset of selection pressure (assuming a similar selection strength) likely happened at a similar time point in the past. Altogether, the parallel gene candidates (Fig. 4 *D–F*) in the

two *Arabidopsis* species likely experienced selection at comparable time scales in all lineages, but the degree of reuse of the same alleles decreased with increasing divergence between parallel lineages, which explained most of the divergence-dependency of gene reuse.

Discussion

By analyzing genome-wide variation over 12 instances of alpine adaptation across Brassicaceae, we found that the degree of gene reuse decreased with increasing divergence between compared lineages. This relationship was largely explained by the decreasing role of allele reuse in a subset of seven thoroughly investigated pairs of *Arabidopsis* lineages. These findings provide empirical support for earlier predictions on genetic parallelism (14, 28) and present a general mechanism that may help explain the tremendous variability in the extent of parallel genome evolution that was recorded across different case studies (1, 13). The decreasing role of allele reuse with divergence agrees with theoretical and empirical findings that the evolutionary potential of a population depends on the availability of preexisting (standing or introgressed) genetic variation (56–58) and that the extent of ancestral polymorphism and gene flow decreases with increasing differentiation between gradually diverging lineages (35, 59). In contrast, the overall low contribution of de novo-originated parallel alleles and generally large and variable outcrossing *Arabidopsis* populations suggest a minor role of mutation limitation, at least within our genomic *Arabidopsis* dataset. In general, our study demonstrates the importance of a quantitative understanding of divergence for the assessment of evolutionary predictability (60) and brings support to the emerging view of the ubiquitous influence of divergence scale on different evolutionary and ecological mechanisms (37).

There are potentially additional, nonexclusive explanations for the observed divergence-dependency of gene reuse, although presumably of much lower impact given the large explanatory power of allele reuse in our system. First, theory predicts that the degree of conservation of gene networks, their functions, and developmental constraints decrease with increasing divergence (14, 28). Diversification of gene networks, however, typically increases at higher divergence scales than addressed here [millions of years of independent evolution (28)] and affects parallelism caused by independent de novo mutations (18). We also did not find any evidence that our gene reuse candidates were under weaker selective constraint than other genic loci genome-wide. Nevertheless, we cannot exclude that changes in constraint contribute to the decreasing probability of gene reuse across Brassicaceae, as was also reported in ref. 61. Second, protein evolution studies reported patterns of diminishing amino acid convergence over time due to the decreasing probability of hemiplasy (i.e., the gene tree discordance caused by incomplete lineage sorting and introgression) (33, 34, 62). As such a pattern reflects neutral processes and is expected to decrease with time, it can confound the assessment of the level of adaptive convergence (33, 34). However, we accounted for this bias in our sampling and analysis design by considering only genes identified as selection candidates in separate divergence scans that contrasted derived alpine populations by their control foothill counterparts. Third, as genetic divergence often corresponds to the spatial arrangement of lineages (63), external challenges posed by the alpine environment at remote locations may differ. Such risk is, however, mitigated at least in our *Arabidopsis* dataset, as the genomically investigated alpine populations share very similar niches (45).

In contrast, no relationship between the probability of gene reuse and divergence was shown in experimental evolution of different populations of yeast (64), raising a question about the generality of our findings. Our study addresses a complex selective agent [a multihazard alpine environment (50)] in order to provide insights into an ecologically realistic scenario relevant

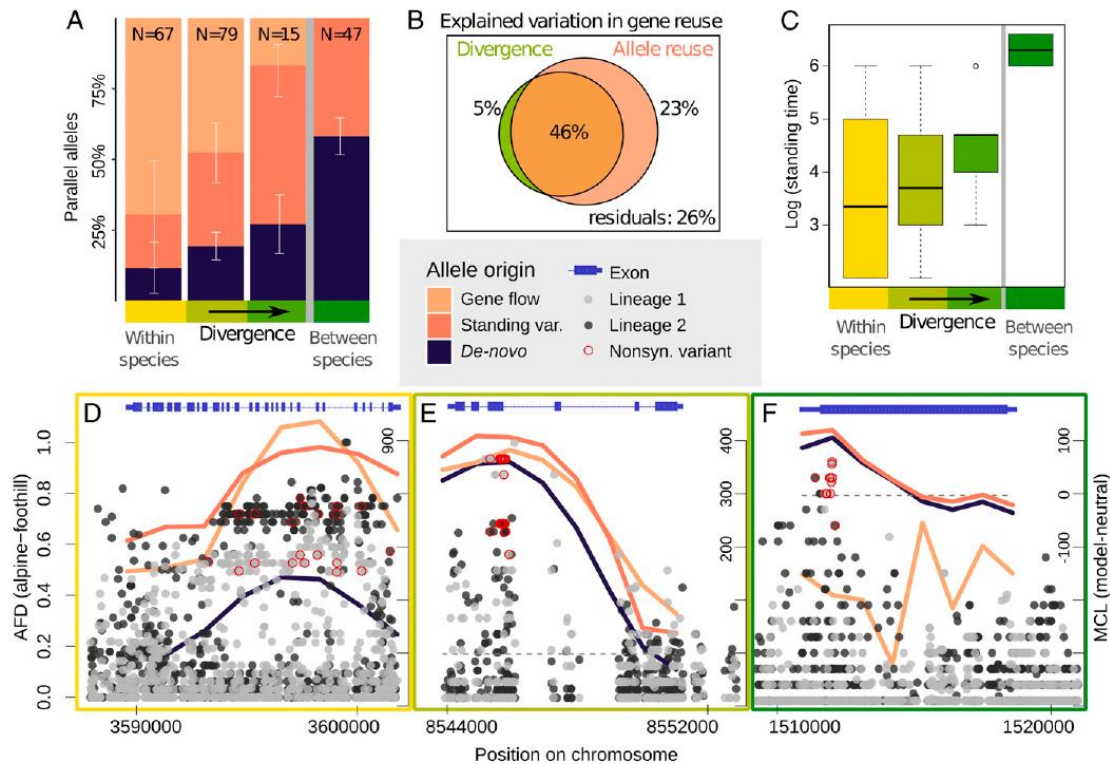


Fig. 4. Decreasing probability of allele reuse with increasing divergence in *A. arenosa* and *A. halleri*. (A) Proportion of parallel candidate gene variants shared via gene flow between alpine populations from different lineages or recruited from ancestral standing variation (together describing the probability of allele reuse) and originated by independent de novo mutations within the same gene. Percentages represent mean proportions for lineages of a particular divergence category (color ramp; total number of parallel gene candidates is given within each plot). (B) Explained variation in gene reuse partitioned by divergence (green circle), allele reuse (orange circle), and shared components (overlaps between them). (C) Maximum composite log-likelihood estimate (MCLE) of median time (generations) for which the allele was standing in the populations prior to the onset of selection. (D–F) Examples of SNP variation and MCL estimation of the evolutionary scenario describing the origin of parallel candidate allele. Two lineages in light and dark gray are compared in each plot. Shown is the entire region of each parallel candidate gene. (D) Parallel selection on variation shared via gene flow on gene *ALA3*, affecting vegetative growth and acclimation to temperature stresses (87). (E) Parallel recruitment of shared ancestral standing variation at gene *AL730950*, encoding heat shock protein. (F) Parallel selection on independent de novo mutations at gene *PKS1*, regulating phytochrome B signaling (88); here, de novo origin was prioritized over standing variation model based on very high MCLE of standing time (*Materials and Methods*). Note that each sweep includes multiple highly differentiated nonsynonymous SNPs (in C and D at the same positions in both population pairs, in line with reuse of the same allele). Dotplot (left y-axis): AFD between foothill and alpine population from each of the two lineages (range 0 to 1 in all plots). Lines (right y-axis): MCL difference from a neutral model assuming no parallel selection (all values above dotted gray line show the difference, higher values indicate higher support for the nonneutral model, and the final model selection is based on the genomic position with the highest likelihood within the gene).

for adaptation in natural environments. Results might differ in systems with a high degree of self-fertilization or recent bottlenecks, as these might decrease the probability of gene reuse even among closely related lineages by reducing the pool of shared standing variation (65, 66). Although this is not the case in our *Arabidopsis* outcrossers, encompassing highly variable and demographically stable populations, drift might have contributed to the low number of overlaps in comparisons involving the less-variable selfer *Arabidopsis thaliana* (43) in our meta-analysis (Fig. 3E). However, considering the supporting evidence from the literature (Fig. 14 and *SI Appendix*, Fig. S1) and keeping the aforementioned restrictions in mind, we predict that our findings are widely applicable. In summary, our study demonstrates divergence-dependency of parallel genome evolution between different populations, species, and genera and identifies allele reuse as the underlying mechanism. This indicates that the availability of

genomic variation preexisting in the species may be essential for (repeated) local adaptation and consequently also for the predictability of evolution, a topic critical for pest and disease control as well as for evolutionary theory.

Materials and Methods

Sampling. *A. arenosa* and *A. halleri* are biennial to perennial outcrossers closely related to the model *A. thaliana*. Both species occur primarily in low to mid elevations (to ~1,000 m above sea level) across Central and Eastern Europe, but scattered occurrences of morphologically distinct populations have been recorded from treeless alpine zones (>1,600 m) in several distinct mountain regions in Central–Eastern Europe (44, 67) that were exhaustively sampled by us (Fig. 1, details provided in *SI Appendix*, *Supplementary Methods*).

Here, we sampled and resequenced genomes of foothill (growing in elevations 460 to 980 m a.s.l.) as well as adjacent alpine (1,625 to 2,270 m a.s.l.) populations from all known foothill–alpine contrasts. In total, we sequenced genomes of 111 individuals of both species and complemented them with 63

published whole-genome sequences of *A. arenosa* (68) totaling 174 individuals and 22 populations (SI Appendix, Table S1). Ploidy of each sequenced individual was checked using flow cytometry following (69).

Sequencing, Raw Data Processing, Variant Calling, and Filtration. Samples were sequenced on Illumina HiSeq X Ten, mapped to reference genome *A. lyrata* (70), and processed following ref. 68. Details are provided in SI Appendix, Supplementary Methods.

Population Genetic Structure. We calculated genome-wide 4d within- [nucleotide diversity (π) and Tajima's D (48)] and between- [F_{ST} (70)] population metrics using python3 ScanTools_ProtEvol pipeline (https://github.com/mbohutinska/ScanTools_ProtEvol) (71). ScanTools_ProtEvol is a customized version of ScanTools, a toolset specifically designed to analyze diversity and differentiation of diploid and autotetraploid populations using SNP data (68). To overcome biases caused by unequal population sizes and to preserve the most sites with no missing data, we randomly subsampled genotypes at each position to six individuals per population.

We quantified divergence between pairs of lineages as average pairwise 4d- F_{ST} between the foothill populations as they likely represent the ancestral state within a given lineage. To control for potential effects of linked selection on our divergence estimates, we also extracted all putatively neutral sites that are unlinked from the selected sites (i.e., sites >5 kb outside genic and conserved regions and sites >1 Mb away from the centromere). As both F_{ST} estimates strongly correlated (Pearson's $r = 0.93$, P value < 0.001), we used only 4d- F_{ST} in further analyses of population structure.

Next, we inferred relationships between populations using allele frequency covariance graphs implemented in TreeMix v. 1.13 (72). We ran TreeMix allowing a range of migration events and presented two and one additional migration edges for *A. arenosa* and *A. halleri*, as they represented points of log-likelihood saturation (SI Appendix, Fig. S4). To obtain confidence in the reconstructed topology, we bootstrapped the scenario with zero events (the tree topology had not changed when considering the migration events), choosing a bootstrap block size of 1,000 bp, equivalent to the window size in our selection scan, and 100 replicates. Finally, we displayed genetic relatedness among individuals using principal component analysis as implemented in adegenet (73).

We further investigated particular hypotheses regarding the demographic history of our system using coalescent simulations implemented in fastsimcoal2 (74). We calculated joint allele frequency spectra (AFS) of selected sets of populations from genome-wide 4d-SNPs and compared their fit to the AFS simulated under different demographic scenarios using the Poisson random field model likelihood. We used wide range of initial parameters (effective population size, divergence times, migration rates; see attached est file, Dataset S10).

Population structure inference was based on a complete dataset of all populations as all the above used methods allow for a combined analysis of diploid and autotetraploid data (further explained in SI Appendix, Supplementary Methods).

Genome-Wide Scans for Directional Selection. To infer SNP candidates, we worked with the full set of SNPs which passed variant filtration (SI Appendix, Table S2). We used a combination of two different divergence scan approaches, both of which are based on population allele frequencies and allow analysis of diploid and autopolyploid populations.

First, we calculated pairwise window-based F_{ST} between foothill and alpine population pairs within each lineage and used minimum sum of ranks to find the candidates. For each population pair, we calculated F_{ST} (75) for 1 kb windows along the genome. Based on the average genome-wide decay of genotypic correlations (150 to 800 bp, SI Appendix, Fig. S13 and Supplementary Methods), we designed windows for the selection scans to be 1 kb (i.e., at least 200 bp larger than the estimated average linkage disequilibrium [LD]). All calculations were performed using ScanTools_ProtEvol and custom R scripts (<https://github.com/mbohutinska/ProtEvol/>). Our F_{ST} -based detection of outlier windows was not largely biased toward regions with low recombination rate [as estimated based on the available *A. lyrata* recombination map (40) and also from our diploid population genomic data; SI Appendix, Figs. S8 and S9]. This corresponds well with outcrossing and high nucleotide diversity that aids divergence outlier detection in our species (76).

Whenever two foothill and two alpine populations were available within one lineage (i.e., aFG, aNT, aVT and aZT populations of *A. arenosa*), we designed the selection scan to account for changes which were not consistent between the foothill and alpine populations (i.e., rather reflected local changes within one environment). Details are provided in SI Appendix, Supplementary Methods. Finally, we identified SNPs which were 5% outliers

for foothill-alpine allele frequency differences in the above-identified outlier windows and considered them SNP candidates of selection associated with the elevational difference in the lineage.

Second, we used a Bayesian model-based approach to detect significantly differentiated SNPs within each lineage, while accounting for local population structure as implemented in BayPass (SI Appendix, Supplementary Methods) (77).

Finally, we overlapped SNP candidate lists from F_{ST} and BayPass analysis within each lineage and considered only SNPs which were outliers in both methods as directional selection candidates. We annotated each SNP candidate and assigned it to a gene using SnpEff 4.3 (78) following *A. lyrata* version 2 genome annotation (79). We considered all variants in 5' untranslated regions (UTRs), start codons, exons, introns, stop codons, and 3' UTRs as genic variants. We further considered as gene candidates only genes containing more than five SNP candidates to minimize the chance of identifying random allele frequency fluctuation in few sites rather than selective sweeps within a gene.

For both selection scans, we used relatively relaxed 95% quantile threshold as we aimed to reduce the chance of getting false negatives (i.e., undetected loci affected by selection) whose extent would be later magnified in overlaps across multiple lineages. At the same time, we controlled for false positives by accepting only gene candidates fulfilling criteria of the two complementary selection scans. Using a more stringent threshold of 1% did not lead to qualitatively different results in regard to the relationship between parallelism and divergence (SI Appendix, Text S4).

GO Enrichment Analysis. To infer functions significantly associated with foothill-alpine divergence, we performed gene ontology enrichment of gene candidates in the R package topGO (80), using *A. thaliana* orthologs of *A. lyrata* genes obtained using biomaRt (81). We used the conservative "elim" method, which tests for enrichment of terms from the bottom of the GO hierarchy to the top and discards any genes that are significantly enriched in descendant GO terms while accounting for the total number of genes annotated in the GO term (80). We used "biological process" ontology and accepted only significant GO terms with more than five and less than 500 genes as very broad categories do not inform about the specific functions of selected genes (false discovery rate [FDR] = 0.05, Fisher's exact test). Reanalysis with "molecular function" ontology led to qualitatively similar results (SI Appendix, Fig. S14).

Quantifying Parallelism. At each level (gene candidates, enriched GO categories), we considered parallel candidates all items that overlapped across at least one pair of lineages. To test for a higher-than-random number of overlapping items per each set of lineages (pair, triplet, etc.), we used Fisher's exact test [SuperExactTest (82) package in R]. Next, we calculated the probability of gene-level parallelism (i.e., gene reuse) and functional parallelism between two lineages as the number of parallel candidate items divided by the total number of candidate items between them (i.e., the union of candidate lists from both lineages). We note that the identification of parallel candidates between two alpine lineages does not necessarily correspond to adaptation to alpine environments as it could also reflect an adaptation to some other trigger or to foothill conditions. However, our sampling and selection scans, including multiple replicates of alpine populations originating from their foothill counterparts, were designed in order to make such an alternative scenario highly unlikely.

Model-Based Inference of the Probability of Allele Reuse. For all parallel gene candidates, we identified whether they indeed support the parallel selection model and the most likely source of their potentially adaptive variant(s). We used the newly developed composite likelihood-based method DMC (49) which uses patterns of hitchhiking at sites linked to a selected locus to distinguish among the neutral model and three different models of parallel selection (considering different sources of parallel variation): 1) on the variation introduced via gene flow, 2) on ancestral standing genetic variation, and 3) on independent de novo mutations in the same gene (at the same or distinct positions). In lineages having four populations sequenced (aVT, aZT, aFG, and aNT), we subsampled to one (best-covered) foothill and one alpine population to avoid combining haplotypes from subdivided populations.

We estimated maximum composite log-likelihoods (MCLs) for each selection model and a wide range of the parameters (SI Appendix, Table S10). We placed proposed selected sites (one of the parameters) at eight locations at equal distance apart along each gene candidate sequence. We analyzed all variants within 25 kb of the gene (both upstream and downstream) to capture the decay of genetic diversity to neutrality with genetic distance from the selected site. We used $N_e = 800,000$ inferred from *A. thaliana* genome-wide mutation rate (83) and nucleotide diversity in our sequence

data (SI Appendix, Table S4) and a recombination rate of 3.7×10^{-8} determined from the closely related *A. lyrata* (40). To determine whether the signal of parallel selection originated from adaptation to the foothill rather than alpine environment, we ran the method assuming that parallel selection acted on 1) two alpine populations or 2) two foothill populations. For the model of parallelism from gene flow, we allowed either of the alpine populations to be the source of admixture.

For each pair of lineages and each gene candidate, we identified the model which best explained our data as the one with the highest positive difference between its MCL and that of the neutral model at the position within each gene with the highest likelihood.

We further simulated data under the neutral model to find out which difference in MCLs between the parallel selection and neutral model is significantly higher than expected under neutrality. For details, reference SI Appendix, Supplementary Methods.

The R code to run the DMC method over a set of parallel population pairs and multiple gene candidates is available at <https://github.com/mbohutinska/DMCloop>.

Statistical Analysis. As a metric of neutral divergence between the lineages within and between the two sequenced species (*A. arenosa* and *A. halleri*), we used pairwise 4d-F_{ST} values calculated between foothill populations. These values correlated with absolute differentiation (D_{X_Y}, Pearson's $r = 0.89$, $P < 0.001$) and geographic separation within species (rM = 0.86 for *A. arenosa*, $P = 0.002$, Fig. 1B) and thus reasonably approximate between-lineage divergence.

To test for a significant relationship between the probability of parallelism and divergence at each level, we calculated the correlation between Jaccard's similarity in the identity of gene/function candidates in each pair of lineages and 1) the 4d-F_{ST} distance matrix (*Arabidopsis* dataset) or 2) the time of species divergence (Brassicaceae meta-analysis). For each pair of lineages, the Jaccard's similarity was calculated as the ratio of intersection in their candidate gene/function lists over their union. Jaccard's similarities calculated for all 21 possible lineage pairs resulted in a similarity matrix which was then correlated with the corresponding matrix of interlineage divergence using Mantel test with 999 replications [ade4 (83) package in R]. We also performed similar test for candidates found in three lineages instead of two and found congruent results showing significant divergence-dependence for gene reuse (Pearson's $r = -0.71$, $P < 0.001$) but not for parallelism by

function (Pearson's $r = 0.04$, $P = 0.82$, $n = 35$; taking average 4d-F_{ST} over the three lineage pairs as a divergence measure).

Then, we tested whether the relative proportion of the two different evolutionary mechanisms of parallel variation (allele reuse versus de novo origin) relate to divergence using GLMs [R package stats (84)] with a binomial distribution of residual variation. We used the 4d-F_{ST} as a predictor variable and counts of the parallel candidate genes assigned to either mechanism as the explanatory variable. Finally, we used multiple regression on distance matrices [R package ecodist (85)] and calculated the fraction of variation in gene reuse that was explained by similarity in allele reuse, divergence, and by their shared component using the original matrices of Jaccard's similarity in gene and allele identity, respectively, following ref. 86.

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 study codes SRP156117 and SRP233571 (see Dataset S9 for individual codes).

ACKNOWLEDGMENTS. This manuscript greatly benefited from constructive feedback of Graham Coop, Michael Nowak, Antonin Machač, Anja Westram, Pádraic Flood, Kristin Lee, Timothy Sackton, Martin Weiser, and Clément Lafon-Placette. We further thank Daniel Bohutínský, Frederick Rooks, Jakub Hojka, Eliška Závěská, and Peter Schönschetter for help with field collections; Gabriela Šrámková, Lenka Flašková, and Aurélie Désamore for help with laboratory work; and Doubravka Pozárová for help with figure editing. This work was supported by the Czech Science Foundation (Project 17-20357Y to F.K.), a student grant of the Charles University Grant Agency (284119 to M.B.), and long-term research development project 67985939 of the Czech Academy of Sciences. This work was also supported by the Science for Life Laboratory, Swedish Biodiversity Program. The Swedish Biodiversity Program has been made available by support from the Knut and Alice Wallenberg foundation. M.F. was supported by a grant from the Swedish Research Council (grant 621-2013-4320 to T.S.). B.L. was supported by a grant from SciLife-Lab. Sequencing was performed by the Norwegian Sequencing Centre, University of Oslo and the SNP&SEQ Technology Platform in Uppsala. The latter facility is part of the National Genomics Infrastructure Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, under the program Projects of Large Research, Development, and Innovations Infrastructures.

- Z. D. Blount, R. E. Lenski, J. B. Losos, Contingency and determinism in evolution: Replaying life's tape. *Science* **362**, eaam5979 (2018).
- S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (Norton, 1989).
- A. A. Agrawal, Toward a predictive framework for convergent evolution: Integrating natural history, genetic mechanisms, and consequences for the diversity of Life. *Am. Nat.* **190** (S1), S1–S12 (2017).
- D. L. Stern, V. Orgogozo, Is genetic evolution predictable? *Science* **323**, 746–751 (2009).
- M. R. Farhat *et al.*, Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **45**, 1183–1189 (2013).
- R. L. Marvig, L. M. Sommer, S. Molin, H. K. Johansen, Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat. Genet.* **47**, 57–64 (2015).
- A. C. Palmer, R. Kishony, Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nat. Rev. Genet.* **14**, 243–248 (2013).
- F. D. Rinkevich, Y. Du, K. Dong, Diversity and convergence of sodium channel mutations involved in resistance to pyrethroids. *Pestic. Biochem. Physiol.* **106**, 93–100 (2013).
- B. E. Tabashnik, T. Brévault, Y. Carrière, Insect resistance to Bt crops: Lessons from the first billion acres. *Nat. Biotechnol.* **31**, 510–521 (2013).
- V. Preite *et al.*, Convergent evolution in *Arabidopsis halleri* and *Arabidopsis arenosa* on calamine metalliferous soils. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180243 (2019).
- N. M. Reid *et al.*, The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* **354**, 1305–1308 (2016).
- S. Lamichhaney *et al.*, Integrating natural history collections and comparative genomics to study the genetic architecture of convergent evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180248 (2019).
- A. Martin, V. Orgogozo, The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013).
- G. L. Conte, M. E. Arnegard, C. L. Peichel, D. Schluter, The probability of genetic parallelism and convergence in natural populations. *Proc. Biol. Sci.* **279**, 5039–5047 (2012).
- N. Gompel, B. Prud'homme, The causes of repeated genetic evolution. *Dev. Biol.* **332**, 36–47 (2009).
- A. Kopp, Metamodels and phylogenetic replication: A systematic approach to the evolution of developmental pathways. *Evolution* **63**, 2771–2789 (2009).
- D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
- S. Yeaman, A. C. Gerstein, K. A. Hodgins, M. C. Whitlock, Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genet.* **14**, e1007717 (2018).
- Z. Zou, J. Zhang, No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
- S. Birkeland *et al.*, Multiple genetic trajectories to extreme abiotic stress adaptation in Arctic Brassicaceae. *Mol. Biol. Evol.* **37**, 2052–2068 (2020).
- K. L. Cooper *et al.*, Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature* **511**, 41–45 (2014).
- A. D. Foote *et al.*, Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
- S. Takuno *et al.*, Independent molecular basis of convergent highland adaptation in maize. *Genetics* **200**, 1297–1312 (2015).
- M. Bohutínská *et al.*, Novelty and convergence in adaptation to whole genome duplication. *Mol. Biol. Evol.* **10.1093/molbev/msab096** (2021).
- M. C. W. Lim, C. C. Witt, C. H. Graham, L. M. Dávalos, Parallel molecular evolution in pathways, genes, and sites in high-elevation hummingbirds revealed by comparative transcriptomics. *Genome Biol. Evol.* **11**, 1552–1572 (2019).
- M. Manceau, V. S. Domingues, C. R. Linnen, E. B. Rosenblum, H. E. Hoekstra, Convergence in pigmentation at multiple levels: Mutations, genes and function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 2439–2450 (2010).
- H. E. Morales *et al.*, Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Sci. Adv.* **5**, eaav9963 (2019).
- T. J. Ord, T. C. Summers, Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evol. Biol.* **15**, 137 (2015).
- J. M. Alves *et al.*, Parallel adaptation of rabbit populations to myxoma virus. *Science* **363**, 1319–1326 (2019).
- Q. Haenel, M. Roesti, D. Moser, A. D. C. MacColl, D. Berner, Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evol. Lett.* **3**, 28–42 (2019).
- Y.-T. Lai *et al.*, Standing genetic variation as the predominant source for adaptation of a songbird. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2152–2157 (2019).
- E. M. Oziolor, *et al.*, Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science* **364**, 455–457 (2019).
- R. A. Goldstein, S. T. Pollard, S. D. Shah, D. D. Pollack, Nonadaptive amino acid convergence rates decrease over time. *Mol. Biol. Evol.* **32**, 1373–1381 (2015).
- F. K. Mendes, Y. Hahn, M. W. Hahn, Gene tree discordance can generate patterns of diminishing convergence over time. *Mol. Biol. Evol.* **33**, 3299–3307 (2016).

Bohutínská *et al.*
Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives

PNAS | 9 of 10
<https://doi.org/10.1073/pnas.2022713118>

35. R. R. Hudson, J. A. Coyne, Mathematical consequences of the genealogical species concept. *Evolution* **56**, 1557–1565 (2002).
36. G. S. Bradburd, P. L. Ralph, Spatial population genetics: It's about time. *Annu. Rev. Ecol. Syst.* **50**, 427–449 (2019).
37. C. H. Graham, D. Storch, A. Machac, Phylogenetic scale in ecology and evolution. *Glob. Ecol. Biogeogr.* **27**, 175–187 (2018).
38. S. Kubota *et al.*, A genome scan for genes underlying microgeographic-scale local adaptation in a wild *Arabidopsis* species. *PLoS Genet.* **11**, e1005361 (2015).
39. C. Rellstab *et al.*, Local adaptation (mostly) remains local: Reassessing environmental associations of climate-related candidate SNPs in *Arabidopsis halleri*. *Heredity* **118**, 193–201 (2017).
40. T. Hämälä, O. Savolainen, Genomic patterns of local adaptation under gene flow in *Arabidopsis lyrata*. *Mol. Biol. Evol.* **32**, 2557–2571 (2019).
41. J. Zhang *et al.*, Genome of plant maca (*Lepidium meyenii*) illuminates genomic basis for high-altitude adaptation in the central Andes. *Mol. Plant* **9**, 1066–1077 (2016).
42. T. Zhang *et al.*, Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7137–7146 (2019).
43. T. Günther, C. Lampe, I. Barilar, K. J. Schmid, Genomic and phenotypic differentiation of *Arabidopsis thaliana* along altitudinal gradients in the North Italian Alps. *Mol. Ecol.* **25**, 3574–3592 (2016).
44. G. Šrámková-Fuxová *et al.*, Range-wide genetic structure of *Arabidopsis halleri* (Brassicaceae): Glacial persistence in multiple refugia and origin of the Northern Hemisphere disjunction. *Bot. J. Linn. Soc.* **185**, 321–342 (2017).
45. A. Knotek *et al.*, Parallel alpine differentiation in *Arabidopsis arenosa*. *Front Plant Sci* **11**, 561526 (2020).
46. G. Wos, M. Bohutinská, J. Nosková, T. Mandáková, F. Kolár, Parallelism in gene expression between foothill and alpine ecotypes in *Arabidopsis arenosa*. *Plant J.* **105**, 1211–1224 (2021).
47. G. Wos *et al.*, Role of ploidy in colonization of alpine habitats in natural populations of *Arabidopsis arenosa*. *Ann. Bot.* **124**, 255–268 (2019).
48. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
49. K. M. Lee, G. Coop, Distinguishing among modes of convergent adaptation using population genomic data. *Genetics* **207**, 1591–1619 (2017).
50. C. Körner, *Alpine Plant Life* (Springer Berlin Heidelberg, 2003).
51. N. Hohmann, E. M. Wolf, M. A. Lysak, M. A. Koch, A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**, 2770–2784 (2015).
52. P. Y. Novikova *et al.*, Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).
53. B. J. Arnold *et al.*, Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8320–8325 (2016).
54. A. Guggisberg *et al.*, The genomic basis of adaptation to calcareous and siliceous soils in *Arabidopsis lyrata*. *Mol. Ecol.* **27**, 5088–5103 (2018).
55. S. Marburger *et al.*, Interspecific introgression mediates adaptation to whole genome duplication. *Nat. Commun.* **10**, 5218 (2019).
56. R. D. H. Barrett, D. Schluter, Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
57. P. L. Ralph, G. Coop, The role of standing variation in geographic convergent adaptation. *Am. Nat.* **186** (suppl. 1), S5–S23 (2015).
58. K. A. Thompson, M. M. Osmond, D. Schluter, Parallel genetic evolution and speciation from standing variation. *Evol. Lett.* **3**, 129–141 (2019).
59. B. Charlesworth, D. Charlesworth, N. H. Barton, The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Ecol. Syst.* **34**, 99–125 (2003).
60. P. K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).
61. C. Rellstab *et al.*, Genomic signatures of convergent adaptation to Alpine environments in three Brassicaceae species. *Mol. Ecol.* **29**, 4350–4365 (2020).
62. Z. Zou, J. Zhang, Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol. Biol. Evol.* **32**, 2085–2096 (2015).
63. S. Ramachandran *et al.*, Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15942–15947 (2005).
64. A. Spor *et al.*, Phenotypic and genotypic convergences are influenced by historical contingency and environment in yeast. *Evolution* **68**, 772–790 (2014).
65. S. Liu, A.-L. Ferchud, P. Grønkrjaer, R. Nygaard, M. M. Hansen, Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Mol. Ecol.* **27**, 4725–4743 (2018).
66. T. Vogwill, R. L. Phillips, D. R. Gifford, R. C. MacLean, Divergent evolution peaks under intermediate population bottlenecks during bacterial experimental evolution. *Proc. Biol. Sci.* **283**, 20160749 (2016).
67. F. Kolár *et al.*, Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model *Arabidopsis arenosa*. *Mol. Ecol.* **25**, 3929–3949 (2016).
68. P. Monahan *et al.*, Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nat. Ecol. Evol.* **3**, 457–468 (2019).
69. F. Kolár *et al.*, Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the *Arabidopsis arenosa* group (Brassicaceae). *Biol. J. Linn. Soc. Lond.* **119**, 673–688 (2016).
70. R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
71. M. Bohutinská *et al.*, De-novo mutation and rapid protein (co-)evolution during meiotic adaptation in *Arabidopsis arenosa*. *Mol. Biol. Evol.*, 10.1093/molbev/msab001 (2021).
72. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
73. T. Jombart, adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
74. L. Excoffier, M. Foll, fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).
75. B. S. Weir, C. C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
76. L. Yant, K. Bomblies, Genomic studies of adaptive evolution in outcrossing *Arabidopsis* species. *Curr. Opin. Plant Biol.* **36**, 9–14 (2017).
77. M. Gautier, Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555–1579 (2015).
78. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
79. V. Rawat *et al.*, Improving the annotation of *Arabidopsis lyrata* using RNA-seq data. *PLoS One* **10**, e0137391 (2015).
80. A. Alexa, J. Rahnenführer, Gene set enrichment analysis with topGO, 10.18129/B9.bioc.topGO (2018). Accessed 8 November 2018.
81. S. Durinck, P. T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
82. M. Wang, Y. Zhao, B. Zhang, Efficient test and visualization of multi-set intersections. *Sci. Rep.* **5**, 16923 (2015).
83. S. Ossowski *et al.*, The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
84. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2019). <https://www.R-project.org/>. Accessed 28 October 2020.
85. S. C. Goslee, D. L. Urban, The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* **22**, 1–19 (2007).
86. J. W. Lichstein, Multiple regression on distance matrices: A multivariate spatial analysis tool. *Plant Ecol.* **188**, 117–131 (2007).
87. S. C. McDowell, R. L. López-Marqués, L. R. Poulsen, M. G. Palmgren, J. F. Harper, Loss of the *Arabidopsis thaliana* P4-ATPase ALA3 reduces adaptability to temperature stresses and impairs vegetative, pollen, and ovule development. *PLoS One* **8**, e62577 (2013).
88. C. Kami *et al.*, Nuclear phytochrome A signaling promotes phototropism in *Arabidopsis*. *Plant Cell* **24**, 566–576 (2012).

Case study 2.

Parallelism in gene expression between foothill and alpine ecotypes
in *Arabidopsis arenosa*



Parallelism in gene expression between foothill and alpine ecotypes in *Arabidopsis arenosa*

Guillaume Wos^{1,*} , Magdalena Bohutínská^{1,2}, Jana Nosková¹, Terezie Mandáková³ and Filip Kolář^{1,2}

¹Department of Botany, Charles University, Prague 128 01, Czech Republic,

²Institute of Botany, The Czech Academy of Sciences, Průhonice 252 43, Czech Republic, and

³Central European Institute of Technology and Faculty of Science, Masaryk University, Brno 625 00, Czech Republic

Received 15 June 2020; revised 13 November 2020; accepted 26 November 2020; published online 1 December 2020.

*For correspondence (e-mail wosg@natur.cuni.cz)

SUMMARY

Parallel adaptation results from the independent evolution of similar traits between closely related lineages and allows us to test to what extent evolution is repeatable. Similar gene expression changes are often detected but the identity of genes shaped by parallel selection and the causes of expression parallelism remain largely unknown. By comparing genomes and transcriptomes of four distinct foothill–alpine population pairs across four treatments, we addressed the genetic underpinnings, plasticity and functional consequences of gene expression parallelism in alpine adaptation. Seeds of eight populations of *Arabidopsis arenosa* were raised under four treatments that differed in temperature and irradiance, factors varying strongly with elevation. Parallelism in differential gene expression between the foothill and alpine ecotypes was quantified by RNA-seq in leaves of young plants. By manipulating temperature and irradiance, we also tested for parallelism in plasticity (i.e., gene–environment interaction, GEI). In spite of global non-parallel patterns transcriptome wide, we found significant parallelism in gene expression at the level of individual loci with an over-representation of genes involved in biotic stress response. In addition, we demonstrated significant parallelism in GEI, indicating a shared differential response of the originally foothill versus alpine populations to environmental variation across mountain regions. A fraction of genes showing expression parallelism also encompassed parallel outliers for genomic differentiation, with greater enrichment of such variants in *cis*-regulatory elements in some mountain regions. In summary, our results suggest frequent evolutionary repeatability in gene expression changes associated with the colonization of a challenging environment that combines constitutive expression differences and plastic interaction with the surrounding environment.

Keywords: parallel evolution, gene expression, alpine adaptation, *Arabidopsis arenosa*, gene–environment interaction, plasticity, common garden experiment.

INTRODUCTION

Parallel evolution of similar phenotypes in repeated environments provides strong evidence for the role of natural selection and, when linked with genomic investigations, reveal genes or pathways particularly important in adaptation under certain changing conditions (Elmer and Meyer, 2011). Parallel evolution usually reflects the independent evolution of similar phenotypes in response to similar selection pressures (Bolnick *et al.*, 2018), and is a widespread phenomenon reported in animals (Elmer *et al.*, 2014; Velotta *et al.*, 2017), plants (Woodhouse and Hufford, 2019) and bacteria (Fong *et al.*, 2005; Lenski, 2017). Studies on parallel evolution have often found a common genetic basis and similar genetic changes, especially in cases of

low divergence between lineages (Stern and Orgogozo, 2009; Conte *et al.*, 2012), suggesting that some genes are more likely to be subject to selection during parallel evolution. Despite the ubiquity of parallel evolution at both the phenotypic and the genetic levels (Martin and Orgogozo, 2013; Blount *et al.*, 2018), further work is needed to understand the factors underlying such parallelism.

To gain insight on the mechanisms governing the predictability of evolution, it is important to understand the genomic underpinnings of parallel evolution (Stern and Orgogozo, 2009). In this regard, variation in gene expression often plays a key role in the evolution of functional traits (Fraser, 2011), and similar gene expression changes, even in a limited number of genes, can be the source of

important phenotypic parallelism (Cooper *et al.*, 2003; Macqueen *et al.*, 2011; Zhao *et al.*, 2015; Jacobs *et al.*, 2020). Therefore, gene expression analyses have great potential to detect genes involved in parallel adaptation, because such genes should exhibit constitutive expression changes in the same direction across multiple populations independently adapted to similar environments, reflecting the role of natural selection. In addition, the transcriptome of a particular individual also reflects its actual surrounding environment. In that sense, phenotypic plasticity has the ability to promote the emergence of ecotypes in response to the environment. In the case of similar environmental challenges, such genotype–environment interactions (GEIs) may also change in a predictable way over parallel populations, leading to parallelism in plasticity (Pfennig *et al.*, 2010). Parallelism in the plastic response in naturally replicated systems remains largely unknown, however (Zhao *et al.*, 2015). Here, we aimed to study to what extent natural selection shapes gene expression parallelism, and tested whether plasticity, a transient adaptive response to environmental changes, contributes to gene expression parallelism.

Although gene expression analysis may reveal genes underlying parallel evolution, some caveats remain about the genetic underpinnings of such parallelism (Sackton and Clark, 2019). Differences in gene expression generally arise from modifications in gene sequences, and it is still unclear to what extent parallel evolution results from repeated changes in different genes of similar functions, in the same gene but not at the same nucleotide, or at the exact same nucleotide (Elmer and Meyer, 2011). So far, studies on the genetic basis of parallel evolution have highlighted the importance of changes located in *cis*-regulatory elements, compared with other genomic elements (i.e., coding elements) (Stern and Orgogozo, 2009; Witkopp and Kalay, 2012). Indeed, *cis*-regulation elements may be affected by selection differently than coding elements. In general, mutations in *cis*-regulation elements showed less deleterious effects because they mainly affect gene regulation and may be limited to a specific tissue or to a developmental stage (Wray, 2007). In addition, mutations in *cis*-regulatory elements are in theory more likely to cause gene expression parallelism through reduced negative pleiotropy, but empirical evidence for this remains scarce and needs to be investigated (Wray, 2007). Parallelism in gene expression was associated with changes in *cis*-regulatory elements, as was recently documented in fish species (Verta and Jones, 2019; Jacobs *et al.*, 2020).

Parallelism in gene expression at the constitutive or plastic level using whole-transcriptome sequencing has been studied in animals, insects (Reed *et al.*, 2011; Zhao *et al.*, 2015) and fish (Derome and Bernatchez, 2006; McGirr and Martin, 2018), in particular, providing evidence for the effects of environment in triggering similar expression

changes. In turn, such studies in plants remain scarce and are generally focused only on the expression patterns of particular genes or metabolic pathways (Streisfeld and Rausher, 2009; Des Marais and Rausher, 2010; Chen *et al.*, 2014). Moreover, transcriptomic inquiries of parallel adaptation in both plants and animals were generally conducted using only two replicates of parallelism (i.e., two pairs of populations or species from inside and outside the selective environment). This, however, limits our inference on the magnitude and general evolutionary drivers of transcriptomic parallelism. Such questions could be efficiently addressed by leveraging comparisons over multiple independent population pairs, as has been demonstrated using both genomic and phenotypic data (Bolnick *et al.*, 2009; Stuart *et al.*, 2017) but rarely using gene expression analyses (Jacobs *et al.*, 2020).

Here, we tested whether there is significant parallelism in gene expression at both the constitutive and plastic level by screening the transcriptomes of parallel alpine lineages of *Arabidopsis arenosa* (Brassicaceae). An alpine environment constitutes an ideal system for investigating adaptive differentiation because it is associated with sharp environmental gradients and the island-like distribution of mountain regions promotes parallel colonization from lower to higher elevations. Although *A. arenosa* thrives mostly in low to mid elevations (up to approx. 1000 m a.s.l., termed the ‘foothill ecotype’ here), scattered occurrences in treeless alpine habitats (approx. 1500–2500 m a.s.l., termed the ‘alpine ecotype’) were reported from several disjunct mountain regions from Central and Eastern Europe. The two ecotypes are morphologically distinct, and a previous study has demonstrated significant phenotypic parallelism in height and floral traits in the alpine ecotypes across the four regions, which persisted after two generations of common garden cultivation (Knotek *et al.*, 2020). In addition, genetic investigations based on range-wide sampled genome resequencing and RADseq data (Monnahan *et al.*, 2019; Knotek *et al.*, 2020) congruently showed populations clustering by region but not by ecotype, thus supporting the parallel origin of the alpine ecotype.

In this study, we used a subset of the populations used in Knotek *et al.* (2020) to investigate parallelism further at the gene expression level. First, we inferred genetic relationships among the eight populations (four foothill–alpine pairs) using neutral markers to verify the parallel alpine differentiation in our populations. Then we compared gene expression of each population pair under the same treatment (‘constitutive level’) as well as across the four treatments (GEI, as a measure of environmentally induced plasticity), and compared these results across the four regions, asking specifically: (i) does the overall direction and magnitude of the difference in gene expression between foothill and alpine ecotypes vary across the regions; (ii) does differential expression in the same (sets

of) genes underlie parallel alpine differentiation and do the differentially expressed genes belong to the same metabolic pathways; (iii) do we also observe parallelism in the plastic response (GEI), suggesting a shared response to environmental variations. Finally, we investigated the underlying genomic basis of expression parallelism by leveraging genome-wide divergence scans of the same populations, asking specifically: does genetic variation in *cis*-regulatory elements drive expression parallelism at the constitutive and/or plastic levels?

RESULTS

Population structure

We sampled one foothill and one alpine population of *A. arenosa* from the following four mountain regions: Făgăraş Mountain in the Southern Carpathians of Romania (FG); Niedere Tauern in the Austrian Alps (NT); and Vysoké Tatry (VT) and Západoé Tatry (ZT) mountains in the Western Carpathians of Slovakia. Although plants from the VT region are diploid, populations from the remaining three regions represent a closely related autotetraploid cytotype of the same species (Arnold *et al.*, 2015). Using the genome resequencing data available for our populations (with seven or eight individuals per population, approx. 130 000 putatively neutral, fourfold degenerated SNPs, average sequencing depth of 16 \times) we inferred genetic relationships between foothill and alpine ecotypes across the different mountain regions. TREEMIX analysis supported the monophyly (100% bootstrap support) of each of the following three regions: NT (Austrian Alps), FG (Southern Carpathians, Romania) and the cluster of VT + ZT regions (the spatially closest pair of regions from the Western Carpathians in Slovakia) (Figure 1b). For the VT and ZT regions, the two alpine populations clustered together, confirming that these two regions were genetically close, despite the difference in ploidy. Bayesian clustering confirmed the TREEMIX results and demonstrated clear grouping according to the three mountain regions, rather than by ecotype, with no indications of further admixture (Figure 1b). Specifically, under $K = 2$ and $K = 3$ (the highest similarity among the runs), the FG and NT regions separated, respectively, and under $K = 4$ we only observed additional substructure within the FG region and no division between the VT and ZT regions. In summary, the phylogenetic and clustering analyses congruently demonstrated differentiation of the populations belonging to the three regions (NT, FG and VT + ZT), regardless of their foothill versus alpine origin, altogether indicating the independent origin of the alpine ecotype in each region, in line with previous range-wide sampling (Monnahan *et al.*, 2019; Knotek *et al.*, 2020). As alpine areas of the VT and ZT regions are represented by distinct ploidy levels, among which extant gene flow is unlikely (Kolár *et al.*, 2016), for the sake of clarity we considered

each region as a separate unit in our analysis. To account for the fact that VT and ZT have greater genetic similarities than the other pairs of regions, however, we also ran an additional analysis based only on tetraploids (i.e., NT, FG and ZT regions).

Effects of treatment, ecotype and region on gene expression

To identify genes that showed parallel differential expression between foothill and alpine populations (hereafter termed 'ecotype') in each mountain region (hereafter termed 'region'), we characterized population variation in whole-leaf transcriptomes of foothill and alpine ecotypes from four regions grown under four different temperature and irradiance treatments (Figure 1a). We sequenced three replicates for each region \times ecotype \times treatment combination for a total of 96 libraries.

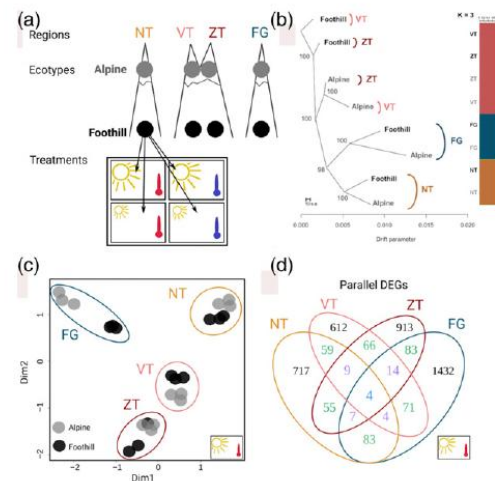


Figure 1. (a) Population structure and the experimental design involving four pairs of foothill (in black) and alpine (in grey) ecotypes from four mountain regions raised in growth chambers under different temperatures and irradiances. (b) Maximum-likelihood TREEMIX graph based on approximately 130 000 genome-wide fourfold degenerated single-nucleotide polymorphisms (SNPs) showing the genetic relationship between eight populations from four mountain regions (numbers at each node indicate bootstrap values) and Bayesian clustering of the same data under $K = 3$. Drift parameter corresponds to $t/2N$, where t is the number of generations separating the two populations, and N is the effective population size. (c) Example of a multidimensional scaling (MDS) plot showing gene expression differences between 24 individuals (coloured by ecotype) raised in one of the four treatments (high temperature and high irradiance), highlighting the strong differentiation in expression between regions and ecotypes. (d) Number of differentially expressed genes (DEGs) between foothill and alpine ecotypes for each region for plants raised in the high temperature and high irradiance treatment. Colours depict the overlaps across two (green), three (purple) and four (blue) regions. Regions: FG, Făgăraş; NT, Niedere Tauern; VT, Vysoké Tatry; ZT, Západoé Tatry.

Treatment, region and ecotype significantly affected the transcriptomic response in *A. arenosa*, showing that parallel alpine colonization consistently resulted in significant changes in gene expression. Treatment explained the highest proportion of variance in gene expression among the 96 samples ($R^2 = 36.2\%$, $F = 31.6$, $P < 0.001$, permutational multivariate analysis of variance, PERMANOVA, test; Figure S1), followed by region and ecotype ($R^2 = 18.9\%$ and 5.20% , $F = 15.5$ and 7.88 , respectively, $P < 0.001$ in both cases, PERMANOVA test; Figure 1c). The overall transcriptomic response of ecotype was consistent across treatments, as indicated by the significant interaction of the treatment \times region terms ($R^2 = 7.48\%$, $F = 1.79$, $P < 0.01$) and the non-significant interaction of treatment \times ecotype terms. When analysing individual parameters separately, the effect of irradiance was significant ($R^2 = 22.70\%$, $F = 27.61$, $P < 0.001$) and explained a greater proportion of the total variance than temperature ($R^2 = 10.25\%$, $F = 10.74$, $P < 0.001$).

Parallelism in gene expression at the constitutive level between foothill and alpine ecotypes

First, we quantified the transcriptome-wide parallelism using a quantitative approach based on vector analysis. Foothill–alpine divergence for each region can be defined by a vector where its length (i.e., magnitude) represents the difference in expression between the foothill and alpine ecotype, and the angle (θ) between the two vectors represents their direction (Figure S2; Table S1). Hence, parallelism between two regions occurred if the angle between two vectors is close to 0° (i.e., point in the same direction) and if they have similar magnitudes.

We found that the values for direction (i.e., the angle between pairs of vectors) ranged from 33.9° to 154.5° and that the values for magnitude (i.e., the difference in the length of pairs of vectors) ranged from 2.57 to 4.40 (log-transformed values), indicating no sign of transcriptome-wide parallelism. Rather, the vectors were ‘acute nonparallel’ or pointed in opposite directions. We tested whether there was a consistent response in the magnitude or direction of the transcriptomic parallelism between pairs of regions across the four treatments using linear models. The magnitude and direction did not significantly differ between pairs of regions (Table S1), as expected from a model of the independent colonization of regions by separate genetic lineages presented previously using genome-wide data (Knotek *et al.*, 2020).

To quantify parallelism on a per-locus basis, which provides a finer estimate of independent adaptation to specific conditions, we identified differentially expressed genes (DEGs; false discovery rate, FDR < 0.05 ; Table S2) between foothill and alpine ecotypes for each region separately and then overlapped them, separately within each of the four treatments (i.e., a ‘common garden’ approach; Figures 1d

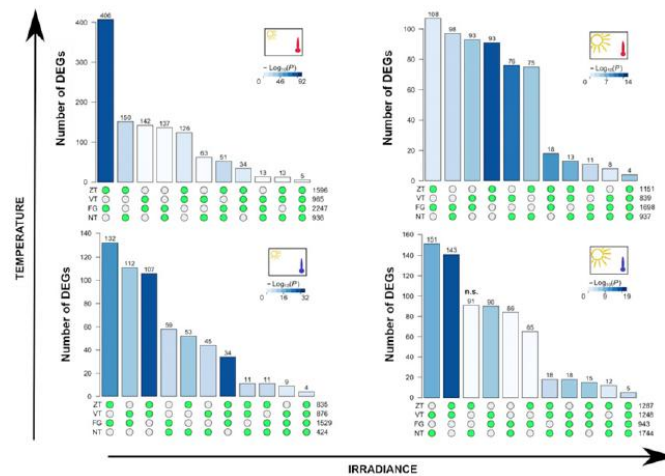
and 2; Table S3). Genes exhibiting parallel changes in expression (parallel DEGs) were those consistently up- or downregulated in alpine ecotypes in at least two regions. We revealed pervasive parallelism at the locus level, as the number of parallel DEGs in each intersection was significantly greater than expected by chance in all but one case; their total number ranged between 390 and 819 across treatments (Figure 2; Table S3). We then ran a Gene Ontology (GO) term enrichment analysis to identify the biological processes in which parallel DEGs were involved, based on all genes showing parallel changes across the four treatments ($n = 2179$) to cover a wide range of conditions in which the ecotypic effect could manifest. GO term enrichment revealed 182 GO terms significantly enriched, with FDR < 0.05 , grouped into 17 categories using REVIGO (Table S4). Most of the enriched GO terms fell into two categories: ‘response to other organisms’, related to the biotic stress response, and ‘secondary metabolism’.

We further focused on the strongest candidates for parallelism: i.e., genes showing overlap across all four regions in at least one treatment (10 genes in total; Table S3). Four of these genes had a known function in *A. thaliana*: three genes were related to the defence response (*BETA-AMYLASE 5*, *FARNESIC ACID CARBOXYL-O-METHYLTRANSFERASE*, LURP-one family) and were all downregulated in alpine ecotypes; one gene was related to protein degradation (*F-BOX/LRR-REPEAT PROTEIN 25-RELATED*) and was upregulated in alpine ecotypes. In the same way, DEGs that overlapped across three regions in at least one treatment ($n = 181$ in total) were also mainly related to defence pathways and response to biotic stress (29 enriched GO terms, 21 of which were related to biotic stress).

To account for the effect of ploidy and to ensure that the significant parallel transcriptomic response was not driven by the closely related VT–ZT pair, we also ran the same analysis using only overlaps between the three regions occupied by tetraploid populations (NT, FG and ZT). GO term enrichment analysis on the total number of genes that overlapped across two and three regions across the four treatments ($n = 1290$) revealed 155 enriched GO terms grouped into 15 categories using REVIGO (Table S5). As previously, most of the enriched GO terms belonged to ‘response to other organisms’, the other GO terms belonged to various metabolic processes (i.e., ‘indole-containing compound metabolism’ or ‘sulfur compound metabolism’).

We further tested whether parallelism may occur not only at the gene level but also at the level of the metabolic pathway (Kyoto Encyclopedia of Genes and Genomes, KEGG). For each treatment and region, we ran a KEGG enrichment analysis on the genes differentially expressed between the foothill and alpine ecotypes and overlapped the enriched KEGG pathways across the four mountain

Figure 2. Number of differentially expressed genes (DEGs) between foothill and alpine ecotypes in each treatment and quantification of constitutive parallelism by overlapping these gene lists over multiple mountain regions (FG, Fägäras; NT, Niedere Tauern; VT, Vysoké Tatry; ZT, Západné Tatry). The significant intersection in DEG lists across mountain regions (i.e., significant parallelism) assessed by Fischer's exact test has been detected in all cases but one (indicated by n.s.; detailed results of the tests are presented in Table S3).



regions (Table S6). We found significant overlap in KEGG pathways across two, three and four regions, indicating significant parallelism at the metabolic pathway level (Table S6). Parallel pathways belonged mainly to secondary metabolism (i.e., anthocyanin or flavonoid biosynthesis), sugar and amino acid metabolism, and photosynthesis and stress response (i.e., glutathione metabolism, cyano-amino acid metabolism or plant-pathogen interaction). The overlaps in metabolic pathways also remained significant after excluding the diploid VT region (Table S6).

Genomic underpinnings of differential parallel expression

Given the importance of mutations in *cis*-regulatory elements to parallel expression changes, we investigated whether genes showing parallelism in differential expression for each region exhibited significant enrichment of highly differentiated single-nucleotide polymorphisms (SNPs; outliers for genome-wide F_{ST} divergence between alpine versus foothill ecotypes in that region among genes containing at least three outlier SNPs), both in their coding sequences and in their *cis*-regulatory elements (Table S7). *Cis*-regulatory regions were defined as sequences containing introns, 5' untranslated regions (5'-UTRs), 3'-UTRs, and a 5-kb segment up- and downstream of a gene, and coding sequences were defined as exons. We found that parallel DEGs were significantly enriched ($P < 0.05$, hypergeometric test) for outlier SNPs in their *cis*-regulatory elements (as compared with the background of all divergence outliers) in two regions: VT and ZT. In turn, we did not find significant enrichment for outlier SNPs in coding elements in any region. We then performed pairwise comparisons between the regions (six pairs in total) to test whether parallel DEGs common to

two regions are more likely to harbour outlier SNPs at the exact same position (i.e., parallel outlier SNPs) as compared with the background of all genes harbouring parallel outlier SNPs (Table 1). For each pair of regions we identified DEGs with shared underlying outlier SNPs, but this number varied greatly, generally corresponding to biogeographic differentiation among the regions (with the lowest shared proportion found for comparisons with the NT region; Table S8). The number of parallel outlier SNPs was significantly enriched in *cis*-regulatory elements only in parallel DEGs of the NT-VT region pair and in coding elements in parallel DEGs of the FG-ZT region pair. By merging all of the parallel DEGs with at least one underlying parallel outlier SNP for the six pairs of regions, GO term enrichment and REVIGO clustering revealed that these genes were mainly related to 'response to other organisms' (biotic stress; Table S8). GO term enrichment remained unchanged when considering only the three tetraploid pairs of NT-FG, NT-ZT and FG-ZT (Table S5).

Parallelism in plastic response of ecotypes

Apart from parallelism at the constitutive level, foothill-alpine pairs across the different mountain regions may also respond in a similar way to an environmental trigger, reflecting parallelism in their plastic response. We thus further tested for transcriptome-wide parallelism in the differential response of alpine and foothill ecotypes to changing environmental conditions simulated by varying both temperature and irradiance. Using vector analysis, we found significant effects of treatment on magnitude ($F = 3.76$, $P = 0.027$) but not on direction (angle) of the vectors over the six regional pairs tested (Table S1). By testing for the specific effects of temperature and irradiance, we found significant effects of irradiance ($F = 12.2$, $P = 0.002$) but not

Table 1 Association between genomic parallelism (outlier single-nucleotide polymorphisms, SNPs, found in both regions) and parallelism in gene expression (parallel differentially expressed genes, DEGs) for each pair of regions

Pairs of regions	Parallel DEGs with parallel SNPs			Background, total no. of parallel outlier SNPs		Fold enrichment and significance	
	No. of genes	No. of parallel outlier SNPs in <i>cis</i> -regulatory elements	No. of parallel outlier SNPs in coding elements	in <i>cis</i> -regulatory elements	in coding elements	<i>cis</i> -regulatory elements	coding elements
NT-FG	7	5	2	519	366	1.22	0.69
NT-VT	5	7	0	992	570	1.57*	0.00
NT-ZT	11	11	4	968	572	1.17	0.72
FG-VT	35	43	31	3052	2348	1.03	0.96
FG-ZT	85	73	88	3062	2416	0.81	1.24**
VT-ZT	90	176	99	11294	7088	1.04	0.93

The table shows the number of genes showing parallelism in differential expression between foothill and alpine ecotypes (within at least one treatment) between each pair of regions containing at least one outlier SNP at the same position, the number of outlier SNPs at the same position in our parallel DEGs located in *cis*-regulatory and coding regions. The background refers to the total number of outlier SNPs located at the same position (in *cis*-regulatory or coding elements) across the entire genome for each pair of regions, the last two columns show the fold enrichment and significance for outlier SNPs in *cis*-regulatory and in coding elements in our parallel DEGs (hypergeometric test). Regions: FG, Făgăraş (Romania); NT, Niedere Tauern (Austria); VT, Vysoké Tatry (Slovakia); ZT, Západné Tatry (Slovakia).

* $P < 0.05$.

** $P < 0.01$.

of temperature on magnitude, with greater magnitudes under low irradiance.

To quantify plasticity in the transcriptomic response on a per-locus basis, we further calculated the GEI between ecotypes and different environmental conditions separately for each region. We then examined parallelism in this measure of plasticity by the degree of overlap of the genes exhibiting significant GEIs (Figure S3; Table 2). We assessed GEI between ecotype and three types of biologically relevant environmental contrasts: (i) conditions characteristic for the foothill and alpine environment (Figure 3; Table S9); (ii) changes of temperature when irradiance was kept constant (under high and low irradiance; Table S10); and (iii) changes of irradiance when temperature was kept constant (under high and low temperature; Table S11). In all environmental contrasts we detected significant overlaps in genes exhibiting GEI across all pairs of mountain ranges as well as in the majority of triple overlaps (Figure S3), altogether demonstrating that our foothill–alpine pairs may respond in the same way to environmental variations.

Then, we examined the potential function of such parallel GEI candidates. For treatments approximating foothill versus alpine conditions, enrichment analysis revealed five enriched GO terms (based on 54 such candidates; Table S9), with four of them related to cell wall modification ('hemicellulose metabolic process', 'cell wall polysaccharide metabolic process', 'xyloglucan metabolic process' and 'cell wall macromolecule metabolic process'). The last GO term enriched was 'response to chemical' and included 11 genes in response to various hormones.

For temperature, we identified 21 (under high irradiance) and 118 (under low irradiance) parallel GEIs (Table S10c).

GO term enrichment analysis on all of the genes affected by changes in temperature (21 + 118 = 139) revealed 56 enriched GO terms grouped into seven categories by REVIGO (Table S10), mainly related to the response of temperature stimulus, flavonoid biosynthesis and secondary metabolism.

Finally, for irradiance, we identified 66 (under low temperature) and 147 (under high temperature) genes with significant ecotype × irradiance interactions in parallel across regions (Table S11). GO term enrichment on all the genes (66 + 147 = 213) revealed 73 associated enriched GO terms. REVIGO grouped them into 10 categories, mainly related to the response to other organisms and to the immune system (Table S11). The results of enrichment analyses were qualitatively similar when we analysed only the overlap of genes showing significant GEIs across the three tetraploid regions. For a comparison of treatment between foothill and alpine conditions, the low number of overlapping genes (33 genes; Table S12) did not lead to significant enrichment, but a classical categorization revealed that most of these genes are classified into 'response to stress' and 'response to chemical' categories. For the two other contrasts (Table S12), GO term enrichment analysis revealed similar enriched categories as with the full design.

In a similar way, we tested whether parallelism in the plastic response may also occur at the level of the metabolic pathway (KEGG). Although in all regions we found enriched KEGG terms related to secondary metabolism (pigment biosynthesis), sugar metabolism and photosynthesis, only a few very general KEGG pathways significantly overlapped across mainly two regions (KEGG enrichment, GEI; Table S13).

Table 2 Plasticity in gene expression estimated as the number of differentially expressed genes (DEGs) showing significant gene–environment interaction (GEI) (FDR < 0.05), and their repeated discovery across the mountain regions demonstrating significant parallelism in expression plasticity. Number of DEGs showing a significant ecotype × alpine environment interaction, ecotype × temperature interaction, under low and high irradiance, and ecotype × irradiance interaction, under low and high temperature, in each of the four mountain regions. The last column shows the number of genes exhibiting GEIs that overlapped across at least two mountain regions (for details, see Figure S3).

Interaction	Mountain Regions					Parallel DEGs
	Niedere Tauern(NT)	Fägäras(FG)	Vysoké Tatro(VT)	Západné Tatro(ZT)		
Ecotype × alpine env.	No. GEI DEGs	191	348	163	103	54
	<i>F</i> (<i>R</i> ²)	4.05*** (21.6%)	4.29*** (20.2%)	3.21*** (19.2%)	4.57*** (20.2%)	
Ecotype × temperature	No. GEI DEGs	132	372	221	394	118
	<i>F</i> (<i>R</i> ²)	3.37*** (20.2%)	6.54*** (24.4%)	3.85*** (20.9%)	5.70*** (25.2%)	
	No. GEI DEGs	109	50	103	138	21
	<i>F</i> (<i>R</i> ²)	4.63*** (22.7%)	3.73** (17.8%)	2.93*** (18.7%)	3.89*** (20.3%)	
Ecotype × irradiance	No. GEI DEGs	122	958	153	240	147
	<i>F</i> (<i>R</i> ²)	4.13*** (21.4%)	5.54*** (20.4%)	3.18*** (17.7%)	4.44*** (15.2%)	
	No. GEI DEGs	207	85	140	356	66
	<i>F</i> (<i>R</i> ²)	4.34*** (17.5%)	3.10*** (16.9%)	3.63*** (17.8%)	4.46*** (24.3%)	

F values for the ecotype × treatment interaction (*df* = 1) for each region tested by PERMANOVA test and percentage of variance explained (*R*²), in parentheses, are listed. Lists of corresponding parallel DEGs are published in Tables S9, S10 and S11. ***P* < 0.01. ****P* < 0.001.

Finally, we investigated whether the parallel genes showing GEIs (all contrasts merged together) exhibit an enrichment of differentiation outlier SNPs in their underlying genomic elements (Table S7). We found significant enrichment for outlier SNPs in *cis*-regulatory elements in NT and VT regions, and marginally significant enrichment in ZT, but an enrichment of SNPs in coding elements in the FG region.

DISCUSSION

Constitutive parallelism in differential gene expression and its genetic underpinnings

Taking advantage from a unique set-up of naturally replicated alpine ecotypes, we investigated intraspecific variation in parallelism in differential gene expression in a wild *Arabidopsis* species. We observed a lack of transcriptome-wide signal of constitutive parallelism in terms of magnitude and direction, reflecting the fact that many genes were specific to the mountain region of origin and that selection operates only on certain functionally relevant genes. Indeed, when looking at the level of individual loci, we found significant overlap in DEGs between foothill and alpine ecotypes across all compared regions, indicating pervasive significant parallelism at the level of individual sets of genes.

The strength of association between parallelism in expression and its genomic underpinnings varied across regions and genomic elements. Specifically, the significant enrichment of outlier SNPs (exhibiting excessive alpine–foothill differentiation, i.e., candidates for directional selection) was detected only in two of the four regions, and exclusively in *cis*-regulatory elements, demonstrating the importance of selection in *cis*-regulatory elements for expression parallelism that may, however, vary across particular instances of parallelism. Pairwise comparisons of regions revealed that candidate parallel outlier SNPs were significantly enriched in *cis*-regulatory elements in one pair of regions (NT–VT), and in coding elements in another

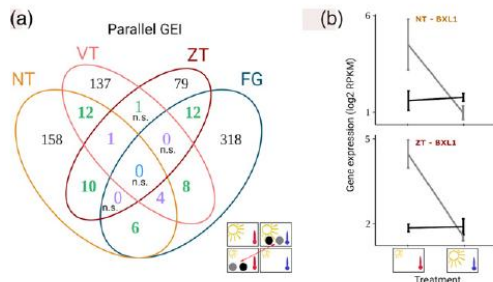


Figure 3. (a) Number of differentially expressed genes showing significant ecotype × alpine environment interaction for each region demonstrating parallelism at the level of transcriptomic plasticity. Colours depict the overlaps across two (green), three (purple) and four (blue) regions. (b) Example of significant gene–environment interaction of the *BXL1* gene (*BETA*-XYLOSIDASE 1, associated with cell wall thickening) for foothill (in black) and alpine (in grey) ecotypes originating from NT and ZT regions raised with high temperature and low irradiance (Ht:Li, approximating low-elevation conditions) and low temperature and high irradiance (Lt:Hi, approximating high-elevation conditions).

regional pair (FG–ZT). Previous studies reported that *cis*-regulatory changes played an important role in gene expression differences between two *Oryza* species (Guo *et al.*, 2016) or that an excess of changes in *cis*-regulatory elements underlined differential expression of some stress-responsive genes between *Arabidopsis* species (He *et al.*, 2016). In light of these results, *cis*-variants appear as potentially important players in shaping parallelism in gene expression, and in ecological adaptation in general, as opposed to variation in coding sequences that had mainly non-significant effects in our analyses. Among our sets of parallel DEGs, however, the overall lack of parallel outlier SNPs suggests that similar expression changes may be achieved by different mutations in *cis*-regulatory elements (Wittkopp and Kalay, 2012) and/or through other processes, such as changes in *trans*-acting regulatory elements (e.g., transcription factors) or in splicing variants, that were not addressable by our study design. Another explanation may involve varying patterns of directional selection between the different mountain regions.

Parallelism at the gene expression level, especially when replicated over more than two environmental transitions, serves as a strong indicator of the role of natural selection in adaptive response to environmental factors (Elmer and Meyer, 2011). Although parallel evolution is not always indicative of adaptive evolution (Losos, 2011), our case put forward the selective scenario as: (i) all alpine populations occupy similar selective environments (Knotek *et al.*, 2020); (ii) significant overlaps in DEGs were found not only between all pairs of regions but also across three or four geographically isolated regions, rendering chance very unlikely; and (iii) based on the high number of different metabolic pathways (182 enriched GO terms) in which parallel DEGs have been involved, genetic correlations or epistatic interactions probably represent a minor contribution to driving such relationships. Indeed, genetic correlations are more likely to occur between genes of the same pathway (Phillips, 2008). Although our study informs on the identity of the parallel candidate loci, the evolutionary source of their alleles remains unclear. Parallel selection may act either on the same allele that emerged in the ancestral non-adapted population, and has been repeatedly swept (standing variation) or introgressed from one adapted population into another (adaptive introgression), or on independent mutations in each region (parallel *de novo* origin; Lee and Coop, 2019). Detailed follow-up studies are necessary for the validation of the adaptive effect of the individual genes as well as for reconstructing the evolutionary source of their variants.

Apart from parallelism in directional selection, similarity in gene expression may reflect historical events such as gene flow and the effect of genetic drift. Although the effect of genetic drift is likely to be minimized by the absence of severe bottlenecks in *A. arenosa* populations

(with high and constant values of synonymous diversity and Tajima's *D* across the species range, disregarding the ecotypes; Monnahan *et al.*, 2019), a previous study has detected the complex reticulated evolution of diploid and tetraploid alpine ecotypes from the Tatra Mountain, VT and ZT (Wos *et al.*, 2019). Such reticulation had not markedly affected our measures of the magnitude of parallelism, however, as the expression similarity among VT and ZT regions in terms of number of parallel DEGs fell well within the range of DEGs that overlapped among all the other regions. Similarly, there was no apparent effect of ploidy on parallelism, as we observed a similar overlap between regions occupied by tetraploids as the overlap observed for the diploid VT region, and excluding the VT region from our analyses did not change our results qualitatively. The VT–ZT pair harboured more outlier SNPs at the same position than any other pairs of regions, however, which may indicate the overall sharing of genetic variation through weak lineage sorting and/or recent gene flow.

Functional implication of gene expression parallelism between the foothill and alpine ecotype

Genes related to biotic stress response, and to fungus and bacteria in particular, showed a strong and consistent signal of parallelism, both at the level of parallel DEGs overall and in the subset of DEGs with underlying parallel outlier SNP variation. Interestingly, we also found significant overlap in KEGG pathways related to biotic stress, especially 'plant–pathogen interaction', describing the defence response to fungus and bacteria, and 'cyano-amino acid metabolism', related to glucosinolate metabolism. Elevation is associated with sharp variations in abiotic and biotic stress factors (Brown *et al.*, 1996; Vetaas, 2002). In general, the importance of abiotic factors increased (Körner, 2003) whereas the importance of biotic factors decreased with elevation (Desprez-Loustau *et al.*, 2010; Rasmann *et al.*, 2014). Accordingly, some studies identified pathways or metabolites related to biotic and abiotic stress associated with an elevational gradient. For instance, antioxidants (glutathione and phenol compounds), photosynthesis and sugar metabolism, and defence proteins (Wildi and Lütz, 1996; Ma *et al.*, 2015) varied strongly with elevation. These results were consistent with our findings, in particular that the overall decrease in pathogen and/or competition pressures with elevation was constitutively manifested at the gene expression and metabolic (KEGG) pathway level.

Despite that the original sites of foothill and alpine ecotypes strongly differ in climatic conditions, we found only a few GO terms directly and consistently associated with abiotic stress across regions among our parallel candidates. Instead, we found some GO terms associated with general processes such as 'metabolism' or 'biological regulation', containing genes involved in responses to

multiple environmental stresses and in developmental aspects from germination to flowering. On one hand, these enriched GO terms may just reflect the major changes that occur along an elevational gradient, such as reduced partial pressure of CO₂ with elevation that affects photosynthesis and the underlying metabolism (Wang *et al.*, 2017). On the other hand, they may also be linked with the observed phenotypic convergence in *A. arenosa* alpine ecotypes in these regions, clearly reflecting abiotic pressures, i.e., reduced height or shifted flowering time (Měsíček and Goliašová, 2002; Knotek *et al.*, 2020). Additionally, we cannot exclude that such major phenotypic differences may result from selection on only a few genes of larger effects that could be missed by enrichment analyses (e.g., *FRIGIDA*, an important determinant of flowering time variation in *Arabidopsis thaliana*; Stinchcombe *et al.*, 2004). Genes involved in developmental processes were lacking among our strongest parallelism candidates (DEGs overlapping across three or four regions), however, ruling out a hypothesis of a single major-effect gene that would stand consistently behind the response to abiotic triggers.

Parallelism in the plastic response of ecotypes

Parallel adaptation may manifest either at the level of constitutive differential expression in a homogeneous environment or in a plastic response of the parallel adaptive candidates to important environmental factors. The plastic response will then manifest as parallelism in the GEI between the adapted ecotype and the manipulated environmental parameters. We detected significant GEIs in response to two prominent abiotic factors affecting plant life at high elevations indicating that foothill and alpine ecotypes responded differently to environmental changes. Importantly, we detected significant parallelism in genes exhibiting such GEIs across most of the regions, although such overlap was considerably lower in absolute terms than parallelism at the constitutive level (i.e., with a lower number of overlapping candidates and with an absence of complete, fourfold, overlaps). Parallel DEGs demonstrating GEIs tended to show greater enrichment for candidate SNPs in *cis*-regulatory elements in all but one region, suggesting a contribution of genetic variation in *cis*-regulatory elements in triggering similar expression changes in response to environmental stress.

In the context of an ecotype × alpine environment interaction, parallel DEGs showing significant GEIs were hormone-related genes and cell wall modification enzymes belonging to the XTH family involved in cell wall strengthening (Cosgrove, 2005). In plants, the cell wall is one of the first mechanical barriers against abiotic and biotic stress, and cell wall thickness was positively correlated with elevation in different species (Kogami *et al.*, 2001; Ma *et al.*, 2015). When environmental variables of temperature and irradiance were analysed separately, DEGs showing

significant GEIs were mainly related to the response to temperature stimulus and plant defence, respectively. The response to temperature stimulus includes many GO terms related to abiotic factors (temperature, oxidative stress and drought stress) known to vary along an elevational gradient (Ma *et al.*, 2015). The over-representation of plant defence genes in response to changes in irradiance may be linked to interactions between the light sensing and plant defence pathways (Karpinski *et al.*, 2003).

Local adaptation is commonly invoked to explain the maintenance of plasticity between populations (Josephs, 2018). In our system, based on the relatively high number of genes showing GEIs, it is likely that foothill and alpine habitats differ in their fitness optima. Our results provided sets of potential candidate genes that are important for local adaptation; however, although changes in abiotic and biotic stress-signalling compounds may be linked to an increased fitness in alpine environments, this requires experimental validation, as does the overall quantification of what extent all the remaining non-parallel genes identified contribute to plant fitness in an alpine environment.

CONCLUSION

Our design, involving a comparison of transcriptomic and genomic profiles across four pairs of foothill and alpine ecotypes, revealed pervasive parallelism at the constitutive level, highlighting the prominent role of natural selection acting on genes and pathways related to biotic stress. On the other hand, we also demonstrated lower, yet still significant levels of parallelism at the plastic level, indicating that populations of different origins may also exhibit shared responses to variation in the same environmental factors. In sum, our study demonstrates that the repeatability of evolution can manifest at various levels of the complex genotype–environmental interface, and that analysis of multiple replicated instances of ecotypic differentiation may help to reveal such processes in natural populations.

EXPERIMENTAL PROCEDURES

Plant material

Arabidopsis arenosa is a perennial out-croser from Europe occurring predominantly at low elevations (from colline to submontane, here termed 'foothill'), but with scattered occurrences at high elevations above the tree line (termed 'alpine'). In Central and South-Eastern Europe, different *A. arenosa* lineages have colonized five mountain regions (hereafter 'regions'), with four regions sampled here (Figure S4): FG, NT, VT and ZT. NT, FG and ZT are occupied by autotetraploid populations and VT is occupied by diploid populations (Kolář *et al.*, 2016). For each region, we collected seeds from 10 maternal plants in one foothill population (600–1000 m a.s.l.) and in one alpine population (1700–2200 m a.s.l.) (hereafter 'ecotype'). In addition, analyses of population genetic structure and coalescent simulations revealed that alpine stands in the NT, FG and ZT + VT regions have been colonized independently from their foothill counterparts, with each foothill–alpine

pair corresponding to a distinct genetic cluster (C. European, S. Carpathian and W. Carpathian, respectively; Knotek *et al.*, 2020; Monnahan *et al.*, 2019). For ZT and VT, each region has been colonized by different ploidy levels; however, the cytotypes are genetically closely related (Wos *et al.*, 2019; see also the Discussion). As alpine areas of the ZT and VT regions have been colonized by distinct ploidy levels, we considered each region as a separate unit in our main analysis for the sake of clarity. To account for the fact that both regions have greater genetic similarities than the other pairs of regions, however, we also ran an additional analysis using only the three tetraploid regions (NT, FG and ZT), excluding the diploid region VT.

Rearing conditions

To reduce potential maternal effects of the original localities, we first raised one generation of plants in growth chambers, using the seeds from the 10 maternal lines collected in the field, under constant conditions (21°C day/18°C night, 16-h day/8-h night, light approx. $300 \mu\text{mol m}^{-2} \text{s}^{-1}$) in pots filled with a mixture of peat and sand (in a ratio of 2:3). For each population, we pollinated 14 flowering plants by a mixture of pollen from the same population to simulate the process of random pollination in nature. Seeds collected from such a first generation (containing full- and half-siblings) were then raised under four experimental treatments that varied in temperature and irradiance. Temperature and irradiance are two environmental parameters associated with elevation that clearly distinguished our foothill and alpine ecotypes (Figure S4). Two levels per factor were combined in a full-factorial design, resulting in four treatments: 'high temperature:high irradiance' (Ht:Hi); 'high temperature:low irradiance' (Ht:Li); 'low temperature:high irradiance' (Lt:Hi); and 'low temperature: low irradiance' (Lt:Li). The two intermediate treatments, Hr:Li and Lt:Hi, were used to mimic the conditions experienced by plants at low and high elevations, respectively. The two extreme treatments, Ht:Hi and Lt:Li, were used to test for an effect of a rise in temperature or irradiance on gene expression.

Seeds were first stratified for 1 week (4°C, constant darkness) and were then germinated under 15 h light/9 h dark, with light intensity $150 \mu\text{mol m}^{-2} \text{s}^{-1}$, at 21°C and a relative humidity of 50%. After 20 days, seedlings were split into four separate growth chambers and exposed to one of the four treatments (seeds from three maternal plants \times two ecotypes \times four regions per treatment; for each population the same maternal lines were used across treatments). For all treatments, the day length was set to 16 h light/8 h dark. Temperature and irradiance were gradually changed to reach 18°C day/13°C night for high temperature treatments or 10°C day/4°C night for low temperature treatments, and from 280 to 980 $\mu\text{mol m}^{-2} \text{s}^{-1}$ during the day in high irradiance treatments or from 50 to 200 $\mu\text{mol m}^{-2} \text{s}^{-1}$ during the day in low irradiance treatments (for details, see Table S14). Temperatures were chosen to reflect average values during the growth period, from April to June (average temperature measured in the Austrian Alps at 600 m a.s.l. = 18°C, at 2000 m a.s.l. = 10°C), and irradiance based on average values reported in a previous study of ecotypic differentiation of alpine plants spanning a similar elevation range (Bertel *et al.*, 2016). Treatments were applied until plants reached the 14-leaf stage so that plant materials for RNA-seq were collected from plants of similar developmental stage.

Sample collection and RNA extraction

Plant materials were collected between 51 and 86 days after plants were transferred in separate growth chambers, depending on the treatment, following different growth rates in different treatments.

For transcriptome analysis we randomly selected one individual per maternal line for each population (eight populations \times three maternal lines \times one individual \times four treatments = 96 plants sequenced in total). For each plant, we collected the seventh rosette leaf at a similar time point for all treatments and leaf samples were immediately snap-frozen in liquid nitrogen. None of the plants had flowered at the time of collection. Total RNA was extracted using the NucleoSpin miRNA kit including a DNase treatment step (Macherey-Nagel, <https://www.mn-net.com>), and the purity and quantity of RNA was assessed with a Nanodrop 2000 spectrophotometer (ThermoFisher Scientific, <https://www.thermofisher.com>) and RNA integrity was assessed with an Agilent 2100 bioanalyzer (Agilent Technologies, <https://www.agilent.com>).

Library preparation, sequencing and data processing

The library was prepared using the Illumina TruSeq Stranded mRNA Kit (RS-122-9004DOC; Illumina, <https://www.illumina.com>). Specific TruSeq adapters were ligated on the cDNA for each individual. Individual sequencing was carried out on Illumina HiSeq 4000 on four lanes (four lanes \times 24 individuals) using 150-bp paired-end reads. After sequencing, raw data were filtered to remove low-quality reads. The data are available from the Sequence Read Archive (<http://www.ebi.ac.uk/ena>) under accession PRJNA575330.

The quality of each individual library was checked using FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Sequencing generated between 8 and 36 million reads per individual. Over-represented sequences, which corresponded to the TruSeq adapter, were trimmed using CUTADAPT (Martin, 2011). As a result of the high similarities between the *A. arenosa* and *Arabidopsis lyrata* genomes, reads for each individual were aligned on the *A. lyrata* reference genome (Hu *et al.*, 2011), as has been successfully applied in previous genomic and transcriptomic studies in *A. arenosa* (Yant *et al.*, 2013; Baduel *et al.*, 2018; Monnahan *et al.*, 2019) with the v2 annotation (Rawat *et al.*, 2015) using HISAT2 2.1.0 (Kim *et al.*, 2019). The number of reads mapped on each gene was counted with FEATURECOUNTS 1.6.3 (Liao *et al.*, 2013) and only the uniquely mapped reads were retained.

Differential gene expression

Read counts were then analysed using EDGER 3.12.0 (Robinson *et al.*, 2010) in R (RStudio Team, 2015). We scaled the library size with the 'calcNormFactors' function, estimated dispersion using 'estimateDisp' and obtained, for each region and treatment, a list of differentially expressed genes between foothill and alpine ecotypes using the 'glmFit' function. *P* values were adjusted for multiple testing with the Benjamini and Hochberg false-discovery rate (FDR) correction. Genes were considered as differentially expressed if $\text{FDR} < 0.05$ and as parallel if they were found to be differentially expressed in the same direction in at least two regions.

We tested for the ecotype \times environment interaction (GEI), as described in Levine *et al.* (2011). We performed three different contrasts using EDGER (Appendix S1): (i) ecotype \times alpine environment interaction, (foothill in Lt:Hi – foothill in Ht:Li) – (alpine in Lt:Hi – alpine in Ht:Li); (ii) ecotype \times temperature interaction, (foothill in Lt:Li – foothill in Ht:Li) – (alpine in Lt:Li – alpine in Ht:Li), and to test for the effects of changes in temperature when irradiance was kept low we ran the same contrast under high irradiance; and (iii) ecotype \times irradiance interaction, (foothill in Lt:Li – foothill in Lt:Hi) – (alpine in Lt:Li – alpine in Lt:Hi), and to test for the effects of changes in irradiance when the temperature was kept low we ran the same contrast under high temperature. GEI-genes were

considered as parallel only if they overlapped across at least two mountain regions.

Enrichment analysis

Gene expression analysis was complemented by a GO term enrichment analysis for biological processes using GO term finder (Boyle *et al.*, 2004) and KEGG enrichment analysis using G:PROFILER (Raudvere *et al.*, 2019). We used the *A. thaliana* annotation, using the definition of orthology provided in the *A. lyrata* v2 annotation (Rawat *et al.*, 2015). GO terms and KEGG terms were considered significantly enriched if the FDR-adjusted *P* value was <0.05. The REVIGO tool was then used to group GO terms based on semantic similarities (Supek *et al.*, 2011). Gene descriptions were obtained from The Arabidopsis Information Resource (TAIR, <https://www.arabidopsis.org>) (Berardini *et al.*, 2015).

Single-nucleotide polymorphism (SNP) calling and candidate SNP detection

In order to investigate genomic underpinnings of our parallel expression candidates, we investigated their association with genomic SNPs exhibiting extreme differentiation between the corresponding foothill–alpine population pair. We complemented our transcriptomes with whole-genome resequencing data from different individuals (sampled in the field) of the same populations used for our experimental plants, partly published previously (Monnahan *et al.*, 2019) and partly as data available at the Sequence Read Archives (project ID PRJNA592307). In total, we retrieved genome-wide SNP variation of eight individuals per population, except for one population (alpine population from NT) with seven individuals (Table S15). Raw sequences were processed and the SNPs were called following the method described by Monnahan *et al.* (2019). Briefly, we used TRIMMOMATIC 0.36 (Bolger *et al.*, 2014) to remove adaptor sequences and low-quality base pairs. Trimmed reads were mapped to the reference genome of *A. lyrata* (Hu *et al.*, 2011) in BWA 0.7.15 (Li and Durbin, 2009), with default settings. Duplicated reads were identified by PICARD 2.8.1 (<https://broadinstitute.github.io/picard>) and discarded, together with reads that showed low mapping quality (<25). We used GATK 3.7 to call and filter reliable variants, following best practices (McKenna *et al.*, 2010). Namely, we used HaplotypeCaller to call variants per individual with respect to its ploidy level and GenotypeGVCFs to aggregate variants for all samples. We selected only biallelic SNPs and removed those that matched the following criteria: Quality by Depth (QD) < 2.0, FisherStrand (FS) > 60.0, RMSMappingQuality (MQ) < 40.0, MappingQualityRankSumTest (MQRS) < -12.5, ReadPosRankSum < -8.0 and StrandOddsRatio (SOR) > 3.0. We further removed variants from sites with average read depths higher than two times the standard deviation and regions with excessive heterozygosity, indicating probable duplicated and paralogous regions, respectively (for further details, see Monnahan *et al.*, 2019). In the final vcf, for each variant we discarded genotypes with read depths lower than 8× and with more than 20% of the genotype missing.

For each region separately, we identified SNPs that were genome-wide differentiation candidates as 5% outliers from genome-wide distribution of F_{ST} (hereafter termed ‘outlier SNPs’). We annotated each outlier SNP and assigned it to a gene, regulatory or intergenic variant, using SNPEFF 4.3 (Cingolani *et al.*, 2012), following *A. lyrata* v2 genome annotation (Rawat *et al.*, 2015). We used the standard terminology of SNPEFF 4.3 to define *cis*-regulatory elements and coding sequences, and considered all SNPs located in 5'-UTRs, introns, 3'-UTRs and in a 5-kb portion up- and downstream of a gene as variants in *cis*-regulatory elements, and all SNPs in exons as variants in coding sequences. We further

restricted our candidate list to genes containing at least three outlier SNPs to minimize the chance of identifying random allele frequency fluctuation rather than selective sweeps within a gene.

Population genetic structure

For population structure analyses, we extracted approximately 130 000 putatively neutral fourfold degenerated SNPs. We inferred a phylogenetic relationship between our eight populations using an allele frequency covariance graph, implemented in TREEMIX 1.3 (Pickrell and Pritchard, 2012). We rooted the tree by the diploid foothill population (foothill VT), as VT diploids represent the lineage that has been inferred as ancestral to all tetraploids by previous range-wide coalescent studies (Arnold *et al.*, 2015; Monnahan *et al.*, 2019). To test for the support for each branch, we bootstrapped the tree by selecting a block size of 1000 bp (equal to the window size in our selection scan) and 100 replicates. Population structure was assessed using FASTSTRUCTURE 1.0 (Raj *et al.*, 2014) for *K* values ranging from 2 (two ecotypes) to 4 (number of regions), running 10 replicates for each *K*. First, we randomly sampled two alleles per tetraploid individual (using a custom script): this approach gives unbiased results in diploid–autotetraploid systems, as has been shown by simulations (Stift *et al.*, 2019), and has also been successfully applied in a range-wide sample of *A. arenosa* (Monnahan *et al.*, 2019). We ran FASTSTRUCTURE analysis using the default parameters.

Statistical analysis

We used permutational multivariate analysis of variance (PERMANOVA) to test for overall differentiation among the transcriptomic profiles categorized according to treatment, ecotype and region. We first created multidimensional scaling (MDS) plots using the ‘MDSplot’ function in EDGER 3.12.0 (Robinson *et al.*, 2010) and extracted the corresponding distance matrix. The distance matrix was then used to compute PERMANOVA test (adonis2 function, VEGAN package; Oksanen *et al.*, 2013, number of permutation = 10000) in R (RStudio Team, 2015) using treatment, ecotype, and their interaction, or treatment, region, and their interaction, as predictors.

We quantified transcriptome-wide parallelism using a vector analysis; R scripts are available in Collyer and Adams (2007). Vector analysis aims at measuring the magnitude and direction of phenotypic evolution, in our case the overall RNA expression profile was treated as a ‘phenotype’. Briefly, we first took the number of reads of each individual that mapped on the reference and normalized the number according to library size and we calculated RPKM (reads per kilobase of transcript per million mapped reads) values using the rpkm function in EDGER. A PERMANOVA test on the 96 individuals using population and individual as variables showed that the effect of population ($R^2 = 8.64\%$, $F = 13.2$, $P < 0.001$) was significant, whereas variation between individuals within populations was non-significant ($R^2 = 5.32\%$, $F = 0.87$, $P = 0.770$). Thus, the variation between individuals was negligible compared with other factors (see the Results for the effects of treatment, ecotype and region). Hence, we characterized each population by a centroid value in the multivariate space constituted by gene expression values of its three individuals within a treatment (Figure S2). Finally, for each region, vectors connecting foothill and alpine ecotypes are compared with each other by calculating the angle (θ) between them and the difference in their lengths (magnitude). In the context of a vector analysis, parallelism is defined as two vectors pointed in the exact same direction: i.e., when θ is close to 0° and the two vectors also have similar magnitudes (Bolnick *et al.*, 2018). If θ lies in the range $0^\circ < \theta < 90^\circ$, vectors are defined as ‘acute nonparallel’, indicating

that the vectors point roughly in the same direction; if θ lies in the range $180^\circ > \theta > 90^\circ$, the vectors point in opposite directions. Then, we tested for the effects of region and treatment on the magnitude and direction using linear models, in which magnitude was log10 transformed to approach a normal distribution.

We tested for significant intersection ($P < 0.05$) (non-random overlap) in gene expression candidate lists and KEGG terms across regions (i.e., significant parallelism) using Fischer's exact test implemented in `SUPEREXACTTEST` (Wang *et al.*, 2015).

In order to investigate the genomic underpinnings of our parallel expression candidates, we tested for over-representation of outlier genomic SNPs (i.e., SNPs exhibiting extreme differentiation between corresponding foothill–alpine population pairs) associated with the parallel expression candidates. We assessed outlier SNPs located in distinct genomic elements separately, namely *cis*-regulatory and coding elements. For each region and genomic element category, we performed hypergeometric tests (using the 'phyper' function; RStudio Team, 2015) to test for outlier SNP enrichment in our lists of parallel expression candidates compared with the background of all divergence outliers (with a total number of 5% outlier SNPs in *cis*-regulatory and coding elements for each region).

In addition, for each pair of regions (six pairs in total with the full design, three pairs considering only tetraploids) we identified parallel expression candidates that harboured at least one outlier SNP at the same position (parallel outlier SNPs). We then ran hypergeometric tests to test for significant enrichment of parallel outlier SNPs in *cis*-regulatory or in coding elements compared with the background (total number of 5% outlier parallel SNPs in *cis*-regulatory and coding elements located at the same position for each pair of regions).

ACKNOWLEDGMENTS

We thank Adam Knotek for his help with collecting plant material for RNA-seq. The Plant Sciences Core Facility of CEITEC MU is acknowledged for the cultivation of the experimental plants used in this paper. Library construction and sequencing were performed at the Norwegian Sequencing Centre, University of Oslo. This work was supported by the Czech Science Foundation (project 17-20357Y to FK). Additional support was provided by the Norwegian Research Council (FRIPRO mobility project 262033 to FK), the Czech Science Foundation (17-13029S to TM), the CEITEC 2020 project (grant no. LQ1601 to TM) and by the Ministry of Education Youth and Sports of the Czech Republic (7AMB18AT022 to GW). Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, under the programme Projects of Large Research, Development, and Innovations Infrastructures.

AUTHOR CONTRIBUTIONS

GW and FK conceived the study; GW, JN and TM performed the experiments; GW, MB and FK analysed and interpreted the data; all authors contributed to writing the article and approved the final version for publication.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest associated with this work.

DATA AVAILABILITY STATEMENT

The RNA-seq data used in this study are available at Sequence Read Archives (project ID PRJNA575330). We

used a subset of the genomic data available at Sequence Read Archives (project ID PRJNA592307) and from a previous study (project ID PRJNA484107; Monnahan *et al.*, 2019) (details on the samples used are available in Table S15).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Multidimensional scaling (MDS) plot showing expression differences between the 96 samples.

Figure S2. Illustrative example of the vector analysis.

Figure S3. Numbers of differentially expressed genes showing significant gene–environment interactions and their overlaps across regions.

Figure S4. Original locations of the eight populations from the four mountain regions.

Table S1. Values for the magnitude and direction of vectors.

Table S2. List of differentially expressed genes between foothill and alpine ecotypes for each region.

Table S3. Lists of differentially expressed genes (DEGs) consistently up- or downregulated in the alpine ecotype across at least two regions for each treatment.

Table S4. Lists of enriched GO terms and REVIGO clustering of parallel genes.

Table S5. Lists of enriched GO terms and REVIGO clustering of parallel genes, excluding the diploid region VT.

Table S6. KEGG enrichment analysis of parallel genes.

Table S7. Association between parallel differentially expressed genes (DEGs) and genomic variation in coding and *cis*-regulatory elements.

Table S8. Number of candidate SNPs (at 5% outlier) at the same position in the parallel DEGs for each pair of regions.

Table S9. List of differentially expressed genes showing significant ecotype \times alpine environment interactions.

Table S10. List of differentially expressed genes showing significant ecotype \times temperature interactions.

Table S11. List of differentially expressed genes showing significant ecotype \times irradiance interactions.

Table S12. Classification and GO term enrichment analysis on parallel genes showing significant gene–environment interactions (GEIs), excluding the diploid region VT.

Table S13. KEGG enrichment analysis on genes showing significant gene–environment interactions (GEIs).

Table S14. Rearing conditions of *Arabidopsis arenosa*.

Table S15. *Arabidopsis arenosa* samples used for the genomic analysis.

Appendix S1. R script: gene–environment interaction (GEI) using EDGER.

REFERENCES

- Arnold, B., Kim, S.-T. and Bomblies, K. (2015) Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. *Mol. Biol. Evol.* **32**, 1382–1395.
- Baduel, P., Hunter, B., Yeola, S. and Bomblies, K. (2018) Genetic basis and evolution of rapid cycling in railway populations of tetraploid *Arabidopsis arenosa*. *PLoS Genet.* **14**, e1007510.
- Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) The *Arabidopsis* information resource: making and

- mining the “gold standard” annotated reference plant genome. *Genesis*, **53**, 474–485.
- Bertel, C., Buchner, O., Schönschwetter, P., Frajman, B. and Neuner, G. (2016) Environmentally induced and (epi-)genetically based physiological trait differentiation between *Heliosperma pusillum* and its polytopically evolved ecologically divergent descendent, *H. veselskyi* (Caryophyllaceae: Sileneae). *Bot. J. Linn. Soc.* **182**, 658–669.
- Blount, Z.D., Lenski, R.E. and Losos, J.B. (2018) Contingency and determinism in evolution: replaying life’s tape. *Science*, **362**, eaam5979.
- Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Bolnick, D.I., Barrett, R.D.H., Oke, K.B., Rennison, D.J. and Stuart, Y.E. (2018) (Non)parallel evolution. *Annu. Rev. Ecol. Syst.* **49**, 303–330.
- Bolnick, D.I., Snowberg, L.K., Patenia, C., Stutz, W.E., Ingram, T. and Lau, O.L. (2009) Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. *Evolution*, **63**, 2004–2016.
- Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004) GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
- Brown, J.H., Stevens, G.C. and Kaufman, D.M. (1996) The geographic range: size, shape, boundaries, and internal structure. *Annu. Rev. Ecol. Syst.* **27**, 597–623.
- Chen, J., Tsuda, Y., Stocks, M. et al. (2014) Clinal variation at phenology-related genes in Spruce: parallel evolution in FTL2 and Gigantea? *Genetics*, **197**, 1025–1038.
- Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, **6**, 80–92.
- Collyer, M.L. and Adams, D.C. (2007) Analysis of two-state multivariate phenotypic change in ecological studies. *Ecology*, **88**, 683–692.
- Conte, G.L., Arnegard, M.E., Peichel, C.L. and Schluter, D. (2012) The probability of genetic parallelism and convergence in natural populations. *Proc. Biol. Sci.* **279**, 5039–5047.
- Cooper, T.F., Rozen, D.E. and Lenski, R.E. (2003) Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *PNAS*, **100**, 1072–1077.
- Cosgrove, D.J. (2005) Growth of the plant cell wall. *Nat. Rev. Mol. Cell Biol.* **6**, 850–861.
- Derome, N. and Bernatchez, L. (2006) The transcriptomics of ecological convergence between 2 limnetic coregoninefishes (*Salmonidae*). *Mol. Biol. Evol.* **23**, 2370–2378.
- Des Marais, D.L. and Rausher, M.D. (2010) Parallel evolution at multiple levels in the origin of hummingbird pollinated flowers in *Ipomoea*. *Evolution*, **64**, 2044–2054.
- Desprez-Loustau, M.-L., Vitasse, Y., Delzon, S., Capdevielle, X., Marçais, B. and Kremer, A. (2010) Are plant pathogen populations adapted for encounter with their host? A case study of phenological synchrony between oak and an obligate fungal parasite along an altitudinal gradient. *J. Evol. Biol.* **23**, 87–97.
- Elmer, K.R., Fan, S., Kusche, H., Spreitzer, M.L., Kautt, A.F., Franchini, P. and Meyer, A. (2014) Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nat. Commun.*, **5**, 1–8.
- Elmer, K.R. and Meyer, A. (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends Ecol. Evol.* **26**, 298–306.
- Fong, S.S., Joyce, A.R. and Palsson, B.Ø. (2005) Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res.* **15**, 1365–1372.
- Fraser, H.B. (2011) Genome-wide approaches to the study of adaptive gene expression evolution. *BioEssays*, **33**, 469–477.
- Guo, J., Liu, R., Huang, L. et al. (2016) Widespread and adaptive alterations in genome-wide gene expression associated with ecological divergence of two *Oryza* species. *Mol. Biol. Evol.* **33**, 62–78.
- He, F., Arce, A.L., Schmitz, G., Koornneef, M., Novikova, P., Beyer, A. and de Meaux, J. (2016) The footprint of polygenic adaptation on stress-responsive cis-regulatory divergence in the *Arabidopsis* genus. *Mol. Biol. Evol.* **33**, 2088–2101.
- Hu, T.T., Pattyn, P., Bakker, E.G. et al. (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481.
- Jacobs, A., Carruthers, M., Yurchenko, A. et al. (2020) Parallelism in ecomorphology and gene expression despite variable evolutionary and genomic backgrounds in a Holarctic fish. *PLoS Genet.* **16**, e1008658.
- Josephs, E.B. (2018) Determining the evolutionary forces shaping G × E. *New Phytol.* **219**, 31–36.
- Karpinski, S., Gabrys, H., Mateo, A., Karpinska, B. and Mullineaux, P.M. (2003) Light perception in plant disease defence signalling. *Curr. Opin. Plant Biol.* **6**, 390–396.
- Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915.
- Knotek, A., Konečná, V., Wos, G., Požárová, D., Šrámková, G., Bohutínská, M., Zeisek, V., Marhold, K. and Kolář, F. (2020) Parallel alpine differentiation in *Arabidopsis arenosa*. *Front. Plant Sci.* **11**. <https://doi.org/10.3389/fpls.2020.561526>
- Kogami, H., Hanba, Y.T., Kibe, T., Terashima, I. and Masuzawa, T. (2001) CO₂ transfer conductance, leaf structure and carbon isotope composition of *Polygonum cuspidatum* leaves from low and high altitudes. *Plant, Cell Environ.* **24**, 529–538.
- Kolář, F., Lucanová, M., Závěská, E. et al. (2016) Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the *Arabidopsis arenosa* group (Brassicaceae). *Biol. J. Linn. Soc.* **119**, 673–688.
- Körner, C. (2003) Alpine plant life: functional plant ecology of high mountain ecosystems, 2nd ed., Berlin Heidelberg: Springer-Verlag. Available at: www.springer.com/gp/book/9783540003472 [Accessed January 29, 2019].
- Lee, K.M. and Coop, G. (2019) Population genomics perspectives on convergent adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **374**, 20180236.
- Lenski, R.E. (2017) Convergence and divergence in a long-term experiment with bacteria. *Am. Nat.* **190**, S57–S68.
- Levine, M.T., Eckert, M.L. and Begun, D.J. (2011) Whole-genome expression plasticity across tropical and temperate *Drosophila melanogaster* populations from Eastern Australia. *Mol. Biol. Evol.* **28**, 249–256.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
- Liao, Y., Smyth, G.K. and Shi, W. (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
- Losos, J.B. (2011) Convergence, adaptation, and constraint. *Evolution*, **65**, 1827–1840.
- Ma, L., Sun, X., Kong, X., Galvan, J.V., Li, X., Yang, S., Yang, Y., Yang, Y. and Hu, X. (2015) Physiological, biochemical and proteomics analysis reveals the adaptation strategies of the alpine plant *Potentilla saundersiana* at altitude gradient of the Northwestern Tibetan Plateau. *J. Proteomics*, **112**, 63–82.
- Macqueen, D.J., Kristjánsson, B.K., Paxton, C.G.M., Vieira, V.L.A. and Johnston, I.A. (2011) The parallel evolution of dwarfism in Arctic charr is accompanied by adaptive divergence in mTOR-pathway gene expression. *Mol. Ecol.* **20**, 3167–3184.
- Martin, A. and Orgogozo, V. (2013) The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution*, **67**, 1235–1250.
- Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* **17**, 10–12.
- McGirr, J.A. and Martin, C.H. (2018) Parallel evolution of gene expression between trophic specialists despite divergent genotypes and morphologies. *Evolution Letters*, **2**, 62–75.
- McKenna, A., Hanna, M., Banks, E. et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* Available at: <http://genome.cshlp.org/content/early/2010/08/04/gr.107524.110> [Accessed November 19, 2018].
- Mesíček, J. and Goliašová, K. (2002). *Cardaminopsis (CA Mey.) Hayek*. In Goliašová, K. and Šipošová, H., eds. Bratislava: Veda, pp. 388–415.
- Monnahan, P., Kolář, F., Baduel, P. et al. (2019) Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nat. Ecol. Evol.* **3**, 457.
- Oksanen, J., Blanchet, F.G., Kindt, R. et al. (2013) Vegan: Community Ecology Package. R Package Version. 2.0-10. CRAN.
- Pennig, D.W., Wund, M.A., Snell-Rood, E.C., Cruickshank, T., Schlichting, C.D. and Moczek, A.P. (2010) Phenotypic plasticity’s impacts on diversification and speciation. *Trends Ecol. Evol.* **25**, 459–467.

- Phillips, P.C. (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867.
- Pickrell, J.K. and Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967.
- Raj, A., Stephens, M. and Pritchard, J.K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.
- Rasmann, S., Pellissier, L., Defosse, E., Jactel, H. and Kunstler, G. (2014) Climate-driven change in plant–insect interactions along elevation gradients. *Funct. Ecol.* **28**, 46–54.
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198.
- Rawat, V., Abdelsamad, A., Pietzenk, B., Seymour, D.K., Koenig, D., Weigel, D., Pecinka, A. and Schneeberger, K. (2015) Improving the annotation of *Arabidopsis lyrata* using RNA-Seq data. *PLoS One*, **10**, e0137391.
- Reed, R.D., Papa, R., Martin, A. et al. (2011) optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, **333**, 1137–1141.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- RStudio Team (2015) *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>
- Sackton, T.B. and Clark, N. (2019) Convergent evolution in the genomics era: new insights and directions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **374**, 20190102.
- Stern, D.L. and Orgogozo, V. (2009) Is genetic evolution predictable? *Science*, **323**, 746–751.
- Stift, M., Kolář, F. and Meirmans, P.G. (2019) Structure is more robust than other clustering methods in simulated mixed-ploidy populations. *Heredity*, **123**, 429–441.
- Stinchcombe, J.R., Weinig, C., Ungerer, M., Olsen, K.M., Mays, C., Halldorsdottir, S.S., Purugganan, M.D. and Schmitt, J. (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. *Proc. Natl. Acad. Sci. USA*, **101**, 4712–4717.
- Streisfeld, M.A. and Rausher, M.D. (2009) Genetic changes contributing to the parallel evolution of red floral pigmentation among *Ipomoea* species. *New Phytol.* **183**, 751–763.
- Stuart, Y.E., Veen, T., Weber, J.N. et al. (2017) Contrasting effects of environment and genetics generate a continuum of parallel evolution. *Nat. Ecol. Evol.* **1**, 1–7.
- Supek, F., Bosnjak, M., Skunca, N. and Šmuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
- Velotta, J.P., Wegrzyn, J.L., Ginzburg, S., Kang, L., Czesny, S., O'Neill, R.J., McCormick, S.D., Michalak, P. and Schultz, E.T. (2017) Transcriptional imprints of adaptation to fresh water: parallel evolution of osmoregulatory gene expression in the Alewife. *Mol. Ecol.* **26**, 831–848.
- Verta, J.-P. and Jones, F.C. (2019) Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks de Meaux, J. and Tautz, D., eds. *eLife*, **8**, e43785.
- Vetaas, O.R. (2002) Realized and potential climate niches: a comparison of four *Rhododendron* tree species. *J. Biogeogr.* **29**, 545–554.
- Wang, H., Prentice, I.C., Davis, T.W., Keenan, T.F., Wright, I.J. and Peng, C. (2017) Photosynthetic responses to altitude: an explanation based on optimality principles. *New Phytol.* **213**, 976–982.
- Wang, M., Zhao, Y. and Zhang, B. (2015) Efficient test and visualization of multi-set intersections. *Sci. Rep.* **5**, 16923.
- Wildi, B. and Lütz, C. (1996) Antioxidant composition of selected high alpine plant species from different altitudes. *Plant, Cell Environ.* **19**, 138–146.
- Wittkopp, P.J. and Kalay, G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69.
- Woodhouse, M.R. and Hufford, M.B. (2019) Parallelism and convergence in post-domestication adaptation in cereal grasses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **374**, 20180245.
- Wos, G., Morkovská, J., Bohutinská, M., Srámková, G., Knotek, A., Lucanová, M., Spaniel, S., Marhold, K. and Kolář, F. (2019) Role of ploidy in colonization of alpine habitats in natural populations of *Arabidopsis arenosa*. *Ann. Bot.* **124**, 255–268.
- Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216.
- Yant, L., Hollister, J.D., Wright, K.M., Arnold, B.J., Higgins, J.D., Franklin, F.C.H. and Bomblies, K. (2013) Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr. Biol.* **23**, 2151–2156.
- Zhao, L., Wit, J., Svetec, N. and Begun, D.J. (2015) Parallel gene expression differences between low and high latitude populations of *Drosophila melanogaster* and *D. simulans*. *PLoS Genet.* **11**, e1005184.

Case study 3.

Interspecific introgression mediates adaptation to whole genome duplication






ARTICLE

<https://doi.org/10.1038/s41467-019-13159-5>

OPEN

Interspecific introgression mediates adaptation to whole genome duplication

Sarah Marburger¹, Patrick Monnahan ¹, Paul J. Seear², Simon H. Martin ³, Jordan Koch¹, Pirta Paajanen¹, Magdalena Bohutínská ^{4,5}, James D. Higgins², Roswitha Schmickl^{4,5*} & Levi Yant^{1,6*}

Adaptive gene flow is a consequential phenomenon across all kingdoms. Although recognition is increasing, there is no study showing that bidirectional gene flow mediates adaptation at loci that manage core processes. We previously discovered concerted molecular changes among interacting members of the meiotic machinery controlling crossover number upon adaptation to whole-genome duplication (WGD) in *Arabidopsis arenosa*. Here we conduct a population genomic study to test the hypothesis that adaptation to WGD has been mediated by adaptive gene flow between *A. arenosa* and *A. lyrata*. We find that *A. lyrata* underwent WGD more recently than *A. arenosa*, suggesting that pre-adapted alleles have rescued nascent *A. lyrata*, but we also detect gene flow in the opposite direction at functionally interacting loci under the most extreme levels of selection. These data indicate that bidirectional gene flow allowed for survival after WGD, and that the merger of these species is greater than the sum of their parts.

¹Department of Cell and Developmental Biology, John Innes Centre, Norwich NR4 7UH, UK. ²Department of Genetics and Genome Biology, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK. ³Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK. ⁴Department of Botany, Faculty of Science, Charles University, Benátská 2, 128 01 Prague, Czech Republic. ⁵The Czech Academy of Sciences, Zámek 1, 252 43 Průhonice, Czech Republic. ⁶Future Food Beacon of Excellence and the School of Life Sciences, University of Nottingham, Nottingham, UK. *email: roswitha.schmickl@gmail.com; levi.yant@nottingham.ac.uk

Whole-genome duplication (WGD) and hybridisation are key drivers of genomic novelty, promoting diversification in all kingdoms of life^{1–3}. Recent progress in evolutionary genomics underscores the high frequency of WGD at both ancient and recent time scales⁴, and population genomic approaches reveal widespread evidence of gene flow between the most diverse species^{5,6}. Both processes have therefore been associated with adaptive benefits. However, WGD and hybridisation are large-effect mutations, often leading to a host of genomic instabilities, including epigenetic shock, perturbed gene expression patterns and meiotic instability, with direct negative impacts on fertility. Perhaps the most challenging issue is the most immediate: that of stable meiotic chromosome segregation following WGD. How nascent polyploids establish meiotic stabilisation remains an unresolved question.

The wild outcrossing members of the *Arabidopsis* genus have recently emerged as fruitful models for the study of genome establishment following WGD⁷. *Arabidopsis arenosa* is a largely biennial outcrossing relative of the model *Arabidopsis thaliana*, which forms distinct lineages of diploids and autotetraploids throughout Central Europe^{8–11}. Initial resequencing of a handful of autotetraploid *A. arenosa* individuals suggested selective sweep signatures at genes involved in genome maintenance, including DNA repair, recombination and meiosis¹². Later, a targeted resequencing effort focused on patterns of differentiation between diploid and autotetraploid *A. arenosa*, revealing evidence of highly localised selective sweeps directly overlapping eight loci whose gene products interact during prophase I of meiosis¹³. These eight loci physically and functionally interact to control crossover designation and interference, strongly implying that a modulation of crossover distribution was essential for polyploid establishment in *A. arenosa*^{14,15}. Cytological evidence of a reduction in crossover numbers in the autotetraploids indicated that the selected alleles had an effect¹³. Similar to its sister species *A. arenosa* (*arenosa* hereafter), *Arabidopsis lyrata* (*lyrata* hereafter) also naturally occurs as diploids and tetraploids across its distribution range^{8,16–18}. Although there is little evidence for gene flow among diploids of each species, there have been reports of gene flow between tetraploid *arenosa* and *lyrata* and, less pronounced, gene flow between diploids and tetraploids^{8,19,20}.

Here we investigate the molecular basis of parallel adaptation to WGD in *lyrata* compared with *arenosa* and the possibility of adaptive gene flow between the two species. Specifically, we ask (1) whether the same or different loci may be involved in adaptation to WGD in *lyrata* as we observed in *arenosa*; and (2) whether these adaptations arose independently or via introgression from one species into the other. Using whole-genome sequence data from 92 individuals of *lyrata*, *arenosa* and outgroup species *Arabidopsis croatica* and *Arabidopsis halleri*, we first analyse population structure and demography, concentrating on assessing admixture and the degree and timing of population divergences. Then, to estimate the relative degree of adaptation to WGD across the ranges of *lyrata* and *arenosa*, we cytologically assess meiotic stability in key populations. Finally, after scanning the *lyrata* genomes for signatures of selective sweeps, we compare the most differentiated regions with those we previously found in *arenosa*^{12,13} and test whether these selective sweep signatures overlap with fine-scale conspicuous introgression signals. Overall, our results reveal the molecular basis by which WGD has been stabilised in both species and indicate that WGD-facilitated hybridisation allowed for stabilisation of meiosis in nascent autotetraploids by highly specific, bidirectional adaptive gene flow.

Results

Population structure and broad-scale admixture. To understand population and species relationships, we analysed the

genomes of 92 individuals from ~30 populations of *lyrata* and *arenosa* throughout Central Europe along with outgroups, sequenced at a depth of ~15× per individual (Supplementary Table 1 and Supplementary Fig. 1). STRUCTURE and principal component analysis (PCA) showed a clear species-specific clustering for diploids, whereas tetraploids exhibited a gradient of relatedness between species (Fig. 1a, b). Admixture was markedly lower in *arenosa* populations than in *lyrata*: consistently, all diploids tested (SNO, KZL, SZI, BEL) and tetraploids from the Western Carpathians (TRE), and most of the Alpine tetraploids (HOC, GUL, BGS) harboured essentially pure *arenosa* genomes (Fig. 1a). Minimal admixture signal (<1%) with *arenosa* was detected in a few *lyrata* genomes, in particular the Austrian diploid (PEQ, PER, VLH), as well as the *lyrata* eastern tetraploid (*Let* hereafter) populations (LIC, MOD) and the tetraploid KAG population (Fig. 1a).

In contrast, many other *lyrata* populations exhibited substantial admixture signals with *arenosa*, varying drastically in degree (Fig. 1a). Several tetraploid *lyrata* populations from the Wachau (SCB, SWA, MAU) displayed only slight admixture with *arenosa* and populations at the Wachau margin (PIL, LOI) showed stronger admixture, probably due to the increased proximity to the Hercynian and Alpine *arenosa* lineages²¹. Compared with the Wachau, where *lyrata* occurs on the slopes and hilltops along the Danube river surrounded by *arenosa* populations outside of the valley, there is a classical hybrid zone in the eastern Austrian Forealps: the parental species are found at the two poles of the zone (diploid and tetraploid *lyrata* in the Wienerwald, tetraploid *arenosa* at higher altitudes to the west and the hybrids between; Supplementary Fig. 1). Populations HAL, ROK, FRE, OCH and KEH are heavily admixed, with KEH appearing more *arenosa*-like, and ROK and FRE being slightly more *lyrata*-like compared with the others (Fig. 1a). Again, proximity of the hybrids to the donor species corresponded with increased admixture. The Hungarian tetraploid *lyrata* population GYE also exhibited admixture signal, suggesting that gene flow between *lyrata* and *arenosa* is not restricted to the Austrian Forealps. PCA was consistent with STRUCTURE findings, with PC1 dividing samples by species (explaining >36% of the variance; Fig. 1b). The second axis (<5% of the variance) separated KZL and SZI from the other diploid *arenosa* populations. These are representatives of the Pannonian lineage, which is the oldest and most distinct diploid *arenosa* lineage²¹. Overall, our results are consistent with previous descriptions of introgression between *lyrata* and *arenosa* in Austria that were based on smaller marker sets and different sampling schemes^{8,22}.

We estimated the population split time without migration between *lyrata* and *arenosa* at 931,000 (931k) generations ago using *fastsimcoal2*²³ (Fig. 1c, Supplementary Figs. 2 and 3, and Supplementary Table 2). This translates to ~2 million years ago (mya), given an average generation time of 2 years, which would coincide with the onset of Pleistocene climate oscillations. This estimate lies within the range of age estimates for this split from ref. 24 with 1.3 mya and from ref. 25 with 8.2 mya, and ref. 22. We estimated the age of WGDs at 81k generations ago for *lyrata* (~160,000 years ago) and 226k generations ago for *arenosa* (~450,000 years ago), which approximately mark periods of glacial maxima²⁶. Noting this, we next asked if either species experienced substantial historical bottlenecks. Using pairwise sequentially Markovian coalescent (PSMC) model²⁷ we could not reach ages as ancient as 130–300 kya (Supplementary Fig. 4), because the recombinant blocks that PSMC measures are too short in these diverse outcrossing species to estimate ancient population histories. Our analysis indicated that diploid *lyrata* had a peak effective population size (N_e) ~25 kya (PER, VHL) and ~20 kya (PEQ), whereas diploid Dinaric *arenosa* (BEL)

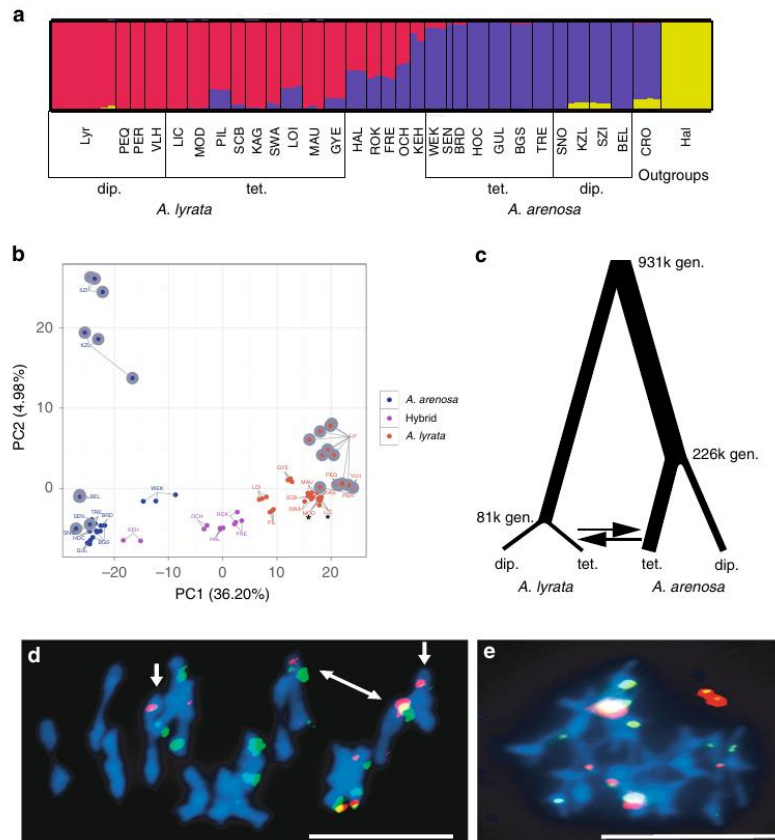


Fig. 1 Ploidy-specific admixture and stable autotetraploid meiosis in *A. lyrata*. **a** A continuous range of admixture specifically in tetraploid populations demonstrated with STRUCTURE analysis of nuclear SNP data (32,256 LD-pruned, 4-fold degenerate SNPs). Populations (in three-letter code) and population groupings (ploidy, species) are displayed. Populations are described in (Supplementary Table 1). **b** PCA shows individuals group on the main (PC1) axis by species and not by ploidy, with hybrid individuals located between *A. lyrata* and *A. arenosa* samples. We refer to all non-pure populations from the hybrid zone in the eastern Austrian Forealps as hybrids (see Supplementary Fig. 1). Diploids are indicated by grey outline. Asterisks (*) are placed under the *Lwt* tetraploid grouping; all other *A. lyrata* tetraploids (except the geographically divergent Pannonian GYE) are in the *Lwt* group. **c** Demographic parameter estimates for *A. lyrata* and *A. arenosa* populations. Line widths are proportional to estimates given in Supplementary Fig. 2. **d, e** Metaphase I chromosome spreads of nuclei from two ROK plants hybridised with 5S rDNA (red) and 45S rDNA (green). **d** MI scored as stable as 16 individual bivalents are observed, even though there are bivalents with unequal probes (white arrows), suggesting non-homologous rearrangements. **e** MI scored as unstable as the majority of chromosomes are connected to each other. Chromosomes are stained with DAPI; bar = 10 μ m. The source data underlying Fig. 1d, e are provided as a Source Data file

peaked earlier, at ~30 kya. Interestingly, diploid Western Carpathian *arenosa* SNO, the population that founded several widespread autotetraploid lineages⁹, gave a strong signal of continuous expansion. These results suggest that diploid *lyrata* and partly *arenosa* underwent a bottleneck after the last glacial maximum 30–19 kya. PSMC does not accommodate autotetraploid data, but using *fastsimcoal2* we detected a strong bottleneck at WGD for *lyrata* and none for *arenosa* (Supplementary Fig. 2).

We next assayed for patterns of gene flow using coalescent modelling with *fastsimcoal2*. Due to model overfitting when using more than two migration edges, we chose the model retaining only two migration edges with the highest support: interspecific gene flow from tetraploid *arenosa* to tetraploid *lyrata* (0.1 alleles/generation) and *lyrata* to *arenosa* gene flow at the same level (0.1 alleles/generation) (Fig. 1c, Supplementary Figs. 2 and 3, and Supplementary Table 2). These results indicate equal amounts of

bidirectional gene flow specifically among the tetraploids of both species.

Stabilisation of *lyrata* meiosis following WGD. Given the very low abundance of *lyrata* tetraploids compared with tetraploid *arenosa* in nature, we assayed whether these tetraploids were indeed meiotically stable. Cytological analysis indicated that in fact *lyrata* tetraploids exhibit similar levels of meiotic stability as *arenosa* tetraploids, as evidenced by relative percentages of stable rod and ring bivalents (Fig. 1d and Supplementary Table 3) vs. less stable multivalents (Fig. 1e and Supplementary Table 3). We were surprised to observe among both species that meiotic stability segregates within populations, typically ranging from <20 to >60% stable metaphase I cells per plant with extremes observed in KAG (0–98%) and consistently higher levels (>80%) in the

Table 1 Genome-wide differentiation between *A. lyrata* diploids and tetraploids, and between tetraploid lineages grouped by biogeography

Contrast	No. of SNPs	AFD	d_{XY}	Fst	Rho	Fixed Diff
Diploid <i>lyrata</i> vs. <i>lyrata</i> eastern tetraploids (<i>Let</i>)	2,904,110	0.14	0.22	0.09	0.19	270
Diploid <i>lyrata</i> vs. <i>lyrata</i> Wachau tetraploids (<i>Lwt</i>)	3,794,257	0.11	0.16	0.07	0.17	64
<i>Lyrata Let</i> tetraploids vs. <i>Lwt</i> tetraploids	4,795,381	0.09	0.16	0.06	0.13	24
<i>Arenosa</i> Hercynian tetraploids vs. <i>arenosa</i> Alpine tetraploids	1,812,223	0.10	0.16	0.03	0.07	0

Differentiation metrics shown are allele frequency difference (AFD), d_{XY} , Fst, Rho and the number of fixed differences (Fixed Diff). Multiple differentiation metrics were used, as the metrics exhibit different sensitivities to diversity and differentiation. Values of all metrics were averaged over pairwise comparisons of populations belonging to that group

arenosa populations. Meiotic stability was also variable within the tetraploid *arenosa* population TBG, which was the population used by Yant et al.¹³ to cytologically assess meiotic stability. A much higher number of chromosome spreads on more individuals and populations in the present study indicates that meiosis is not universally stable among autotetraploids across these populations. Overall, these results indicate that meiotic stability is broadly segregating within tetraploid populations of both species.

Selective sweep signatures in *lyrata*. To gain insight into the processes underlying adaptation to WGD in *lyrata* tetraploids, we performed a population-based genome scan for selection. We quantified differentiation between *lyrata* ploidies by calculating d_{XY} ²⁸, Fst²⁹ and Rho³⁰ in adjacent windows along the genome between diploids and tetraploids. Fst is influenced by within-population diversity and lacks sensitivity in cases of low differentiation. Therefore, we used additional differentiation metrics. d_{XY} does not take within-population diversity into account, whereas Rho is a divergence measure that is independent of ploidy level and double reduction in autopolyploids. We focused on the non-admixed *lyrata* tetraploid populations LIC and MOD (*lyrata* eastern tetraploids; *Let*), which by STRUCTURE and PCA analyses exhibited the lowest levels of admixture (Fig. 1a) and clustered with *lyrata* diploids, distant from the *arenosa* tetraploids or the broadly admixed *lyrata* tetraploids (Fig. 1b). Overall, genome-wide differentiation levels between *lyrata* diploids and the tetraploids indicate shallow divergence between all groups (with mean Rho in the most differentiated contrast between ploidies = 0.19; Table 1 and Supplementary Table 4 for additional population contrasts), consistent with our previous studies in *arenosa*^{12,13,21,31}.

To identify the most robust signals of selection in the tetraploid *lyrata* populations, we performed genome scans on two different *lyrata* tetraploid population groups and then focussed on the genes that were repeatedly in the extreme 1% outlier windows in both contrasts. This identified 196 genes (0.6% of gene-coding loci in the genome; Supplementary Dataset 1). First, contrasting the *lyrata* diploids and the *Let* tetraploids, we partitioned the genome into gene-sized windows and identified outliers for allele frequency differences (AFDs), d_{XY} , Fst, Rho and the number of fixed differences. Although the comparison of the most pure *lyrata* tetraploid populations, represented by the *Let* group, to *lyrata* diploids is the most stringent test of which loci are under selection in a purely *lyrata* genomic context, we extended our tetraploid *lyrata* sampling to populations from the Wachau, which frequently showed admixture with *arenosa* (*lyrata* Wachau tetraploids, *Lwt* hereafter: PIL, SCB, KAG, SWA, LOI and MAU; GYE was excluded due to distant geographic grouping in Pannonia). As the *Let* and part of the *Lwt* populations grow in contrasting edaphic conditions (*Let* on limestone, *Lwt* on siliceous bedrock), we used this approach to maximise our chances of capturing differentiation specifically related to ploidy

and not local adaptation. In addition, we observed that differentiation between these two tetraploid *lyrata* groups is stronger than differentiation between the tetraploid *arenosa* lineages studied here (Table 1), suggesting that there is stronger genetic structure within *lyrata* than *arenosa*, as was observed by ref.³², and supporting a degree of independence between the *Let* and *Lwt* divergence scans.

Gene Ontology (GO) enrichment analysis of these 196 genes identified significant overrepresentations in categories related to meiotic and homologous chromosome segregation, but also diverse processes including epidermal cell differentiation, trichoblast maturation, root hair cell and epidermal differentiation, root hair cell development and elongation, and others such as indole-containing compound metabolic process and mRNA catabolic process (Supplementary Fig. 5 and Supplementary Dataset 2). These results indicate that evolutionary change may occur throughout a broad array of processes during adaptation to WGD, beyond meiotic chromosome segregation.

Comparing this set of outliers to those found under selection upon WGD in *arenosa*¹³, 20 gene-coding loci exhibited the highest levels of differentiation in both studies (Table 2). These included those meiosis-related loci reported above (*PRD3*, *ASY1*, *ASY3* and *SYN1*), as well as the endopolyploidy genes *CYCA2;3* and *MEE22*, and the global transcriptional regulator *TFIIF*, among others. We observed selective sweep signatures at the majority (6/11) of coding loci of known function that were found as the very top outliers in *arenosa* (0.5% outliers for all three metrics used in that study) having primary functions of mediating meiosis, endopolyploidy and transcription. In particular, outlier loci participating in meiotic crossover formation, including *ASY1*, *ASY3*, *PDS5-like*, *PRD3*, and *SYN1* exhibited tight peaks of divergence directly over single gene-coding loci (an example is given in Fig. 2), a divergence signal we have broadly seen in this system^{13,21,31}. In addition, the meiosis loci important for crossover formation reported by Yant et al.¹³ *ZYP1b* and *PDS5* were outliers in the *Lwt* contrast. The paralog *ZYP1b* was differentiated in the *Let* group also, but was not among the 1% top outliers; *PDS5* showed no differentiation between the *Let* and *lyrata* diploids. *SMC3*, a top outlier in *arenosa*, showed only moderate differentiation in the *Lwt* and no differentiation in the *Let* scan. Taking this most restrictive list representing the overlap of three genome scans, GO enrichment analysis identified significant overrepresentations only in categories related to meiotic chromosome segregation (Supplementary Dataset 3). These results further support the notion that these same loci were under the highest levels of selection following the more recent WGD event in *lyrata* as were under selection following the independent, earlier WGD (Fig. 1c) in *arenosa*.

Apart from loci encoding meiosis-related genes, we detected extreme differentiation at loci belonging to other functional categories clearly related to the challenges attendant to WGD, including loci involved in endoreduplication and transcriptional regulation: *CYCA2;3*, *PAB3*, *NAB*, *TFIIF* and *GTE6*. WGD

Table 2 Overlap list of the top 1% outliers from the genome scans

Lyrata ID	Name	Description	Let scan Outlier	Lwt scan Outlier
AL1G10680	PRD3	Involved in meiotic double strand break formation	Yes	Yes
AL1G27690	CYCA2;3	Negatively regulates endocycles and acts as a key regulator of ploidy levels in endoreduplication	Yes	Yes
AL1G36300	PBP3	Putative poly(A) binding protein	Yes	Yes
AL2G25520	SWEETIE	Involved in trehalose metabolic process	Yes	Yes
AL2G25920	ASY1	ASYNAPTIC 1 mediates meiotic crossovers	Yes	Yes
AL2G37810	PDS5-like	ARM repeat superfamily protein	Yes	Yes
AL2G40680	CMT1	Chromomethylase 1 DNA methyltransferase	Yes	Yes
AL4G29630	NAB	Nucleic acid-binding, OB-fold-like protein	Yes	Yes
AL4G29650	Unknown		Yes	Yes
AL4G30770	MEE22	Involved in endoreduplication and cell fate	Yes	Yes
AL4G46460	ASY3	ASYNAPTIC 3 required for normal meiosis	Yes	Yes
AL5G13440	ASF	Asparagine synthase family protein	Yes	Yes
AL5G32850	PSF	Pseudouridine synthase family protein	Yes	Yes
AL5G32860	TFIIF	Functions in RNA polymerase II activity	Yes	Yes
AL5G32870	GTE6	Bromodomain containing nuclear-localised protein involved in leaf development	Yes	Yes
AL5G39280	NRPA1	Subunit of RNA polymerase I (aka Pol A)	Yes	Yes
AL6G15380	SYN1	A RAD21-like gene essential for meiosis	Yes	Yes
AL7G35790	unknown		Yes	Yes
AL8G25590	DYAD, SWI1	Involved in meiotic chromosome organisation	Yes	Yes
AL8G25600	TPR-like	Tetratricopeptide repeat (TPR) protein	Yes	Yes
AL1G35730	ZYP1a, b	Transverse filament of meiotic synaptonemal complex	No	Yes
AL4G20920	SMC3	Member of the meiotic cohesin complex	No	No
AL8G10260	PDS5	Member of the meiotic cohesin complex	No	Yes

Overlap list of the top 1% outliers from the genome scan of diploid *A. lyrata* vs. *Let* and diploid *A. lyrata* vs. *Lwt* overlapped with the outliers of the *A. arenosa* diploid-tetraploid scan of Yant et al.¹³. The overlap between the diploid/*Let* and diploid/*Lwt* contrasts yielded 196 genes, which is approximately a third of the genes identified in each scan. The overlap of those two scans with the *A. arenosa* scan gave 20 genes in common. Core meiosis genes found in Yant et al.¹³, which were found in only one or none of the two *lyrata* scans, are stated in the bottom part of this list

increases the ploidy of all cells, whereas endopolyploidy occurs in single cells during their differentiation, and this cell- and tissue-specific ploidy variation is important in plant development^{33–35}. Thus, given the instantly doubled organism-wide nuclear content following WGD, we postulate that the degree of endopolyploidy would be modulated in response, with accumulating support for this notion^{36–38}. Our findings bolster the idea that there may be a link between organism-wide polyploidization, and that of single cells within an organism. Research about the effect of WGD-induced dosage responses of the transcriptome is still in its infancy^{3,39} and emerging studies on allopolyploids support incomplete dosage compensation.

Highly specific introgression at sweep genes. Finally, we sought to confirm whether the strong observed signals of selective sweep were the products of localised interspecific introgression. To confirm candidate introgressed regions at high resolution, we used *Twisst*⁴⁰, performing two independent analyses, with either *Let* or *Lwt* representing tetraploid *lyrata*. The consensus species phylogeny, topology 3, represented the overwhelmingly dominant genome-wide topology (Fig. 3a). Topologies consistent with introgression (6, 11 and 14, which group tetraploids of the two species together) all had comparatively low values, but also showed multiple narrow peaks across the genome. Twelve peaks had weightings >0.7 and nine of these overlapped with our divergence scan outliers (Fig. 3b and Supplementary Dataset 4). Similarly, 61 had a weighting >0.5 and 21 (34%) of these overlapped with gene-coding loci that were positive in both the *Let* and *Lwt* divergence scans (Fig. 3b and Supplementary Dataset 4). This degree of overlap of the loci found under selection in our genome scans is dramatically greater than expected by chance (0.6%), which we confirmed by performing permutation tests (Supplementary Fig. 6). By contrast to the introgression-indicative topologies, those consistent with incomplete lineage sorting (ILS) alone (7, 8, 9, 12 and 13, which group diploid

arenosa with tetraploid *lyrata* or vice versa) were low genome-wide with only two peaks reaching above 0.5 (Fig. 3a).

Similar to the divergence outlier windows, *Twisst*-positive windows were narrow, which might be an indication that genomic differentiation following divergence between *lyrata* and *arenosa* is advanced and introgression occurred fairly deep in the past, similar to the numerous narrow genomic regions of introgression in the case of gene flow between *Populus alba* and *Populus tremula*⁴¹. However, we have recently estimated linkage disequilibrium (LD) in this system²¹, finding a very rapid reduction specifically in the autotetraploid cytotype (50% lower mean correlations at 1 kb distance), suggesting that tight introgression signals may be formed rapidly in this system.

Although sharing of adaptive alleles between tetraploid populations can also be explained by ILS, the symmetrical design of our study allows us to reject ILS in most cases. Under ILS, we expect two divergent alleles to have existed in the ancestor of both species, which would lead to topology 6 after sorting of these alleles into diploids and tetraploids, respectively. Of the top 12 *Twisst* peaks, only one represents topology 6. The others represent topologies 11 and 14, in which the tetraploid alleles are nested within the diploids, implying that they arose after speciation and subsequently introgressed (Fig. 3a). A slight majority of these loci with introgression signal appear to have a *lyrata* origin (30/53 *Twisst* peaks), but among those with the highest levels of selection in *Lwt*, *Let*, and *arenosa*, a majority (11/16 where direction can be inferred) harbour evidence of an *arenosa* origin (Supplementary Dataset 4). GO enrichment analysis of the gene-coding loci in the windows where direction could be inferred found only enrichment for categories related to meiosis (Supplementary Datasets 5 and 6).

Taking these results together, we observe that four meiosis-related loci were outliers in all *Twisst* and divergence scans: *PRD3*, *ASY3*, *SYN1* and *DYAD*, and four did not show a signal in both *Twisst* analyses as well as both divergence scans: *ASY1*, *ZYP1a*,

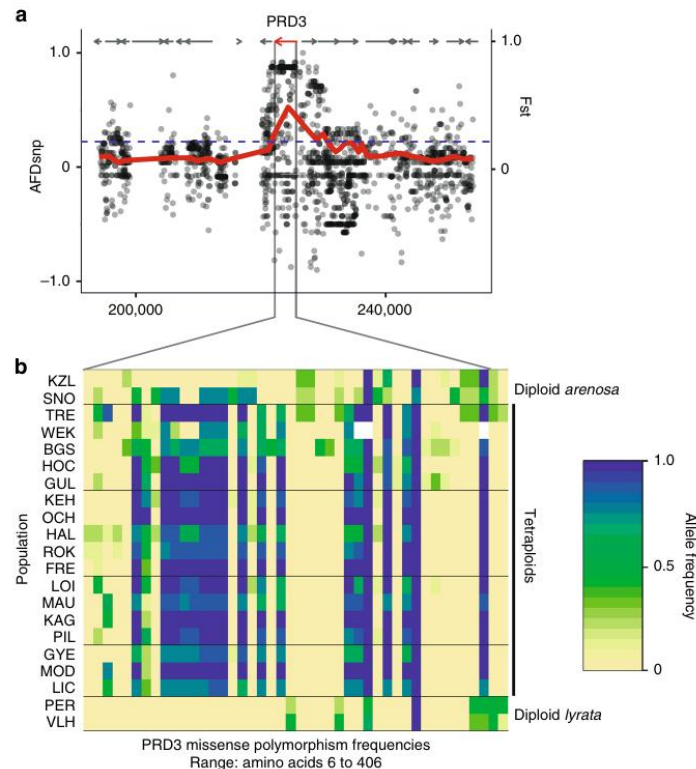


Fig. 2 Selective sweeps and missense polymorphism frequencies by population. **a** Selective sweep example in *PRD3*, a gene involved in meiotic double strand break formation. X-axis gives chromosome 1 position in base pairs. Left Y-axis gives allele frequency differences between diploid and tetraploid *A. lyrata* and at single-nucleotide polymorphisms (dots). Right Y-axis (and red line) gives local F_{st} . Arrows indicate gene models. Red arrow indicates selective sweep candidate with localised differentiation. The dotted line gives the 99th percentile of genome-wide F_{st} values. **b** Zoom-in on *PRD3* coding changes. Heatmap represents allele frequencies of missense polymorphisms. Frequencies 0–100% follow yellow to green, to blue. Derived diploid *A. arenosa*-specific missense polymorphisms are driven to high frequency in the tetraploids, whereas diploid *A. lyrata* alleles are absent, implicating diploid *A. arenosa* origin to this selected allele in the tetraploids

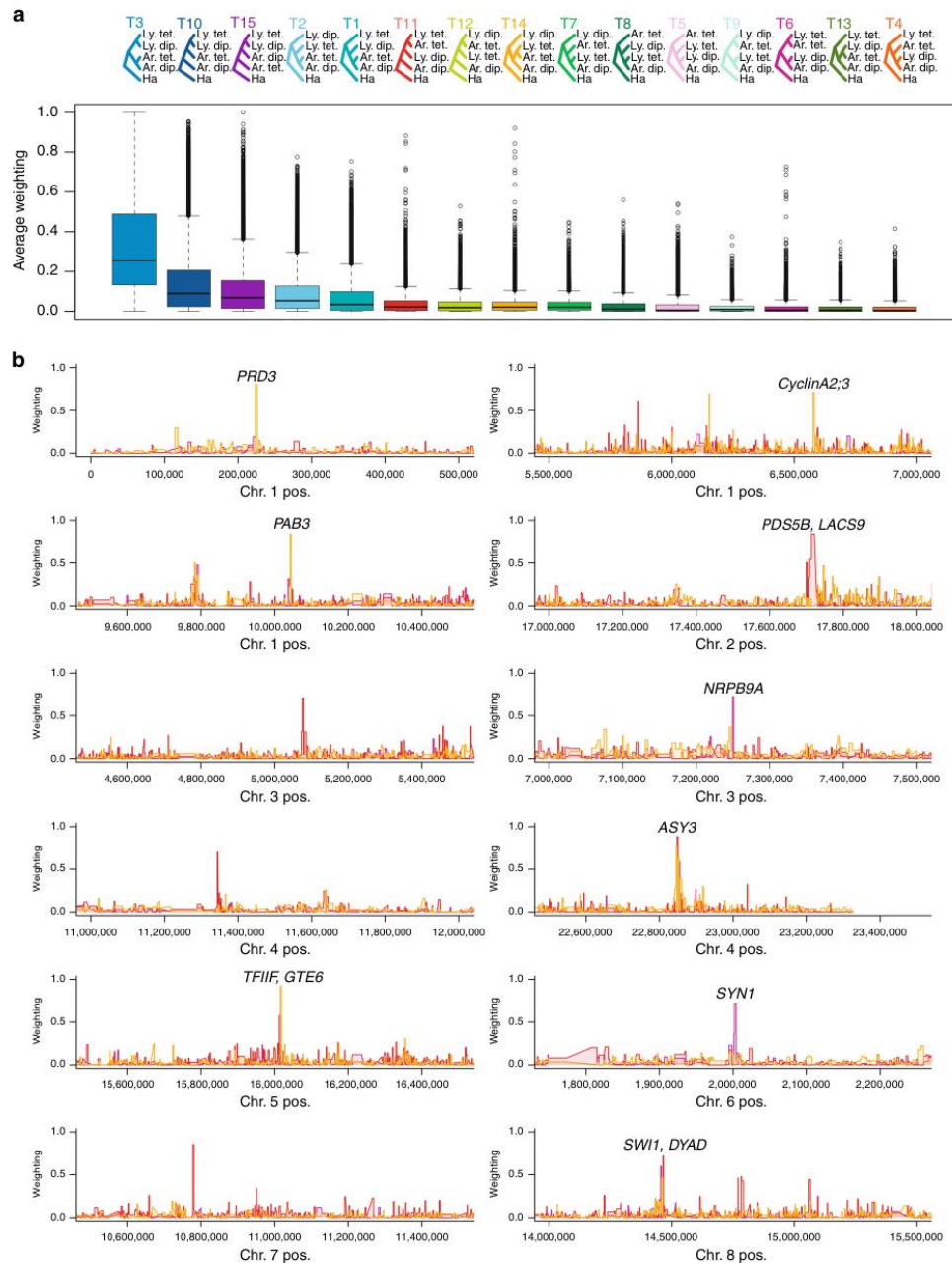
ZYP1b and *PDS5*. Given that *arenosa* is the much older tetraploid (2.5 times as ancient as *lyrata*; Fig. 1c) and is much more widespread, we hypothesise that *arenosa*-sourced alleles were under selection for stable polyploid meiosis longer, providing pre-adapted alleles to the nascent *lyrata* tetraploid, although this hypothesis needs to be functionally tested in dedicated studies. Introgression of optimised alleles from an older to a younger species has been indicated for high-altitude adaptive alleles from Denisovans and Tibetan *Homo sapiens*⁴². Taking *PRD3* as a clear example (Fig. 2b), derived *arenosa*-specific missense polymorphisms in the diploid population SNO are driven to high frequency in the tetraploids of both species, whereas diploid *lyrata* alleles are absent, strongly implicating a specific diploid *arenosa* origin in this case. At the same time, we detect specific signals of gene flow from *lyrata* into *arenosa*. In addition to meiosis-related genes, we see introgression signal at the endopolyploidy gene *CYCA2;3* and the global transcriptional regulator *TFIIIF*, but very few other loci exhibit both persistent signatures of extreme selection as well as introgression (Supplementary Datasets 3 and 4).

Our findings suggest that introgression of particular alleles of meiosis-related genes might stabilise polyploid meiosis, with the less effective alleles of one species being replaced by introgression of alleles from orthologous loci in the other tetraploid.

Introgression of alleles optimised for adaptation to WGD could be especially beneficial in hybrid zones such as this one, which spans a climatic gradient from a warmer, Pannonian climate at its eastern margin to harsher conditions in the eastern Alps. Meiosis is a temperature-sensitive process¹⁵ and we hypothesise substantial levels of meiosis-related allele–environment associations with variable temperature. Allele–environment associations with climatic variables across a hybrid zone have been observed in spruce⁴³.

Discussion

For these newly formed tetraploids, WGD appears to be both a blessing and a curse. Although WGD appears to have opened up access to the allelic diversity of a sister species, as well as provided population genomic benefits²¹, it also presents new challenges to the establishment of optimal allelic combinations. As the gene products at the meiotic loci under the most extreme selection across this hybrid zone functionally and physically interact, we expect that efficient evolved polyploid meiosis requires the harmonious interactions of multiple selected, introgressed alleles in concert. However, relatively high levels of residual masking of genetic load in autotetraploids^{21,44} will tend to extend the duration that deleterious alleles segregate in populations, with negative



phenotypic consequences. This is consistent with our observation that polyploid meiosis exhibits wide degrees of within-population variability in stability. This observed diversity suggests that the optimal combination of meiosis alleles is yet segregating, which may also be the result of the recent age of these WGDs. Dedicated molecular investigation of whether the measured within-population meiotic stability is associated with particular allele combinations is the focus of ongoing functional analyses.

In this study, we investigated the population genetic basis of adaptation to WGD in congeners that, due to an endosperm-based postzygotic barrier²⁰, hybridise only as tetraploids. We found that many of the same loci exhibit the most conspicuous signatures of selective sweep in *lyrata* following WGD that we observed in *arenosa*, and further, that the strongest signals of interspecific introgression occur precisely at many of these same loci. Using whole-genome sequence data from 30 populations, we

Fig. 3 Highly specific introgression events across species boundaries. **a** Topologies from *Twisst* analysis of *Lwt*: Although topology 3 is the dominant species tree, topologies 11, 14 and 6 indicate localised gene flow between tetraploids. Box plots give relative weightings of all topologies across the genomes analysed. It is noteworthy that the extreme outliers concentrate specifically on the introgression-indicative topologies 11, 14 and 6. The bold line indicates the median. The box spans the first and third quartiles, and the whiskers extend to the most extreme point within 1.5 times the interquartile range from the box. Source data are provided as a Source Data file. **b** Introgression events revealed by *Twisst* analysis are highly localised at loci encoding genes controlling meiosis, endopolyploidy, and transcriptional control. All gene-coding loci under a given narrow peak are labelled; many of the indicated loci are divergence scan outliers in both the *Let* and *Lwt* divergence scans in addition to being *Twisst* outliers. The genome-wide dominant topology 3 weightings are omitted in **b** for clarity. The colours in **b** correspond to topologies 11, 14 and 6 in **a**. The weighting quantifies the extent to which each 50 SNP window tree matches a given topology, accounting for the fact that each taxon is represented by multiple individuals that each have 2 (for diploids) or 4 (for tetraploids) tips in the tree. A weighting of 1 indicates that all individuals cluster in the same way, such that all possible subtrees match the same topology. Weightings >0 but <1 indicate that different subtrees match different topologies. Source data are provided as a Source Data file

probed complex population structure and patterns of gene flow. Interestingly, we observed cytologically that the degree of meiotic stability varied dramatically, even within populations of both species, suggesting that stability has not been completely established, or that other, perhaps epigenetic or environmental factors influence meiotic stability in still unknown ways. At the same time, populations exhibited admixture signals that contrast dramatically in degree, indicating a complex introgression landscape. We present evidence that the molecular basis by which WGD was stabilised in *lyrata* and *arenosa* is shared. Our data further suggest that WGD-facilitated hybridisation allowed for stabilisation of meiosis in nascent autotetraploids by specific, bidirectional adaptive gene flow, tightly overlapping loci known to be essential for processes that are impacted by WGD: meiotic stability, endopolyploidy, and transcription, and others. It is curious that the very process that rescues fitness in these species, hybridisation, is potentiated by the same phenomenon to which the resultant adaptive gene flow responds: WGD.

Methods

Sample design and sequencing. Individual plants were collected from field sites across Central Europe (Supplementary Fig. 1). Cytotypes were determined by flow cytometry from these populations in ref. 8 and ref. 21, no triploids have been detected in these populations, nor have we found any evidence in the flow cytometry or cytology data that any of these populations consist of mixed-ploidy subpopulations.

Central European tetraploid *lyrata* has its largest distribution in eastern Austria, in two biogeographic regions: the Wienerwald (*lyrata* eastern tetraploids/*Let* hereafter: LIC, MOD), and the Wachau (*lyrata* Wachau tetraploids/*Lwt* hereafter: PIL, SCB, KAG, SWA, LOI, MAU). We found an additional tetraploid *lyrata* population in Hungary (GYE) and included it in this study. Diploid *lyrata* populations were chosen from the Wienerwald (PEQ, PER, VLH), which are the geographically closest diploid populations to the *Let* and *Lwt*, and therefore likely serve as source populations.

For *arenosa*, representative populations of tetraploids from the Hercynian (WEK, SEN, BRD) and Alpine lineages (HOC, GUL, BGS) were selected, as well as additional *arenosa* populations from the Western Carpathians (diploid: SNO; tetraploid: TRE), which is the centre of *arenosa* genetic diversity and the region of origin of the tetraploid cytotype⁹. For breadth, we selected several more diploid *arenosa* populations from the Pannonian (KZL, SZI) and Dinaric (BEL) lineages, as well as the following populations from the hybrid zone in the eastern Austrian Forealps: HAL, ROK, FRE, OCH and KEH. To complement our sampling with diploid *lyrata* from across its entire distribution range, we selected samples from the Hercynian (SRR2040791, SRR2040804), arctic-Eurasian (SRR2040796, SRR2040798, SRR2040805) and arctic-North American lineages (DRR054584, SRR2040769, SRR2040770, SRR2040789). *A. croatica* (CRO) and *A. halleri* (SRR2040780, SRR2040782, SRR2040783, SRR2040784, SRR2040785, SRR2040786, SRR2040787) were included as outgroups³². The majority of *lyrata* and hybrid samples were collected as seeds, cultivated and flash-frozen prior to DNA extraction, whereas samples for three populations (LIC, MOD, HAL) were collected and silica-dried. Silica-dried material from GYE was obtained from Marek Šlenker and Karol Marhold. *Arenosa* samples were collected and sequenced as part of a different study²¹. In addition, 16 accessions were downloaded from the NCBI Sequence Read Archive (SRA), bringing the total sample number to 92 (Supplementary Table 1). DNA of the *lyrata* and hybrid samples was extracted and purified from frozen or silica-dried leaf and/or flower tissue using the Epicentre MasterPure DNA extraction kit. DNA concentration measurements were performed with the Qubit 3.0 fluorometer (Invitrogen/Life Technologies, Carlsbad, California, USA). Genomic libraries for sequencing were prepared using the

Illumina TRUSeq PCR-free library kit with 500 ng to 1 µg extracted DNA as input. We multiplexed libraries based on the Qubit concentrations, and those multiplexed mixes were run on an initial quantification lane. According to the yields for each sample, loading of the same multiplex mix on several lanes was increased to achieve a minimum of 15× coverage. Samples that had less than our target coverage were remixed and run on additional lanes. Libraries were sequenced as 125 bp paired-end reads on a HiSeq2000 by the Harvard University Bauer Core Facility (Cambridge, MA, USA).

Data preparation and genotyping. Newly generated sequencing data and SRA accessions were processed together from raw fastq reads. We first used Cutadapt⁴⁵ to identify and remove adapter sequences with a minimum read length of 25 bp and a maximum error rate of 0.15. We then quality trimmed reads using TRIM-MOMATIC⁴⁶ (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:50). Samples sequenced on several lanes were then concatenated using custom scripts. Reads were deduplicated using MarkDuplicates in picard v.1.103. Broadin and readgroup names were adjusted utilising AddorReplaceReadGroups within the picard package. Reads were then mapped to the North American *lyrata* reference genome (v.2³⁷) using bwa-mem in the default paired-end mode⁴⁸. Indels were realigned using the Genome Analysis Toolkit (GATK) IndelRealigner. Prior to variant discovery, we excluded individuals that had fewer than 40% of bases <8× coverage (assessed via GATK's DepthOfCoverage with the restriction to a minimum base quality of 25 and a minimum mapping quality of 25). Our final dataset for analysis contained 92 individuals.

Variant calling was performed using the GATK HaplotypeCaller (`--min_base_quality_score 25 --min_mapping_quality_score 25 -rf DuplicateRead -rf BadMate -rf BadCigar -ERC BP_RESOLUTION -variant_index_type LINEAR -variant_index_parameter 128000 --pcrindel_model NONE`), followed by GenotypeGVCFs for genotyping. 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 GVCf and then ran GenotypeGVCFs to jointly genotype these GVCf, 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 be biallelic. Additional quality filtering was performed using the GATK VariantFiltration tool (`QD < 2, MQ < 40.00, FS > 60.0, SOR > 4.0, MQRankSum < - 8.0, ReadPosRankSum < - 8.0, DP < 8`). Then we masked sites that had excess read depth, which we defined as 1.6× the second mode (with the first mode being heterozygous deletions or mismapping) of the read depth distribution.

Population structure. All analyses dedicated to reveal population structure and demography were based on putatively neutral fourfold degenerate (4dg) single-nucleotide polymorphisms (SNPs) only. We used the 4dg filter generated for *arenosa* from ref. 9. After quality filtering, these analyses were based on a genome-wide dataset consisting of 4,380,806 4dg SNPs, allowing for a maximum of 10% missing alleles per site (1.2% missing data) at a 5× coverage minimum for a given individual sample.

Although we expected fastSTRUCTURE⁴⁹ to be superior in recognising admixture compared with STRUCTURE⁵⁰, running fastSTRUCTURE on our dataset resulted in poor performance, in that the result did not coincide with the STRUCTURE results or other analysis. This misbehaviour was probably due to the inclusion of polyploid data, as fastSTRUCTURE does not accommodate polyploid genotypes. We had randomly subsampled two alleles per each tetraploid site, similar to ref. 32, using a custom script. However, evidently such a subsampling strategy dissolves the fine-scale differences in admixture between populations at this scale. Hence, STRUCTURE was preferred, and was run on all samples and both ploidy levels. As STRUCTURE accepts only uniform ploidy as input, with one row per each ploidy, we added two rows of missing data for our diploid samples, making them pseudo-tetraploid. In addition, input data were LD-pruned and singletons removed using custom scripts. Window size was set to 500 with a distance of 1000 between windows, allowing for 10% missing data, which resulted in a dataset of 32,256 SNPs genome wide. We performed ten pruning replicates

using the admixture model with uncorrelated allele frequencies, and then ran each for K -values 2–10 with a burn-in period of 50,000 and 500,000 Markov Chain Monte Carlo (MCMC) replicates. We conducted PCA using the `glPca` function in the `adeigenet R` package⁵¹.

Demographic parameters and reconstruction of gene flow. We next performed demographic analyses with `fastsimcoal2`²³ on 4dg sites. A minimum of two individuals per each population was required. Custom python scripts (`FSC2input.py` at <https://github.com/pmonnahan/ScanTools/>) were used to obtain the multi-dimensional allele frequency (DSFS) spectrum as well as bootstrap replicates of the DSFS for confidence interval estimation. For the bootstrap replicates, the genome was divided into 50 kb segments and segments were resampled with replacement until recreating a DSFS of equivalent size as the genome. Ultimately, we aimed to estimate demographic parameters and confidence intervals for a four-population tree corresponding to diploid and tetraploid *lyrata* and *arenosa*. For computational efficiency, three-population trees were initially used to establish the presence/absence of migration edges by comparing models with a single migration edge to a null model with no migration. Additional migration edges would then be added and compared with the initial simple model. For each model, 50 replicates were performed and values kept for the replicate with the highest likelihood. For each replicate, we allowed for 40 optimisation cycles and 100,000 simulations in each step of each cycle for estimation of the expected site frequency spectrum. Although the above process identified the key migration edges, it resulted in a four-population tree that was overly complex; the exercise suggested six migration edges in total (Supplementary Fig. 3). Overfitting was evidenced by highly imprecise and nonsensical estimates for a subset of parameters (Supplementary Table 2). For example, the ancestral population size for *lyrata* was estimated to be greater than 5 million with individual replicate estimates ranging from <100,000 to over 10 million. Estimates for population fusion times were also drastically greater than observed in previous three-population trees. We therefore opted for a simpler model, retaining only the two migration edges with the highest support: bidirectional migration between tetraploids. Parameter estimates for each of the 100 bootstrap replicates were obtained using the scheme described above, and 95% confidence intervals were calculated using the 2.5th and 97.5th percentiles of the resulting distribution of each parameter.

Changes in effective population size over time. PSMC model v.0.6.4 was used to infer changes in effective population size (N_e) through time using information from whole-genome sequences of *lyrata* and *arenosa* diploids²⁷. We generated plots of the most deeply sequenced representative of each of the diploid *lyrata* and *arenosa* populations, with the exception of distinct *arenosa* KZL and SZL. A consensus fastq sequence was created using `samttools v.1.2` and `bcftools v.1.2` using `samttools mpileup -C50 -Q 30 -q 30` with the *lyrata* v.2 genome as the reference. The reference was masked at all sites at which read depth was more than twice the average read depth across the genome. `samttools mpileup` was followed with `bcftools call -c and vcftools.pl vcf2fq -d 5 -D 34 -Q 30` to create a fastq reference file. Using PSMC, this was changed to a format that was required with PSMC by `fq2psmcfa -q20`, and `psmc` was run with parameters `psmc -N25 -t15 -r5 -p "4 + 25*2 + 4 + 6"` and `psmc_plot.pl -R -g 2 -u 3.7e-8` to get a text file that could be plotted with `R`. We used the mutation rate estimate $\mu = 3.7 \times 10^{-8}$ and a generation time of 2 years for both species, as *arenosa* is mainly biennial, and we estimate that *lyrata* generates the highest number of propagules in its second year after germination (R.S., personal observation).

Cytological assessment of meiotic stability. Individual tetraploid *lyrata* and *arenosa* plants were germinated in 7 cm pots with Levington® Advance Seed and Modular Compost Plus Sand soil with 16 h light/8 h dark cycles at 20 °C constant temperature. Once rosettes had formed, plants were vernalised for six weeks with 8 h light (6 °C)/16 h dark (4 °C) cycles. Plants were then grown in 16 h light (13 °C)/8 h dark (6 °C) cycles to encourage flowering. Buds were collected, fixed and anthers dissected for basic cytology as described in⁵² except that 50 mg (30 Gelatine Digestive Units) Zygote® Bromelain were added to the enzyme mixture, and incubation time was increased to 75 min. The prepared slides were stained and mounted with 7 μ l 4',6-diamidino-2-phenylindole (DAPI) in Vectashield (Vector Lab) and metaphase I chromosomes visualised using a Nikon 90i Eclipse fluorescent microscope with NIS Elements software. Chromosome spreads with all rod and/or ring bivalents were scored as stable meiosis (Fig. 1d), whereas multivalents with multiple chiasmata were scored as unstable meiosis (Fig. 1e). FISH was performed as in⁵², except 62 °C was used as the chromosome denaturing temperature. The 5S rDNA probe was generated by directly incorporating biotin into a PCR product (Jenna Biosciences) using primers 5SF 5'-AACCGAAATTGCGTGCA TAG-3' and 5SR 5'-AAACGGGAGGTGAGAGGAG-3' with *Mimulus guttatus* cloned genomic DNA that shares 96% nucleotide identity with *A. lyrata* in this region and the 45S pTa71 clone (Gerlach and Bedbrook, 1979) by nick translation with digoxigenin (Jenna Biosciences). Streptavidin Dylight 594 and anti-digoxigenin Dylight 488 (Vector laboratories) were used as secondary fluorophores. Chromosomes were stained with DAPI in Vectashield (Vector Laboratories).

Differentiation scans for signatures of selective sweeps. We grouped populations by ploidy level, species or hybrid affiliation, and affiliation to a biogeographic region in case of tetraploid *lyrata*. We calculated the following metrics in adjacent nonoverlapping genomic windows: AFD, d_{XY} , F_{ST} , R_{ho} ³⁰ and the number of fixed differences between the *lyrata* diploids and the two *lyrata* tetraploid groups (*Let* and *Lwt*). We identified selective sweep candidates as the 1% outliers of the empirical distribution for each metric. To maximise our chances of capturing differentiation truly related to ploidy and not local adaptation, we selected the overlap between these two independent scans wherein the tetraploids contrast by edaphic (soil) preference and then focused on outliers that were identified in a highly stringent genome scanning approach in *arenosa*¹³.

To obtain insight into differentiation between population groups, AFD, d_{XY} , F_{ST} , R_{ho} , and the number of fixed differences were calculated for additional populations. *Arenosa* populations were grouped by lineage, as identified in refs. 9,21, as *arenosa* Hercynian tetraploids (*Aht*) and *arenosa* Alpine tetraploids (*Aat*), which also corresponds to biogeographic groupings.

GO enrichment analysis. We performed gene function enrichment tests for each contrast using the CLUEGO app version 2.5.4 in CYTOSCAPE version 3.7.2 using GO information associated with orthologous *A. thaliana* gene identifiers. We retained levels 3–8 for biological process (Benjamini-Hochberg correction $p \leq 0.05$).

Visualisation of allele frequencies. We visualised allele frequencies of amino acid substitutions in form of a heatmap. Pre-processed VCF files were annotated using `SnPEff`⁵⁴ (10.4161%2Ff) with the manually added *lyrata* v.2 reference annotation⁵⁵ (10.1371/journal). Variants annotated as missense (i.e. amino acid substitutions) were extracted using `SnPSift`⁵⁴. Gene-coding loci were extracted from the whole-genome annotated VCF and per-population allele frequencies for each amino acid substitution calculated using GATK's `SelectVariants`. Alternative allele frequencies (polarised against the *lyrata* reference) were visualised using the `heatmap.2` function in the `gplots` package in `R` (Warnes et al., 2016, <https://CRAN.R-project.org/package=gplots>).

Identification of differentiated and introgressed regions. To investigate how the relationships among diploid and tetraploid populations of the two species vary across the genome, we used topology weighting by iterative sampling of subtrees (*Twisst*)⁴⁰ [www.github.com/simonmartin/twisst/]. *Twisst* provides a quantitative measure of the relationships among a set of taxa when each taxon is represented by multiple individuals and the taxa are not necessarily reciprocally monophyletic. This provides a naive means to detect both introgression and ILS, and how these vary across the genome. We first inferred genealogies for 50 SNP windows across the whole genome using the BIONJ algorithm⁵⁶ as implemented in `PHYML`⁵⁷. As each individual carries two (for diploids) or four (for tetraploids) distinct haplotypes that represent different tips in the genealogy, it is necessary to first separate the haplotypes by phasing heterozygous genotypes. We used a heuristic approach to estimate phase that maximises the average extent of LD among all pairs of polymorphic sites in the window. This approach iteratively selects the best genotype configuration for each site, beginning with the site that has the most heterozygous genotypes. At each step, the optimal configuration is that which maximises the average LD between the target site and all previous target sites. This allows simultaneous phasing of diploids and tetraploids. We investigated the accuracy of this phasing approach using simulated sequences generated using the coalescent simulator `msms`⁵⁸ and `seq-gen`⁵⁹, following⁴⁰, but here adding steps to randomise phase and then apply phase inference. As *Twisst* is robust to within-taxon phasing errors⁴⁰, the relevant question here is the extent to which imperfect phasing would affect the estimated topology weightings. We therefore applied *Twisst* to the simulated data and compared the results with (i) perfect phase, (ii) randomised phase and (iii) randomised and then inferred phase. This confirmed that our heuristic phasing algorithm led to an improvement in the accuracy of the weightings, giving results that approached what is achieved with perfect phase information.

For running *Twisst* on the empirical data, we combined samples into four ingroup populations: diploid *lyrata*, tetraploid *lyrata*, diploid *arenosa* and tetraploid *arenosa*, and included *A. halleri* as outgroup. These five taxa give fifteen possible rooted taxon topologies (Fig. 3a). Although *Twisst* does not consider rooting when computing topology weightings, the inclusion of an outgroup improves the interpretation of the results, allowing the direction of introgression to be inferred in some cases⁴⁰. In all analyses, topology weightings were computed exactly for all window trees that could be simplified to $\leq 2,000$ remaining haplotype combinations (see ref. 40 for details). In cases where this was not possible, approximate weightings were computed by randomly sampling combinations of haplotypes until the 95% binomial confidence interval for all fifteen topology weightings was below 0.05. Confidence intervals were computed using the Wilson method implemented in the package `binom` in `R` (R Core Team 2015).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. The datasets generated and analysed during the current study are available from the corresponding author upon request. All sequence data are freely available in the European Nucleotide Archive through accession code PRJEB34247. The source data underlying Figs. 1D, 1E, and 3 are provided as a Source Data file.

Code availability

Custom programmes and scripts used in this study are available at GitHub: <https://github.com/pmonnahan/ScanTools/> and <https://github.com/simonmartin/twist/>.

Received: 1 May 2019; Accepted: 24 October 2019;

Published online: 18 November 2019

References

- Abbott, R. et al. Hybridization and speciation. *J. Evolution Biol.* **26**, 229–246 (2013).
- Selmecki, A. M. et al. Polyploidy can drive rapid adaptation in yeast. *Nature* **519**, 349–351 (2015).
- Doyle, J. J. & Coate, J. E. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.* **180**, 1–52 (2019).
- Cui, L. et al. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
- Schmickl, R., Marburger, S., Bray, S. & Yant, L. Hybrids and horizontal transfer: Introgression allows adaptive allele discovery. *J. Exp. Bot.* **68**, 5453–5470 (2017).
- Mallet, J., Besansky, N. & Hahn, M. W. How reticulated are species? *Bioessays* **38**, 140–149 (2015).
- Yant, L. & Bomblies, K. Genomic studies of adaptive evolution in outcrossing *Arabidopsis* species. *Curr. Opin. Plant Biol.* **36**, 9–14 (2017).
- Schmickl, R. & Koch, M. A. *Arabidopsis* hybrid speciation processes. *Proc. Natl Acad. Sci. USA* **108**, 14192–14197 (2011).
- Arnold, B., Kim, S. T. & Bomblies, K. Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. *Mol. Biol. Evol.* **32**, 1382–1395 (2015).
- Kolář, F. et al. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model *Arabidopsis arenosa*. *Mol. Ecol.* **25**, 3929–3949 (2016).
- Baduel, P., Hunter, B., Yeola, S. & Bomblies, K. Genetic basis and evolution of rapid cycling in railway populations of tetraploid *Arabidopsis arenosa*. *PLoS Genet.* **14**, e1007510–e1007526 (2018).
- Hollister, J. D. et al. Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* **8**, e1003093 (2012).
- Yant, L. et al. Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr. Biol.* **23**, 2151–2156 (2013).
- Bomblies, K. & Madlung, A. Polyploidy in the *Arabidopsis* genus. *Chromosome Res.* **22**, 117–134 (2014).
- Bomblies, K., Higgins, J. D. & Yant, L. Meiosis evolves: Adaptation to external and internal environments. *New Phytol.* **208**, 306–323 (2015).
- Clauss, M. J. & Mitchell-Olds, T. Population genetic structure of *Arabidopsis lyrata* in Europe. *Mol. Ecol.* **15**, 2753–2766 (2006).
- Ross-Ibarra, J. et al. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* **3**, e2411 (2008).
- Ansell, S. W. et al. Population structure and historical biogeography of European *Arabidopsis lyrata*. *Heredity* **105**, 543–553 (2010).
- Jørgensen, M. H., Ehrlich, D., Schmickl, R., Koch, M. A. & Brysting, A. K. Interspecific and interloidal gene flow in Central European *Arabidopsis* (Brassicaceae). *BMC Evol. Biol.* **11**, 346 (2011).
- Lafon-Placet, C. L. & Köhler, C. Endosperm-based postzygotic hybridization barriers: developmental mechanisms and evolutionary drivers. *Mol. Ecol.* **25**, 2620–2629 (2016).
- Monnahan, P. et al. Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nat. Ecol. Evol.* **3**, 1–15 (2019).
- Hohmann, N. & Koch, M. A. An *Arabidopsis* introgression zone studied at high spatio-temporal resolution: interglacial and multiple genetic contact exemplified using whole nuclear and plastid genomes. *BMC Genomics* **18**, 1–18 (2017).
- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
- Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. A time-calibrated road map of brassicaceae species radiation and evolutionary history. *Plant Cell* **27**, 2770–2784 (2015).
- Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **107**, 18724–18728 (2010).
- Ehlers, J., Gibbard, P. L. & Hughes, P. D. *Quaternary Glaciations and Chronology. Past Glacial Environments* Chapter 4, 75–102 (Elsevier Ltd, 2017). <https://doi.org/10.1016/B978-0-08-100524-8.00003-8>.
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
- Weir, B. S. *Genetic Data Analysis II* (International Biometric Society, 1997). <https://doi.org/10.2307/2533134>.
- Ronfort, J., Jenczewski, E., Bataillon, T. & Rousset, F. Analysis of population structure in autotetraploid species. *Genetics* **150**, 921–930 (1998).
- Arnold, B. J. et al. Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl Acad. Sci. USA* **113**, 8320–8325 (2016).
- Novikova, P. Y. et al. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).
- Barow, M. Endopolyploidy in seed plants. *Bioessays* **28**, 271–281 (2006).
- Breuer, C., Braidwood, L. & Sugimoto, K. Endocycling in the path of plant development. *Curr. Opin. Plant Biol.* **17**, 78–85 (2014).
- Scholes, D. R. & Paige, K. N. Plasticity in ploidy: a generalized response to stress. *Trends Plant Sci.* **20**, 165–175 (2015).
- Albertin, W. et al. Autopolyploidy in cabbage (*Brassica oleracea* L.) does not alter significantly the proteomes of green tissues. *Proteomics* **5**, 2131–2139 (2005).
- Stupar, R. M. et al. Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics* **176**, 2055–2067 (2007).
- del Pozo, J. C. & Ramirez-Parra, E. Deciphering the molecular bases for drought tolerance in *Arabidopsis* autotetraploids. *Plant, Cell Environ.* **37**, 2722–2737 (2014).
- Coate, J. E. & Doyle, J. J. Variation in transcriptome size: are we getting the message? *Chromosoma* **124**, 27–43 (2014).
- Martin, S. H. & Van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429–438 (2017).
- Christe, C. et al. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol. Ecol.* **26**, 59–76 (2016).
- Huerta-Sánchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
- Hamilton, J. A., la Torre, De, A. R. & Aitken, S. N. Fine-scale environmental variation contributes to introgression in a three-species spruce hybrid complex. *Tree Genet. Genomes* **11**, 95–114 (2014).
- Ronfort, J. The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. *Genet. Res.* **74**, 31–42 (1999).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
- Bolger, A., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Hu, T. T. et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv (2013) arXiv:1303.3997
- Raj, A., Stephens, M. & Pritchard, J. K. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
- Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208–R215 (2010).
- Jombart, T. & Ahmed, I. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
- Higgins, J. D., Wright, K. M., Bomblies, K. & Franklin, F. C. H. Cytological techniques to analyze meiosis in *Arabidopsis arenosa* for investigating adaptation to polyploidy. *Front. Plant Sci.* **4**, 546 (2014).
- Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
- Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
- Rawat, V. et al. Improving the annotation of *Arabidopsis lyrata* using RNA-Seq data. *PLoS ONE* **10**, e0137391–12 (2015).
- Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).
- Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

58. Ewing, G. & Hermisson, J. MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065 (2010).
59. Rambaut, A. & Grassly, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Appl. Biosci.* **CABIOS** **13**, 235–238 (1997).

Acknowledgements

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [grant number ERC-SiG 679056 HOTSPOT], via a grant to L.Y.; and the Biotechnology and Biological Sciences Research Council [grant number BB/P013511/1], via a grant to the John Innes Centre (L.Y.). JDH was funded via BBSRC New Investigator grant BB/M01973X/1. Additional support was provided by the Charles University Grant Agency (GAUK 228716 to M.B.). Computational resources were partly provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, under the programme Projects of Large Research, Development, and Innovations Infrastructures. The authors thank Jeff Doyle for critical reading of the manuscript.

Author contributions

L.Y. and R.S. conceived the study. S.M., P.M., P.J.S., S.H.M., J.K., P.P. and M.B. performed analyses with input from L.Y., R.S. and J.D.H. P.J.S. and S.M. performed laboratory experiments. L.Y., R.S. and S.M. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-019-13159-5>.

Correspondence and requests for materials should be addressed to R.S. or L.Y.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Case study 4.

De-novo mutation and rapid protein (co-)evolution during meiotic adaptation in *Arabidopsis arenosa*.



De Novo Mutation and Rapid Protein (Co-)evolution during Meiotic Adaptation in *Arabidopsis arenosa*

Magdalena Bohutínská,^{1,2} Vinzenz Handrick,³ Levi Yant,³ Roswitha Schmickl,^{1,2} Filip Kolář,^{1,2,4} Kirsten Bomblies,^{*,3,5} and Pirita Paajanen ^{*,3}

¹Department of Botany, Faculty of Science, Charles University, Prague, Czech Republic

²Institute of Botany of the Czech Academy of Sciences, Průhonice, Czech Republic

³Department of Cell and Developmental Biology, John Innes Centre, Norwich, United Kingdom

⁴Department of Botany, University of Innsbruck, Innsbruck, Austria

⁵Plant Evolutionary Genetics, Department of Molecular Plant Biology, ETH Zürich, Zurich, Switzerland

*Corresponding authors: E-mails: kirsten.bomblies@biol.ethz.ch; pirita.paajanen@jic.ac.uk.

Associate editor: Jian Lu

Abstract

A sudden shift in environment or cellular context necessitates rapid adaptation. A dramatic example is genome duplication, which leads to polyploidy. In such situations, the waiting time for new mutations might be prohibitive; theoretical and empirical studies suggest that rapid adaptation will largely rely on standing variation already present in source populations. Here, we investigate the evolution of meiosis proteins in *Arabidopsis arenosa*, some of which were previously implicated in adaptation to polyploidy, and in a diploid, habitat. A striking and unexplained feature of prior results was the large number of amino acid changes in multiple interacting proteins, especially in the relatively young tetraploid. Here, we investigate whether selection on meiosis genes is found in other lineages, how the polyploid may have accumulated so many differences, and whether derived variants were selected from standing variation. We use a range-wide sample of 145 resequenced genomes of diploid and tetraploid *A. arenosa*, with new genome assemblies. We confirmed signals of positive selection in the polyploid and diploid lineages they were previously reported in and find additional meiosis genes with evidence of selection. We show that the polyploid lineage stands out both qualitatively and quantitatively. Compared with diploids, meiosis proteins in the polyploid have more amino acid changes and a higher proportion affecting more strongly conserved sites. We find evidence that in tetraploids, positive selection may have commonly acted on de novo mutations. Several tests provide hints that coevolution, and in some cases, multinucleotide mutations, might contribute to rapid accumulation of changes in meiotic proteins.

Key words: de novo mutations, standing variation, coevolution, meiosis, polyploidy.

Article

Introduction

Sometimes an abrupt change in circumstances forces a rapid evolutionary response. As populations face new challenges, positive selection can act on alleles recruited from standing variation or on de novo mutations (Barrett and Schluter 2008). Though in long-term macroevolution, de novo mutations clearly play a role, evolution from standing variation may be especially important in facilitating rapid adaptation, because it eliminates the waiting time needed for novel mutations (Hermisson and Pennings 2005; Prezeworski et al. 2005; Barrett and Schluter 2008). There are numerous reports of rapid adaptation to novel environments that utilize standing genetic variation (Jones et al. 2012; Van Belleghem et al. 2018; Haenel et al. 2019; Lai et al. 2019), whereas reports of de novo mutations in such instances are rare and often include loss of function mutations (Messer and Petrov 2013; Exposito-Alonso et al. 2018; Wu et al. 2018; Xie et al. 2019). However, it is also predicted that de novo variants may have stronger

phenotypic effects than standing variants (Matuszewski et al. 2015). Thus, the relative importance of de novo mutations may be greater when extensive functional restructuring is needed.

Whole-genome duplication, which leads to polyploidy, is an example of a situation where the cellular context suddenly and substantially shifts, necessitating a rapid adaptive response (Comai 2005; Bomblies and Madlung 2014). Previous studies on the genetic basis of adaptation to genome duplication in the diploid–autotetraploid species, *Arabidopsis arenosa* (fig. 1A), identified a set of genes showing strong evidence of positive selection in its tetraploid lineage (Hollister et al. 2012; Yant et al. 2013; Wright et al. 2015). Many of these genes encode interacting proteins important for meiosis, which is consistent with the fact that meiosis is particularly challenged by genome duplication (Comai 2005; Cifuentes et al. 2010; Stenberg and Saura 2013; Bomblies and Madlung 2014; Bomblies et al. 2016). That at least some of these changes are adaptive is supported by the observation

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

1980

Mol. Biol. Evol. 38(5):1980–1994 doi:10.1093/molbev/msab001 Advance Access publication January 27, 2021

Downloaded from <https://academic.oup.com/mbe/article/38/5/1980/6120800> by Department of Plant Physiology, Faculty of Science, Charles University user on 11 June 2021

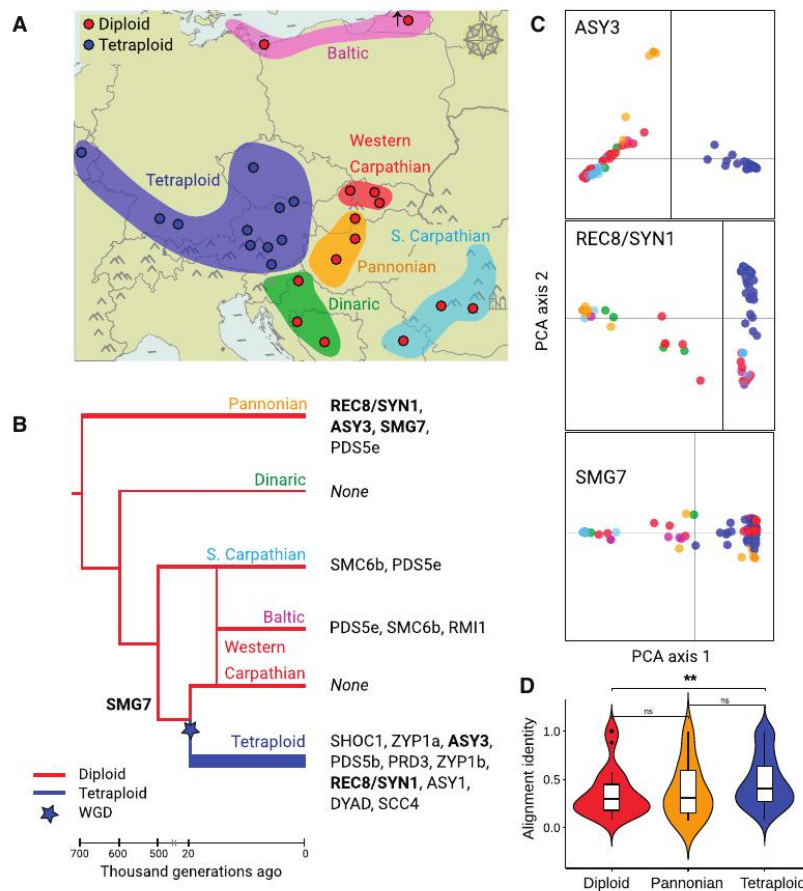


FIG. 1. Meiosis proteins showing signatures of positive selection in *Arabidopsis arenosa* lineages. (A) Our sampling of *A. arenosa* populations in Europe. Dots show 14 diploid (red) and 11 tetraploid populations without signs of introgression from diploids (blue) studied here. Distribution ranges of all known *A. arenosa* lineages are shown as colored areas, indicating that our sampling covers a complete diversity of diploid lineages (based on Kolář et al. 2016; Monnahan et al. 2019). The tetraploid distribution range covers areas occupied by populations without signs of introgression from diploids (Monnahan et al. 2019). (B) Phylogeny of *A. arenosa* (based on Kolář et al. 2016; Monnahan et al. 2019) with candidate meiosis proteins placed on the branch where they exhibit signatures of selective sweeps (identified as F_{ST} and FineMAV overlap, see the main text). Width of the branches corresponds to the number of meiosis proteins that are identified as positive selection candidates. Only Pannonian and tetraploid lineages had more meiosis proteins showing signatures of positive selection than expected by chance. Lineages with no evidence for positive selection on meiosis proteins are indicated as "None." Proteins are ordered from those having the highest number of candidate AASs to the lowest (supplementary tables S5 and S6, Supplementary Material online). Three proteins found independently as candidates in parallel in two lineages are written in bold. Time axis below the tree indicates median estimates of lineage divergence times (based on Arnold et al. 2015; Kolář et al. 2016). (C) Principal component analysis based on allele frequencies of candidate AASs in the three parallel candidate meiosis proteins. Each dot represents one individual, colored based on lineage in panel A. (D) positive selection targeted more conserved amino acids in tetraploids (blue) than in diploids (red; summarizing candidate AASs identified in all but the Pannonian diploid lineage—orange). Each violin plot summarizes alignment identity (calculated across 17 plant reference genomes, higher value indicate more conserved site) over all candidate AASs identified in the corresponding lineage. ** $P = 0.002$, Wilcoxon rank sum test.

that the derived alleles of two of the genes, which encode interacting meiotic axis proteins, have been experimentally shown to affect meiotic traits relevant to tetraploid meiotic stability (Morgan et al. 2020).

Meiosis is a structurally conserved process that is periodically challenged and driven to evolve in diploids as well

(Heyting 1996; Kumar et al. 2010; Grishaeva and Bogdanov 2014; Bomblies et al. 2015; Baker et al. 2017; Brand et al. 2019). But what was striking in the *A. arenosa* tetraploids, and remains unexplained, is that although two independent estimates suggest the tetraploids are likely only about 20,000–30,000 generations old (fig. 1B; Arnold et al. 2015; Monnahan

et al. 2019), a surprisingly large number of amino acid changes differentiates ancestral diploid and derived tetraploid alleles in the subset of meiosis genes that have signatures of positive selection. Meanwhile, the rest of the genome, including other meiosis genes, remains largely undifferentiated (Hollister et al. 2012, Yant et al. 2013). Another study showed that positive selection on meiosis is not unique to the tetraploid *A. arenosa* lineage: Signatures of selection were also found in two of the same meiosis genes (different alleles) in a distinct diploid *A. arenosa* lineage (Wright et al. 2015). This raised the possibility that rapid evolution of meiosis genes might be a common feature of *A. arenosa* lineages regardless of ploidy, and this is one of the ideas we test here.

The above observations leave many questions about the evolution of meiosis in *A. arenosa* lineages unanswered, which also have wider implications for understanding rapid evolutionary adaptation of essential cellular processes. Remaining questions include: Is the evolution of meiosis in the tetraploid lineage more likely to have targeted functionally important sites than in diploids? Were the variants that selection acted on in the tetraploid lineage already present as standing variation in diploids? If not, what might drive the rapid accumulation of multiple amino acid changes in these proteins? To address such questions, we analyzed a range-wide data set of 145 diploid and tetraploid *A. arenosa* genome sequences (fig. 1A; Monnahan et al. 2019), sampling four additional diploid lineages not previously included, complemented with newly generated assemblies for the diploid and tetraploid that allowed us to define haplotypes more reliably. We found that although evidence of positive selection on meiosis proteins is not unique to the tetraploid lineage, the extent of meiotic protein remodeling is. Moreover, we found evidence that selection likely acted at least in part on de novo mutations not present in the diploid gene pool. We also find support for the idea that coevolution of proteins and the accumulation of multinucleotide mutations could contribute to the de novo accumulation of many amino acid variants in the tetraploid lineage.

Results and Discussion

Meiosis Protein Evolution in *A. arenosa* Lineages

We investigated the patterns of evolution of meiosis proteins across all currently known *A. arenosa* lineages (fig. 1A), including samples of four additional diploid lineages in which meiosis protein evolution was not investigated in our previous study (Wright et al. 2015). This additional sampling allowed us to ask whether positive selection commonly targets meiosis in different diploid and tetraploid lineages (i.e., whether selection on meiosis is the rule rather than the exception). This sampling also allowed us to investigate whether the patterns in the tetraploid lineage are qualitatively or quantitatively unusual. We did this using a published data set of single nucleotide polymorphism (SNP) variation that includes range-wide sampling of diploid and tetraploid whole-genome resequenced individuals (Monnahan et al. 2019; see supplementary table S1, Supplementary Material online), complemented with two new genome assemblies of diploid and tetraploid individuals using

the 10× genomics Chromium platform and supernova assembler (Weisenfeld et al. 2017; see supplementary table S2, Supplementary Material online). These new assemblies allowed us to extract diploid- and tetraploid-specific haplotypes for candidate genes (see Materials and Methods for details). We focused on protein sequence evolution, as this allows us to capitalize on the availability of tests that can help assess which changes are likely to be functional.

We first asked whether evidence of selection on meiosis genes is unique to the two lineages, it was previously reported in (the tetraploid and Pannonian diploids; Wright et al. 2015), or is consistently seen across *A. arenosa* lineages (i.e., to ask if this is a ubiquitous feature of meiotic protein evolution). We focused on a list of 78 meiosis-specific proteins (supplementary table S3, Supplementary Material online) selected by refining available lists (Sánchez-Morán et al. 2005; Yant et al. 2013) using the Pathway Interaction Database (PID; Schaefer et al. 2009), AraNet (Lee et al. 2015), and TAIR databases (Berardini et al. 2015). We also confirmed that diploid and tetraploid populations included in our analyses had similar genetic diversity and allele frequency spectra (Monnahan et al. 2019, supplementary table S4, Supplementary Material online), indicating a lack of severe demographic change such as recent population expansions or bottlenecks that could otherwise have had a confounding effect on our analyses.

To identify potential targets of positive selection among the set of 78 meiosis proteins, we first scanned sequences for amino acid substitutions (AASs) between 1) all five previously defined diploid lineages (Kolář et al. 2016; Monnahan et al. 2019: Pannonian, Dinaric, Baltic, Southeastern Carpathian, and Western Carpathian; fig. 1A and B) and 2) comparing all diploid individuals as a group with the tetraploid lineage, using a subsampled data set of 120 individuals to ensure comparable sample sizes across ploidies and lineages (see Materials and Methods for details, supplementary table S1, Supplementary Material online). We identified outlier differentiated AASs as those exceeding the 99% F_{ST} genomewide quantile. We then narrowed this set to those changes predicted to also have functional effects, by selecting the overlap with 1% genomewide outliers identified using the FineMAV method (Szpak et al. 2018; modified to use Grantham and SIFT scores that predict potential functional impact of each AAS, Grantham 1974; Kumar et al. 2009, see Materials and Methods for details). The overlap of F_{ST} and FineMAV outliers identified 56 AAS outliers, in seven meiosis proteins, among the pairwise diploid contrasts, and 171 AAS outliers, in 11 meiosis proteins, in the diploid/tetraploid contrast (below, these are termed “candidate selected AASs” and the proteins they occur in as “candidate selected proteins”; supplementary tables S5–S7, Supplementary Material online). We inferred which are the derived variants of each AAS by comparing with three Arabidopsis outgroup species.

To further test for evidence of positive selection on meiosis genes in the tetraploids, we used McDonald–Kreitman test (McDonald and Kreitman 1991, Smith and Eyre-Walker 2002; see Materials and Methods for details). In this method, we calculated alpha, the proportion of divergences driven by positive selection (supplementary table S8, Supplementary

Material online). Overall, we found evidence of a significantly increased genomewide proportion of divergence values that show evidence of having been driven by positive selection, between diploids and tetraploids of *A. arenosa* ($\alpha = 0.44$, P value < 0.001). Among the candidate meiosis proteins, values of α exceeded the neutrality value of zero in all but three cases (supplementary table S8, Supplementary Material online), the exceptions being ZYP1b, ASY3, and SMG7. For five meiosis proteins (PRD3, ASY1, PDS5b, REC8/SYN1, and DYAD), α estimates exceeded the genomewide value of 0.44 (α between 0.5 and 1, mean = 0.71, P values > 0.05 due to the low number of divergences), suggesting that these proteins evolved under positive selection. In summary, despite the biases that could arise due to the low divergence between the lineages studied here (Monnahan et al. 2019), the results of McDonald–Kreitman test nevertheless support our FineMAV and F_{ST} -scan results, supporting the idea that positive selection targeted meiosis proteins during the divergence of diploids and tetraploids.

When analyzing genomewide patterns, Pannonian diploids and tetraploids both had significant excess proportions of meiosis proteins among all candidate positively selected proteins genomewide ($P = 0.02$ and < 0.001 , respectively, Fisher's exact test, fig. 1B and C, supplementary table S7, Supplementary Material online). This was not the case in any other populations or lineages (supplementary table S7, Supplementary Material online). These results show that signatures of positive selection are only prevalent in the two lineages in which we previously identified them and are not a ubiquitous feature of meiotic protein evolution in *A. arenosa*. In addition to confirming previously identified genes, we identified several new candidate meiotic genes that show evidence of having been under positive selection. We discuss these and their functional implications further in supplementary text 1, Supplementary Material online.

We next wished to test if the candidate-selected AASs are likely to affect conserved or potentially functional sites, and whether this propensity differs among lineages. To do this, we first estimated the potential constraint on particular amino acids by calculating pairwise amino acid identity at all candidate-selected AAS sites across the proteomes of 17 Malvaceae species with sequenced genomes available (see Materials and Methods). In tetraploids, AASs differentiated from diploids were significantly more likely to affect amino acids that are conserved across plant evolution than AASs that show differentiation among the different diploid lineages (P value = 0.002, Wilcoxon rank sum test, fig. 1D and supplementary text 2, supplementary fig. S1, Supplementary Material online). Even though multiple meiosis genes also show evidence of positive selection in the Pannonian diploid lineage, in contrast to the tetraploids, this lineage does not differ significantly from other diploid lineages in the proportion of differentiated AASs in meiosis genes that affect conserved sites (P value > 0.05 , $n = 56$, Wilcoxon rank sum test, fig. 1C). We also found that the differentiated AASs in tetraploids are predicted to cause secondary protein structure variation (supplementary fig. S2 and supplementary text 3, Supplementary Material online). This is interesting in light of

the evidence that 3D structures of meiosis proteins are strongly conserved across even wide evolutionary distances, though the underlying primary sequences can vary substantially even among closely related species (Grishaeva and Bogdanov 2014; Rosenberg and Corbett 2015). These results suggest that tetraploids have both a higher total number of candidate-selected AASs and show evidence that positive selection also targeted more conserved amino acids. This observation supports the hypothesis that greater functional readjustment occurred in the meiotic machinery in the tetraploids than in the diploids.

Positive Selection in the Tetraploids Acted at Least in Part on De Novo Mutations

The high number of potentially functional amino acid changes in multiple interacting proteins in the tetraploids is striking given their relatively recent origin. We thus hypothesized that at least some of the candidate-selected alleles were likely selected from standing variation that existed in diploids. To explore this, we first examined standing variation present in diploids for amino acid changes that characterize tetraploid alleles. We analyzed 10 of the 11 meiosis proteins that show evidence of positive selection in tetraploids (one, ZYP1a with 26 candidate AASs, was removed due to poor mapping of the gene to the reference genome). We analyzed the full available data set of 105 individuals from 14 genome-resequenced diploid populations (including 23 individuals from the Western Carpathian lineage, the most closely related diploids to the tetraploids, fig. 1A and B; Arnold et al. 2015; Monnahan et al. 2019). A rarefaction analysis of *A. arenosa* diploids implied that such sampling is sufficient to converge on the full diploid diversity (supplementary fig. S3, Supplementary Material online).

We found that 63% of tetraploid differentiated AASs were not present in any of the diploid individuals sampled (71 out of 113 AASs; however, we note that this is likely an overestimate of the proportion of amino acids absent from the standing variation as some of the variants might be too rare to be sampled, or may have been originally present in diploids, but went extinct after the divergence of the tetraploids). The remaining 42 ploidy-differentiated AASs were found in our diploid samples, indicating a contribution from standing variation. Most of the “standing” AASs occurred in three proteins (PRD3, ZYP1b, and SHOC1; supplementary fig. S4, Supplementary Material online). An additional 32 candidate-selected AASs, not included in the 113 AASs above, showed parallel differentiation in tetraploids and Pannonian diploids (in proteins SMG7, ASY3, and REC8/SYN1, fig. 1B and C). Whether this pattern is due to incomplete lineage sorting or gene flow between these lineages is not clear.

Since all proteins with evidence of positive selection contain multiple highly differentiated amino acid polymorphisms, we asked if there are instances where full haplotypes of linked candidate-selected AASs exist as standing variation in diploids. We reconstructed the most likely haplotypes across tetraploid-differentiated AASs (supplementary table S6, Supplementary Material online) using allele frequency information complemented with haplotype phasing

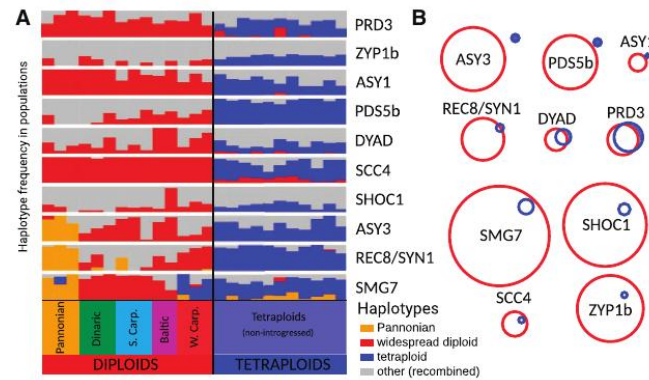


FIG. 2. Limited standing variation across *Arabidopsis arenosa* diploids in protein candidates for tetraploid meiotic adaptation. (A) Lack of tetraploid-specific haplotypes in diploid populations sampled across the total range of *A. arenosa*. Haplotypes were combined across linked candidate AASs within each protein. A set of bar plots for each of ten candidate proteins (horizontal lines) shows frequencies of diploid, Pannonian (if different from widespread diploid) and tetraploid-specific haplotypes (y axis) in each of 14 diploid and 11 tetraploid populations (x axis, grouped to lineages and ploidies). Frequencies of minor frequency haplotypes found in either or both ploidies are summed in a gray column. (B) A hypothetical maximal variation among haplotypes of meiosis proteins in diploids and tetraploids, quantified by Hamming distances. The diameter of the red and blue circles denotes the full range of potential variability of haplotypes reconstructed by all combinations of AASs among all diploid and tetraploid individuals, respectively. The relative distance of the red and blue circles denotes the genetic distance between the diploid and tetraploid haplotypes. Overlap of both circles suggests that it is plausible that the tetraploid haplotype could have existed within the observed variation in diploids, even if the exact tetraploid haplotype was not found in our diploid sampling. Filled area of the tetraploid circle, non-overlapping with diploid, represents the tetraploid haplotype space that cannot be explained by, and would not be expected to exist, within extant diploid AAS variation. The upper six proteins show evidence that their tetraploid haplotypes most likely accumulated additional mutations after diploid/tetraploid divergence.

data in the respective diploid or tetraploid genome assembly (see Materials and Methods). Apart from the SMG7 haplotype, which was found in one population in the Pannonian lineage and in three populations in the Western Carpathian lineage, none of the complete haplotypes predominant in tetraploids were found in any diploid population (fig. 2A). Taken together, these findings suggest that although some AASs in each case likely originated as standing variation in diploid populations, additional de novo changes likely accumulated in each of the meiosis genes to generate the extant tetraploid alleles.

The findings above cannot rule out that the full haplotypes were originally present in diploids, but lost after divergence of the tetraploids, or that they were present, but too rare to have been sampled. Thus, we quantified whether an unsampled haplotype allele as different from other diploid variants as the current tetraploid allele is, could plausibly have existed within the range of variation in our sampled portion of the diploid gene pool. If not, this would suggest that additional amino acids likely accumulated postdivergence. To do this, we compared the Hamming distance (which quantifies the number of sites in which diploid and tetraploid alleles differ in nucleotide sequence) to the Hamming diameter of each gene pool (which is the maximum pairwise distance among alleles within a set, see Materials and Methods, Robinson 2003). If the Hamming distance between diploids and tetraploids is lower than the Hamming diameter within diploids, it is considered plausible that the tetraploid haplotype could have

existed within the diploid pool of genetic variation, even if not sampled. This was the case for four proteins (fig. 2B and supplementary table S9, Supplementary Material online). For six meiotic proteins, however, the tetraploid haplotype was differentiated beyond the diploid variation and thus likely not available within the original pool of standing diploid variation (fig. 2B and supplementary table S9, Supplementary Material online). This includes two meiotic axis proteins (ASY1 and ASY3) whose diploid and tetraploid variants have recently been shown to have distinct functional effects in meiosis (Morgan et al 2020).

The proportion of meiosis proteins with likely de novo changes as identified by Hamming distances was only slightly higher than that of the other candidate-selected proteins genomewide (proportion of de novo candidates = 0.60 and 0.52 for meiosis proteins and other proteins genomewide, respectively), suggesting that selection on de novo mutations might be a general feature of positive selection in polyploids. However, meiosis proteins do show an excess relative to other proteins of “de novo” candidate-selected AASs per protein (11.1 for meiosis proteins and 5.4 for other proteins genomewide; $P = 0.001$, Wilcoxon rank sum test), suggesting that meiosis as a process underwent more extensive de novo restructuring than most other processes that show evidence of having been under positive selection in the tetraploid genome.

The above analysis cannot completely rule out allele extinction. However, we note that selection from standing

variation followed by allele extinction at multiple independent loci in diploids is not the most parsimonious explanation. We would have to imagine that, six times independently, a standing variant that is more different than any other allele sampled from the present gene pool came under positive selection in the tetraploids and was subsequently lost in diploids. Thus, we believe that although some amino acids characteristic of tetraploid alleles do come from standing variation a considerable fraction of the observed differences accumulated de novo in the tetraploid lineage after divergence.

The Accumulation of Amino Acid Changes in the Tetraploids

Given that positive selection predominantly from standing variation is an unlikely explanation for the pattern of amino acid divergence in tetraploids, we explored whether rapid protein evolution might be driven by compensatory evolution and coevolution, as previously proposed for autotetraploid *A. arenosa* (Hollister et al. 2012). Compensatory coevolution of interacting proteins can speed the accumulation of novel changes because if a change in one protein causes even a subtle shift in structure or stability, this will lead to selection for compensatory mutations that return the structure or stability of the protein, or an entire complex, to its optimal state (DePristo et al. 2005; Szamecz et al. 2014; Rojas Echenique et al. 2019). Because compensatory mutations have a large mutational target, as any number of amino acid changes can readjust the stability or shape of a protein, they can accumulate rapidly relative to changes that must target particular functional sites (DePristo et al. 2005; Szamecz et al. 2014). Empirical data support this idea, for example, work in bacteria has shown that this kind of compensatory evolution can lead to the rapid accumulation of AASs in groups of interacting proteins (Moura de Sousa et al. 2017). Since meiotic proteins are well known to interact (e.g., Zickler and Kleckner 1999), compensatory evolution and coevolution might be one cause of rapid evolution of amino acid changes (Maisnier-Patin et al. 2002; Davis et al. 2009). Thus, we asked if a process of protein coevolution might have promoted the extensive accumulation of de novo amino acid changes in tetraploids.

We found hints in our data that support the idea that compensatory evolution may contribute to the observed differentiation. First, all six proteins that likely accumulated multiple de novo amino acid changes after divergence of the tetraploids and diploids, are interacting cohesin and axis components, suggesting that changes in one could plausibly affect essential interactions with the others (fig. 3A). Second, we examined the relative ages of the selective sweeps (i.e., the likely order in which the tetraploid alleles of the six proteins rose in frequency). Under a coevolution scenario, we might expect positive selection to have acted sequentially on the different cointeracting proteins, rather than all alleles having been targeted at the same time, or that selection acted episodically on each protein as changes occurred in its partners. We estimated the relative sweep age as a ratio of number of SNPs accumulated in the selected haplotype, and its length. For each meiosis protein we counted the number of

polymorphisms normalized to the length of the haplotype between first and last candidate positively selected AAS as a proxy for sweep age. The oldest sweeps were inferred to have occurred in PRD3 and REC8/SYN1, followed by ASY1 and PDS5b, with ASY3 and DYAD being the youngest (fig. 3A and B and supplementary table S10, Supplementary Material online). Age estimates of this sort are error prone (Messer and Neher 2012; Ormond et al. 2016; Smith et al. 2018), but the potentially staggered origin of selected alleles hints that changes in one may have provided a context that favored changes in another (e.g., positive epistasis; Pedruzzi et al. 2018).

We also searched for hints of mechanistic evidence of coevolution, for example, predicted structural differences in binding sites of the candidate proteins. We did this using our diploid and tetraploid genome assemblies, for the subset of proteins with known structures: the cohesin subunit REC8/SYN1, the cohesin regulator PDS5b, and the meiotic axis components ASY1 and ASY3, together with the cohesin component SCC3 (which does not show strong evidence of selection, but where we identified a medium-frequency premature stop codon in tetraploids, supplementary text 2, Supplementary Material online). Using PSIPRED secondary structure predictions, which calculate which of the three local amino acid interactions, helix, sheet or coil elements, are most likely for each position in the amino acid chain, we found clusters of predicted structural changes in the interaction surfaces of REC8/SYN1 and SCC3 and to lesser degree of PDS5b and REC8/SYN1. This finding suggests that these proteins may be coevolving (fig. 3C). Whether the structural changes generate novel interaction dynamics, or preserve ancestral ones in the face of other functional changes to the cohesin complex, remains to be tested. Though they are not definitive, the above tests for the expected coevolution of the candidate meiosis proteins are consistent with the idea that coevolution of interacting proteins might indeed have been involved in promoting the accumulation of at least some of the amino acid changes observed.

A potential nonselective explanation for the large number of differentiated AASs in some proteins could be that they arose in single multi-nucleotide mutation (MNM) events, which can give rise to multiple closely linked substitutions in a single instance. A hallmark of MNMs is that substitutions are closely spaced, and commonly also have a significant excess of transversions relative to transitions (Schridder et al. 2011; Harris and Nielsen 2014; Besenbacher et al. 2016). We therefore scanned for these features in genes encoding the candidate-selected meiosis proteins. We found patterns suggestive of MNM events in derived alleles of ASY3 and SMG7, which both had a higher than random proximity of AASs (the median distance = 26 and 46 bp for ASY3 and SMG7, respectively, whereas for other proteins genomewide the distance is 61 bp; $P < 0.01$ in both cases, Wilcoxon rank sum test). Derived alleles in both genes also have a significant excess of transversions relative to transitions compared with genomewide rates ($P < 0.01$, two-sample z test). We observed a similar transversion/transition bias in derived alleles of four other proteins in the tetraploid (REC8/SYN1, ASY1, PRD3,

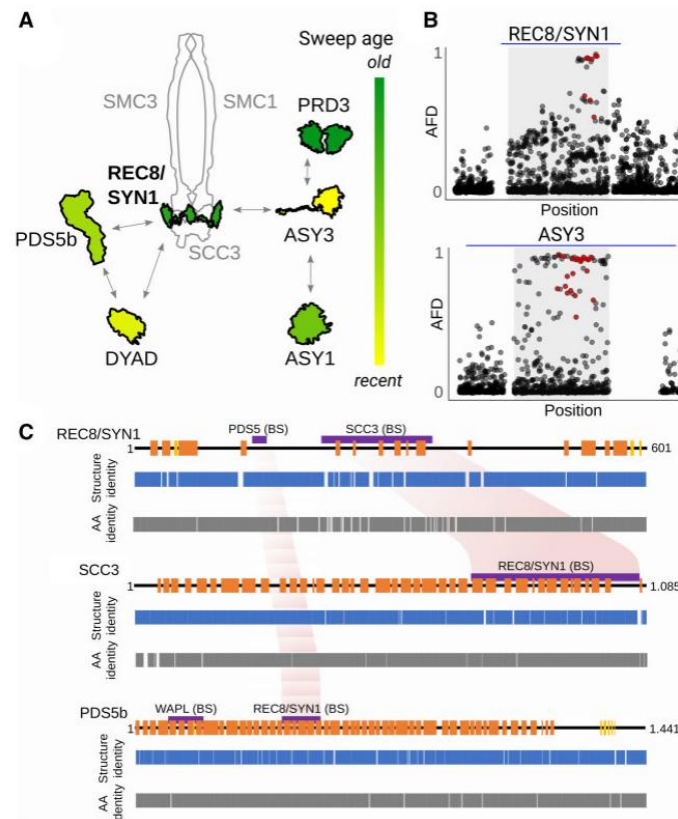


Fig. 3. Evidence for meiosis protein coevolution in tetraploids. (A) Cartoon of the cohesin complex with associated proteins and variability in relative order of their selection sweeps inferred from haplotype length and number of accumulated SNPs (see Materials and Methods for details). Shown are schemes of candidate protein structures (outlined in black) and other core complex protein structures for illustration (gray). We propose that REC8/SYN1 (bolt) might be the core driver of coevolution as it is the central protein with one of the oldest sweeps. (B) Illustrative examples of pattern of allele frequency decay at locus with old (REC8/SYN1) and young (ASY3) selection sweep (as inferred in A). Plotted is AFD between diploid and tetraploid individuals for all genic variants in and around the gene. Red dots are candidate AASs identified here; blue line corresponds to 10 kb. (C) Coordinated structural changes in protein-binding sites. Cartoons of secondary protein structures from diploid *A. arenosa* meiosis proteins (upper lane; in orange = helix elements, in yellow = sheet elements, and black line = disordered protein regions). The pairwise comparison of predicted secondary protein structures from sequences of diploid and tetraploid *A. arenosa* lineages (middle lane, Structure identity plots) and the identity of their amino acid sequences (lower lane, AA identity plots). Gaps are sites with zero identity. Protein-binding sites and functional domains identified in other eukaryotes are shown as violet bars above the secondary structure plot. Reciprocal structure identity changes in corresponding binding sites of REC8/SYN1 and SCC3 and to lesser degree REC8/SYN1- and PDS5b-binding sites might indicate coevolution of these proteins—highlighted in light red.

and ZYP1a) and REC8/SYN1 in the *Pannonian* diploid ($P < 0.01$, two-sample z test), but these latter examples lacked the close spacing of mutations characteristic of MNMs.

Conclusions

Here we investigated the evolution of meiosis proteins in *A. arenosa* using a rangewide sampling of diploid and tetraploid lineages (Monnahan et al. 2019). Since many meiosis proteins are thought to evolve rapidly (Heyting 1996; Kumar et al. 2010; Grishaeva and Bogdanov 2014; Baker et al. 2017; Brand et al. 2019), we reasoned that signals of positive

selection might be common, and therefore not unique to the two *A. arenosa* lineages where they were found previously (Hollister et al. 2012; Yant et al. 2013; Wright et al. 2015). However, we found from sampling four additional lineages that strongly differentiated AASs were found almost exclusively in these two lineages (the tetraploid and the *Pannonian* diploid), suggesting that positive selection on these genes is situational and not ubiquitous. The pattern in the tetraploid lineage is especially striking; despite its tender evolutionary age (~20,000–30,000 generations; Arnold et al. 2015; Monnahan et al. 2019), it has the largest number of proteins with excessive differentiation, the largest number of

differentiated AAs, a higher proportion of AAs in conserved sites, amino acid changes occurring in the interaction surfaces of the meiotic cohesin alpha-kleisin (REC8/SYN1) and its interacting partners, and predicted structural shifts. Thus, the shift in the tetraploid meiotic machinery appears to be far more substantial than what occurred in any diploid lineage of *A. arenosa* including the Pannonian lineage. This fits with the idea that genome duplication is an especially strong challenge for meiosis, likely necessitating a rapid evolutionary response (Bomblies et al. 2015, 2016).

We also find evidence that a considerable proportion of amino acids that are differentiated between the diploid and tetraploid lineages likely arose de novo on alleles that already contained some polymorphisms that preexisted as standing variation. This likely high contribution from de novo variation might come as a surprise, given the theoretical prediction and empirical evidence that rapid evolution is greatly facilitated by the availability of preexisting genetic variation (Jones et al. 2012; Olson-Manning et al. 2012; Ralph and Coop 2015; Van Belleghem et al. 2018; Alves et al. 2019; Haenel et al. 2019; Lai et al. 2019; Oziolor et al. 2019; Thompson et al. 2019). We suggest three nonmutually exclusive explanations for why this might be: 1) Meiosis is a conserved multiprotein process whose need for restructuring after polyploidization (Comai 2005; Bomblies and Madlung 2014) requires variants that are perhaps deleterious in the diploid background. 2) The considerable contribution of novel mutation to rapid adaptation may be a more common feature of autopolyploid evolution, perhaps due to their higher effective population size and/or lower homozygosity (Parisod et al. 2010) or the sudden novel physiological context of polyploids (Doyle and Coate 2019; Bomblies 2020). 3) Empirical literature may be biased toward reports of adaptation from standing variation (Barrett and Schluter 2008), as it is easier to detect a presence of genetic variants than to exclude it.

In summary, our study supports the idea that both standing and de novo variation may be important sources of adaptive variants in multiple interacting meiosis proteins in autotetraploid *A. arenosa*. It will be interesting to see whether this is a particularly prominent feature of polyploid evolution, or a more common pattern for the evolutionary modification of conserved multiprotein processes that occurs when populations must adapt to sudden novel circumstances that challenge these processes.

Materials and Methods

Data Sets

Sampling and Population Genetic Structure of the Genomic Data Set

To study the evolution of meiosis proteins in both diploid and tetraploid *Arabidopsis arenosa* populations, we reanalyzed a rangewide genomic data set previously described in (Monnahan et al. 2019). This data set originally consists of sequences from 15 diploid and 25 tetraploid genome resequenced *A. arenosa* populations (287 individuals, seven individuals per population on average, supplementary table S1, Supplementary Material online). We first aligned the short-

read sequences to the *Arabidopsis lyrata* version 2 (LyV2) reference genome (Hu et al. 2011), called variants and filtered as previously (Monnahan et al. 2019) using the Genome Analysis Toolkit (GATK 3.5 and 3.6, McKenna et al. 2010) and finally called SNPs with GATK HaplotypeCaller. For most analyses described below (except where noted), we used a subset of the full data set consisting of 80 diploid individuals (16 samples with the highest depth of coverage of sequences from each of the five major lineages) and 40 tetraploid individuals from populations unaffected by secondary introgression from diploid lineages (i.e., sampling from C European, Alpine, and Swabian lineages as defined in Monnahan et al. (2019)). Such subsampling gave us a balanced number of 160 high-quality haploid genomes of each ploidy suitable for unbiased scans for positive selection, which was also unaffected by later unidirectional interploidy introgression (supplementary table S1, Supplementary Material online). Finally, we filtered each subsampled data set for genotype read depth >8 and maximum fraction of missing genotypes <0.5 in each lineage to be confident about the variant calling.

We used our total diploid sampling (105 individuals, supplementary table S1, Supplementary Material online) in a separate analysis aimed to screen for standing variation of tetraploid alleles in the total diploid sample (fig. 2 and supplementary fig. S4, Supplementary Material online). This yielded the total number of 145 resequenced individuals used throughout our analyses.

To avoid polarization toward a single reference species genome, we repolarized the variants using a collection of individuals across three closely related diploid *Arabidopsis* species, European *A. lyrata*, *A. croatica*, and *A. halleri*, following procedure described in Monnahan et al. (2019). We further confirmed the repolarization using frequencies of the variants across the data set (considering the minor frequency allele overall as derived).

We calculated genomewide nucleotide diversity (π) and Tajima's D (Tajima 1989) for each lineage, all diploids and all tetraploids using putatively neutral 4-fold degenerate sites. In agreement with the previous study (Monnahan et al. 2019), the per-population genomewide synonymous diversity (π) was similar between ploidies (π values ranging between 0.028 and 0.032 in five diploid lineages, 0.036 for all diploids and 0.034 for tetraploids, supplementary table S4, Supplementary Material online) and total range of Tajima's D over synonymous sites (-0.34 to $+0.34$, supplementary table S4, Supplementary Material online) was far from the accepted threshold of nonneutrality (± 2 ; Tajima 1989). Calculations were performed using python3 ScanTools pipeline (github.com/mbohutinska/ScanTools_ProtEvol), a modification of ScanTools, a toolset specifically designed to analyze diploid–autotetraploid data sets.

Novel Diploid and Tetraploid Genome Assemblies

We created two *A. arenosa* draft reference assemblies, to investigate the haplotypes of meiosis proteins and differences in secondary structure prediction in a sufficient detail, as well as

to remap the areas in the *A. lyrata* genome, where the *A. arenosa* reads did not map well (7 out of the 78 loci, see the next section for details). We assembled genome of one diploid (from *Western Carpathian* population SNO) and one tetraploid individual (population TBG). The diploid assembly is also described in (Liu et al. 2020), but we include it here for completeness.

First, fresh leaf material was sent to Earlham Institute, where DNA was extracted using the BioNano plant protocol from the tetraploid *A. arenosa* and using CTAB DNA extraction protocol from *A. arenosa* diploid (as in Paajanen et al. 2019). Second, to construct the 10× library, DNA material was diluted to 0.5 ng/μl with EB (Qiagen) and checked with a QuBit Fluorometer 2.0 (Invitrogen) using the QuBit dsDNA HS Assay kit. The Chromium User Guide was followed as per the manufacturer's instructions (10× Genomics, CG00043, Rev A). The final library was quantified using quantitative polymerase chain reaction (qPCR, KAPA Library Quant kit [Illumina], ABI Prism qPCR Mix, Kapa Biosystems). Sizing of the library fragments were checked using a Bioanalyzer (High Sensitivity DNA Reagents, Agilent). Samples were pooled based on the molarities calculated using the two QC measurements. The library was clustered at 8 pM with a 1% spike in of PhiX library (Illumina). The pool was run on a HiSeq2500 150 bp Rapid Run V2 mode (Illumina). The following run metrics were applied: Read 1: 250 cycles, Index 1: 8 cycles, Index 2: 0 cycles, and Read 2: 250 cycles.

Sample TBG was sequenced on HiSeq2500 Rapid Run V2 mode (Illumina, on 150-bp sequences). About 58.49 M (121.71 M) reads were created. These were assembled on Supernova 2.0.0 giving raw coverage 27.66× and effective coverage 22.07×. The molecule length was 57.19 kb. The assembly size, counting only scaffolds longer than 10 kb was 58.84 Mb, and the Scaffold N50 was 33.92 kb.

Sample SNO was sequenced on HiSeq2500 Rapid Run V2 mode (Illumina, on 150-bp sequences). About 82.10 M reads were created. These were assembled on Supernova 2.0.0 giving raw coverage 57.91× and effective coverage 45.30×. The molecule length was 26.58 Kb. The assembly size, counting only scaffolds longer than 10 kb was 127.02 Mb and the Scaffold N50 was 2.19 Mb (supplementary table S2, Supplementary Material online).

We analyzed the gene content using BUSCO, and the results showed that the gene space of the diploid *A. arenosa* assembly was nearly complete with 97.5% of the plant specific BUSCOs present and 1.4% missing completely. Of these, 4.7% were duplicate copies.

With the tetraploid *A. arenosa* assembly, we captured 98.5% of the core plant genes and had 1.3% missing. Since the plant is a tetraploid, the rate of duplicate genes was high in the assembly, and total of 82.8% of the core plant genes were found as duplicates. This is not surprising, especially since the plant was from the TBG population that is in the railway lineage and hence shows secondary admixture from a diploid *A. arenosa* lineage (Monnahan et al. 2019). Thus when working with the TBG fragmented assembly, we always checked the variation among all diploid and nonadmixed tetraploid populations for confirmation which of the two

cooccurring haplotypes is dominating our tetraploid sampling.

Detecting Signatures of Positive Selection Acting on Meiosis Proteins

Meiosis Protein Identification, Processing, and Annotation

We annotated each SNP in the genomewide data set and assigned it to a gene using SnpEff 4.3 (Cingolani et al. 2012) and following *A. lyrata* version 2 genome annotation (Rawat et al. 2015). Annotated variants genomewide were extracted from vcf format to table using SnpSift, part of SnpEff 4.3, with flags "CHROM POS REF ALT AC AN 'ANN[*].HGVS_P'" and these tables were used as the basis for the subsequent analysis of positive selection. Next, we identified a list of 78 proteins related to meiosis was based on Yant et al. (2013) and updated by searching PID, AraNet (Probabilistic Functional Gene Network of *A. thaliana*) and *A. thaliana* orthologs in TAIR database (Berardini et al. 2015) and using the list of meiosis proteins from (Sánchez-Morán et al. 2005). ZYP1A, which is not present in the *A. lyrata* version 2 annotation, was added manually based on gene model available from the previous study (Yant et al. 2013). We assigned it with ID AL1G35725 to place it in the correct order into the reference .gff3 file. We further validated that the meiosis genes were expressed in *A. arenosa* using an available RNASeq data set (supplementary text 4, Supplementary Material online).

We found seven meiosis genes (SHOC1, SCC1, SCC2, SCC3, SCC4, MSH4, and SMC6A), where duplicated regions mapped to the same reference loci or where the reads were mis-mapped when aligning to the *A. lyrata* reference (Hu et al. 2011). To overcome this problem, we realigned these loci separately to our own *A. arenosa* diploid reference. To do so, we took the *A. arenosa* reference sequence and found the *A. lyrata* genes in the assembly using bwa 0.7.12 (Li 2013). We extracted 20 kb upstream and downstream from the gene and created a new reference with just these seven genes. Then we mapped the raw reads from each of the 291 samples back to this reference, following the same procedure which we used for mapping to *A. lyrata*. The heterozygosity and coverage of newly remapped genes stayed within the genomewide average. The commands that were used are available at (github.com/paajanen/meiosis_protein_evolution/). We built a separate *A. arenosa* database for these mismapped genes using our *A. arenosa* reference sequence and gff3 files made manually based on *A. lyrata* V2 gff3 using Geneious 11.0.3. The SnpEff analyses then followed the above outlined procedure and the total list of all 78 meiosis genes was analyzed jointly hereafter.

Scans for Positive Selection with Likely Functional Consequences Acting on Meiosis Proteins

To infer candidate AASs within our data set of 78 meiosis genes, highly differentiated between lineages and with likely impact on protein function, we combined a differentiation-based positive selection scan (F_{ST} , Hudson et al. 1992) with genome scanning method accounting for theoretical functional consequence of each AAS (modified FineMAV,

Szpak et al. 2018). Both methods are well suited to infer signatures of recent (within species) positive selection (Oleksyk et al. 2010; Vitti et al. 2013). We used both approaches based on population allele frequencies, allowing joint analysis of diploid and autopolyploid populations. We screened for positive selection 1) among the five diploid lineages (fig. 1A) and 2) between all diploids and tetraploids. We considered only AASs that were outliers in both selection scans as putative positive selection candidates. For these analyses, we worked with six lineages in total, covering a full known distribution range of *A. arenosa* (fig. 1A and B; Kolář et al. 2016; Monnahan et al. 2019): *Pannonian*, *Dinaric*, *Baltic*, *Southeastern Carpathian*, and *Western Carpathian* (diploid lineages, subsampled to 32 chromosomes each) and tetraploid (subsampled to 160 chromosomes and contrasted to the sum of all 160 diploid chromosomes). A reanalysis of diploid–tetraploid selection scans using 16 diploid and 16 tetraploid individuals (comparable with the sample size of diploid) did not yield qualitatively different results. First, for each lineage pair, we calculated F_{ST} for all nonsynonymous SNPs (i.e., AASs) across the 78 meiosis proteins. We used Hudson's F_{ST} estimator, which is suitable for a single variant calculations (Bhatia et al. 2013). Next, we calculated distribution of F_{ST} over all synonymous (i.e., putatively functionally neutral) SNPs genome-wide. We used the 99th quantile of this “neutral” distribution as a threshold for identification of outlier AASs. The neutral synonymous F_{ST} quantiles did not differ significantly from those derived from nonsynonymous SNPs (supplementary table S11, Supplementary Material online, Wilcoxon rank sum test, $W = 69.5$, P value = 0.58, $n = 11$). However, the quantile values were consistently slightly lower for nonsynonymous SNPs (supplementary table S11, Supplementary Material online), making the use of synonymous F_{ST} quantiles more conservative. All calculations were performed using ScanTools_ProtEvol, and custom R scripts (github.com/mbohutinska/ProtEvol/).

Second, we adopted the Fine-Mapping of Adaptive Variation (FineMAV, Szpak et al. 2018) and modified it to fit the resources available for *A. lyrata* reference genome. Specifically, we replaced CADD, the functional score available for human reference (Szpak et al. 2018; Rentzsch et al. 2019), by 1) the Grantham score (Grantham 1974), which is a purely theoretical AAS value, encoded in the Grantham matrix, where each element shows the differences of physicochemical properties between two amino acids and 2) the SIFT annotation score (Kumar et al. 2009), which estimated the effect of amino acid change based on sequence homology across available reference sequences and physical properties of amino acids. To estimate the SIFT scores specifically for our data set, we created a SIFT annotation of our vcf-file using *A. lyrata* database v.1.0.23 from SIFT website (https://sift.bii.a-star.edu.sg/sift4g/, last accessed February 3, 2021). The annotation was done using SIFT4G algorithm (command `java -jar SIFT4G_Annotator_v2.4.jar -c -i input.vcf -d ./Lyrata_db/v.1.0.23/-r annotated`). We rescaled the SIFT score to be 1 when it is most deleterious and 0 when it is most tolerated. Next, we estimated the population genetic component of FineMAV (see Szpak et al. 2018 for details on calculations)

using allele frequency information at each site (considering minor frequency allele as derived) and DAP parameter of 3.5. Finally, for each AAS, we assigned Grantham scores and SIFT scores, together with population genetic component of FineMAV, using a custom scripts in Python 2.7.10 and the Biopython 1.69 package. By rescaling the SIFT scores, we ensured that for both functional score, higher value indicate more likely impact of the AASs to the protein function. Finally, we identified the overlap of top 1% outlier AASs identified in the FineMAV analysis with SIFT scores and with Grantham scores and considered these double outlier AASs as a final candidate identified in FineMAV analysis. All the calculations were performed using code available at (github.com/paajanen/meiosis_protein_evolution).

We note that the SIFT database was developed for *A. lyrata* annotation version 1, and do not contain all meiosis proteins from our list. Thus, we did not obtain any SIFT score for SCC3, MSH4, SMC6A, and ZYP1a and we only considered Grantham scores for them (supplementary tables S5 and S6, Supplementary Material online).

Finally, we controlled for the presence of differentiated indel variants in all candidate meiosis proteins by inspecting their alignment files of the RNA-Seq mapping and screening their gene sequences in the newly generated diploid and tetraploid draft assemblies. We identified only three indel variants differentiated between diploids and tetraploids and neither of them was a frameshift mutation affecting any of our candidate AASs. Thus, the indel variants should not affect the interpretations of our SNP-based selection scans.

Finally, to further assess selection acting on meiotic proteins, we conducted a McDonald–Kreitman test, which is a powerful approach for detecting selection in proteins (McDonald and Kreitman 1991, Smith and Eyre-Walker 2002). We calculated alpha, which quantifies the proportion of divergence driven by positive selection and is defined as $\alpha = 1 - (D_S P_N) / (D_N P_S)$, where D_S and D_N are the numbers synonymous and nonsynonymous substitutions per gene, respectively, and P_S and P_N are the numbers of synonymous and nonsynonymous polymorphisms per gene. The divergence between diploids and tetraploids of *A. arenosa* is too recent to satisfy the assumption of fixation of nucleotide substitutions within species. We thus estimated nucleotide divergence values (D_S , D_N) using the upper 1% outliers of allele frequency differences (AFD) between diploids and tetraploids (upper 1% AFD outlier threshold = 0.53). It has also been suggested that it is important to exclude rare polymorphisms to minimize the impact of slightly deleterious mutations on the estimate of adaptive evolution (Charlesworth and Eyre-Walker 2008). Thus, we excluded variants with overall allele frequency lower than 0.15 (following Fay et al. 2001; Zhang 2005).

Ortholog Search and Analysis of Evolutionary Conservation of Candidate AASs

To examine the tendency of candidate AASs to affect conserved sites, we compared levels of pairwise alignment identity (PAI, mean pairwise identity over all pairs in the

alignment column) of the 78 meiosis protein sequences across the proteomes of 17 Malvaceae reference genomes. To do so, we downloaded *A. lyrata* sequences of the meiosis proteins from Phytozome12.1 database (www.phytozome.jgi.doe.gov, last accessed August 7, 2018) and used as query sequences to identify orthologs of 17 Malvaceae species proteomes. Species included in the search were *Arabidopsis halleri*, *A. thaliana*, *Boechera stricta*, *Capsella grandiflora*, *Capsella rubella*, *Eutrema salsugineum*, *Brassica rapa*, *Brassica oleracea*, *Populus trichocarpa*, *Salix purpurea*, *Theobroma cacao*, *Manihot esculenta*, *Gossypium raimondii*, *Carica papaya*, *Citrus clementina*, *Citrus sinensis*, and *Linum usitatissimum*. We performed searches using the BlastP program in Phytozome with proteome as target type, e-threshold -1 and BLOSUM62 comparison matrix. In case of identification of multiple orthologs (i.e., multiple hits for the same species), only the ortholog with the lowest e -value was considered. The number of sequences in protein alignments ranged 13–17 (16.5 on average, [supplementary table S12, Supplementary Material online](#)). We aligned protein sequences of all identified orthologs using MUSCLE as implemented in Geneious v11 ([Kearse et al. 2012](#)), with default settings (UPGMB clustering method, terminal gaps full penalty, gap open score -1 , window size five). PAI was extracted for each reference (*A. lyrata*) amino acid and we tested the difference in the PAI of diploid and tetraploid candidate AASs sites using Wilcoxon rank sum test (R package stats, [R Core Team 2018](#)).

Distinguishing between Positive Selection on De Novo Mutations and Standing Variation

We used a three-step procedure to distinguish whether positive selection in each candidate meiosis protein likely acted on de novo mutations or standing variation: 1) search for the presence of candidate tetraploid-differentiated AASs across full sampling of individuals from all known diploid lineages of *A. arenosa*, 2) search for the presence of tetraploid-differentiated haplotypes across these diploid individuals, and 3) study of uniqueness of tetraploid haplotypes by comparing their differentiation from diploids to their overall diploid diversity.

In order to conclude that positive selection in a candidate meiosis protein likely acted on de novo variation, we requested that all three of these criteria pointed toward de novo origin in tetraploids; that is, that at least some of its candidate tetraploid-differentiated AASs were not found in any diploid individual, the complete tetraploid haplotype was not found in any diploid individual, and the tetraploid haplotype divergence from the diploid exceeds the overall diploid diversity (diploid–tetraploid Hamming distance exceeding diploid Hamming diameter).

The Presence of Candidate Tetraploid-Differentiated AASs in Diploid Lineages

To identify possible standing variation for the tetraploid alleles, we searched for the presence of each candidate tetraploid-differentiated AASs in diploids. We analyzed the full sampling of all 105 individuals from the 14 diploid

populations, covering all known lineages of *A. arenosa* ([fig. 1A, Kolář et al. 2016; Monnahan et al. 2019](#)). The rarefaction analysis implies that our sample of 105 individuals is sufficient to converge on the true diversity of *A. arenosa* diploids. In fact, the rarefaction curve ([supplementary fig. S3, Supplementary Material online](#)) suggests that as little as 40 diploid individuals sampled across the *A. arenosa* species range would be enough to cover most of its diploid diversity.

Reconstructed Haplotypes across Linked Candidate AASs

To search for the presence of tetraploid haplotypes in diploids, we reconstructed lineage-specific haplotypes and their allele frequencies across the sets of linked candidate AASs within each candidate protein in tetraploids ([supplementary table S6, Supplementary Material online](#)). We used this simplified procedure as we were not able to use standard phasing procedures reliably, due to the fact that we were using short reads and working with tetraploids ([Kyriakidou et al. 2018](#)).

For each protein, with n candidate AAS sites in the data set of 145 individuals consisting of 105 diploids and 40 tetraploids, we defined M_i to be the major allele frequency at the candidate AAS site i , given that the sample consists of 160 tetraploid haplotypes, and 210 diploid haplotypes, this major allele frequency is going to be dominated by the diploid haplotype, thus we define the ancestral (i.e., diploid) haplotype allele frequency as $HAF_d = \min\{M_i\}$, and consequently, we define the derived (i.e., tetraploid) HAF as $HAF_a = 1 - \max\{M_i\}$. We further define the frequency of all other haplotypes, which result from recombination of the two previous, as $HAF_r = 1 - HAF_a - HAF_d$.

We checked for reliability of our approach by extracting haplotypes from our diploid and tetraploid assemblies. Extracted diploid and tetraploid haplotypes of candidate meiosis proteins were consistent with the diploid and tetraploid haplotypes combined based on the allele frequencies at candidate AAS sites.

For all calculations, we used our in-house R script (github.com/mbohutinska/ProtEvol).

Hamming Distance and Diameter

In order to study the uniqueness of the tetraploid haplotypes, we defined a measure based on maximum pairwise Hamming distance within a sample ([Robinson 2003](#)). In our setting, the Hamming distance compares distances between genotypes, for diploids we first define a distance between alleles such that if the genotypes of two different plants at a given loci is AA aa or aa AA, the genotypic distance is 1, and for pairs AA Aa, Aa aa, Aa Aa, Aa AA, AA AA, aa Aa, the genotypic distance is 0. For tetraploids, we define the genotypic distance to be 1 if the pairs of genotypes are AAAA aaaa, AAAa aaaa, AAAA Aaaa, aaaA AAAA, aaaa AAAa, aaaa AAAA and 0 otherwise. For diploid/tetraploid comparison, we define the genotypic distance to be 1 for the pairs AA aaaa, AA Aaaa, aa AAAA, aa AAAa and 0 otherwise.

The Hamming distance is the sum over all positions that are different. The maximum pairwise numbers are called the Hamming diameter. If the Hamming distance between

diploids and tetraploids exceeds Hamming diameter within diploids, it becomes plausible that the AASs forming the tetraploid haplotypes originated de novo. This is a conservative indication of possible de novo origin of the tetraploid haplotype, as the fact that all the AASs forming the tetraploid haplotype are standing in the diploids does not imply that the complete tetraploid haplotype preexisted in any diploid individual.

The code used for the calculations is available in github (https://github.com/paajanen/meiosis_protein_evolution/).

Compensatory Evolution and Coevolution

Timing of Sweeps Using the Haplotype Information

Assuming hard sweep, the sweeping allele initially clears variation on the swept haplotype in a population, but over time, new variants accumulate. In addition, recombination causes the length of swept haplotypes to decline over time (Ormond et al. 2016; Stephan 2019). We thus combined these two metrics to infer the relative age of selection sweeps within the subset of six candidate meiosis proteins with signs of de novo origin of the selected haplotype. For each meiosis protein, we used the haplotype interval between first and last candidate AASs. We took the length of the haplotype in base pairs and measured how many new mutations had appeared in the set of the tetraploid genomes between the first and the last candidate AAS, excluding the candidate AASs, and normalized this count by the length of the haplotype. Finally, we considered the protein with the highest proportion of accumulated mutations in the selected haplotype as the oldest. Note that the short-read population genomic tetraploid data did not allow for reliable phasing so we could not use any method relying on haplotype length decay across individuals.

Secondary Structure Prediction in a Subset of Candidate Meiosis Proteins

Coding sequences of candidate meiotic genes were extracted from our diploid and tetraploid *A. arenosa* reference genomes. Open-reading frames were translated into amino acid sequences using Geneious v11 (Kearse et al. 2012). The presence of characteristic amino acid polymorphisms found in this study, conserved in diploid and tetraploid *A. arenosa*, could be confirmed in the extracted sequences. The online tool PSIPRED was used to predict secondary protein structures (Jones 1999; bioinf.cs.ucl.ac.uk/psipred/, last accessed February 3, 2021). The PSIPRED algorithm calculates the likelihood of local amino acid interactions including coil (C; disordered), helix (H), or sheet structures (E) for every amino acid position. The folding of amino acid chains into 3D structures is influenced by local forces (interactions between close amino acid residues, connected neighbors), which determine the secondary structure, and nonlocal forces (topological neighbors), which lead to the tertiary structure. The PSIPRED algorithm includes two feed-forward neural networks that perform an analysis of the output of PSI-Blast (position-specific iterated-Blast), which in turn is based on an alignment of multiple protein sequences. To compare secondary structures of meiotic proteins with each other,

sequences of secondary structures from diploid and tetraploid *A. arenosa* were pairwise aligned using the Geneious alignment tool with default settings. Structure identity scores (0; 1) were extracted and plotted together with the identities of the amino acid sequences. Binding sites were identified by literature search: PDS5b-binding site in REC8/SYN1 (Muir et al. 2016), SCC3-binding site in REC8/SYN1- and REC8/SYN1-binding sites in SCC3 (Roig et al. 2014; Orgil et al. 2015), and WAPL-binding site in PDS5b (Ouyang et al. 2016).

Evidence for MNMs

We observed that in some of our candidate proteins, the candidate AASs were <20 bp apart (supplementary tables S5 and S6, Supplementary Material online), a common rough way how to define MNMs in human or *Drosophila* (Schridter et al. 2011; Besenbacher et al. 2016). Thus, we tested if distances between our candidate AASs in tetraploids were significantly shorter than distances between sites harboring missense SNPs in genes genomewide. We repeated the analysis over tetraploid individuals from the subsampled data set and the results were consistent, so we report results for the individual with the highest coverage SWA_002_1. We used Wilcoxon rank sum test to compare distances between candidate AASs within candidate meiosis proteins and any SNPs genomewide (R package stats, R Core Team 2018).

Another evidence for MNMs is a significant excess of transversions relative to transitions compared with genomewide counts. Thus, for each SNP, we determined if it is a transition or transversion using SnpEff (Cingolani et al. 2012) and tested for excess of transversions relative to transitions in our candidate proteins compared with genomewide counts using z test (R package stats, R Core Team 2018).

Code Availability

Custom scripts used in this paper are available at the following github repositories <https://github.com/mbohutinska/ProtEvol>, https://github.com/mbohutinska/ScanTools_ProtEvol, https://github.com/paajanen/meiosis_protein_evolution.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Jonathan Kirby, Jakub Vlček, Pierre Baduel, and Patrick Monnahan for valuable discussions. This work was supported by a European Research Council Consolidator (CoG EVO-MEIO 681946 to K.B.), European Research Council Starter (StG HOTSPOT 679056 to L.Y.), Charles University (Project 284119, grant agency to M.B.), and by the BBSRC via grant BB/P013511/1 to the John Innes Centre. Additional support was provided by Czech Science Foundation (project 20-22783S to F.K.) and by the long-term research development project no. RVO 67985939 of the Czech Academy of Sciences. Computational resources were supplied by the project “e-Infrastruktura CZ” (e-INFRA

LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures and by the NBI Computing infrastructure for Science (CIS) group through the use of HPC and storage facilities.

Data Availability

The diploid and tetraploid assemblies and raw reads used to generate them are freely available in the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena>) with the project ID PRJEB37828. The RNA-Seq data are freely available from ENA with the project ID PRJEB34382. The resequenced reads have been deposited in the Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>) with the primary accession code PRJNA484107 (available at <http://www.ncbi.nlm.nih.gov/bio-project/484107>).

References

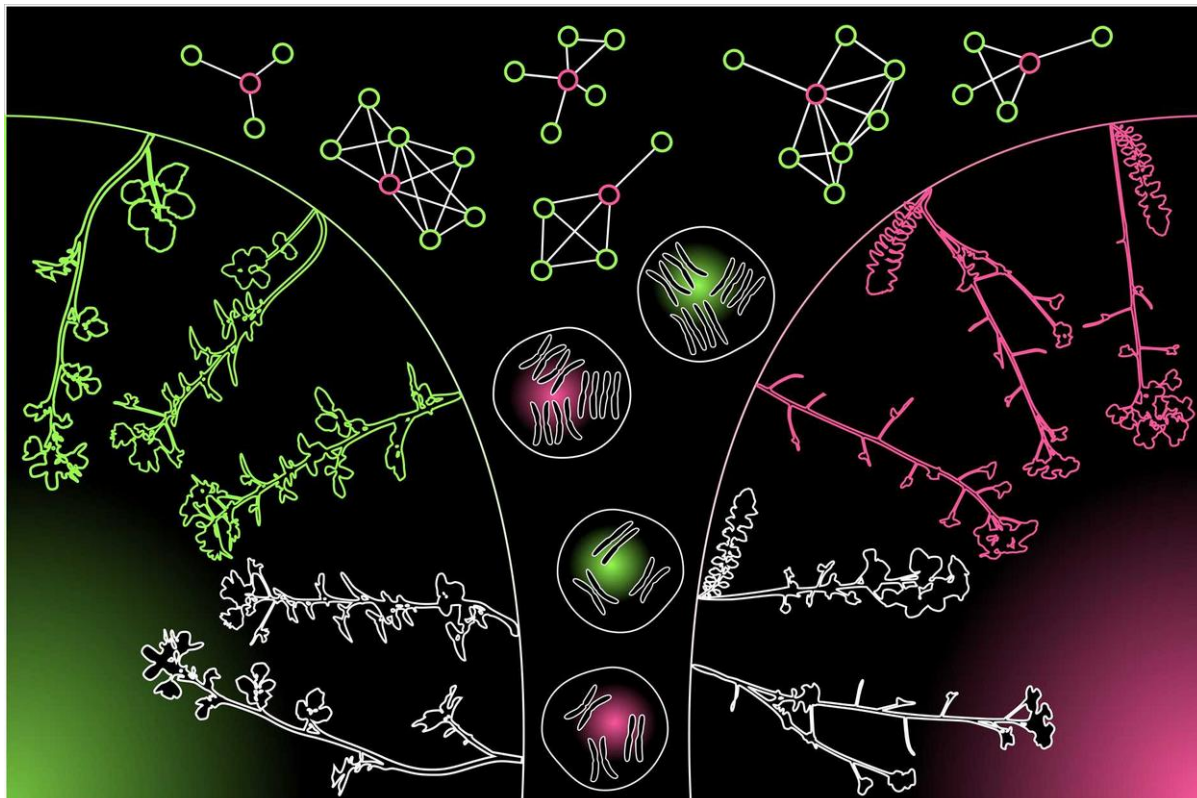
- Alves JM, Carneiro M, Cheng JY, de Matos AL, Rahman MM, Loog L, Campos PF, Wales N, Eriksson A, Manica A, et al. 2019. Parallel adaptation of rabbit populations to myxoma virus. *Science* 363(6433):1319–1326.
- Arnold B, Kim S-T, Bomblies K. 2015. Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by inter-ploidy admixture. *Mol Biol Evol.* 32(6):1382–1395.
- Baker Z, Schumer M, Haba Y, Bashkurova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *Elife* 6:e24133.
- Barrett RDH, Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol Evol.* 23(1):38–44.
- Van Belleghem SM, Vangestel C, De Wolf K, De Corte Z, Möst M, Rastas P, De Meester L, Hendrickx F. 2018. Evolution at two time frames: polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS Genet.* 14(11):e1007796.
- Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The *Arabidopsis* Information Resource: making and Mining the “Gold Standard” annotated reference plant genome. *Genes* 53(8):474–485.
- Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjánsson H, Jonasdóttir A, Jonasdóttir A, Magnusson OT, Thorsteinsdóttir U, Masson G, et al. 2016. Multi-nucleotide de novo mutations in humans. *PLoS Genet.* 12(11):e1006315.
- Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* 23(9):1514–1521.
- Bomblies K. 2020. When everything changes at once: finding a new normal after genome duplication. *Proc R Soc Lond B.* 287(1939):20202154.
- Bomblies K, Higgins JD, Yant L. 2015. Meiosis evolves: adaptation to external and internal environments. *New Phytol.* 208(2):306–323.
- Bomblies K, Jones G, Franklin C, Zickler D, Kleckner N. 2016. The challenge of evolving stable polyploidy: could an increase in “crossover interference distance” play a central role? *Chromosoma* 125(2):287–300.
- Bomblies K, Madlung A. 2014. Polyploidy in the *Arabidopsis* genus. *Chromosome Res.* 22(2):117–134.
- Brand CL, Wright L, Presgraves DC. 2019. Positive selection and functional divergence at meiosis genes that mediate crossing over across the *Drosophila* phylogeny. *G3 (Bethesda)* 9:3201–3211.
- Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman Test and Slightly Deleterious Mutations. *Mol Biol Evol.* 25(6):1007–1015.
- Cifuentes M, Grandont L, Moore G, Chèvre AM, Jenczewski E. 2010. Genetic regulation of meiosis in polyploid species: new insights into an old question. *New Phytol.* 186(1):29–36.
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.
- Comai L. 2005. The advantages and disadvantages of being polyploid. *Nat Rev Genet.* 6(11):836–846.
- Davis BH, Poon AFY, Whitlock MC. 2009. Compensatory mutations are repeatable and clustered within proteins. *Proc Biol Sci.* 276(1663):1823–1827.
- DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet.* 6(9):678–687.
- Doyle J, Coate J. 2019. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *Int J Plant Sci.* 180(1):1–52.
- Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagemann J, Grimm DG, Chen J, et al. 2018. The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* 14(2):e1007155.
- Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.
- Grishaeva TM, Bogdanov YF. 2014. Conservation and variability of synaptonemal complex proteins in phylogenesis of eukaryotes. *Int J Evol Biol.* 2014:1–16.
- Haenel Q, Roesti M, Moser D, MacColl ADC, Berner D. 2019. Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evol Lett.* 3(1):28–42.
- Harris K, Nielsen R. 2014. Error-prone polymerase activity causes multi-nucleotide mutations in humans. *Genome Res.* 24(9):1445–1454.
- Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169(4):2335–2352.
- Heyting C. 1996. Synaptonemal complexes: structure and function. *Curr Opin Cell Biol.* 8(3):389–396.
- Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K. 2012. Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* 8(12):e1003093.
- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgrén N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet.* 43(5):476–481.
- Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132(2):583–589.
- Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 292(2):195–202.
- Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484(7392):55–61.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647–1649.
- Kolář F, Fuxová G, Závěská E, Nagano AJ, Hyklová L, Lučanová M, Kudoh H, Marhold K. 2016. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model *Arabidopsis arenosa*. *Mol Ecol.* 25(16):3929–3949.

- Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 4(7):1073–1082.
- Kumar R, Bourbon HM, De Massy B. 2010. Functional conservation of Mei4 for meiotic DNA double-strand break formation from yeasts to mice. *Genes Dev.* 24(12):1266–1280.
- Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömviik MV. 2018. Current strategies of polyploid plant genome sequence assembly. *Front Plant Sci.* 9:1660.
- Lai Y-T, Yeung CKL, Omland KE, Pang E-L, Hao Y, Liao B-Y, Cao H-F, Zhang B-W, Yeh C-F, Hung C-M, et al. 2019. Standing genetic variation as the predominant source for adaptation of a songbird. *Proc Natl Acad Sci U S A.* 116(6):2152–2157.
- Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, Shim JE, Shim H, Kim H, Kim C, et al. 2015. AraNet v2: an improved database of co-functional gene networks for the study of *Arabidopsis thaliana* and 27 other nonmodel plant species. *Nucleic Acids Res.* 43:D996–D1002.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v1* [q-bio.GN]. Available from: <http://github.com/lh3/bwa>.
- Liu Z, Cheema J, Vigouroux M, Hill L, Reed J, Paajanen P, Yant L, Osbourn A. 2020. Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat Commun.* 11:1–11.
- Maisnier-Patin S, Berg OG, Liljas L, Andersson DI. 2002. Compensatory adaptation to the deleterious effect of antibiotic resistance in *Salmonella typhimurium*. *Mol. Microbiol.* 46(2):355–366.
- Matuszewski S, Hermisson J, Kopp M. 2015. Catch me if you can: adaptation from standing genetic variation to a moving phenotypic optimum. *Genetics* 200(4):1255–1274.
- McDonald J, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328):652–654.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297–1303.
- Messer PW, Neher RA. 2012. Estimating the strength of selective sweeps from deep population diversity data. *Genetics* 191(2):593–605.
- Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol.* 28(11):659–669.
- Monnahan P, Kolář F, Baduel P, Sailer C, Koch J, Horvath R, Laenen B, Schmickl R, Paajanen P, Šrámková G, et al. 2019. Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nat Ecol Evol.* 3(3):457–468.
- Morgan C, Zhang H, Henry CE, Franklin FCH, Bomblies K. 2020. Derived alleles of two axis proteins affect meiotic traits in autotetraploid *Arabidopsis arenosa*. *Proc Natl Acad Sci USA.* 117(16):8980–8988.
- Moura de Sousa J, Balbontin R, Durão P, Gordo I. 2017. Multidrug-resistant bacteria compensate for the epistasis between resistances. *PLoS Biol.* 15(4):e2001741.
- Muir KW, Kschonsak M, Li Y, Metz J, Haering CH, Panne D. 2016. Structure of the Pds5-Scc1 complex and implications for cohesin function. *Cell Rep.* 14(9):2116–2126.
- Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc B.* 365(1537):185–205.
- Olson-Manning CF, Wagner MR, Mitchell-Olds T. 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. *Nat Rev Genet.* 13(12):867–877.
- Orgil O, Matityahu A, Eng T, Guacci V, Koshland D, Onn I. 2015. A conserved domain in the Scc3 subunit of cohesin mediates the interaction with both Mcd1 and the cohesin loader complex. *PLoS Genet.* 11(3):e1005036.
- Ormond L, Foll M, Ewing GB, Pfeifer SP, Jensen JD. 2016. Inferring the age of a fixed beneficial allele. *Mol Ecol.* 25(1):157–169.
- Ouyang Z, Zheng G, Tomchick DR, Luo X, Yu H. 2016. Structural basis and IP6 requirement for Pds5-dependent cohesin dynamics. *Mol Cell.* 62(2):248–259.
- Oziolor EM, Reid NM, Yair S, Lee KM, Guberman VerPloeg S, Bruns PC, Shaw JR, Whitehead A, Matson CW. 2019. Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science* 364(6439):455–457.
- Paajanen P, Kettleborough G, Opez-Girona EL, Giolai M, Heavens D, Baker D, Lister A, Cugliandolo F, Wilde G, Hein I. 2019. A critical comparison of technologies for a plant genome sequencing project. 8:1–12.
- Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. *New Phytol.* 186(1):5–17.
- Pedruzzi G, Barlukova A, Rouzine IM. 2018. Evolutionary footprint of epistasis. *PLoS Comput Biol.* 14(9):e1006426.
- Prezeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution (N. Y.)* 59(11):2312–2323.
- R Core Team. 2018. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: <https://www.r-project.org/>
- Ralph PL, Coop G. 2015. The role of standing variation in geographic convergent adaptation. *Am Nat.* 186(5):S5–S23.
- Rawat V, Abdelsamad A, Pietzenek B, Seymour DK, Koenig D, Weigel D, Pecinka A, Schneeberger K. 2015. Improving the annotation of *Arabidopsis lyrata* using RNA-Seq data. *PLoS One.* 10(9):e0137391.
- Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47(D1):D886–D894.
- Robinson DJS. 2003. An introduction to abstract algebra. Berlin (Germany): Walter de Gruyter
- Roig MB, Löwe J, Chan K-L, Beckouët F, Metson J, Nasmyth K. 2014. Structure and function of cohesin's Scc3/SA regulatory subunit. *FEBS Lett.* 588(20):3692–3702.
- Rojas Echenique JI, Kryazhimskiy S, Nguyen Ba AN, Desai MM. 2019. Modular epistasis and the compensatory evolution of gene deletion mutants. *PLoS Genet.* 15(2):e1007958.
- Rosenberg SC, Corbett KD. 2015. The multifaceted roles of the HOR MA domain in cellular signaling. *J. Cell Biol.* 211(4):745–755.
- Sánchez-Morán E, Mercier R, Higgins JD, Armstrong SJ, Jones GH, Franklin FCH. 2005. A strategy to investigate the plant meiotic proteome. *Cytogenet Genome Res.* 109(1–3):181–189.
- Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the Pathway Interaction Database. *Nucleic Acids Res.* 37(Suppl 1):D674–D679.
- Schrider DR, Hourmozdi JN, Hahn MW. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol.* 21(12):1051–1054.
- Smith J, Coop G, Stephens M, Novembre J. 2018. Estimating time to the common ancestor for a beneficial allele. *Mol Biol Evol.* 35(4):1003–1017.
- Smith N, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415(6875):1022–1024.
- Stenberg P, Saura A. 2013. Meiosis and its deviations in polyploid animals. *Cytogenet Genome Res.* 140(2–4):185–203.
- Stephan W. 2019. Selective sweeps. *Genetics* 211(1):5–13.
- Szamecz B, Boross G, Kalapis D, Kovács K, Fekete G, Farkas Z, Lázár V, Hrtzyan M, Kemmeren P, Groot Koerkamp MJA, et al. 2014. The genomic landscape of compensatory evolution. *PLoS Biol.* 12(8):e1001935.
- Szpak M, Mezzavilla M, Ayub Q, Chen Y, Xue Y, Tyler-Smith C. 2018. FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. *Genome Biol.* 19(1):18.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.
- Thompson KA, Osmond MM, Schluter D. 2019. Parallel genetic evolution and speciation from standing variation. *Evol Lett.* 3(2):129–141.
- Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet.* 47(1):97–120.
- Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res.* 27(5):757–767.

- Wright KM, Arnold B, Xue K, Surinov M, O'connell J, Bomblies K, Wright S. 2015. Selection on meiosis genes in diploid and tetraploid *Arabidopsis arenosa*. *Mol Biol Evol.* 32(4):944–955.
- Wu M, Kostyun JL, Hahn MW, Moyle LC. 2018. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Mol Ecol.* 27(16):3301–3316.
- Xie KT, Wang G, Thompson AC, Wucherpennig JI, Reimchen TE, MacColl ADC, Schluter D, Bell MA, Vasquez KM, Kingsley DM. 2019. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* 363(6422):81–84.
- Yant L, Hollister J, Wright K, Arnold B, Higgins J, Franklin F, Bomblies K. 2013. Meiotic Adaptation to Genome Duplication in *Arabidopsis arenosa*. *Curr Biol.* 23(21):2151–2156.
- Zhang L. 2005. Human SNPs Reveal No Evidence of Frequent Positive Selection. *Mol Biol Evol.* 22(12):2504–2507.
- Zickler D, Kleckner N. 1999. Meiotic chromosomes: integrating structure and function. *Annu Rev Genet.* 33(1):603–754.

Case study 5.

Novelty and convergence in adaptation to whole genome duplication.



Novelty and convergence in adaptation to whole genome duplication

Magdalena Bohutínská^{1,2,*}, Mark Alston³, Patrick Monnahan³, Terezie Mandáková⁴, Sian Bray^{5,6}, Pirita Paaanen³, Filip Kolář^{1,2,7}, and Levi Yant^{5,8*}

1. Department of Botany, Faculty of Science, Charles University, Prague, Czech Republic
2. Institute of Botany, The Czech Academy of Sciences, Průhonice, Czech Republic
3. Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park, Norwich, UK
4. CEITEC – Central European Institute of Technology, and Faculty of Science, Masaryk University, Kamenice, Czech Republic
5. Future Food Beacon of Excellence, University of Nottingham, Nottingham, UK
6. School of Biosciences University of Nottingham, Nottingham, UK
7. Natural History Museum, University of Oslo, Oslo, Norway
8. School of Life Sciences University of Nottingham, Nottingham, UK

***Authors for correspondence:** Levi Yant (levi.yant@nottingham.ac.uk)

and Magdalena Bohutínská (magdalena.holcova@natur.cuni.cz)

Keywords: polyploidy; convergence; genome duplication; adaptation

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Whole genome duplication (WGD) can promote adaptation but is disruptive to conserved processes, especially meiosis. Studies in *Arabidopsis arenosa* revealed a coordinated evolutionary response to WGD involving interacting proteins controlling meiotic crossovers, which are minimised in an autotetraploid (within-species polyploid) to avoid mis-segregation. Here we test whether this surprising flexibility of a conserved essential process, meiosis, is recapitulated in an independent WGD system, *Cardamine amara*, 17 million years diverged from *A. arenosa*. We assess meiotic stability and perform population-based scans for positive selection, contrasting the genomic response to WGD in *C. amara* with that of *A. arenosa*. We found in *C. amara* the strongest selection signals at genes with predicted functions thought important to adaptation to WGD: meiosis, chromosome remodelling, cell cycle, and ion transport. However, genomic responses to WGD in the two species differ: minimal ortholog-level convergence emerged, with none of the meiosis genes found in *A. arenosa* exhibiting strong signal in *C. amara*. This is consistent with our observations of lower meiotic stability and occasional clonal spreading in diploid *C. amara*, suggesting that nascent *C. amara* autotetraploid lineages were preadapted by their diploid lifestyle to survive while enduring reduced meiotic fidelity. However, in contrast to a lack of ortholog convergence, we see process-level and network convergence in DNA management, chromosome organisation, stress signalling, and ion homeostasis processes. This gives the first insight into the salient adaptations required to meet the challenges of a WGD state and shows that autopolyploids can utilize multiple evolutionary trajectories to adapt to WGD.

Introduction

Whole genome duplication (WGD) is both a massive mutation and a powerful force in evolution. The opportunities and challenges presented by WGD emerge immediately, realised in a single generation. As such, WGD comes as a shock to the system. Autopolyploids, formed by within-species WGD (without hybridization), result from the chance encounter of unreduced gametes (with diverse underlying factors, see Mason and Pires 2015). Thus, they typically harbour four full haploid genomes that are similar in all pairwise combinations, resulting in a lack of pairing partner preferences at meiosis. This, combined with multiple crossover events per chromosome pair, can result in multivalents among three or more homologs at anaphase, increasing the likelihood of mis-

segregation or chromosome breakage, leading to aneuploidy (Bomblies and Madlung 2014; Bomblies et al. 2016). Beyond this, WGD presents a suddenly transformed intracellular landscape to the conserved workings of the cell, such as altered ion homeostasis and a host of nucleotypic factors related to cell size, volume, and cell cycle progression (Chao et al. 2013; Yant and Bomblies 2015; Doyle and Coate 2019; Bomblies 2020).

Despite this, some lineages survive this early trauma and successfully speciate, with direct empirical evidence of the increased adaptability of autopolyploid lineages from *in vitro* evolutionary competition experiments in yeast (Selmecki et al. 2015). With increased ploidy, genetic variability can be maintained in a masked state, with evidence of young WGD lineages further recruiting diverse alleles by gene flow across ploidies, and indeed, species (Arnold et al. 2016; Marburger et al. 2019; Monnahan et al. 2019). At the genomic level, recent detailed understanding of gene flow following WGD supports the idea that WGD can cause the breakdown of species barriers present in diploids. Evidence for this has come from both plants (*A. arenosa*/*Arabidopsis lyrata* (Schmickl and Koch 2011)) and animals (the frog genus *Neobatrachus* (Novikova et al. 2020), reviewed in Schmickl and Yant 2021). In both examples WGD led to niche expansion (Molina-Henao and Hopkins 2019; Novikova et al. 2020) and the invasion of particularly challenging environments relative to the diploid: in the case of polyploid frogs, the desert (Novikova et al. 2020) and polyploid *A. arenosa*, metal-contaminated mines and serpentine barrens (Arnold et al. 2016; Preite et al. 2019; Konečná et al. 2021). Thus, while clear challenges must be overcome to function as a polyploid (Bomblies et al. 2015; Yant and Bomblies 2015; Baduel, et al. 2018), novel population genomic and ecological opportunities await a lineage that successfully adapts to a WGD state (Yant and Bomblies, 2015; Baduel *et al.*, 2018).

The functional and genomic basis for adaptation to WGD has been closely investigated in *A. arenosa*, which exists as both diploid and young autotetraploid lineages (~20,000 generation old; Kolář *et al.*, 2016; Arnold *et al.* 2015). Population genomic scans for selection using a diversity of metrics have shown the strongest signals of positive selection following WGD in *A. arenosa* as sharp, single-gene peaks over 10 genes that physically and functionally interact to control meiotic chromosome crossovers (Hollister et al. 2012; Yant et al. 2013; Bohutínská et al. 2021). During early meiotic chromosome crossover formation in an autotetraploid, the four copies of each chromosome are impossible to distinguish. Thus, crossovers can occur haphazardly in any pairwise manner. If more

than one crossover per chromosome pair is allowed to occur, multivalent associations can result, leading to aneuploidy at anaphase. Thus a reduction in the number of meiotic crossovers to one per chromosome pair stands as the leading candidate process mediating adaptation to WGD (Bomblies et al. 2016). In the young *A. arenosa* autotetraploids harbouring these derived alleles, we observed a decrease in meiotic crossover number as well as fewer multivalents relative to synthetic autopolyploids with ancestral-like diploid alleles (Yant et al. 2013). Recent work found that the closely related sister species *Arabidopsis lyrata*, which contains a younger autotetraploid lineage, also harbours many of the same selected alleles discovered in *A. arenosa* (Marburger et al. 2019). Moreover, from a joint population genomic analysis of both species across an established natural hybrid zone between *A. arenosa* and *A. lyrata*, clear gene sized signals of directional adaptive gene flow and positive selection emerge precisely at these alleles specifically between the two tetraploids (Marburger et al. 2019; Seear et al. 2020), indicating that *A. lyrata* and *A. arenosa* WGD stabilisation events are not fully independent. Among these candidate adaptive alleles at least one has been functionally shown to modulate adaptive decreases in crossover numbers (Morgan et al. 2020; Seear et al. 2020),

Here we use an independent system, ~17 million years diverged from both *A. arenosa* and *A. lyrata* (Huang et al. 2020), to test the hypothesis that this solution of meiosis gene evolution is repeated, and if not, whether changes in other genes from analogous processes are associated with adaptation to WGD. Given the clear results in *A. arenosa* and *A. lyrata*, we hypothesised that the adaptive trajectories which are available to mediate adaptation to a WGD state are constrained, leading to repeated selection of the same suite of meiosis genes. Such a result would offer a striking case of convergent evolution in core cellular processes. To test this hypothesis, we take advantage of a well-characterised model, *Cardamine amara* (Brassicaceae, tribe Cardamineae). A large-scale cytotyping survey of over ~3,300 individuals in 302 populations and genetic analysis detail the demographic relationships of this diploid/tetraploid complex in the Eastern and Central Alps (Zozomová-Lihová et al., 2015). Comparison of genotyping results of this study with simulations indicates a single autotetraploid origin. Importantly, *C. amara* is a perennial herb harbouring a high level of genetic diversity and shares with *A. arenosa* a similar distribution range and evolutionary history, with a likely single geographic origin, followed by autotetraploid expansion associated with glacial oscillations (Marhold et al. 2002; Zozomová-Lihová et al. 2015).

To test our hypothesis that gene-level evolutionary convergence is likely following WGD, we performed genome scans for positive selection in both *C. amara* and *A. arenosa*, contrasting natural autotetraploid and diploid populations in both species. Because there was no reference genome available for *C. amara*, we first generated a novel quality reference. We then tested for convergence in the evolutionary response to WGD at the level of the ortholog, process, and network in a sampling of 100 *C. amara* and 120 *A. arenosa* individuals from well-assessed ranges (Arnold et al. 2015; Zozomová-Lihová et al. 2015; Kolář et al. 2016; Monnahan et al. 2019). Overall, we found that the evolutionary response to WGD in *C. amara* is very different to that of *A. arenosa*, with none of the orthologous meiosis-related genes that control meiotic chromosome crossovers in *A. arenosa* under strong selection in *C. amara*. In contrast, we find a clear signal of process-level convergence in core pathways controlling DNA management and chromosome organisation.

Results and Discussion

Reference genome, population selection, sampling and genetic structure. Because *C. amara* is ~17 million years diverged from *A. arenosa* (Huang et al. 2020), using the same reference genome for mapping reads of both species would result in unacceptably low mapping efficiencies and missing data. We therefore first generated a novel reference genome for *C. amara* (N50 = 1.82 mb, 95% complete BUSCOs; see Methods). We then resequenced in triplicate four populations of contrasting ploidy, sampling 100 individuals: two diploid (LUZ, VRK) and two autotetraploid (CEZ, PIC; Fig. 1a; Supplementary Table 1). We chose these populations based on a comprehensive cytological and demographic survey of ~3,300 *C. amara* samples throughout the Czech Republic (Zozomová-Lihová et al. 2015). Sampled plants were spaced at least 3 m apart, as this distance was sufficient to avoid resampling of identical clones in that study. We chose populations to represent core areas of each cytotype, away from potential hybrid zones and distant from any triploid-containing populations based on (Zozomová-Lihová et al. 2015). Further, we performed flow cytometry on every sample sequenced to verify expected ploidy.

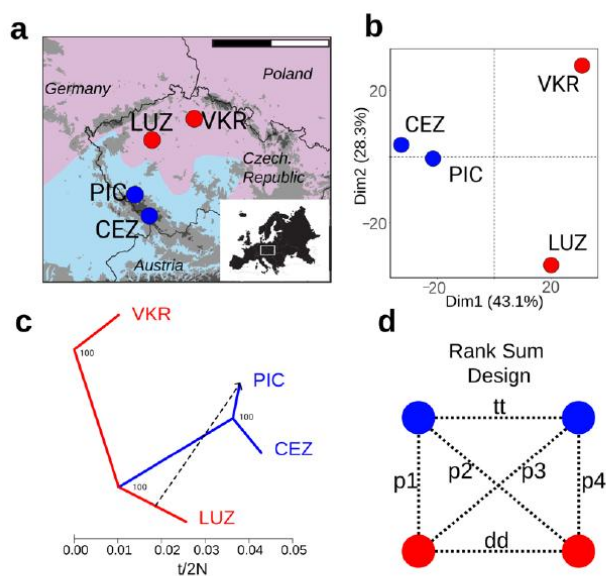


Figure 1. Sampling and population structure of *Cardamine amara*. **a**, Locations of diploid (red) and autotetraploid (blue) *C. amara* populations sampled. Scale bar corresponds to 200 km; shaded area represents each cytotype range in (Zozomová-Lihová et al., 2015). **b**, Population differentiation represented by Principal Component Analysis of ~124,000 fourfold degenerate SNPs. **c**, Phylogenetic relationships and migration events between populations inferred by TreeMix analysis. X-axis shows the drift estimation, corresponding to the number of generations separating the two populations (t), and effective population size (N) (Pickrell and Pritchard 2012). Node labels show bootstrap support, and the arrow indicates the most likely migration event (migration weight, which can be interpreted as a moderate degree of admixture = 0.18, similar to *A. arenosa*, shown in Supplementary Fig. 1). Additional migration events did not improve the model likelihood. **d**, Rank Sum design used in divergence scans to minimise potential bias of population-specific divergence. $p1$ to $p4$ represent the between-ploidy contrasts used for the rank sum calculations. dd and tt represent within-ploidy contrasts used to subtract signal of local population history within each cytotype.

To obtain robust population allele frequency estimates across genomes, we performed a replicated pooled sequencing approach. From every population we pooled DNA from 25 individuals and generated on average 31 million reads per sample (for all samples we generated triplicate DNA preps, pooling, and sequencing to control for potential sampling error: details in Methods) and mapped reads to our new *C. amara* assembly (mean coverage per population = 86, Supplementary Table 2). After mapping, variant calling and quality filtration, we obtained a final dataset of 2,477,517 SNPs.

The first PCA axis dominantly explained 43% of variation (Fig. 1b) and was consistent with differentiation primarily by geographic distribution or ploidy (which coincide), followed by differentiation between the two diploid populations from each other (second axis explaining 28% of variation). The two autotetraploid populations clustered together in the TreeMix graph (Fig. 1c) and had the lowest genetic differentiation of all contrasts ($F_{st} = 0.04$, mean allele frequency difference = 0.06, Table 1) and lacked any fixed SNP difference whatsoever (Table 1). This high genetic similarity and spatial arrangement (the populations represent part of a continuous range of the autotetraploid cytotype), suggest that both autotetraploid populations represent the outcome of a single polyploidization event, in line with previous assessments (Marhold *et al.*, 2002; Zozomová-Lihová *et al.*, 2015), although multiple tetraploid origins cannot be ruled out. The absence of individual-level genotype information did not allow for exact dating, but nearly identical levels of interploidy divergence in both *C. amara* and *A. arenosa* (average F_{st} between diploids and autotetraploids = 0.10 and 0.11, respectively) and comparable drift estimates in TreeMix (Supplementary Fig. 1), suggested that the polyploidization may be roughly the same age (Table 1). Supporting this, both WGD events were estimated to correspond with the end of the last European glaciation (Marhold *et al.*, 2002; Arnold *et al.*, 2015; Zozomová-Lihová *et al.*, 2015).

Table 1. Measures of genome-wide differentiation between *C. amara* and *A. arenosa* populations

Populations	Ploidies	Mean AFD	Fixed diffs	Mean F_{st}	# SNPs
PIC - VKR	4x - 2x	0.09	30	0.09	2,326,315
PIC - LUZ	4x - 2x	0.09	2	0.08	2,314,229
CEZ - VKR	4x - 2x	0.11	120	0.12	2,333,538
CEZ - LUZ	4x - 2x	0.11	86	0.11	2,335,004
CEZ - PIC	4x - 4x	0.06	0	0.04	2,297,229
LUZ - VKR	2x - 2x	0.1	6	0.09	2,018,892
<i>A. arenosa</i> tetraploids - <i>A. arenosa</i> diploids	4x - 2x	0.05	21	0.11	7,106,848

Note: Differentiation metrics shown are genome-wide mean allele frequency difference between populations (Mean AFD), the number of fixed differences (Fixed diffs) and mean F_{st} (Nei 1972). In the case of *A. arenosa*, F_{st} in diploids is calculated as a mean over all pairwise F_{st} measurements between the five previously characterised diploid lineages (Monnahan *et al.* 2019).

Selection specifically associated with WGD in *C. amara*. To minimise false positives due to local population history we leveraged a quartet-based design (Vijay et al. 2016), consisting of two diploid and two autotetraploid populations (details in Methods). The mean number of SNPs per population contrast was 2,270,868 (Table 1). We calculated F_{st} for 1 kb windows with a minimum 20 SNPs for all six possible population contrasts (Fig. 1d), and ranked windows based on F_{st} values. To focus on WGD-associated adaptation, we first assigned ranks to each window based on the F_{st} values in each of four possible pairwise diploid-autotetraploid contrasts and identified windows in the top 1% outliers of the resultant combined rank sum (Fig. 1d, contrasts p1-p4). We then excluded any window which was also present in the top 1% F_{st} outliers in diploid-diploid or autotetraploid-autotetraploid population contrasts to avoid misattribution caused by local population history (Fig. 1c, contrasts tt and dd). By this approach, we identified 440 windows that intersected 229 gene coding loci (Supplementary Dataset 1; termed WGD adaptation candidates below). To control for possible biases due to suboptimal window size selection, we recalculated F_{st} on a SNP-by-SNP basis, considering genes with 5 or more SNPs. This approach resulted in the comparable candidate list to the window-based analysis (see Methods). Larger windows (50kb) failed to detect peaks of divergence.

Among these 229 gene coding loci, a Gene Ontology (GO) term analysis yielded 22 significantly enriched biological processes (Fisher's exact test with conservative 'elim' method, $p < 0.05$, Supplementary Table 3). To further control for false positives and refine this candidate list to putatively functional candidates, we complemented these differentiation measures with a quantitative estimate that incorporates potential functional impact of encoded derived amino acid changes, following the FineMAV method (Szpak et al. 2018) (see Methods for a full description). In short, as an orthogonal complement to F_{st} scans above, FineMAV assigns SNPs a score based on the predicted functional consequences of resultant amino acid substitutions using Grantham scores, and amplifies these by the per-cytotype allele frequency difference between the two amino acids (Szpak *et al.*, 2018, Bohutinska et al. 2021). This allowed us to focus on radical amino acid changes driven to high frequency specifically in the autotetraploids. From our 229 F_{st} window-based WGD adaptation candidates, 120 contained at least one 1% FineMAV outlier amino acid substitution (Supplementary Datasets 1 and 2).

DNA maintenance (repair, chromosome organisation) and meiosis under selection in *C. amara*. Of the 22 significantly enriched GO processes, the most enriched by far was DNA metabolic process (p-value = 6.50E-08, vs 0.00021 for the next most confident enrichment), although there was also enrichment for chromosome organization and meiotic cell cycle. The 40 genes contributing to these categories showed highly localised peaks of differentiation (Fig. 2), as well as 1% FineMAV outlier SNPs in coding regions (Fig. 2, Supplementary Datasets 1 and 2). These genes also clustered in STRING interaction networks, suggesting coevolutionary dynamics driving the observed selection signals (Supplementary Fig. 2; see Methods). The largest cluster comprised of *MSH6*, *PDS5e*, *SMC2*, *MS5*, *PKL*, *HDA18*, *CRC*, and homologs of two uncharacterised, but putative DNA repair related loci *AT1G52950* and *AT3G02820* (containing SWI3 domain). *MutS Homolog 6 (MSH6)* is a component of the post-replicative DNA mismatch repair system. It forms a heterodimer with MSH2 which binds to DNA mismatches (Culligan and Hays 2000; Wu et al. 2003), enhancing mismatch recognition. *MutS* homologs have also been shown to control crossover number in *A. thaliana* (Lu et al. 2008). The *C. amara* ortholog of *AT1G15940* is a close homolog of *PDS5*, a protein required in fungi and animals for formation of the synaptonemal complex and sister chromatid cohesion (Panizza et al. 2000). *Structural Maintenance Of Chromosomes 2 (SMC2/TTN3)* is a central component of the condensin complex, which is required for segregation of homologous chromosomes at meiosis (Siddiqui et al. 2003) and stable mitosis (Liu and Meinke 1998). *PICKLE (PKL)* is a SWI/SWF nuclear-localized chromatin remodelling factor (Ogas et al., 1999; Shaked et al., 2006) that also has highly pleiotropic roles in osmotic stress response (Perruc et al., 2007), stomatal aperture (Kang et al. 2018), root meristem activity (Aichinger et al. 2011), and flowering time (Jing et al., 2019). Beyond this cluster, other related DNA metabolism genes among our top outliers include *DAYSLEEPER* (Fig. 2), a domesticated transposase that is essential for development, first isolated as binding the *Kubox1* motif upstream of the DNA repair gene *Ku70* (Bundock and Hooykaas 2005). The complex Ku70/Ku80 regulate non-homologous end joining (NHEJ) double-strand break repair (Tamura et al. 2002). Consistent with this, *DAYSLEEPER* mutants accumulate DNA damage (Knip 2012), but the exact role of *DAYSLEEPER* in normal DNA maintenance is not yet understood. Interesting also is the identification of *MALE-STERILE 5 (MS5/TDM1)*, which is required for cell cycle exit after meiosis II. As the name implies, MS5 mutants are male sterile, with pollen tetrads undergoing an extra round of division after meiosis II without chromosome replication (Glover et al. 1998). *MS5/TDM1* may be an APC/C component whose function is to ensure meiosis termination at the end of meiosis II (Cifuentes et al.

2016). Together, this set of DNA management loci exhibiting the strongest signals of selection points to widespread modulation of DNA repair and chromosome management following WGD in *C. amara*.

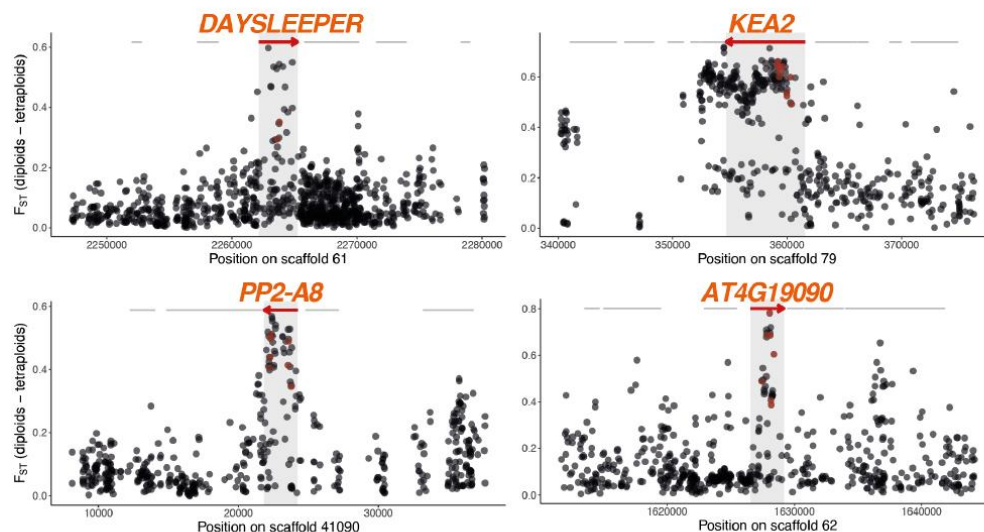


Figure 2. Selective sweep signatures at DNA management and ion homeostasis loci. Examples of selective sweep signatures among four candidate loci (red arrows). X-axis gives scaffold position in base pairs. Y-axis gives F_{ST} values at single nucleotide polymorphisms (dots) between diploid and autotetraploid *C. amara*. Red dots indicate FineMAV outlier SNPs. Red arrows indicate gene models overlapping top 1% F_{ST} windows and grey lines indicate neighbouring gene coding loci.

Evolution of stress signalling and ion homeostasis genes. The remainder of the enriched GO categories in *C. amara* revolved around a diversity of cellular processes, including stress response, protein phosphorylation, root development, ABA signalling, and ion homeostasis. The intersection of these processes was often represented by several genes. For example, two of the top 20 highest-scoring SNPs in the genome-wide FineMAV analysis reside in SNF1-related protein kinase SnRK2.9 (Supplementary Dataset 2). SnRKs have been implicated in osmotic stress and root development (Fujii *et al.*, 2011; Kawa *et al.*, 2020), and their activity also mediates the prominent roles of Clade A protein phosphatase 2C proteins in ABA and stress signalling (Cutler *et al.* 2010). Interesting in this respect is a strong signature of selection in *HIGHLY ABA-INDUCED PP2C GENE 1*, a clade A PP2C protein (Supplementary Dataset 1). Stress-related phosphoinositide phosphatases are represented by *SAC9*, mutants of which exhibit constitutive stress responses (Williams *et al.* 2005). Diverse other genes

related to these categories exhibit the strongest signatures of selection, such as *PP2-A8* (Meyers *et al.*, 2002) and *AT4G19090*, a transmembrane protein strongly expressed in young buds (Klepikova *et al.* 2016) (Fig. 2).

Given the observed increase in potassium and dehydration stress tolerance in first generation autotetraploid *Arabidopsis thaliana* (Chao *et al.* 2013), it is very interesting that our window-based outliers included an especially dramatic selective sweep at *K⁺ Efflux Antiporter 2 (KEA2)*, Fig. 2), a K⁺ antiporter that modulates osmoregulation, ion, and pH homeostasis (Kunz *et al.* 2014). Recent evidence indicates that *KEA2* is important for eliciting a rapid hyperosmotic-induced Ca²⁺ response to water limitation imposed by osmotic stress (Stephan *et al.* 2016). The *KEA2* locus in autotetraploid *C. amara* features an exceptional ten FineMAV-outlier SNPs (Fig. 2, Supplementary Datasets 1 and 2), indicating that the sweep contains a run of radical amino acid changes at high allele frequency difference between the ploidies, pointing to a potential functional change. We also detected *cation-chloride co-transporter 1 (HAP 5)* a Na⁺, K⁺, Cl⁻ co-transporter, involved in diverse developmental processes and Cl⁻ homeostasis (Colmenero-Flores *et al.* 2007). These cellular processes map well onto increasingly recognized changes that occur in polyploids, most comprehensively reviewed by (Bomblies, 2020).

Limited gene ortholog-level convergence between *C. amara* and *A. arenosa*. We hypothesized that WGD imposed strong, specific selection pressures leading to convergent directional selection on the same genes or at least on different genes playing a role in the same process (ortholog- or function-level convergence, respectively) between *C. amara* and *A. arenosa*. To test for this, we complemented our *C. amara* genome scan with an analysis of *A. arenosa* divergence outliers based on an expanded sampling relative to the original *A. arenosa* genome scan studies. We selected the 80 diploid and 40 autotetraploid individuals sequenced most deeply in a recent range-wide survey (Monnahan *et al.*, 2019, subsampling following Bohutinska *et al.* 2021) of genomic variation in *A. arenosa* (mean coverage depth per individual = 18; 160 haploid genomes sampled of each ploidy), and scanned for Fst outliers in 1 kb windows, as we did for *C. amara*. We identified 696 windows among 1% Fst outliers, overlapping 452 gene-coding loci (Supplementary Dataset 3), recovering results similar to (Yant *et al.* 2013, Bohutinska *et al.* 2021), including the interacting set of loci that govern meiotic chromosome crossovers, despite radically different sampling in each of the *A. arenosa* studies. From

this entire list of 452 *A. arenosa* WGD adaptation candidates, only six orthologous loci were shared with our 229 *C. amara* WGD adaptation candidates (Table 2). While it is possible that these six genes may be convergently evolving in each species, this degree of overlap was not significant ($p = 0.42$, Fisher's exact test), indicating no excess convergence at the level of orthologous genes beyond the quantity expected by chance. Re-analysis with candidate genes detected using the SNP-by-SNP divergence scan did not identify any additional convergent gene. Similarly, there was no excess overlap among genes which harbour at least one candidate FineMAV substitution (3 overlapping candidate genes out of 120 in *C. amara* and 303 in *A. arenosa*; $p = 0.27$, Fisher's exact test). This lack of excess convergence at the ortholog level may come as a surprise given the expected shared physiological challenges attendant to WGD (Yant and Bomblies 2015; Baduel, Bray, et al. 2018; Bomblies 2020).

Table 2. WGD adaptation candidates in both *A. arenosa* and *C. amara*.

<i>C. amara</i> ID	<i>A. thaliana</i> ID	<i>A. arenosa</i> ID	Name	Function (TAIR)
<i>C</i> Ag1480	AT1G16460	AL1G28600	MST2/RDH2	embryo/seed development
<i>C</i>Ag20214	AT2G45120	AL4G44210	C2H2-like zinc finger	stress response
<i>C</i>Ag11103	AT3G42170	AL3G27110	DAYSLEEPER	DNA repair
<i>C</i> Ag16465	AT3G62850	AL1G11960	zinc finger-like	unknown
<i>C</i>Ag4024	AT5G05480	AL6G15370	Asparagine amidase A	growth and development
<i>C</i> Ag5641	AT5G23570	AL6G34840	SGS3	posttranscriptional gene silencing

Note: The number of genes does not exceed random expectations for the overlap of candidate gene lists from each species, indicating a lack of gene-level convergence. Genes in bold also harbour at least one candidate FineMAV SNP in both species.

To determine whether we may have failed to detect convergent loci due to missing data or if top outliers in *A. arenosa* had few, but potentially functionally-implicated, differentiated SNPs in *C. amara*, we performed a targeted search in *C. amara* for the interacting set of meiosis proteins found to exhibit the most robust signatures of selection in *A. arenosa* (Yant *et al.*, 2013, Bohutinska *et al.* 2021) (Supplementary Table 4). All meiosis-related orthologs in *C. amara* that exhibit selection signatures in *A. arenosa* (13 in total) passed our data quality criteria and were included in our analyses. Only three showed any signal by FineMAV analysis: *PDS5b* harbours an unusually high three

fineMAV outlier SNPs, although it is not a Fst outlier. *ASY3*, which controls crossover distribution at meiosis, has only one FineMAV outlier SNP. Finally, a regulator of endoreduplication, *CYCA2;3*, also harbours a single FineMAV 1% outlier in *C. amara*, although it was not included in the Fst window analysis (the window overlapping it contained only 7 SNPs, below the 20 SNP minimum cut-off for inclusion in the Fst window analysis). However, these 7 SNPs exhibited high mean Fst (0.55). Thus, while we detect varying signal in these three meiosis-related genes following WGD (Supplementary Table 4), we do not see widespread signals of selection in the set of interacting crossover-controlling genes that were so conspicuous in *A. arenosa* (Yant et al. 2013).

Meiotic stability in *C. amara*. Despite our broad overall analysis of selection in *C. amara*, as well as a targeted assessment of particular meiosis genes, we did not detect strong signal of selection in meiosis genes in *C. amara* (Supplementary Table 4). The *C. amara* autotetraploid is a fertile, outcrossing, well-established lineage, but we still wondered if some contrast in meiotic behaviour underlies this difference in specific loci under selection. We therefore cytologically assessed the degree of male meiotic stability in *C. amara* (Fig. 3a). A reduction in crossover number to one per bivalent is indicated as a leading mechanism for meiotic diploidization in autopolyploids because this limits multivalent associations (which increase the propensity toward breakage and aneuploidy vs bivalents (Cifuentes et al. 2010; Le Comber et al. 2010; Bomblies et al. 2016)), so we use proportion of bivalents to multivalents as our estimator (Methods). This revealed a highly variable degree of stability in both *C. amara* cytotypes (mean proportion stable metaphase I cells in diploid maternal seed lines = 0.38 – 0.69, n = 133 scored cells; in tetraploids = 0.03 – 0.38; n = 348 scored cells; Supplementary Table 5). Indeed, while still highly variable, the overall degree of stability was lower in autotetraploids vs. diploids (differing proportion of stable to unstable meiotic cells for each ploidy; $D = 62.7$, $df = 1$, $p < 0.0001$, GLM with binomial errors; Fig. 3b, Supplementary Table 5), corresponding with the lack of selection signal in crossover-controlling meiosis genes. Interestingly, the broad variation in stability estimates within both cytotypes suggests widespread standing variation controlling this trait. In contrast, higher frequencies of stable metaphase I cells (>80%) have been commonly observed for diploid and autotetraploid *A. arenosa* (Marburger et al. 2019), although wider estimates of meiotic variation have also been observed in populations hybridising with *A. lyrata* (Seear et al., 2019). Taken together with the observation of occasional clonal spreading of *C. amara* (Hejný et al. 1992; Tedder et al. 2015; Zozomová-Lihová et al. 2015), this indicates an ability to

maintain stable populations, thus perhaps decreasing the immediate necessity to fully stabilise meiosis in either cytotype. Vegetative reproduction is often seen in polyploids (Herben *et al.*, 2017; Van Drunen and Husband, 2019) and in turn may have facilitated the establishment of the autotetraploid cytotypes. We note finally that the tetraploid populations are still highly fertile, consistent with observations across the range (Koch *et al.* 2003).

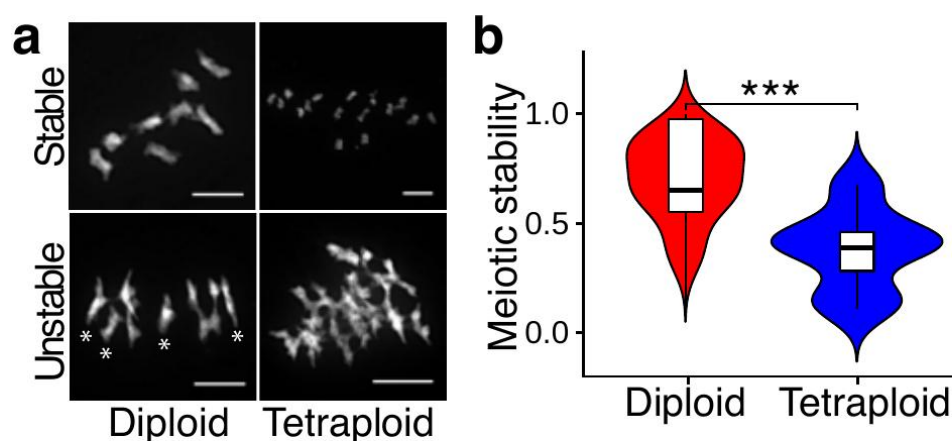


Figure 3. Variable meiotic stability in *C. amara*. **a**, An example of stable and unstable diploid and autotetraploid DAPI-stained meiotic chromosomes (diakinesis and metaphase I). Unstable meiosis is characterised by multivalent formation and interchromosomal connections, so we use the proportion of bivalents to multivalents as a proxy to estimate stability. In this example, the stable and unstable diploids (left panels) pictured contain 8 and 4 bivalents, respectively, while the stable and unstable tetraploids (right panels) show 16 and 0 bivalents, respectively. Thus all chromosomes pictured in these ‘Stable’ examples are present as bivalents, while in the ‘Unstable’ examples, only the four with astrisks (*) are bivalents, while the rest are multivalents. Scale bar corresponds to 10 μ m. For a complete overview of all scored chromosome spreads see Supplementary Fig. 5. **b**, Distribution of meiotic stability (calculated as proportion of stable and partly stable to all scored meiotic spreads) in diploid and autotetraploid individuals of *C. amara*. *** - $p < 0.001$, GLM with binomial errors.

Evidence for process-level convergence. While we found no excess convergence at the level of orthologous genes under selection, we speculated that convergence may occur nevertheless at the level of functional processes. To test this, we used two complementary approaches: overlap of GO term enrichment and evidence of shared protein function from interaction networks. First, of 73

significantly ($p < 0.05$) enriched GO terms in *A. arenosa* (Supplementary Table 6), we found that five were identical to those significantly enriched in *C. amara*, which is more than expected by chance ($p < 0.001$, Fisher's exact test; Table 3). In addition, some processes were found in both species, but were represented by slightly different terms, especially in the case of meiosis ("meiotic cell cycle" in *C. amara*, "meiotic cell cycle process" in *A. arenosa*: Supplementary Tables 3 and 6). Remarkably, the relative ranking of enrichments of all five convergent terms was identical in both *C. amara* and *A. arenosa* (Table 3). This stands in strong contrast to the fact that *A. arenosa* presented an obvious set of physically and functionally interacting genes in the top two categories (DNA metabolic process and chromosome organisation), while the genes in these categories in *C. amara* are implicated in more diverse DNA management roles.

Table 3. Convergent processes under selection in both *C. amara* and *A. arenosa* following WGD

GO ID	Term	p-value (<i>C. amara</i>)	p-value (<i>A. arenosa</i>)	Enrichment (<i>C. amara</i>)	Enrichment (<i>A. arenosa</i>)
GO:0006259	DNA metabolic process	6.50E-08	8.20E-04	3.72	2.46
GO:0051276	chromosome organization	0.019	2.10E-04	1.98	2.01
GO:0009738	abscisic acid-activated signalling pathway	0.032	0.022	2.54	2.10
GO:0071215	cellular response to abscisic acid stimulation	0.048	0.04	2.30	1.90
GO:0097306	cellular response to alcohol	0.048	0.04	2.30	1.90

Note: p-values given are Fisher's exact test, which tests for enrichment of terms from the GO hierarchy.

Enrichment refers to fold enrichment.

Second, we sought for evidence that genes under selection in *C. amara* might interact with those found under selection in *A. arenosa*, which would further support process-level convergence between the species. Thus, we took advantage of protein interaction information from the STRING database, which provides an estimate of proteins' joint contributions to a shared function (Szklarczyk et al. 2015). For each *C. amara* WGD adaptation candidate we searched for the presence of STRING interactors among the *A. arenosa* WGD adaptation candidates, reasoning that finding such an association between candidates in two species may suggest that directional selection has targeted the same processes in both species through different genes. Following this approach, we found that out of the 229 *C. amara* WGD adaptation candidates, 90 were predicted to interact with at least one of

the 452 WGD adaptation candidates in *A. arenosa*. In fact, 57 likely interacted with more than one *A. arenosa* candidate protein (Fig. 4 and Supplementary Table 7). This level of overlap was greater than expected by chance ($p = 0.001$ for both "any interaction" and "more-than-one interaction", as determined by permutation tests with the same database and 1000 randomly generated candidate lists).

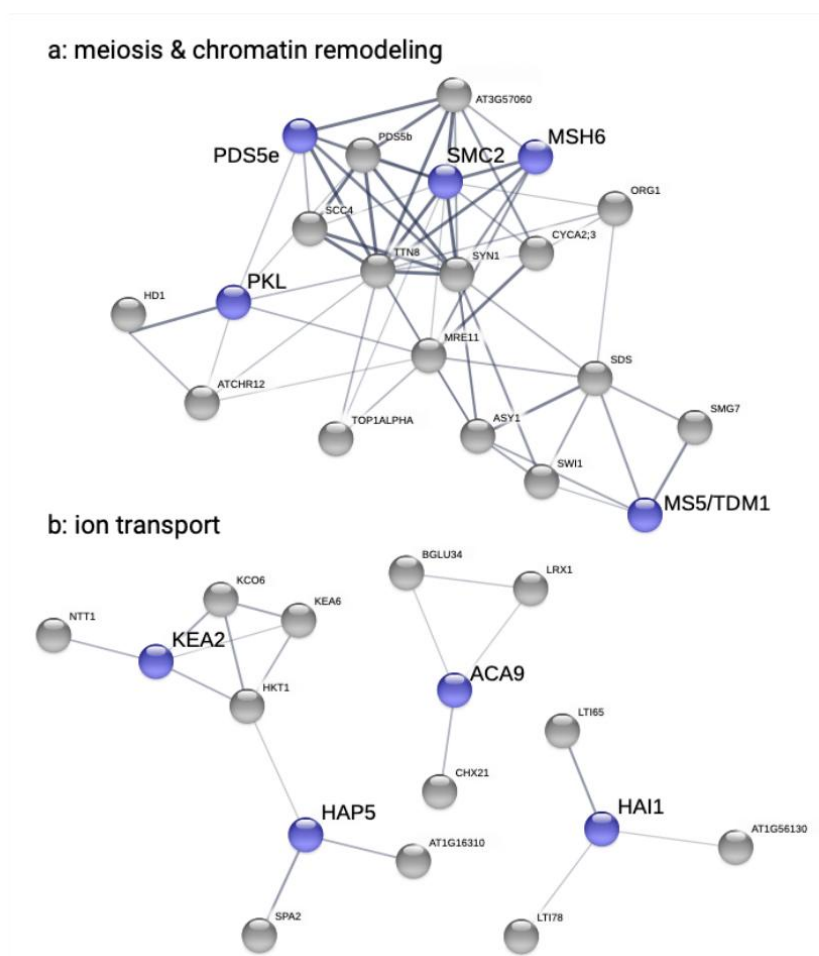


Figure 4. Evidence for functional convergence between *C. amara* and *A. arenosa* following independent WGDs. Plots show *C. amara* candidate genes in blue and STRING-associated *A. arenosa* candidate genes in grey. We used only medium confidence associations and higher (increasing thickness of lines connecting genes indicates greater confidence). **a**, meiosis- and chromatin remodelling-related genes. **b**, ion transport-related genes.

Several large STRING clusters were evident among WGD adaptation candidates in *C. amara* and *A. arenosa* (Fig. 4). The largest of these clusters centre on genome maintenance, specifically meiosis and chromatin remodelling (Fig. 4a), and ion homeostasis (especially K⁺ and Ca²⁺), along with stress (ABA) signalling (Fig. 4b), consistent with the results of GO analysis. Taken together, both STRING and GO analyses support our hypothesis of functional convergence of these processes following WGD in *C. amara* and *A. arenosa*.

Conclusions

Given the expected shared challenges attendant to WGD in *C. amara* and *A. arenosa*, we hypothesised at least partially convergent evolutionary responses to WGD. While we found obvious convergent recruitment at the level of functional processes, we did not detect excess convergence at the gene level. This was consistent with the probable absence of shared standing variation between these species (Hudson and Coyne 2002), which are 17 million years diverged. Nevertheless, we note that if any shared variation has persisted, it was not selected upon convergently in both young autotetraploids, thus strengthening the conclusion that the genes selected in response to WGD are not highly constrained.

The most prominent difference we observed here is the lack of an obvious coordinated evolutionary response in genes stabilizing early meiotic chromosome segregation in *C. amara*, relative to the striking coevolution of physically and functionally interacting proteins governing crossover formation in *A. arenosa*. This might be explained to some extent by our observation that in *C. amara* both diploids and autotetraploids are somewhat less meiotically stable than either cytotype in *A. arenosa*, and this instability may preadapt the autotetraploids to enjoy a less strict reliance on the generation of a high percentage of euploid gametes, by forcing occasional reliance on vegetative reproduction, as has been observed (Herben et al. 2017). This then may allow a decoupling of crossover number reduction from broader changes across meiosis and other processes we observe. This is not to say that we see no signal of WGD adaptation in *C. amara*: factors governing timing during later meiosis, especially the exit from meiotic divisions as evidenced by the interacting trio of *SMG7*, *SDS* and *MS5*, along with other chromatin remodelling factors and DNA repair-related proteins, such as *MSH6* and

DAYSLEEPER give very strong signals. The convergent functions we did detect (other meiosis processes, chromosome organisation/chromatin remodelling, ABA signalling and ion transport) provide first insights into the salient challenges associated with WGD. We note also that tetraploid populations of both *C. amara* and *A. arenosa* are found in slightly colder environments than conspecific diploids (Zozomová-Lihová et al. 2015; Molina-Henao and Hopkins 2019), so some of these processes (e.g. ABA signalling and ion transport) might be linked to ecological adaptation following WGD.

Overall, our results provide contrast to widespread reports of gene-level convergence (reviewed e.g. in (Elmer and Meyer, 2011; Martin and Orgogozo, 2013; Blount *et al.*, 2018)) and support the idea that pathway-level convergence becomes dominant when the divergence between species is high (Takuno et al. 2015; Birkeland et al. 2020; Bohutínská et al. 2020). This could be due to the absence of shared low-frequency alleles (acquired via gene flow or from standing variation) in species diverging millions of years ago, as was shown in alpine adaptation of different Brassicaceae species (Bohutínská et al. 2020). Alternatively, WGD provides complex multi-factorial challenge (Bomblies et al. 2015; Baduel, Bray, et al. 2018; Bomblies 2020) and the possible solutions to overcome such challenge may in fact be diverse. The result would be multiple alternative genetic paths to adaptation, with limited gene-level convergence due to the low diversity constraints (Yeaman et al. 2018). Finally, we note that the lack of gene-level convergence in meiosis genes suggests that the genomic changes associated with meiosis stabilization after WGD might not be as constrained as would be expected based on its functional conservation across eukaryotes (Grishaeva and Bogdanov 2014; Rosenberg and Corbett 2015; Baker et al. 2017).

We conclude that evolutionary solutions to WGD-associated challenges vary strongly from case to case, suggesting less functional constraint than one may expect based on the fact that these processes are conserved and essential. This may help explain how many species manage to thrive following WGD and, once established as polyploids, experience evolutionary success. In fact, we envision that the meiotic instability experienced by some WGD lineages, such as *C. amara*, could serve as a diversity-generating engine promoting large effect genomic structural variation, as has been observed in aggressive polyploid gliomas (Yant and Bomblies 2015).

Methods

Reference Genome Assembly and Alignment

We generated a *de novo* assembly using the 10x Genomics Chromium approach. In brief, a single diploid individual from pop LUZ (Supplementary Table 8) was used to generate one Chromium library, sequenced using 250PE mode on an Illumina sequencer, and assembled with Supernova version 2.0.0. This assembly had an overall scaffold N50 of 1.82mb. An assessment of genome completeness using BUSCO (version 3.0.2) (Seppey *et al.*, 2019) for the 2,251 contigs ≥ 10 kb was estimated at 94.8% (1365/1440 BUSCO groups; Supplementary Table 9).

BioNano Plant Extraction protocol

Fresh young leaves of the *C. amara* accession LUZ were collected after 48-hour dark treatment. DNA was extracted by the Earlham Institute's Platforms and Pipelines group following an IrysPrep "FixnBlend" Plant DNA extraction protocol supplied by BioNano Genomics. First 2.5 g of fresh young leaves were fixed with 2% formaldehyde. After washing, leaves were disrupted and homogenized in the presence of an isolation buffer containing PVP10 and BME to prevent polyphenol oxidation. Triton X-100 was added to facilitate nuclei release. Nuclei were then purified on a Percoll cushion. The nuclear phase was taken and washed in isolation buffer before embedding into low melting point agarose. Two plugs of 90 μ l were cast using the CHEF Mammalian Genomic DNA Plug Kit (Bio-Rad 170-3591). Once set at 4°C the plugs were added to a lysis solution containing 200 μ l proteinase K (QIAGEN 158920) and 2.5 ml of BioNano lysis buffer in a 50 ml conical tube. These were put at 50°C for 2 hours on a thermomixer, making a fresh proteinase K solution to incubate overnight. Samples were then removed from the thermomixer for 5 minutes before 50 μ l RNase A (Qiagen158924) was added and the tubes incubated for a further hour at 37°C. Plugs were then washed 7 times in the Wash Buffer supplied in the Chef kit and 7 times in 1xTE. One plug was removed and melted for 2 minutes at 70°C followed by 5 minutes at 43°C before adding 10 μ l of 0.2 U/ μ l of GELase (Cambio Ltd G31200). After 45 minutes at 43°C the melted plug was dialysed on a 0.1 μ M membrane (Millipore VCWP04700) sitting on 15 ml of 1xTE in a small petri dish. After 2 hours the sample was removed with a wide bore tip and mixed gently and left overnight at 4°C.

10X library construction

DNA material was diluted to 0.5 ng/ul with EB (Qiagen) and checked with a QuBit Fluorometer 2.0 (Invitrogen) using the QuBit dsDNA HS Assay kit (Supplementary Table 8). The Chromium User Guide was followed as per the manufacturer's instructions (10X Genomics, CG00043, Rev A). The final library was quantified using qPCR (KAPA Library Quant kit [Illumina] and ABI Prism qPCR Mix, Kapa Biosystems). Sizing of the library fragments was checked using a Bioanalyzer (High Sensitivity DNA Reagents, Agilent). Samples were pooled based on the molarities calculated using the two QC measurements. The library was clustered at 8 pM with a 1% spike in of PhiX library (Illumina). The pool was run on a HiSeq2500 250bp Rapid Run V2 mode (Illumina).

Sequencing and assembly

Reads were subsampled to 90 M reads and assembled with Supernova 2.0.0 (10x Genomics), giving a raw coverage of 60.30x and an effective coverage of 47.43x. The estimated molecule length was 44.15 kb. The assembly size, considering only scaffolds longer than 10kb was 159.53 Mb and the Scaffold N50 was 1.82 MB. Genome size estimate by kmer analysis was 225.39 MB, hence we estimate we are missing 16.61% from the assembly. Because the diploid individual used for reference genome sequencing was not homozygous, we sought to confirm whether the assembly harboured evidence of uncollapsed haplotypes by using a reciprocal BLAST (BLAST 2009) best hits approach. A small proportion (1.7%) of scaffolds exhibited substantial homology (90% or greater identity to another scaffold over 90% of their length), indicating that very few alternate alleles at heterozygous loci were misinferred as separate genomic loci in the diploid assembly. Manual investigation of a suite of meiosis-related loci indicated no cases of false negatives in the data set caused by alternate alleles aligning to separate scaffolds. We further scaffolded the assembly using the published *Cardamine hirsuta* genome using *graphAlign* (Spalding and Lammers 2004) and *Nucmer* (Marçais et al. 2018).

Gene Calling and Annotation

The plants set database *embryophyta_odb9.tar.gz* was downloaded from <http://busco.ezlab.org/> and used to assess orthologue presence/absence in our *C. amara* genome annotation. Running BUSCO gave Augustus (Stanke and Waack 2003) results via BUSCO HMMs to infer where genes lie in the assembly and to infer protein sequences. Augustus was used to generate a gff annotation file using 'arabidopsis' as the training option. A BLAST (v. 2.2.4) database was built for Brassicales (taxid: 3699) by downloading ~ 1.26M protein sequences from <https://www.ncbi.nlm.nih.gov/taxonomy/> and the

Augustus-predicted proteins were annotated via Interproscan (Quevillon et al. 2005) and blast2go (Conesa and Götz 2008).

Functional Annotation of *C. amara* genes

To functionally annotate *C. amara* genes we performed an orthogrouping analysis using Orthofinder version 2.3.3 (Emms and Kelly 2018), inferring orthologous groups (OGs) from four species (*C. amara*, *A. lyrata*, *A. thaliana*, *Cochlearia pyrenaica*). A total of 21,618 OGs were found. Best reciprocal blast hits (RBHs) for *C. amara* and *A. thaliana* genes were found using BLAST version 2.9.0.

Cardamine amara genes were then assigned an *A. thaliana* gene ID for GO enrichment analysis via the following protocol: 1) if the *C. amara* gene was in an OG with only one *A. thaliana* gene, that *A. thaliana* ID was used; 2) if the *C. amara* gene was in an OG with more than one *A. thaliana* gene, then the RBH, provided it was in the same OG with the *C. amara* gene, was used; 3) if the *C. amara* gene was in an OG that contained more than one *A. thaliana* gene, none of which was the RBH, then the *A. thaliana* gene from that OG with the lowest BLAST E-value was taken; 4) if the *C. amara* gene was in an OG group that lacked *A. thaliana* genes, then the RBH was taken instead; 5) Finally, if the *C. amara* gene was in an OG group without any *A. thaliana* genes and there was no RBH, then the gene with the lowest E-value in a BLASTs versus the TAIR10 database was used. BLASTs versus the TAIR10 database were performed during December 2019.

Sampling, sequencing and genetic structure analysis

Sampling

A total of 100 plants were sampled from four populations (Fig. 1d): 25 individuals for each of CEZ (4x), PIC (4x), VKR (2x), and LUZ (2x). Sampled plants were spaced at least 3 m apart, as such distance was enough to avoid resampling of identical clones according to analysis in a study sampling ~3,300 individuals across the *C. amara* range, including these populations (Zozomová-Lihová et al. 2015).

Flow Cytometry

All plants used for DNA extraction were verified for expected ploidy by flow cytometry. Approximately 1 square cm of leaf material was diced alongside an internal reference using a razor blade in 1 ml ice cold extraction buffer (45 mM MgCl₂, 30 mM sodium citrate, 20 mM MOPS, 1% Triton-100, pH 7 with

NaOH). The resultant slurry was then filtered through a 40- μ m nylon mesh before the nuclei were stained with the addition of 1 ml staining buffer (either CyStain UV precise P [Sysmex, Fluorescence emission: 435 nm to 500 nm] for relative ploidy, or Otto 2 buffer [0.4 M Na₂HPO₄·12H₂O, Propidium iodide 50 μ g/mL, RNase 50 μ g/mL], for absolute DNA content). After 1 minute of incubation at room temperature the sample was run for 5,000 particles on either a Partec PA II flow cytometer or a BD FACS Melody. Histograms were evaluated using FlowJo software version 10.6.1.

DNA isolation, library preparation and sequencing

A replicated approach was used for the DNA isolation, pooling, and sequencing to reduce variation that may be associated with Pool-Seq data. DNA isolations were performed in triplicate for every plant and then each replicate was pooled with samples from the other 24 replicates in each population, generating three independently extracted and pooled replicates for every population. DNA was extracted with the RNeasy Plant Mini Kit (Qiagen). Each of the 12 resultant pools for the 4 populations was used as input for library construction with the Illumina Truseq kit (Illumina, Inc.), and then sequenced on an Illumina NextSeq (150 bp paired end specification).

Data preparation, alignment, and genotyping

Fastq files from two runs on the Illumina NextSeq concatenated to give an average of 30.5 million reads per sample. Adapter sequences were removed using cutadapt (version 1.9.1) (Martin 2011) and quality trimmed via Sickle (version 33) (Joshi and Fass 2011) to generate only high-quality reads (Phred score \geq 30) of 30bp or more, resulting in an average of 27.9 million reads per sample. Reads were then aligned with (Li et al. 2009) BWA (version 0.7.12) (Li and Durbin 2009) and processed with Samtools (version 1.7) (Li et al. 2009). Using Picard (version 1.134) (Broad Institute 2009), duplicate reads were removed via MarkDuplicates followed by the addition of read group IDs to the bam files via AddOrReplaceReadGroups. Finally, to handle the presence of indels, GATK (version 3.6.0) (McKenna et al. 2010) was used to realign reads using IndelRealigner.

Variant Calling

Variants were called for the 12 bam files (three replicates per population) using Freebayes (version 1.1.0.46) (Garrison and Marth 2012) to generate a single VCF output file. Freebayes was run with default parameters, except we specified "--pooled-discrete" to indicate samples were pooled, "--use-

best-n-alleles 2" to restrict to biallelic sites, and "--no-indels" to exclude indels. The resultant VCF was then filtered with BCFtools (version 1.8) (Narasimhan et al. 2016) to remove sites where the read depth was less than 10, or greater than 1.6x the second mode (determined as $1.6 \times 31 = 50$, Supplementary Fig. 3) in order to remove from the analysis regions exhibiting heterozygous deletions or where multiple genomic regions may have mapped to the reference due to e.g. paralogous duplications in the sequenced individuals.

Population genetic structure

We first calculated genome-wide between-population metrics (Nei's F_{st} (Nei 1972) and allele frequency difference). The Allele Frequency (AF) in individual replicate pools was calculated as the fraction of the total number of reads supporting the alternative allele (Anand et al. 2016). For each population the average AF was then calculated from the three replicates and used for all further calculations. We used the python3 PoolSeqBPM pipeline, designed to input pooled data (<https://github.com/mbohutinska/PoolSeqBPM>). Then we inferred relationships between populations over putatively neutral four-fold degenerate SNPs using PCA as implemented in *adegenet* (Jombart and Ahmed 2011). Finally, we inferred relationships between populations using allele frequency covariance graphs implemented in TreeMix (Pickrell and Pritchard 2012). We ran TreeMix allowing a range of migration events; and presented one additional migration edge, as it represented points of log-likelihood saturation. To obtain confidence in the reconstructed topology, we bootstrapped the scenario with zero events (the tree topology had not changed when considering the migration events) choosing a bootstrap block size of 1000 bp, equivalent to the window size in our selection scan, and 100 replicates.

Genome scans for selection

To detect signals of selection, we used a combination of two different selection scan approaches. First, we calculated pairwise window-based F_{st} between diploid and polyploid populations and used minimum sum of ranks between informative contrasts in a quartet design (below). To further control for false positives and refine the gene list to putatively functional candidates we complemented these differentiation measures with a functional score estimate following the FineMAV method (below). Both approaches are based on population allele frequencies and allow analysis of diploid and autopolyploid populations.

Window-based selection scan using a quartet design

We performed a window-based F_{st} (Nei 1972) scan for directional selection in *C. amara*, taking advantage of quartet sampling of two diploid and two autotetraploid populations (Fig. 1d). Using this design we identified top candidate windows for selective sweeps associated with ploidy differentiation, while excluding differentiation patterns private to a single population or ploidy-uninformative selective sweeps. Thus comparisons between populations of the same ploidy constitute a null model for shared heterogeneity in genetic differentiation arising through processes unrelated to WGD (following an approach successfully applied in Vijay et al. 2016). To do this, we calculated F_{st} for 1 kb windows with minimum 20 SNPs for all six population pairs in the quartet (Fig. 1d) and ranked windows based on their F_{st} value. We excluded windows which were top 1% outliers in diploid-diploid (dd in Fig. 1d) or autotetraploid-autotetraploid (tt) population contrasts, as they represent variation inconsistent with diploid-autotetraploid divergence but rather signal local differentiation within a cytotype. Next, we assigned ranks to each window based on the F_{st} values in four diploid-autotetraploid contrasts and identified windows being top 1% outliers of minimum rank sum.

Because candidate detection could be biased by arbitrary window size choice, we re-analysed our differentiation scans changing two parameters: 1) using a SNP-by-SNP basis (requiring at least five SNPs per gene for inclusion); and 2) using larger, 50 kb windows. Doing this, we found that SNP-level and 1 kb-window scans resulted in comparable candidate gene lists, while 50kb windows were too wide to identify local peaks of differentiation. Thus, we decided to use scans with a window size of 1 kb, which best corresponded to the average length of selective sweep signatures in differentiation plots (e.g. Fig. 2), and allowed to locate the candidate selected region while still providing enough polymorphisms to robustly estimate differentiation between ploidies.

To account for possible confounding effect of comparing windows from genic and non-genic regions, we calculated the number of base pairs overlapping with any gene within each window. There was no relationship between the proportion of genic space within a window and F_{st} (Pearsons $r = -0.057$, Supplementary Fig. 4), indicating that our analyses were unaffected by unequal proportion of genic space in a window.

In *A. arenosa*, we performed window-based F_{st} scan for directional selection using the same criteria as for *C. amara* (1kb windows, min 20 SNPs per window). We did not use the quartet design as the range-wide dataset of 80 diploid and 40 autotetraploid individuals drawn from the entire *A. arenosa* range (15 diploid and 24 autotetraploid populations) assured power to detect genomic regions with WGD-associated differentiation. This *A. arenosa* analysis gave very similar results to (Yant et al. 2013), which used only 2 diploid and 4 autotetraploid populations, indicating minimal dependence on sampling to detect these strongest signatures of selection in the *A. arenosa* system.

FineMAV

We adopted the approach, Fine-Mapping of Adaptive Variation, FineMAV (Szpak et al. 2018), using our *C. amara* annotation (following approach successfully applied to non-human genome in Bohutinska, et al. 2021). To functionally annotate each amino acid change, we used the Grantham score (Grantham 1974), a theoretical amino acid substitution value, encoded in the Grantham matrix, where each element shows the differences of physicochemical properties between two amino acids. We used SnpEff (version 4.3) (Cingolani et al. 2012) to annotate our SNP dataset by applying our Augustus-generated *C. amara* annotation ('Gene Calling and Annotation,' above). We estimated the population genetic component of FineMAV (see (Szpak et al. 2018) for details on calculations) using allele frequency information at each site (considering minor frequency alleles as derived) and derived allele purity (DAP) parameter of 3.5, a measure of population differentiation, which describes how unequally the derived allele is distributed among populations. The advantage is that DAP can summarize differentiation across many populations in a single measure for each variant. Finally, for each amino acid substitution, we assigned Grantham scores, together with population genetic component of FineMAV, using custom scripts in Python 2.7.10 and Biopython version 1.69. We identified the top 1% outliers as FineMAV candidates. All calculations were performed using code available at (github.com/paajanen/meiosis_protein_evolution).

***Arabidopsis arenosa* population genomic dataset**

Our selection analysis in *A. arenosa* was based on an expanded sampling (Monnahan et al. 2019) relative to (Yant et al. 2013), who sampled 24 individuals (from 2 diploid and 4 tetraploid populations, sourced from a fraction of now known lineages). This expanded sampling covered all known lineages,

across the entire range of the species, including 39 populations: 15 diploid populations (105 individually resequenced plants) and 24 tetraploid populations (182 individually resequenced plants) (Monnahan et al. 2019). We aligned PE Illumina data to the *A. lyrata* reference (Hu et al. 2011), called variants and filtered as previously (Monnahan et al. 2019) using GATK 3.5 (McKenna et al. 2010). We used a subset of the data consisting of 80 diploid individuals and 40 tetraploid individuals from populations unaffected by secondary introgression from diploid lineages (following Bohutinska et al. 2021; samples selected based on the highest mean depth of coverage). Such sub-sampling gave us a balanced number of 160 high-quality haploid genomes of each ploidy suitable for selection scans. Finally, we filtered each subsampled dataset for genotype read depth > 8 and maximum fraction of missing genotypes < 0.5 in each lineage. We calculated *F_{st}* using python3 ScanTools pipeline (github.com/mbohutinska/ScanTools_ProtEvol). All subsequent analyses were performed following the same procedure as with *C. amara* data.

GO enrichment analysis

To infer functions significantly associated with directional selection following WGD, we performed a gene ontology enrichment on the gene list using the R package topGO (Tilford and Siemers 2009), using *A. thaliana* orthologs of *C. amara*/*A. lyrata* genes, obtained using biomaRt (Smedley et al. 2009). We used Fisher's exact test with conservative 'elim' method, which tests for enrichment of terms from the bottom of the GO hierarchy to the top and discards any genes that are significantly enriched in a descendant GO terms (Grossmann et al. 2007). Re-analysis with the 'classic' method did not identify any additional convergently enriched GO terms. We used biological process ontology with minimum node size of 150 genes.

Protein associations from STRING database

We searched for potential functional associations among *C. amara* and *A. arenosa* candidate genes using STRING (Szklarczyk et al. 2015). Genes were assigned an *A. thaliana* gene ID as described above. We used the 'multiple proteins' search in *A. thaliana*, with text mining, experiments, databases, co-expression, neighbourhood, gene fusion and co-occurrence as information sources. We used minimum confidence 0.4 and retained only 1st shell associations (proteins that are directly associated with the candidate protein: i.e., immediately neighbouring network circles).

Quantifying convergence

We considered convergent any candidates or enriched GO categories that overlapped across both species. Convergent candidate genes had to be members of the same orthogroups (Emms and Kelly 2018). To test for higher than random number of overlapping items we used Fisher's Exact Test for Count Data in R (R Development Core Team 2011).

Cytological assessment of meiotic stability

We cytologically estimated the degree of male meiotic stability in *C. amara* by counting the number of bivalent chromosome associations in each metaphase event. A lower number of bivalents and a higher number of multivalents is taken as a proxy for reduced meiotic stability. The reasoning behind this is that a reduction in crossover number to one per bivalent is strongly indicated as a leading mechanism for meiotic diploidization in autopolyploids as this limits multivalent associations (which increase the propensity toward breakage and aneuploidy vs bivalents (Cifuentes *et al.*, 2010; Le Comber *et al.*, 2010; Bomblies *et al.*, 2016)).

Chromosome preparation

Whole young inflorescences were fixed in freshly prepared ethanol:acetic acid (3:1) overnight, transferred into 70% ethanol and stored at -20 °C until use. Meiotic chromosome spreads were prepared from anthers according to (Mandáková *et al.* 2014). Briefly, after washing in citrate buffer (10 mM sodium citrate, pH 4.8), selected flower buds were digested using a 0.3% mix of pectolytic enzymes (cellulase, cytohelicase, pectolyase; Sigma-Aldrich Corp., St. Louis, MO) in citrate buffer for 3 hours. Individual anthers were dissected and spread in 20 µl of 60% acetic acid on a microscope slide placed on a metal hot plate (50°C), fixed by ethanol:acetic acid (3:1) and the preparation was dried using a hair dryer. Slides were postfixed in freshly prepared 4% formaldehyde in distilled water for 10 min and air-dried. The preparations were stained with 4',6-diamidino-2-phenylindole (DAPI; 2 µg/ml) in Vectashield (Vector Laboratories, Peterborough, UK). Fluorescence signals were analysed using an Axioimager Z2 epifluorescence microscope (Zeiss, Oberkochen, Germany) and CoolCube CCD camera (MetaSystems, Newton, MA).

Meiotic stability assessments

In diploids, chromosome spreads with 8 bivalents were scored as "stable meiosis", 7-6 as "partly stable", 5-4 as "partly unstable", and <4 as "unstable". In autotetraploids, chromosome spreads with 16 bivalents were scored as "stable meiosis", 14-12 as "partly stable", 10-8 as "partly unstable", and <8 as "unstable". We report a mean value of meiotic stability for each ploidy calculated over "stable meiosis" and over sum of "stable meiosis" and "partly stable" categories. Differences in meiotic stability between diploids and autotetraploids (Fig. 3b) are reported for the sum of "stable" and "partly stable" categories. However, considering only the "stable meiosis" category does not qualitatively affect the results (i.e. the degree of meiotic stability is significantly lower in tetraploids, $D = 125.7$, $df = 1$, $p < 0.0001$, GLM with binomial errors). Photos of all spreads scored are supplied in Supplementary Fig. 4.

Data Availability Statement

Sequence data are available at the European Nucleotide Archive at <http://www.ebi.ac.uk/ena/data/view/PRJEB39872> (*C. amara*) and the Sequence Read Archive at <https://www.ncbi.nlm.nih.gov/sra/>, study code SRP156117 (*A. arenosa* data). All scripts are available at <https://github.com/mbohutinska/PoolSeqBPM> (Fst-based selection scans and all following analyses) and https://github.com/paajanen/meiosis_protein_evolution (FineMAV scan).

Acknowledgements

The authors thank Veronika Konečná for assistance with the map figure, Doubravka Požárová and Paolo Bartolić for help with field collections and Karol Marhold for useful discussions. We also thank the anonymous reviewers for their constructive comments. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [grant number ERC-StG 679056 HOTSPOT], via a grant to LY. Additional support was provided by Czech Science Foundation (project 20-22783S to FK, 19-03442S to TM and 19-06632S), by Charles University (project Primus/SCI/35 to FK), by the long-term research development project No. RVO 67985939 of the Czech Academy of Sciences and by the CEITEC 2020 project (grant LQ1601). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations

Infrastructures. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

LY conceived of and directed the study. MB, SB, MA, PP, TM and PM performed analyses. PM and TM performed laboratory experiments. PM, FK, SB, and MB performed field collections. LY and MB wrote the manuscript with input from all authors. All authors approved of the final manuscript.

Competing Interests statement

The authors declare no competing interests.

Materials & Correspondence

All requests should be addressed to Levi Yant (levi.yant@nottingham.ac.uk) or Magdalena Bohutínská (magdalena.holcova@natur.cuni.cz)

References

- Aichinger E, Villar CBR, di Mambro R, Sabatini S, Köhler C. 2011. The CHD3 chromatin remodeler PICKLE and polycomb group proteins antagonistically regulate meristem activity in the Arabidopsis root. *Plant Cell* 23:1047–1060.
- Anand S, Mangano E, Barizzone N, Bordoni R, Sorosina M, Clarelli F, Corrado L, Boneschi FM, D’Alfonso S, De Bellis G. 2016. Next generation sequencing of pooled samples: Guideline for variants’ filtering. *Sci. Rep.* 6.
- Arnold B, Kim ST, Bomblies K. 2015. Single geographic origin of a widespread autotetraploid Arabidopsis arenosa lineage followed by interploidy admixture. *Mol. Biol. Evol.* 32:1382–1395.
- Arnold BJ, Lahner B, DaCosta JM, Weisman CM, Hollister JD, Salt DE, Bomblies K, Yant L. 2016. Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 113:8320–8325.
- Baduel P, Bray S, Vallejo-Marin M, Kolář F, Yant L. 2018. The “Polyploid Hop”: Shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Front. Ecol. Evol.* 6.
- Baduel P, Hunter B, Yeola S, Bomblies K. 2018. Genetic basis and evolution of rapid cycling in railway populations of tetraploid Arabidopsis arenosa. *PLoS Genet.* 14.
- Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *Elife* 6.
- Birkeland S, Gustafsson ALS, Brysting AK, Brochmann C, Nowak MD, Purugganan M. 2020. Multiple Genetic Trajectories to Extreme Abiotic Stress Adaptation in Arctic Brassicaceae. *Mol. Biol. Evol.* 37:2052–2068.
- BLAST. 2009. Nucleotide BLAST: Search nucleotide databases using a nucleotide query. *Basic Local Alignment Search Tool* [Internet]. Available from: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome
- Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life’s tape. *Science* 362.
- Bohutínská M, Handrick V, Yant L, Schmickl R, Kolář F, Bomblies K, Paajanen P. 2021. De-novo mutation and rapid protein (co-)evolution during meiotic adaptation in Arabidopsis arenosa. *Mol. Biol. Evol.*
- Bohutínská M, Vlček J, Yair S, Laenen B, Konečná V, Fracassetti M, Slotte T, Kolář F. 2020. Genomic basis of parallel adaptation varies with divergence in Arabidopsis and its relatives. *bioRxiv*.
- Bomblies K. 2020. When everything changes at once: Finding a new normal after genome duplication: Evolutionary response to polyploidy. *Proc. R. Soc. B Biol. Sci.* 287.
- Bomblies K, Higgins JD, Yant L. 2015. Meiosis evolves: Adaptation to external and internal environments. *New Phytol.* 208:306–323.
- Bomblies K, Jones G, Franklin C, Zickler D, Kleckner N. 2016. The challenge of evolving stable polyploidy: could an increase in “crossover interference distance” play a central role? *Chromosoma* 125:287–300.
- Bomblies K, Madlung A. 2014. Polyploidy in the Arabidopsis genus. *Chromosom. Res.* 22:117–134.
- Broad Institute. 2009. Picard Tools - By Broad Institute. *GitHub*.
- Bundock P, Hooykaas P. 2005. An Arabidopsis hAT-like transposase is essential for plant development. *Nature* 436:282–284.
- Chao DY, Dilkes B, Luo H, Douglas A, Yakubova E, Lahner B, Salt DE. 2013. Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science* 341:658–659.
- Cifuentes M, Grandont L, Moore G, Chèvre AM, Jenczewski E. 2010. Genetic regulation of meiosis in

- polyploid species: New insights into an old question. *New Phytol.* 186:29–36.
- Cifuentes M, Jolivet S, Cromer L, Harashima H, Bulankova P, Renne C, Crismani W, Nomura Y, Nakagami H, Sugimoto K, et al. 2016. TDM1 Regulation Determines the Number of Meiotic Divisions. *PLoS Genet.* 12.
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 6:80–92.
- Colmenero-Flores JM, Martínez G, Gamba G, Vázquez N, Iglesias DJ, Brumós J, Talón M. 2007. Identification and functional characterization of cation-chloride cotransporters in plants. *Plant J.* 50:278–292.
- Le Comber SC, Ainouche ML, Kovarik A, Leitch AR. 2010. Making a functional diploid: From polysomic to disomic inheritance. *New Phytol.* 186:113–122.
- Conesa A, Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008.
- Culligan KM, Hays JB. 2000. Arabidopsis MutS homologs - AtMSH2, AtMSH3, AtMSH6, and a novel AtMSH7 - Form three distinct protein heterodimers with different specificities for mismatched DNA. *Plant Cell* 12:991–1002.
- Cutler SR, Rodriguez PL, Finkelstein RR, Abrams SR. 2010. Abscisic acid: Emergence of a core signaling network. *Annu. Rev. Plant Biol.* 61:651–679.
- Doyle JJ, Coate JE. 2019. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.* 180:1–52.
- Van Drunen WE, Husband BC. 2019. Evolutionary associations between polyploidy, clonal reproduction, and perenniality in the angiosperms. *New Phytol.* 224:1266–1277.
- Elmer KR, Meyer A. 2011. Adaptation in the age of ecological genomics: Insights from parallelism and convergence. *Trends Ecol. Evol.* 26:298–306.
- Emms DM, Kelly S. 2018. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *bioRxiv* 466201. Available from: <https://www.biorxiv.org/content/10.1101/466201v1>
- Fujii H, Verslues PE, Zhu JK. 2011. Arabidopsis decuple mutant reveals the importance of SnRK2 kinases in osmotic stress responses in vivo. *Proc. Natl. Acad. Sci. U. S. A.* 108:1717–1722.
- Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio R Dello, Hofhuis H, Pieper B, Cartolano M, Neumann U, et al. 2016. The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. *Nat. Plants* 2.
- Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. Available from: <http://arxiv.org/abs/1207.3907>
- Glover J, Grelon M, Craig S, Chaudhury A, Dennis E. 1998. Cloning and characterization of MS5 from Arabidopsis: A gene critical in male meiosis. *Plant J.* 15:345–356.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.
- Grishaeva TM, Bogdanov YF. 2014. Conservation and Variability of Synaptonemal Complex Proteins in Phylogenesis of Eukaryotes. *Int. J. Evol. Biol.* 2014:1–16.
- Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. *Bioinformatics* 23:3024–3031.
- Hejny S, Slavik B, Kirschner J, Krísa B. 1992. Květena České republiky 3. Academia
- Herben T, Suda J, Klimešová J. 2017. Polyploid species rely on vegetative reproduction more than diploids: A re-examination of the old hypothesis. *Ann. Bot.* 120:341–349.
- Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K. 2012. Genetic Adaptation Associated with Genome-Doubling in Autotetraploid Arabidopsis arenosa. *PLoS Genet.* 8.
- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome

- size change. *Nat. Genet.* 43:476–483.
- Huang XC, German DA, Koch MA. 2020. Temporal patterns of diversification in Brassicaceae demonstrate decoupling of rate shifts and mesopolyploidization events. *Ann. Bot.* 125:29–47.
- Hudson RR, Coyne JA. 2002. Mathematical consequences of the genealogical species concept. *Evolution (N. Y.)* 56:1557–1565.
- Jing Y, Guo Q, Lin R. 2019. The chromatin-remodeling factor pickle antagonizes polycomb repression of FT to promote flowering. *Plant Physiol.* 181:656–668.
- Jombart T, Ahmed I. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070–3071.
- Joshi N, Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at <https://github.com/najoshi/sickle>:2011.
- Kang X, Xu G, Lee B, Chen C, Zhang H, Kuang R, Ni M. 2018. HRB2 and BBX21 interaction modulates Arabidopsis ABI5 locus and stomatal aperture. *Plant Cell Environ.* 41:1912–1925.
- Kawa D, Meyer AJ, Dekker HL, Abd-El-Halim AM, Gevaert K, Van De Slijke E, Maszkowska J, Bucholc M, Dobrowolska G, De Jaeger G, et al. 2020. SnRK2 protein kinases and mRNA decapping machinery control root development and response to salt. *Plant Physiol.* 182:361–371.
- Klepikova A V., Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. 2016. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.* 88:1058–1070.
- Knip M (Leiden U. 2012. Daysleeper : from genomic parasite to indispensable gene.
- Koch M, Huthmann M, Bernhardt KG. 2003. Cardamine amara L. (Brassicaceae) in dynamic habitats: Genetic composition and diversity of seed bank and established populations. *Basic Appl. Ecol.*
- Kolář F, Fuxová G, Závěská E, Nagano AJ, Hyklová L, Lučanová M, Kudoh H, Marhold K. 2016. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model Arabidopsis arenosa. *Mol. Ecol.* 25:3929–3949.
- Konečná V, Bray S, Vlček J, Bohutínská M, Požárová D, Choudhury RR, Bollmann-Giolai A, Flis P, Salt DE, Parisod C, et al. 2021. Parallel adaptation in autopolyploid Arabidopsis arenosa is dominated by repeated recruitment of shared alleles. *bioRxiv* [Internet]:2021.01.15.426785. Available from: <http://biorxiv.org/content/early/2021/01/17/2021.01.15.426785.abstract>
- Kunz HH, Gierth M, Herdean A, Satoh-Cruz M, Kramer DM, Spetea C, Schroeder JI. 2014. Plastidial transporters KEA1, -2, and -3 are essential for chloroplast osmoregulation, integrity, and pH regulation in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 111:7480–7485.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
- Liu CM, Meinke DW. 1998. The titan mutants of Arabidopsis are disrupted in mitosis and cell cycle control during seed development. *Plant J.* 16:21–31.
- Lu X, Liu X, An L, Zhang W, Sun J, Pei H, Meng H, Fan Y, Zhang C. 2008. The Arabidopsis MutS homolog AtMSH5 is required for normal meiosis. *Cell Res.* 18:589–599.
- Mandáková T, Marhold K, Lysak MA. 2014. The widespread crucifer species Cardamine flexuosa is an allotetraploid with a conserved subgenomic structure. *New Phytol.* 201:982–992.
- Marburger S, Monnahan P, Seear PJ, Martin SH, Koch J, Paaajanen P, Bohutínská M, Higgins JD, Schmickl R, Yant L. 2019. Interspecific introgression mediates adaptation to whole genome duplication. *Nat. Commun.* [Internet] 10. Available from: <https://doi.org/10.1038/s41467-019-13159-5>
- Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* 14.
- Marhold K, Huthmann M, Hurka H. 2002. Evolutionary history of the polyploid complex of

- Cardamine amara (Brassicaceae): Isozyme evidence. *Plant Syst. Evol.* [Internet] 233:15–28. Available from: <http://www.jstor.org/stable/23644306>
- Martin A, Orgogozo V. 2013. The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution (N. Y.)*. 67:1235–1250.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10.
- Mason AS, Pires JC. 2015. Unreduced gametes: Meiotic mishap or evolutionary mechanism? *Trends Genet.* 31:5–10.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.
- Meyers BC, Morgante M, Michelmore RW. 2002. TIR-X and TIR-NBS proteins: Two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. *Plant J.* 32:77–92.
- Molina-Henao YF, Hopkins R. 2019. Autopolyploid lineage shows climatic niche expansion but not divergence in Arabidopsis arenosa. *Am. J. Bot.* 106:61–70.
- Monnahan P, Kolář F, Baduel P, Sailer C, Koch J, Horvath R, Laenen B, Schmickl R, Paajanen P, Šrámková G, et al. 2019. Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. *Nat. Ecol. Evol.* 3:457–468.
- Morgan C, Zhang H, Henry CE, Franklin CFH, Bomblies K. 2020. Derived alleles of two axis proteins affect meiotic traits in autotetraploid Arabidopsis arenosa. *Proc. Natl. Acad. Sci. U. S. A.* 117:8980–8988.
- Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. 2016. BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32:1749–1751.
- Nei M. 1972. Genetic Distance between Populations. *Am. Nat.* 106:283–292.
- Novikova PY, Brennan IG, Booker W, Mahony M, Doughty P, Lemmon AR, Lemmon EM, Dale Roberts J, Yant L, de Peer Y Van, et al. 2020. Polyploidy breaks speciation barriers in Australian burrowing frogs Neobatrachus. *PLoS Genet.* 16.
- Ogas J, Kaufmann S, Henderson J, Somerville C. 1999. PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 96:13839–13844.
- Panizza S, Tanaka T, Hochwagen A, Eisenhaber F, Nasmyth K. 2000. Pds5 cooperates with cohesion in maintaining sister chromatid cohesion. *Curr. Biol.* 10:1557–1564.
- Van De Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18:411–424.
- Perruc E, Kinoshita N, Lopez-Molina L. 2007. The role of chromatin-remodeling factor PKL in balancing osmotic stress responses during Arabidopsis seed germination. *Plant J.* 52:927–936.
- Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* 8.
- Preite V, Sailer C, Syllwasschy L, Bray S, Ahmadi H, Krämer U, Yant L. 2019. Convergent evolution in Arabidopsis halleri and Arabidopsis arenosa on calamine metalliferous soils. *Philos. Trans. R. Soc. B Biol. Sci.* 374.
- Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: Protein domains identifier. *Nucleic Acids Res.* 33.
- R Development Core Team R. 2011. R: A Language and Environment for Statistical Computing. Available from: <http://www.r-project.org>
- Rentsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47:D886–D894.

- Rosenberg SC, Corbett KD. 2015. The multifaceted roles of the HOR MA domain in cellular signaling. *J. Cell Biol.* 211:745–755.
- Schmickl R, Koch MA. 2011. Arabidopsis hybrid speciation processes. *Proc. Natl. Acad. Sci. U. S. A.* 108:14192–14197.
- Schmickl R, Yant L. 2021. Adaptive introgression: how polyploidy reshapes gene flow landscapes. *New Phytol.*
- Seear P, France M, Gregory C, Heavens D, Schmickl R, Yant L, Higgins J. 2020. A novel allele of ASY3 promotes meiotic stability in autotetraploid Arabidopsis lyrata. *PLoS Genet.* [Internet] 16(7). Available from: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008900>
- Selmecki AM, Maruvka YE, Richmond PA, Guillet M, Shores N, Sorenson AL, De S, Kishony R, Michor F, Dowell R, et al. 2015. Polyploidy can drive rapid adaptation in yeast. *Nature* 519:349–351.
- Seppy M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing genome assembly and annotation completeness. In: *Methods in Molecular Biology*. Vol. 1962. p. 227–245.
- Shaked H, Avivi-Ragolsky N, Levy AA. 2006. Involvement of the arabidopsis SWI2/SNF2 chromatin remodeling gene family in DNA damage response and recombination. *Genetics* 173:985–994.
- Siddiqui NU, Stronghill PE, Dengler RE, Hasenkampf CA, Riggs CD. 2003. Mutations in Arabidopsis condensin genes disrupt embryogenesis, meristem organization and segregation of homologous chromosomes during meiosis. *Development* 130:3283–3295.
- Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009. BioMart - Biological queries made easy. *BMC Genomics* 10.
- Spalding JB, Lammers PJ. 2004. BLAST Filter and GraphicAlign: Rule-based formation and analysis of sets of related DNA and protein sequences. *Nucleic Acids Res.* 32.
- Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. In: *Bioinformatics*. Vol. 19.
- Stephan AB, Kunz HH, Yang E, Schroeder JI. 2016. Rapid hyperosmotic-induced Ca²⁺ responses in Arabidopsis thaliana exhibit sensory potentiation and involvement of plastidial KEA transporters. *Proc. Natl. Acad. Sci. U. S. A.* 113:E5242–E5249.
- Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43:D447–D452.
- Szpak M, Mezzavilla M, Ayub Q, Chen Y, Xue Y, Tyler-Smith C. 2018. FineMAV: Prioritizing candidate genetic variants driving local adaptations in human populations. *Genome Biol.* 19.
- Takuno S, Ralph P, Swart K, Elshire RJ, Glaubitz JC, Buckler ES, Hufford MB, Ross-Ibarra J. 2015. Independent molecular basis of convergent highland adaptation in maize. *Genetics* 200:1297–1312.
- Tamura K, Adachi Y, Chiba K, Oguchi K, Takahashi H. 2002. Identification of Ku70 and Ku80 homologues in Arabidopsis thaliana: Evidence for a role in the repair of DNA double-strand breaks. *Plant J.* 29:771–781.
- Tedder A, Helling M, Pannell JR, Shimizu-Inatsugi R, Kawagoe T, Van Campen J, Sese J, Shimizu KK. 2015. Female sterility associated with increased clonal propagation suggests a unique combination of androdioecy and asexual reproduction in populations of Cardamine amara (Brassicaceae). *Ann. Bot.* 115:763–776.
- Tilford CA, Siemers NO. 2009. Gene set enrichment analysis. *Methods Mol. Biol.* 563:99–121.
- Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* 7.
- Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf JBW. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow

- species complex. *Nat. Commun.* 7.
- Williams ME, Torabinejad J, Cohick E, Parker K, Drake EJ, Thompson JE, Hortter M, DeWald DB. 2005. Mutations in the Arabidopsis phosphoinositide phosphatase gene SAC9 lead to overaccumulation of PtdIns(4,5)P2 and constitutive expression of the stress-response pathway. *Plant Physiol.* 138:686–700.
- Wu SY, Culligan K, Lamers M, Hays J. 2003. Dissimilar mismatch-recognition spectra of Arabidopsis DNA-mismatch-repair proteins MSH2·MSH6 (MutS α) and MSH2·MSH7 (MutS γ). *Nucleic Acids Res.* 31:6027–6034.
- Yant L, Bomblies K. 2015. Genome management and mismanagement—cell-level opportunities and challenges of whole-genome duplication. *Genes Dev.* 29:2405–2419.
- Yant L, Hollister JD, Wright KM, Arnold BJ, Higgins JD, Franklin FCH, Bomblies K. 2013. Meiotic adaptation to genome duplication in Arabidopsis arenosa. *Curr. Biol.* 23:2151–2156.
- Yeaman S, Gerstein AC, Hodgins KA, Whitlock MC. 2018. Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genet.* 14.
- Zozomová-Lihová J, Malánová-Krásná I, Vít P, Urfus T, Senko D, Svitok M, Kempa M, Marhold K. 2015. Cytotype distribution patterns, ecological differentiation, and genetic structure in a diploid–tetraploid contact zone of *Cardamine amara*. *Am. J. Bot.* 102:1380–1395.

Supplementary figure captions

Supplementary Fig. 1: Comparable phylogenetic relationships and migration events between diploid and tetraploid populations of *C. amara* and *A. arenosa* inferred by TreeMix. X-axis shows the drift estimation, corresponding to the number of generations separating the two populations (t), and effective population size (N) (Pickrell and Pritchard, 2012). Node labels show bootstrap support and the arrow indicates the most likely migration event (migration weight, which can be interpreted as admixture proportion, = 0.18 and 0.19 for *C. amara* and *A. arenosa*, respectively). Additional migration events did not improve the model likelihood.

Supplementary Fig. 2: *C. amara* candidate meiosis gene associations as identified by STRING analysis. We used only medium confidence associations and higher (shown as thickness of lines connecting genes).

Supplementary Fig. 3: Distribution of read depth over all sequenced samples.

Supplementary Fig. 4: Relationship between the proportion of genic space within a window and F_{st} .

Supplementary table captions

Supplementary Table 1. GPS coordinates of population localities.

Supplementary Table 2. Mean depth of coverage (MDOC) per pool of individuals from each population.

Supplementary Table 3. GO terms enriched in *C. amara* WGD candidate genes.

Supplementary Table 4. Targeted search for patterns suggesting directional selection in *C. amara* orthologs of candidate *A. arenosa* meiosis genes.

Supplementary Table 5. Chromosome stability scoring of individual diploid ($2n = 16$) and autotetraploid ($2n = 32$) plants of *C. amara* at meiotic diakinesis and metaphase I.

Supplementary Table 6. GO terms enriched in *A. arenosa* WGD candidate genes.

Supplementary Table 6. Chromosome stability scoring of individual diploid ($2n = 16$) and autotetraploid ($2n = 32$) plants of *C. amara* at meiotic metaphase I.

Supplementary Table 7. *C. amara* candidate genes that have more than one associated protein among *A. arenosa* candidates by STRING analysis.

Supplementary Table 8. Quality checks of DNA isolated from LUZ.

Supplementary Table 9. Assessment of genome completeness using BUSCO.

Other supplementary material

Supplementary Datasets (separate excel file consisting of three worksheets)

Supplementary Dataset 1. Genes in the top 1% of Fst scores (1000 bp windows) in *C. amara*. Note: red lines denote six genes which are candidates also in *A. arenosa*.

Supplementary Dataset 2. Top 1% of amino acid substitutions with the highest fineMAV score.

Supplementary Dataset 3. Genes in the top 1% of Fst scores (1000 bp windows) in *A. arenosa*.

Supplementary Figure 5. (separate image file)

Case study 6.

Convergence and novelty in adaptation to whole genome duplication
in three independent polyploids.



Convergence and novelty in adaptation to whole genome duplication in three independent polyploids

Sian M. Bray¹, Eva M. Wolf², Min Zhou³, Silvia Busoms^{4,5}, Magdalena Bohutínská⁶,
Pirita Paajanen¹, Patrick Monnahan¹, Jordan Koch¹, Sina Fischer⁵, Marcus A. Koch²,
and Levi Yant^{7,*}

1. Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park NR4 7UH, Norwich, UK
2. Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, Im Neuenheimer Feld 345, 69120 Heidelberg, Germany
3. Chengdu Institute of Biology, Chinese Academy of Sciences, No.9, section 4 of South RenMin Road, Wuhou District, Chengdu 610041, Sichuan, China
4. Department of Plant Physiology, Universitat Autònoma de Barcelona 08193 Barcelona, Spain
5. Future Food Beacon and Division of Plant and Crop Sciences, School of Biosciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom
6. Institute of Botany, The Czech Academy of Sciences, Zámek 1, 252 43 Průhonice, Czech Republic
7. Future Food Beacon and School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom

***Author for correspondence:** levi.yant@nottingham.ac.uk; Tel: +44 7 966 731 125

Keywords: polyploidy; evolution; population genomics; convergent evolution

Abstract

Convergent evolution is observed broadly across the web of life, but the degree of evolutionary constraint during adaptation of core intracellular processes is not known. High constraint has been assumed for conserved processes, such as cell division and DNA repair, but reports of nimble evolutionary shifts in these processes have confounded this expectation. Whole genome duplication (WGD) necessitates the concerted adjustment of a wide range of fundamental intracellular functions but nevertheless has been repeatedly survived in all kingdoms. Given this repeated adaptation to WGD despite obvious intracellular challenges to core processes such as meiosis, we asked: how do lineages not only survive WGD, but sometimes ultimately thrive? Are the solutions employed constrained or diverse? Here we detect genes and processes under selection following WGD in the *Cochlearia* species complex by performing a scan for selective sweeps following WGD in a large-scale survey of 73 resequenced individuals from 23 populations across Europe. We then contrast our results from two independent WGDs in *Arabidopsis arenosa* and *Cardamine amara*. We find that while WGD does require the adaptation of particular functional processes in all three cases, the specific genes recruited to respond are highly flexible. We also observe evidence of varying degrees of convergence between different cases. Our results point to a polygenic basis for the distributed adaptive systems that control meiotic crossover number, ionomic rewiring, cell cycle control, and nuclear regulation. Given the sheer number of loci under selection post-WGD, we surmise that this polygenicity may explain the general lack of convergence between these species that are ~30 million years diverged. Based on our results, we speculate that adaptive processes themselves – such as the rate of generation of structural genomic variants—may be altered by WGD in nascent autopolyploids, contributing to the occasionally spectacular adaptability of autopolyploids observed across kingdoms.

Biologists have long been fascinated by the convergent evolution of similar traits in distant lineages¹. Given the tractability of population-based resequencing studies to detect candidate mechanisms underlying adaptations, the genomic underpinnings of convergence are increasingly coming to light². Diverse examples can be found in all kingdoms, from the convergent basis for the loss in flight in birds³, to the evolution of toxin-resistant herbivory in insects⁴, to drug resistance in pathogens⁵. These studies ask the fundamental question: to what degree is evolution constrained along a given adaptive trajectory? That is, do convergent traits, when detected, have a basis in similar or identical genetic changes? Ultimately, this drives at a bigger question: is evolution predictable? Overwhelmingly, these works focus on adaptation to external selection pressures. Here, in contrast, we investigate convergence in genomic responses to an array of internal physiological pressures resulting from whole genome duplication (WGD).

The duplication of an entire genome is a dramatic mutation that disrupts the most fundamental of cellular processes, yet is full of promise for those that can adapt to a transformed WGD state⁶⁻⁸. Immediately following WGD in autopolyploids (within species WGD, with no hybridisation), a series of novel challenges arise. The most obvious concerns meiosis: the instant doubling of chromosome homologs complicates their neat pairing during meiosis⁹, with consequences that directly reduce fitness. If a chromosome engages in crossovers with more than one other homolog, the likelihood of entanglement dramatically rises, along with breakage upon anaphase. Simultaneously, WGD throws evolved cellular equilibria off balance, including ion homeostasis, protein expression and cell size regulation^{6,7,10,11}. These challenges are so severe that they are insurmountable for many young autopolyploids, although established polyploid populations persist in nature for many species, indicating that the early challenges can be overcome.

To date there have been two genome-scale investigations probing the genomic basis of adaptation to WGD. The first evidence for specific adaptive signatures in response to WGD comes from *Arabidopsis arenosa*^{12,13}. Population resequencing studies scanning for divergent selection across the genomes of this young autotetraploid revealed a set of physically and functionally interacting proteins exhibiting the strongest signatures of selection post-WGD. While a range of processes was under selection, 8 of the 18 most robust signatures of selective sweep directly overlapped genes that appear to have coevolved to decrease chromosome crossover rates during prophase I of meiosis¹³. This same suite of alleles was found to be shared in a sister species, *Arabidopsis lyrata*, with which *A. arenosa* hybridises in the wild, specifically between the autotetraploid cytotypes of each species¹⁴. The importance of these genes was highlighted by the discovery that these same alleles were shared by both young autotetraploids. Specific signatures of gene flow at these same 8 alleles indicated that the two cases were not independent,

and that these 8 alleles that cooperatively function had their origins in separate diploid species, coming together across species barriers only when the two autotetraploids hybridised^{15,16}. More recently, a pool-seq-based genome scan for divergence outliers following WGD in *Cardamine amara*, a Brassicaceae ~17 million years diverged from *A. arenosa*, was unable to detect excess convergence beyond that expected by chance among individual loci under selection at the gene orthologue level¹⁷. However, a modest degree of convergence on the level of functional pathways was detected, in particular for genes that control meiosis. Despite this convergence at the process level, the marked coevolution of functionally and physically interacting chromosome crossover-governing genes discovered in *A. arenosa* was largely absent in *C. amara*¹⁷.

Here we test whether this convergence signal is further abrogated in a more distantly related, independent WGD event. With the addition of a third case we can better estimate whether adaptation of this set of interacting meiosis proteins is the exception or the rule. We focus here on the *Cochlearia* species complex, which is ~20-25 million years diverged from *A. arenosa*¹⁸⁻²¹. The *Cochlearia* genus exhibits two ploidy series with diploid base chromosome number $n=6$ and $n=7$, with the $n=6$ series consisting of diploid, tetraploid, hexaploid, octoploid and eventually heptaploid cytotypes, which broadly hybridise in nature²²⁻²⁸. This cytotype richness is magnified by the presence of B chromosomes in some populations²⁹. *Cochlearia* is found across Europe, from Spain to the Arctic, in a wide range of habitats including freshwater springs, coastal cliffs, sand dunes, salt marshes, metal contaminated sites and roadside grit^{22,24,33-38,26-31,31,32}. A broad habitat differentiation is evident by ploidy, with diploids typically found in upland freshwater springs, tetraploids on coasts, often directly adjacent to seawater or continuously submerged, and hexaploids in similarly extreme saline conditions. In fact, the hexaploid *C. danica* is one of the most rapidly spreading invasive species in the UK and Continental Europe, specifically invading salted roadways since the 1970's^{39,40}, and thriving the most highly saline road grit conditions^{41,42}.

We first assess *Cochlearia* demography and scan for selective sweeps post-WGD by individually resequencing 76 individuals from 23 diploid, tetraploid, and hexaploid populations sampled from across Europe, focusing analysis on diploids and tetraploids in the UK. We find evidence of convergence at the functional level across all three species and partial convergence at the gene ortholog level between *Cochlearia* and *A. arenosa*, but not *C. amara*. This suggests that the same core set of cellular functions must adapt in response to WGD, but that the specific genes that can be utilized to this end are not fully constrained, though there is some evidence of ortholog-level convergence.

Results and Discussion

Population sampling and genome assembly. To determine optimal populations to contrast for WGD-specific signatures of selection, we first sampled populations across the reported range of the *Cochlearia* species complex throughout Europe^{24,26,41,43,44,27,28,30,31,31-34} and then conducted a flow cytometry-based survey of genome size variation (Dataset S1). Measurements were normalised against the diploid population with the most stable individual within-population genome size estimates, WOL. Because *Cochlearia* species extensively hybridise and exhibit considerable phenotypic plasticity, we here primarily designate populations by demonstrated ploidy rather than taxonomic names: in general diploids = *Cochlearia pyrenaica*, tetraploids = *Cochlearia officinalis*, hexaploids = *Cochlearia danica* (coastal dunes and roadside) or octoploids = *Cochlearia anglica* (immediately coastal, marshes).

We constructed a *de novo* genome assembly of one diploid individual from the NEN population using 10x Genomics Chromium linked-read sequencing assembled with Supernova 2.0.0 (91% complete BUSCOs; contig n50=40kbp; Table S1, Table S2 and Methods). We next choose for population-level genome resequencing 116 individuals from 25 populations across Europe and re-sequenced these using Illumina PE format (genome-wide average sequencing depth = 15x; Figure 1A and Table S3). After retaining only individuals with a minimum average genome-wide sequencing depth (average=21x; min=4x), this cohort consisted of 76 individuals from 23 populations. The final dataset consisted of 6,020,948 SNPs (quality and depth filtered; see Methods).

Demographic structure and WGD selection scan population choice. To assess demographic structure, we first performed principal component analysis (PCA) on 415,139 (putatively neutral) 4-fold-degenerate SNPs. The first two axes indicate that geographic origin dominates over ploidy: PC1 (9.5% of variability) differentiates populations by geographic location, and PC2 (7.5% of variability) differentiates by ploidy (figure 1B). The PC1/PC2 distribution resembles a three-pointed star (concave hexagon), where each point represents one ploidy over a gradient of geographic distance, with the exception of individuals from the Hull estuary on the east of Northern England, where individuals of all ploidies intermingle, suggesting extensive local interploidy introgression and a complex reticulate system. Global ancestry estimation with fastSTRUCTURE was consistent with this observation (Fig. 1E), showing obvious clustering by ploidy and geographic location as well as interploidy admixture, especially in the Hull region.

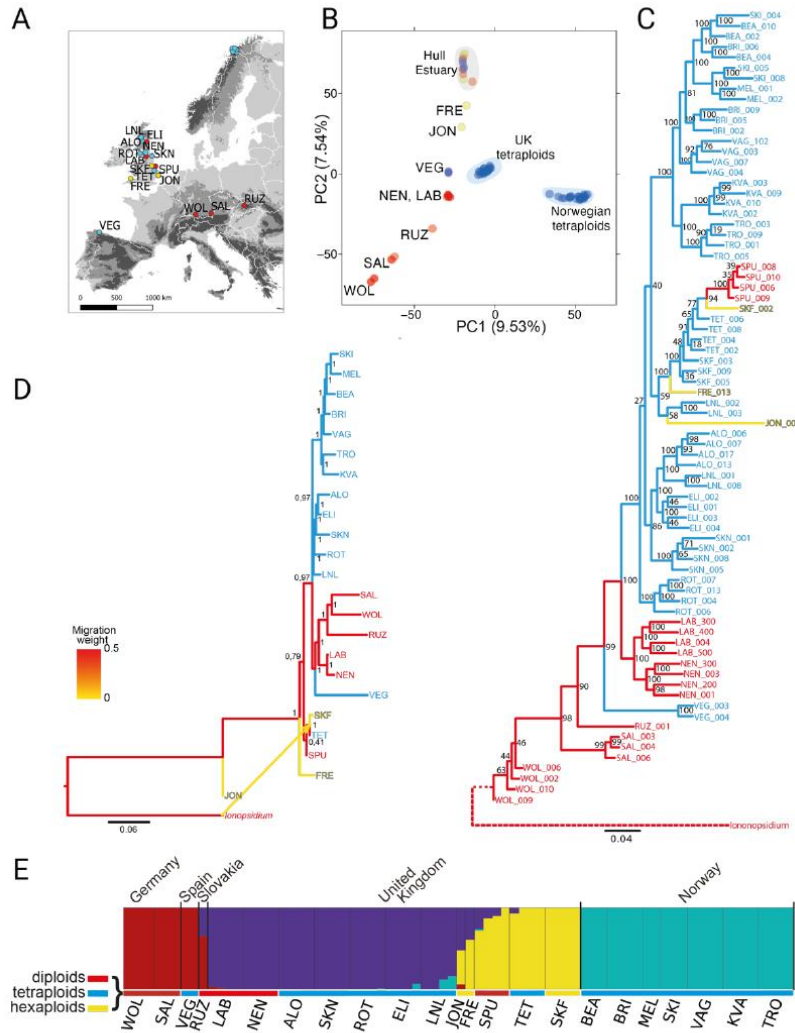


Fig. 1 | Geographic distribution and genetic structure of *Cochlearia*. **A**, distribution of the 23 *Cochlearia* populations (red labels, diploids; blue, tetraploids; yellow, hexaploids; locations are given in Table S3). **B**, Principal Component Analysis (PCA) of all populations. **C**, Rooted phylogeny of 76 individuals from 23 populations created with RAxML, with outgroup *Ionopsidium*. **D**, TreeMix graph of populations, indicating migration edge from *Ionopsidium* to SKF/TET ancestor. **E**, fastSTRUCTURE analysis ($k=4$, min alleles=8) of all *Cochlearia* individuals, with regions, populations, and ploidy indicated.

Next, we performed a phylogenetic analysis in RAxML based on 72,641 4-fold-degenerate SNPs (using the relative *Ionopsidium* to root the tree; Figure 1C). Bootstrap values were strong for major groupings (e.g. the clustering of British diploids vs. tetraploids or the Norwegian tetraploids), although backbone resolution was weaker, with the positioning of groups flipped between trees. Such a pattern could result from high levels of introgression or rapid radiation. Introgression seems likely, as there is known interploidy hybridisation in *Cochlearia*^{23-25,28}. This observation is supported by the admixture seen by fastSTRUCTURE and SplitsTree analyses (Figure 1E and 2C). This could also be consistent with a rapid inter- or peri-glacial radiation¹⁸ and postglacial migration and introgression scenarios such as found in other *Cochlearia* species²⁸. To further assess demographic history and admixture patterns, we performed TreeMix modeling (on 52,186 biallelic, fourfold-degenerate SNPs), which represents genome relationships through a graph of ancestral populations⁴⁵. Here, the optimal number of migration edges was determined to be a single one (based on the Evanno method) which revealed an admixture signal from the outgroup *Ionopsidium* to British *Cochlearia* hexaploids SKF (Figure 1D). Taken together, these demographic analyses indicate complex patterns of ancestry and/or hybridization among the hexaploids, but simpler groupings between diploids and tetraploids.

To focus on adaptation following WGD, we selected six sequenced populations of British *Cochlearia* tetraploids (27 individuals; average depth >20x) and two populations of British *Cochlearia* diploids (8 individuals; average depth >20x; Fig. 2A). This resulted in a dataset containing 6,020,948 SNPs after quality and depth filtering. To assess population relationships in this simplified demographic scenario, we first performed a PCA using 361,981 4-fold-degenerate SNPs. The first axis (11.7% of variability explained) clearly separated the samples by ploidy, while the second (8.1% of variability explained) separated the LNL and TET individuals that showed signs of admixture with either the Norwegian tetraploids or British hexaploids respectively (Fig. 2B-D, Fig. 1E).

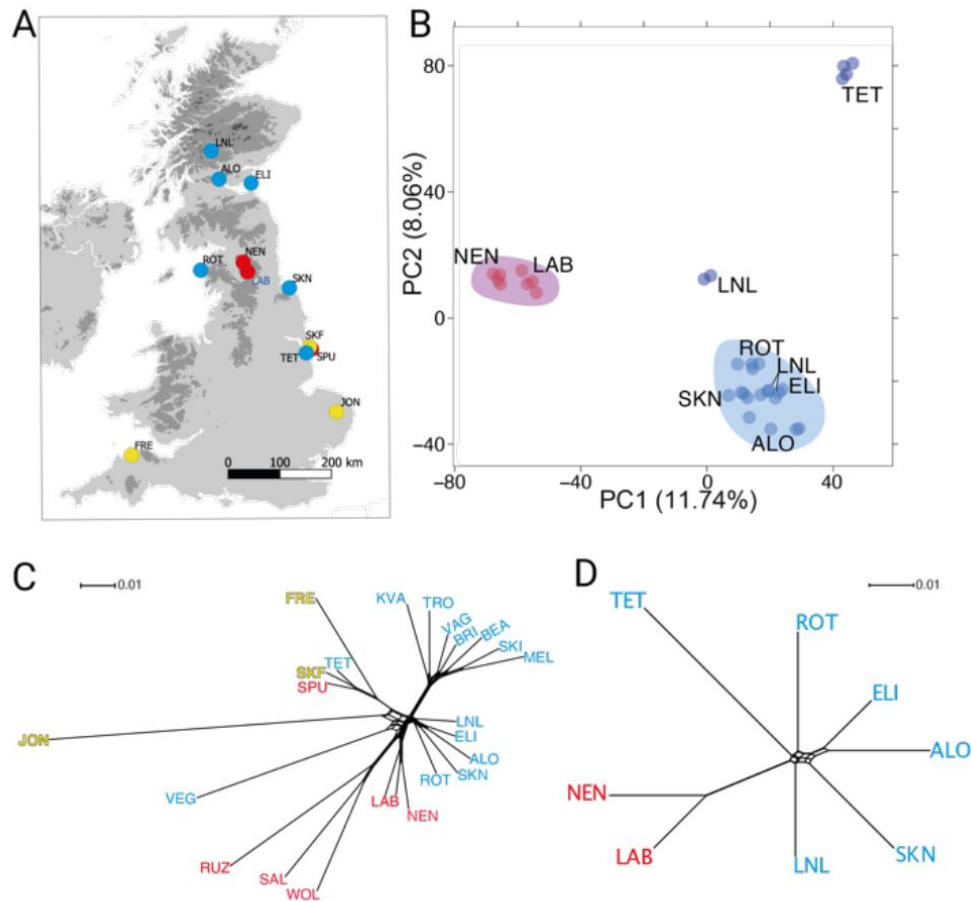


Fig. 2 | British *Cochlearia* demographic structure. **A**, British *Cochlearia* populations used for selective sweep scan (red labels, diploids; blue, tetraploids; yellow, hexaploids). **B**, PCA of British samples only, excluding hexaploids. **C** and **D**, Nei's genetic distances visualised in a SplitsTree representation for all sequenced populations (C), and only British diploids and tetraploids (B).

Selective sweeps associated specifically with WGD. To determine which genes exhibit the strongest signatures of selection following WGD in *Cochlearia* tetraploids, we contrasted allele frequencies across the genomes of our British diploids and tetraploids, calculating differentiation metrics (Rho⁴⁶, Hudson's Fst⁴⁷, Nei's Fst⁴⁸, Weir-Cochran's Fst⁴⁹, Dxy⁵⁰, number of fixed differences and average groupwise allele frequency difference) for all 1 kb windows genome-wide which contain a minimum of 20 post-filter SNPs, a minimum average depth of 8x and a maximum of 20% missing data. The number of SNPs in this contrast was 3,024,896 residing in 44,968 windows, covering

39,594 of the 44,023 predicted genes, or 90% of all gene coding loci.

To determine which differentiation metric most reliably identified genomic regions that exhibit peaks in allele frequency difference (AFD) above local background levels, we performed a quantitative inspection of all AFD plots in the outlier tails of empirical distributions of each differentiation metric (see Methods). Based on superior performance in this assessment, we used Hudson's F_{st} ⁴⁷, which brings the added benefit of robustness for unequal population sizes and presence of rare variants⁵¹. Given that F_{st} -based selective sweep scans have met with success in other diploid/autopolyploid systems^{12,13,15,17,52}, this also makes our current results directly comparable to previous works. We extracted windows in the top 1% of the F_{st} distribution as empirical outliers, consisting of 450 1kb windows, overlapping 296 gene-coding loci (Dataset S2). This list was further refined using a fineMAV-like method^{17,53}, which uses Grantham scores to predict the potential functional impact of each SNP that encodes a non-synonymous amino acid change. This approach then amplifies the severity of each predicted amino acid change by the AFD between the SNPs encoding the change. Out of the 448,625 non-synonymous-encoding SNPs assigned a MAV score, the 1% outliers from the empirical distribution were reserved (4,486 SNPs; Dataset S3) and intersected with our 296 F_{st} outlier windows, yielding a refined list of 144 gene coding loci, containing 406 MAV SNPs (bold in Dataset S2). A selection of AFD peaks for these candidate genes exhibiting gene-localised, ploidy-specific selective sweep signatures is given in Fig. 3.

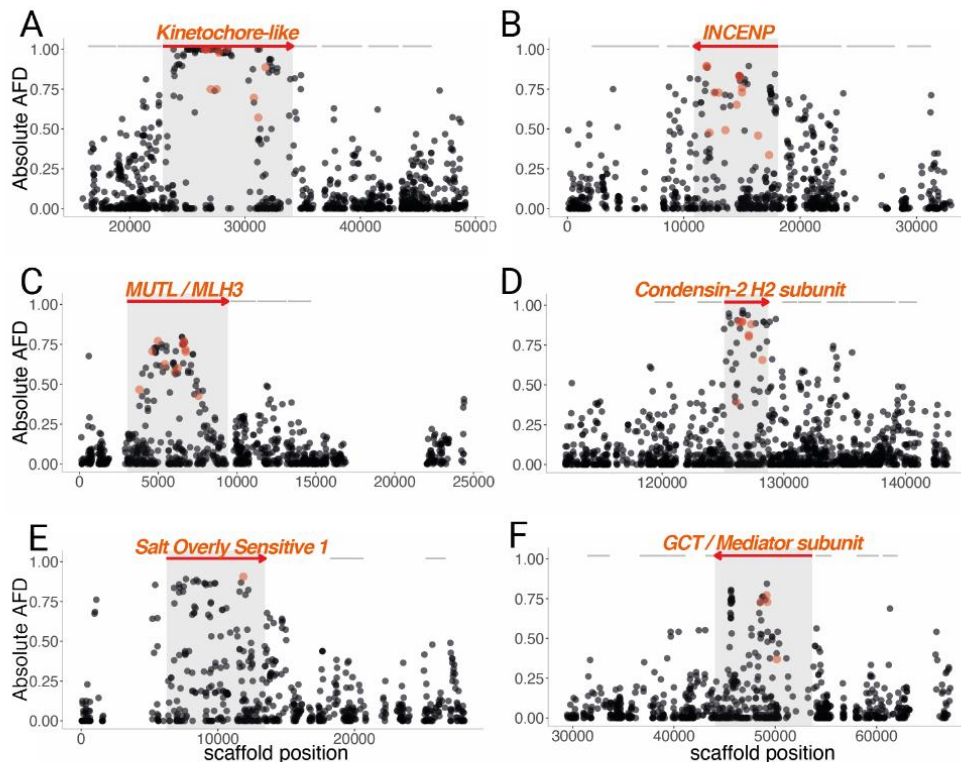


Fig. 3 | Selective sweep signatures at DNA management and ion homeostasis loci. Examples of selective sweep signatures among 6 Fst outlier genes (red arrows). The X axis gives genome position in base pairs. The Y axis gives AFD values at single SNPs (dots) between diploid and tetraploid *Cochlearia*. Red arrows indicate genes overlapping top 1% Fst windows and grey lines indicate neighboring genes. Light red dots indicate MAV outlier SNPs present in the candidate selected locus.

Functional processes under selection post-WGD in *Cochlearia*. Using our 296 WGD-specific selective sweep candidates, gene ontology (GO) enrichment analysis yielded 100 significantly enriched biological processes (using a conservative ‘elim’ Exact Test $p < 0.05$, Table S4; see Methods). Taking higher level categories (minimum 150 annotated genes/category) yielded 30 significantly enriched biological processes (Table S4, bold rows). Among the most significantly enriched categories we observe DNA integration, cell division, meiotic and sister chromatid segregation, mitosis, DNA repair, and recombination represented. Other processes such as response to salt stress, organelle fission, stomatal complex development, and cellular cation homeostasis were also significantly enriched. Overall, these processes are remarkably well aligned with

physiological changes post-WGD suspected to be important challenges to the establishment of nascent autopolyploids^{8,11}.

Evolution of DNA movement and related processes. Eleven of the top sixteen most enriched higher-level GO categories were terms closely related to cell division, organelle fission, cell cycle and DNA metabolic processes, including DNA recombination and DNA repair, and cellular response to DNA repair (Table S4). Candidate genes contributing to these enrichments have predicted roles in cell division, specifically meiosis, or chromosome movement more generally. For example, a particularly dramatic selective sweep signature overlaps directly *CPg20426* (the orthologue of *AT3G10180*) and includes 12 MAV outlier SNPs, 6 of which are fixed in the tetraploid relative to the diploid (Fig. 3A). This gene encodes a kinetochore-like protein, which consists of a kinesin domain and a binding domain with homology to a TATA binding protein, separated by a Structural Maintenance of Chromosomes (SMC) coiled coil domain. Taken together, this suggests that the encoded protein may be part of a complex that mediates chromosome movement. The category ‘formation of the centromere’ is represented by three further genes, *CPg14121*, *CPg15411* and *CPg17406*. *CPg14121* is the ortholog of *CENP-C*, an essential kinetochore component^{54,55} and *CPg15411* is the ortholog of the inner centromere protein *INCENP* (Fig. 3B). Implicated in the control of chromosome segregation and cytokinesis in yeast and animals, plant *INCENP* has a role in the development and differentiation of the gametophytes⁵⁶⁻⁵⁸. Both of these loci contain 1% outlier-MAV SNPs, with the *INCENP* ortholog harboring a remarkable 14 MAV outlier SNPs, the greatest number of MAV outlier SNPs of any gene in the entire genome. Beyond these, there are a diversity of proteins with putative roles in cell division, such as *CPg25053* and *CPg25685* whose *A. thaliana* homologs both regulate cell cycle progression^{59,60}.

Evolution of DNA repair. Several of the most confident signals of selective sweep are found in DNA repair-related genes. In particular *CPg6863*, the ortholog of *A. thaliana* *MLH3* (*MutL*) is both a 1% *Fst* outlier and contains 12 MAV SNPs, joint third most (Fig. 3C; equal with the kinetochore protein *CPg20426* as the gene with the third highest genome-wide quantity of MAV SNPs). This mismatch repair gene controls the number of meiotic crossovers in *A. thaliana*, with mutants exhibiting a 60% reduction in crossover number⁶¹, making this a particularly strong candidate for modulating post-WGD meiosis, where a reduction in crossover number stands as the strongest candidate mechanism for re-establishing fertility in *A. arenosa*^{13,62}.

Two other genes directly implicated in DNA repair stand out: *CPg1416*, the ortholog of the condensin-2 H2-subunit required for proper DNA double-strand break repair in *A.*

thaliana and humans^{63,64}, is an Fst outlier with 8 MAV SNPs (Fig. 3D). Furthermore condensin-2 has been recently implicated in controlling the association and dissociation of centromeres⁶⁵. Additionally, we find an Fst outlier peak over CPg36347, homologue of *DAYSLEEPER*, an essential domesticated transposase⁶⁶. *DAYSLEEPER* was first isolated as binding the Kubox1 motif upstream of the DNA repair gene *Ku70*. The complex Ku70/Ku80 regulate non-homologous end joining double-strand break repair, the only alternative to homologous recombination⁶⁶⁻⁶⁸.

Selection on proteins involved in global transcription. In a polyploid the total DNA content doubles but the protein content and cell size does not scale accordingly, so we predict that the control of gene expression, like meiosis, should undergo adaptive evolution post-WGD. This is supported by empirical findings in *A. arenosa*^{12,13,69}. Here we confirm this finding in *Cochlearia*, with 11 predicted DNA/RNA polymerase-associated genes (*CPg1405*, *CPg1875*, *CPg5061*, *CPg12069*, *CPg16591*, *CPg17556*, *CPg21554*, *CPg26775*, *CPg26954*, *CPg28073* and *CPg31859*) and 3 putative ribosomal genes (*CPg28891*, *CPg34724* and *CPg35322*) among our selective sweep outliers. These include *CPg1405*, the ortholog of *NRPB9A*, an RNA polymerase subunit that is implicated in transcription initiation, processivity, fidelity, proofreading and DNA repair⁷⁰⁻⁷⁴. We also detect an ortholog of *MED13*, *CPg28073*, a component of the mediator complex, which is essential for the production of nearly all cellular transcripts.

Selection on ion transport and stress signaling. The ionic equilibrium of the cell is immediately altered upon WGD¹⁰. We see signatures of selection that may represent a response to this, including stress response genes that are triggered in response to environmental ionic stressors. For example, the ortholog of *SALT OVERLY SENSITIVE 1*, a membrane Na⁺/H⁺ transporter that removes excessive Na⁺ from the cell⁷⁵ and is central to salt tolerance, exhibits a selective sweep signature (Fig 3E). We also find the ortholog of *DEAD-BOX RNA HELICASE 25*, identified in *A. thaliana* as a repressor of stress signaling for salt, osmotic, and cold stress^{76,77}. This gene also controls freezing tolerance⁷⁸, a phenotype relevant to the likely cold-loving demographic history of *Cochlearia*. Similarly, we see *CPg16997*, the ortholog of drought response gene *AtHB7*⁷⁹. To confirm that this was not the result of ecotype differences we performed a salt tolerance experiment on diploid and tetraploid plants. Surprisingly, given their divergent ecotype preferences, with tetraploids found in more saline conditions, we found that the diploid *Cochlearia* are in fact more salt tolerant than the tetraploids (p-value = 2.178e-05; See Supplementary Text 1 and Table S5.). This finding also contrasts strongly to observations of increased salinity tolerance in neotetraploid *Arabidopsis thaliana*¹⁰.

Stomata, plastid-related, and other categories under selection. Several genes

involved in stomatal function were outliers post-WGD, such as the ortholog of *OST2* (*OPEN STOMATA2*; *CPg30015*), which encodes the AHA1 protein, the major H⁺ ATPase in the plasma membrane that drives hyperpolarization and initiates stomatal opening. This protein is a target of ABA stress signaling to close the stomata during drought response⁸⁰. We confirmed this functionally, detecting increases in both stomatal conductance and net photosynthetic rate under drought conditions in tetraploid *Cochlearia* populations relative to diploids (See Supplementary Text 2, Figure S1; Tables S6-S8). Another equilibrium disrupted by WGD, and that has not been discussed in previous WGD adaptation genome scans^{12,13,15,17}, is that between the chloroplast, mitochondrion and nuclear genomes. Many genes related to the function of such organelles are divergence outliers in tetraploid *Cochlearia*. For example, we detect 11 genes annotated as linked to the function of plastids/chlorophyll (*CPg1559*, *CPg1878*, *CPg2251*, *CPg16297*, *CPg18478*, *CPg19736*, *CPg21595*, *CPg26364*, *CPg30068*, *CPg30733* and *CPg33711*) and five linked to mitochondrial function (*CPg2266*, *CPg24404*, *CPg26437*, *CPg28878*, *CPg17406*). Notable also is that four selective sweep candidates encode myosins (*CPg10091*, *CPg22983*, *CPg6763* and *CPg35628*), suggesting a WGD associated adaptation in cellular organization, a category that encompasses organelle localization, cytoskeletal dynamics and nuclear shape¹¹.

Ortholog-level convergence. To test for convergence in adaptation to WGD at the ortholog level, we determined orthogroups using the three genomes gene annotations with OrthoFinder⁸¹, giving a total of 21,619 orthogroups (Methods). Top 1% Fst gene-coding outliers for *A. arenosa* (n=452; Dataset S4), *C. amara* (n=229; Dataset S5), and *Cochlearia* (n=296; Dataset S2) were extracted and then considered orthologues if they were part of the same orthogroup in the genome-wide orthofinder analysis. By this analysis, not a single orthogroup was represented in all three independent WGDs. However, there were a handful of orthogroups common to any two WGD adaptation events: 6 orthogroups were identified in both *C. amara* and *A. arenosa* outliers, while 11 were identified in both *Cochlearia* and *A. arenosa* outlier lists (Table S9). No orthogroups were common in both *Cochlearia* and *C. amara* lists. Consistent with our previous work¹⁷, this overlap was not significant for *C. amara* vs. *A. arenosa* (SuperExactTest P=0.23). In contrast, however, we did detect a significantly greater number of overlaps at the gene ortholog level between Fst outliers in *Cochlearia* and those in *A. arenosa* (SuperExactTest P=0.013). Gene coding loci under selection post-WGD in both *Cochlearia* and *A. arenosa* have inferred roles in DNA and RNA polymerisation (either *DNA pol V* or *nuclear DNA-directed RNA polymerase NRPB9* orthologs, respectively), potassium homeostasis (the *HIGH-AFFINITY K⁺ TRANSPORTER 1* ortholog), and stomatal control (the ortholog of *OPEN STOMATA 2*) (Table S9). All of these genes are involved in processes that have been implicated in adaptation to WGD^{7,8,11}. These loci

therefore stand as strong candidates in salient challenges to nascent polyploids. We note, however, that the degree of this convergence is low, consisting of only 3.7% of genes exhibiting the strongest signatures of selection in *Cochlearia*, and only 2.4% of those under strong selection in *A. arenosa*. While we expect that this likely represents a bona fide lack of convergence in these adaptations at the level of orthologous loci, we note that we focused here on genic signal, and there are many levels upon which selection may act beyond the scope of this study. For example, many gene regulatory changes would escape the notice of our scan, given that we required any outlier divergence window to at least partially overlap a gene coding locus⁶⁶.

Functional process-level convergence. While we detected minimal (or no) convergence in any of our pairwise contrasts at the gene level, we reasoned that there may be similarities in processes under selection between the three independent WGDs, representing process level, or functional convergence. To estimate this, we performed GO enrichment analysis on each outlier set and overlapped the results from each GO WGD (Table 1). Three high-level terms were significantly enriched in all three species: ‘DNA metabolic process’, ‘cellular response to abscisic acid stimulus,’ and ‘cellular response to alcohol’. Additionally, there were six subcategories that were enriched in *Cochlearia* and one of the other two species. These included DNA recombination, drought tolerance, mitosis and meiosis. Taken together, these results strongly imply that adaptive evolution in response to WGD is focused on particular functions, but that there is a high degree of stochasticity in which exact genes evolve.

Table 1. Enriched GO terms common among independent WGD events

		<i>Cochlearia</i>	<i>C. amara</i>	<i>A. arenosa</i>
GO:0006259	DNA metabolic process	0.01764	6.50E-08	0.00082
GO:0071215	cellular response to ABA	0.03738	0.04813	0.04013
GO:0097306	cellular response to alcohol	0.03738	0.04813	0.04013
GO:0051301	cell division	0.00044	-	0.0032
GO:1903047	mitotic cell cycle process	0.00797	-	0.00148
GO:0006310	DNA recombination	0.01543	-	0.02109
GO:0000280	nuclear division	0.03186	-	4.10E-08
GO:0051276	chromosome organization	-	0.01902	0.00021
GO:0009738	ABA-activated signalling	-	0.03186	0.02214
GO:0051321	meiotic cell cycle	0.0395	0.02571	-
GO:0009414	water deprivation response	0.04327	0.04782	-

Note: GO Fisher’s Exact test elim values are given. This conservative method tests for enrichment of terms from the bottom of the GO hierarchy to the top and discards any genes that are significantly enriched in a descendant GO term.

Conclusion

Following WGD, an instantly changed intracellular environment drives the evolution of a suite of distinct processes. These processes include meiotic chromosome segregation, ion homeostasis, and nucleotypic factors revolving around cell size, volume, and cell cycle control^{7,10,11}. The broad array of relevant genes exhibiting signatures of selection in our data suggest that adaptations of each of these processes have a polygenic basis. Further, we observed the footprint of WGD-associated selection in genome-wide shifts in the frequencies of many alleles, replicated in three independent WGDs. This work, along with others^{13,15-17}, shows that these processes exhibit genomic signatures of adaptive evolution consistent with observations of altered phenotype upon WGD consistently in independent adaptation events. A primary, well-established challenge occurs during meiotic chromosome segregation, and we here illustrated in *Cochlearia* other genomic and physiological changes. These physiological changes include increases in drought resilience and stomatal conductance, concordant with findings in other studies that concentrated on phenotype rather than genome-wide signal^{10,82,83}.

Given this complex genomic architecture, we found that convergence in three species was minimal at the ortholog level. Even pairwise comparisons between species gave only a handful of common orthologs under selection (Table S9). However, at the level of functional processes, we observe evidence of convergent evolution. DNA management, as a high level process stood out, and in *Cochlearia* we saw a substantial shift relative to *A. arenosa*: in *Cochlearia* we observe enrichment in DNA repair and later meiotic processes, as opposed to the prophase I-oriented signal previously reported in *A. arenosa*³. It is not yet clear why particular solutions are favored in one species relative to another. A degree of stochasticity depending on available standing variation can be expected. But we also expect that an important role may also be played by difference in species histories, which may offer preadaptations that steer evolution in a particular direction. For example, our analysis of salinity tolerance in *Cochlearia* provided the surprising result the diploid cytotype was at least as tolerant to extreme levels of salt (600 mM, seawater concentrations) as the tetraploid, even though the diploid is found exclusively inland, with the tetraploids in seawater or directly on coasts. This cryptic salinity adaptation may be related to genetically connected polyploid coastal populations along the Atlantic coast from Portugal towards coastal systems in northern Scandinavia and the UK with glacial coastal refugia in the southern regions of Europe^{44,84}. Furthermore diploid coastal populations from salt-affected habitat occur in Spain and a postglacial and boreal spread of the diploid towards the UK is possible^{39, new}, with salinity tolerance developing along the way, altering the genomic substrate upon which selection acted in response to WGD-associated ionic challenge^{12,13}.

While we have not directly addressed the cause of the commonly observed adaptability of polyploids in this work, our results may suggest a hypothesis for one potential contributing factor to the occasional dramatic niche shifts observed following WGD. We observed a large quantity of DNA metabolism and repair genes under selection in all species, and especially *Cochlearia*. This may signal a temporarily increased susceptibility to DNA damage, due to suboptimal function of DNA repair genes during the process of adaptive evolution, resulting in a relative ‘mutator phenotype’ in young polyploids. Such a mutator phenotype has been plainly observed in aggressive polyploid human cancers, which not only exhibit SNP-level hypermutator phenotypes, but also dramatic structural variation in malignant aneuploid swarms that are associated with cancer progression⁷. It could be that a parallel to this exists following other WGD events, even in plants. Whether or not this hypothesis is supported by future discoveries, the cross-kingdom importance of WGD to fields from evolution, to ecology, to agriculture and medicine, underscores the importance of understanding the processes mediating adaptation to—and perhaps by—WGD.

Methods

Plant material. We first located 89 populations throughout Europe and collected population samplings of plants from each, aiming for at least 10 plants per population, with each sampled plant a minimum of 2 meters from any other. Of these, we selected 12 representative populations from British Isles and 12 populations from the rest of the European range for population resequencing (Table S3). An average of 4 individuals per population were sequenced with the exception of two hexaploid lineages that fall outside the central focus of this study, but were included for the demographic analysis, coastal *Cochlearia danica* (JON), which is invasive at inland stands, and coastal *Cochlearia anglica*-like FRE, for which only one individual was sequenced. A total of 116 individuals were initially sequenced, which was narrowed down by a cutoff of 4x, leaving 76 individuals from 23 populations in the final analysed dataset.

Ploidy Determination. DNA content and ploidy were inferred for populations using Flow Cytometry (Dataset S1). Approximately 1 square cm of leaf material was diced alongside an internal reference using razor blades in 1 ml ice cold extraction buffer (either 45 mM MgCl₂, 30 mM sodium citrate, 20mM MOPS, 1% Triton-100, pH 7 with NaOH for relative staining or 0.1 M citric acid, 0.5% Tween 20 for absolute measurements). The resultant slurry was then filtered through a 40-µm nylon mesh before the nuclei were stained with the addition of 1 ml staining buffer (either CyStain UV precise P [Sysmex, Fluorescence emission: 435nm to 500nm] for relative ploidy, or Otto 2 buffer [0.4 M Na₂HPO₄·12H₂O, Propidium iodide 50 µg mL⁻¹, RNase 50 µg mL⁻¹], for absolute DNA content). After 1 minute of incubation at room temperature the sample was run for 5,000 particles on either a Partec PA II flow cytometer or a BD FACS Melody. Histograms were evaluated using FlowJo software version 10.6.1.

Reference Genome Assembly and Alignment

We generated a synthetic long read-based *de novo* genome assembly using 10x Genomics Chromium linked read technology.

CTAB DNA extraction method. A total of 0.4 g *Cochlearia pyrenaica* leaf material from one individual plant in the NEN population was ground using liquid nitrogen before the addition of 10 ml of CTAB DNA extraction buffer (100 mM Tris-HCl, 2% CTAB, 1.4 M NaCl, 20 mM EDTA, and 0.004 mg/ml Proteinase K). The mixture was incubated at 55°C for 1 hour then cooled on ice before the addition of 5 ml Chloroform. This was then centrifuged at 3000 rpm for 30 minutes and the upper phase taken, this was added to 1X volume of phenol:chloroform:isoamyl-alcohol and spun for 30 minutes at 3000 rpm. Again, the upper phase was taken and mixed with a 10% volume of 3M NaOAc and 2.5X volume of 100% ethanol at 4 °C. This was incubated on ice for 30 minutes before being centrifuged for 30 minutes at 3000 rpm and 4 °C. Three times the pellet was washed in 4ml 70% ethanol at 4 °C before being centrifuged again for 10 minutes at 3000 rpm and 4°C. The pellet was then air dried and resuspend in 300 ul nuclease-free water containing 0.0036 mg/ml RNase A. The DNA concentration was checked on a QuBit Fluorometer 2.0 (Invitrogen) using the QuBit dsDNA HS Assay kit. Fragment sizes were assessed using a Q-card (OpGen Argus) and the Genomic DNA TapeStation assay (Agilent).

10X library construction. DNA material was diluted to 0.5 ng/ μ l with EB (Qiagen) and checked with a QuBit Fluorometer 2.0 (Invitrogen) using the QuBit dsDNA HS Assay kit. The Chromium User Guide was followed as per the manufacturer's instructions (10X Genomics, CG00043, Rev A). The final library was quantified using qPCR (KAPA Library Quant kit (Illumina), ABI Prism qPCR Mix, Kapa Biosystems). Sizing of the library fragments were checked using a Bioanalyzer (High Sensitivity DNA Reagents, Agilent). Samples were pooled based on the molarities calculated using the two QC measurements. The library was clustered at 8 pM with a 1% spike of PhiX library (Illumina).

Sequencing and assembly and assembly QC. The sample was sequenced on HiSeq2500 Rapid Run V2 mode (Illumina). The first run on 150 bp sequences gave 101.29 M reads. A second run was carried out on 250 bp sequences, bringing the total number of reads up to 269.58 M, total coverage to 123x and effective coverage to 51x. These were subsampled to 135 M reads and assembled on Supernova 2.0.0 giving raw coverage of 63x and effective coverage 35x. The molecule length was 16.6 kb. Two assemblies were kept, with a minimum contig size of 10 Kb or 3 Kb, with an assembly size of 174.47 Mb and 219.69 Mb respectively. The k-mer estimate for the genome size was 528.26 Mb and the flow cytometry estimate of the genome size was 656 Mb (0.67 pg \pm 0.07) for the NEN diploid and 1,352 Mb (1.38 pg \pm 0.12) for the tetraploids, consistent with previous reports^{85,86}. The final 3kb minimum contig length assembly had 13,302 contigs, an N50 of 39.7 kb. Assessment of gene space completeness gave 91.3% complete, single copy core eukaryotic 'BUSCO genes' (1315/1440 BUSCO groups; Table S1; BUSCO version 3.0.2).⁸⁷ Uncollapsed haplotypes were detected in this assembly: to purge these we identified uncollapsed haplotypes (defined as ID > 99%, coverage > 99%; consisting ~28 mb of the genome) and one scaffold was randomly selected to use in alignments (consisting ~12 mb of the genome), while the rest were excluded.

Gene Calling and Annotation. The genome was annotated with gene model predictions produced by AUGUSTUS (version 3)⁸⁸ which had been trained on the *Arabidopsis thaliana* genome. A total of 44,023 putative genes were identified.

Population Resequencing and Analysis.

Library preparation and sequencing. DNA was prepared using the commercially available DNeasy Plant Mini Kit from Qiagen. DNA libraries were made using TruSeq DNA PCR-free Library kit from Illumina as per the manufacturer's instructions and were multiplexed based on concentrations measured with a QuBit Fluorometer 2.0 (Invitrogen) using the QuBit dsDNA HS Assay kit. Sequencing was carried out on either NextSeq 550 (Illumina) in house (4 runs) or sent to Novogene for Illumina HiSeq X, PE150 sequencing (2 runs).

Data preparation, alignment, and genotyping. Adapter sequences were removed using cutadapt (version 1.9.1)⁸⁹ and quality trimmed with Sickle (version 1.2)⁹⁰ to generate only high-quality reads (Phred score \geq 30) of 30bp or more. The reads were aligned to the reference using

bwa (v. 0.7.12)⁹¹ and further processed with samtools (v. 1.3)⁹². Duplicate reads were removed and read group IDs added to the bam files using Picard (version 1.134)⁹³. Indels were realigned with GATK (version 3.8)⁹⁴. Samples were first genotyped individually with “HaplotypeCaller” and were then genotyped jointly using “GenotypeGVCFs” in GATK (version 3.8)⁹⁴. The resulting VCF files were then filtered for biallelic sites and mapping quality (QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, HaplotypeScore < 13.0). The VCF was then filtered by depth. To prevent a single individual from dominating the mean depths the three individuals with the most coverage (VEG_003, SKF_009 and VEG_004) were removed and a depth histogram was created for the remaining 73 individuals. Based on this distribution a depth cutoff of 2,469 was applied to the VCF containing the 73 individuals and this was then used as a mask for the final VCF containing all individuals.

Demographic analysis. We inferred relationships between populations as genetic distances using principal component analysis (PCA) implemented in *adegenet*⁹⁵. To further interrogate their relationships we then ran fastSTRUCTURE⁹⁶. Since fastSTRUCTURE does not handle polyploid genomes we randomly subsampled two alleles from tetraploid and hexaploid populations using a custom script and used this dataset in fastStructure. We have previously demonstrated that results generated in this way are directly comparable to results generated with the full dataset in STRUCTURE⁹⁷. We calculated Nei’s distances among all individuals in stamp and visualised these using SplitsTree⁹⁸.

Phylogenetic analysis. We constructed a maximum likelihood phylogeny using RAxML version 8.1.16⁹⁹ under a GTR + G model of evolution and with an ascertainment bias correction (--asc-corr=lewis) in order to account for unsampled invariant sites in SNP datasets. Sites with more than 10% of missing data were excluded from a set of 79,252 fourfold-degenerate SNPs using the GATK tool “SelectVariants” (GATK version 3.8), and the python script *ascbias.py* (https://github.com/btmartin721/raxml_ascbias) was used to remove sites considered as invariable by RAxML. The maximum likelihood analysis was performed with 1000 rapid bootstrap replicates. Additionally, we used TreeMix version 1.13⁴⁵ to generate a population maximum likelihood phylogeny allowing for migration events (admixture) between populations. The input file was generated using the script *vcf2treemix.py* (https://github.com/CoBiG2/RAD_Tools/blob/master/vcf2treemix.py), thereby excluding multiallelic sites from the set of fourfold-degenerate variants with a maximum of 10% missing data. We tested for up to 10 migration edges (*M*) and performed 10 initial replicate runs for every *M*. We then determined the optimal number of migration edges based on the Evanno method using the R package *optM*¹⁰⁰. Hereafter, we performed 100 bootstrap replicates for the best-fitting *M* and finally, a consensus tree was inferred from the resulting 100 maximum likelihood trees using *sumtrees.py* version 4.10¹⁰¹.

Orthogrouping and Reciprocal Best Blast Hits. We performed an orthogroup analysis using Orthofinder version 2.3.3⁸¹. to infer orthologous groups (OGs) from four species (*C. amara*, *A. lyrata*, *A. thaliana*, *C. pyrenaica*). A total of 21,618 OGs were found. Best reciprocal blast hits

(RBHs) for *Cochlearia* and *A. thaliana* genes were found using BLAST version 2.9.0. *Cochlearia* genes were then assigned an *A. thaliana* gene ID for GO enrichment analysis in one of five ways. First if the genes' OG contained only one *A. thaliana* gene ID, that gene ID was used. If the OG contained more than one *A. thaliana* gene ID then the RBH was taken. If there was no RBH then the OG gene with the lowest E-value in a BLAST versus the TAIR10 database was taken. If no OG contained the *Cochlearia* gene then the RBH was taken. Finally, if there was no OG or RBH then the gene with the lowest E-value in a BLAST versus the TAIR10 database was taken. BLASTs versus the TAIR10 database were performed during December 2019.

GO Enrichment Analysis. To infer functions significantly associated with directional selection following WGD, we performed gene ontology enrichment of candidate genes in the R package TopGO v.2.32¹⁰², using *A. thaliana* orthologs of *Cochlearia* genes and an *A. thaliana* universe set. We tested for overrepresented Gene Ontology (GO) terms within the three domains Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) using Fisher's exact test with conservative 'elim' method, which tests for enrichment of terms from the bottom of the GO hierarchy to the top and discards any genes that are significantly enriched in a descendant GO term¹⁰³. We used 'biological process' ontology with minimum node size 150 genes and FDR = 0.05. A significance cut-off of 0.05 and all processes represented by a single gene were removed (25 in total).

Window-based scan for selective sweep signatures. We performed a window-based divergence scan for selection consisting of 1 kb windows that contained at least 20 SNPs. The data was filtered as described above and in addition was filtered for no more than 20% missing data and a depth of $\geq 8x$. We calculated metrics: Rho, Nei's Fst, Weir-Cochran's Fst, FstH, Dxy, number of fixed differences and average groupwise allele frequency difference (AFD). To determine the best metric to use we performed a quantitative analysis of AFD plot quality for all 1% outliers of each metric. Each window was given a score of 0-4, with 0 being the lowest quality and 4 the highest. Scores were based on two qualities: peak height and peak specificity. For peak height one point was awarded if the window contained one SNP of $AFD > 0.5 < 0.7$, and two points were awarded for any SNP of $AFD > 0.7$. Likewise, for peak specificity two points were awarded for an AFD peak that was restricted to a single gene and one point was awarded for a peak that was restricted to 2-3 genes. The top 1% outliers from the metric FstH⁴⁷ was selected as, compared to all other single 1% outlier lists and all permutations of overlapped 1% outlier lists, it maximized the number of '4' and '3' scores while minimizing the number of '1' and '0' scores. This is consistent with the good performance of Fst in our previous studies^{12,13,17,52}.

MAV analysis. A FineMAV⁵³-like analysis was carried out on all biallelic, non-synonymous SNPs passing the same filters as the window-based selection scan. SNPs were assigned a Grantham score according to the amino acid change and this was scaled by the AFD between ploidies. The top 1% outliers of all these MAV-SNPs were then overlapped with the genes in our 1% Fst outlier windows to give a refined list of candidate genes that contain potentially functionally significant non-synonymous mutations at high AFD between cytotypes. The code outlining this

can be found at https://github.com/paajanen/meiosis_protein_evolution/tree/master/FAAD.

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 PRJNAXXXXXX (available at <http://www.ncbi.nlm.nih.gov/bioproject/XXXXX>) and will be released following peer review and publication.

Acknowledgements

The authors thank Kirsten Bomblies, Lara Hebberecht-Lopez, Filip Kolar, Mary Bray and Nigel Bray and for their assistance collecting plant material. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [grant number ERC-StG 679056 HOTSPOT], via a grant to LY.

Author Contributions

LY and SMB conceived the study. SMB, EW, MB, SB, SF, PP, MK and LY performed analyses. SMB, SB, MZ, and SF performed laboratory experiments. LY, SMB, PM, and JK performed field collections. LY and SMB wrote the manuscript with primary input from all authors. All authors edited and approved 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. Darwin, C. *On the Origin of the Species*. John Murray, 1859.
2. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: New insights and directions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374** (2019).
3. Sackton, T. B. *et al.* Convergent regulatory evolution and the origin of flightlessness in palaeognathous birds. *Science*. **364**, 74–78 (2019).
4. Zhen, Y., Aardema, M. L., Medina, E. M., Schumer, M. & Andolfatto, P. Parallel molecular evolution in an herbivore community. *Science*. **337**, 1634–1637 (2012).
5. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **45**, 1183–1189 (2013).
6. Van De Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18** 411–424 (2017).
7. Yant, L. & Bomblies, K. Genome management and mismanagement—cell-level opportunities and challenges of whole-genome duplication. *Genes Dev.* **29**, 2405–2419 (2015).
8. Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F. & Yant, L. The ‘Polyploid Hop’: Shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Front. Ecol. Evol.* **6**, (2018).
9. Lloyd, A. & Bomblies, K. Meiosis in autopolyploid and allopolyploid *Arabidopsis*. *Current Opinion in Plant Biology* **30** 116–122 (2016).
10. Chao, D. Y. *et al.* Polyploids exhibit higher potassium uptake and salinity tolerance in *Arabidopsis*. *Science*. **341**, 658–659 (2013).
11. Doyle, J. J. & Coate, J. E. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* **180** 1–52 (2019).
12. Hollister, J. D. *et al.* Genetic Adaptation Associated with Genome-Doubling in Autotetraploid *Arabidopsis arenosa*. *PLoS Genet.* **8**, (2012).
13. Yant, L. *et al.* Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr. Biol.* **23**, 2151–2156 (2013).
14. Lafon-Placette, C. *et al.* Endosperm-based hybridization barriers explain the pattern of gene flow between *Arabidopsis lyrata* and *Arabidopsis arenosa* in Central Europe. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1027–E1035 (2017).
15. Marburger, S. *et al.* Interspecific introgression mediates adaptation to whole genome duplication. *Nat. Commun.* (2019) doi:10.1038/s41467-019-13159-5.
16. Seear, P. J. *et al.* A novel allele of *ASY3* promotes meiotic stability in autotetraploid *Arabidopsis lyrata*. *bioRxiv* 2019.12.25.888388 (2019) doi:10.1101/2019.12.25.888388.
17. Bohutínská, M. *et al.* Genomic novelty versus convergence in the basis of adaptation to whole genome duplication. *bioRxiv* (2020) doi:10.1101/2020.01.31.929109.
18. Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**,

- 2770–2784 (2015).
19. Guo, X. *et al.* Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* **18**, (2017).
 20. Huang, X. C., German, D. A. & Koch, M. A. Temporal patterns of diversification in Brassicaceae demonstrate decoupling of rate shifts and mesopolyploidization events. *Ann. Bot.* **125**, 29–47 (2020).
 21. Huang, C. H. *et al.* Resolution of brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* **33**, 394–412 (2016).
 22. Brummitt, R. K. & Stace, C. A. *Hybridization and the Flora of the British Isles*. *Kew Bulletin* vol. 32 (Academic Press, 1977).
 23. Saunte, L. H. Cytogenetical studies in the *Cochlearia officinalis* complex. *Hereditas* (1955) doi:10.1111/j.1601-5223.1955.tb03006.x.
 24. Fearn, G. M. A morphological and cytological investigation of *Cochlearia* on the Gower Peninsula, Glamorga. *New Phytol.* (1977) doi:10.1111/j.1469-8137.1977.tb02226.x.
 25. Koch, M., Hurka, H. & Mummenhoff, K. Chloroplast DNA restriction site variation and RAPD-analyses in *Cochlearia* (Brassicaceae): Biosystematics and speciation. *Nord. J. Bot.* (1996) doi:10.1111/j.1756-1051.1996.tb00276.x.
 26. Gill, B., McAllister, M. & Fearn, G. Cytotaxonomic studies on the *Cochlearia officinalis* L. group from inland stations in Britain. *Watsonia* **21**, 15–21 (1978).
 27. Gill, J. J. B. Cytogenetic Studies in *Cochlearia* L. *Ann. Bot.* (1971).
 28. Gill, E. Conservation genetics of the species complex *Cochlearia officinalis* L. sl in Britain. (University of Edinburgh, 2008).
 29. Gupta, P. P. Suppression of multivalent formation by B chromosomes in natural and artificial autopolyploids of scurvy-grass (*Cochlearia* L.). *Theor. Appl. Genet.* **59**, 221–223 (1981).
 30. Gill, J. J. B. Cytogenetic studies in *Cochlearia* L. (Cruciferae). The chromosomal constitution of *C. danica* L. *Genetica* (1990) doi:10.1007/BF00122520.
 31. Gill, J. J. B. Cytogenetic studies in *Cochlearia* L. (Cruciferae). The origins of *C. officinalis* L. and *C. micacea* Marshall. *Genetica* (1973) doi:10.1007/BF00119107.
 32. Gill, J. J. B. Cytogenetic Studies in *Cochlearia* L. The chromosomal homogeneity within both the $2n = 12$ diploids and the $2n = 14$ diploids and the cytogenetic relationship between the two chromosome levels. *Ann. Bot.* (1971) doi:10.1093/oxfordjournals.aob.a084558.
 33. Koch, M., Dobeš, C., Bernhardt, K. G. & Kochjarová, J. *Cochlearia macrorrhiza* (Brassicaceae): A bridging species between *Cochlearia* taxa from the Eastern Alps and the Carpathians? *Plant Syst. Evol.* (2003) doi:10.1007/s00606-003-0048-4.
 34. Koch, M. Genetic differentiation and speciation in prealpine *Cochlearia*: Allohexaploid *Cochlearia bavarica* Vogt (Brassicaceae) compared to its diploid ancestor *Cochlearia pyrenaica* DC. in Germany and Austria. *Plant Syst. Evol.* (2002) doi:10.1007/s006060200025.
 35. Koch, M., Mummenhoff, K. & Hurka, H. Molecular phylogenetics of *Cochlearia* (Brassicaceae) and allied genera based on nuclear ribosomal ITS DNA sequence

- analysis contradict traditional concepts of their evolutionary relationship. *Plant Syst. Evol.* **216**, 207–230 (1999).
36. Koch, M. A. Mid-miocene divergence of *Ionopsidium* and *Cochlearia* and its impact on the systematics and biogeography of the tribe Cochlearieae (Brassicaceae). *Taxon* (2012) doi:10.1002/tax.611006.
 37. Pegtel, D. M. Effect of ploidy level on fruit morphology, seed germination and juvenile growth in scurvy grass (*Cochlearia officinalis* L. s.l., Brassicaceae). *Plant Species Biol.* **14**, 201–215 (1999).
 38. Nawaz, I., Iqbal, M., Blied, M. & Schat, H. Salt and heavy metal tolerance and expression levels of candidate tolerance genes among four extremophile *Cochlearia* species with contrasting habitat preferences. *Sci. Total Environ.* **584–585**, 731–741 (2017).
 39. Koch, M. A. Kurznotiz zur südlichen Ausbreitung des Dänischen Löffelkrauts (*Cochlearia danica* L.) in Nordrhein-Westfalen. *Flor. Rundbr* **30**, 136–138 (1997).
 40. Koch, M. A. Zur Ausbreitung des Dänisches Löffelkrautes (*Cochlearia danica* L.) als Küstensippe in das Niedersächsische Binnenland. *Flor. Rundbr* **30**, 20–23 (1996).
 41. Scott, N. E. & Davison, A. W. De-icing salt and the invasion of road verges by maritime plants. *Watsonia* **14**, 41–52 (1982).
 42. Fekete, R., Mesterházy, A., Valkó, O. & Molnár, A. V. A hitchhiker from the beach: The spread of the maritime halophyte *Cochlearia danica* along salted continental roads. *Preslia* **90**, 23–37 (2018).
 43. Brandrud, M. K., Paun, O., Lorenzo, M. T., Nordal, I. & Brysting, A. K. RADseq provides evidence for parallel ecotypic divergence in the autotetraploid *Cochlearia officinalis* in Northern Norway. *Sci. Rep.* (2017) doi:10.1038/s41598-017-05794-z.
 44. Koch, M., Huthmann, M. & Hurka, H. Isozymes, speciation and evolution in the polyploid complex *Cochlearia* L. (Brassicaceae). *Bot. Acta* (1998) doi:10.1111/j.1438-8677.1998.tb00727.x.
 45. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* **8**, (2012).
 46. Ronfort, J., Jenczewski, E., Bataillon, T. & Rousset, F. Analysis of population structure in autotetraploid species. *Genetics* **150**, 921–930 (1998).
 47. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
 48. Nei, M. Genetic Distance between Populations. *Am. Nat.* (1972) doi:10.1086/282771.
 49. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N. Y.)* **38**, 1358 (1984).
 50. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
 51. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521

- (2013).
52. Arnold, B. J. *et al.* Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8320–8325 (2016).
 53. Szpak, M. *et al.* FineMAV: Prioritizing candidate genetic variants driving local adaptations in human populations. *Genome Biol.* **19**, (2018).
 54. Ogura, Y., Shibata, F., Sato, H. & Murata, M. Characterization of a CENP-C homolog in *Arabidopsis thaliana*. *Genes Genet. Syst.* **79**, 139–144 (2004).
 55. Sandmann, M. *et al.* Targeting of arabisidopsis KNL2 to centromeres depends on the conserved CENPC-K motif in its C terminus. *Plant Cell* **29**, 144–155 (2017).
 56. Kirioukhova, O. *et al.* Female gametophytic cell specification and seed development require the function of the putative *Arabidopsis* INCENP ortholog WYRD. *Development* **138**, 3409–3420 (2011).
 57. Ruchaud, S., Carmena, M. & Earnshaw, W. C. Chromosomal passengers: Conducting cell division. *Nature Reviews Molecular Cell Biology* **8** 798–812 (2007).
 58. Vagnarelli, P. & Earnshaw, W. C. Chromosomal passengers: The four-dimensional regulation of mitotic events. *Chromosoma* **113** 211–222 (2004).
 59. Wang, W., Sijacic, P., Xu, P., Lian, H. & Liu, Z. *Arabidopsis* TSO1 and MYB3R1 form a regulatory module to coordinate cell proliferation with differentiation in shoot and root. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3045–E3054 (2018).
 60. Cheng, Y. *et al.* Down-regulation of multiple CDK inhibitor ICK/KRP genes promotes cell proliferation, callus induction and plant regeneration in *Arabidopsis*. *Front. Plant Sci.* **6**, (2015).
 61. Jackson, N. *et al.* Reduced meiotic crossovers and delayed prophase I progression in AtMLH3-deficient *Arabidopsis*. *EMBO J.* **25**, 1315–1323 (2006).
 62. Bomblies, K., Jones, G., Franklin, C., Zickler, D. & Kleckner, N. The challenge of evolving stable polyploidy: could an increase in “crossover interference distance” play a central role? *Chromosoma* **125** 287–300 (2016).
 63. Sakamoto, T. *et al.* Condensin ii alleviates DNA damage and is essential for tolerance of boron overload stress in arabisidopsis. *Plant Cell* **23**, 3533–3546 (2011).
 64. Wood, J. L., Liang, Y., Li, K. & Chen, J. Microcephalin/MCPH1 associates with the condensin II complex to function in homologous recombination repair. *J. Biol. Chem.* **283**, 29586–29592 (2008).
 65. Sakamoto, T., Sugiyama, T., Yamashita, T. & Matsunaga, S. Plant condensin II is required for the correct spatial relationship between centromeres and rDNA arrays. *Nucleus* **10**, 116–125 (2019).
 66. Bundock, P. & Hooykaas, P. An *Arabidopsis* hAT-like transposase is essential for plant development. *Nature* **436**, 282–284 (2005).
 67. Chang, W. C. *et al.* Regulation of Ku gene promoters in *Arabidopsis* by hormones and stress. *Funct. Plant Biol.* **35**, 265–280 (2008).
 68. Tamura, K., Adachi, Y., Chiba, K., Oguchi, K. & Takahashi, H. Identification of Ku70 and Ku80 homologues in *Arabidopsis thaliana*: Evidence for a role in the repair of DNA double-strand breaks. *Plant J.* **29**, 771–781 (2002).

69. Bomblies, K., Higgins, J. D. & Yant, L. Meiosis evolves: Adaptation to external and internal environments. *New Phytol.* **208**, 306–323 (2015).
70. Li, S., Ding, B., Chen, R., Ruggiero, C. & Chen, X. Evidence that the Transcription Elongation Function of Rpb9 Is Involved in Transcription-Coupled DNA Repair in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **26**, 9430–9441 (2006).
71. Hull, M. W., McKune, K. & Woychik, N. A. RNA polymerase II subunit RPB9 is required for accurate start site selection. *Genes Dev.* **9**, 481–490 (1995).
72. Tan, E. H., Blevins, T., Ream, T. S. & Pikaard, C. S. Functional Consequences of Subunit Diversity in RNA Polymerases II and V. *Cell Rep.* **1**, 208–214 (2012).
73. Hemming, S. A. *et al.* RNA polymerase II subunit Rpb9 regulates transcription elongation in vivo. *J. Biol. Chem.* **275**, 35506–35511 (2000).
74. Nesser, N. K., Peterson, D. O. & Hawley, D. K. RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 3268–3273 (2006).
75. Shi, H., Ishitani, M., Kim, C. & Zhu, J. K. The *Arabidopsis thaliana* salt tolerance gene SOS1 encodes a putative Na⁺/H⁺ antiporter. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6896–6901 (2000).
76. Kant, P., Kant, S., Gordon, M., Shaked, R. & Barak, S. Stress Response Suppressor1 and Stress Response Suppressor2, two Dead-box RNA helicases that attenuate *Arabidopsis* responses to multiple abiotic stresses. *Plant Physiol.* **145**, 814–830 (2007).
77. Khan, A. *et al.* The *Arabidopsis* STRESS RESPONSE SUPPRESSOR DEAD-box RNA helicases are nucleolar- and chromocenter-localized proteins that undergo stress-mediated relocalization and are involved in epigenetic gene silencing. *Plant J.* **79**, 28–43 (2014).
78. Kim, J. S., Kim, K. A., Oh, T. R., Park, C. M. & Kang, H. Functional characterization of DEAD-box RNA helicases in *Arabidopsis thaliana* under abiotic stress conditions. *Plant Cell Physiol.* **49**, 1563–1571 (2008).
79. Ré, D. A., Capella, M., Bonaventure, G. & Chan, R. L. *Arabidopsis* AtHB7 and AtHB12 evolved divergently to fine tune processes associated with growth and responses to water stress. *BMC Plant Biol.* **14**, (2014).
80. Merlot, S. *et al.* Constitutive activation of a plasma membrane H⁺-ATPase prevents abscisic acid-mediated stomatal closure. *EMBO J.* **26**, 3216–3226 (2007).
81. D.M., E. & S., K. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv* 466201 (2018) doi:10.1101/466201.
82. Allario, T. *et al.* Tetraploid Rangpur lime rootstock increases drought tolerance via enhanced constitutive root abscisic acid production. *Plant, Cell Environ.* **36**, 856–868 (2013).
83. Del Pozo, J. C. & Ramirez-Parra, E. Deciphering the molecular bases for drought tolerance in *Arabidopsis* autotetraploids. *Plant, Cell Environ.* **37**, 2722–2737 (2014).
84. Koch, M. A. *et al.* The Quaternary evolutionary history of Bristol rock cress (*Arabis scabra*, Brassicaceae), a Mediterranean element with an outpost in the

- north-western Atlantic region. *Ann. Bot.* (2020) doi:10.1093/aob/mcaa053.
85. Lysak, M. A., Koch, M. A., Beaulieu, J. M., Meister, A. & Leitch, I. J. The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol. Evol.* **26**, 85–98 (2009).
 86. Krisai, R. & Greilhuber, J. Cochlearia pyrenaica DC , das Löffelkraut , in Oberösterreich (mit Anmerkungen zur Karyologie und zur Genomgröße). *Beitr. Naturk. Oberösterreichs* **5**, 151–160 (1997).
 87. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. in *Methods in Molecular Biology* vol. 1962 227–245 (2019).
 88. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. in *Bioinformatics* vol. 19 (2003).
 89. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
 90. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at <https://github.com/najoshi/sickle>. 2011 (2011).
 91. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
 92. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
 93. Broad Institute. Picard Tools - By Broad Institute. *Github* (2009).
 94. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
 95. Jombart, T. & Ahmed, I. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
 96. Raj, A., Stephens, M. & Pritchard, J. K. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
 97. Monnahan, P. *et al.* Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. *Nat. Ecol. Evol.* **3**, 457–468 (2019).
 98. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* vol. 23 254–267 (2006).
 99. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu033.
 100. Fitak, R. R. optM: an R package to optimize the number of migration edges using threshold models. *J. Hered.* (2019).
 101. Sukumaran, J. & Holder, M. T. DendroPy: A Python library for phylogenetic computing. *Bioinformatics* (2010) doi:10.1093/bioinformatics/btq228.
 102. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor Improv.* (2007).
 103. Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis.