PNAS

Main Manuscript for

Repeated global adaptation across plant species

Gabriele Nocchi[1], James R. Whiting[1], Samuel Yeaman[1]

1. Department of Biological Sciences, University of Calgary, Calgary, Canada T2N 1N4

Corresponding author: Gabriele Nocchi

Email: gabriele.nocchi@ucalgary.ca

Author Contributions: GN analysed the data and wrote the manuscript. JW performed SNP calling, developed the methods to assess pleiotropy and provided help and suggestions for the analyses. SY conceived the project, provided overall guidance, and helped with writing the manuscript.

Competing Interest Statement: NA

Classification: Biological Sciences - Evolution

Keywords: global adaptation, selective sweep, pleiotropy, repeated adaptation

This PDF file includes:

Main Text
Figures 1 to 4

Significance

Global adaptation occurs when a species undergoes selection toward a common optimum throughout its natural range. While instances of global adaptation are widespread in the literature, there is a shortage of comparative studies aimed at understanding its genetic architecture and how it contrasts with that of local adaptation. This research compares global selective sweeps across 17 plant species to uncover the attributes of the genetic loci repeatedly involved in adaptation. We show that global adaptation tends to rely on genes with reduced pleiotropy and is characterized by increased levels of gene duplication. This finding contrasts with recent observations of increased pleiotropy in genes driving local adaptation, reflecting the opposing dynamics underlying these two evolutionary processes.

Abstract

Global adaptation occurs when all populations of a species undergo selection toward a common optimum. This can occur by a hard selective sweep with the emergence of a new globally advantageous allele that spreads throughout a species' natural range until reaching fixation. This evolutionary process leaves a temporary trace in the region affected, which is detectable using population genomic methods. While selective sweeps have been identified in many species, there have been few comparative and systematic studies of the genes involved in global adaptation. Building upon recent findings showing repeated genetic basis of local adaptation across independent populations and species, we asked whether certain genes play a more significant role in driving global adaptation across plant species. To address this question, we scanned the genomes of 17 plant species to identify signals of repeated global selective sweeps. Despite the substantial evolutionary distance between the species analysed, we identified several gene families with strong evidence of repeated positive selection. These gene families tend to be enriched for reduced pleiotropy, consistent with predictions from Fisher's evolutionary model and the cost of complexity hypothesis. We also found that genes with repeated sweeps exhibit elevated levels of gene duplication. Our findings contrast with recent observations of increased pleiotropy in genes driving local adaptation, consistent with predictions based on the theory of migration-selection balance.

Introduction

Plant species occupy a wide range of niches, adopt very different life history strategies, and inhabit environments with drastically different biotic pressures (1). Due to their sessile nature, plants must contend with the biotic and abiotic stresses they encounter directly in their environment, therefore phenotypic plasticity and genomic adaptation are of critical importance in the plant kingdom (1). To no surprise, local adaptation has been more commonly detected in plants than in animals (2,3,4,5).

Local adaptation occurs when a species inhabits a heterogeneous environment with spatial variability in the optimal phenotype. This can lead to the evolution of spatially differentiated genotypes along environmental gradients that exhibit fitness trade-offs when transplanted, with local genotypes providing higher fitness than foreign (4,5,6). By contrast, global adaptation occurs when all populations of a species experience selection towards the same optimum, resulting in the gradual refinement of a trait that is advantageous throughout all the environments inhabited by a species, such as the evolution of opposable thumbs in ancestral humans (6). While both global and local adaptation involve positive selection, the tension between migration and divergent natural selection inherent in local adaptation can result in different predictions about the evolution of genetic architecture. Local adaptation will tend to involve alleles with larger effects and more tightly linked than global adaptation, as such architectures are better able to resist the homogenizing effect of migration (6). However, while this is a clear theoretical prediction, it has not been extensively tested using empirical data.

At the molecular level, global adaptation can occur via hard or soft selective sweeps or through more subtle shifts in allele frequency at many loci (7). With a hard selective sweep (8,9,10) a new globally advantageous mutation rapidly spreads throughout a species' natural range until it reaches fixation. During this process, the affected genomic region displays a distinctive signature marked by diminished genetic diversity and increased homozygosity, a shift of the site frequency spectrum (SFS) toward low and high-frequency variants (11,12), and particular patterns of linkage disequilibrium (LD) characterized by elevated LD on each side of the selected site and decreased LD between sites located on different sides (13). This trace can be detected by scanning intraspecific genetic data, and various population genomics methods have been developed to detect sweeps (10,14,15). After fixation of the beneficial variant, new mutations and recombination in the region slowly decay the genomic signature typical of a hard selective sweep, therefore genomic scans can only detect hard sweeps within a restricted time frame (14). Global adaptation can also arise through positive selection on standing genetic variation, which tends to result in soft selective sweeps (16) or more subtle allele frequency shifts at many loci (7). When a beneficial variant is already segregating in a population before being subjected to positive selection, the selective process does not affect linked neutral polymorphisms as much as in hard-selective sweeps, making the detection of soft-sweeps much more challenging (10,16).

While selective sweeps have been reported in many species across both plants and animal kingdoms (17,18,19,20,21,22,23,24,25,26,27,28,29,30,31), there has been limited comparative and systematic genome-wide study of the repeatability of gene involvement in global adaptation across multiple species. On the other hand, several studies have demonstrated that the same genetic loci are observed repeatedly driving local adaptation in different populations or species (32,33,34,35,36,37,38,39,40,41). Whether adaptation is local or global, one important factor affecting the extent of repeatability is genotypic redundancy. If a given adaptive trait is characterized by limited genotypic redundancy (6), few loci are available to produce the mutations needed to reach the phenotypic optimum, therefore more repeatability is expected. In such cases, adaptive loci usually have large effects, a pattern that has recently emerged for several cases of local adaptation (42). On the other hand, if a trait is highly polygenic and driven by numerous alleles of small and interchangeable effect (6,43), a multitude of genotypes could potentially yield the same optimum phenotype. In such cases, fewer loci are expected to exhibit repeated contributions to adaptation in independent bouts of evolution across species or lineages.

As repeated genetic basis of adaptation has been identified among different populations and in closely related species, it is intriguing to assess whether repeatability is observed at greater phylogenetic distances (41, 44). It is expected that more recently diverged lineages will demonstrate a higher degree of shared gene utilization in adaptation, owing to limited functional differentiation, increased allele sharing, comparable genomic architecture, and similar life histories/adaptive strategies (41). Also, while factors such as pleiotropy, mutability, and average mutation effect size likely vary among genes and would be theoretically predicted to affect the repeatability of adaptation, there has been little systematic study of the importance of such factors in empirical datasets.

To address these questions, we retrieved publicly available high-quality whole-genome sequencing (WGS) data from thousands of individuals from 17 natural plant and forest trees species distributed across four continents, including woody and herbaceous angiosperms species (Fig. 1 and SI Appendix, Table S1). We employed linkage disequilibrium-based genomic scans for selective sweeps (45) within each species to identify putative genes under positive selection, and then used *PicMin* (46) to identify genes that are enriched for strong sweep signatures across multiple species. We explored the attributes of genes with repeated global sweeps by examining the pleiotropic potential and gene duplication levels of the repeatedly swept gene families identified. Our assessment of pleiotropy leveraged available gene expression data from *Arabidopsis thaliana* and *Medicago truncatula* to generate different pleiotropy measures based on gene tissue specificity and position within gene co-expression networks. We also assessed levels of gene duplication based on our orthology assignment. We contextualized these findings in view of Fisher's model of evolution and the cost of complexity hypothesis (47, 48), as well as recent theoretical work on migration-selection balance (6). To contrast our results for the architecture of global adaptation with that of local adaptation, we

compare our observations to results from a related study on local adaptation (49), which employed the same bioinformatic methods and incorporated 14 of the 17 datasets used in our analysis.

Results

Datasets assembly and filtering

Whole-genome sequencing data (WGS) from 17 non-domesticated and non-invasive angiosperm species, including both herbaceous plants and forest trees, was retrieved from SRA/ENA (Fig. 1 and SI Appendix, Table S1). These datasets included exclusively natural populations and ranged from closely related sister species known to hybridize (50) to species separated by up to 160 million years of evolution (Fig. 1B).

Single nucleotide polymorphisms (SNPs) were identified by applying a uniform pipeline to all datasets for consistency, with filtering steps to exclude singleton SNPs and individuals with high relatedness (SI Appendix, Fig. S1; see *Methods* for details). The SNP calling and filtering process resulted in between 20 and 565 individuals per dataset, and between 667,641 and 50,268,965 SNPs (Fig. 1D and SI Appendix, Table S1). We assessed population structure and found a wide range of patterns, with some species showing nearly panmictic structure across wide spatial scales (e.g. *Eucalyptus albens*; SI Appendix, Fig. S2) and others exhibiting substantial sub structuring, even across small spatial scales (e.g. *Helianthus argophyllus*) (Fig. 1C and SI Appendix, Fig. S2). In most species, the range of samples included in our study is sufficient to either cover multiple populations with detected genetic structure (SI Appendix, Fig. S2) or span a substantial portion of the overall species range (>20%), as estimated using data from GBIF (SI Appendix, Fig. S3). The only exception is *P. tremula*, which exhibited a panmictic genetic structure and had samples spanning approximately 11.3% of its range (SI Appendix, Fig. S2 and Fig. S3). Despite this, it included several sampling sites scattered over approximately 1,100 km of latitude comprising a large number of individuals (SI Appendix, Fig. S2 and Table S1), so it still should be seen as representative of capturing patterns of adaptation across a broad range of space. It is important to note that the native range estimation of the species was likely overestimated due to the possible presence of erroneous observations, observations of planted specimens and observations within invasive ranges in the GBIF data.

Orthology inference

The phylogenetic relationship among the genes of the 12 species with reference genome assemblies was inferred with *Orthofinder2* (51), which was used to group genes into orthology groups, or orthogroups. These included both orthologous genes, which are homologous genes separated by a speciation event, and paralogous genes, which are homologous genes diverged from a within-species duplication event. In total, *Orthofinder* assigned 376,881 genes out of 415,552 total genes to 31,582 orthogroups. These included 9,521 orthogroups specific to individual species, 7,919 orthogroups with representation from all species, and 633 single-copy orthogroups. Subsequently, this assignment was refined by excluding cases where a given species had more than 10 paralogues within a single orthogroup (but retaining the orthogroup in other species with fewer than 10 paralogues). We also excluded any gene with insufficient sequencing coverage within a given species, and excluded any orthogroups with representation in fewer than 7 species. These filtering steps resulted in a final set of 13,268 orthogroups for subsequent analysis, which exhibited low levels of paralogy and high species inclusion (Fig. 2). The mean and median number of paralogues per orthogroup was 1.72 and 1.67 per species respectively, while mean and median species number per orthogroup was 15.37 and 15.42 (per species). Notably, *B. stricta* yielded results for significantly fewer orthogroups compared to the other species, owing to insufficient sequencing coverage of the data (52) across many genic regions (Fig. 2 and SI Appendix,Table S1). Overall, the topology of the tree inferred with *Orthofinder2* largely matched the species tree derived from *TimeTree* (https://timetree.org/), with the exception of *M. truncatula* and *E. grandis. Orthofinder2* placed *M. truncatula* closer to the Brassicaceae family, relative to *E. grandis* (Fig. 2D and SI Appendix, Fig. S4). It is worth noting that *TimeTree* estimates divergence

between species using a simple average across published time estimates from scientific literature (53), therefore a minor mismatch with the protein sequences-based tree inferred with *Orthofinder2* is not a reason of concern.

Repeated selective sweeps

We used *OmegaPlus* (45) to scan for global selective sweeps within species followed by *PicMin* (46) to identify orthogroups with repeated sweep signatures across several species. After performing a False Discovery Rate (FDR) correction to the p-values from *PicMin* based on the number of orthogroups tested, we detected 33 orthogroups with significant signatures of repeated sweeps, at $q < 0.1$ (Fig. 3A).

Our application of *PicMin* tests for an enrichment of strong sweep signatures across multiple species but does not perform a test of which species contribute to the signal. To explore any patterns in the species driving these results, we classify any gene with an empirical p-value of < 0.1 as a "driving gene". Contribution towards the signatures of repeatability in the 33 OGs varied between species, ranging from approximately 14% (*B. stricta*) driving genes to about 68% (*E. sideroxylon*) (Fig. 3B). The repeatability signal is scattered throughout the phylogeny, with low *OmegaPlus* empirical p-values present in every species tested (Fig. 3C). However, two clusters of closely related species (*Eucalyptus spp.* and *Helianthus spp.*) were top contributors based on the count of driving genes and displayed a denser heat signature indicating an enrichment of driving genes (Fig. 3B, Fig. 3C and SI Appendix, Fig. S5). Despite this, results for only a single orthogroup appeared to be driven solely by species from the *Eucalyptus* genus, OG0010509, and none was solely driven by the *Helianthus* genus (Fig. 3C and SI Appendix, Fig. S5). However, in four orthogroups (OG0017633, OG0019910, OG0022308, OG0017585), driving genes were found exclusively in multiple *Eucalyptus* species in conjunction with multiple *Helianthus* species (Fig. 3C and SI Appendix, Fig. S5).

It is biologically expected that closely related species might evolve similarly and therefore might be more likely to experience sweeps in the same genes or genes families, particularly if they occupy an overlapping habitat and experience the same, or similar, selective pressures (41, 54). Alternatively, sweeps detected on the same gene of closely related species of the same genus could have originated from a single sweep that was then introgressed across these hybridizing species. However, it is also possible that such signatures are bioinformatic artefacts, as a single reference genome was used to call variants in the species within each of the *Helianthus* and *Eucalyptus* genera, which means that any assembly errors could yield anomalous LD signatures that might confound *OmegaPlus* in each of the species. Consequently, the same errors could systematically propagate across closely related species sharing the same reference and potentially drive some of the observed *PicMin* repeatability.

To test whether this was a source of error, we assessed whether phylogenetic distance correlated with the amount of overlap in the driving genes between pairs of species, which we calculated as the ratio of observed versus expected overlap, according to the expectation of a hypergeometric distribution, and found no significant correlation (*Pearson's r* = 0.05, *p-value* = 0.57; *Spearman's rho* = -0.05, *p-value* = 0.6) (SI Appendix, Fig. S6A). Overall, the observed overlap in driving genes between closely related species did not appear to deviate from the expectation differently than that between more distantly related species. Further assessment of the average phylogenetic distance between driving genes species showed that all the candidates, except one (OG0010509), are characterized by considerable mean phylogenetic distances between the species with driving genes (SI Appendix, Fig. S6B). This confirms that despite an apparent enrichment of low empirical p-values in *Helianthus* and *Eucalyptus* species, the overall repeatability signal is driven by genes spread throughout the analysed phylogeny.

To preclude spurious contributions from closely related species arising from bioinformatic artefacts, we conducted follow-up repeatability testing by running *PicMin* with only one species from each of these two genera and taking the union of significant hits (FDR < 0.5) across runs, evaluating all nine of the 13-species combinations [1 *Helianthus* + 1 *Eucalyptus* + all other species]. While this restricted

analysis cannot detect true signals of repeated sweeps within a genus, it does not suffer from the risk of being driven by bioinformatic errors propagating due to closely related species sharing the same reference genome. Despite reduced power due the overall lower number of species included, this follow-up analysis successfully retrieved 15 of the 33 originally identified candidate orthogroups (FDR < 0.1) (SI Appendix, Table S2).

We explored the functional characteristics of this highest confidence set of 15 orthogroups with repeated sweep signatures, using the annotations of *A. thaliana* (55) and *M. truncatula* (56) genes (SI Appendix, Table S2). These orthogroups included several genes with pivotal roles in biotic and abiotic stress responses. For instance, TPS14 (TERPENE SYNTHASE 14) is a key protein involved in terpene metabolism. Terpenes are metabolites involved in plant defence against both pathogens and herbivores, and are responsible for the attraction of beneficial organisms such as pollinators and seed dispersers (57, 58). Similarly, TGSL12 (CALLOSE SYNTHASE 3) is a protein that participates in callose metabolism, a polysaccharide synthesized in the cell wall of a variety of higher plants in response to pathogens infections and abiotic stress (59). The extensively studied cytochrome P450, belongs to a large family of proteins involved in NADPH- and O_2-dependent hydroxylation reactions across all domains of life and is closely linked to environmental stress response in plants (60). Notably ATERF019 (ERF019, ERF19), is a critical transcription factor involved in the response to a range of stressors, such as drought, osmotic, and oxidative stress, as well as bacterial and fungal infection (61, 62). Intriguingly, this gene has been found to negatively regulate plant resistance to *Phytophthora parasitica*, underscoring its significance in mitigating the impact of destructive plant pathogens causing significant crop losses worldwide (61). Another notable protein group identified was the Plant U-box type E3 ubiquitin ligase (PUB62, PUBs), which encompasses proteins with diverse functions in abiotic and biotic stress responses and is also associated with pollen self-incompatibility (63).

In addition to stress responses, several orthogroups with signatures of repeated sweeps play roles in plant growth and development. For instance, the clathrin adaptor complexes medium subunit family protein is involved in clathrin-mediated endocytosis (CME). This process regulates various aspects of plant development, such as hormone signaling, and has also been linked to response against environmental stresses (64). Interestingly, the machinery of CME has been shown to be evolutionarily conserved in plants (64). Furthermore, we identified the p300/CBP acetyltransferase-related protein, which is linked to cell differentiation, growth, and homeostasis. Notably, this protein has a 600 amino acid C-terminal segment which appears highly conserved across plants and animals, suggesting an essential role in multicellular organisms (65). Other key players in growth and development included ATSPLA2-ALPHA (PHOSPHOLIPASE A2-ALPHA), which is involved in various growth-related processes (66), and ROXY1/ROXY2 of the CC-type glutaredoxin (ROXY) family, which is a group of proteins crucial for organogenesis, particularly anther development (67). Lastly, among our top candidates was ATIPT4 (ARABIDOPSIS THALIANA ISOPENTENYLTRANSFERASE 4), a protein that plays a significant role in cytokinin biosynthesis. Cytokinins are hormones essential for regulating various aspects of plant growth and development (68).

Spatial scale of adaptation

Selective sweeps are typically interpreted as evidence of a mutation spreading throughout the range of an entire species (i.e. global adaptation), but they can also occur within a restricted portion of the range due to local adaptation. It is possible that some of the signatures of repeated sweeps we observed could be driven by strong local adaptation within a subsection of the species range, rather than global adaptation. To explore this possibility, we estimated F_{ST} for each SNP within each species and took the average across all SNPs within each gene. Driving genes in orthogroups with repeated sweep signatures displayed low to average F_{ST} within species, with few exceptions, suggesting that the signatures of repeated selective sweeps were driven by global, rather than local adaptation (SI Appendix, Fig. S7 and Fig. S8). Even though we observed few driving genes with high F_{ST} (8 genes out of 78 fell within the top 10% of F_{ST} values genome wide; SI Appendix, Fig. S8), it is worth noting that a mutation making a global selective sweep can generate transient genetic differentiation as F_{ST}

outliers, so even the relatively rare cases of high F_{ST} that we observed within species could still have been driven by global adaptation (69).

Genes with repeated sweeps have low pleiotropy

We explored the pleiotropic characteristics of orthogroups with repeated sweep signatures to assess the genetic architecture of global adaptation in comparison to the classical theoretical expectation based on Fisher's geometric model of universal pleiotropy (47,48), other empirical studies (32,33,34,35,36,37,38,39,70,71,72,73,74), as well as more recent findings on the architecture of local adaptation derived from a large comparative study across plant species sharing the same methods and majority of datasets, including the same SNPs sets (49).

To evaluate the amount of pleiotropy in genes with repeated selective sweeps, we utilized gene expression data for *A. thaliana* and *M. truncatula* genes sourced from Expression Atlas (75) and ATTED-II (76). We estimated two kinds of gene property related to pleiotropy: A) tissue specificity of gene expression (77) and B) statistics describing a gene's position/importance within a gene co-expression network (78). For tissue specificity (A), we employed the τ metric (79) of *A. thaliana* genes, which should be inversely proportional to pleiotropy; highly tissue-specific genes are likely to influence fewer processes compared to genes expressed across several tissues. Tissue specificity has been found to be positively associated with rates of molecular evolution (80). Using gene co-expression networks (B), we estimated pleiotropy based on four centrality measures (see *Methods* for details): degree, strength, closeness, and betweenness (81). Node degree signifies the number of nodes connected to a gene and thus the number of co-expressed genes; node strength represents the combined weights of these connections; closeness reflects a node's capacity to interact with all other nodes, even those not directly connected to it, and hence its co-expression capability across the entire network; betweenness indicates a node's ability to act as a bridging node in the network, linking co-expression subnetworks together. Each of these measures should be positively correlated to pleiotropy (82,83,84). Genes with high centrality measures in co-expression networks are hubs; hence, their expression affects many other genes and processes, potentially imposing an evolutionary constraint as changes in them can more likely be lethal or have strong effects. Conversely, genes with low centrality measures can be considered less pleiotropic, with changes in them affecting fewer processes (82,83,84).

To evaluate the amount of pleiotropy in the candidate orthogroups with repeated sweep signatures, each of the above measures was calculated for each orthogroup based on the metrics described above, using data from *A. thaliana* and *M. truncatula* (see *Methods* for details). We assessed pleiotropy in the candidate set including 33 Orthogroups (FDR < 0.1), as well as within a larger set of 107 orthogroups with a more relaxed false discovery rate cutoff (FDR < 0.2). While using a more relaxed FDR results in more false positives, it also allows a larger number of true positives, which improves the power to test patterns in pleiotropy for genes involved in global adaptation. Consistent with the Fisher-Orr model of evolution (47,48), both sets of results showed the same pattern: pleiotropy appeared significantly lower in the candidate orthogroups compared to random expectation (Fig. 4A, Fig. 4B). All measures used as proxies for pleiotropy, except *Medicago* betweenness, showed either a significant ($p < 0.05$) association with decreased pleiotropy or tended strongly towards this direction. It is noteworthy that another study focusing on local (rather than global) adaptation found the opposite relationship with increased pleiotropy for genes driving repeated local adaptation (49), using many of the same datasets and the same methods to identify repeated adaptation and pleiotropy.

We further assessed pleiotropy within the more conservative set of 15 orthogroups identified by intersecting the 33 candidate OGs identified in the main analysis (FDR < 0.1) with results derived from nine additional *PicMin* omitting closely related *Eucalyptus* and *Helianthus* species. The correlation between global adaptation and decreased pleiotropy largely persisted ($p < 0.05$) when the assessment was restricted to this small set, indicating a robust association between genes driving global adaptation and decreased pleiotropy (Fig. 4C).

Duplications are enriched in genes with repeated sweeps

Newly duplicated genes may experience selective sweeps due to positive selection if the resulting copies undergo sub- or neo-functionalization. Additionally, sub-functionalization reduces the pleiotropy of the original gene by distributing the function of the parent gene among new copies (85). This process may aid adaptation by easing the evolutionary constraints associated with a more pleiotropic ancestral gene (85,86). Considering this, we utilized the results obtained from *Orthofinder2* to investigate the involvement of gene duplications in recurrent global adaptation. We counted the number of paralogs within orthogroups as a representation of the number of duplication events and compared our candidates with repeated sweeps against a set of randomly chosen orthogroups, using the same statistical approach as the pleiotropy assessment. The 33 orthogroups with repeated sweeps (FDR < 0.1) exhibited a significant enrichment in gene duplications ($p < 0.05$), with a 3.3-fold-increase relative to the overall average for all orthogroups (Fig. 4D).

Next, we investigated whether the excess of duplicated genes observed in the recurrently globally adapted orthologous groups originated from small-scale single-copy gene duplications or from large-scale duplications due to polyploidization, as genes resulting from these two types of events evolve very differently (87,88). For this assessment, we retrieved lists of polyploidy-produced duplicated genes available for *Arabidopsis thaliana*, for which there is evidence of at least three ancient whole-genome duplication events (α, β and γ) (87,88,89).

Among the six driving genes identified in *A. thaliana* (*OmegaPlus* empirical p-values < 0.1 for the 33 *PicMin* orthogroups with FDR < 0.1, Fig. 3), two appear to have originated from whole-genome duplication events: ROXY1/ROXY2 (AT5G14070/OG0002218), which is dated back to the most recent duplication event α (86-14.5 mya), and a protein of unknown function (AT5G11000/OG0005857), which is dated to the β event (235-112 mya) (88,89). Extending this assessment to those orthogroups out of the 33 identified (*PicMin* FDR < 0.1) with a representative gene from *A. thaliana* (even though not necessarily an *A. thaliana* driving gene), we found that 9 genes in total (out of 22) appear to be polyploidy duplicated genes retained after one of the three ancient tetraploidy events (88,89). Between those with available functional annotation there were several other key developmental genes in addition to ROXY1/ROXY2, such as p300/CBP (OG0004755/ AT1G16705), ATIPT4 (OG0001035/AT4G24650), NPY (AT1G67900/OG0002632) and SLY1 (AT2G17980/OG0007322) as well as a few genes with key roles in the response to biotic stimuli, such as ERF19 (OG0006228/AT1G22810) and G-type lectin S-receptor-like serine/threonine-kinase (OG0002608/AT1G11340).

Possible biases and artefacts

PicMin identifies repeated orthogroups based on statistical support across different species or lineages (46). To test *PicMin's* robustness to unexpected biases, we performed 1000 additional *PicMin* analyses, each time randomly shuffling the *OmegaPlus* empirical p-values within each species. After correcting for multiple tests, the mean number of orthogroups with FDR < 0.1 per-permutation was 0.024. These 24 orthogroups with FDR < 0.1 arose within 9 out of the 1000 permutations (SI Appendix, Fig. S9). Generally, the *PicMin* analyses based on permuted data reported few orthogroups with false discovery rates below 0.9 compared to the actual *PicMin* analysis, and rarely any below 0.1 (SI Appendix, Fig. S9 and Fig. S10). The FDR distributions derived from *PicMin* based on the shuffled data appear clearly different from that of the actual data, with the latter exhibiting a markedly longer left tail (SI Appendix, Fig. S11).

As *PicMin* identifies orthogroups with consistently extreme signatures across multiple species, any source of bias arising from some characteristic of a gene can also bias *PicMin* if the characteristic tends to be conserved over deep time. Gene length is one such characteristic that might contribute to bias, based on how we ran *OmegaPlus* using one scan near the centre of each gene. This analysis choice might lead to reduced detection power in longer genes, as the signal of a selective sweep near a gene edge would be more attenuated at the middle of the gene due to the more extensive decay of

linkage disequilibrium, compared to shorter genes. Consequently, *OmegaPlus* might exhibit reduced sensitivity for larger genes. Our investigation revealed that this concern was unfounded, as there were generally low correlations between empirical p-values for the sweep signatures and gene length within each species (SI Appendix, Fig. S12). No consistent pattern was observed in the relationship between gene length and empirical p-value across species, as 10 species had $r > 0$, while 7 had $r < 0$; (SI Appendix, Fig. S13). Furthermore, when assessing the average gene length in the 33 orthogroups candidates for repeated global adaptation (FDR < 0.1), we found no significant difference in mean gene length compared to random expectations. If anything, our candidate orthogroups exhibit a tendency toward increased gene length (SI Appendix, Fig. S14), counter to the expectation favouring shorter genes under this type of bias.

To strengthen our confidence in the results, we next assessed the candidates derived from each of three additional approaches to implementing sweep scans using *OmegaPlus*, with variations in settings (refer to *Methods* for details). All three approaches showed high correlations between the resulting *OmegaPlus* empirical p-values and the empirical p-values obtained with the main approach implemented above (*OmegaPlus* with minimum window of 500 bp) (SI Appendix, Table S3). Remarkably, all approaches yielded consistent results in the association with pleiotropy, highlighting a robust relationship between global adaptation and low pleiotropy (SI Appendix, Fig. S15). While our analysis of pleiotropy assumes that estimates of tissue specificity from *A. thaliana* are representative of patterns in other species, specificity in expression has been shown to decline slowly among orthologues (90). In addition, our findings are robust to different measures of pleiotropy (Fig. 4).

Another potential source of bias is in the recombination landscape around a gene, as regions with lower recombination rates may harbour higher LD and lower diversity (91), which could lead to an enrichment of low *OmegaPlus* empirical p-values in these regions. It is also possible that sweep signatures may take longer to degrade in regions of low recombination. However, recombination events play a crucial role in generating the typical selective sweep linkage disequilibrium patterns recognized by *OmegaPlus*, therefore recombination must be substantial to obtain higher ω scores (13,15). This was largely reflected in our assessment based on available recombination data (see *Methods*), which showed higher density of lower *OmegaPlus* empirical p-values in regions with moderate to high recombination, with the exception of *H. argophyllus* (SI Appendix, Fig. S16). Overall, *OmegaPlus* demonstrated robustness to misidentifying regions with lower recombination rates as sweeps (SI Appendix, Fig. S16). In addition, we estimated average gene recombination rates de novo for each species based on observed patterns of linkage disequilibrium (see *Methods* for details). The average recombination rates of the driving genes (*OmegaPlus* emp-p < 0.1) of the 33 orthogroups with repeated signals of global selection identified with *PicMin* (FDR < 0.1) showed significant variability and no consistent pattern (SI Appendix, Fig. S17). If the recombination rates of the driving genes are randomly distributed relative to the background genomic rates, we would expect up to 13 driving genes in total to fall within either 5% tail of the distribution of genes recombination rates within species (95% confidence limit of a binomial distribution, based on 160 driving genes with available estimated recombination rate across all species). Our results do not significantly exceed this expectation, as we found 11 driving genes in the top 5% of the recombination rate distribution within species and 12 in the bottom 5% (SI Appendix, Fig. S18).

Finally, even though a previous study has shown that background selection (BGS) does not generate the LD signature typical of selective sweeps and does not tend to cause any deviation in *OmegaPlus* ω scores relative to neutrality (92), we assessed whether BGS could have driven our repeatability results. Using nucleotide diversity as a proxy for assessing BGS, we found that the driving genes of the 33 orthogroups identified with *PicMin* exhibit significant variability in nucleotide diversity, showing no consistent pattern, indicating the absence of this potential bias (SI Appendix, Fig. S19). If the nucleotide diversity of the driving genes is randomly distributed relative to the background gene diversity, we would expect up to 14 driving genes in total to fall within either 5% tail of the distribution of genes nucleotide diversity within species (95% confidence limit of a binomial distribution, based on 172 driving genes across all species). Our results do not significantly exceed this expectation, as we found eight driving genes in the top 5% of the recombination rate distribution within species and four in the bottom 5% (SI Appendix, Fig. S20).

Discussion

Here, we have uncovered evidence of 33 orthogroups (FDR < 0.1) with repeated global selective sweeps in multiple species (Fig. 3C). This discovery is noteworthy given the considerable evolutionary divergence among the analysed plant species and their diverse life history strategies and growth forms. These range from the herbaceous perennial monocot *P. hallii*, to eudicots encompassing both herbaceous annuals (*A. thaliana*, *C. rubella*, *M. truncatula*, *A. tuberculatus*, *Helianthus spp*.) and herbaceous perennials (*A. halleri*), to long-lived trees such as *Betula spp*., *Eucalyptus spp*., and *Populus spp*. (Fig. 1). While our study stands out for its focus on global adaptation and the exploration of more extensive phylogenetic distances, other studies have identified repeated signatures of local adaptation across independent populations and species residing in similar habitats or experiencing analogous selective pressures (32,33,34,35,36,37,38,39,40,41,49). Taken together, these studies suggest that while adaptation may commonly be polygenic and involve only subtle changes in allele frequency across many loci (93), it can also involve the contribution of key genes that appear to play particularly important roles, given their repeated involvement. In attempting to understand what determines the propensity of a gene to contribute to adaptation, it is important to consider both the biological characteristics of the gene and the influence of the methods used to detect selection.

Our set of genes with repeated sweeps includes many with demonstrated roles in abiotic and biotic stress responses in plants, as well as a few candidates with important functions related to growth and developmental processes (SI Appendix, Table S2). This echoes prior research based on rates of sequence substitution showing that protein function significantly influences these signatures of selection (94,95,96,97), as found in *Drosophila* (98,99), *Arabidopsis* (100), hominids (101), and other mammals (102,103). There is strong support for the hypothesis that host-pathogen interactions drive particularly rapid protein evolution, as immune and stress/defence response genes are often identified as targets of natural selection (94,95,96,97). Similarly, genes linked to defence responses, such as cytochrome P450 proteins (OG0001991), have been previously reported for their higher rates of sequence evolution (94,95). These findings are derived from methods based on high rates of nonsynonymous substitution that have very limited power to detect evolution that involves only a small number of impactful mutations and will therefore only tend to capture particularly strong and persistent selection pressure that drives many recurrent sweeps. By contrast, our sweep-based methods only detect more recent adaptation, as genomic signatures of selective sweeps degrade with time (14) but are able to detect the signatures arising from individual selective sweeps that may be missed by substitution-based methods. It is interesting that methods suited to detecting selection operating at very different timescales both tend to find genes involved in biotic interactions and defence as the targets of repeated natural selection.

The role of gene duplication in adaptation

Gene duplication is thought to provide a source of genetic flexibility and facilitate of adaptation, through the means of sub and neo functionalization (85,86). Indeed, candidates for convergent local adaptation to temperature in two distantly related conifers, lodgepole pine (*Pinus contorta*) and interior spruce (*Picea glauca*, *Picea engelmannii*, *Picea glauca x Picea engelmannii*), were enriched for duplicated genes (34), but this was not found in the recent study of repeated local adaptation across plants (49). In this study we found a strong association between the orthogroups with signatures of repeated global adaptation and increased duplication number, with a 3.3x-fold increase relative to the average, despite our method penalizing duplicated genes with a conservative paralogue correction per orthogroup (Fig. 4D). However, we note that pleiotropy and number of duplications tend to be inversely correlated, therefore it can be tricky to separate the causative effects (SI Appendix, Table S4). Gene duplication may facilitate global adaptation by decreasing pleiotropy of a parent gene via sub-functionalization (85). Alternatively, if sub/neo-functionalization result in a new selective landscape for the novel copy of a gene, repeated sweeps would be expected as it evolves to improve

this novel function. In either case, if the propensity for gene duplication is conserved over deep time, these mechanisms associated with duplication could partly explain our findings.

As polyploidy is a potential source of gene duplication, we also assessed whether the gene duplicates involved in repeated adaptation were derived from whole-genome or localized duplication. We found that a large portion of genes present in the repeated orthogroups (9 out of 22 genes in *A. thaliana*) may have originated during ancient polyploidization events, highlighting the important evolutionary role of whole-genome duplications in angiosperms. In *A. thaliana*, there is compelling evidence of at least three past tetraploidizations followed by subsequent re-diploidizations, with the oldest of these events dated before the moncot – dicot divergence, the second oldest dated before *A. thaliana* divergence from all the other dicots included in our study and the most recent event dated before the Brassicaceae radiation (89). During re-diploidizations some of the polyploidy-duplicated genes were retained and previous studies have shown that this process was highly biased towards certain gene functional classes, with some gene categories being expanded exclusively through these large-scale polyploidization–diploidization cycles (87,88). This is the case for several transcription factors and signal transducers involved in responses to biotic stimuli and secondary metabolisms, which are critical for survival (87,88), such as ERF19 (OG0006228/AT1G22810) and receptor-like protein kinases (RLKs) (OG0002608/AT1G11340), both identified here. Another category that has shown increased gene retention following whole-genome duplication events are developmental genes (87,88). Our analysis identified several *A. thaliana* genes belonging to this category present within the repeatedly globally selected orthogroups (*PicMin* FDR < 0.1), including ROXY1/ROXY2 (AT5G14070/OG0002218), p300/CBP (OG0004755/AT1G16705), ATIPT4 (OG0001035/AT4G24650), NPY (AT1G67900/OG0002632), and SLY1 (AT2G17980/OG0007322). Altogether, given that ancient whole genome duplications likely played a major role in the evolution of angiosperms by expanding certain gene functional classes to enhance growth and development as well as survival strategies against pathogens and other biotic threats (87,88,89), our results are consistent with an ongoing effect of this on the process of adaptation.

Contrasting effects of pleiotropy in global vs. local adaptation

One characteristic of a gene that appears particularly important for adaptation is pleiotropy, where a single genetic locus influences multiple phenotypic traits (104). Pleiotropy is thought to constrain adaptation due to the possible detrimental effects of a mutational change affecting multiple biological processes (42,105). Fisher's model of evolution provides a mathematical representation of this effect: the greater the number of trait dimensions, the lower the chance that a random mutation is beneficial, thus posing a reduced adaptive potential under higher pleiotropy (47). Similarly, if all genes have the same pleiotropy, larger mutations are more likely to incur negative fitness costs, due to the increased likelihood of overshooting the optimal trait value in some dimensions. Kimura (106) re-assessed Fisher's model by incorporating the probability of fixation and concluded that mutations of intermediate effect would be most likely to contribute to adaptation. Later, Orr (107,108), considered the distribution of mutations fixed over an adaptive walk and found this would be approximately exponential, with alleles of larger effect fixing only earlier in the process, and later stages of evolution dominated by mutations of smaller effect, a prediction which has found empirical support in stickleback (109). This family of models therefore predicts that the rate of adaptation will be slower for organisms with more traits due to the greater amount of pleiotropy, and that most mutations driving adaptation will be of small effect (48).

Consistent with this "cost of complexity" hypothesis, genes and mutations with low pleiotropy have often been found as the target of parallel genetic and phenotypic evolution (42,105). Similarly, our research has pinpointed genes exhibiting low pleiotropy as recurrent targets of global adaptation across multiple plant species (Fig. 4). Furthermore, this evolutionary model has also found support in previous empirical studies, such as in yeast, where mutations affecting more phenotypic traits showed higher fitness costs hence implying a negative relationship between pleiotropic and fitness effects of mutations (70). Similarly, a study on Norwegian graylings populations showed that gene pleiotropy constrains both plastic and adaptive gene expression responses and highlighted the importance of genes with low pleiotropy in evolution (71). Taken together, these findings align with Fisher's view of

evolution, suggesting that functional changes that favour one process are more likely to have deleterious trade-offs on others if pleiotropy is high.

Further theoretical work on the importance of pleiotropy was motivated by empirical observations of substantial modularity (i.e. most genes affect few traits) along with larger per-trait mutation effects in more pleiotropic genes (110,111). These observations were incorporated into models showing that organisms with intermediate levels of pleiotropy would have the fastest adaptive rates, mitigating the cost of complexity (110). These species-level models have subsequently been interpreted as predicting the greatest contribution to adaptation within a species by genes with intermediate pleiotropy (73), which is reasonable but worthy of further directed theoretical study. Consistent with this prediction, empirical studies in *A. thaliana* (73), stickleback (32), ragweed (72), and *Heliconius* butterflies (74) have found evidence of intermediate pleiotropy driving repeated adaptation.

Another factor that can affect the impact of pleiotropy on adaptation is the spatial pattern of natural selection. If natural selection favours different trait optima in different locations of a species range (i.e., local adaptation), then migration will tend to counteract divergence, resulting in a general advantage for alleles of larger effect, as they can withstand the homogenizing effect of migration (6). To the extent that pleiotropic effects of a mutation are aligned with the direction of divergence in phenotypic optima, which seems common for modular traits (110,111), we would therefore expect more pleiotropic genes to contribute to local adaptation more readily, due to their larger effects overcoming the homogenizing effect of migration (112). By contrast, when a species adapts to a similar phenotypic optimum across its range (i.e., global adaptation), there is no tension with migration and therefore no additional advantage for alleles of larger effect (6). In accordance with this prediction (6,112), we found a strong association between decreased pleiotropy and genes involved in global adaptation (Fig. 4), which is also in line with Fisher's and Orr's models of evolution and with evidence of reduced pleiotropy in rapidly evolving adaptive genes reported in other studies (70,80,82,83). In further agreement with migration-selection theory predictions (6,112), the opposite pattern was found for genes driving local adaptation (i.e. increased pleiotropy) by another study using the same bioinformatic pipeline and statistical tests, and conducted on many of the same species, but studying signatures of local, rather than global adaptation (49). Taken together, these two studies provide the first controlled comparison of the genetics of repeated global vs. local adaptation. Observed differences in the pleiotropy of genes driving adaptive responses are consistent with population genetic predictions from models of spatially varying vs. spatially uniform selection. This therefore suggests that local adaptation may often look quite different than global adaptation, underlining the importance of considering the spatial pattern of selection when studying how evolution works.

Materials & Methods

Dataset selection

We downloaded raw sequencing data of 17 angiosperm plants and forest trees whole-genome sequencing (WGS) datasets from SRA and ENA (Fig. 1A and SI Appendix, Table S1).

The choice of WGS as the sole sequencing method for inclusion was imposed by the need for high quality dense SNP data by the software used in this study, *OmegaPlus* (45). Additionally, the datasets were chosen based on the following specific criteria: they encompassed natural populations in their native habitats, comprised non-invasive and non-domesticated species, included a minimum of 20 unrelated individuals sampled from five or more locations, and featured a high-quality reference genome, or one of a closely related species. We estimated each species natural range based on 1000 random observations downloaded from the Global Biodiversity Information Facility (http://www.gbif.org) using the *R* packages *rgbif* and *gbif.range* (113).

Finally, we explored the phylogenetic relationship between reference genome species using *TimeTree* (https://timetree.org/). Phylogenetic distances between species were calculated based on the *TimeTree* phylogeny using the R package *ape* (function: cophenetic).

SNP calling

We applied a uniform SNP calling pipeline to all datasets for consistency. This pipeline was derived from a previous study and was selected as it optimizes the trade off between SNP quality and processing times (114). Raw *fastq* files were trimmed using *fastp* (115) with default settings. Clean reads were then aligned to reference genomes with *bwa-mem* (v0.7.17-r1188) (116), using 12 distinct reference genomes to map 17 datasets. If a species reference genome was not available, we used that of a closely related species. Three clusters of closely related species were mapped to the same reference genome (*B. pendula* and *B. platyphylla* mapped to *B. pendula*; *E. magnificata, E. sideroxylon* and *E. albens* to *E. grandis*; *H. annuus, H. petiolaris* and *H. argophyllus* to *H. annuus*). Following mapping, *samtools* was used to convert the alignment files from sequence alignment map (SAM) format to sorted, indexed binary alignment map (BAM) files, while discarding any alignment with mapping quality below 10 (-q 10) (117). The *MarkDuplicates* tool (118) from *Picard tools* was used to remove potential PCR duplicates and to set read groups. Indels were realigned using *GATK RealignerTargetCreator* followed by *GATK IndelRealigner* (119). After indel realignment, SNP calling was performed using *BCFTtools mpileup*, computing genotype likelihoods based on alignments with a minimum mapping quality of 5 (-q 5), followed by *BCFtools call* to identify single nucleotide polymorphisms (SNPs) from the pileup output and generate VCFs (117). Finally, we filtered raw VCF files with *VCFtools* (120) to retain only biallelic SNPs genotyped in at least 70% of the individuals, SNPs with quality value above 30 (--minQ 30), genotype quality above 20 (--minGQ 20) and minimum read depth above 5 (--minDP 5).

For downstream analyses, we retained SNP present at all minor allele frequencies except singletons. This retention of frequencies, often overlooked, was critical since the method employed to detect selective sweeps relies upon evaluating the patterns of linkage disequilibrium across genic regions (45). Therefore, filtering by allele frequency would introduce a bias by distorting LD patterns and could significantly decrease the detective power of the analysis, as excess of low frequency variants constitutes the main signature of a selective sweep (121). However, singletons are often bioinformatic artifacts and distinguishing them from real mutations can be challenging, hence their exclusion (122).

Finally, each dataset was filtered based on genomic relatedness (Fig. 1D). This step was taken to remove closely related individuals, as cryptic relatedness can potentially distort the estimation of regions under selection similarly to how it can confound GWAS (123). We used *Plink* (--genome function) to calculate relatedness (124), and the R package *PlinkQC* for filtering (125). *PlinkQC* aims to find the minimum number of individuals to remove to keep relatedness between any pairs below a chosen threshold (125). Individuals were systematically excluded from each species dataset to ensure that no relatedness scores between pairs exceeded 0.2, effectively removing any first and second-degree relatedness.

Orthology inference

To assign genes to orthogroups, we first retrieved the amino acid sequences (proteomes) for each of the 12 reference species (Table S1). For each gene, we selected a primary transcript according to the longest isoform using custom scripts. Sequences were sorted by amino acid length, and each protein was given a name corresponding to the genomic coordinates of its gene. A Perl script was used to scan the proteins FASTA files. Upon encountering a duplicate sequence (i.e., a sequence with the same header), the script retained the first occurrence (the longest) and discarded subsequent occurrences. The decision is made based on whether the header has been encountered before. If it's the first time encountering a particular header, the script writes both the header and sequence to the output file. Finally, *Orthofinder2* (51) was run with default settings using as input the 12 filtered proteomes, including only a single transcript per gene.

Detection of selective sweeps using OmegaPlus

We used *OmegaPlus* (45) to scan each species dataset for global selective sweeps. *OmegaPlus* searches for specific LD patterns characteristic of recent hard selective sweeps and outputs the ω-statistic (13). LD-based methods have been shown to outperform SFS-based methods in the search for hard selective sweeps resulting in higher true positive rate (TPR), and *OmegaPlus* (45) has been consistently reported as the most sensitive tool to detect potential hard selective sweeps (10,15, 120). For our analysis of repeated sweeps across multiple species using *PicMin*, it is preferrable to use the most sensitive method at the cost of a higher false positive rate, in order to detect the highest number of true positive sweeps. This is not a reason of concern in comparative studies using *PicMin*, as the same false positives are unlikely to arise from independent analyses in different species (46). We extracted genes from each species' VCF and added 1000 bp flanking regions on either side to include potential promoter and regulatory regions. Subsequently, *OmegaPlus* was run on each gene using a grid size of 3, resulting in 3 measurements: one at the first SNP, one at the last SNP and one equidistant between those 2 measurements. The minimum and maximum sizes of the subregion around a position that was included in the calculation of the ω-statistic were set to 500 and 100,000 base pairs, respectively. For each gene we retained the second scan, which was expected to lie approximately near the center of genes.

Following *OmegaPlus*, we ranked genes within each species by converting ω-scores to empirical p-values. An empirical p-value corresponds to the rank of a given gene's selection score relative to all other genes from that species, therefore it reflects the strength of evidence against a null hypothesis of no selection (46). Gene empirical p-values were further summarized within each orthogroup by taking the lowest empirical p-value (i.e. the strongest selection evidence for a sweep) and applying a Dunn-Sidak correction to account for the number of paralogues within the orthogroup. This method of correcting for multiple comparisons within orthogroups only represents the contribution of the member gene with the strongest sweep signature within a given species (*i.e.* if two genes experience sweeps within an orthogroup in a given species, only the gene with the stronger signal contributes to the test).

We excluded orthogroups with more than 10 paralogues within a species, as they would be heavily penalized by the Dunn-Sidak correction and suffer from low power. Furthermore, we removed orthogroups with genes from less than 7 species from the analysis, as *PicMin* sensitivity would be reduced under such conditions. Finally, after these exclusions, we re-ranked the empirical p-values within each species to ensure uniform distributions of orthogroups empirical p-values for *PicMin*.

Detection of selective sweeps using OmegaPlus: additional approaches

OmegaPlus was tested using the methodology outlined in the preceding section, with variations in the minimum window settings to explore robustness of our results. We experimented with both 200 and 1000 as minimum window sizes before ultimately opting for the intermediate value of 500. This was crucial, as research indicates that this setting can potentially significantly impact results and introduce increased stochasticity, particularly when employing very small minimum windows (121).

Additionally, we explored a genome-wide scan strategy, covering the entire genomes at intervals of 1000 bases. Subsequently, we extracted the scans falling within genes and aggregated these measurements on a per-gene basis by calculating the average for each gene, before applying the same methodology as before. However, we deemed this latter approach unsuitable due to its bias towards shorter genes, which tended to report more extreme average ω-scores.

Nevertheless, we evaluated the correlation between all different *OmegaPlus* runs per dataset and systematically examined the pleiotropy of the candidate orthogroups derived from each approach to enhance our confidence in the results.

Testing for repeated global sweeps - PicMin

We used *PicMin* (46) to test for repetitive selective sweeps in 13,268 orthogroups across 17 species. *PicMin* uses order statistics to perform hypothesis tests on a set of ranked values, in this case

empirical p-values derived from *OmegaPlus* ω scores, and identifies orthogroups enriched for large numbers of genes with low empirical p-values (46). Lower *OmegaPlus* empirical p-values correspond to higher ω scores, and indicate genes with stronger evidence of selection. For each orthogroup, *PicMin* provides a p-value representing the probability of generating ranks as extreme or more extreme than the observation under the null hypothesis of random genetic drift driving the *OmegaPlus* scores within each species. The method works as follows: for an orthogroup with genes in *n* species or lineages (17 in this case), under the null hypothesis the *n* empirical p-values representing the strength of evidence for selective sweeps in each species should follow a uniform distribution. If we order the empirical p-values within the orthogroup, the theory of order statistics shows they have marginal distributions that belong to the beta distribution, which *PicMin* uses to compute one-sided p-values. If the xth rank in the list of ordered empirical p-values is low relative to the beta distribution, this indicates that the *x* species with the lowest empirical p-values all have stronger signatures of selective sweeps than would be expected by chance. Because some genes will always have low a rank within one species, we ignore the $x = 1^{st}$ (lowest) ranked empirical p-value and consider all remaining higher ranks, effectively conducting tests of repeatability across two or more species. *PicMin* applies a multiple comparisons correction to the minimum p-value across all $x = \{2 \ldots n\}$ contrasts, based on the methods by Tippett (126), Dunn and Sidak (127,128). This results in a final p-value that reflects the evidence that a particular orthogroup exhibits repeated adaptation. Finally, a multiple testing correction to account for the number of tested orthogroups was applied to the final per orthogroup p-values according to the Benjamini & Hochberg (1995) (129) formula, implemented in the R function *p.adjust* (method = “fdr”).

Population structure assessment

For population structure assessment, SNPs with minor allele frequencies < 0.05 were discarded (*VCFtools*) (120) and each dataset was pruned by linkage disequilibrium ($r^2 > 0.4$) using the *indep-pairphase* function of *Plink* in windows of 50 and step of 5 (124). Population structure was explored using principal component analysis (PCA) with *Plink* and ancestry inference with *fastSTRUCTURE* (130), testing Ks from 1 to 10. The representative admixture model for each dataset was determined using *fastSTRUCTURE* built-in *chooseK* function, which selects the model that maximises the log-marginal likelihood lower bound (LLBO) of the data and best explain strong population structure (130).

Weir & Cockerham (1984) F_{ST} (131) was calculated between the populations identified with *fastSTRUCTURE* on a per SNP site basis with the *vcftools* function *weir-fst-pop* (120). Within each species, individuals were assigned to populations according to the best K model Q coefficient, using as threshold of inclusion Q > 0.9. Average F_{ST} per gene was calculated by taking the mean across F_{ST} values of SNPs within each gene.

Gene length

We assessed the association between gene length and *OmegaPlus* empirical p-values by calculating their Pearson correlation in each species, and examining the consistency of any patterning across species. We tested whether the mean gene length of the orthogroups with repeated sweep signatures identified with *PicMin* differed significantly from the expectation for randomly chosen genes. This was done by taking 10,000 random orthogroup draws of the same size as the significant set (33 orthogroups) and calculating the mean gene length of each draw. We then compared the mean of our candidate set against this null distribution.

Recombination landscape

We tested the correlation between gene recombination rate and evidence for selective sweeps derived from the *OmegaPlus* analysis. To achieve this, we downloaded recombination rate data for *Arabidopsis thaliana* (132) and *Helianthus annuus* (133), corresponding to four distinct datasets in our

analysis (*A. thaliana*, *H. annuus*, *H. petiolaris*, *H. argophyllus*). We then constructed density plots (in *R*) comparing *OmegaPlus* per-gene empirical p-values against the ranked average gene recombination rate.

Furthermore, we estimated recombination rates from patterns of linkage disequilibrium for each dataset using the R package *FastEPRR* (134). The final filtered VCFs used for the main analysis were further refined to include only SNPs with minor allele frequency above 0.1 located on the major scaffolds (excluding short/fragmented scaffolds to avoid inaccurate phasing), and to retain no more than one SNP every 2500 bp, with *VCFtools* (120). This filtering step ensured a manageable SNP set for subsequent computations. Each species' VCF was then split by chromosome and phased using *Beagle* (135). Recombination rates were estimated with *FastEPRR* (134) in non-overlapping 100,000 bp windows, and average per-gene recombination rates were calculated and then converted into empirical p-values in *R*. We used the R function *qbinom* to calculate the total number of driving genes expected to fall in either 5% tail of the distribution of genes' recombination rates across all species.

Nucleotide diversity

To determine if BGS could have influenced our repeatability results, we calculated genome-wide nucleotide diversity at each SNP site for each species using *VCFtools* (120). Next, we computed the average nucleotide diversity per gene for all genes in every species. We then compared the driving genes' average nucleotide diversity relative to the average gene nucleotide diversity distribution within each species. Finally, we converted per-gene nucleotide diversity estimates into empirical p-values in *R*, and used the *R* function *qbinom* to calculate the total number of driving genes expected to fall in either 5% tail of the distribution of genes' nucleotide diversity across all species.

Estimation of pleiotropy

We estimated the amount of pleiotropy for each gene using two characteristics based on gene expression data: A) specificity of expression across tissues (77) and B) centrality within co-expression networks (78). To measure tissue specificity (A), we obtained *Arabidopsis thaliana* tissue expression data from Expression Atlas, accession *E-MTAB-7978* (75). This dataset includes tissue expression (transcripts per million - TPM) across developmental stages, tissue types and sub-tissue types. We computed the mean TPM across all developmental stages and sub-tissue types within each tissue type, to result in the mean TPM for each of the 23 tissue types. The tissue specificity metric τ was determined following the method by Yanai et al. (2005) (79) as:

$$\tau = \frac{\sum_{i=1}^{N} (1 - x_i)}{N - 1}$$

where, for a given gene, x_i corresponds to the mean TPM for a given tissue type normalised by the maximum mean TPM across N tissue types.

We converted τ scores into per-gene empirical p-values, treating higher τ estimates (corresponding to higher tissue specificity) as higher empirical p-values. We then applied the same methodology used to summarize *OmegaPlus* per-gene empirical p-values into orthogroups values, retaining the minimum empirical p-value (corresponding to minimum τ and therefore maximum pleiotropy) per orthogroup and applying a Dunn-Šidák correction to correct for the number of paralogues. Finally, we transformed per-orthogroup empirical p-values into Z-scores with a mean of 0 and standard deviation of 1 across all orthogroups.

This approach avoids assuming that specificity/pleiotropy is maintained across paralogues, which should be more representative than taking the mean τ per-orthogroup. In fact, calculating the mean τ per orthogroup significantly decreased the prevalence of high τ values in the genome-wide distribution. This suggests that tissue specificity within orthogroups varies significantly among paralogues, possibly due to neofunctionalization or sub functionalization. Additionally, we explored the

same methodology in the reverse direction, considering lower τ estimates as lower empirical p-values. However, this alternative approach did not yield any substantial differences in the results.

To calculate the four centrality measures of co-expression networks (B), we built two networks using co-expression data from ATTED-II for *A. thaliana* and *M. truncatula*. The ATTED-II database summarises gene co-expression data derived from RNA-seq and microarray sources in a condition-independent manner, which is given as a standardised z-score between a given pair of genes (76). To construct co-expression networks, genes were treated as nodes and z-scores as edges, where positive z-scores denote positive co-expression and vice-versa for negative z-scores. Co-expression gene tables were downloaded for both species: *A. thaliana* = Ath-u.c3-0 and *M. truncatula* = Mtr-u.c3-0. We discarded all edges with $-5 < z < 2.33$ following the recommendations for significant negative/positive co-expression.

The *A. thaliana network* included 18,570 genes, while *M. truncatula* network included 17,786 genes. Networks were generated using the *igraph* package in R. Node betweenness and closeness were calculated using the *estimate_betweenness()* and *closeness()* functions respectively. Node degree and strength were calculated as the number of edges and the sum of all edge absolute z-scores respectively. We then condensed the resulting gene centrality measures into per-orthogroup Z-scores with mean 0 and standard deviation of 1 across all orthogroups, with the same approach used for tissue specificity τ scores and testing both directions for the initial conversion of centrality scores into empirical p-values.

To test whether genes with repeated sweep signatures had high/low values of pleiotropy, for both the tissue specificity (A) and centrality measures (B) we used a bootstrapping method to compare their values to those of randomly chosen genes. For each measure, we calculated the mean z-score of the candidate set for repeatability, including 33 OGs. We then performed 10,000 random draws, each comprising the same number of orthogroups as the candidate set and calculated the mean of each random draw. Finally, we assessed whether the mean of the candidate set fell within the 95% confidence interval of the means from the10,000 random draws. For the smaller sets of candidates, we assessed pleiotropy by calculating Stouffer's Z score (136) for tissue specificity and centrality measures using the following formula:

$$Z = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$$

Gene duplication

We used the output from *Orthofinder2* to count duplication events within orthogroups, including both terminal and nonterminal nodes. We then used this data to test whether the candidate set with repeated sweep signatures was significantly enriched for duplications, using the same bootstrapping method used for pleiotropy (described in the previous section).

Data Availability

The scripts for SNP calling are available at:
https://github.com/GabrieleNocchi/snp_calling_bcftools_slurm

The scripts for the population structure analysis are available at:
https://github.com/GabrieleNocchi/population_structure_analysis

The scripts for the main analyses are available at: https://github.com/GabrieleNocchi/RepSweeps

References and links to the genomic resources for each dataset are available in SI Appendix, Table S1.

Acknowledgements

Funding was provided by NSERC Discovery and Alberta Innovates, with computational resources and support provided by the Digital Research Alliance of Canada.

References

1. P.J. Flood, A.M. Hancock, The genomic basis of adaptation in plants. *Curr Opin Plant Biol* 36, 88-94 (2017).
2. J. Hereford, A quantitative survey of local adaptation and fitness trade-offs. *Am Nat* 173, 579-588 (2009).
3. E. Sanford, M.W. Kelly, Local adaptation in marine invertebrates. *Ann Rev Mar Sci* 3, 509-535 (2011).
4. O. Savolainen, M. Lascoux, J. Merilä, Ecological genomics of local adaptation. *Nat Rev Genet* 14, 807-820 (2013).
5. O. Savolainen, T. Pyhäjärvi, T. Knürr, Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics* 38, 595–619 (2007).
6. S. Yeaman, Evolution of polygenic traits under global vs local adaptation. *Genetics* 220, iyab134 (2022).
7. I. Höllinger, P.S. Pennings, J. Hermisson, Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS Genet* 15, e1008035 (2019).
8. J.M. Smith, J. Haigh, The hitch-hiking effect of a favourable gene. *Genet Res* 89, 391-403 (2007).
9. R. Nielsen et al., Genomic scans for selective sweeps using SNP data. *Genome Res* 15, 1566-1575 (2005).
10. P. Pavlidis, N. Alachiotis, A survey of methods and tools to detect recent and strong positive selection. *J Biol Res* (Thessalon) 24, 7 (2017).
11. J.M. Braverman, R.R. Hudson, N.L. Kaplan, C.H. Langley, W. Stephan, The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140, 783-796 (1995).
12. J.C. Fay, C.I. Wu, Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405-1413 (2000).
13. Y. Kim, R. Nielsen, Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167, 1513-1524 (2004).
14. H. Weigand, F. Leese, Detecting signatures of positive selection in non-model species using genomic data. *Zool. J. Linn. Soc.* 184, 528-583 (2018)
15. A. Koropoulis, N. Alachiotis, P. Pavlidis, Detecting Positive Selection in Populations Using Genetic Data. *Methods Mol Biol* 2090, 87-123 (2020).
16. W. Stephan, Selective Sweeps. *Genetics* 211, 5-13 (2019).
17. H.M.T. Vy, Y.J. Won, Y. Kim, Multiple Modes of Positive Selection Shaping the Patterns of Incomplete Selective Sweeps over African Populations of Drosophila melanogaster. *Mol Biol Evol* 34, 2792-2807 (2017).
18. S. Ihle, I. Ravaoarimanana, M. Thomas, D. Tautz, An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol Biol Evol* 23, 790-797 (2006).
19. C.D. Huber, M. DeGiorgio, I. Hellmann, R. Nielsen, Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol Ecol* 25, 142-156 (2016).
20. T.R. Booker, B.C. Jackson, P.D. Keightley, Detecting positive selection in the genome. *BMC Biol* 15, 98 (2017).
21. J.M. Akey et al., Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* 2, e286 (2004).
22. E.C. Andersen et al., Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat Genet* 44, 285-290 (2012).

23. S. Glinka, L. Ometto, S. Mousset, W. Stephan, D. De Lorenzo, Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. *Genetics* 165, 1269-1278 (2003).
24. Y. Kim, W. Stephan, Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765-777 (2002).
25. D.J. Orengo, M. Aguadé, Detecting the footprint of positive selection in a european population of Drosophila melanogaster: multilocus pattern of variation and distance to coding regions. *Genetics* 167, 1759-1766 (2004).
26. C.J. Rubin et al., Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587-591 (2010).
27. M.S. Wang et al., Positive selection rather than relaxation of functional constraint drives the evolution of vision during chicken domestication. *Cell Res* 26, 556-573 (2016).
28. Y. Yuan et al., Selective sweep with significant positive selection serves as the driving force for the differentiation of japonica and indica rice cultivars. *BMC Genomics* 18, 307 (2017).
29. Z. Zhang et al., Whole-genome resequencing reveals signatures of selection and timing of duck domestication. *Gigascience* 7, giy027 (2018).
30. J.Y. Choi et al., Natural variation in plant telomere length is associated with flowering time. *Plant Cell* 33, 1118-1134 (2021).
31. K. Wei, G.A. Silva-Arias, A. Tellier, Selective sweeps linked to the colonization of novel habitats and climatic changes in a wild tomato species. *New Phytol* 237, 1908-1921 (2023).
32. D.J. Rennison, C.L. Peichel, Pleiotropy facilitates parallel adaptation in sticklebacks. *Mol Ecol* 31, 1476-1486 (2022).
33. L.R. Moreira, B.T. Smith, Convergent genomic signatures of local adaptation across a continental-scale environmental gradient. *Sci Adv* 9, eadd0560 (2023).
34. S. Yeaman et al., Convergent local adaptation to climate in distantly related conifers. *Science* 353, 1431-1433 (2016).
35. S. Soudi et al., Repeatability of adaptation in sunflowers: genomic regions harbouring inversions also drive adaptation in species lacking an inversion. *eLife*, 12:RP88604 (2023).
36. L. Wang et al., Molecular Parallelism Underlies Convergent Highland Adaptation of Maize Landraces. *Mol Biol Evol* 38, 3567-3580 (2021).
37. H.A. Poore et al., Repeated genetic divergence plays a minor role in repeated phenotypic divergence of lake-stream stickleback. *Evolution* 77, 110-122 (2023).
38. I.S. Magalhaes et al., Intercontinental genomic parallelism in multiple three-spined stickleback adaptive radiations. *Nat Ecol Evol* 5, 251-261 (2021).
39. M. Bohutínská et al., Genomic basis of parallel adaptation varies with divergence in Arabidopsis and its relatives. *Proc Natl Acad Sci U.S.A* 118, e2022713118 (2021).
40. G. Montejo-Kovacevich et al., Repeated genetic adaptation to altitude in two tropical butterflies. *Nat Commun* 13, 4676 (2022).
41. M. Bohutinska, C. L. Peichel, Divergence time shapes gene reuse during repeated adaptation. *Trends Ecol. Evol.* (2023).
42. A. Martin, V. Orgogozo, The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* 67, 1235-1250 (2013).
43. H.A. Orr, J.A. Coyne, The genetics of adaptation: a reassessment. *Am Nat* 140, 725-742 (1992).
44. 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).
45. N. Alachiotis, A. Stamatakis, P. Pavlidis, OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* 28, 2274-2275 (2012).
46. T.R. Booker, S. Yeaman, M.C. Whitlock, Using genome scans to identify genes used repeatedly for adaptation. *Evolution* 77, 801-811 (2023).
47. R.A. Fisher, The Genetical Theory of Natural Selection (Oxford, Clarendon Press, 1930).
48. H.A. Orr, Adaptation and the cost of complexity. *Evolution* 54, 13-20 (2000).

49. J. Whiting et al., Core genes driving climate adaptation in plants. *Nature Ecology and Evolution* (2024; *in press*).
50. G. Nocchi et al., Genomic signals of local adaptation and hybridization in Asian white birch. *Mol Ecol* 32, 595-612 (2023).
51. D.M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20, 238 (2019).
52. B. Wang et al., Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in Boechera stricta. *Genome Biol* 20, 126 (2019).
53. S. Kumar et al., TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* 39, msac174 (2022).
54. D.L. Stern, The genetic causes of convergent evolution. *Nat Rev Genet* 14, 751-764 (2013).
55. T. Z. Berardini et al., The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* 53, 474-485 (2015).
56. F. Cunningham et al., Ensembl 2022. *Nucleic Acids Res* 50, D988-D995 (2022).
57. D. Tholl, S. Lee, Terpene Specialized Metabolism in Arabidopsis thaliana. *Arabidopsis Book* 9, e0143 (2011).
58. F. Zhou, E. Pichersky, The complete functional characterisation of the terpene synthase family in tomato. *New Phytol* 226, 1341-1360 (2020).
59. N. Li et al., The multifarious role of callose and callose synthase in plant development and environment interactions. *Front Plant Sci* 14, 1183402 (2023).
60. B. A. Pandian, R. Sathishraj, M. Djanaguiraman, P. V. V. Prasad, M. Jugulam, Role of Cytochrome P450 Enzymes in Plant Stress Response. *Antioxidants* (Basel) 9, 454 (2020).
61. W. Lu et al., The Arabidopsis thaliana gene AtERF019 negatively regulates plant resistance to Phytophthora parasitica by suppressing PAMP-triggered immunity. *Mol Plant Pathol* 21, 1179-1193 (2020).
62. P. Y. Huang et al., NINJA-associated ERF19 negatively regulates Arabidopsis pattern-triggered immunity. *J Exp Bot* 70, 1033-1047 (2019).
63. M. Trujillo, News from the PUB: plant U-box type E3 ubiquitin ligases. *J Exp Bot* 69, 371-384 (2018).
64. S. Di Rubbo et al., The clathrin adaptor complex AP-2 mediates endocytosis of brassinosteroid insensitive1 in Arabidopsis. *Plant Cell* 25, 2986-2997 (2013).
65. L. Bordoli, M. Netsch, U. Lüthi, W. Lutz, R. Eckner, Plant orthologs of p300/CBP: conservation of a core domain in metazoan p300/CBP acetyltransferase-related proteins. *Nucleic Acids Res* 29, 589-597 (2001).
66. H.J. Kim et al., Endoplasmic reticulum- and Golgi-localized phospholipase A2 plays critical roles in Arabidopsis pollen development and germination. *Plant Cell* 23, 94-110 (2011).
67. S. Xing, S. Zachgo, ROXY1 and ROXY2, two Arabidopsis glutaredoxin genes, are required for anther development. *Plant J* 53, 790-801 (2008).
68. J. Sun et al., The Arabidopsis AtIPT8/PGA22 gene encodes an isopentenyl transferase that is involved in de novo cytokinin biosynthesis. *Plant Physiol* 131, 167-176 (2003).
69. T.R. Booker, S. Yeaman, M. C. Whitlock, Global adaptation complicates the interpretation of genome scans for local adaptation. *Evol Lett* 5, 4-15 (2021).
70. T.F. Cooper, E. A. Ostrowski, M. Travisano, A negative relationship between mutation pleiotropy and fitness effect in yeast. *Evolution* 61, 1495-1499 (2007).
71. S. Papakostas et al., Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions. *Nat. Commun* 5, 4071 (2014).
72. T. Hämälä, A. J. Gorton, D. A. Moeller, P. Tiffin, Pleiotropy facilitates local adaptation to distant optima in common ragweed (Ambrosia artemisiifolia). *PLoS Genet* 16, e1008707 (2020).
73. L. Frachon et al., Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nat Ecol Evol* 1, 1551-1561 (2017).

74. J.J. Lewis et al., Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. *Proc Natl Acad Sci U.S.A.* 116, 24174-24183 (2019).
75. I. Papatheodorou et al., Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* 46, D246-D251 (2018).
76. T. Obayashi, H. Hibara, Y. Kagaya, Y. Aoki, K. Kinoshita, ATTED-II v11: A Plant Gene Coexpression Database Using a Sample Balancing Technique by Subagging of Principal Components. *Plant Cell Physiol* 63, 869-881 (2022).
77. K. Watanabe et al., A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* 51, 1339-1348 (2019).
78. S.R. Proulx, D.E. Promislow, P.C. Phillips, Network thinking in ecology and evolution. *Trends Ecol Evol* 20, 345-353 (2005).
79. I. Yanai et al., Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification*. Bioinformatics* 21, 650-659 (2005).
80. K.L. Mack, M. Phifer-Rixey, B. Harr, M. W. Nachman, Gene Expression Networks Across Multiple Tissues Are Associated with Rates of Molecular Evolution in Wild House Mice. *Genes* (Basel) 10, 225 (2019).
81. F.J. Azuaje, Selecting biologically informative genes in co-expression networks with a centrality score. *Biol Direct* 9, 12 (2014).
82. M. W. Hahn, A.D. Kern, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 22, 803-806 (2005).
83. N. Mähler et al., Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genet* 13, e1006402 (2017).
84. K.C. Wollenberg Valero, Aligning functional network constraint to evolutionary outcomes. *BMC Evol Biol* 20, 58 (2020).
85. J. A. Birchler, H. Yang, The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* 34, 2466-2474 (2022).
86. S. Ohno, Evolution by Gene Duplication (Springer-Verlag, 1970).
87. S. Maere et al., Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* 102, 5454-5459 (2005).
88. B.C. Thomas, B. Pedersen, M. Freeling, Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* 16, 934-946 (2006).
89. J.E. Bowers, B.A. Chapman, J. Rong, A.H. Paterson, Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433-438 (2003).
90. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Comput Biol* 12, e1005274 (2016).
91. W. Stephan, Genetic hitchhiking versus background selection: the controversy and its implications. *Philos Trans R Soc Lond B Biol Sci* 365, 1245-1253 (2010).
92. D.R. Schrider, Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps. *Genetics* 216, 499-519 (2020).
93. N. Barghi, J. Hermisson, C. Schlötterer, Author Correction: Polygenic adaptation: a unifying framework to understand positive selection. *Nat Rev Genet* 21, 782 (2020).
94. A.F. Moutinho, F.F. Trancoso, J.Y. Dutheil, The Impact of Protein Architecture on Adaptive Evolution. *Mol Biol Evol* 36, 2013-2028 (2019).
95. A. F. Moutinho, T. Bataillon, J.Y. Dutheil, Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology* 34, 315–338 (2019).
96. A.F. Moutinho, A. Eyre-Walker, J.Y. Dutheil, Strong evidence for the adaptive walk model of gene evolution in Drosophila and Arabidopsis. *PLoS Biol* 20, e3001775 (2022).
97. T.B. Sackton, Studying Natural Selection in the Era of Ubiquitous Genomes. *Trends Genet* 36, 792-803 (2020).
98. T.B. Sackton et al., Dynamic evolution of the innate immune system in Drosophila. *Nat Genet* 39, 1461-1468 (2007).

99. D.J. Obbard, J. J. Welch, K. W. Kim, F. M. Jiggins, Quantifying adaptive evolution in the Drosophila immune system. *PLoS Genet* 5, e1000698 (2009).
100. T. Slotte et al., Genomic determinants of protein evolution and polymorphism in Arabidopsis. *Genome Biol Evol* 3, 1210-1219 (2011).
101. R. Nielsen et al., A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3, e170 (2005).
102. C. Kosiol et al., Patterns of positive selection in six Mammalian genomes. *PLoS Genet* 4, e1000144 (2008).
103. F.J. Alberto et al., Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun* 9, 813 (2018).
104. F.W. Stearns, One hundred years of pleiotropy: a retrospective. *Genetics* 186, 767-773 (2010).
105. D.L. Stern, V. Orgogozo, The loci of evolution: how predictable is genetic evolution. *Evolution* 62, 2155-2177 (2008).
106. Kimura M., The neutral theory of molecular evolution. Cambridge: Cambridge University Press, (1983).
107. H.A. Orr, The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* 52, 935-949 (1998).
108. H.A. Orr, The evolutionary genetics of adaptation: a simulation study. *Genet Res* 74, 207-214 (1999).
109. S.M. Rogers et al., Genetic signature of adaptive peak shift in threespine stickleback. *Evolution* 66, 2439-2450 (2012).
110. G.P. Wagner, J. Zhang, The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet* 12, 204-213 (2011).
111. Z. Wang, B.Y. Liao, J. Zhang, Genomic patterns of pleiotropy and the evolution of complexity. *Proc Natl Acad Sci U.S.A.* 107, 18034-18039 (2010).
112. P. Battlay, S. Yeaman, K.A. Hodgins, Impacts of pleiotropy and migration on repeated genetic adaptation. *Genetics* (2024; *in press*).
113. Y. Chauvier et al., gbif.range - An R package to generate species range maps based on ecoregions and a user-friendly GBIF wrapper. *EnviDat* (2022).
114. R.J. Jasper et al., Evaluating the accuracy of variant calling methods using the frequency of parent-offspring genotype mismatch. *Mol Ecol Resour* 22, 2524-2533 (2022).
115. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884-i890 (2018).
116. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760 (2009).
117. P. Danecek et al., Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008 (2021).
118. "Picard Toolkit." Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute (2018).
119. M.A. DePristo et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-498 (2011).
120. P. Danecek et al., The variant call format and VCFtools. *Bioinformatics* 27, 2156-2158 (2011).
121. N. Alachiotis, P. Pavlidis, Scalable linkage-disequilibrium-based selective sweep detection: a performance guide. *Gigascience* 5, 7 (2016).
122. R.C. Edgar, H. Flyvbjerg, Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 31, 3476-3482 (2015).
123. E. Uffelmann et al., Genome-wide association studies. *Nat. Rev. Methods Primers* 1, 59 (2021).
124. C.C. Chang et al., Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).
125. H.V. Meyer, plinkQC: Genotype quality control in genetic association studies. Zenodo. https://doi.org/10.5281/zenodo.3934294 (2020).
126. L. Tippett, The methods of statistics: An introduction mainly for workers in the biological sciences. (Williams & Norgate Ltd., 1931).

127. O.J. Dunn, Estimation of the means of dependent variables. *Ann. Math. Stat.* 29, 1095–1111 (1958).
128. Z. Sidak, Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.* 62, 626 (1967).
129. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat.* 57, 289–300 (1995).
130. A. Raj, M. Stephens, J. K. Pritchard, fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573-589 (2014).
131. B. S. Weir, C. C. Cockerham, Estimating f-statistics for the analysis of population structure. *Evolution* 38, 1358-1370 (1984).
132. B. A. Rowan et al., An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics* 213, 771-787 (2019).
133. K. Huang et al., Mutation Load in Sunflower Inversions Is Negatively Correlated with Inversion Heterozygosity. *Mol Biol Evol* 39, msac101 (2022).
134. F. Gao, C. Ming, W. Hu, H. Li, New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. *G3 (Bethesda)* 6, 1563-1571 (2016).
135. B.L. Browning, X. Tian, Y. Zhou, S.R. Browning, Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* 108, 1880-1890 (2021).
136. S. Stouffer, L. DeVinney, E. Suchmen, The American Soldier: Adjustment During Army Life, Vol. 1. (Princeton University Press, Princeton, NJ, 1949).

Figures and Tables

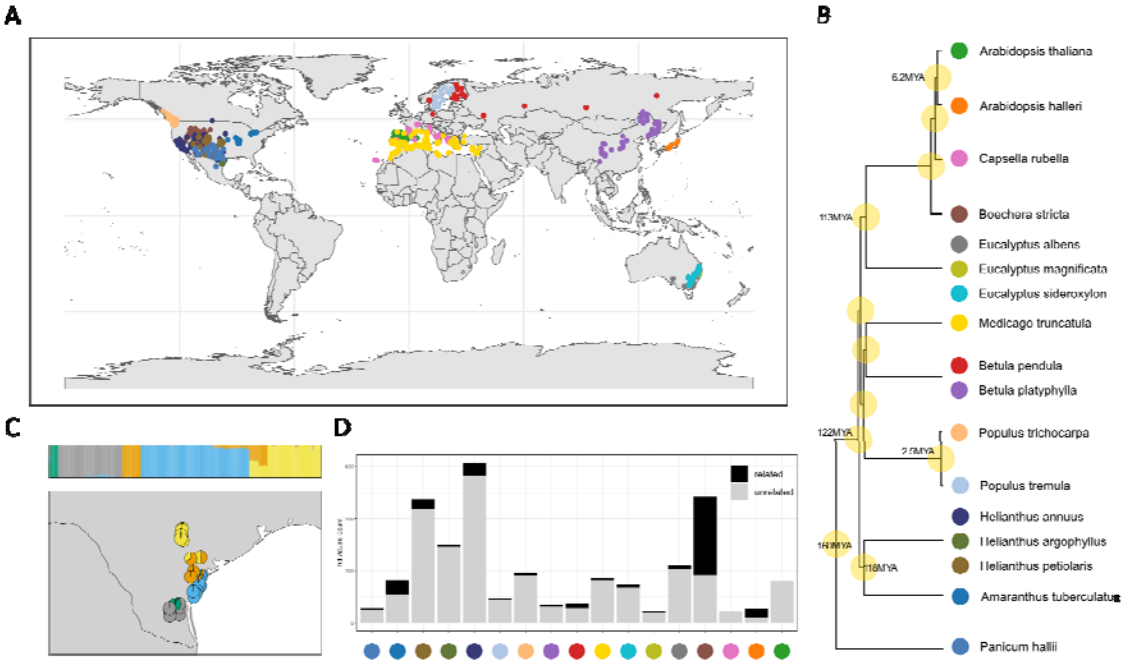


Fig. 1. Geographical distribution and relatedness within and among the study species. (A) Sampling locations of the 17 datasets included in the study. (B) Time-calibrated phylogenetic tree (retrieved from https://timetree.org/) of the 17 datasets, based on 12 reference species. (C) *fastSTRUCTURE* ancestry pie plot (K = 5) of the *Helianthus argophyllus* dataset in Texas (USA), showing substantial sub structuring. (D) Relatedness filtering summary bar plot by dataset. Datasets labelled by corresponding colour from B.

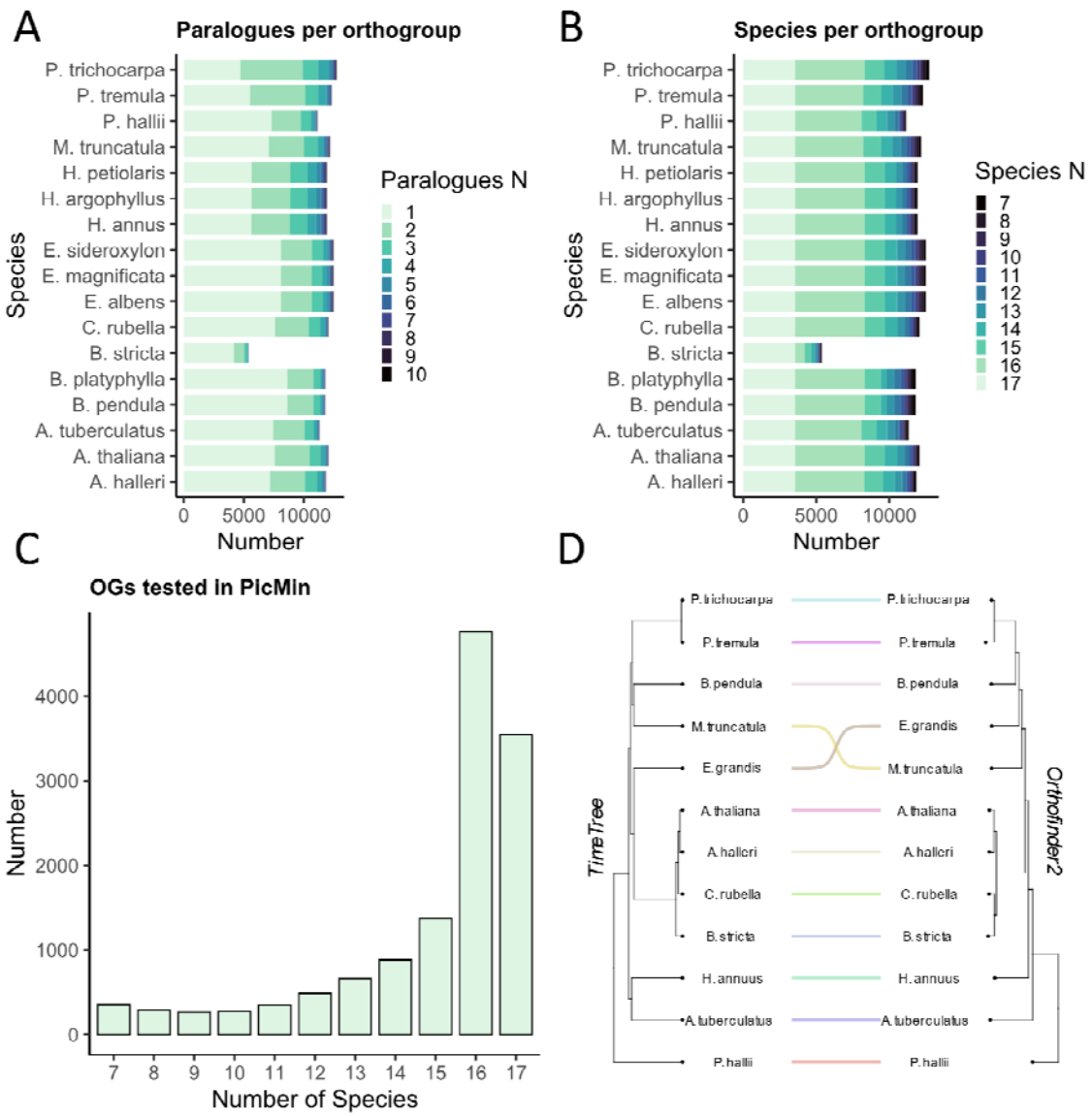


Fig. 2. Orthology assignment summary of the final set of 13,268 orthogroups. (A) Bar plot of the number of paralogues per orthogroup for each species. (B) Bar plot of the number of species included in each orthogroup, for the orthogroups of each species. (C) Distribution of the 13,268 tested orthogroups across species number. (D) Comparison between the *TimeTree* and *Orthofinder2* phylogenies, each based on 12 reference species.

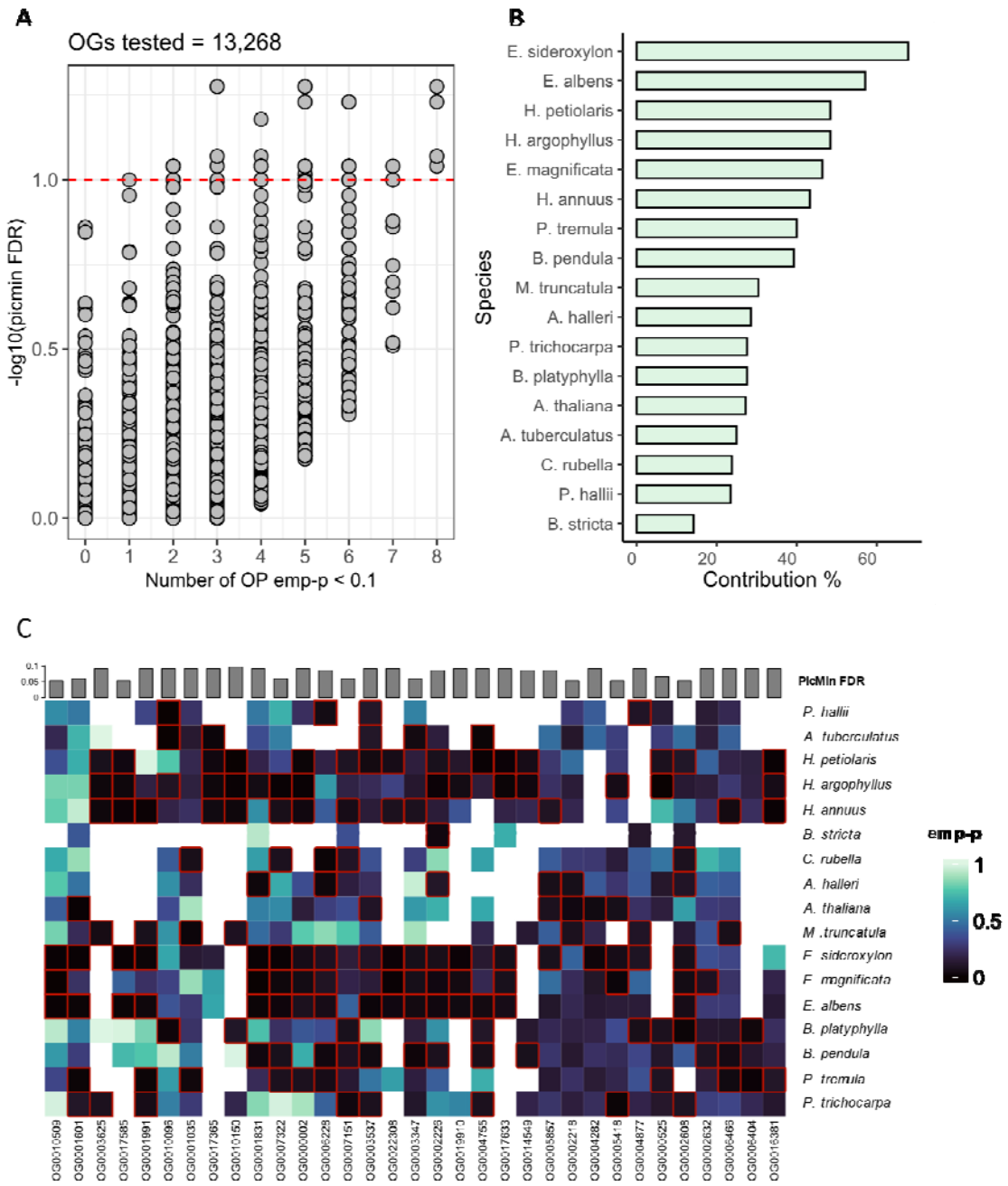


Fig. 3. Evidence for repeated selective sweeps across multiple species. (A) Distribution of *PicMin* FDR (-log10 *PicMin* FDR on the Y axis) for the 13,268 tested orthogroups, ordered by number of putative driving genes on the X axis (number of *OmegaPlus* emp-p < 0.1). Points above the red line have FDR < 0.1. (B) Species contribution to *PicMin* top candidates (FDR < 0.1), calculated for each species as: [number of empirical p-values < 0.1 in the *PicMin* FDR < 0.1 orthogroups]/[total number of orthogroups with *PicMin* FDR < 0.1 tested]. (C) Heatmap of *OmegaPlus* empirical p-values for the 33 candidate orthogroups (*PicMin* FDR < 0.1); driving genes cells (emp-p < 0.1 – black cells) are outlined in red, species are ordered by phylogenetic distance along the Y axis. A white cell indicates that an orthogroup was not tested in a species.

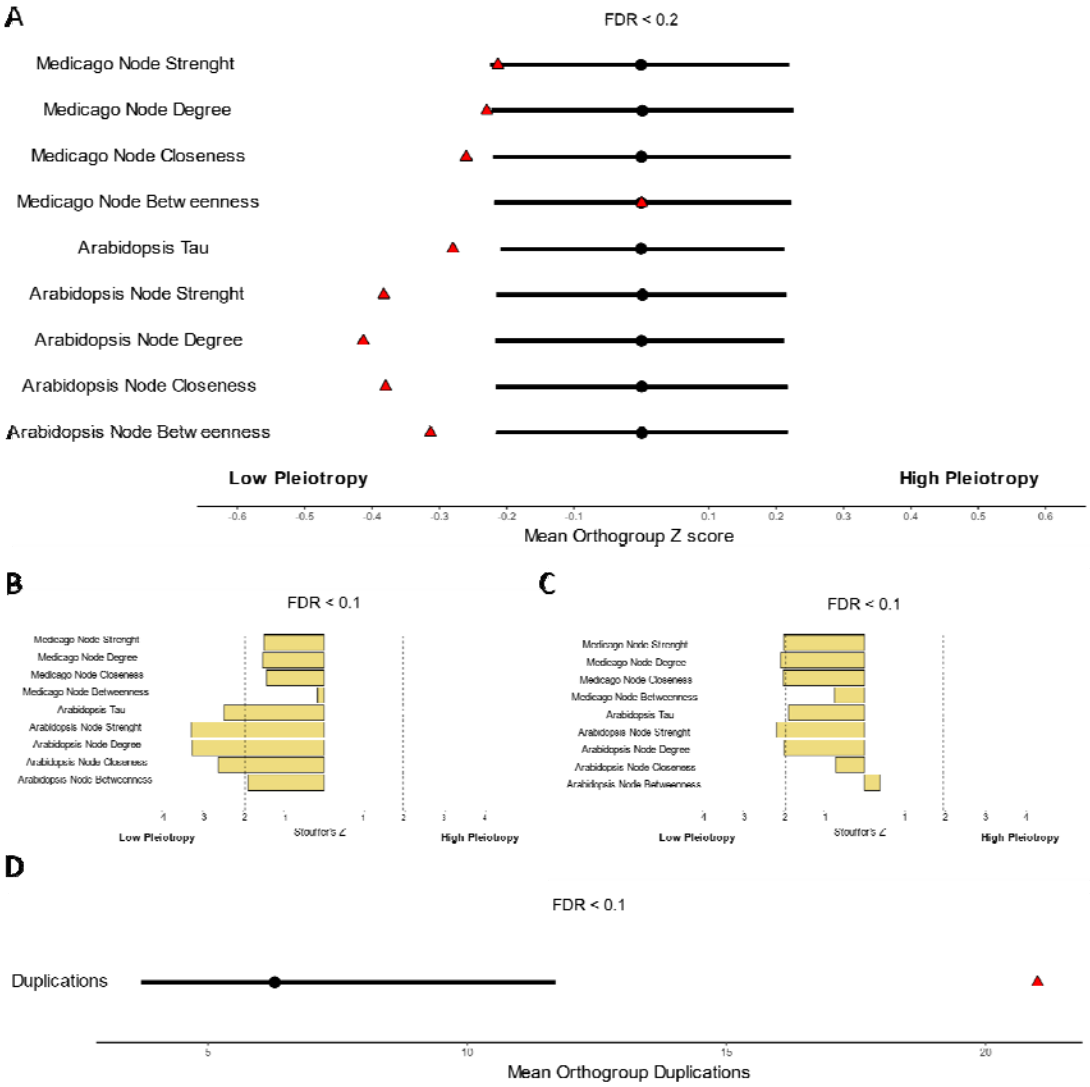
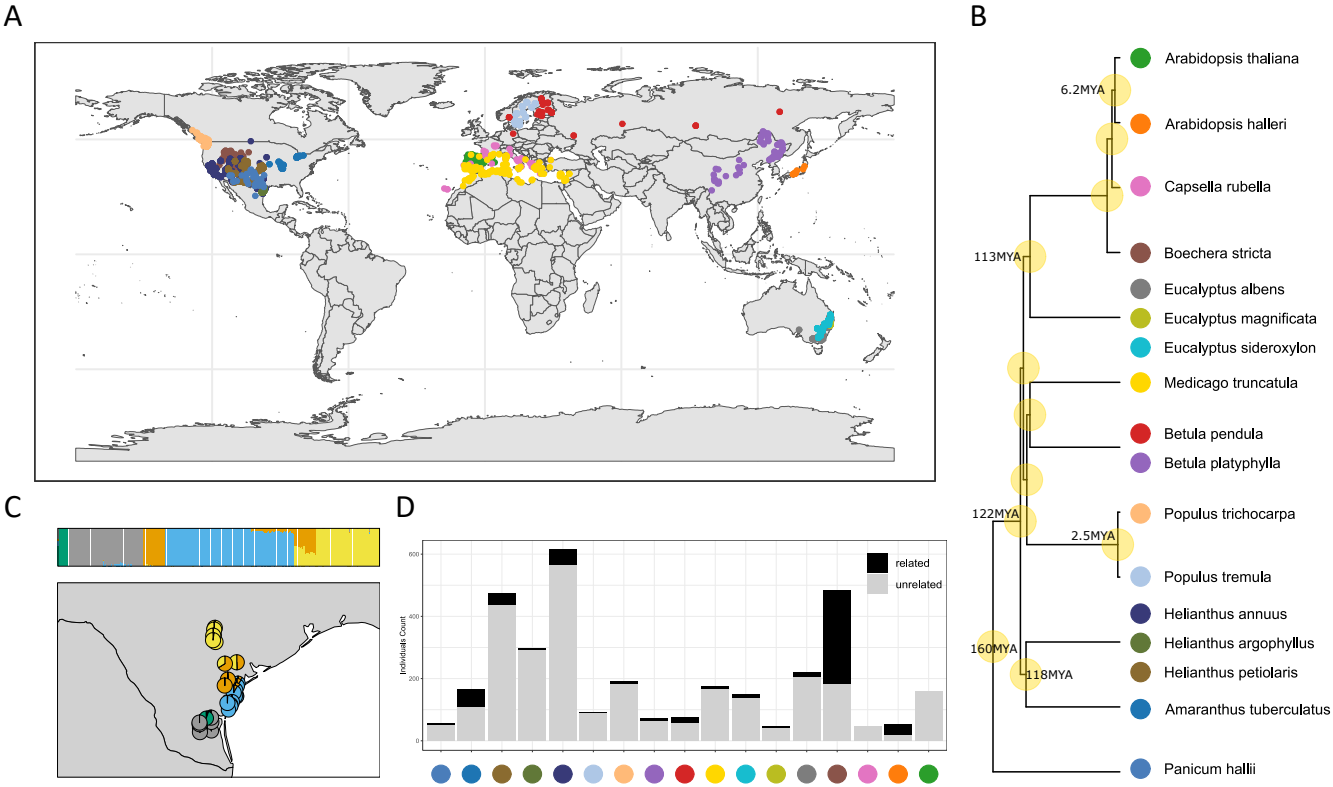
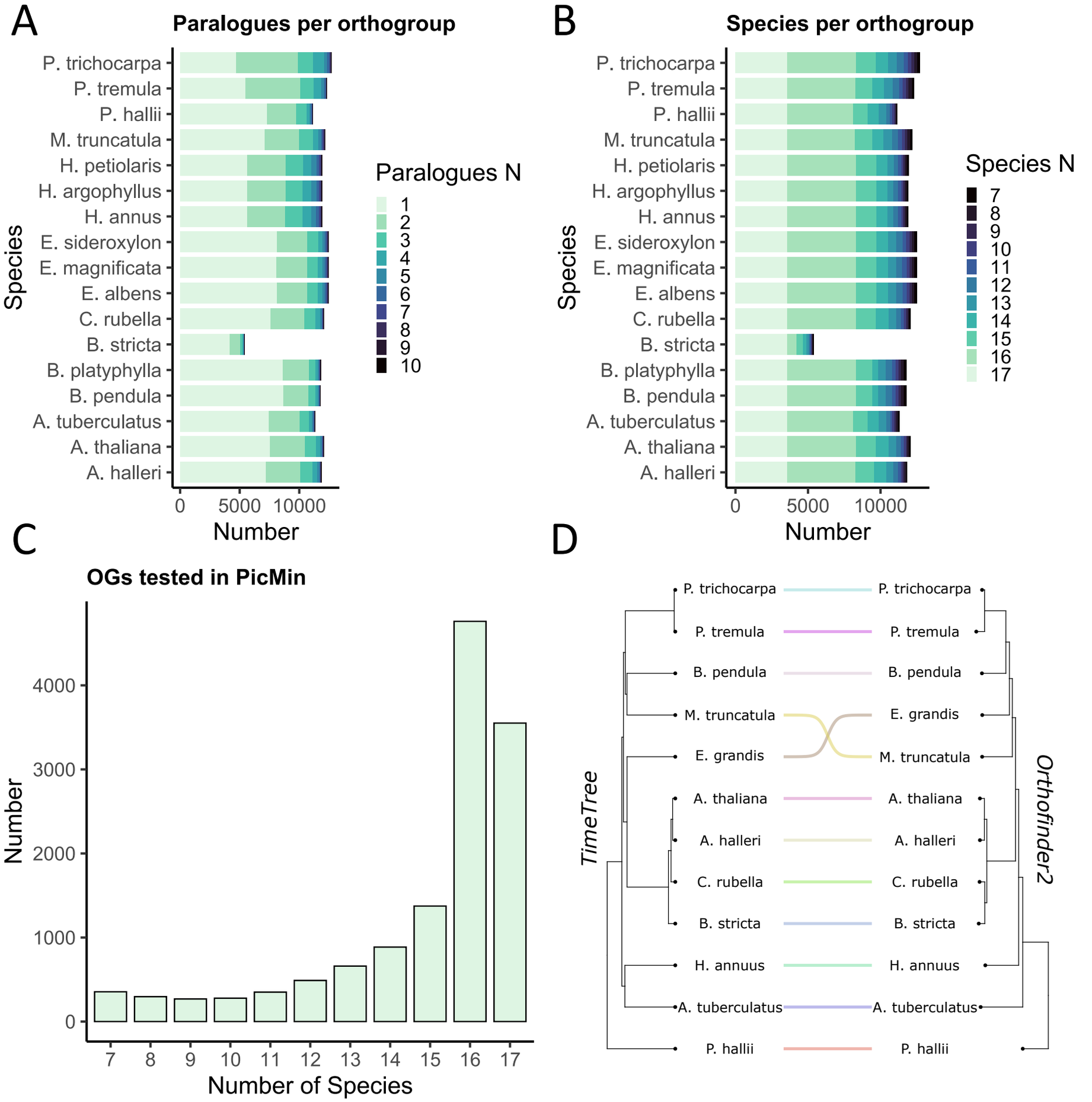
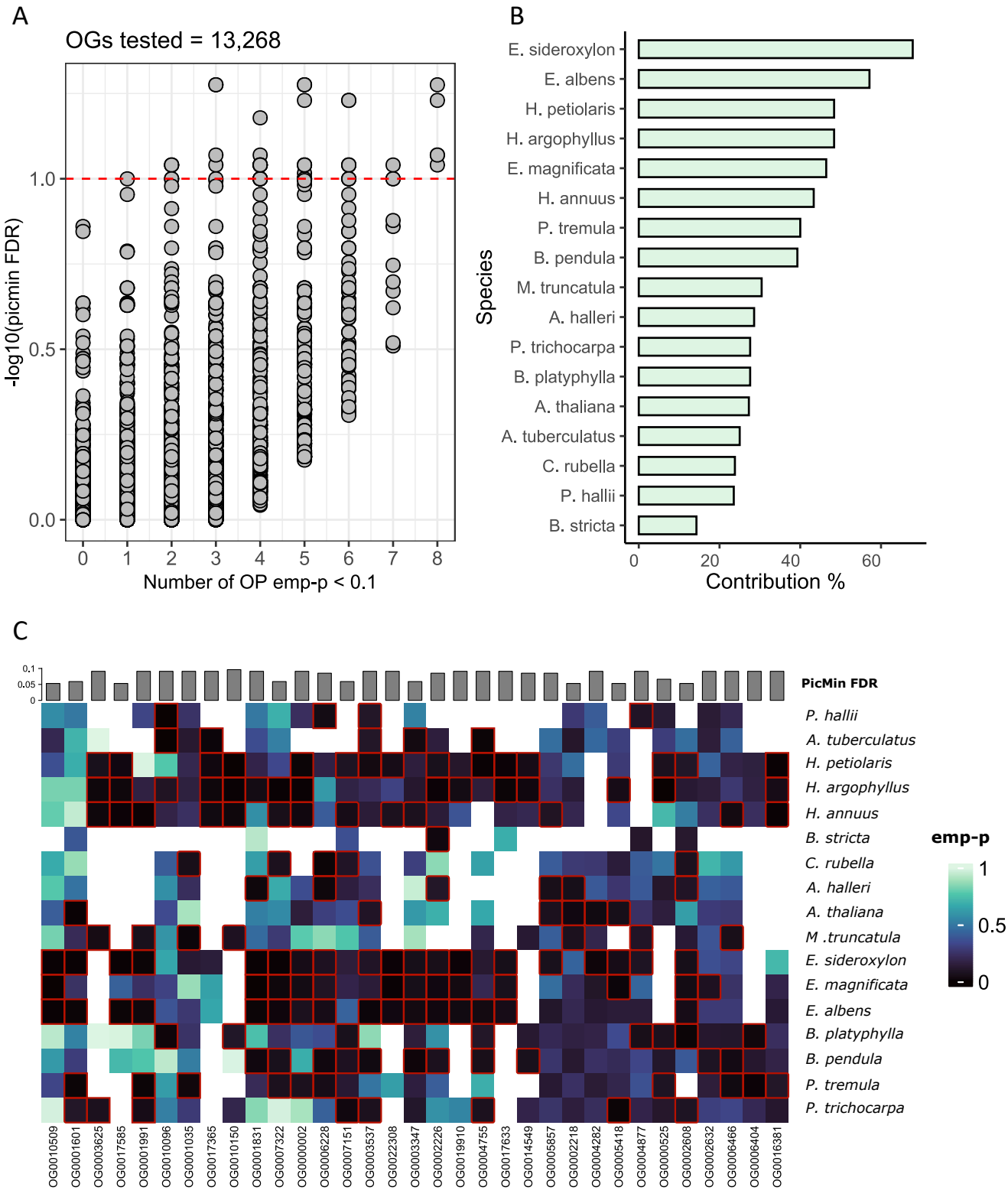
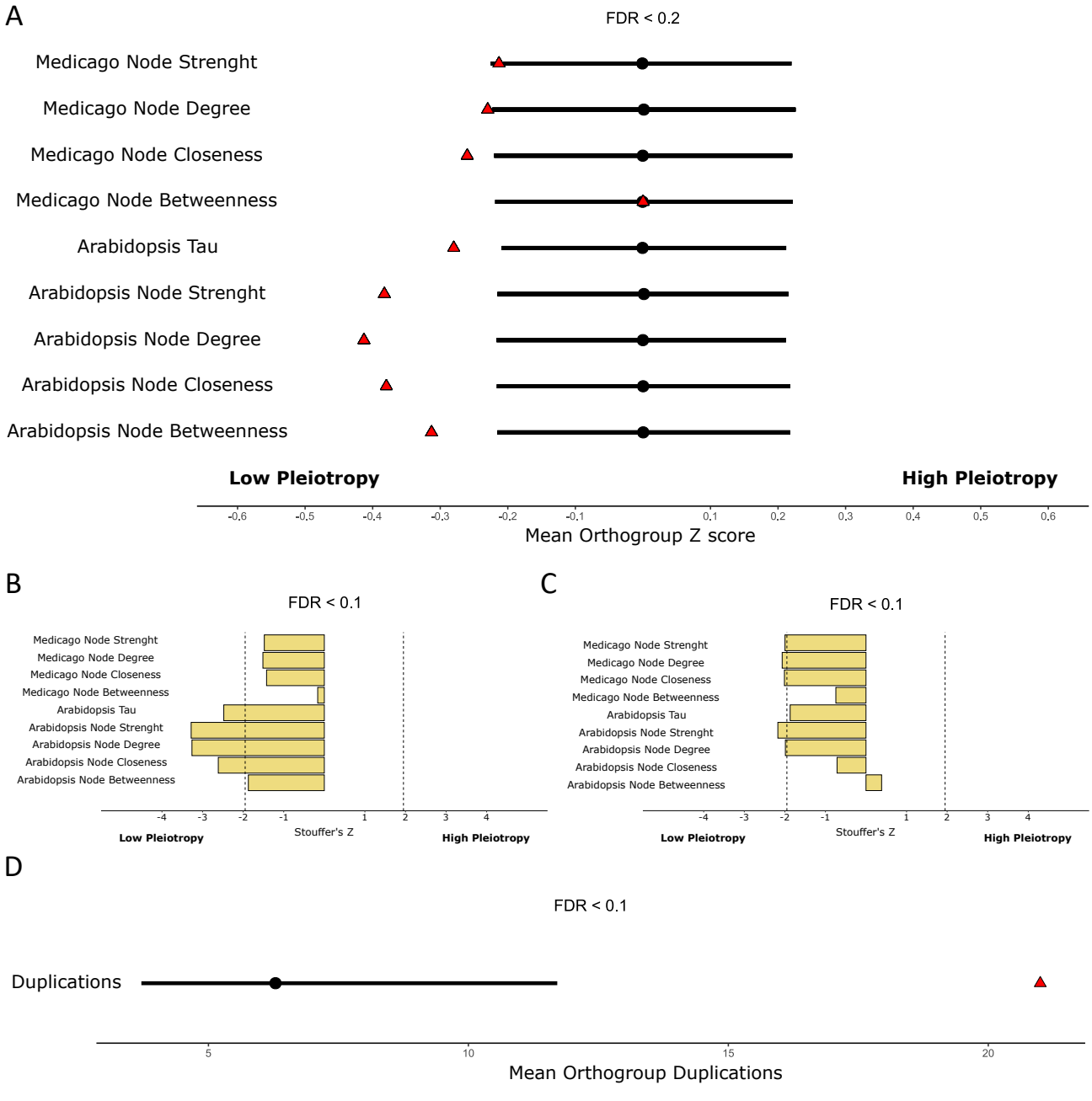


Fig. 4. Orthogroups with repeated sweeps tend to be less pleiotropic and harbour more duplications. (A) Pleiotropy assessment of the relaxed set of 107 orthogroups with FDR < 0.2, against 10,000 random draws. Red triangles represent the average pleiotropy of the candidate set, while black circles represent the mean of 10,000 random draws. Black lines represent the 95% interval of the random draws means. Each row represents a different pleiotropy measure, labelled on the left. (B) Pleiotropy Stouffer's Z of the top 33 orthogroups (FDR < 0.1), calculated using different pleiotropy measures labelled on the left. Dotted lines delimit the 95% confidence interval. (C) Pleiotropy Stouffer's Z for the more conservative set of 15 orthogroups identified by intersecting the 33 candidate OGs of the main analysis (FDR < 0.1) with results derived from nine additional *PicMin* omitting closely related *Eucalyptus* and *Helianthus* species. Dotted lines delimit the 95% confidence interval. (D) Duplication bootstrapping assessment of the top 33 orthogroups (FDR < 0.1) against 10,000 random draws. The red triangle represents the average duplication number in the candidate set of orthogroups, while the black circle and line represent mean and 95% interval of 10,000 random draws respectively.


A
B
Arabidopsis thaliana
6.2MYA
Arabidopsis halleri
Capsella rubella
113MYA
Boechera stricta
Eucalyptus albens
Eucalyptus magnificata
Eucalyptus sideroxylon
Medicago truncatula
Betula pendula
Betula platyphylla
122MYA
Populus trichocarpa
2.5MYA
Populus tremula
Helianthus annuus
160MYA
Helianthus argophyllus
118MYA
Helianthus petiolaris
Amaranthus tuberculatus
Panicum hallii
C
D
related
unrelated
Individuals Count


A
Paralogues per orthogroup
P. trichocarpa
P. tremula
P. hallii
M. truncatula
H. petiolaris
H. argophyllus
H. annus
E. sideroxylon
E. magnificata
E. albens
C. rubella
B. stricta
B. platyphylla
B. pendula
A. tuberculatus
A. thaliana
A. halleri
Species
0
5000
10000
Number
Paralogues N
1
2
3
4
5
6
7
8
9
10
B
Species per orthogroup
Species
Species N
7
8
9
10
11
12
13
14
15
16
17
C
OGs tested in PicMin
Number
0
1000
2000
3000
4000
7
8
9
10
11
12
13
14
15
16
17
Number of Species
D
TimeTree
Orthofinder2
P. trichocarpa
P. tremula
B. pendula
M. truncatula
E. grandis
A. thaliana
A. halleri
C. rubella
B. stricta
H. annuus
A. tuberculatus
P. hallii


A
OGs tested = 13,268
-log10(picmin FDR)
1.0
0.5
0.0
0
1
2
3
4
5
6
7
8
Number of OP emp-p < 0.1
B
Species
E. sideroxylon
E. albens
H. petiolaris
H. argophyllus
E. magnificata
H. annuus
P. tremula
B. pendula
M. truncatula
A. halleri
P. trichocarpa
B. platyphylla
A. thaliana
A. tuberculatus
C. rubella
P. hallii
B. stricta
0
20
40
60
Contribution %
C
PicMin FDR
0.1
0.05
0
P. hallii
A. tuberculatus
H. petiolaris
H. argophyllus
H. annuus
B. stricta
C. rubella
A. halleri
A. thaliana
M .truncatula
E. sideroxylon
E. magnificata
E. albens
B. platyphylla
B. pendula
P. tremula
P. trichocarpa
emp-p
1
0.5
0
OG0010509
OG0001601
OG0003625
OG0017585
OG0001991
OG0010096
OG0001035
OG0017365
OG0010150
OG0001831
OG0007322
OG0000002
OG0006228
OG0007151
OG0003537
OG0022308
OG0003347
OG0002226
OG0019910
OG0004755
OG0017633
OG0014549
OG0005857
OG0002218
OG0004282
OG0005418
OG0004877
OG0000525
OG0002608
OG0002632
OG0006466
OG0006404
OG0016381


A
FDR < 0.2
Medicago Node Strenght
Medicago Node Degree
Medicago Node Closeness
Medicago Node Betweenness
Arabidopsis Tau
Arabidopsis Node Strenght
Arabidopsis Node Degree
Arabidopsis Node Closeness
Arabidopsis Node Betweenness
Low Pleiotropy
High Pleiotropy
-0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.1 0.2 0.3 0.4 0.5 0.6
Mean Orthogroup Z score
B
FDR < 0.1
Medicago Node Strenght
Medicago Node Degree
Medicago Node Closeness
Medicago Node Betweenness
Arabidopsis Tau
Arabidopsis Node Strenght
Arabidopsis Node Degree
Arabidopsis Node Closeness
Arabidopsis Node Betweenness
-4 -3 -2 -1 1 2 3 4
Stouffer's Z
Low Pleiotropy
High Pleiotropy
C
FDR < 0.1
Medicago Node Strenght
Medicago Node Degree
Medicago Node Closeness
Medicago Node Betweenness
Arabidopsis Tau
Arabidopsis Node Strenght
Arabidopsis Node Degree
Arabidopsis Node Closeness
Arabidopsis Node Betweenness
-4 -3 -2 -1 1 2 3 4
Stouffer's Z
Low Pleiotropy
High Pleiotropy
D
FDR < 0.1
Duplications
5 10 15 20
Mean Orthogroup Duplications