nature ecology & evolution

https://doi.org/10.1038/s41559-018-0787-9

The global biogeography of polyploid plants

Anna Rice[1], Petr Šmarda[2], Maria Novosolov[3,4], Michal Drori[1], Lior Glick[1], Niv Sabath[1], Shai Meiri[3,4], Jonathan Belmaker[3,4] and Itay Mayrose[1]*

Deciphering the global distribution of polyploid plants is fundamental for understanding plant evolution and ecology. Many factors have been hypothesized to affect the uneven distribution of polyploid plants across the globe. Nevertheless, the lack of large comparative datasets has restricted such studies to local floras and to narrow taxonomical scopes, limiting our understanding of the underlying drivers of polyploid plant distribution. We present a map portraying the worldwide polyploid frequencies, based on extensive spatial data coupled with phylogeny-based polyploidy inference for tens of thousands of species. This allowed us to assess the potential global drivers affecting polyploid distribution. Our data reveal a clear latitudinal trend, with polyploid frequency increasing away from the equator. Climate, especially temperature, appears to be the most influential predictor of polyploid distribution. However, we find this effect to be mostly indirect, mediated predominantly by variation in plant lifeforms and, to a lesser extent, by taxonomical composition and species richness. Thus, our study presents an emerging view of polyploid distribution that highlights attributes that facilitate the establishment of new polyploid lineages by providing polyploids with sufficient time (that is, perenniality) and space (low species richness) to compete with pre-adapted diploid relatives.

Polyploidy is a key feature of extant organismal diversity, particularly in plants[1–4], occurring during the evolution of all seed plant lineages[1], with approximately 35% of extant flowering plant species being of recent polyploid origin[5]. Due to the strong reproductive barrier between newly arisen polyploids and their parental diploid populations, polyploidy is frequently considered an instant speciation mechanism[6,7]. Specifically, allopolyploids are formed through the hybridization of two differentiated genomes while reproductive barriers of autoploypoids, formed through genome duplication within a single species, may be less severe due to recurrent gene flow from their diploid progenitors[8] (but see ref. [9]). However, to affect plant communities and floras, polyploids must become locally established and propagate once formed. The nearly complete post-zygotic reproductive barrier posed between individuals of different ploidy levels may lead to the extinction of the polyploid—not because it is less fit, but because it is less common. This minority cytotype exclusion[10] serves as an initial selective sleeve that disproportionally allows the survival of emerging polyploid lineages bearing certain properties (for example, high selfing rates[11]). Indeed, successfully established polyploids, especially allopolyploids, often differ from their diploid progenitors in morphological, physiological, life history and reproductive characteristics[3,12–15], while some autopolyploids may remain cryptic due to morphological similarities with surrounding diploids[9,16]. The potential phenotypic differences between polyploids and their diploids may have contributed to the differential establishment and success of polyploids in certain ecological, geographical, edaphic and climatic settings[17].

The biogeography of polyploidy has intrigued scientists for decades. Polyploid frequencies (especially in the northern hemisphere) have often been shown to increase with latitude[14,18–20]. Several hypotheses were raised to explain this phenomenon, ranging from mutational and mechanistic perspectives to niche availability and selective processes. First, polyploid abundance could be correlated with the frequency of unreduced gamete production, the major mechanism of polyploidy formation[21]. Many studies noted an increased frequency of unreduced gametes under cold treatments and in harsh and fluctuating environments that are more typical of polar latitudes[8,22,23]. Polyploids are further thought to be more frequent in disturbed and vacant habitats, such as newly deglaciated or anthropogenically disturbed regions, due to high niche availability and lower competition[18,23–25]. Furthermore, according to the secondary contact hypothesis[26] allopolyploids should be more abundant in formerly glaciated regions due to increased mating opportunities between divergent allopatric populations that were brought into contact following ice retreat. The resulting hybridizations could subsequently have led to the formation and establishment of new allopolyploid lineages. The duplicated genomic content of polyploids may result in a higher capacity for adaptation[17,25,27–29] and could enable polyploids to inhabit more extreme environments, but could impose a greater demand for nucleic acids[3], restricting polyploid abundance in regions with low availability of phosphorus and nitrogen[30,31]. This could potentially explain the low polyploid abundance in phosphorus-poor soils, such as those found in the tropics[32]. Furthermore, the variation in polyploid frequency across the major plant groups[33,34], together with the disproportional abundance of certain taxonomical groups across the globe[35], could also affect the spatial distribution of polyploids.

Finally, life history traits could strongly influence polyploid distribution as polyploids are disproportionally more frequent among perennial herbs[14]. This phenomenon has been attributed to the longer life cycle of perennials, which provides more time for polyploids to find compatible mating partners and to produce fertile offspring[36,37]. Alternatively, the lower frequency of polyploidy among perennial woody plants has been attributed to ecological and historical factors that may have retarded the establishment of new polyploid woody lineages. These factors may include a reduced rate of new habitat emergence in regions where woody plants are enriched (for example, the tropics), such that the potential benefits of polyploids were not realized[14]. Taken together, because perennial herbs predominate in colder regions and higher latitudes[38], the

[1]School of Plant Sciences and Food Security, Tel-Aviv University, Tel-Aviv, Israel. [2]Department of Botany and Zoology, Masaryk University, Brno, Czech Republic. [3]School of Zoology, Tel-Aviv University, Tel-Aviv, Israel. [4]Steinhardt Museum of Natural History, Tel-Aviv University, Tel-Aviv, Israel. *e-mail: itaymay@post.tau.ac.il

higher frequency of polyploidy in these regions may be explained by the proportions of different lifeforms in the respective floras[14,39].

Being fundamental to plant evolution, much effort has been made to decipher the patterns underlying polyploid biogeography. However, broad generalizations are limited owing, in part, to the absence of large comparative data with sufficiently broad taxonomical and geographical scopes. Here, we assembled extensive worldwide distributional data coupled with phylogenetically based ploidy inferences for tens of thousands of angiosperm species, to provide a global map of polyploid distribution. This allowed us to revisit previous hypotheses and to unravel new aspects regarding the conditions that govern the distribution of polyploid plants.

Results and Discussion

A global map of polyploid distribution. We present a map portraying the distribution of polyploid frequency across the globe (hereafter, 'polyploid frequency' is referred to as the relative proportion of polyploid species out of all species with ploidy estimates). To this end, we compiled an extensive dataset of angiosperm genera in which ploidy levels were estimated from cytological and phylogenetic data. Specifically, we reconstructed the phylogeny of each angiosperm genus using publicly available sequence data (see section on Phylogeny reconstruction, below). Chromosome numbers were then mapped to terminal taxa (mostly species) and ploidy shifts were inferred within each genus, using a probabilistic model of chromosome number evolution[40]. This allowed us to classify an extant taxon as (neo)polyploid if it had undergone a polyploidization event since divergence from its generic ancestor, and as diploid otherwise. Georeferenced species distributions were obtained, and genera with adequate coverage and occurrence precision were retained (see section on Geographical and climatic data, below). These procedures resulted in a database consisting of > 25 M occurrence records spanning 1,287 genera and 26,599 terminal taxa, with thousands of inferred polyploidization events. In agreement with previous estimates[5], approximately 33% (8,786) of the species in the database were inferred as polyploids. We further used data and text-mining procedures and assembled an extensive database of plant lifeforms that included information on > 140,000 angiosperm species (Supplementary Table 1; see Methods). In this dataset, the frequency of polyploids is 28% within annual species, 39% within perennial herbs and 22% within woody species (Supplementary Table 2).

Fig. 1a presents the global distribution of polyploid frequencies at the coarse resolution of the 14 terrestrial biomes (that is, major ecological land classifications as defined by the World Wildlife Fund (http://www.worldwildlife.org/); Table 1). A strong latitudinal pattern is evident: polyploid frequency increases away from the equator, particularly toward the northern pole, as previously reported at smaller scales[14,18]. The highest polyploid frequency (51%) is found in the tundra biome, in the far northern hemisphere; the lowest percentages are found in the tropical and subtropical biomes. The large taiga biome, located south of the tundra and characterized by harsh climatic conditions, contains the second highest proportion of polyploid species (47%). Relatively high polyploid frequencies are found in the temperate zones (38–40%) and in the montane grasslands (39%). The latter are characterized by high altitude with mostly cool and wet conditions[41] and frequently appear as islands of high polyploid frequency within the surrounding low-altitudinal biomes (for example, the Ethiopian Highlands, Andes and much of New Guinea).

Next, we examined polyploid distribution at the finer resolution of ecoregions[42], representing ~800 geographically distinct natural assemblages and environmental conditions (Fig. 1b; Supplementary Table 3). A similar latitudinal trend was observed, with ecoregions exhibiting the highest polyploid frequencies located near the poles. Without exception, all ecoregions belonging to the taiga, and particularly to the tundra, are polyploid rich (containing ≥ 38% and ≥ 50% polyploids, respectively). Polyploid-poor ecoregions are generally found in tropical and subtropical regions (see Supplementary Fig. 1 for ecoregions with highest and lowest polyploid frequencies). Interestingly, the polyploid frequencies of 39 ecoregions are markedly different from those of their respective biomes ($P < 0.05$; binomial test followed by Bonferroni correction; Supplementary Table 4). These include tropical ecoregions that are particularly polyploid rich, such as several ecoregions in the Andes and Hawaii, an archipelago known to be polyploid rich[43]. Many ecoregions where polyploid frequency is surprisingly high for their biome belong to temperate zones, and include mountainous areas (for example, the Rocky Mountains and the Alps). The Magellanic subpolar forests, the southernmost ecoregion in South America that has markedly de-glaciated since the Last Glaciation Maximum (LGM; 22,000 years ago), also belongs to a temperate biome but contains 68% polyploids. Notably, only two ecoregions exhibit polyploid frequencies markedly lower than their respective biomes. The first is the Montane Fynbos and Renosterveld ecoregion, which is part of the Cape flora, a region known to be exceptionally polyploid poor[44]. The second ecoregion is the Daba Mountains in China, in which the low polyploid frequency may be driven by the high proportion of woody species dominating it.

Global drivers of polyploid distribution. The large dataset assembled here enabled us to test whether biogeographic patterns of polyploid diversity, which were previously identified and discussed at smaller scales, are globally evident. Initially, we used a generalized linear model (GLM) to assess the relationships between polyploid frequencies at the ecoregion level and individual eco-geographical explanatory variables (Fig. 2; Supplementary Table 5 provides results of each variable separately; all results are robust to spatial autocorrelation as determined using bootstrap sampling, see Methods). Our analysis confirms that polyploid frequency increases towards the poles, as it correlates quadratically with latitude (McFadden pseudo-R^2 (R^2_{MF}) = 50%, $P \ll 0.01$). Polyploid frequency across ecoregions is correlated, to some extent, with all bioclimatic variables, particularly temperature variables. Polyploids tend to prevail in cooler climates ($R^2_{\mathrm{MF}} = 45\%$, $P \ll 0.01$; annual mean temperature), a phenomenon previously discussed in depth[22,23]. Precipitation has a milder effect, with the most noticeable pattern being that polyploids are less frequent in regions with greater rainfall seasonality ($R^2_{\mathrm{MF}} = 19\%$, $P \ll 0.01$). We further tested whether change-in-climate (see Methods) is correlated with polyploid frequency. An overall lower association was found compared to current climate conditions. Comparing ice coverage between LGM and the present, we found a positive relationship between the extent of de-glaciation and polyploid abundance ($R^2_{\mathrm{MF}} = 19\%$, $P \ll 0.01$), as mentioned previously[16]. Additionally, ecoregions with lower net primary productivity (NPP), and particularly those with lower species richness, have a higher proportion of polyploids ($R^2_{\mathrm{MF}} = 10$ and 28%, respectively; $P \ll 0.01$). Polyploid frequency is associated, to some extent, with the taxonomical composition of local floras (categorized into six groups, hereafter termed TaxComp; see Methods), correlating positively with the percentage of commelinids (represented mainly by the grasses) and negatively with the percentage of rosids ($R^2_{\mathrm{MF}} = 25\%$ for both attributes, $P \ll 0.01$). Finally, polyploid distribution is strongly associated with lifeform characteristics (as described in ref. [45]). Polyploids are more frequent in ecoregions with a high frequency of perennial herbs ($R^2_{\mathrm{MF}} = 44\%$, $P \ll 0.01$) and where woody species are rarer ($R^2_{\mathrm{MF}} = 30\%$, $P \ll 0.01$). Other prevailing distributional hypotheses, such as the association with phosphorus soil availability, areas of anthropogenic disturbance, altitude and elevation amplitude, had a low explanatory power ($R^2_{\mathrm{MF}} < 2\%$). However, it is possible that the ecoregion level is too crude to reveal meaningful altitudinal effects. Alternatively, while there are ample documented examples of differences in altitudinal position between populations of polyploids and

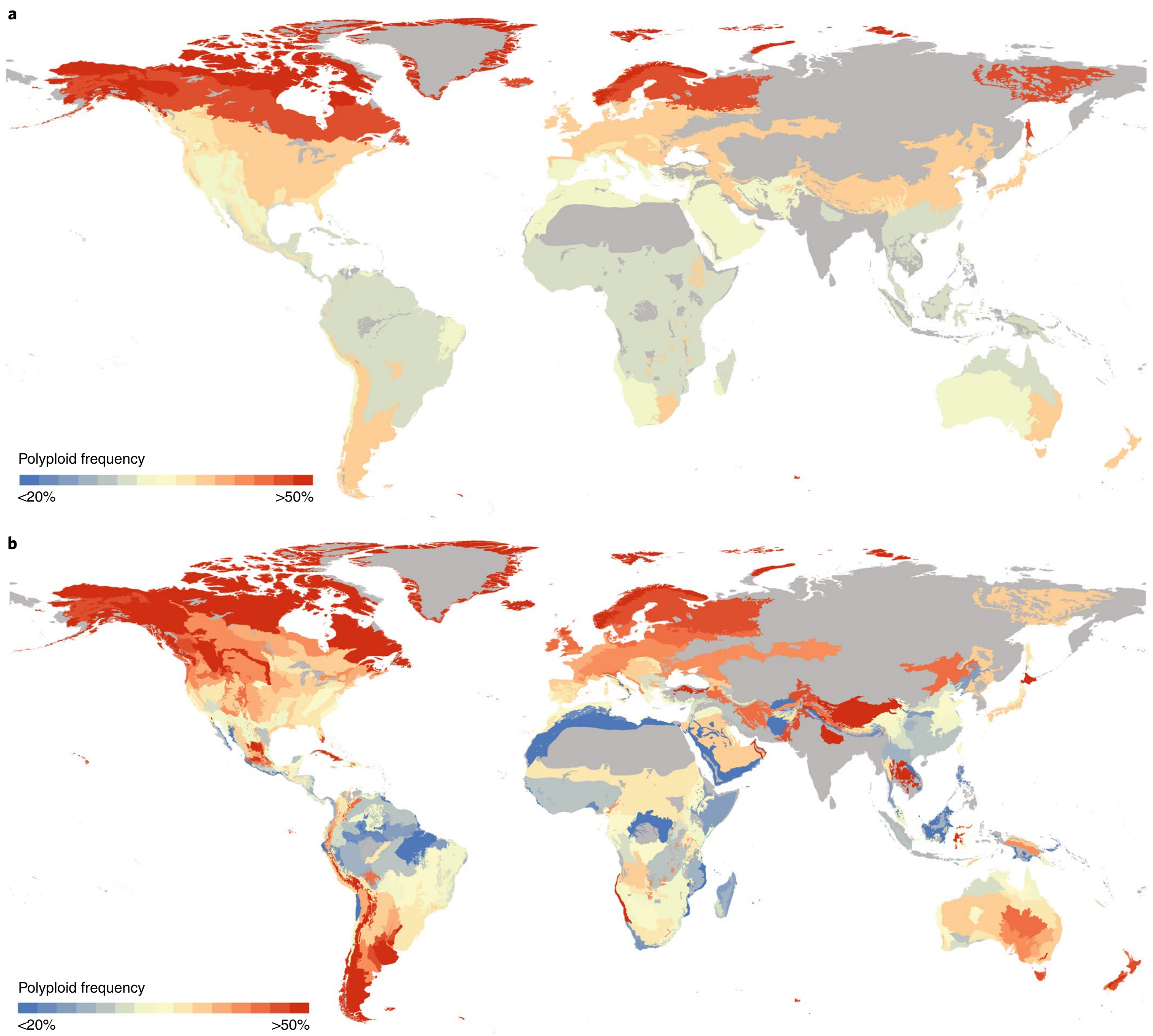

Fig. 1 | The global distribution of polyploid frequency. a,**b**, The percentage of polyploid species is presented for (**a**) 14 terrestrial biomes and (**b**) terrestrial ecoregions following the colour scale at bottom left. In both panels, ecoregions with insufficient data (see Methods) are coloured grey, resulting in 528 coloured ecoregions. To keep both maps at the same colouring scale, all ecoregions with <20% polyploids are coloured blue and those with >50% are coloured dark red.

their diploid counterparts, this phenomenon could be driven by a strong selective force for ecological divergence rather than by a consistent trend towards higher or lower altitudes[46]. Additionally, the previously reported observation of an altitudinal trend (for example, ref. [18]) may have been driven by a confounding attribute—for example, if examined mostly in formerly glaciated mountains or in areas with large temperature gradients.

To reduce the dimensionality of the parameter space, particularly for multiple variables belonging to climate, taxonomical and lifeform attributes, we examined whether the use of composite variables could serve as a meaningful predictor for each variable group. To this end, principle component analysis (PCA) was applied to each of the following variable groups: climate (19 BIOCLIM variables), change-in-climate (19 variables), TaxComp (six variables), and lifeform (three variables). The first three components of climate explained 0.82 of the total variance in polyploid frequency across ecoregions. The loadings of all principal components were low to moderate (for example, the highest squared loading of PC1 was $\leq$ 0.1), rendering inferences based on their analysis hard to interpret. The first component (PC1) of TaxComp and lifeform explained 0.36 and 0.73 of the total variance, respectively. PC1 of TaxComp was mostly influenced by the percentages of rosids, basal angiosperms and basal dicots with loadings of 0.54, 0.51 and −0.50, respectively. The loadings of PC1 for lifeform were −0.67, 0.60 and 0.44 for percentages of woody, perennial herbs and annual species, respectively (see Supplementary Table 6 for other PCA results). Examining the association between these principal components and polyploid frequencies revealed that PC1 of TaxComp explained a greater variance than any of its individual attributes ($R^2_{\mathrm{MF}}=29\%$, $P\ll0.01$). On the other hand, in climate and lifeform composition, several of the

Table 1 | Proportions of polyploids and perennial herbs in the 14 major terrestrial biomes of the world

Biome	Number of species[a]	Polyploids (%)	Perennial herbs (%)[b]
Tropical and subtropical moist broadleaf forests	7,667	30.05	12.54
Mangroves	683	30.31	13.50
Tropical and subtropical dry broadleaf forests	2,132	31.19	12.68
Tropical and subtropical grasslands	4,079	31.45	17.45
Mediterranean forests, woodlands and scrub	8,605	32.68	24.91
Deserts and xeric shrublands	5,659	32.73	20.93
Tropical and subtropical coniferous forests	1,803	37.10	20.87
Temperate conifer forests	7,415	37.65	42.23
Flooded grasslands and savannas	606	38.28	26.63
Temperate broadleaf and mixed forests	10,179	38.81	30.61
Montane grasslands and shrublands	3,737	38.85	27.99
Temperate grasslands, savannas and shrublands	4,617	39.68	34.78
Boreal forests/taiga	2,014	46.67	47.22
Tundra	1,341	50.78	57.79

[a] Number of species was calculated from the database containing only species with ploidy-level inference ($n=26,599$). [b] The percentage of perennial herbs was calculated from the larger database including species with georeferenced data, with or without ploidy-level inference ($n=134,769$).

original attributes of each group of variables had a higher explanatory power than their PC1 ($R^2_{\mathrm{MF}}=32\%$ and $R^2_{\mathrm{MF}}=28\%$, respectively; $P\ll0.01$). This indicates that the central predictor of lifeform on polyploid distribution is the relative proportion of perennial herbs, while the percentages of woody and annual species may counterbalance each other such that a composite lifeform component has a lower predictive value on polyploid frequency.

While many of the individual predictors were individually highly associated with polyploid frequency, the GLM results did not indicate their relative importance. To this end, hierarchical partitioning[47] was applied within a multiple predictor model to decompose the variance of polyploid frequency across the globe (again, examined at the ecoregion level) to the individual contributions of several variable groups. To focus on a succinct set of predictors, and for computational tractability, composite variables were used, representing taxonomical and lifeform compositions, climate and change-in-climate. Species richness, NPP, phosphorus retention, altitude, elevation amplitude and anthropogenic disturbance were included as single predictors. The total variance explained by the combined model was 74% (Fig. 3; see Supplementary Table 7 for results with bootstrap sampling), with the strongest impact on polyploid proportions ascribed to climatic attributes, followed by TaxComp, lifeform, species richness and change-in-climate (explaining 26, 16, 14, 8 and 8% of the variance, respectively). The contributions of all single predictors were negligible, and thus these were excluded from any subsequent analyses.

The relative importance analysis on each variable group was conducted separately (climate, TaxComp, lifeform and change-in-climate; Supplementary Fig. 2). A model that included only the BIOCLIM attributes explained 54% of the total variance in polyploid frequencies across ecoregions, with the highest contribution ascribed to temperature measurements (for example, BIO1: mean annual temperature and BIO10: temperature of warmest quarter), followed by precipitation seasonality (BIO15), temperature variability (for example, BIO4: temperature seasonality) and precipitation (for example, BIO14: precipitation of driest month). This reinforces the importance of temperature in driving polyploid distribution.

The effect of climate on polyploid distribution. While many of the predictor variables were each strongly associated with polyploid frequency, association does not imply causation and it is challenging to determine whether a specific attribute has a direct or an indirect effect on the response variable. Therefore, we used path analysis, evaluating the interplay between the major attributes affecting polyploid distribution, with an emphasis on the crucial role of climate, and particularly that of temperature (following the results above; Fig. 2b and Supplementary Fig. 2). We focused on establishing the associations between the most influential variable groups: climate, TaxComp, lifeform and species richness. Lifeform, TaxComp and climate were each represented by a single predictor: climate by mean annual temperature (BIO1), TaxComp by its PC1 and lifeform by the percentage of perennial herbs (these representatives were chosen based on the GLM and relative importance results; Supplementary Table 5 and Supplementary Fig. 2); results obtained using other representations for climate were qualitatively similar (Supplementary Text 5).

First, we established an a priori model to distinguish between the direct effect of climate on polyploid distribution and its indirect effects, mediated through the three other main predictors (Fig. 4). Because the effect of climate on species richness[48,49], taxonomic composition[50,51] and plant lifeform[52,53] is well established on a global scale, we specified a direct path from climate to each of these predictors. Following previous observations, we additionally included direct paths from TaxComp to lifeform[54] and to species richness[55,56], as both of these attributes are dependent on the phylogenetic composition of the flora. Finally, we specified a direct effect of each of the four predictors on polyploid frequency.

Applying this model to the data indicated a good fit between the observed covariance matrix and that predicted by the model ($\chi^2=0.02$, d.f. = 1, $P=0.90$). Accordingly, the inferred path coefficients indicated that the variable with the highest direct effect on polyploid distribution was perennial herbs percentage (path coefficient of 0.45), while the direct effect of climate was weaker (path coefficient of −0.17). Species richness and TaxComp also exhibited relatively weak direct associations (path coefficients of −0.20 and 0.06, respectively; TaxComp being non-significant with $P=0.29$). Considering the cumulative effect of each variable (see Methods), climate was inferred as the most influential, followed by perennial herbs percentage, species richness and TaxComp (cumulative effects of −0.57, 0.45, −0.20 and −0.14, respectively). Thus, considering the strong direct effect of climate on each of the other three variables, particularly perennial herbs percentage and TaxComp (path coefficients of −0.52 and 0.66, respectively), the contribution of climate to worldwide polyploid distribution seems to be predominantly mediated through its effect on the floral composition of ecoregions.

Conclusion

Geographic variation in polyploid frequencies among plants is striking. However, no global description of this variation has been available thus far. The geographic and lifeform databases assembled here, together with the large-scale ploidy inferences, offer the means to examine prominent hypotheses regarding worldwide polyploid distribution, and to suggest a few extensions. Our results demonstrate that the predominant factor shaping polyploid distribution is climate, particularly temperature (for example, through

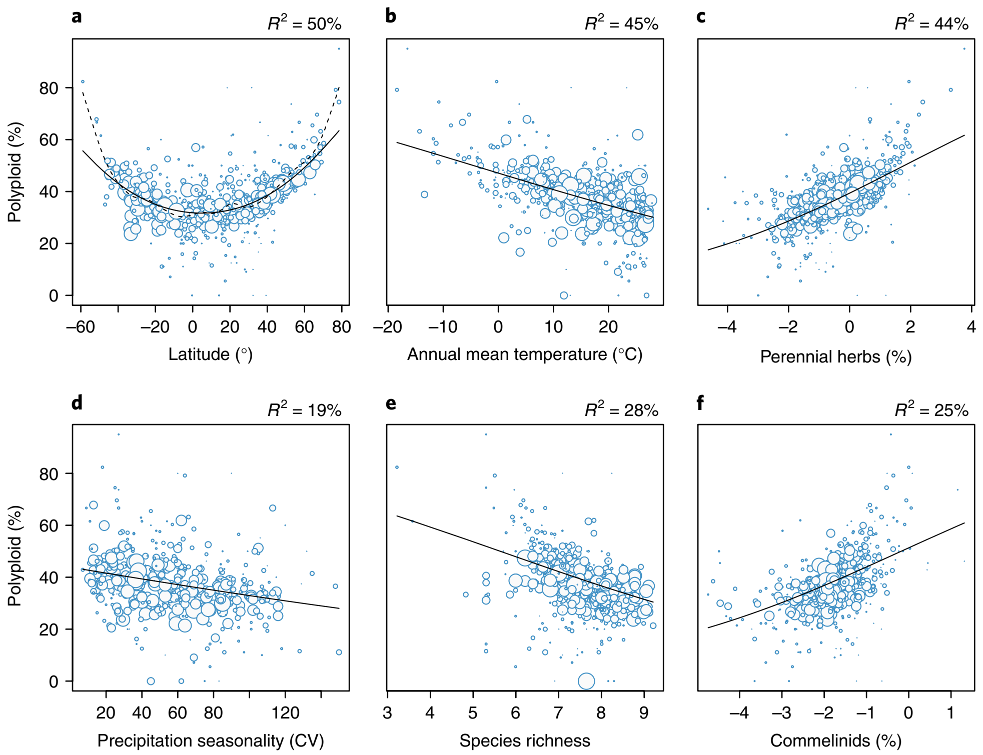

Fig. 2 | GLM results for different predictors associated with polyploid percentage. a–e, Polyploid percentage per ecoregion (n = 528) as a function of (**a**) latitude, (**b**) annual mean temperature, (**c**) perennial herbs percentage, (**d**) precipitation seasonality, (**e**) species richness and (**f**) commelinid percentage. The x axis for **e** is shown following log transformation and for **c** and **f** following logit transformation. In **d**, CV stands for coefficient of variation. The size of the circles corresponds to the number of species analysed for which ploidy data per ecoregion are available. The solid fitted line represents the GLM fit. Squared term was used in **a**, together with the cubic spline fit (dashed line), as suggested by Schluter[90]. The R^2 value associated with each predictor is displayed on the top right corner of each panel.

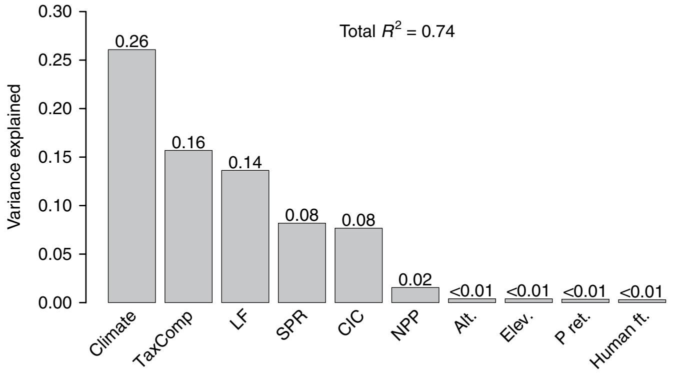

Fig. 3 | The relative importance of predictors to polyploid frequency. Relative importance was assessed as the average contribution of each predictor to the variance in polyploid frequency across ecoregions (n = 528) computed over all possible models. Four predictors (climate, TaxComp (taxonomical composition), LF (lifeform) and CIC (change-in-climate)) represent groups of attributes, while the other six (SPR (species richness), NPP, Alt. (altitude), Elev. (elevation amplitude), P ret. (phosphorus retention) and Human ft. (human footprint)) were treated as individual predictors. The numbers above the bars correspond to the variance explained by each predictor separately and sum to the total variance explained by the model.

environmentally dependent rates of unreduced gamete formation[8]). Nevertheless, our analyses suggest that most of its influence is mediated via other local attributes, particularly the lifeform composition of the flora. As such, much of the frequently highlighted association of polyploid frequency and latitudinal gradients, and exceptions thereof[26], could be attributed to the higher abundance of perennial herbs at high latitudes, as has been exemplified for plant mating systems[57]. Taken together, the emergent view of polyploid distribution highlights the importance of attributes that favour the establishment of new polyploid lineages. This is true for attributes that weaken the minority cytotype disadvantage by providing polyploids with time and space (that is, perenniality and low species richness) to find a compatible mating partner. Furthermore, we found that taxonomic composition impacts polyploid frequency—polyploids are relatively common where commelinids are abundant and rosids are rare. In the case of the former, this may be attributed to variables that were not incorporated into our analysis, such as the tendency of this clade to reside in places that are favourable for polyploid formation (for example, areas prone to secondary contacts, as suggested by Stebbins[24] for the grasses). Additionally, it is possible that diploidization is exceptionally rapid in this clade, making recurrent polyploidy more likely.

Finally, we should acknowledge possible limitations that are inherent in the large-scale analysis presented here. First, the statistical framework used to infer duplication events is unable to distinguish between allopolyploids and autopolyploids. While the categorization of species to ploidy levels is expected to remain accurate even with reticulate evolution, since a large degree of the signal used to infer ploidy levels is derived from the inferred chromosome number at the root of the phylogeny, we could not address any of the hypotheses regarding the possible differential distribution of allopolyploids and autopolyploids. This topic should be of prime interest in future research. Our results may be partially influenced

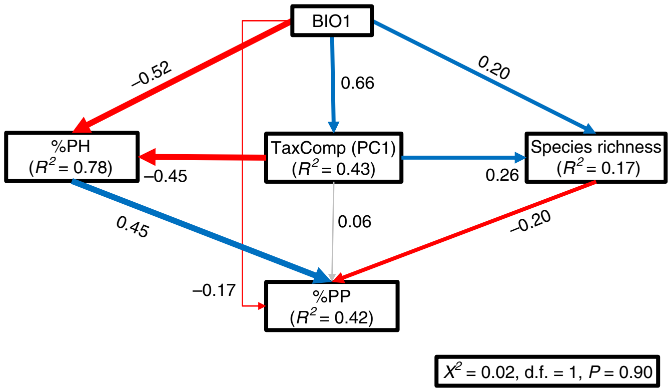

Fig. 4 | Path analysis. A path diagram representing the relationships between the putative predictors and percentage of polyploids (%PP) within an ecoregion ($n = 528$). The model includes the four most influential predictors, including climate (represented by BIO1), lifeform (represented by percentage of perennial herbs; PH%), taxonomical composition (represented by its PC1) and species richness. Red and blue paths refer to the nature of the relationship (negative and positive, respectively). Non-significant associations ($P > 0.05$) are in grey. The thickness of the paths is scaled based on the magnitude of the standardized path coefficients, which are given by the numbers adjacent to arrows. The R^2 value associated with each exogenous variable is presented within each box, and the data fit measures are presented below the figure.

by species delimitation biases. While some taxonomists tend to treat allopolyploids as distinct species, others regard variations in chromosome number as taxonomically unimportant, particularly in autopolyploids, even when they do represent distinct evolutionary lineages[9,15]. This could have underestimated the relative number of autopolyploid species used in our analyses. Second, the phylogenetic relationship between species has not been considered when examining the drivers of polyploid distribution. This could better account for possible taxonomic biases (for example, when polyploid frequency of a certain ecoregion is based on few dominating clades). Third, our analyses are based on current distributional patterns, which mainly reflect established plant populations. Thus, the drivers that govern rates of polyploid formation (for example, production of unreduced gametes) may be different to those reported here. Fourth, other putatively confounding attributes (for example, differential geographical abundance of sexual and mating systems[57,58], which are known to be associated with polyploidy[8,59]) were not considered. Data on these attributes currently exist for only a small fraction of our database, and thus await future compilations. Furthermore, all our assembled databases (ploidy, species occurrences and lifeform) obviously suffer from sampling biases, which could affect our conclusions. Indeed, while our analyses were conducted at a global scale, they still did not include some of the most extreme regions of the world (for example, Siberia and the South Saharan steppe). Notwithstanding, our results are robust to bootstrap resampling, indicating that the findings are likely to prove valid when future compendia are assembled. Lastly, our analyses were conducted at the ecoregion scale and therefore higher-resolution alterations in potentially important factors (for example, soil composition or altitude) were overlooked. Nonetheless, with the accumulation of more data, approaches such as that presented here will allow the discovery of increasingly refined phenomena governing the biogeographic patterns of polyploids.

Methods

Phylogeny reconstruction. To enable large-scale comparative analyses, we used a variant of the OneTwoTree[60] pipeline to automatically reconstruct the phylogeny of each angiosperm genus using publicly available sequence data. Briefly, sequence data for all taxa within a genus (including all species and intraspecific taxa) were automatically retrieved from GenBank[61]. This step was followed by a name resolution procedure, which matched species names as they appeared in GenBank to their currently accepted taxonomic names according to The Plant List (V1.1; http://www.theplantlist.org/). Sequences were then clustered to orthology groups and an appropriate outgroup was chosen and added to the corresponding clusters. The sequences in each cluster were then aligned, filtering out unreliably aligned sequences and positions, and concatenated to form a single partitioned alignment. Next, a relaxed clock model was applied to reconstruct a set of ultrametric Bayesian phylogenies, using the best-fitting nucleotide substitution model for each cluster. A detailed description of this procedure is given in Supplementary Text 1.

Ploidy-level inferences. It is currently accepted that all angiosperm species have experienced one or more polyploidization events in their history[1], and so all angiosperms are in fact palaeopolyploids that have diploidized to some extent. Thus, polyploids must be defined with respect to a reference time point. In this study, polyploids are explicitly defined as those taxa that have undergone a polyploidization event since divergence from the most recent common ancestor of their genus. Accordingly, a lineage that had undergone polyploidy event after divergence from the most recent common ancestor, but has since diploidized, is still classified as a polyploid. Thus, our approach of using chromosome number variation within a genus to assign ploidy levels largely limits the detection of polyploidy to groups with un-diploidized polyploids, as with many previous studies (for example, ref. [5]).

Chromosome numbers for all species within a genus were extracted from the Chromosome Counts Database[34]. When multiple chromosome numbers were reported for a given taxon, we used the median count. Taxa-specific, ploidy-level inferences were based on the ChromEvol program (v2.0) that inferred duplication events for each genus separately[40]. This likelihood-based method accounts for various types of chromosome number transitions and estimates the expected number of polyploidy and dysploidy (that is, single chromosome changes) transitions along each branch of the phylogeny, thereby allowing categorization of terminal taxa as either diploid or polyploid relative to other taxa in the genus. To increase reliability, ploidy-level inferences were compared across 100 trees, thus accounting for phylogenetic uncertainties. Additionally, we used a parametric bootstrapping approach (described in ref. [40]) to detect taxa with ploidy-level inference of low reliability. Taxa for which ploidy levels could not be reliably inferred under both approaches were omitted from further analysis. Subsequent to ChromEvol analyses, additional ploidy-level inferences were obtained for species that were absent from the phylogeny but for which chromosome number data exist, using a threshold approach that accounted for ChromEvol inference of the respective genus. Full details of the ploidy-level inferences and reliability assessment are given in Supplementary Text 2. Of note, whether taxa appeared on the phylogeny or not, all inferences were made only for taxa that had an assigned chromosome number.

The robustness of the results to various alterations of the ploidy-level inferences was examined using two alternative strategies. First, we used a mega-phylogeny recently reconstructed by Smith & Brown[62] (termed here SB_TREE) to examine whether ploidy assignments are robust to alternative procedures used to reconstruct the phylogenies. Specifically, in accordance with the main analysis, we partitioned the SB_TREE into smaller subtrees such that each corresponds to a single genus. We then applied our ploidy inference pipeline to each of these phylogenies and compared the results (note that since the SB_TREE was inferred using maximum likelihood, ChromEvol was applied to a single phylogeny rather than multiple Bayesian phylogenies). For most genera, the number of species with assigned chromosome numbers found on the SB_TREE was fewer than that reconstructed using our pipeline. Thus, in this comparison we used the 155 genera that had full correspondence between the two taxa sets. The overall consent between the ploidy-level inferences of two alternative phylogenetic pipelines was >98% (the inferences for 2,110 out of 2,147 species were identical), indicating that the use of alternative phylogenies has little effect on our ploidy-level inferences.

Second, because our probabilistic inference is conducted at the genus level, it might underestimate the number of polyploidy events. This could occur, for example, if a genome duplication has occurred just before the diversification of the genus or if all ancestral diploid species have gone extinct, leading to all species being inferred as diploids. We thus repeated the analyses while setting as polyploids all species whose assigned chromosome number is higher than 15. The results obtained with this threshold approach were very similar to those obtained with the fully probabilistic approach (Supplementary Table 9).

We further examined possible biases in the ploidy-level inference pipeline by comparing our inferences to those manually curated in the Plant DNA C-values database, housed at Kew Royal Botanical Gardens[63]. We found a very high correspondence in the ploidy inferences of the two databases (>91%), with most discrepancies attributed to differences in chromosome number assignments (that is, many species have multiple reported counts in the Chromosome Counts Database and their median was different to that reported in Kew) and in the assumed base number of the genus (for example, for several genera multiple base numbers were reported in the literature; the inferences in Kew are based on one of these, while the likelihood calculations of ChromEvol consider many possible assignments of the root state weighted by their probability of generating the data).

Geographical and climatic data. Species-specific geographical information was based on occurrence data available through the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/). To this end, more than 113 million occurrence data points available in the Plantae kingdom were downloaded (20 October 2016) and processed locally, using only angiosperm data. These occurrence data were subject to filtering criteria before their download (for example, certain *invalid*, *unlikely* and *mismatch* issues as defined by GBIF; a full list of the issues omitted can be found in Supplementary Text 3). To ensure a uniform taxonomic classification—both across GBIF records and between GBIF data and other data sources (for example, sequence data as obtained through GenBank)—all records passed a name resolution process (see Supplementary Text 1).

The following steps were carried out using R[64] scripts to filter unreliable data points: (1) We used the R package CoordinateCleaner[65] (along with tidyverse[66], rgbif[67] and rnaturalearthdata[68]), which discards problematic records (see Supplementary Text 3 for a full description of this process). (2) Occurrences whose coordinates had fewer than two decimal digits were discarded due to lack of precision. (3) Similarly, occurrences were removed in case the GBIF 'coordinate uncertainty' field was >10 km. (4) Occurrences whose 'Basis of records' GBIF field was *literature* or *living specimen* were discarded (generally, these refer to the location of the collection for occurrences sampled in museums or ex situ collections, respectively). (5) Crop species were discarded, since occurrence data of such taxa may not represent natural habitats. The list of crop species was obtained from the Food and Agricultural Organization of the United Nations website (http://faostat3.fao.org/home/E), Crop Index database of Purdue University (https://www.hort.purdue.edu/newcrop/Indices/index_ab.html) and a list compiled by Meyer et al.[69]. (6) Lastly, species with fewer than five occurrences were excluded.

To partly overcome sampling biases and to smooth potential inexact geographical measurements, we mapped each georeferenced data point into a wider eco-geographical scale defined by distinctive biodiversity and environmental conditions. For this purpose, each georeferenced point was assigned to its corresponding biome, representing one out of 14 major habitat types as defined by the World Wildlife Fund and terrestrial ecoregions (867 possible assignments[42]). These mappings were obtained using the R packages raster[70] and rgdal[71]. Each species was assigned to all biomes and ecoregions it inhabits (aggregation of species occurrences was done using the R package plyr[72]). This resulted in an initial database that included 134,769 species (disregarding whether they have ploidy-level inferences), encompassing 686 ecoregions with occurrences data. All ecoregions that were subsequently analysed in this study ($n = 528$; Supplementary Table 4) had at least ten species with georeferenced data points within them, each having at least five occurrences and at least five of these species with ploidy inferences. These filtering steps were performed to ensure that any assignment of species composition attribute in an ecoregion is based on a sufficient amount of data and not on a random observation (for example, a small ecoregion containing only one species, which happens to be a polyploid, would be assigned with 100% polyploid frequency).

Predictors of polyploid distribution. To test various hypotheses regarding the relationships between ploidy-level distribution and a set of possible explanatory variables, we determined the polyploid frequency within each ecoregion (that is, the number of polyploid species out of the total number of species with ploidy inference in each region) and assigned each ecoregion a set of environmental and life history variables. Unless stated otherwise, the median value over all pixels within the ecoregion was calculated per variable.

1. Geographical coordinates and altitude.The centroid latitude of each ecoregion was extracted using ArcGIS[73]. This attribute served as a descriptive statistic and was used only in the GLM analysis. The median altitude of each ecoregion was extracted using the R package raster[70] together with the elevation amplitude, calculated as the difference between the 90th and 10th percentiles.
2. Climate. Bioclimatic attributes were downloaded from the WorldClim Global Climate Data[74] at ten arc-minutes resolution. For each ecoregion, the 19 BIOCLIM variables, representing the major temperature and precipitation attributes of a given locality, were extracted using the R package raster[70].
3. Change-in-climate. Estimates of palaeoclimatic conditions present at the LGM (~22,000 years ago) were obtained from WorldClim[74], and were used to calculate change-in-climate attributes as the difference between each of the 19 bioclimatic variables and their corresponding palaeoclimate attributes.
4. De-glaciation. The extent of de-glaciation since the LGM was estimated by comparing the present ice coverage of each ecoregion[75] to that inferred for the LGM[76]. The result is expressed as the percentage of de-glaciation per ecoregion. When both past and present glaciation extents were missing, the ecoregion was discarded from this analysis. Therefore, this attribute was examined for the 207 ecoregions that had some ice coverage data but not across all ecoregions.
5. NPP. NPP values, as estimated by Imhoff et al.[77], were extracted using the R package raster[70].
6. Lifeform composition. We assembled an extensive lifeform database to estimate the lifeform composition of each ecoregion (that is, the percentages of annual, perennial herb and woody species). This database combined information from six available resources (Encyclopaedia of Life[78], Index to Chromosome number in Asteraceae[79], International Legume database & Information Service http://www.ildis.org/, Missouri Botanical Garden http://www.missouribotanicalgarden.org/, World Checklist of Selected Plant Families[80] and a database compiled by Zanne et al.[81]). Additionally, we used text-mining procedures to extract lifeform information from thousands of botanical descriptions that appear in efloras[82]. The resulting database contains >188,000 taxa with growth form and/or life cycle information (of which > 140,000 have both) and is available in Supplementary Table 1. A full description of the database construction process is detailed in Supplementary Text 4.
7. Soil phosphorus. Phosphorus soil retention potential[83] was extracted per ecoregion. This layer comprises 12 levels of phosphorus availability that were converted to a discrete numerical scale.
8. Taxonomical composition (TaxComp). We categorized species into six clades of angiosperm as presented in Wood et al.:[5] higher monocots (commelinids), basal monocots (non-commelinid monocots), basal angiosperms, basal dicots (non-asterid + non-rosid dicots), dicots: core rosids, and dicots: core asterids. Genera were classified into each of the six clades following the classification of National Center for Biotechnology Information taxonomy[84], as obtained using the R package Geiger[85], and a manually curated look-up table (Supplementary Table 10). Genera for which we could not retrieve taxonomic classifications automatically were classified manually through eFloras.org[82]. Finally, the percentage of species from each taxonomical group was calculated for each ecoregion.
9. Human footprint. As a measure for anthropogenic disturbance of ecosystems (that is, the amount of land or sea necessary to support the consumption of local human activity, including population density, land transformation, accessibility and electrical power infrastructure), we used a data layer assembled by Sanderson et al.[86].
10. Species richness. The estimated number of vascular plant species in each ecoregion was obtained from Kier et al.[87]. These estimates of species richness were extracted from larger data sources than those used here, and thus include data also on species without ploidy inferences.

Species richness, rainfall amounts (except for precipitation seasonality; both in present and change-in-climate attributes), altitude and elevation amplitude were scaled up by 1 unit and then log transformed. Additionally, lifeform and taxonomy compositions, which are presented as percentages, were transformed using logit transformation with an adjustment of 0.5. De-glaciation percentage was transformed with an adjustment of 0.025.

Statistical analyses. The effect of each variable on the global distribution of ploidy levels was first examined independently using binomial GLM using a logit link function with the number of polyploid species considered as 'success' and the number of diploid species considered as 'failure'. Thus, this model weighs each ecoregion by its sample size (that is, inferred number of species with ploidy estimates). First, each predictor was tested independently, followed by a Bonferroni multiple testing correction. Based on the results obtained by GLM analysis, several variables were omitted from subsequent analyses: (1) difference in glaciation was discarded due to low data coverage, with only 207 ecoregions having sufficient data; and (2) altitude and elevation amplitude were omitted as their relationship with polyploid distribution was weak and non-significant in the initial GLM analysis. All variables were tested with their linear fit, except for latitude for which we used its squared term to fit its bipolar pattern.

Due to the high dependency within climate, TaxComp, lifeform and change-in-climate attributes (for example, the 19 BIOCLIM attributes strongly correlate with each other), we also used PCA to represent each group by a succinct set of orthogonal variables. For standardization, scaling and centring of the variables was performed before application of the PCA.

To compute the relative importance of each variable we used the R package relaimpo[47] (v2.2–2) with a model by Lindeman et al.[88], which calculates for each variable the average sequential R^2 obtained from all possible sets of predictors. We used the weighting option to provide each observation with its relative weight as determined through the number of species with ploidy inferences. To reduce the large number of attributes into a manageable number, most attributes were clustered into the following four groups: (1) climate—comprising the 19 BIOCLIM variables; (2) TaxComp—comprising the relative proportion of the six major plant groups; (3) lifeform—comprising the relative proportion of the three major distinct plant lifeform categories; and (4) change-in-climate—comprising the difference between each of the 19 BIOCLIM variables and their corresponding palaeoclim attributes. The remainder of the attributes (that is, species richness, NPP, altitude, elevation amplitude, phosphorus retention and anthropogenic disturbance) were included as additional independent variables.

Path analysis was used to compare the direct and indirect effect of the main factors affecting polyploid frequency and to evaluate the support of several alternative hypotheses. Furthermore, the cumulative effect, measured as the combined coefficients of all direct and indirect paths originating from each variable, was evaluated. In this analysis, we used the four most influential factors as inferred in the preceding analyses: climate, lifeform, TaxComp and species richness. First, we devised an a priori model in which climate was represented by BIO1 (models with other representations of climate, using either one or two principal components or by the two most influential BIOCLIM variables of temperature and precipitation, are detailed in Supplementary Text 5). This choice was based on the GLM and relative importance analyses, in which annual mean temperature had high explanatory power on polyploid frequency. All path analyses were conducted using the R package lavaan[89] (v0.5–23.1097). We note that climate, lifeform and TaxComp could be represented as composite variables (for example, a composite lifeform variable consisting of the percentages of perennial herb, woody and annual species in each ecoregion). Nonetheless, using composite variables resulted in a poor fit of the model to the data. We thus represented these predictors with one variable each: climate was represented by BIO1, lifeform by the percentage of perennial herbs and TaxComp by PC1, as this was found to serve as a better predictor than each of the individual taxonomical variables (see GLM results).

To test the robustness of all analyses to spatial autocorrelation between ecoregions, we used an iterative sampling procedure that generated 100 different datasets comprising 200 ecoregions each. In this approach, an ecoregion was chosen randomly and then added to the sampled set only if it shared <50% of its terrestrial border with the ecoregions that were already in the sampled set. The total terrestrial border length per ecoregion was obtained using the intersect tool in ArcGIS[73]. This procedure was iteratively performed until 200 ecoregions were sampled. The median and standard deviation of each bootstrap distribution were then used as summary statistics. The bootstrap procedure was performed for the GLM, relative importance and path analysis.

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

Data and code availability. The phylogenies, ploidy assignments, chromosome counts, lifeform data, predictors analysed in the current study and raw GBIF data are available in figshare (https://doi.org/10.6084/m9.figshare.c.4306004). Custom code related to the analyses can be found on GitHub (https://github.com/MayroseLab/GlobPoly).

Received: 30 March 2018; Accepted: 11 December 2018;
Published online: 29 January 2019

References

1. Jiao, Y. N. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–113 (2011).
2. Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L. & Vandepoele, K. The flowering world: a tale of duplications. *Trends. Plant. Sci.* **14**, 680–688 (2009).
3. Leitch, A. R. & Leitch, I. J. Genomic plasticity and the diversity of polyploid plants. *Science* **320**, 481–483 (2008).
4. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).
5. Wood, T. E. et al. The frequency of polyploid speciation in vascular plants. *Proc. Natl Acad. Sci. USA* **106**, 13875–13879 (2009).
6. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
7. Rieseberg, L. H. & Willis, J. H. Plant speciation. *Science* **317**, 910–914 (2007).
8. Ramsey, J. & Schemske, D. W. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* **29**, 467–501 (1998).
9. Soltis, D., Soltis, P. & Schemske, D. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* **56**, 13–30 (2007).
10. Levin, D. Minority cytotype exclusion in local plant populations. *Taxon* **24**, 34–43 (1975).
11. Barringer, B. C. Polyploidy and self-fertilization in flowering plants. *Am. J. Bot.* **94**, 1527–1533 (2007).
12. Levin, D. Polyploidy and novelty in flowering plants. *Am. Nat.* **122**, 1–25 (1983).
13. Ramsey, J. & Schemske, D. W. Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Syst.* **33**, 589–639 (2002).
14. Stebbins, G. L. *Chromosomal Evolution in Higher Plants* (Edward Arnold, London, 1971).
15. Spoelhof, J. P., Soltis, P. S. & Soltis, D. E. Pure polyploidy: closing the gaps in autopolyploid research. *J. Syst. Evol.* **55**, 340–352 (2017).
16. Ramsey, J. & Ramsey, T. S. Ecological studies of polyploidy in the 100 years following its discovery. *Philos. Trans. Royal Soc. B. Biol. Sci.* **369**, 1–20 (2014).
17. te Beest, M. et al. The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* **109**, 19–45 (2011).
18. Brochmann, C. et al. Polyploidy in arctic plants. *Biol. J. Linn. Soc.* **82**, 521–536 (2004).
19. Hagerup, O. Über polyploidie in beziehung zu klima, ökologie und phylogenie. *Hereditas* **16**, 19–40 (1931).
20. Martin, S. L. & Husband, B. C. Influence of phylogeny and ploidy on species ranges of North American angiosperms. *J. Ecol.* **97**, 913–922 (2009).
21. Bretagnolle, F. & Thompson, J. D. Gametes with the somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. *New Phytol.* **129**, 1–22 (1995).
22. De Storme, N. & Geelen, D. Sexual polyploidization in plants-cytological mechanisms and molecular regulation. *New Phytol.* **198**, 670–684 (2013).
23. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
24. Stebbins, G. L. Polyploidy, hybridization, and the invasion of new habitats. *Ann. Missouri Bot. Gard.* **72**, 824–832 (1985).
25. Parisod, C., Holderegger, R. & Brochmann, C. Evolutionary consequences of autopolyploidy. *New Phytol.* **186**, 5–17 (2010).
26. Stebbins, G. L. Polyploidy and the distribution of the Arctic-Alpine flora - new evidence and a new approach. *Bot. Helv.* **94**, 1–13 (1984).
27. Soltis, P. S. & Soltis, D. E. The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl Acad. Sci. USA* **97**, 7051–7057 (2000).
28. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).
29. Doyle, J. J. et al. Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461 (2008).
30. Guignard, M. S. et al. Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. *New Phytol.* **210**, 1195–1206 (2016).
31. Šmarda, P. et al. Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytol.* **200**, 911–921 (2013).
32. Johnston, A. E., Poulton, P. R., Fixen, P. E. & Curtin, D. Phosphorus: its efficient use in agriculture. *Adv. Agron.* **123**, 177–228 (2014).
33. Husband, B., Baldwin, S. & Suda, J. in *Plant Genome Diversity*, Vol. 2 (eds Greilhuber J., Jaroslav D. & Wendel J. F.) 255–276 (Springer, Vienna, 2013).
34. Rice, A. et al. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytol.* **206**, 19–26 (2015).
35. Stevens, P. Angiosperm Phylogeny Website v.13 (2012); http://www.mobot.org/MOBOT/research/APweb
36. Müntzing, A. The evolutionary significance of autopolyploidy. *Hereditas* **21**, 363–378 (1936).
37. Stebbins, G. L. Cytological characteristics associated with the different growth habits in the dicotyledons. *Am. J. Bot.* **25**, 189–198 (1938).
38. Engemann, K. et al. Patterns and drivers of plant functional group dominance across the Western Hemisphere: a macroecological re-assessment based on a massive botanical dataset. *Bot. J. Linn. Soc.* **180**, 141–160 (2016).
39. Ehrendorfer, F. in *Polyploidy: Biological Relevance* (ed. Lewis, W. H.) 45–60 (Springer, Boston, 1980).
40. Glick, L. & Mayrose, I. ChromEvol: assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Mol. Biol. Evol.* **31**, 1914–1922 (2014).
41. WWF. *Montane grasslands and shrublands* https://www.worldwildlife.org/biomes/montane-grasslands-and-shrublands.
42. Olson, D. M. et al. Terrestrial ecoregions of the world: a new map of life on earth. *Bioscience* **51**, 933–938 (2001).
43. Carr, G. D. in *Evolution and Speciation of Island Plants* (eds Stuessy, T. F. & Ono, M.) (Cambridge Univ. Press, 2007).
44. Oberlander, K. C., Dreyer, L. L., Goldblatt, P., Suda, J. & Linder, H. P. Species-rich and polyploid-poor: Insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. *Am. J. Bot.* **103**, 1336–1347 (2016).
45. Gustafsson, Å. Polyploidy, life-form and vegetative reproduction. *Hereditas* **34**, 1–22 (1948).
46. Levin, D. *The Role of Chromosomal Change in Plant Evolution* (Oxford Univ. Press, New York, 2002).
47. Grömping, U. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* **17**, 1–27 (2006).
48. Currie, D. J. et al. Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecol. Lett.* **7**, 1121–1134 (2004).
49. Kreft, H. & Jetz, W. Global patterns and determinants of vascular plant diversity. *Proc. Natl Acad. Sci. USA* **104**, 5925–5930 (2007).
50. Donoghue, M. J. A phylogenetic perspective on the distribution of plant diversity. *Proc. Natl Acad. Sci. USA* **105**, 11549–11555 (2008).
51. Andrew Jones, F., Sobkowiak, B., Orme, D. et al. in *Early Events in Monocot Evolution* (eds. Wilkin, P. & Mayo, S. J.) 99–117 (Cambridge Univ. Press, New York, 2011).
52. Huggett, R. J. in *Fundamentals of Biogeography* (ed. Gerrard, J.) 77–80 (Routledge, London, 2004).
53. Raunkiaer, C. *The Life Forms of Plants and Statistical Plant Geography* (Oxford Univ. Press, Oxford, 1934).
54. FitzJohn, R. G. et al. How much of the world is woody? *J. Ecol.* **102**, 1266–1272 (2014).
55. Ricklefs, R. E. & Renner, S. S. Species richness within families of flowering plants. *Evolution* **48**, 1619–1636 (1994).

56. Soltis, D. E. et al. *Phylogeny and Evolution of the Angiosperms: Revised and Updated Edition* (Univ. of Chicago Press, Chicago, 2018).
57. Moeller, D. A. et al. Global biogeography of mating system variation in seed plants. *Ecol. Lett.* **20**, 375–384 (2017).
58. Vamosi, J. C. & Vamosi, S. M. Key innovations within a geographical context in flowering plants: towards resolving Darwin's abominable mystery. *Ecol. Lett.* **13**, 1270–1279 (2010).
59. Glick, L., Sabath, N., Ashman, T. L., Goldberg, E. & Mayrose, I. Polyploidy and sexual system in angiosperms: Is there an association? *Am. J. Bot.* **7**, 1223–1235 (2016).
60. Drori, M. et al. OneTwoTree: An online tool for phylogeny reconstruction. *Mol. Ecol. Resour.* **18**, 1492–1499 (2018).
61. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, 36–42 (2013).
62. Smith, S. A. & Brown, J. W. Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.* **105**, 302–314 (2018).
63. Bennett, M. D. & Leitch, I. J. *Plant DNA C-values Database* (Royal Botanic Gardens, 2012); http://www.kew.org/cvalues/
64. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2013); http://www.R-project.org/
65. Zizka, A. CoordinateCleaner: Automated Cleaning of Occurrence Records from Biological Collections (2018); https://CRAN.R-project.org/package=coordinatecleaner
66. Wickham, H. Tidyverse: Easily Install and Load the 'Tidyverse' (2017); https://CRAN.R-project.org/package=tidyverse
67. Chamberlain, S. and Boettiger C. R Python, and Ruby clients for GBIF species occurrence data. *PeerJ PrePrints* (2017); https://CRAN.R-project.org/package=rgbif
68. South, *A.* rnaturalearthdata: World Vector Map Data from Natural Earth Used in 'rnaturalearth' (2017); https://CRAN.R-project.org/package=rnaturalearth
69. Meyer, R. S., DuVal, A. E. & Jensen, H. R. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* **196**, 29–48 (2012).
70. Hijmans, R. J. Raster: Geographic Data Analysis and Modeling, R package v.2.2-12 (2014); http://CRAN.R-project.org/package=raster
71. Rowlingson, R. B. and T. K. and B. Rgdal: Bindings for the Geospatial Data Abstraction Library (2014); http://cran.r-project.org/package=rgdal
72. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).
73. ArcGIS Desktop: Release 10 (Environmental Systems Research Institute, 2011); http://www.esri.com
74. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
75. Raup, B. H. et al. Global Land Ice Measurements from Space (GLIMS) Database at NSIDC. *AGU Fall Meeting Abstracts* **1**, 837 (2003).
76. Ehlers, J., Gibbard, P. & Hughes, P. in *Quaternary Glaciations – Extent and Chronology*, Part IV (Elsevier, Oxford, 2011).
77. Imhoff, M. L. et al. Global patterns in human consumption of net primary production. *Nature* **429**, 870–873 (2004).
78. Wilson, E. O. The encyclopedia of life. *Trends Ecol. Evol.* **18**, 77–80 (2003).
79. Watanabe, K. *Index to Chromosome Numbers in Asteraceae* http//www.asteraceae.cla.kobe-u.ac.jp/index.html (2002).
80. WCSP. *World Checklist of Selected Plant Families* (Royal Botanic Gardens, Kew, 2016); http://apps.kew.org/wcsp/ https://doi.org/10.1163/_q3_SIM_00374
81. Zanne, A. E. et al. Three keys to the radiation of angiosperms into freezing environments. *Nature* **506**, 89–92 (2014).
82. Brach, A. R. & Song, H. eFloras: new directions for online floras exemplified by the Flora of China Project. *Taxon* **55**, 188–192 (2006).
83. Batjes, N. H. *Global Distribution Of Soil Phosphorus Retention Potential* http://www.isric.org/data/global-assessment-soil-phosphorus-retention-potential (2011).
84. Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **37**, D5 (2009).
85. Harmon, L. J, Weir, J. T, Brock, C. D, Glor, R. E. & Challenger, W. GEIGER: investigating evolutionary radiations. *Bioinformatics* 129–131 (2008).
86. Sanderson, E. W. et al. The human footprint and the last of the wild. *Bioscience* **52**, 891 (2002).
87. Kier, G. et al. Global patterns of plant diversity and floristic knowledge. *J. Biogeogr.* **32**, 1107–1116 (2005).
88. Lindeman, R., Merenda, P. & Gold, R. *Introduction to Bivariate and Multivariate Analysis* (Scott Foresman, London, 1980).
89. Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36 (2012).
90. Schluter, D. Estimating the form of natural selection on a quantitative trait. *Evolution* **42**, 849–861 (1988).

Acknowledgements

We thank S. P. Otto for instructive comments and G. Berman and U. Roll for technical assistance. This work was supported by the Israel Science Foundation (grant number 961/17) to I.M., by the Czech Science Foundation (grant number P505 14–30313S) to P.S. and by a fellowship provided by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University to A.R.

Author contributions

A.R. and I.M. designed the study. A.R., M.N. and L.G. collected and processed the data. M.D. and N.S. reconstructed phylogenies and provided ploidy-level inferences. A.R. and P.S. analysed data. A.R., P.S., S.M. and J.B. analysed results. A.R. and I.M. drafted the manuscript. All authors provided comments and helped to improve the manuscript. I.M supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0787-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to I.M.

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

nature research

Corresponding author(s): Itay Mayrose

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	□	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All software packages (including details, such as version numbers) are provided in the Methods section.
Data analysis	All software packages (including details, such as version numbers) are provided in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The phylogenies, ploidy assignments, chromosome counts, lifeform data, predictors analysed in the current study, and raw GBIF data are available in figshare (DOI 10.6084/m9.figshare.c.4306004). Custom code related to the analyses can be found on GitHub (https://github.com/MayroseLab/GlobPoly).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We present the global map of polyploid frequency and describe the main ecological and life-history attributes affecting this distribution using several statistical tests (GLM, PCA, relative importance, and SEM).
Research sample	Initial data collection was performed on all available spatial data for angiosperm.
Sampling strategy	All angiosperms. Various filtering processes are described in the Methods section.
Data collection	Spatial data was downloaded from GBIF and eco-climatic attributes were collected from WorldClim. Chromosome numbers were retrieved from CCDB. Sequence data for phylogeny inference were retrieved from NCBI.
Timing and spatial scale	Data from GBIF were downloaded on October the 20th 2016.
Data exclusions	We omitted from all analyses ecoregions without sufficient data to avoid the distortion of results due to poorly-sampled regions.
Reproducibility	We tested for robustness of results using bootstrap analyses in the statistical analyses (GLM, relative importance and SEM). Bootstrap results are reported in the Supplementary Materials.
Randomization	The bootstrap sampling included randomly choosing ecoregions that do not share more than a certain cutoff of shared border together.
Blinding	Blinding is not relevant to this study as we used all available data.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	□ Unique biological materials
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging