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

Magdalena Bohutínská[1,2*], Jakub Vlček[1,3,4], Sivan Yair[5], Benjamin Laenen[6], Veronika Konečná[1,2], Marco Fracassetti[6], Tanja Slotte[6], Filip Kolář[1,2,7*]

[1]Department of Botany, Faculty of Science, Charles University, Benátská 2, 128 01, Prague, Czech Republic.

[2]Institute of Botany, The Czech Academy of Sciences, Zámek 1, 252 43, Průhonice, Czech Republic.

[3]Biology Centre, The Czech Academy of Sciences, Branišovská 31, České Budějovice, Czech Czech Republic.

[4]Department of Zoology, Faculty of Science, University of South Bohemia, Branišovská 1670 České Budějovice, Czech Republic.

[5]Center for Population Biology, University of California, Davis, CA 95616, USA.

[6]Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden.

[7]Natural History Museum, University of Oslo, NO-0164 Oslo, Norway.

*correspondence: filip.kolar@natur.cuni.cz (Filip Kolář) and Magdalena Bohutínská (magdalena.holcova@natur.cuni.cz)

Date of submission: 24 March 2020

Word count (Introduction, Results & Discussion): 2962

Keywords: Genomic parallelism, evolutionary predictability, genetic divergence, gene reuse, allele reuse

Running title: Genomic parallelism varies with divergence

Abstract

Understanding the predictability of evolutionary change is of major importance in biology. Parallel adaptation provides unique insights through replicated natural experiments, yet mechanisms governing the ample variation in genomic parallelism remain unknown. Here, we investigate them using multi-scale genomic analyses of parallel evolution across populations, species and genera. By resequencing genomes of seven independent alpine lineages from two *Arabidopsis* species, we found that the degree of gene reuse decreases with increasing divergence between lineages. This relationship is well predicted by decreasing frequency of allele reuse, suggesting that availability of preexisting genetic variation is the prime mechanism. A meta-analysis demonstrated that the relationship further continues within the Brassicaceae family. Thus, we found empirical support for a long-standing hypothesis that the genetic basis of adaptive evolution is more predictable in closely related lineages while it becomes more contingent over larger evolutionary distances.

Introduction

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 common 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[4,13–18]. In contrast, we know little about 'genomic' parallelism, i.e., what fraction of the genome evolves in parallel 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], reuse of a limited number of genes[22–24] to abundant parallelism at both gene and functional levels[25,26]. Yet, there is a limited 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 promising candidate[14,27,28]. So far, phenotype-oriented meta-analyses suggest that both phenotypic convergence[27] and reuse of genes underlying parallel phenotypic traits[14] decreas with increasing age to the common ancestor. AtAt the genomic level, such a targeted multi-scale comparison is lacking. We reviewed a subset of published studies of genetic parallelism (109 cases, Supplementary Dataset 1) and found that gene reuse indeed tends to scale with divergence (Fig. 1a, Supplementary Fig. 1). Moreover, we found that allele reuse (repeated selection of the same haplotype, shared either via gene flow or from standing genetic variation) frequently underlies parallel adaptation between closely related lineages[29–32], while parallelism from de-*novo* mutations dominates between distantly related taxa[13]. This suggests that the degree of allele reuse may be the primary mechanism behind the hypothesized divergence-dependency of parallel genome evolution, possibly reflecting either genetic (weak hybridization barriers, widespread ancestral polymorphism between closely related lineages[33]) or ecological reasons (lower niche differentiation and geographical proximity[34,35]). However, the generally restricted focus of individual studies of genomic parallelism to a single level of divergence does not lend itself to a unified comparison across divergence scales. Although different ages of compared lineages affects a variety of evolutionary-ecological processes such as diversification rates, community structure or niche conservatism[34] and its role in shaping parallelism is intuitive, the hypothesis that genomic parallelism scales with divergence has not yet been tested.

Here, we aimed to test this hypothesis and investigate whether allele reuse is a major mechanism. We used replicated instances of adaptation to a challenging alpine environment, spanning a range of divergence from populations to tribes within the plant family Brassicaceae[36–41] (Fig. 1a). First, we took advantage of a unique naturally multi-replicated setup in the plant model genus *Arabidopsis*. We sequenced genomes of alpine and foothill populations from seven lineages and documented ubiquitous genomic parallelism at levels of single nucleotide polymorphisms (SNPs), genes and functional pathways. We thus tested whether the degree of gene reuse decreases with increasing divergence between the compared lineages (Fig. 1a) and whether this trend corresponds with a decrease in the fraction of genes exhibiting allele reuse. Finally, by extending towards deeper divergences using published candidate lists from five additional species across Brassicaceae, we asked whether there are limits to gene reuse above the species level. Overall, our empirical analysis sheds light on the expected degree of gene reuse under a certain level of divergence and identifies allele reuse as the dominant underlying mechanism.

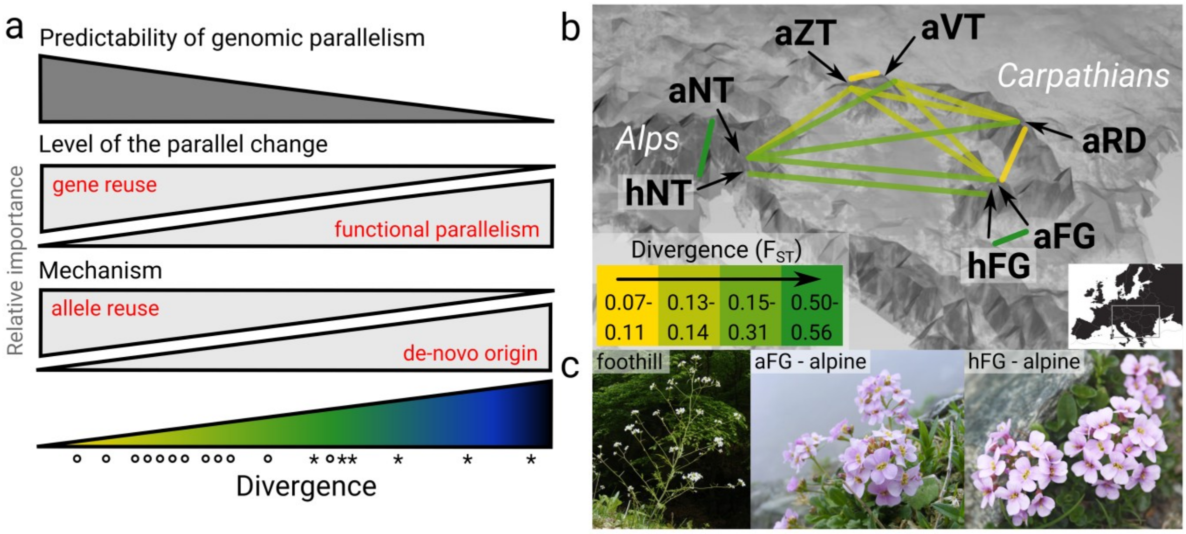


Fig. 1 Hypotheses regarding relationships between genomic parallelism and divergence and the *Arabidopsis* system used to address these hypotheses. (a) Based on our literature review we propose that genetically closer lineages adapt to a similar challenge more frequently by gene reuse, sampling suitable variants from the shared pool (allele reuse), which makes their adaptive evolution more predictable. Color ramp symbolizes rising divergence between the lineages (~0.02-18 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 independent alpine colonization events within the two *Arabidopsis* outcrosses from central Europe: *A. arenosa* (aNT, aZT, aVT, aRD and aFG) and *A. halleri* (hNT and hFG). Note that only two of the nine between-species pairs (dark green) are shown to aid visibility. The color scale corresponds to left part of a ramp used in a). (c) Representative phenotypes of foothill *A. arenosa* (aFG) and alpine *A. arenosa* (aFG) and *A. halleri* (hFG) demonstrating phenotypic convergence.

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 where alpine populations of *Arabidopsis arenosa* and *A. halleri* occur (a set of populations from one mountain region is further referred to as a 'lineage', Fig. 1b, Supplementary Fig. 2, Supplementary Table 1 and 2). Within each species, population structure analyses based on genome-wide nearly-neutral four-fold degenerate (4d) SNPs demonstrated clear grouping according to lineage but not alpine environment, suggesting parallel alpine colonization of each mountain region by a distinct genetic lineage (Supplementary Figs. 3 and 4). This was in line with separation histories between diploid populations of *A. halleri* estimated in Relate (Supplementary Fig. 5) and previous coalescent simulations on broader population sampling of *A. arenosa*[42]. 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 post-zygotic (ploidy difference) barriers between the alpine populations from these two lineages[43] 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, Supplementary Text 1).

We observed a gradient of neutral differentiation between the seven lineages (quantified as average pairwise 4d-Fst between foothill populations from each lineage, range 0.07 – 0.56, Supplementary Table 3). All populations showed high levels of 4d-nucleotide diversity (average of 0.023), as expected for strict outcrossers and no remarkable deviation from neutrality (the range of 4d-Tajima's D was -0.16 – 0.66, Supplementary Table 4). 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 positive 4d-Tajima's D (Supplementary Fig. 6).

Ubiquitous SNP, gene and function-level parallelism

Leveraging the seven natural replicates, we inferred the extent of parallel directional selection associated with alpine colonization at the level of SNP, gene and function and tested whether the degree of gene reuse decreases with increasing divergence between pairs of lineages. First, we identified candidate SNPs for alpine adaptation (hereafter 'SNP candidates') for each lineage. Using a conservative approach taking the intersection of F_{ST}-based divergence scans and candidate detection under a Bayesian framework (BayPass), we identified 3270 – 13753 SNP candidates in genic regions in each of the seven lineages (Fig. 2a). Of these, a significant number (12 – 793 SNPs, p < 0.05, Fisher's exact test, Supplementary Dataset 2) overlapped between two lineages (i.e., 'parallel SNP candidates') in 13 out of the possible 21 (62 %) pairwise comparisons (Supplementary Table 5). Notably, the overlaps were significant for all 11 pairwise comparisons among the lineages within a species but only in seven out of 10 pairwise comparisons across species. Parallel candidates were enriched for nonsynonymous over synonymous SNPs compared to the non-parallel candidate SNPs (p < 0.001, Fisher's exact test; Supplementary Table 6), suggesting that parallel changes were more likely at functionally relevant sites than non-parallel.

Then, we quantified parallelism at the level of genes and gene functions. We annotated SNP candidates into genes and identified significant overlaps between the candidate gene lists (p < 0.05, Fisher's exact test, Fig. 2b) among 15 out of 21 (71 %) lineage pairs

(Supplementary Dataset 3). All pairs within *A. arenosa* were significant while the comparison within *A. halleri* as well as half of the comparisons between lineages across the species resulted in non-significant overlaps (Supplementary Table 5). We then annotated the functions of gene candidates (using biological process terms in the GO) and extracted only significantly enriched functions. Of these, we found significant overlaps ($p < 0.05$, Fisher's exact test) among 17 out of 21 (81 %) pairs of lineages; nine out of 11 possible combinations within species (Fig. 2c, Supplementary Table 5). As an indirect validation of our approach we found that the parallel gene candidates encode proteins which likely facilitate adaptation to alpine stress via antioxidant production (*PAP1*), timing of flowering (*FT*), flower development (*AT5G49770*, *SEU*, *SLK3*), freezing tolerance (*ACC1*, *FAR5*) or defense against pathogens (*MSL5*). The majority of parallel function candidates was related to ion transport, defense, life cycle regulation and response to light/radiation (Supplementary Dataset 4). We did not find any evidence that our parallel candidate genes were subjected to weaker selective constraint (approximated by ratio of their nonsynonymus to synonymous diversity, see Supplementary Text 2, Supplementary Table 7 and Supplementary Fig. 7 for details) compared to the rest of the genome, nor did they cluster together in the regions of low recombination rates (Supplementary Figs. 8, 9).

Finally, for each pair of lineages we quantified the degree of parallelism as the proportion of overlapping SNP, gene and function candidates out of all candidates identified for these two lineages. The degree of parallelism significantly varied across the SNP, gene and function- level (mean proportion of parallel items across all pairwise comparisons = 0.022, 0.045 and 0.063, respectively, $p < 0.001$, GLM with binomial errors). Importantly, the degree of parallelism at both the SNP and gene-level (i.e., gene reuse) decreased with increasing divergence between the lineages (negative relationship between Jaccard's similarity in candidate SNP / gene identity among lineages and 4d-Fst; Mantel rM = -0.68 / -0.71, respectively, p-value < 0.01 in both cases, Fig. 2d, e). In contrast, the degree of parallelism by function did not correlate with divergence (rM = 0.06, p = 0.6, n = 21, Fig. 2f). These results suggest that while there are likely similar functions associated with alpine adaptation in each lineage, the molecular mechanism underlying adaptation differs more with increasing divergence among the lineages.

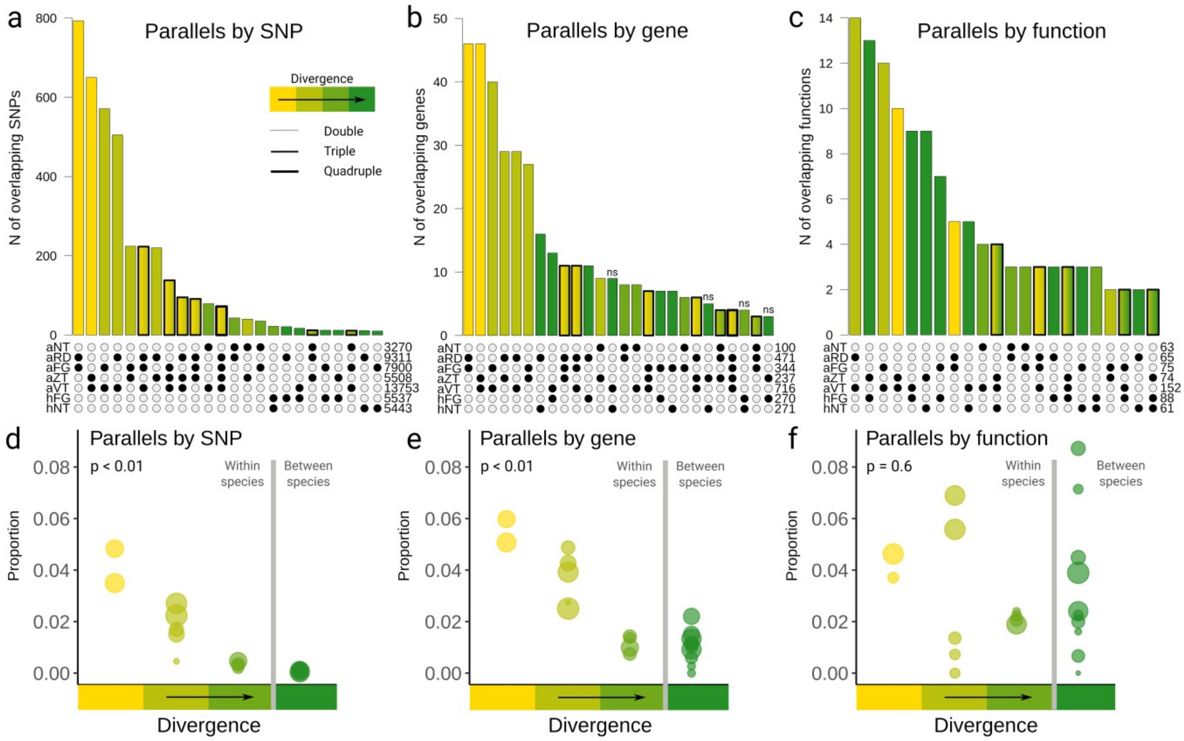


Fig. 2 Variation in SNP, gene and function-level parallelism and its relationship with divergence in *Arabidopsis arenosa* and *A. halleri*. Number of parallel genic SNPs (a), genes (b) and GO categories (c; functional classes) colored by increasing divergence between the compared lineages. Only overlaps of > 9 SNPs, > 2 genes and > 1 function are shown (for a complete overview see Supplementary Datasets 2-4). Numbers in the bottom-right corner of each panel show the total number of candidates in each lineage. Unless indicated ("ns") the categories 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. Proportions of parallel SNPs (d), genes (e; gene reuse) and functions (f) among all candidates identified within each pair of lineages (dot) binned into categories of increasing divergence (bins correspond to Fig. 1b) with significance levels inferred by Mantel test (see the text). Size of the dot corresponds to the number of parallel items.

Probability of allele reuse

Alleles of the genes reused by selection in different lineages could be either repeatedly recruited from a shared pool of variants (i.e., allele reuse, sampled from ancestral standing variation or acquired via gene flow) or may reflect independent *de-novo* mutations at the same locus (i.e., use of distinct alleles)[44]. We quantified the relative importance of these mechanisms in each pair of lineages and tested whether the probability of allele reuse, classified as the proportion of alleles repeatedly recruited from a shared pool out of all significant parallel gene candidates, decreases with increasing divergence.

For each parallel gene candidate, we inferred the most likely source of its selected variant(s) by using a recently developed likelihood-based approach that investigates patterns of shared hitchhiking from allele frequency covariance at positions surrounding the selected site[44]. In line with our expectations, the degree of allele reuse decreased with divergence ($p < 0.001$, GLM with binomial errors; Fig. 3a). This was driven mainly by a steeply decreasing proportion of variants acquired via gene flow. In contrast, the proportion of variants sampled from standing variation remained high even at the deepest interspecific comparison (43%; Fig. 3a, Supplementary Fig. 10) and the absolute number of *de-novo* originated variants was low across all divergence levels. This corresponds to predictions about a substantial amount of shared variation between related species with high genetic diversity[33] and frequent adaptive trans-specific polymorphism in *Arabidopsis*[11,45–47]. Absence of interspecific parallelism 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 (Supplementary Fig. 11).

Allele reuse covered the vast majority of the variation in gene reuse that was explained by divergence (variance partitioning by multiple regression on distance matrices, Fig. 3b) suggesting allele reuse as a likely mechanism underlying the observed divergence-dependency of gene reuse. Finally, we 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 $r = 0.83$, $p < 0.001$) (Fig. 3c). 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. 3d–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 rising divergence between parallel lineages, mostly because gene flow became less likely.

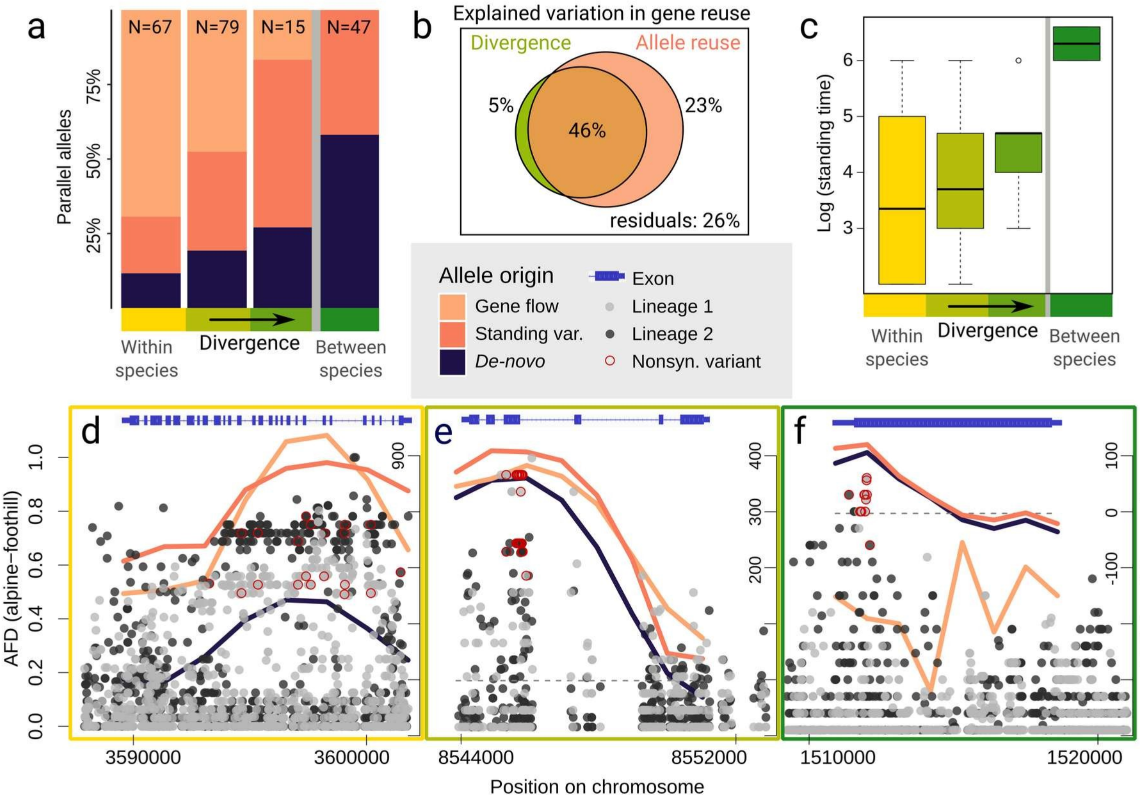


Fig. 3 Decreasing probability of allele reuse with increasing divergence in *Arabidopsis arenosa* and *A. halleri*. (a) Proportion of parallel candidate gene variants shared via gene flow between alpine populations from different lineages or recruited from ancestral standing variation (together describing 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 between lineages 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. (d) Parallel selection on variation shared via gene flow on gene ALA3, affecting vegetative growth and acclimation to temperature stresses[95]. (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[96]; here, *de-novo* origin was prioritized over standing variation model based on very high MCLE of standing time (see 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): allele frequency difference (AFD) between foothill and alpine population from each of the two lineages (range 0 – 1 in all plots), lines (right y-axis): maximum composite log-likelihood (MCL) difference from a neutral model assuming no parallel selection (all values above dotted grey line show the difference, higher values indicate higher support for the non-neutral model).

Parallelism at deeper divergence scale

To address our hypothesis at deeper phylogenetic scales, we performed a meta-analysis of gene- and function-level parallelism by complementing our data with gene candidate lists from six genome-wide studies of alpine adaptation from the Brassicaceae family[36–41] (species diverging 0.5 – 18 millions of years ago[48,49], Supplementary Tables 8, 9). We found significant parallelism both at the level of candidate genes (overlaps between 19 % of species pairs were significant, all within the *Arabidopsis* genus, p < 0.05, Fisher's exact test, Fig. 4a, Supplementary Dataset 5) and functions (33% significant overlaps, spread along all but the highest divergence category, p < 0.05, Fisher's exact test, Fig. 4b, Supplementary Dataset 6). The degree of parallelism by function was significantly higher than that of gene reuse across all pairwise comparisons between lineages (mean 0.032 vs. 0.021, respectively; p < 0.001, GLM with binomial errors). The relationship with divergence was non-significant both for gene- and function-level parallelism (Mantel, rM = 0.08 and 0.22, respectively, p > 0.05 in both, n = 21). However, the degree of gene reuse was significantly higher for comparisons within a genus (*Arabidopsis*) than between genera (p = 0.003, GLM with binomial errors) while such a trend was absent for parallel function candidates (p = 0.28). This demonstrates that while gene reuse was detectable only up to a within-genus level, functional parallelism was present also at deeper divergence levels, between distinct tribes of Brassicaceae.

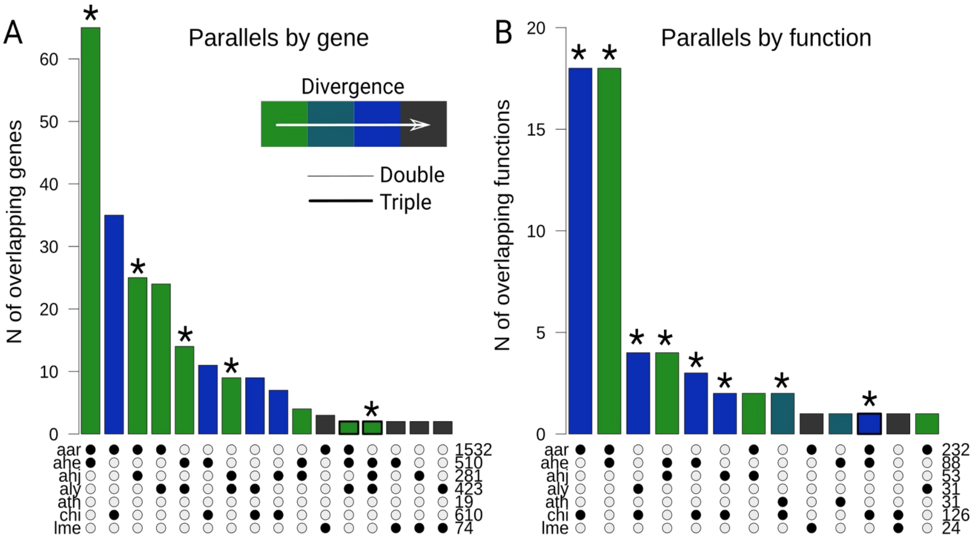


Fig. 4 Parallelism in alpine adaptation across species from Brassicaceae family. Number of genes and function candidates (GO categories) identified in parallel among species, colored by divergence between the species. Numbers in the bottom-right corner of each panel show the total number of candidate items (genes, functions) in each lineage; only overlaps of > 1 candidate item are shown (for a complete overview see Supplementary Datasets 5, 6). Categories indicated by an asterisk exhibited higher than random overlap of the candidates (p < 0.05, Fisher's exact test). Codes: aar: our data on *Arabidopsis arenosa* ahe: our data on *A. halleri* combined with *A. halleri* candidates from Swiss Alps[37], ahj:

Arabidopsis halleri subsp. *gemmifera* from Japan[41], aly: *Arabidopsis lyrata* from Northern Europe[39], ath*: Arabidopsis thaliana* from Alps[38], chi: *Crucihimalaya himalaica*[36], lme: *Lepidium meyenii*[40].

Discussion

By analyzing genome-wide variation over twelve 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 bring empirical support for earlier predictions on genetic paralellism[14,27], and present a general mechanism that may 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 well-reflects theoretical and empirical findings that the evolutionary potential of a population is dependent on availability of preexisting (standing or introgressed) genetic variation[50–52] and that the extent of ancestral polymorphism and gene flow decreases with increasing differentiation between gradually diverging lineages[33,53]. 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[54] and brings support to the emerging view of ubiquitous influence of divergence scale on different evolutionary and ecological mechanisms[34].

There are potentially additional, non-exclusive explanations for the observed divergence-dependency of gene reuse, although of 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[14,27]. Diversification of gene networks, however, typically increases at higher divergence scales than addressed here (millions of years of independent evolution[27]) 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. Second, as genetic divergence often corresponds to the spatial arrangement of lineages[55], 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[42]

In contrast, no relationship between the probability of gene reuse and divergence was shown in experimental evolution of different populations of yeast[56] raising a question about the generality of our findings. Our study addresses a complex selective agent (a multi-hazard alpine environment[57]) in order to provide insights into an ecologically realistic scenario relevant for adaptation in natural environments[14,53,54].. 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[58,59]. 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*[38] in our meta-analysis (Fig. 4a). However, considering the supporting evidence from the literature (Fig. 1a, Supplementary Fig. 1), and keeping the aforementioned restrictions in mind, we predict that our novel 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 prime 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 predictability of evolution, a topic critical for pest and disease control as well as for evolutionary theory.

Online Methods

Sampling

Arabidopsis arenosa and *A. halleri* are biannual to perennial outcrossers closely related to the model *Arabidopsis thaliana*. Both species occur primarily in low to mid elevations (to ~ 1000 m a.s.l.) across Central and Eastern Europe but scattered occurrences of morphologically distinct populations have been recorded from treeless alpine zones (> 1600 m) in several distinct mountain regions in Central-Eastern Europe[60,61]. Diploid *A. arenosa* populations colonized alpine stands only in Vysoké Tatry (aVT) mountains in Slovakia; more widespread autotetraploids (with random pairing among the four homologous chromosomes[62]) colonized alpine stands in Západné Tatry (aZT) mountains in Slovakia, Făgăraș (aFG) in central Romania, Rodna (aRD) in northern Romania (all these regions are part of the Carpathian Mts.) and Niedere Tauern (aNT) in South-Eastern Alps in Austria[42]. Exclusively diploid *A. halleri* species colonized high elevations in Făgăraș (hFG) in Romania and in South-Eastern Alps in Austria (hNT)[61].

The alpine populations in all cases occupy similar habitats (alpine screes and rocky outcrops in glacial cirques) and are separated by at least a 500 m altitudinal gap from their foothill counterparts, that also corresponds with timberline[42]. Alpine forms of both species are morphologically very distinct from foothill ones but similar together, with lower stature, cushion-like growth form, small, waxy and less hairy leaves and large, usually pinkish petals (Fig. 1c). Although the alpine populations resemble each other phenotypically, they are genetically more closely related to the foothill populations from the same region, suggesting parallel colonization in each range combined with phenotypic convergence[42,61]. Moreover, the widespread occurrence of both species in the foothills vs. rarity in alpine habitats and the fact that the alpine zone of European mountains was previously glaciated suggests that the colonization proceeded from low elevations to the alpine environment[42,61].

Here, we sampled and re-sequenced foothill (growing in elevations 460-980 m a.s.l.) as well as adjacent alpine (1625-2270 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*[63]*.* Altogether our dataset consists of 174 individuals and 22 populations of both species (Supplementary Table 1). Ploidy of each sequenced individual was checked using flow cytometry following[64]. The final dataset contained 2-4 populations per mountain range, eight individuals per population (only seven individuals were used in populations SCH and INE, one sample from each was excluded due to low quality).

Sequencing, raw data processing, variant calling and filtration

We extracted DNA from plant material stored in silica gel by CTAB protocol (for details see[64]). Each sample was individually barcoded with dual barcodes during library prep using Illumina TruSeq PCR free kit (Supplementary Dataset 7), ~ 15 samples were pooled and sequenced on one Illumina HiSeq X-ten lane in a pair-end mode (2x 150 bp) in Norwegian Sequencing Centre, University of Oslo and SNP&SEQ platform in Uppsala.

We used trimmomatic-0.36[65] to remove adaptor sequences and low quality base pairs (< 15 PHRED quality score). Trimmed reads longer than 100bp were mapped to reference genome *Arabidopsis lyrata*[66] in bwa-0.7.15[67] with default settings. Duplicated reads were identified by picard-2.8.1[68] and discarded together with reads that showed low mapping quality (< 25). Afterwards we used GATK (v. 3.7) to call and filter reliable variants and invariant sites according to best practices[69]. Namely, we used HaplotypeCaller to call variants per individual with respect to its ploidy level. Then we aggregated variants for all samples per species by GenotypeGVCFs.

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, StrandOddsRatio (SOR) > 3.0. We called invariant sites using the GATK pipeline similarly to how we did with variants, and we removed sites where QUAL was lower than 15. Both variants and invariants were masked for sites with average read depth higher than 2 times the standard deviation as these sites are most likely located in duplicated regions (duplicated in the genome of our target not in the reference) and regions with excessive heterozygosity, indicating likely paralogous regions mapped on top of each other (genes with > 5 sites which showed fixed heterozygosity in > 2 diploid populations; following[63]). In the final vcf we discarded genotypes with read depth lower than 8 and with more than 20% genotypes missing. Altogether, the dataset contained 11390267 and 6713051 SNPs after variant filtration in *A. arenosa* and *A. halleri*, respectively (Supplementary Table 2), and the average depth of coverage over both datasets was 32^x (Supplementary Dataset 7).

Population genetic structure

We calculated genome-wide 4d within- (nucleotide diversity (π) and Tajima's D[70]) and between- (F_{ST}[71]) population metrics using python3 ScanTools_ProtEvol pipeline (github.com/ mbohutinska/ScanTools_ProtEvol). 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[63]. 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). We found out that the genome-wide pairwise inter-population F_{ST} calculated using such sites strongly correlated with 4d-F_{ST} (Pearson's r = 0.93, p-value < 0.001) and therefore 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[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 (Supplementary Fig. 4). 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. Finally, we displayed genetic relatedness among individuals using principal component analysis (PCA) as implemented in adegenet[73].

Coalescent simulations in fastsimcoal2

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. We used wide range of initial parameters (effective population size, divergence times, migration rates, see attached *est* file, Supplementary Dataset 8) and an assumed mutation rate of 4.3e^{-8} inferred for *A. arenosa*[62].

First, we tested for the presence of a bottleneck associated with alpine colonization by comparing models with and without a bottleneck. For this analysis we focused on the outlier population with the highest positive 4d-Tajima's D (i.e., indicative of a bottleneck; LAC from aFG region) and constructed joint AFS using its foothill counterparts from the same (aFG) and next closest (aRD) region (Supplementary Fig. 6). Second, we tested whether *A. halleri* and *A. arenosa* alpine populations growing in the same mountain regions (i.e. aNT-hNT and aFG-hFG) had experienced recent interspecific gene flow since the last glacial maximum (LGM; approx. 10,000 generations ago assuming a generation time of two years), since their current alpine range was de-glaciated. We constructed the joint AFS from population pairs, iterating over all four combinations of sympatric alpine populations and compared models with recent (post-LGM) migration, ancient (pre-LGM) migration, migration in both periods and without migration (Supplementary Fig. 11; for *tpl* and *est* files see Supplementary Dataset 8).

For each scenario and joint AFS, we performed 50 independent fastsimcoal runs. We then extracted the highest likelihood partition for each fastsimcoal run, calculated Akaike information criterion (AIC) and summarized them across the 50 different runs, over the scenarios and different population pairs/trios. The scenario with the lowest mean AIC values was selected.

Coalescent rates in Relate

We used program Relate[75] to further refine the demographic history of our system by leveraging haplotype data of diploid populations of *A. arenosa* (aVT lineage) and *A. halleri* (hFG and hNT lineages). First, we phased diploid populations (separately for each species) using the program Shapeit (ver. 2)[76]. We used read aware phasing that accounts for phase information present in the filtered sequencing reads. We also used genetic map of *A. lyrata* as input information[77]. We took the phased data from Shapeit and oriented derived alleles based on the polarization table (adopted from[63]). Then we used the command RelateFileFormats to generate a distance file that accounts for uncalled sites. We ran the program Relate per individual chromosome (--mode All) to estimate genome wide genealogies with population size parameter -N set to 3.2e^{6} and mutation rate -m set to 4.3e^{-}

[8]. We used the output of the previous command (anc and mut) to estimate coalescence rates between and within populations (script EstimatePopulationSize.sh) in nine iterations with generation time set to 2 years. We ran the commands for each scaffold separately and then we did joint estimation based on information from all the scaffolds. Relative pairwise coalescence was calculated as the proportion of cross-coalescence rate between populations and intra-coalescence rate of one of the populations. All of these estimates were analyzed and visualized in the R script plot_population_size_modif.R.

Strength of genotypic associations

To design optimal window sizes for the F_{ST} scans, we calculated the strength of genotypic associations (a proxy of linkage disequilibrium, LD) over a range of distances between 10 bp and 50 kb for diploid populations of *A. arenosa* and *A. halleri*. We used PLINK[78] version v1.9 and function r2 (plink --vcf data/scaffold_*.vcf.gz --r2 --ld-window-kb 50 --ld-window 10 --ld-window-r2 0.001 --maf 0.05 --out results/scaffold_* --threads 4 --allow-extra-chr) and summarized the output using custom R scripts. We estimated the average LD as the distance at which genotypic correlation became constant.

Genome-wide scans for directional selection

To infer SNP candidates, 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}[79] for 1 kb windows along the genome. Based on the average genome-wide decay of genotypic correlations, (150-800 bp, Supplementary Fig. 12), we designed windows for the selection scans to be 1 kb, i.e., at least 200 bp larger than the estimated average LD. All calculations were performed using ScanTools_ProtEvol, and custom R scripts (github.com/mbohutinska/ProtEvol/selScans.R). Our F_{ST}-based detection of outlier windows was not largely biased towards regions with low recombination rate (as estimated based on the available *A. lyrata* recombination map[39] and also from our diploid population genomic data; Supplementary Text 2, Supplementary Figs. 8, 9). This corresponds well with outcrossing and high nucleotide diversity that aids divergence outlier detection in our species[80].

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). To do so, we divided the six possible contrasts among the four populations to four positive (alpine population contrasted with foothill) and two negative (alpine-alpine and foothill-foothill) contrasts. We assigned a rank to each window in each positive contrast, based on the value of F_{ST} (windows with the highest F_{ST} were assigned with the lowest rank) and summed up the ranks over the four positive contrasts per window. We identified windows with the lowest sum of ranks (top 5% outliers of minimum rank sum) as candidates for directional selection. To exclude the local changes uninformative about selection between the foothill and alpine environment, we further identified windows which were the top 5% outliers of F_{ST} in either of the negative contrasts and removed them from the candidate list. For three lineages with only one population pair available (aRD, hNT, hFG), we considered only positive contrast to detect 5% minimum rank outlier windows. We did not observe a decrease in numbers of parallel candidates for between-lineage comparisons including aRD in the same divergence category in *A. arenosa*, suggesting that using two, instead of four populations did not bias our detection of directional selection candidates (Fig. 2a, b).

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[81]. First, we extracted all variable sites for all populations within a particular lineage and calculated reference and alternative allele counts at each site in each population. Then we ran a core model of BayPass which estimates a covariance matrix (approximating the neutral structure among the populations) and differentiation (XtX measure) between populations. We used the default settings; 5000 iterations for the burn-in period, 25000 iterations after burn-in, recorded each 25th (i.e. size of thinning was 25) and 20 pilot runs to adjust the parameters of the MCMC proposal distributions. Then we simulated a set of “neutral” allele counts of 1000 alleles based on our estimated covariance matrix (function simulate.baypass in baypass_utils.R script from BayPass) and ran the same model on the simulated data. We estimated 95% quantile of “neutral” XtX statistics calculated from simulated data and used it to identify excess differentiation SNP candidates in our real dataset.

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. 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 SNP candidates fulfilling criteria of the two complementary selection scans. Using a more stringent threshold of 1% did not lead to qualitatively different results in regards to the relationship between parallelism and divergence (see Supplementary Text 3).

Annotation of SNP candidate variants

We annotated each SNP candidate and assigned it to a gene using SnpEff 4.3[82] following *A. lyrata* version 2 genome annotation[83]. We considered all variants in 5' 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.

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[84], using *A. thaliana* orthologs of *A. lyrata* genes obtained using biomaRt[85]. 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[84,86]. We used 'biological process' ontology and accepted only significant GO terms with more than five and less than 200 genes as very broad categories do not inform about the specific functions of selected genes (FDR = 0.05, Fisher's exact test). Re-analysis with 'molecular function' ontology led to qualitatively similar results (Supplementary Fig. 13).

Quantifying parallelism

At each level (SNP candidates, 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[87] package in R). Only genes with at least five SNPs were included in the background set as by our criteria (see above) no gene with less than five SNPs could be identified as a gene candidate. Next, we calculated the probability of SNP and 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. But 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 the most likely source of their potentially adaptive variant(s). We used the newly developed composite likelihood-based method DMC (Distinguishing among Modes of Convergent adaptation)[44] which leverages 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): (i) on the variation introduced via gene flow, (ii) on ancestral standing genetic variation and (iii) 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 (Supplementary Table 10). 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 Ne = 800 000 inferred from *A. thaliana* genome-wide mutation rate[88] and nucleotide diversity in our sequence data (Supplementary Table 4) and a recombination rate of 3.7e^{-8} determined from the closely related *A. lyrata*[39]. 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 (i) two alpine populations or (ii) 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.

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. To do so, we generated a distribution of differences between selection model MCLs and the neutral MCL by analyzing neutral data sets, simulated with ms[89], that had similar numbers of segregating sites and demographic history as our real data. We considered the MCL difference between a parallel and neutral model significant if it was higher than the maximum of the distribution of the differences from the simulated data (i.e. 21, a conservative estimate, Supplementary Fig. 14). We further focused on cases of parallel adaptation in alpine populations by interpreting only cases (loci) for which the model of parallel selection in the alpine environment had a higher MCL than in the foothill environment.

Then, for all genes with significant parallel alpine selection (208 in total) we inferred the most likely source of the parallel variant. Analytical theory and simulations show that the ability to distinguish between the standing variation model from the *de-novo* mutation and gene flow models based on MCL is limited at specific parameter combinations[44]. Specifically, this depends on the parameter in the standing variation model that specifies the amount of time the allele was standing between the divergence of selected populations and the onset of selection. One cannot distinguish the standing variation model from the *de-novo* mutation model when the maximum likelihood (ML) standing time is much longer than the divergence time of parallel lineages. The standing variation and gene flow models are indistinguishable when the ML standing time is much shorter than the divergence time between the populations experiencing selection[44]. Therefore, to categorize all genes into one of the three models, we assigned all genes showing highest support for the standing variation model with a ML standing time of more than 100 thousand and 2 millions of generations for within and between species, respectively, to the *de-novo* mutation model of parallelism. These borderline times were selected as three times the mean estimated times of divergence between *A. halleri* and *A. arenosa*[48] and between *A. arenosa* lineages[60,62]. Similarly, all genes showing highest support for the standing variation model with ML standing time less than 1000 and 1 million generations for within and between species, respectively, were assigned to the gene flow model of parallelism. These borderline times were selected as the lowest non-zero standing time parameters at which the models converged.

The borderline times used gave biologically meaningful results as genes with inferred parallel selection from *de-novo* mutations usually included highly differentiated SNPs at different positions in parallel populations while genes under likely parallel selection from standing variation contained highly differentiated SNPs that were identical (Fig. 3d, e). Moreover, using higher or lower cut-off time (Supplementary Table 10), did not lead to qualitatively different results in regards to the relationship between parallelism and divergence (Supplementary Text 4).

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

Integration of candidate lists from *Brassicaceae*

We gathered lists of gene candidates from all available studies focused on alpine adaptation in Brassicaceae[36–41], i.e., adaptation towards treeless high-elevation habitats addressed by studies of whole genome sequence data (Supplementary Table 8). We unified (merged) the gene candidate lists if multiple altitudinal gradients were screened within a species (except for *A. halleri* from Europe and Japan which were kept as separate units due to extraordinarily high spatial and genetic divergence of these two lineages)[61]. We obtained 19 – 1531 *A. thaliana* orthologues and annotated them into functions using gene ontology enrichment in the same way as described earlier.

We note that gene sets identified by the revisited studies may partly differ also due to the different genome properties of diverged species and varying methods used to detect candidate genes in them (for details see Supplementary Table 8). However, a dramatic rearrangement of gene number and/or gene functions genome-wide is unlikely among species from the same family[90]. Moreover, the total numbers of identified candidate genes in individual studies did not decrease with increasing divergence from *Arabidopsis* (e.g. analysis of distantly related *Crucihimalaya himalaica* identified a higher number of candidate genes than analyses of most of the *Arabidopsis* species, Fig. 4a). Thus the pattern of divergence-driven decrease of probability of gene-level parallelism does not reflect the mere limits of candidate detection methods. Still, keeping such potential technical limitations in mind, we did not interpret any specific values obtained from the analysis.

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_{XY}, Pearson $r = 0.89$, $p < 0.001$) 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 SNP / gene / function candidates in each pair of lineages and (i) the 4d-F_{ST} distance matrix (*Arabidopsis* dataset) or (ii) the time of species divergence (Brassicaceae meta-analysis) using the mantel test with 999 replications (ade4[91] package in R). Then, we tested whether the relative proportion of the two different evolutionary mechanisms of parallel variation (allele reuse vs. *de-novo* origin) relate to divergence using generalized linear models (R package stats[92]) 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[93]) 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, and following[94].

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 (Supplementary Dataset 7)

References

1. Blount, Z. D., Lenski, R. E. & Losos, J. B. Contingency and determinism in evolution: Replaying life's tape. *Science.* **362**, (2018).
2. Gould, S. J. *Wonderful life : the Burgess Shale and the nature of history*. (Norton, 1989).
3. Stern, D. L. & Orgogozo, V. Is genetic evolution predictable? *Science* **323**, 746–751 (2009).
4. Agrawal, A. A. Toward a Predictive Framework for Convergent Evolution: Integrating Natural History, Genetic Mechanisms, and Consequences for the Diversity of Life. *Am. Nat.* **190**, S1–S12 (2017).
5. Marvig, R. L., Sommer, L. M., Molin, S. & Johansen, H. K. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nat. Genet.* **47**, 57–64 (2015).
6. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45**, 1183–1189 (2013).
7. Palmer, A. C. & Kishony, R. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nat. Rev. Genet.* **14**, 243–248 (2013).
8. Rinkevich, F. D., Du, Y. & Dong, K. Diversity and Convergence of Sodium Channel Mutations Involved in Resistance to Pyrethroids. *Pestic. Biochem. Physiol.* **106**, 93–100 (2013).
9. Tabashnik, B. E., Brévault, T. & Carrière, Y. Insect resistance to Bt crops: Lessons from the first billion acres. *Nat. Biotechnol.* **31**, 510–521 (2013).
10. Reid, N. M. *et al.* The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science.* **354**, 1305–1308 (2016).
11. Preite, V. *et al.* Convergent evolution in Arabidopsis halleri and Arabidopsis arenosa on calamine metalliferous soils. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, (2019).
12. Lamichhaney, S. *et al.* Integrating natural history collections and comparative genomics to study the genetic architecture of convergent evolution. *Philos. Trans. R. Soc. B* **374**, (2019).
13. Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots

of phenotypic variation. *Evolution.* **67**, 1235–1250 (2013).

14. Conte, G. L., Arnegard, M. E., Peichel, C. L. & Schluter, D. The probability of genetic parallelism and convergence in natural populations. *Proc. R. Soc. B* **279**, 5039–5047 (2012).
15. Gompel, N. & Prud'homme, B. The causes of repeated genetic evolution. *Dev. Biol.* **332**, 36–47 (2009).
16. Kopp, A. Metamodels and phylogenetic replication: a systematic approach to the evolution of developmental pathways. *Evolution.* **63**, 2771–2789 (2009).
17. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
18. Yeaman, S., Gerstein, A. C., Hodgins, K. A. & Whitlock, M. C. Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genet.* **14**, e1007717 (2018).
19. Zou, Z. & Zhang, J. No Genome-Wide Protein Sequence Convergence for Echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
20. Cooper, K. L. *et al.* Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature* **511**, 41–45 (2014).
21. Birkeland, S. *et al.* Multiple genetic trajectories to extreme abiotic stress adaptation in Arctic Brassicaceae. *Mol. Biol. Evol.* (2020).
22. Takuno, S. *et al.* Independent Molecular Basis of Convergent Highland Adaptation in Maize. *Genetics* **200**, 1297–312 (2015).
23. Foote, A. D. *et al.* Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, (2015).
24. 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.
25. Manceau, M., Domingues, V. S., Linnen, C. R., Rosenblum, E. B. & Hoekstra, H. E. Convergence in pigmentation at multiple levels: mutations, genes and function. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365**, 2439–50 (2010).
26. Lim, M. C. W., Witt, C. C., Graham, C. H. & Dávalos, L. M. Parallel Molecular Evolution in Pathways, Genes, and Sites in High-Elevation Hummingbirds Revealed by Comparative Transcriptomics. *Genome Biol. Evol.* **11**, 1573–1585 (2019).
27. Ord, T. J. & Summers, T. C. Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evol. Biol.* **15**, 137 (2015).
28. Morales, H. E. *et al.* Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Sci. Adv.* **5**, (2019).
29. Oziolor, E. M. *et al.* Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science* **364**, 455–457 (2019).

30. Haenel, Q., Roesti, M., Moser, D., MacColl, A. D. C. & Berner, D. 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).

31. Lai, Y.-T. *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).

32. Alves, J. M. *et al.* Parallel adaptation of rabbit populations to myxoma virus. *Science.* **363**, 1319–1326 (2019).

33. Hudson, R. R. & Coyne, J. A. Mathematical consequences of the genealogical species concept. *Evolution (N. Y).* **56**, 1557–1565 (2002).

34. Graham, C. H., Storch, D. & Machac, A. Phylogenetic scale in ecology and evolution. *Glob. Ecol. Biogeogr.* **27**, 175–187 (2018).

35. Bradburd, G. S. & Ralph, P. L. Spatial Population Genetics: It's About Time. *Annu. Rev. Ecol. Evol. Syst.* **50**, 427–449 (2019).

36. Zhang, T. *et al.* Genome of Crucihimalaya himalaica, a close relative of Arabidopsis, shows ecological adaptation to high altitude. *Proc. Natl. Acad. Sci.* **116**, 7137–7146 (2019).

37. Rellstab, C. *et al.* Local adaptation (mostly) remains local: Reassessing environmental associations of climate-related candidate SNPs in Arabidopsis halleri. *Heredity (Edinb).* **118**, 193–201 (2017).

38. Günther, T., Lampei, C., Barilar, I. & Schmid, K. J. Genomic and phenotypic differentiation of *Arabidopsis thaliana* along altitudinal gradients in the North Italian Alps. *Mol. Ecol.* **25**, 3574–3592 (2016).

39. Hämälä, T. & Savolainen, O. Genomic Patterns of Local Adaptation under Gene Flow in Arabidopsis lyrata. *Mol. Biol. Evol.* **32**, 2557–2571 (2019).

40. Zhang, J. *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).

41. Kubota, S. *et al.* A Genome Scan for Genes Underlying Microgeographic-Scale Local Adaptation in a Wild Arabidopsis Species. *PLoS Genet.* **11**, e1005361 (2015).

42. Knotek, A. *et al.* Parallel alpine differentiation in Arabidopsis arenosa. *bioRxiv* (2020) doi:10.1101/2020.02.13.948158.

43. Wos, G. *et al.* Role of ploidy in colonization of alpine habitats in natural populations of Arabidopsis arenosa. *Ann. Bot.* **124**, 255–268 (2019).

44. Lee, K. M. & Coop, G. Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data. *Genetics* **207**, 1591–1619 (2017).

45. Marburger, S. *et al.* Interspecific introgression mediates adaptation to whole genome duplication. *Nat. Commun.* **10**, 5218 (2019).

46. Arnold, B. J. *et al.* Borrowed alleles and convergence in serpentine adaptation. *Proc.*

Natl. Acad. Sci. U. S. A. **113**, 8320–5 (2016).

47. Guggisberg, A. *et al.* The genomic basis of adaptation to calcareous and siliceous soils in Arabidopsis lyrata. *Mol. Ecol.* **27**, 5088–5103 (2018).
48. 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).
49. 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).
50. Ralph, P. L. & Coop, G. The Role of Standing Variation in Geographic Convergent Adaptation. *Am. Nat.* **186**, S5-23 (2015).
51. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
52. Thompson, K. A., Osmond, M. M. & Schluter, D. Parallel genetic evolution and speciation from standing variation. *Evol. Lett.* **3**, 129–141 (2019).
53. Charlesworth, B., Charlesworth, D. & Barton, N. H. The Effects of Genetic and Geographic Structure on Neutral Variation. *Annu. Rev. Ecol. Evol. Syst.* **34**, 99–125 (2003).
54. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, (2020).
55. Ramachandran, S. *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**, 15942–7 (2005).
56. Spor, A. *et al.* Phenotypic and genotypic convergences are influenced by historical contingency and environment in yeast. *Evolution* **68**, 772–790 (2014).
57. Körner, C. *Alpine Plant Life*. (Springer Berlin Heidelberg, 2003). doi:10.1007/978-3-642-98018-3.
58. Liu, S., Ferchaud, A.-L., Grønkjaer, P., Nygaard, R. & Hansen, M. M. Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Mol. Ecol.* **27**, 4725–4743 (2018).
59. Vogwill, T., Phillips, R. L., Gifford, D. R. & Maclean, R. C. Divergent evolution peaks under intermediate population bottlenecks during bacterial experimental evolution. *Proc. R. Soc. B Biol. Sci.* **283**, (2016).
60. 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).
61. Šrámková-Fuxová, G. *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).

62. 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).

63. Monnahan, P. *et al.* Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. *Nat. Ecol. Evol.* **3**, 457–468 (2019).

64. Kolář, F. *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.* **119**, 673–688 (2016).

65. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

66. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–81 (2011).

67. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. http://github.com/lh3/bwa. (2013).

68. Broad Institute. Picard Tools. *Broad Institute, GitHub repository*.

69. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).

70. Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585–595 (1989).

71. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).

72. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* **8**, (2012).

73. Jombart, T. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).

74. Excoffier, L. & Foll, M. fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).

75. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

76. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).

77. Aalto, E. A., Koelewijn, H. P. & Savolainen, O. Cytoplasmic male sterility contributes to hybrid incompatibility between subspecies of arabidopsis lyrata. *G3 Genes, Genomes, Genet.* **3**, 1727–1740 (2013).

78. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

79. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N. Y).* **38**, 1358 (1984).

80. Yant, L. & Bomblies, K. Genomic studies of adaptive evolution in outcrossing Arabidopsis species. *Curr. Opin. Plant Biol.* **36**, 9–14 (2017).

81. Gautier, M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555–1579 (2015).

82. 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).

83. Rawat, V. *et al.* Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLoS One* **10**, e0137391 (2015).

84. Alexa, A. & Rahnenführer, J. *Gene set enrichment analysis with topGO*. http://www.mpi-sb.mpg.de/~alexa (2018).

85. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

86. Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. *Bioinformatics* **23**, 3024–3031 (2007).

87. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5**, (2015).

88. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in arabidopsis thaliana. *Science (80-.).* **327**, 92–94 (2010).

89. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).

90. Panchy, N., Lehti-Shiu, M. & Shiu, S. H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).

91. Dray, S. & Dufour, A. B. The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).

92. R Core Team. R: A Language and Environment for Statistical Computing. (2018).

93. Goslee, S. C. & Urban, D. L. The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* **22**, 1–19 (2007).

94. Lichstein, J. W. Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecol.* **188**, 117–131 (2007).

95. McDowell, S. C., López-Marqués, R. L., Poulsen, L. R., Palmgren, M. G. & Harper, J.

F. Loss of the Arabidopsis thaliana P4-ATPase ALA3 Reduces Adaptability to Temperature Stresses and Impairs Vegetative, Pollen, and Ovule Development. *PLoS One* **8**, (2013).

96. Kami, C. *et al.* Nuclear Phytochrome A Signaling Promotes Phototropism in *Arabidopsis*. *Plant Cell* **24**, 566–576 (2012).

Acknowledgements

This manuscript greatly benefited from constructive feedback of Graham Coop, Mike Nowak, Antonín Machač, Anja Westram, Pádraic Flood, Kristin Lee, Timothy Sackton, Martin Weiser and Clément Lafon-Placette. The authors further thank Daniel Bohutínský, Frederick Rooks, Jakub Hojka, Eliška Záveská and Peter Schönswetter for help with field collections; Gabriela Šrámková and Lenka Flašková for help with laboratory work and Doubravka Požárová for help with figure editing. This work was supported by the Czech Science Foundation (GACR project 17-20357Y to FK) and a student grant of the Charles University Grant Agency (GAUK 284119 to MB). MF was supported by a grant from the Swedish Research Council (grant #621-2013-4320 to TS). BL was supported by a grant from SciLifeLab to TS. 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 (NGI) 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.

Author Contributions

FK and MB conceived the study. FK and MB performed field collections. MB, JV, SY, BL and VK performed analyses with input from FK, MF and TS. MB and FK wrote the manuscript with primary input from JV, SY and TS. All authors edited and approved the final manuscript.

Competing Interests statement

The authors declare no competing interests.

Supplementary Dataset 1: Case studies of parallel genetic evolution from isolated model systems categorized based on divergence between parallel lineages.

Supplementary Dataset 2: The list of SNP candidates within each *Arabidopsis* lineage and their overlap among lineages.

Supplementary Dataset 3: The list of gene candidates within each *Arabidopsis* lineage and their overlap among lineages.

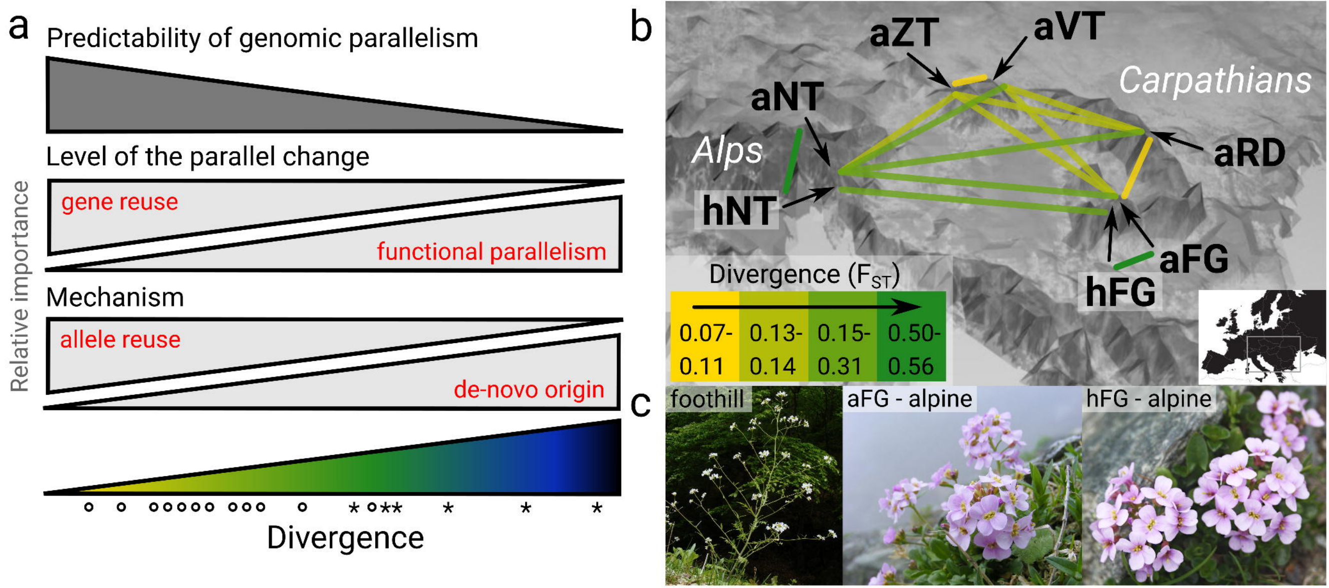
Supplementary Dataset 4: The list of function candidates within each *Arabidopsis* lineage and their overlap among lineages.

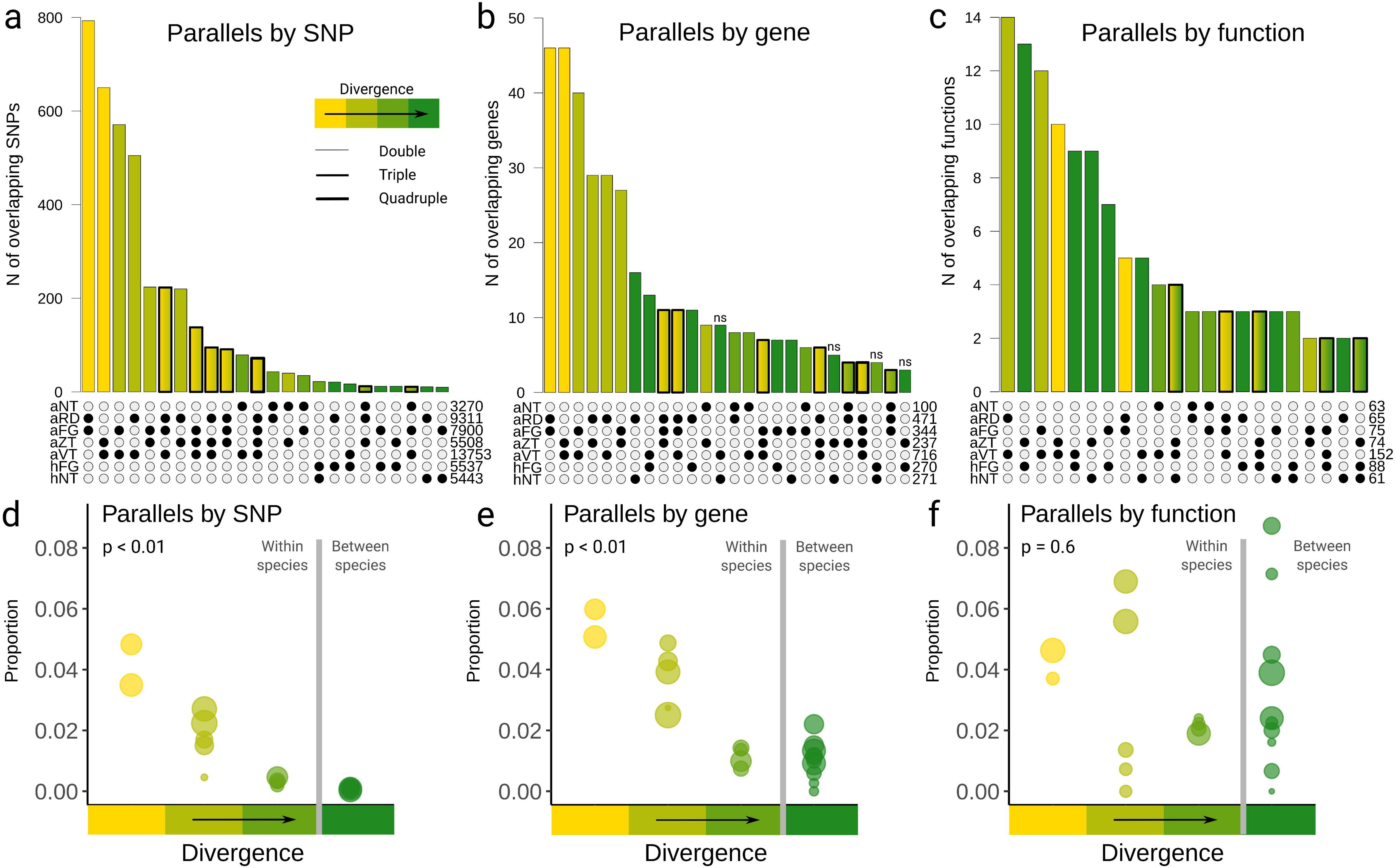
Supplementary Dataset 5: The list of gene candidates from each species from the family Brassicaceae and their overlap.

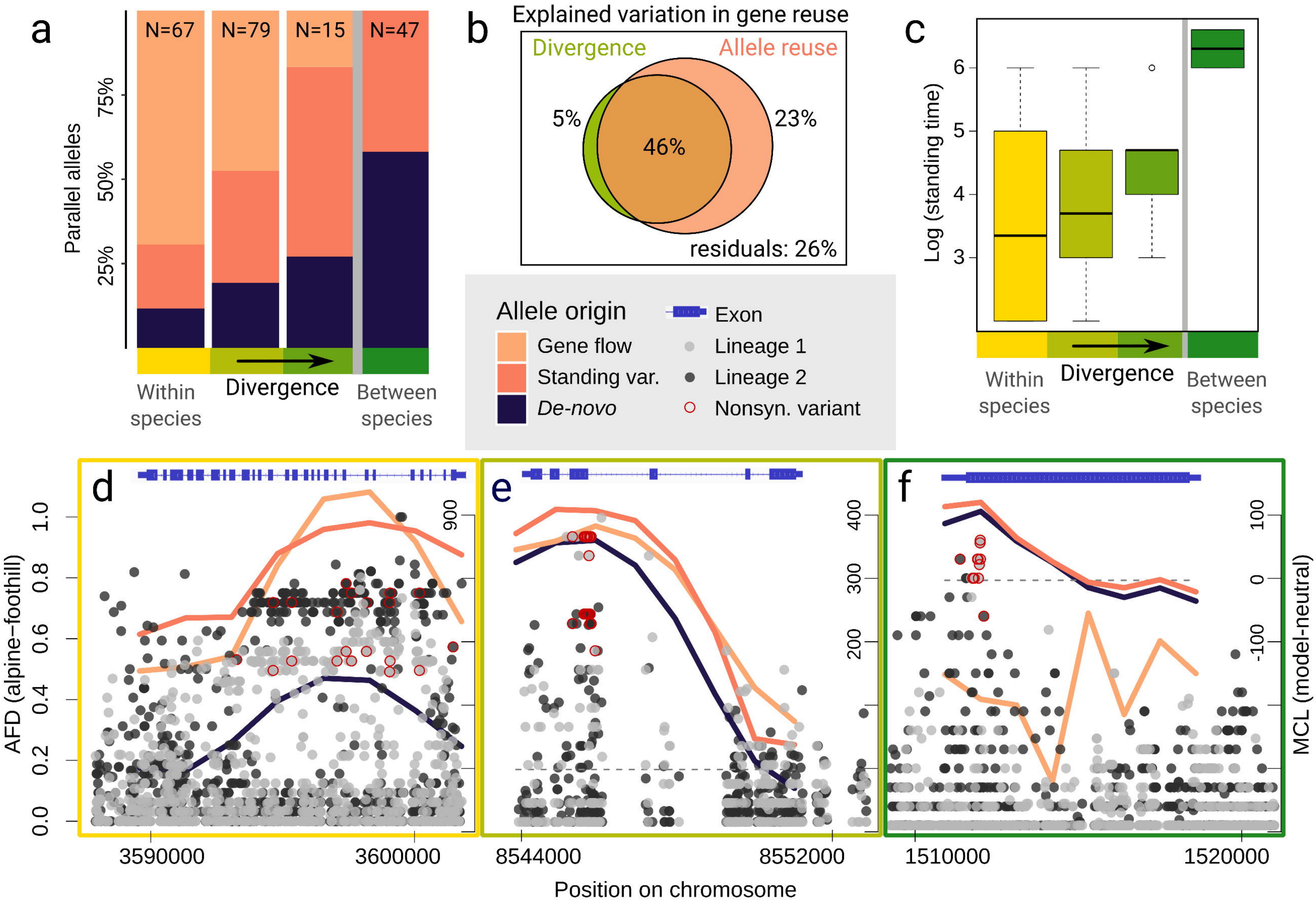
Supplementary Dataset 6: The list of function candidates from each species from the family Brassicaceae and their overlap.

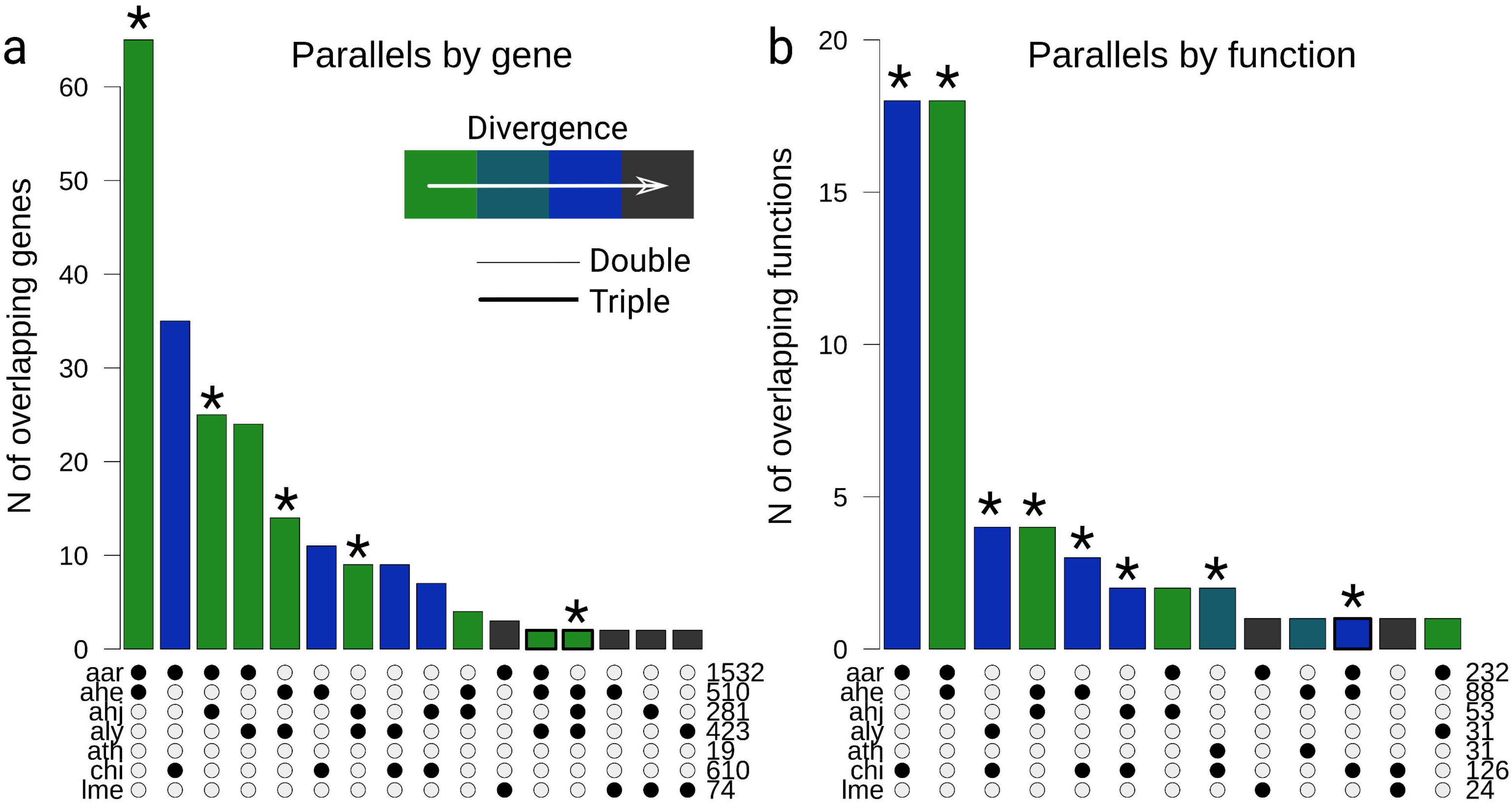
Supplementary Dataset 7: Sequence processing quality assessment of each re-sequenced individual.

Supplementary Dataset 8: 'Template' and 'distribution' (*tpl* and *est*) files used in coalescent simulations.


a
Predictability of genomic parallelism
Level of the parallel change
gene reuse
functional parallelism
Mechanism
allele reuse
de-novo origin
Relative importance
° ° °°°°° °°° ° *°** * * *
Divergence
b
aZT
aVT
Carpathians
aNT
Alps
aRD
hNT
aFG
hFG
Divergence (F_{ST})
0.07-0.11
0.13-0.14
0.15-0.31
0.50-0.56
c
foothill
aFG - alpine
hFG - alpine


a
Parallels by SNP
N of overlapping SNPs
0
200
400
600
800
Divergence
Double
Triple
Quadruple
aNT 3270
aRD 9311
aFG 7900
aZT 5508
aVT 13753
hFG 5537
hNT 5443
b
Parallels by gene
N of overlapping genes
0
10
20
30
40
50
ns
ns
ns
ns
aNT 100
aRD 471
aFG 344
aZT 237
aVT 716
hFG 270
hNT 271
c
Parallels by function
N of overlapping functions
0
2
4
6
8
10
12
14
aNT 63
aRD 65
aFG 75
aZT 74
aVT 152
hFG 88
hNT 61
d
Parallels by SNP
p < 0.01
Within species
Between species
Proportion
0.00
0.02
0.04
0.06
0.08
Divergence
e
Parallels by gene
p < 0.01
Within species
Between species
Proportion
0.00
0.02
0.04
0.06
0.08
Divergence
f
Parallels by function
p = 0.6
Within species
Between species
Proportion
0.00
0.02
0.04
0.06
0.08
Divergence


a
N=67
N=79
N=15
N=47
Parallel alleles
75%
50%
25%
Within species
Divergence
Between species
b
Explained variation in gene reuse
Divergence
Allele reuse
5%
46%
23%
residuals: 26%
c
Log (standing time)
6
5
4
3
Within species
Divergence
Between species
Allele origin
Gene flow
Standing var.
De-novo
Exon
Lineage 1
Lineage 2
Nonsyn. variant
d
e
f
AFD (alpine−foothill)
1.0
0.8
0.6
0.4
0.2
0.0
900
400
300
200
100
0
-100
MCL (model-neutral)
3590000
3600000
8544000
8552000
1510000
1520000
Position on chromosome


a
Parallels by gene
N of overlapping genes
0
10
20
30
40
50
60
Divergence
Double
Triple
*
aar 1532
ahe 510
ahj 281
aly 423
ath 19
chi 610
lme 74
b
Parallels by function
N of overlapping functions
0
5
10
15
20
aar 232
ahe 88
ahj 53
aly 31
ath 31
chi 126
lme 24