Genome size mediates the effect of environmental filtering in determining plant β-diversity across temperate grasslands

Hai-Yang Zhang[1], Xiaotao Lü[2], cunzheng wei[3], Jeff Powell[4], Xiaobo Wang[5], Dingliang Xing[6], and Xing-Guo Han[7]

[1]Western Sydney University Hawkesbury Institute for the Environment
[2]Institute of Applied Ecology, Chinese Academy of Sciences
[3]Institute of Botany Chinese Academy of Sciences
[4]Hawkesbury Institute for the Environnment
[5]Lanzhou University College of Pastoral Agriculture Science and Technology
[6]School of Ecological and Environmental Sciences, East China Normal University
[7]State Key Laboratory of Vegetation and Environmental Change

May 12, 2021

Abstract

Elucidating mechanisms underlying community assembly and biodiversity patterns is central to ecology and evolution. Genome size (GS, i.e. nuclear DNA content) determines species' capacity to tolerate environmental stress or to exploit new environments and therefore potentially drive community assembly. However, its role in driving β-diversity (i.e., the site-to-site variability in species composition) remains unclear. We measured GS for 169 plant species and investigated their occurrences within plant communities across 52 sites spanning a 3200-km transect in the temperate grasslands of China. We found environmental factors showed larger effects on β-diversity of large-GS than that of small-GS species. Community weighted mean GS increased with mean annual precipitation, soil total nitrogen and phosphorus concentrations, but decreased with mean annual temperature, suggesting a negative selection against species with large GS in resources-limited or warmer climates. These findings highlight the roles for GS in driving community assembly and predicting species responses to climate change.

Introduction

Disentangling the drivers of β-diversity (the site-to-site variability in species composition) provides insights into the processes that govern community assembly (Chase 2010; Kraft *et al.* 2011). β-diversity can arise from community assembly processes involving deterministic selection, when environmental heterogeneity creates different niches that shape the occurrences of species in a community, and stochastic aspects related to dispersal limitation and ecological drift (Laliberté*et al.* 2014; Mori *et al.* 2018). Interspecific variations in plant traits determine the capacity for individuals to grow, reproduce and disperse within and among environments and, therefore, play important roles in determining the relative importance of deterministic selection (McGill *et al.* 2006; Blonder 2018). Traits that have been considered thus far are largely related to ecological strategy axes along which plant species vary in their abilities to acquire and allocate resources to different aspects of their life histories (Adler *et al.* 2014; Salguero-Gómez *et al.* 2016). As a fundamental trait that significantly varies within angiosperm (2400-fold) and precisely correlates with diverse phenotypic characters at cellular and organismal level, genome size (the amount of DNA contained in the nucleus), has received relatively little attention in the context of its role in community assembly. Genome size is generally constant within an organism, which has functional consequences for species' environmental tolerances, capacity for dispersal and interactions with other species (Knight & Ackerly 2002; Herben *et al.* 2012). The potential impact of

plant genome size on community assembly processes is starting to be recognized (Greilhuber & Leitch 2013; Pellicer *et al.* 2018) but to date, its importance in shaping β-diversity has not been tested empirically.

A key determinant of β-diversity is environmental filtering. The strength of the relationship between local environmental conditions and species environmental requirements affects the establishment and persistence of species (Laliberté *et al.* 2014; van Breugel*et al.* 2019). Environmental filtering is hypothesized to differ for plants with large vs. small genome size. First, plants with small genomes grow faster due to short cell cycle duration and are subject to less material costs for packing DNA, allowing them to achieve optimal growth across a wider range of environments (Knight & Beaulieu 2008; Hessen *et al.* 2010). Second, according to the 'large genome constraint hypothesis', the optimal growth for plants with large genomes is only achievable under stress-free conditions (Knight *et al.*2005) or high resource availability. It has been hypothesized that in nutrient-depleted soils there will be selection for species with small genomes as a way to reduce the biochemical cost of synthesising DNA, which is rich in nitrogen (N) and phosphorus (P) (Leitch & Leitch 2008; Hessen *et al.* 2010). Indeed, a recent study showed that plants with large genomes became more dominant when there are more available nutrients such as N and P (Guignard *et al.* 2016). Thus, we expect that environmental filtering would have stronger effects on plant species with large genomes than for those with small genome size.

To assess the roles that genome size plays in plant community assembly, we used data from 520 plant communities in 52 sites (10 quadrats of plant communities per site) along a 3200-km transect in the temperate grasslands of northern China (Fig. S1). We then measured plant genome size (the amount of DNA in a gamete nucleus or 1C-value, for details see Methods; Table S1-2) for 169 herbaceous species occurring along the transect (Fig. S2). Generalized dissimilarity models [GDMs (Ferrier*et al.* 2007; Fitzpatrick *et al.* 2013)] were used to quantify the effects of genome size, environmental variation and geographical distance on β-diversity along the entire gradient. We also estimated the importance of genome size as a continuous trait in driving species distribution with two model approaches: (i) a joint species distribution model (Hierarchical Modelling of Species Communities [HMSC (Ovaskainen *et al.* 2017)] and (ii) phylogenetic generalized linear mixed model [PGLMM (Li *et al.* 2020)]. Both approaches can simultaneously model the environmental responses of multiple species accounting for shared habitats and evolutionary histories.

We hypothesised that the relative importance of the ecological processes (i.e. environmental filtering and dispersal limitation) in driving spatial turnover of plant communities and species would differ between species with larger- and small- genomes. We also hypothesised that, due to higher resource requirements, the plant species with larger genomes would be more constrained by environmental effects compared to species with smaller genomes.

Materials and Methods

Study sites

The study was conducted across a 3200 km transect of northern China's grasslands (Wang *et al.* 2017), extending from the Xinjiang Uygur Autonomous Region in the west to eastern Inner Mongolia (83.45° E to 120.36° E, 42.89° N to 49.19° N; Fig. S1). There are four vegetation types along this transect, including alpine-, desert-, typical- and meadow-steppe from west to east. Dominant soil types are classified as aeolian chestnut soil in the east to brown calcic soil, grey desert soil and sandy soil to the west.

Plant community and soil sampling

We conducted transect sampling during July and August in 2012 near the period of the peak aboveground biomass. To collect plant samples at the same period of phenology, we started to sample from the west to the east along this transect since this coincides roughly with a decreasing temperature trend (i.e., delayed growing peak period). A total of 60 sites with an interval of 50˜100 km was investigated along the transect, but only data from 52 sites were selected in the final analysis (details see below; Fig. S1). At each site, ten 1 m x 1 m quadrats were established. For each quadrat, we clipped the aboveground tissues of living plants, sorted these by species and stored them in paper bags. Soil samples from each quadrat were collected with

five soil cores (2.5 cm diameter x 10 cm depth) from the upper 10 cm layer. Five soil cores were collected from four corners 10 cm away from the edge plus one from the centre of the quadrat. Soils from the five soil cores were combined for each quadrat and sieved through a 2.0 mm mesh to remove roots and rocks, homogenized by hand, and preserved for subsequent chemical analysis.

Environmental variables

At each site, spatial geographical coordinates and altitudes were recorded using a handheld GPS (eTrex Venture, Garmin, Olathe, Kansas, USA). Climate attributes, including mean annual temperature (MAT) and precipitation (MAP) of each sampling site were obtained from the global climate data set WorldClim 2.0 (1-km spatial resolution) (Fick & Hijmans 2017). Total N concentration of soil samples was determined using wet oxidation and a modified Kjeldahl procedure, and total P concentration was measured by colorimetric analysis with ammonium molybdate and persulfate oxidation (Murphy & Riley 1962). We obtained altitude estimates from each site using GPS since climate conditions, resource availability and other environmental conditions vary with altitude (Korner 2007).

Plant species selection for genome size measurements

The name of all plant species was standardized according to The Plant List (version 1.1; www.theplantlist.org). We recorded 286 herbaceous species in the transect; of these we were able to obtain measurements of genome size from 169 species during subsequent visits to representative study sites belonging to the Chinese Grassland Long-term Research Station (at least one site for each grassland type: alpine grassland, desert steppe, typical steppe and meadow steppe). These stations represented the typical vegetation and species pool for each grassland type and were convenient to re-visit. Samples were collected to measure genome sizes during the growing seasons of 2017-2019 (from July to September). We sampled plant species that occurred at each grassland type of the transect, focusing primarily on (but not limited to) the more common.

For each species, three individuals were selected and leaf samples from each individual plant were measured for the genome size and the averaged value was used. For each individual, leaves from at least three fresh samples were analysed by flow cytometer (BD LSRFortessa, USA) (Doležel*et al.* 2007). The pg of DNA per nucleus in each peak is estimated by comparing measurements to a known internal reference standard, which consisted of seeds of species obtained from Centre of Plant Structural and Functional Genomics of the Institute of Experimental Botany, AS, CR in 2017, including *Kerria japonica* (1C = 0.50 pg),*Lycopersicon esculentum* (1C = 0.98 pg), *Glycine max* (1C = 1.25 pg), *Zea mays* (1C = 2.72 pg), *Viburnum dilatatum* (1C = 3.97 pg), *Pisum sativum* (1C = 4.55 pg), *Secale cereale*(1C = 8.10 pg), *Hosta plantaginea* (1C = 11.33 pg). The mean, standard deviation, and individual measurements of the studied species and the used internal standards are reported in Table S1 (Doležel*et al.* 2007).

Among the 169 species sampled for genome size measurements, there were 24 species that occurred more than once across the resampling sites, where we could check intraspecific variation of ploidy levels (Table S2). Most of these species showed quite stable 1C DNA content values across the transect, except *Artemisia frigida* and *Agropyro cristatum* . 1C DNA value (pg) for *A. frigida* was 3.25 in meadow steppe while 5.26 in the other three grassland types, indicating the potential for polyploidy in those other grassland types. Similarly, the value for *A. cristatum* was 6.84 in the meadow while 13.30 in other grassland types. Thus, when scaling up to the transect plant community, we chose a different value for each of the two species according to their occurring grassland type.

For the species that were able to obtain genome size measurements, we assessed their contributions to biomass and species richness of the total plant community at each site. Sites were included in further analyses if, for that specific site, the species that had genome size values contributed to more than 85% of biomass and richness (52 sites of the 60 meet this requirement). Thus, for the final analysis, we used data for plant communities containing 169 species from 52 sites along the transect. Overall, the plant genome size (1C DNA content) varied 260-fold from the smallest *Astragalus scaberrimus* (0.12 pg) to the largest *Allium ramosum* (31.5 pg), with a median and mean of 1.57 and 4.02 pg, respectively (Fig. S2; Table S1).

Phylogenetic trees

We constructed the plant phylogenetic tree using an updated version of the mega-phylogeny published by (Zanne *et al.* 2014) as the backbone. For those genera and species in our dataset that were absent from the meta-phylogeny, we used S.PhyloMaker (available at https://github.com/jinyizju/) to add them to their respective families (in the case of genera) and genera (in the case of species) in the mega-phylogeny under Scenario 3 (Qian & Jin 2016). The polytomies were resolved by the multi2di function in the ape package in R. Finally, our phylogenetic tree had 169 species (following APG IV, Fig. S2). We estimated phylogenetic signals for genome size by employing Blomberg's K and Pagel's λ tests, with the significance being estimated with randomization and likelihood ratio tests in the package phylosignal in R (Keck *et al.* 2016) (Fig. S2).

Data analysis

To match with the site-level environmental data, we pooled the plant community collected from all quadrats from each study site to obtain the site-level species pool (10 field plots per site) for each of the 52 sites. Similarly, the corresponding soil property data were averaged for each study site. Principal components analysis was applied to the bioclimatic variables for the sites and the five studied environmental variables (i.e. MAP, MAT, altitude, soil N and P), which PC1 and PC2 explained 88% of variation among sites (Fig. S1b-g).

Beta diversity partitioning

Spatial variation in plant community composition was estimated from two components, species replacement (turnover component, β_{turnover}) and changes in species richness (nestedness component, $\beta_{\text{nestedness}}$), which together contribute to the total β-diversity (β_{total}). To test for effects on β-diversity that are independent of species richness differences, we first used the Simpson index of dissimilarity to observe how β_{turnover} changed along the environmental gradient. We also calculated the β_{total} with the Sørensen index to observe changes of the total β-diversity. We calculated the two β-diversity components via the betapart package in R (Baselga & Leprieur 2015). Dissimilarity was computed 1000 times for randomly sampled subsets of 10 sites (command beta.sample in R package betapart), and the resulting distributions of β_{turnover} and β_{total} values across the 1000 samples were used to empirically assess the proportion of β_{turnover} from the total β-diversity. All these analyses were based on species presence/absence data.

Generalized Dissimilarity Modelling

We used the GDM approach to analyze β-diversity patterns along environmental gradients. GDMs are widely used to identify important environmental drivers for β-diversity and to test the independent significance of these drivers (using permutation tests). The advantage of GDMs is that they can allow the rate of compositional turnover to vary along an environmental gradient, i.e. allow the relationships between dissimilarity and distance to be nonlinear (Ferrier *et al.* 2007; Fitzpatrick *et al.* 2013). The environmental matrix in our study included habitat variables (MAT, MAP, altitude, soil N, and soil P) and geographical distances (i.e. spatial distance from latitude, and longitude) between sites.

We plotted the partial effect of each predictor against the level of a given predictor to visualize the results of each GDM (holding all other predictors constant). The maximum height of the line shows how the relative importance of the studied predictor in explaining the variation of β-diversity in the model. The shape of the line shows how β-diversity varies along each environmental or spatial gradient, i.e. how the effect of a given predictor on β-diversity varies at a given level of that predictor. Furthermore, we also determined the proportion of deviance uniquely attributable to environment or distance. We did this by comparing the deviance explained by a GDM containing all of the variables and a GDM with all variables except environment or distance, respectively. The unique deviance explained by environment or distance was calculated as the difference in deviance explained by these models. We then converted this to a percentage by dividing the deviance explained by the full GDM. GDMs were fitted to the β-diversity for turnover (β_{turnover}) and total component (β_{total}), separately, using the gdm function in the gdm library (Manion *et al.* 2017). The results were generally similar for β_{turnover} vs β_{total} (Fig. 1-2 vs. Fig. S3-4). This was understandable as

the β-diversity of plant assemblages was dominated by the turnover component (contributing 90% to β_{total}) across the transect (Fig. S5).

Modelling species responses with HMSC

We applied HMSC from the Hmsc R package (Tikhonov *et al.* 2020) to fit a joint species distribution model to plant community data simultaneously including information on traits (genome size), environmental covariates, and phylogenetic relationships in a single model (Pollock *et al.* 2012; Warton *et al.* 2015). We included MAP, MAT, altitude, soil N, and soil P as fixed effects (i.e., predictors) and used the sampling site as the sampling unit and the random effect. HMSC estimates species responses (slope parameters) to environmental covariates and uses these responses as the species' functional niche to estimate the strength and direction in which these niches moderate multiple species responses to environmental filtering. To account for the phylogenetic correlations among all the species, we included a phylogenetic correlation matrix (construction details of the phylogenetic trees see above) in the model's covariance structure.

We fitted the model to the plant community (with a probit-link for the presence/absence data) with Bayesian inference, using the posterior sampling scheme (Ovaskainen *et al.* 2016). We used four Markov Chain Monte Carlo (MCMC) chains, each of which consisted of 15,000 iterations, out of which we discarded the first 5,000 as the burn-in and thinned the remaining by 10 to yield in total 1000 posterior samples per chain. We assessed the convergence of the MCMC chains by examining the distribution of the potential scale reduction factor over the parameters that measure the responses of the species to the fixed effects included in the model. The MCMC convergence of the HMSC models was satisfactory as the potential scale reduction factors (psrf) are close to one (the psrf of the ?-parameters that measure the response of species to environmental covariates (Ovaskainen *et al.* 2017) were on average 1.16). We examined the explanatory and predictive powers of the HMSC models through species-specific Tjur's R^2 values (Tjur 2009). Twofold cross-validation was performed to assess the predictive power of the model.

Based on the output of HMSC, we first characterized species' responses to each of the environmental variables as a mean parameter and its 95% posterior probability. Second, we predicted the community weighted mean genome size based on 169 species considered in the model as responses to the included environmental covariates. Third, we used HMSC to capture the species-to-species associations (positive or negative co-occurrences) with latent variables. We considered statistical support for these associations based on 90% posterior probabilities. Association patterns were classified as positive or negative, with this result used to estimate the frequency of two taxa co-existing (or not) compared to the frequency expected based on shared habitats or random occurrences.

Phylogenetic Generalized Linear Mixed Model (PGLMM)

Ignoring phylogenetic relationships among species when estimating relationships between traits and environments can lead to inflated type I errors (false positives) (Li & Ives 2017). To address this, we also fitted a PGLMM to account for the confounding effects of shared evolutionary histories among species using the R package *phyr* (Li*et al.* 2020) to check whether environment, genome size and their interactions can affect occurrence for the entire ecological community. We used the Bayesian framework with the default uninformative INLA priors (Rue *et al.* 2009). We scaled explanatory variables to have a mean of zero and standard deviation of one to ensure comparable model effect sizes across variables. All analyses were conducted in R software 3.6.2 (R Core Team 2019).

Results and discussion

Γενομε σιζε ανδ β-διερσιτψ

Pairwise species turnover among communities (β_{turnover}) showed a positive relation with increasing environmental dissimilarity (Fig. 1a), in line with many previous findings (Robroek *et al.*2017; Mori *et al.* 2018). Plant species within each community were then divided into two subsets based on whether their genome was greater or less than 2.5 pg (the median genome size for angiosperms globally (Leitch & Leitch 2013)) and variation in community composition partitioned to environmental variables and geographic distance among

communities. Consistent with our hypothesis, environmental variables explained more variation in β_{turnover} for species with large genomes than that for those with small genomes (Fig. 1b, d-h; Fig. 2a), while geographic distance explained more variation in β_{turnover} for species with small genomes than those with large genomes (Fig. 1b-c). Environmental predictors (altitude, soil N, soil P, and MAP [marginally]) explained more variation in β_{turnover} for species with large genomes.

These results indicate that the relative importance (or effects) of environmental heterogeneity and geographic distance (which can indicate, in part, dispersal limitation) on plant β-diversity depend greatly on genome size of plants. We assessed the sensitivity of these results to different thresholds in genome size (median and mean values for the present study) [1.57 pg, 4.03 pg; Fig. 2] or globally for angiosperms [2.5 pg, 5.90 pg] (Leitch & Leitch 2013) for GDMs that used either environment, geographical distance, or both as predictor variables. We observed qualitatively similar results, with the environment explaining the most variation for species with large genomes and geographic distance explaining the most variation for species with small genomes (Fig. 2a-c). The thresholds differed mainly in how variation was partitioned for species classified to the large genome size category, with environment explaining more variation and geographic distance explaining less variation as the threshold increased, indicating that the species with the largest genomes were most strongly impacted by environmental filtering (Fig. 2).

We found spatial distances explained less variation in β-diversity for plants with large genomes. While we are unable to determine the underlying mechanisms, our results can be potentially explained by the competition-dispersal trade-off framework. This framework is often invoked to explain species coexistence (Levins & Culver 1971; Hastings 1980; Yawata *et al.* 2014), where species with an inferior ability to use resources can compensate by migrating more frequently and over a longer distance (Yawata *et al.* 2014; Gude *et al.* 2020). Here we showed evidence that large genome size plants are likely inferior competitors for resources given their decreased frequencies when resources were more limiting (Fig. 3). Under this framework, those species with large genomes should have greater fitness if they can disperse their offspring further to avoid competition and increase their chances of finding resource-rich environments. Our results combine genome size into this framework and suggest hypotheses for how plants with large- and small-genome sizes can co-exist. This will have important implications in biodiversity conservation under climate change and habitat fragmentation given that these pressures are likely to differentially impact species depending on their genome size.

Genome size and species distribution: HMSC model

The HMSC model showed a good fit to the data, with the mean Tjur R^2 being on average 62% for explanatory power and 14% for predictive power. We observed both positive and negative associations with species occurrences for each variable (Fig. 3a). To check whether phylogenetic relatedness matters for their occurrence in response to environments, HMSC model can indicate the strength of the phylogenetic signal based on parameter ϱ (varying from 0 to 1) (Ovaskainen *et al.* 2017). In our HMSC model, ϱ was 0.08 in median, with the 95% confidence interval between 0.01 and 0.21, suggesting that there was a phylogenetic influence.

We used species occurrences as predicted by the HMSC model to calculate the community weighted mean (CWM) of genome size and mapped this against environmental covariates (Fig. 3b). We used predicted occurrences because using CWMs based on raw occurrence data in regressions can be subject to type I error (Miller *et al.* 2019). Here we used the model-based approach to analyse trait–environment associations with community data, which maximises power and information contained in the data (Brown *et al.* 2014; Miller *et al.* 2019). For example, our HMSC model took phylogenic relativeness into account to model species occurrences. CWM of genome size was greater in communities from those sites with higher levels of precipitation and soil N and P availability, supporting our hypothesis that alleviating resource (water and nutrient) limitation would favour species with large genomes. Such result was consistent with previous studies that found that species with large genome size became more dominant when soil nutrients (N and P) were added (Šmarda *et al.* 2013; Guignard *et al.* 2016).

Water is the most limiting resource in the studied temperate grasslands (Bai *et al.* 2004), here we showed that reduced precipitation constrained the abundance of plants with large genomes within the community. It has been shown previously that genome size was positively correlated with cell size while negatively correlated with tissue elasticity (Castro-Jimenez *et al.* 1989). Increased elastic tissue and smaller cell size are important for turgor maintenance in plants in drought-stressed environments, suggesting that natural selection should favor plants with smaller genomes in xeric environments (Castro-Jimenez*et al.* 1989). Similarly, a positive correlation was found in Californian angiosperms between 2C DNA content and annual precipitation (Knight & Ackerly 2002). However, inconsistent results have also been observed in the context of polyploidy. Previous study observed that polyploid species have a higher drought tolerance due to their greater hydraulic safety than their congeneric diploid species (Zhang *et al.* 2017). In contrast, analysis of leaf anatomy in fireweed (*Chamerion angustifolium*) revealed morphological adjustments for tolerating water deficiency in diploids in the form of closely packed mesophyll cells and small conduits in the midvein, making diploid *C. angustifolium* more tolerant to drought than hexaploid plants (Guo *et al.* 2016). In comparing with others that focus generally at the species level, our work was the first to show that, depending on genome size, water stress can shift the plant dominance at the community level.

We also showed that MAT was negatively correlated with CWM of genome size, possibly because low temperatures favour plants with large genomes (Grime & Mowforth 1982; MacGillivray *et al.* 1995). For instance, species with large genomes gain an advantage in frost resistance (MacGillivray *et al.* 1995). Altitude was a poor predictor for CWM of genome size, with variation among sites better explained by other properties of those sites; for example, alpine grasslands with low MAT were occupied by more species with large genomes. A similar altitude-genome size pattern has been reported previously (Albach & Greilhuber 2004). Together, results from HMSC model support the "large genome constraint hypothesis" – that species with large genomes are less tolerant of environmental stress and resource limitation (Knight & Ackerly 2002; Knight *et al.* 2005).

Genome size and species distribution: PGLMM

To account for shared evolutionary history for the whole community over the transect, we used a PGLMM (Ives & Helmus 2011) to estimate the effects of genome size and environment in shaping species' distributions. For the responses of species occurrence to environmental changes, we found similar results as the prediction from HMSC, i.e. MAP and soil N generally have positive while MAT had negative effects on species occurrences, although only the effect of MAP was statistically significant (Fig. 4a). We also found that genome size had a significant positive effect on species occurrence, suggesting species with larger genomes tended to have a higher occurrence probability. Interestingly, there was a negative interaction between genome size and MAT (Fig. 4b), suggesting that increasing MAT tends to select species with smaller genomes. Because MAT negatively affected species occurrence, one can expect that species with large genomes would be excluded as MAT increases, in line with the HMSC prediction, supporting that species with a large genome can be more sensitive to rising temperature. Our results suggest that even after removing the cofounding effect of phylogeny, genome size still affected species occurrences.

Genome size and species co-occurrence

Biotic interactions among species that share a common set of environmental tolerances, in whole or in part, can also be an important driver of variation in species composition (Cornell & Lawton 1992; Tilman 1994; Stephens *et al.* 2020). We evaluated the direction and magnitude of these species' interactions, representative of facilitative and competitive interactions among species, via the residual species-to-species association matrices derived from the HMSC analysis (Fig. S6). We found that the frequencies of both positive and negative interactions (90% posterior probability) were greater when both members of the interacting pair had large genomes, compared with when one or both members possessed a small genome. In addition, this difference became larger as the threshold used to differentiate large and small genome size increased (Fig. S6). In other words, both facilitation and competition were more likely to occur among plants with large genomes while those with small genomes were less likely to engage in either facilitation or competition. An outcome of these interactions could be that facilitation among species with large genomes could lead

to communities dominated by those species in resource-rich environments, while competition among those species under resource-limiting conditions may be responsible for the loss of species with large genomes. Regardless, our results indicate that the importance of biotic filtering in structuring β-diversity was higher for species with larger genomes.

Genome size, polyploids and other traits

Variation in genome size arises as a consequence of the amplification/deletion of transposable elements (Lee & Kim 2014) or polyploidy (i.e., whole-genome duplication), the latter often being accompanied by interspecific hybridization (allopolyploidy) (Leitch*et al.* 2008). Polyploid individuals usually experience faster niche differentiation (Baniaga *et al.* 2020) because they are better at exploiting different environments where competition with the parent species is relaxed (Weiss-Schneeweiss *et al.* 2013). Meanwhile, polyploids allow species to expand their geographic ranges (Otto & Whitton 2000; Sheth *et al.* 2020). For example, allopolyploids that form via hybridization between two species often exhibit asexual reproduction and thereby have broader environmental tolerances via reproductive assurance, greater adaptive potential due to mutations, and greater dispersal distance over a large spatial scale. As our study mainly focused on genome size per se, additional studies are needed to determine whether observed effects of genome size in driving β-diversity arise immediately as a result of polyploidization or changes of transposable elements (Sheth *et al.* 2020).

Genome size is a fundamental trait of organisms that evolved over long timescales and correlates with phenotypic traits at the cellular or whole-plant level, including physiological processes (e.g. photosynthesis) (Roddy *et al.* 2020), nutrient acquisition (Forrester *et al.* 2020), and plant growth rate (Knight *et al.* 2005; Herben *et al.* 2012; Greilhuber & Leitch 2013). It should be noted that selection pressures act on genome size alone and/or through functional relationships between genome size and other traits and that either of these mechanisms can contribute to the observed relationships with environmental factors. As a fundamental trait in driving many phenotypic traits at the species level and the emergence of new species over evolutionary time scales, we conclude that that genome size should play a key role in driving community assembly, species distribution, and species interactions.

Conclusion

One of the central questions in macroecology and biogeography is to understand how species composition differs among different sites. Equally important is to understand how genomic evolution can drive the divergence of species and the patterns of biodiversity. Our results indicate genome size can be used to bridge the ecological and evolutionary processes together for a deeper understanding of how biodiversity evolves over a large spatial scale (Segraves 2017; Pellicer*et al.* 2018). Here we demonstrate the fundamental role of genome size in the plant community assembly processes, including environmental and biotic filtering and potentially dispersal limitation. This also opens a new avenue to understand how genome size contributes to community shifts along other gradients of environmental change (including human-induced climate change) (Marquet *et al.* 2019) and to gather more mechanistic and predictive insights into community assembly processes. For instance, given that climate models consistently predict higher temperatures and increased aridity in the temperate steppe (Day *et al.* 2018), species with large genomes may be more threatened by global change and should, therefore, receive more attention in conservation efforts.**References**

Adler, P.B., Salguero-Gómez, R., Compagnoni, A., Hsu, J.S., Ray-Mukherjee, J., Mbeau-Ache, C. *et al.* (2014). Functional traits explain variation in plant life history strategies. *Proc. Natl. Acad. Sci. U.S.A.* , 111, 740-745.

Albach, D.C. & Greilhuber, J. (2004). Genome size variation and evolution in Veronica. *Ann. Bot.* , 94, 897-911.

Bai, Y., Han, X., Wu, J., Chen, Z. & Li, L. (2004). Ecosystem stability and compensatory effects in the Inner Mongolia grassland. *Nature* , 431, 181-184.

Baniaga, A.E., Marx, H.E., Arrigo, N. & Barker, M.S. (2020). Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecol. Lett.* , 23, 68-78.

Baselga, A. & Leprieur, F. (2015). Comparing methods to separate components of beta diversity. *Methods Ecol. Evol.* , 6, 1069-1079.

Blonder, B. (2018). Hypervolume concepts in niche-and trait-based ecology. *Ecography* , 41, 1441-1455.

Brown, A.M., Warton, D.I., Andrew, N.R., Binns, M., Cassis, G. & Gibb, H. (2014). The fourth-corner solution–using predictive models to understand how species traits interact with the environment. *Methods Ecol. Evol.* , 5, 344-352.

Castro-Jimenez, Y., Newton, R., Price, H. & Halliwell, R. (1989). Drought stress responses of Microseris species differing in nuclear DNA content. *Am. J. Bot.* , 76, 789-795.

Chase, J.M. (2010). Stochastic community assembly causes higher biodiversity in more productive environments. *Science* , 328, 1388-1391.

Cornell, H.V. & Lawton, J.H. (1992). Species interactions, local and regional processes, and limits to the richness of ecological communities: a theoretical perspective. *J. Anim. Ecol.* , 61, 1-12.

Day, J.A., Fung, I. & Liu, W. (2018). Changing character of rainfall in eastern China, 1951–2007. *Proc. Natl. Acad. Sci. U.S.A.* , 115, 2016-2021.

Doležel, J., Greilhuber, J. & Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* , 2, 2233.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* , 13, 252-264.

Fick, S.E. & Hijmans, R.J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* , 37, 4302-4315.

Fitzpatrick, M.C., Sanders, N.J., Normand, S., Svenning, J.-C., Ferrier, S., Gove, A.D. *et al.* (2013). Environmental and historical imprints on beta diversity: insights from variation in rates of species turnover along gradients. *Proc. Royal Soc. B.* , 280, 20131201.

Forrester, N.J., Rebolleda-Gómez, M., Sachs, J.L. & Ashman, T.L. (2020). Polyploid plants obtain greater fitness benefits from a nutrient acquisition mutualism. *New Phytol.*

Greilhuber, J. & Leitch, I.J. (2013). Genome size and the phenotype. In: *Plant Genome Diversity, Volume 2, Physical Structure, Behaviour and Evolution of Plant Genomes* (eds. Leitch, IJ, Greilhuber, J, Doležel, J & Wendel, JF). Springer, pp. 323-344.

Grime, J. & Mowforth, M. (1982). Variation in genome size—an ecological interpretation. *Nature* , 299, 151.

Gude, S., Pinçe, E., Taute, K.M., Seinen, A.-B., Shimizu, T.S. & Tans, S.J. (2020). Bacterial coexistence driven by motility and spatial competition. *Nature* , 578, 588-592.

Guignard, M.S., Nichols, R.A., Knell, R.J., Macdonald, A., Romila, C.A., Trimmer, M. *et al.* (2016). Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. *New Phytol.* , 210, 1195-1206.

Guo, W., Yang, J., Sun, X.-D., Chen, G.-J., Yang, Y.-P. & Duan, Y.-W. (2016). Divergence in ecophysiological responses to drought mirrors the distinct distribution of *Chamerion angustifolium* cytotypes in the Himalaya–Hengduan mountains region. *Front. Plant Sci.* , 7, 1329.

Hastings, A. (1980). Disturbance, coexistence, history, and competition for space. *Theor. Popul. Biol.* , 18, 363-373.

Herben, T., Suda, J., Klimešová, J., Mihulka, S., Říha, P. & Šímová, I. (2012). Ecological effects of cell-level processes: genome size, functional traits and regional abundance of herbaceous plant species. *Ann. Bot.* , 110, 1357-1367.

Hessen, D.O., Jeyasingh, P.D., Neiman, M. & Weider, L.J. (2010). Genome streamlining and the elemental costs of growth. *Trends Ecol. Evol.* , 25, 75-80.

Ives, A.R. & Helmus, M.R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecol. Monogr.* , 81, 511-525.

Keck, F., Rimet, F., Bouchez, A. & Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol* , 6, 2774-2780.

Knight, C.A. & Ackerly, D.D. (2002). Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecol. Lett.* , 5, 66-76.

Knight, C.A. & Beaulieu, J.M. (2008). Genome size scaling through phenotype space. *Ann. Bot.* , 101, 759-766.

Knight, C.A., Molinari, N.A. & Petrov, D.A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* , 95, 177-190.

Körner, C. (2007). The use of 'altitude'in ecological research. *Trends Ecol. Evol.* , 22, 569-574.

Kraft, N.J.B., Comita, L.S., Chase, J.M., Sanders, N.J., Swenson, N.G., Crist, T.O. *et al.* (2011). Disentangling the drivers of beta diversity along latitudinal and elevational gradients. *Science* , 333, 1755-1758.

Laliberté, E., Zemunik, G. & Turner, B.L. (2014). Environmental filtering explains variation in plant diversity along resource gradients. *Science* , 345, 1602-1605.

Lee, S.-I. & Kim, N.-S. (2014). Transposable elements and genome size variations in plants. *Genomics Inform* , 12, 87-97.

Leitch, A. & Leitch, I. (2008). Genomic plasticity and the diversity of polyploid plants. *Science* , 320, 481-483.

Leitch, I.J., Hanson, L., Lim, K., Kovarik, A., Chase, M., Clarkson, J. *et al.* (2008). The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann. Bot.* , 101, 805-814.

Leitch, I.J. & Leitch, A.R. (2013). Genome size diversity and evolution in land plants. In: *Plant Genome Diversity, Volume 2, Physical Structure, Behaviour and Evolution of Plant Genomes* (eds. Leitch, IJ, Greilhuber, J, Doležel, J & Wendel, JF). Springer, pp. 307-322.

Levins, R. & Culver, D. (1971). Regional coexistence of species and competition between rare species. *Proc. Natl. Acad. Sci. U.S.A.* , 68, 1246-1248.

Li, D., Dinnage, R., Nell, L.A., Helmus, M.R. & Ives, A.R. (2020). phyr: an R package for phylogenetic species-distribution modelling in ecological communities. *Methods Ecol. Evol.* , 11, 1455-1463.

Li, D. & Ives, A.R. (2017). The statistical need to include phylogeny in trait-based analyses of community composition. *Methods Ecol. Evol.* , 8, 1192-1199.

MacGillivray, C., Grime, J. & Team, T.I.S.P. (1995). Testing predictions of the resistance and resilience of vegetation subjected to extreme events. *Funct. Ecol.* , 9, 640-649.

Manion, G., Lisk, M., Ferrier, S., Lugilde, K.M., Fitzpatrick, M.C., Fitzpatrick, M.M.C. *et al.* (2017). Package 'gdm'. *A toolkit with functions to fit, plot, and summarize Generalized Dissimilarity Models: CRAN Repository, R* .

Marquet, P.A., Naeem, S., Jackson, J.B. & Hodges, K. (2019). Navigating transformation of biodiversity and climate. *Science Advances* , 5, eaba0969.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends Ecol. Evol.* , 21, 178-185.

Miller, J.E., Damschen, E.I. & Ives, A.R. (2019). Functional traits and community composition: A comparison among community-weighted means, weighted correlations, and multilevel models. *Methods Ecol. Evol.* , 10, 415-425.

Mori, A.S., Isbell, F. & Seidl, R. (2018). β-diversity, community assembly, and ecosystem functioning. *Trends Ecol. Evol.* , 33, 549-564.

Murphy, J. & Riley, J.P. (1962). A modified single solution method for the determination of phosphate in natural waters. *Anal. Chim. Acta* , 27, 31-36.

Otto, S.P. & Whitton, J. (2000). Polyploid incidence and evolution.*Annu. Rev. Genet.* , 34, 401-437.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* , 7, 549-555.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D. *et al.* (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* , 20, 561-576.

Pellicer, J., Hidalgo, O., Dodsworth, S. & Leitch, I.J. (2018). Genome size diversity and its impact on the evolution of land plants.*Genes* , 9, 88.

Pollock, L.J., Morris, W.K. & Vesk, P.A. (2012). The role of functional traits in species distributions revealed through a hierarchical model.*Ecography* , 35, 716-725.

Qian, H. & Jin, Y. (2016). An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *J Plant Ecol* , 9, 233-239.

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Robroek, B.J., Jassey, V.E., Payne, R.J., Martí, M., Bragazza, L., Bleeker, A. *et al.* (2017). Taxonomic and functional turnover are decoupled in European peat bogs. *Nat. Commun.* , 8, 1-9.

Roddy, A.B., Théroux-Rancourt, G., Abbo, T., Benedetti, J.W., Brodersen, C.R., Castro, M. *et al.* (2020). The scaling of genome size and cell size limits maximum rates of photosynthesis with implications for ecological strategies. *Int. J. Plant Sci.* , 181, 75-87.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Series B Stat Methodol J R STAT SOC B* , 71, 319-392.

Salguero-Gómez, R., Jones, O.R., Jongejans, E., Blomberg, S.P., Hodgson, D.J., Mbeau-Ache, C. *et al.* (2016). Fast–slow continuum and reproductive strategies structure plant life-history variation worldwide. *Proc. Natl. Acad. Sci. U.S.A.* , 113, 230-235.

Segraves, K.A. (2017). The effects of genome duplications in a community context. *New Phytol.* , 215, 57-69.

Sheth, S.N., Morueta-Holme, N. & Angert, A.L. (2020). Determinants of geographic range size in plants. *New Phytol.* , 226, 650-665.

Šmarda, P., Hejcman, M., Březinová, A., Horová, L., Steigerová, H., Zedek, F. *et al.* (2013). 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.

Stephens, C.R., Gonzalez-Salazar, C., Villalobos, M. & Marquet, P. (2020). Can Ecological Interactions be Inferred from Spatial Data?*Biodivers. inform.* , 15, 11-54.

Tikhonov, G., Opedal, O.H., Abrego, N., Lehikoinen, A., de Jonge, M.M., Oksanen, J. *et al.* (2020). Joint species distribution modelling with the r-package Hmsc. *Methods Ecol. Evol.* , 11, 442-447.

Tilman, D. (1994). Competition and biodiversity in spatially structured habitats. *Ecology* , 75, 2-16.

Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *Am Stat* , 63, 366-372.

van Breugel, M., Craven, D., Lai, H.R., Baillon, M., Turner, B.L. & Hall, J.S. (2019). Soil nutrients and dispersal limitation shape compositional variation in secondary tropical forests across multiple scales. *J. Ecol.* , 107, 566-581.

Wang, X.-B., Lu, X.-T., Yao, J., Wang, Z.-W., Deng, Y., Cheng, W.-X.*et al.* (2017). Habitat-specific patterns and drivers of bacterial β-diversity in China's drylands. *ISME J.* , 11, 1345-1358.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. *et al.* (2015). So many variables: joint modeling in community ecology. *Trends Ecol. Evol.* , 30, 766-779.

Weiss-Schneeweiss, H., Emadzade, K., Jang, T.-S. & Schneeweiss, G.M. (2013). Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenet. Genome Res.* , 140, 137-150.

Yawata, Y., Cordero, O.X., Menolascina, F., Hehemann, J.-H., Polz, M.F. & Stocker, R. (2014). Competition–dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations.*Proc. Natl. Acad. Sci. U.S.A.* , 111, 5622-5627.

Zanne, A.E., Tank, D.C., Cornwell, W.K., Eastman, J.M., Smith, S.A., FitzJohn, R.G. *et al.* (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature* , 506, 89-92.

Zhang, W., Song, J., Wang, M., Liu, Y.-Y., Li, N., Zhang, Y.-J. *et al.* (2017). Divergences in hydraulic architecture form an important basis for niche differentiation between diploid and polyploid*Betula* species in NE China. *Tree Physiol.* , 37, 604-616.

Acknowledgements

We are grateful to all the team members in 2012-Transect-Investigation in the northern grassland of China. We thank Otso Ovaskainen, Nerea Abrego, Øystein Opedal, and others for organizing the workshop:*Multivariate modelling in ecology and joint species distribution models* . This work was supported by National Key Research and Development Program of China (2016YFC0500703) and Natural Science Foundation of China (31822006).

Figure legends

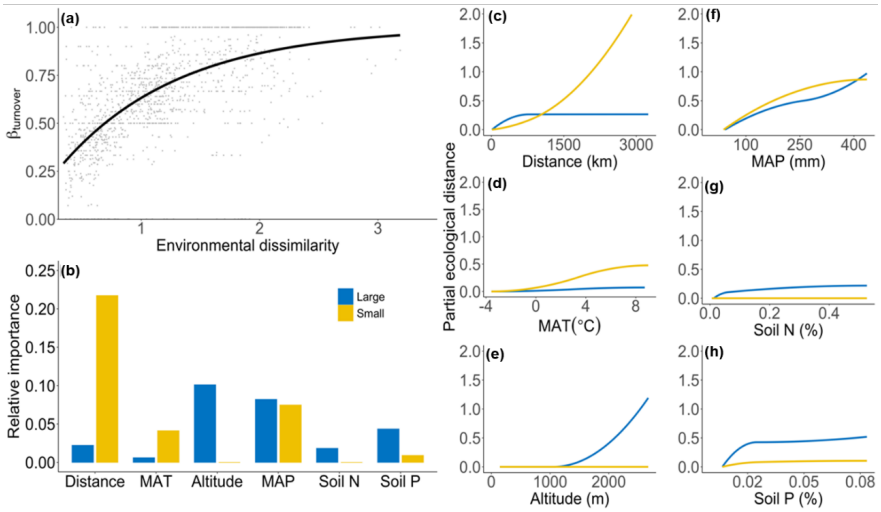

Fig. 1 Παιρωισε ςομποσιτιοναλ δισσιμιλαριτψ αμονγ πλαντ ςομμυνιτιες (τυρνοερ ςομπονεντ, $\beta_{\text{τυρνοερ}}$, βασεδ ον Σιμπσον ινδεξ) αλονγ τηε 3200-κμ ενιρονμενταλ γραδιεντ. These results were generated from the generalized dissimilarity model (GDM, for details see methods) based on presence/absence data. (a) Relationship between β_{turnover} and environmental dissimilarity based on all the environmental variables (MAP, MAT, altitude, soil N, and soil P). (b) Relative importance of variables for explaining variation in β_{turnover}. (c-h) Partial ecological distances (i.e. effects on β_{turnover}) showing the individual effects of each variable on β_{turnover} for species with large (blue) and small (yellow) genomes. The separation plants by genome size was based on the global median genome size value of terrestrial plants (2.5 pg/1Cx) in the Kew database (see*Materials and Methods*). Locations on each line associated with larger values on the y-axis indicate an increased likelihood of observing that genome-size group at that value along the x-axis, and higher maxima in curves indicate larger effects associated with that variable overall.

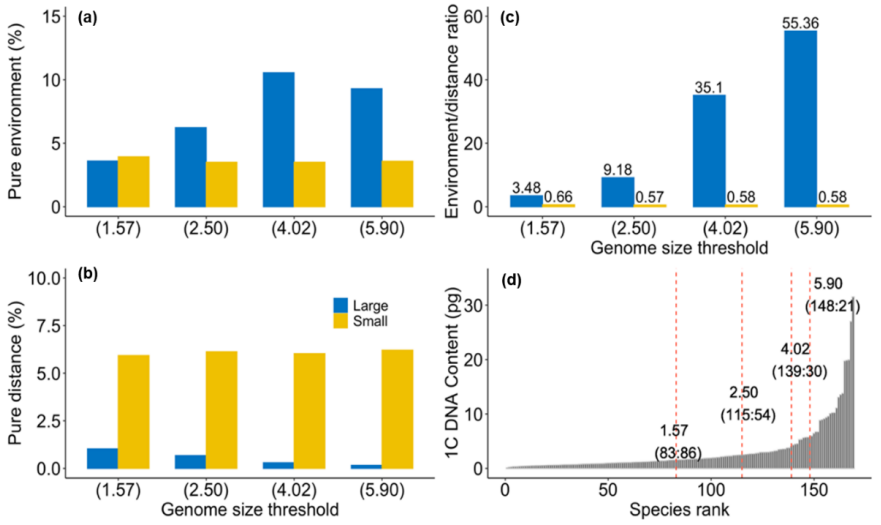

Fig. 2 Τηε προπορτιον οφ αριανςε (% αριατιον) ιν ςομποσιτιοναλ δισσιμιλαριτψ αμονγ πλαντ ςομμυνιτιες ($\beta_{\text{τυρνοερ}}$, βασεδ ον Σιμπσον ινδεξ) ιν εαςη συβσετ (λαργε/σμαλλ γενομε σιζε) ας εξπλαινεδ βψ ενιρονμενταλ αριαβλες (ινςλυδινγ ΜΑΠ, ΜΑΤ, αλτιτυδε, σοιλ Ν, ανδ σοιλ Π) ανδ γεογραπηιςαλ διστανςε υσινγ διφφερεντ γενομε σιζε τηρε-

σηολδς. Plant communities was separated into the large- and small- genome size (1C DNA content, pg) subsets by either the median (1.57 pg) or mean (4.02 pg) genome size for the 169 species of the present study or from the median (2.50 pg) or mean (5.90 pg) genome size of the global terrestrial plants in the Kew database (see *Materials and Methods*). Generalized dissimilarity models (GDMs) were fitted using presence/absence data for the large- (blue) and small- (yellow) genome size subsets. (a) Pure environmental variation without a spatial component represents the effect of environmental filtering. (b) Pure distance variation without an environmental component can represent the effect of dispersal limitation if all important environmental drivers are included in the environmental component. (c) The ratio of variation explained by pure environmental variation versus geographic distance was calculated, with estimates also shown numerically on top of each bar. (d) Measured genome size for the 169 species ranked from small-to-large along the x-axis, with red dashed lines indicating thresholds used in the analyses; the numbers inside parentheses indicate the number of species in each group.

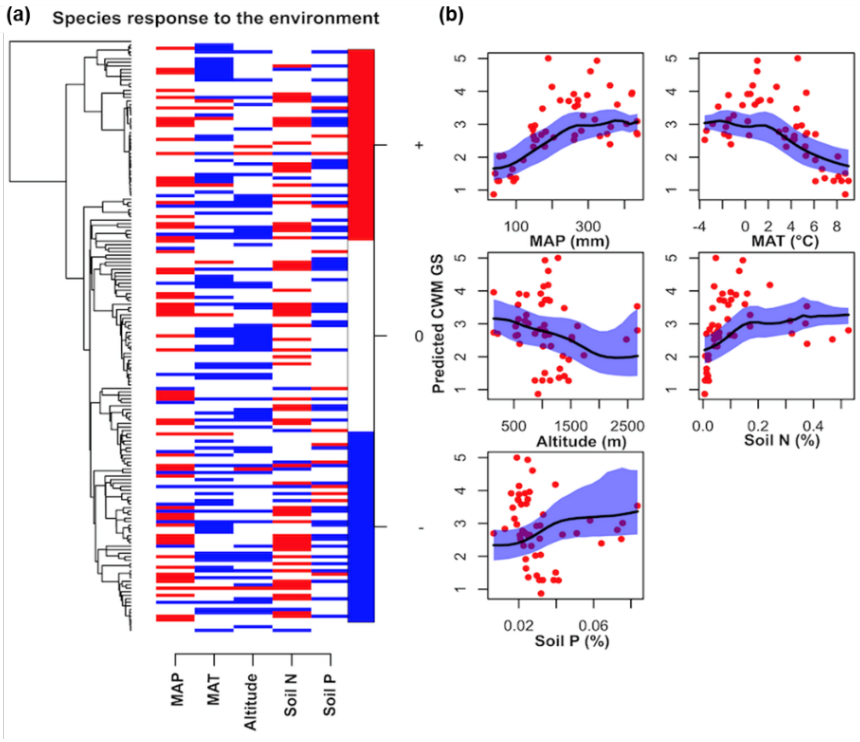

Fig. 3 Results of HMSC models showing effects of environmental filtering based on species' presence/absence data. (a) HSMC-based estimates of species responses to the environmental covariates (MAP, MAT, altitude, soil N and P), i.e β parameter. Plant species were listed according to their phylogenetic relationships (a high-resolution tree with detailed genome size (GS, 1C DNA content, pg) was supplied in Supplementary Fig. S2). Positive and negative responses to the environmental covariates, based on posterior statistical support value (posterior probability at least 95%), are shown in red, and blue, respectively. (b) Model-based community weighted means (CWMs) of genome size and its response to the environmental covariates. The CWM of genome size was calculated based on the prediction from the HMSC model.

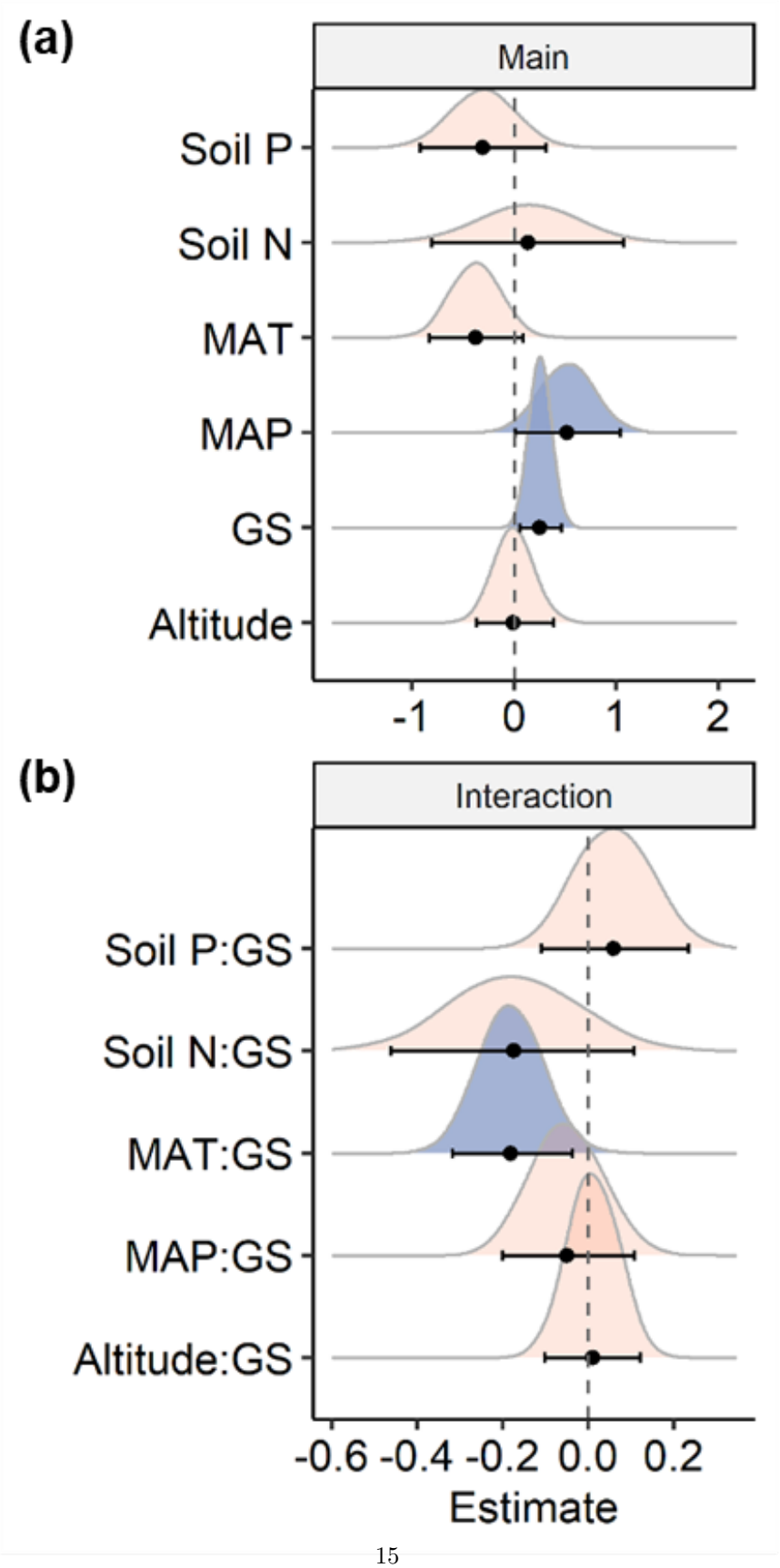

Fig. 4. Estimated PGLMM coefficients for the effects of genome size (GS, 1C DNA content, pg) and environmental predictors (MAP, MAT, altitude, soil N and P) on the probability of plant species occurrence. Panel (a) depicts the main effects of predictors while panel (b) illustrates estimates interactive effects of GS and environments. Positive estimates indicate a positive relationship between predictors and the probability of occurrence. Blue (or pink) density distribution indicates estimates that are (or are not) significantly different from zero (95% credible interval does not overlap with zero).