Changes in the British and Irish flora – the role of genome size

Marie Christine Henniges

School of Biological and Behavioural Sciences,

Queen Mary University of London,

Mile End Road,

London E1 4NS, United Kingdom

Submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy

December 2022

Statement of originality

I, Marie Christine Henniges, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Marie Christine Henniges

Date: 15 December 2022

Details of collaboration and publications are specified at the beginning of each chapter where applicable.

Acknowledgements

I am grateful for the constant and steadfast support of my supervisors, Ilia Leitch, Andrew Leitch and Alex Antonelli. It has been a great honour to work with such remarkable people who continuously amazed me with their knowledge, insight and passion. Beyond their professionalism, I was lucky to have such a supportive and understanding supervisory team who never failed to encourage me and spur me on with their own enthusiasm and dedication. Thank you.

This research would not have been possible without the opportunities provided by the London NERC DTP. I thank NERC for funding my research (grant no. NE/L002485/1) and the DTP for providing training, advice and career development.

The work undertaken here builds on the knowledge generated and kindly shared with me by countless experts, notably Clive A Stace, Richard J Gornall, Kevin J Walker, Oliver L Pescott, Olaf Booy, Tom A Humphrey, Maarten Christenhusz, Alex Twyford, Peter Hollingsworth, Natasha de Vere, Jaume Pellicer and Rafaël Govaerts. Their support and input have been vital to the development of this research.

I would also like to thank my panel members, Richard Nichols and Rob Knell, for their interest in my work and animated discussions that always led to new angles to consider. The experience of my time as a PhD student has been shaped and made vastly more enjoyable by the lovely team I was able to be a part of; Robyn Powell, Luca Pegoraro, Joseph Morton, Emily Gu, Ana Franco, Lubna Faizullah, Lauren Pollitt, Rory Donnellan, Elaine Wang and Tahir Shah – I thank you all for your contributions (scientific and personal).

Many thanks are owed to my friends, both old and new, for their constant support and for bearing with me when I got snowed under with work. The greatest thanks are owed to my family, especially my wonderful parents, Marion and Martin, whose unconditional love and support underlies every of my achievements. Last but certainly not least I am extremely grateful to my husband Michael who has been in the frontline, enduring the highs and lows of this journey with me and has been my greatest support and strength. Thank you all.

Abstract

Unprecedented anthropogenic changes are causing drastic shifts in biodiversity, species ranges and the survival of plants. Understanding which attributes put plants at risk is of vital importance for safeguarding the natural world. Genome size is a fundamental plant attribute with strong links to a variety of plant traits and its study opens novel areas of ecological research, leading to a new understanding of plant responses to environmental changes.

The aim of this thesis is to consider the role that genome size plays at landscape scales. To achieve this aim, I assembled an inventory of the flora of Britain and Ireland and analysed species distribution patterns within the flora over time, together with information on land use, climate and nutrient deposition changes across the past three decades.

Distinctive spatial patterns of mean genome size per hectad of Britain and Ireland were found across time, with a steady increase in mean genome size since the 1980s. A particular driver of the patterns appears to be land use, with areas especially impacted by humans containing plant communities characterised by larger mean genome sizes.

Genome size, along with a set of functional traits and niche descriptors, were all informative characters in a random forest algorithm predicting species trends, achieving 70% prediction accuracy. The effect of genome size was found to be indirect, mediated via its influence on functional traits, which in turn lead to differing niche requirements and temporal trends.

The results suggest that the effects of genome size on plant growth, fitness and response to the abiotic environment impacts landscape scale species compositions. Genome size emerges as an important meta-trait to consider when monitoring and anticipating biodiversity changes in response to environmental change and could be used in models that guide conservation efforts.

Glossary

Category	Term	Description
Native status	Native	Species which colonised the study region naturally since the last glaciation or that were present before that point
	Non-native/alien	Species which were most likely introduced by human activity, they are further subdivided into archaeophytes and neophytes
	Archaeophyte	Non-natives that were introduced by human activity before the year 1500, further subdivided into colonist, cultivated and denizen
	- colonist	Weedy species occurring on open ground
	- cultivated	Deliberately cultivated species
	- denizen	Species with near-native behaviour, able to compete with natives
	Neophyte	Non-natives that were introduced by human activity since the year 1500
	- casual	Not naturalised, persist only for a short time
	- naturalised	Established and self-perpetuating
	- survivor	Not naturalised, but able to persist for long times, often as relics in locations where they were planted
	Neonative	Species that arose from natural hybridisation between either a native and a non-native or between two non-native taxa, or that evolved from another neonative or non-native species within Britain & Ireland
Genome size	Genome size	The amount of DNA in an unreplicated nucleus as estimated by flow cytometry, given as 1 C (haploid nucleus) and 2 C (diploid nucleus), measured in picograms (pg) or mega base pairs (Mbp)
Realised niche	Ellenberg indicator values	Ordinal data for the preference of a species within an environmental gradient; data given for light, moisture, soil acidity, soil fertility, salt and temperature (each species is assigned a value (typically from 1 to 9) depending on its predicted preference within the environmental gradient); concept developed by Ellenberg (1974)
Life strategy	CSR strategy	Functional classification of each species' propensity for being a competitor (C), stress-tolerator (S) or ruderal (R); developed by Grime (1974)

Life-form *sensu* Raunkiær (1934)	Hydrophyte	Aquatic herb, buds are submerged in water or in soil underneath water, leaves may float or be submerged, flowering parts may emerge (= ‘aquatics’)
	Helophyte	Buds are fully submerged in water or within water-saturated soil, flowers and leaves emerge fully (= ‘emergents’)
	Geophyte	Above ground parts die outside the growing season, plant survives as a bulb, rhizome, tuber or root bud
	Hemicryptophyte	Herbaceous stems that tend to die back outside the growing season, buds survive on or just under the soil level, includes many biennial and perennial herbs
	Therophyte	Life cycle is completed within one growing season, surviving as a seed until the next growing season (= ‘annuals’)
	Chamaephyte	Herbaceous or woody stems, buds above soil, but not exceeding 50 cm (= ‘shrubs’)
	Phanerophyte	Persistent, woody stems, buds usually 3 m or more above ground, trees and larger shrubs (= ‘trees’)

Contents

Each appendix begins with a table of contents.

List of figures

List of tables

Chapter 1 General introduction

An introduction to genome size and ploidy

Setting the scene

Since the molecular revolution, modern biology has been dominated by sequence-based studies of genomes. Consequently, while the content of genomes is increasingly well understood, comparatively less attention has been paid to a fundamental characteristic of every living being's genetic material: its amount. Indeed, genome size is often considered useful only in the context of estimating costs of sequencing projects (Li & Harkness, 2018), and frequently not even then. However, the 'nucleotype hypothesis' (Bennett, 1972) established the idea that the size of the genome itself, rather than just the information encoded within it, might have fundamental effects on the phenotype. While across eukaryotes most genomes are small, ranges in genome size are staggeringly large in a few groups, as exemplified in the diverse clade of flowering plants, where genome sizes range at least 2,400-fold (Pellicer *et al.,* 2018). Given this span, the notion that genome sizes might affect plant evolution, physiology and ecology in a fundamental way suggests itself and questions regarding the impacts of genome size on plants' abilities to establish, adapt and dominate are gaining more traction (e.g. Guignard *et al.,* 2016; Simonin & Roddy, 2018; Suda *et al.,* 2015).

Some definitions

Defining genome size is not an easy matter, with its terminology being unclear until Greilhuber *et al.* (2005) gave stable definitions for the terms used in the context of DNA amounts. Since then, genome size has been defined as the total amount of DNA within an organism's unreplicated gametic nucleus, based on chromosome numbers and measured in units of mega base pairs (Mbp) or picograms (pg); one pg equals 978 Mbp (Doležel *et* *al.,* 2003). The C-value, often used synonymously with genome size (Bennett & Smith, 1976), can be considered the DNA amount typical for a specific genotype (Bennett & Leitch, 2005), with the numeric value attributed to it corresponding to the amount of DNA in the nucleus as the cell progresses through the cell cycle (i.e. 1C, 2C and 4C corresponding respectively to the amount of DNA in the nucleus of (i) a gamete, (ii) a somatic cell following fertilisation, and (iii) a cell that has undergone DNA replication (S phase of cell cycle) but not yet entered mitosis or meiosis). The C of C-value was clarified by the inventor of the abbreviation (Swift, 1950) to stand for ‘constant’, but a suite of genetic processes means it is not constant for a species over evolutionary time, nor indeed need it be constant within species. Nevertheless, in large parts the C-value of a species is a good indication of the genome size in the majority of individuals belonging to it.

The size of a genome itself is determined by genetic processes in the species’ ancestry, including whole genome multiplications (especially in plants; Van de Peer, Mizrachi & Marchal, 2017; Wendel, 2015), the multiplication of repetitive, non-essential DNA sequences (often termed ‘junk’ or ‘selfish’ DNA), caused – for the most part – by transposable elements (Leitch & Leitch, 2013; Chénais *et al.,* 2012; Elliott & Gregory, 2015), and the frequency of recombination-based DNA removal (Schubert & Vu, 2016; Pellicer *et al.,* 2018).

The role of genome size and ploidy in angiosperm evolution

Charles Darwin famously considered the rapid radiation of the angiosperm clade an ‘abominable mystery’, as he lamented to JD Hooker in 1879 (Darwin, 1903; Davies *et al.,* 2004; Buggs, 2021). Genome duplication events are believed to be one driving force behind this burst in diversification (Wendel, 2015; Escudero & Wendel, 2020; Fox *et al.,* 2020), since they create opportunities for example for subfunctionalisations and neofunctionalisations of duplicated genes (Wood *et al.,* 2009; Tank *et al.,* 2015; Van de Peer *et al.,* 2017; Landis *et al.,* 2018; Ren *et al.,* 2018; Sandve, Rohlfs & Hvidsten, 2018). Fossil and genomic data show that all modern angiosperms have polyploid ancestors arising from one or multiple whole genome duplication events, even if they are now considered to be diploid (Masterson, 1994; Wood *et al.,* 2009; Jenczewiski *et al.,* 2013). They are thus palaeopolyploids (Van de Peer, Maere & Meyer, 2009; Jiao *et al.,* 2011; Paterson *et al.,* 2012). Roughly a third of modern angiosperms and nearly all economically important crops are polyploids and it is estimated that c. 15% of all angiosperm speciation events involve ploidy changes (Wood *et al.,* 2009).

Partly caused by this history of genome duplications, the 1C-values in flowering plants range at least 2,400-fold, with the smallest known genome containing a mere 0.07 pg/1C of DNA in *Genlisea tuberosa* (Fleischmann *et al.,* 2014) and the largest, of *Paris japonica* (Franch. & Sav.) Franch., measured at 152.23 pg/1C (Pellicer *et al.,* 2010). The abundance of whole genome duplications in angiosperm lineages (Van de Peer, Maere & Meyer, 2009; Jiao *et al.,* 2011; Paterson *et al.,* 2012) might suggest a prevalence of large genome sizes, but there is a considerable skew towards smaller genomes instead (Pellicer *et al.,* 2018; Fig. 1.1). This skew is further unexpected considering the constant pressure towards genome expansion caused by the amplification of transposable elements and other repetitive DNA sequences (e.g. tandem repeats) which can comprise up to 90% of the genome (Novák *et al.,* 2020), and it has been suggested that there might be inherent disadvantages for larger genomes, leading to a universal limit in genome size, at around 150 pg/1C (Hidalgo *et al.,* 2017). The skew in the distribution of genome sizes towards small genomes indicates various mechanisms of DNA deletion, leading to genome downsizing, processes that also contribute to the diploidisation of genomes following an episode of polyploidy (Leitch &

Bennett, 2004; Wendel, 2015; Pellicer *et al.*, 2018; Zenil-Ferguson, Ponciano & Burleigh, 2016; Wang *et al.*, 2021).

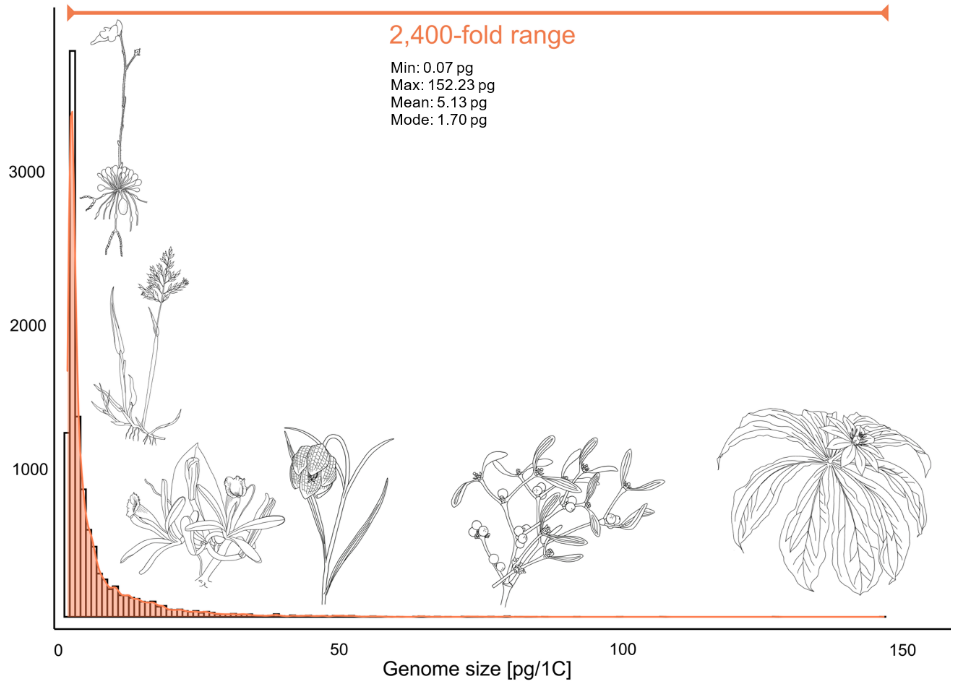


Fig. 1.1 Histogram and smoothed density of genome size data in angiosperms showing skew towards smaller genomes. The number of species is plotted by genome size in picograms [pg] per haploid genome [1C] based on 10,770 estimates. Examples for plants are represented along the histogram, close to their genome size. They are from left to right: *Genlisea tuberosa* Rivadavia, Gonella & A.Fleischm. (0.07 pg/1C = minimum), *Holcus lanatus* L. (1.70 pg/1C = mode), *Vanilla × tahitensis* J.W.Moore (2.62 pg/1C ≈ mean), *Fritillaria meleagris* L. (47.30 pg/1C), *Viscum album* L. (88.90 pg/1C), *Paris japonica* (Franch. & Sav.) Franch. (152.23 pg/1C = maximum). Minimum, maximum, mean and mode genome size are given at the top and species are chosen to represent approximations of those values.

Ecological consequences of large genomes and polyploidy

Given the strong tendency of genomes to be small, the question as to whether large genomes are detrimental has been explored. The 'large genome constraint hypothesis' (Knight, Molinari & Petrov, 2005) highlights a number of physiological routes that might disadvantage or exclude plants with large genomes from certain habitats and growth strategies. This hypothesis has since been refined and supported in some experimental settings (e.g. Faizullah *et al.*, 2021; Guignard *et al.*, 2016).

Larger genomes impose limits on trait space and strategies

A multitude of studies has found correlations between genome size and a range of plant traits with ramifications at all levels of plant form and function (see e.g. Knight & Beaulieu, 2008; Šímová & Herben, 2012; Greilhuber & Leitch, 2013; Doyle & Coate, 2019). Greater amounts of genetic material have been shown to be associated with longer cell cycles and hence longer generation times, constraining plants with very large genomes to slow-growing, perennial life strategies (Bennett, 1971; Bennett, 1987; Veselý, Bureš & Šmarda, 2013). It has been shown that larger genomes impact minimum cell size through constraints imposed by DNA packing. However, while the relationship holds for some cell types including meristematic cells, it is not apparent for all cell types (Cavalier-Smith, 2005; Knight & Beaulieu, 2008), largely due to variances in vacuole sizes (Greilhuber & Leitch, 2013). Beyond the cell, Beaulieu *et al. (*2007) show that larger genome sizes are correlated with increases in seed size, which in some species may lead to shorter maximum dispersal distances (Jenkins *et al.,* 2007), although in those species that exploit large herbivores (e.g. elephants) or water (e.g. coconuts) for dispersal, such relationships would certainly break down.

Of particular interest to recent research and to this thesis is the role of genome size in shaping the water and nutrient requirements of plants. The hypothesised role of genome size in water use efficiency is complex and coupled intimately with photosynthetic efficiency and nutrient acquisition (Faizullah *et al.,* 2021). Genome size has been found to correlate positively with stomatal guard cell length and negatively with the density of stomatal pores (Beaulieu *et al.,* 2008). Larger stomatal openings are associated with slower reactions to changes in water availability which can occur rapidly with fluctuations in weather patterns, and their low density may lead to suboptimal gas exchange within plant tissues impacting photosynthesis and water use efficiency (Franks & Farquhar, 2001;

Lawson & Vialet-Chabrand, 2019). Stomatal conductance handles the delicate balance between limiting water loss and allowing sufficient CO_2 uptake for efficient photosynthetic rates, leading to potentially detrimental effects of large genomes, especially in arid environments and under CO_2 limitation (Roddy *et al.,* 2020). Adding to this, larger genome size has also been demonstrated to negatively affect CO_2-diffusion within cells and leaves by increasing cell sizes, thus altering surface to volume ratios of cells and changing the mesophyll structure, further impairing photosynthetic rates (Cavalier-Smith, 2005; Théroux-Rancourt *et al.,* 2021). Simonin & Roddy (2018) suggest that the competitive success of early angiosperm lineages is a direct consequence of genome downsizing which allowed for smaller, more densely packed stomata and consequently for more efficient gas exchange and photosynthesis. While support for this link between genome size and the trade-off between water use efficiency and photosynthetic productivity is well supported by correlation studies, the causation remains to be proven in experimental settings.

Experimental support exists to a greater degree for the effect of genome size on nutrient requirements. Nucleic acids are inherently expensive molecules, particularly demanding high levels of nitrogen and phosphorus, which contribute 14.5% and 8.7% respectively to their make-up (Sterner & Elser, 2002; Hessen, Elser & Sterner, 2013). Competition for nutrient allocation between genomes and essential proteins suggest another direct detriment to plants that need to maintain excessively large genomes at the cost of efficient growth (Hessen *et al.,* 2010). This constraint becomes particularly drastic when nutrients are limited or biologically unavailable (Elser *et al.,* 2007); biologically available phosphorus is a sparse resource especially in tropical soils (Vitousek *et al.,* 2010; Chadwick *et al.,* 1999), while mineralised nitrogen limitation increases towards the poles (Houlton *et al.,* 2008; Menge *et al.,* 2017; Deng *et al.,* 2018; Du *et al.,* 2020). Controlled field experiments have consolidated our understanding of the limiting effect of a lack of nutrients for plants with large genomes. Šmarda *et al. (*2013), Guignard *et al.* (2016) and Peng *et al.* (2022)

demonstrated in grassland settings that in both the short and long term, combined fertilisation with nitrogen and phosphorus allows communities comprising species with higher mean genome size to develop, suggesting that restrictions imposed by nutrient deficits on genomes were lifted. Another interesting potential drawback associated with the higher nutrient content in species with large genomes lies in their apparent appeal for herbivores; rabbits may preferentially consume plants with larger genome sizes (Guignard *et al.,* 2019), potentially due to their higher nutrient content and/or the lower cell wall to cytoplasm ratio making them more succulent, or they recover more slowly following herbivore damage due to their longer cell cycle times and hence slower growth rates.

Certain plant life forms appear to relax some of the constraints placed on genome size by the environment. Many plants with extremely large genomes are geophytes (e.g. *Paris japonica* (Franch. & Sav.) Franch., Fig. 1.1) or parasitic plants (e.g. *Viscum album* L., Fig. 1.1), both of which may be less dependent on environmental nutrient limitation. In the case of parasitic plants, nutrients and water can be obtained from their respective host plants, potentially allowing an upward drift of genome sizes in the absence of selection pressures imposed by nutrient limitation on free living species (Hibberd & Jeschke, 2001; Veselý, Bureš & Šmarda, 2013). Geophytes are characterised by storage organs, such as bulbs or tubers. Such storage capacity allows for the accumulation of nutrients and pre-division of cells during dormancy periods, or over long periods of time, rendering those plants more independent from fluctuations in soil nutrient availability and enabling them to grow speedily by cell expansion in spite of long replication times for larger genomes (Grime & Mowforth, 1982; Grime, 1983; Greilhuber, 1995; Veselý, Bureš & Šmarda, 2013).

The supposition that large genomes are detrimental has also received support from a study of species at risk of extinction, where threatened plant species were demonstrated to have larger genomes on average than less vulnerable related species (Vinogradov, 2003). The situation is complicated, however, by the fact that polyploidy is often associated with increased performance and vigour caused by fixed heterozygosity (Soltis & Soltis, 2000; Birchler, 2015; Dodsworth, Chase & Leitch, 2016). This, along with a tendency towards selfing tolerance in polyploids (Dodsworth, Chase & Leitch, 2016), perhaps contributes towards explaining the prominent role of polyploid species in plant breeding (Sattler, Carvalho & Clarindo, 2016). While genome size has been found to be smaller in invasive species which typically have fast growth rates and excellent dispersal abilities, it has also been shown that polyploidy and higher chromosome numbers are positively correlated with invasiveness (Pandit, White & Pocock, 2014; Suda *et al.,* 2015). These data suggest that genome size and ploidy should be considered together in order to gain full insight into their effects on species performance.

Genome size and environmental change

There appears to be an emerging pattern suggesting that plants with large genomes might face limitations (see above) that preclude them from some ecological strategies (such as short-lived annual lifestyles; Bennett, 1972), whereas species with smaller genomes may have a wider range of options open to them. Existing data suggest that plants with large genomes are excluded from extreme environments (e.g. by Knight & Ackerly, 2002), e.g. where fast reproduction cycles and tolerance of pollution, radiation or nutrient and water limitation are advantages (Vidic *et al.,* 2009; Temsch *et al.,* 2010; Sparrow & Miksche, 1961; Einset & Collins, 2018; Knight, Molinari & Petrov, 2005).

The hypothesised decreases in water use efficiency with increasing genome size, as noted above, suggest that increasingly arid conditions should select against plants with large genomes, but studies attempting to show such effects in plant distribution data have led to varied results and only partial support for the hypothesis (synthesised in Knight, Molinari & Petrov, 2005). It has been suggested that the relatively small spatial scales at which such studies have been performed and the use of predominantly linear methods to test for correlation might obscure patterns of genome size and climatic preferences that actually do exist (Knight & Ackerly, 2002; Knight, Molinari & Petrov, 2005). Recent data are suggesting a tendency towards species with smaller genomes in the tropics and larger genomes towards the poles, although species occupying areas above latitudes c. 50-60 N exhibit decreasing genome sizes (Bureš *et al.,* 2022 (in press)), and a study of palm genome size demonstrated selection pressure against genome expansion under water stress (Schley *et al.,* 2022). Should the expected hypothesis of disadvantages of large genomes in arid conditions hold true, the effects of climate change could have disproportionate effects on plants with larger genomes, especially in areas where increasing temperatures and more frequent drought events are to be expected under unmitigated climate change scenarios (Ritchie *et al.,* 2019).

While species with larger genome sizes tend to have a decreased tolerance to heavy metal pollution (Vidic *et al.,* 2009; Temsch *et al.,* 2010) and radiation (Sparrow & Miksche, 1961; Einset & Collins, 2018), one very prevalent pollution type might actually favour plants with large genomes. Nutrient pollution (e.g. from agricultural fertilisation) might favour species with large genomes, as observed in field experiments (Šmarda *et al.,* 2013; Guignard *et al.,* 2016), but is associated with decreasing biodiversity leading to diminishing ecosystem services (Peng *et al.,* 2022; Lambers *et al.,* 2011; Carpenter *et al.,* 2009; Rohr *et al.,* 2016; Stevens *et al.,* 2016).

Given the above, genome size is expected to have a role to play in shaping species distributions in response to climate change and anthropogenic pressures. The hypothesised links between the environment and genome size are illustrated in Fig. 1.2.

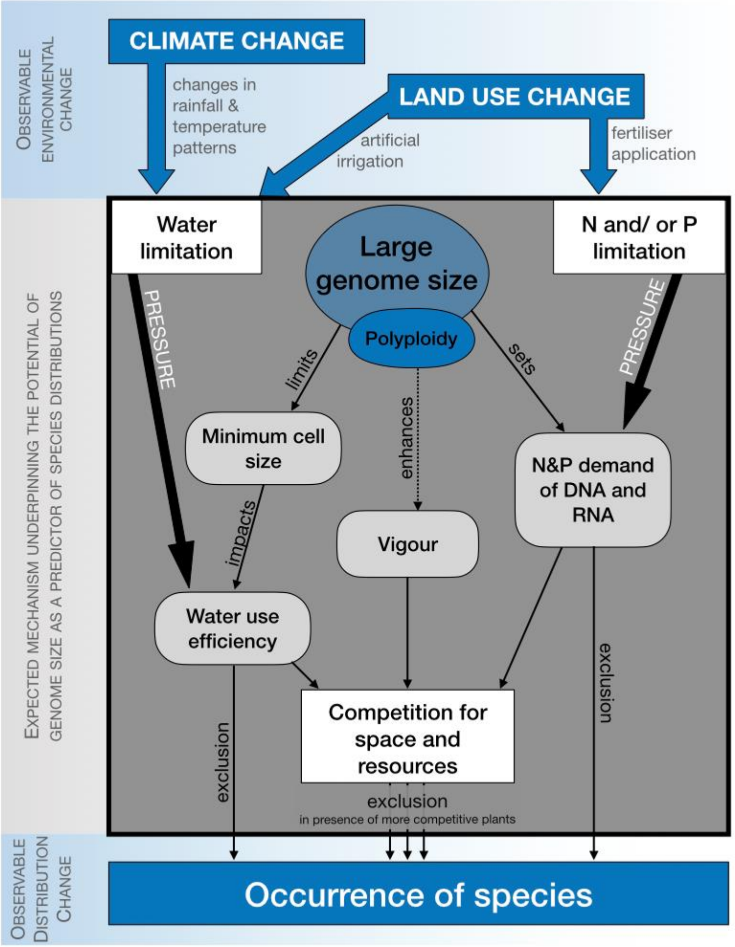

OBSERVABLE ENVIRONMENTAL CHANGE
CLIMATE CHANGE
changes in rainfall & temperature patterns
LAND USE CHANGE
artificial irrigation
fertiliser application
EXPECTED MECHANISM UNDERPINNING THE POTENTIAL OF GENOME SIZE AS A PREDICTOR OF SPECIES DISTRIBUTIONS
Water limitation
Large genome size
Polyploidy
N and/ or P limitation
PRESSURE
limits
sets
enhances
PRESSURE
Minimum cell size
N&P demand of DNA and RNA
impacts
Vigour
Water use efficiency
Competition for space and resources
exclusion
exclusion
exclusion
in presence of more competitive plants
OBSERVABLE DISTRIBUTION CHANGE
Occurrence of species

Fig. 1.2 Proposed mechanism by which genome size might link observable environmental change with the occurrence of species. The immediate effects of genome size modulate the ability of plants to withstand pressures posed by nutrient and water limitation as well as competition with other species.

Only connect – the flora of Britain and Ireland

The knowledge base regarding plant genomes is ever increasing and well-accessible; the Chromosome Count Database (Rice *et al.*, 2015) compiles information on chromosome counts made on land plants, while the Plant DNA C-values database, established in 2001 (Bennett & Leitch, 2001) and most recently updated in release 7.1, represents a central hub for genome size and polyploid-level estimates that currently houses information for 12,273 species of land plants and algae (Pellicer & Leitch, 2019). An even greater wealth of information exists for functional traits and characters associated with plants worldwide, with the TRY Plant Trait database (Kattge *et al.*, 2020) and local floras (e.g. Chytrý *et al.*, 2021; Falster *et al.,* 2021) at the forefront of the collation, curation and dissemination of trait data.

The flora of Britain and Ireland offers itself as the setting for a case study of the local role of genome size, chromosome counts and ploidy level in shaping species distributions in the context of plants' overall trait space. Its history of repeated colonisations shaped by glaciations (Clark *et al.,* 2012; Ingrouille, 2012) and local isolation as a consequence of rising sea levels (Ingrouille, 2012), followed by pervasive and high levels of human disturbance (Fig. 1.3), high levels of eutrophication (Smart *et al.,* 2003; Firbank *et al.,* 2000) and current climate change (Ritchie *et al.,* 2019), make the area of particular interest in exploring how genome characters impact the distribution dynamics of native species (see glossary) and new arrivals in a system.

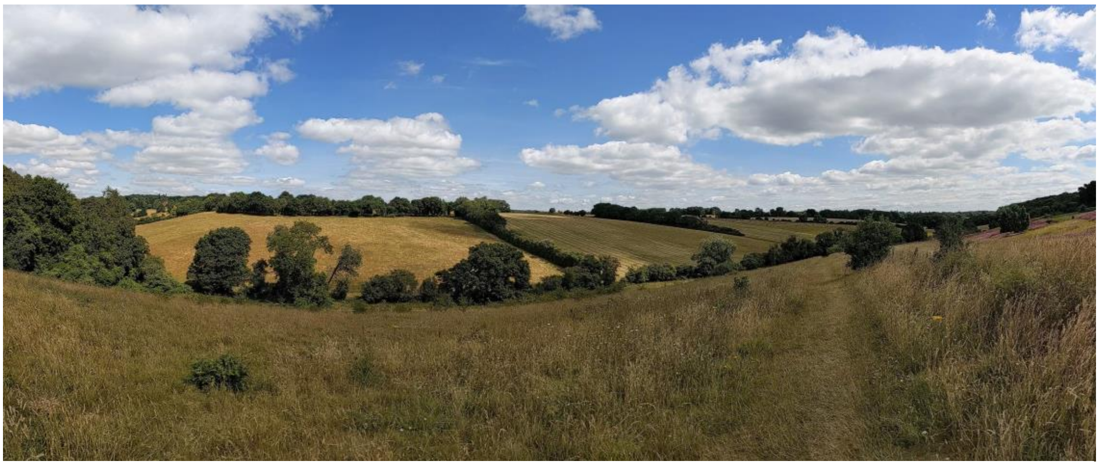


Fig. 1.3 Characteristic landscape panorama of Britain and Ireland. The high levels of human disturbance throughout history have created a landscape characterised by a patchwork of arable field, grazing grounds and settlements, interspersed with more natural environments. Image taken near Henley-on-Thames, Oxfordshire, in summer 2022.

Spatio-temporal changes in the British and Irish flora have been remarkably well documented for centuries, with keen interest in botany resulting in comprehensive species lists as early as the 1690s (Ray, 1690) and continued recording by passionate expert and amateur botanists alike continuing to this day (Pescott *et al.,* 2019a). High quality distribution information for the flora of Britain and Ireland is curated and made available by the Botanical Society of Britain and Ireland and presented in the *Atlas of the British*

Flora (Preston, Pearman & Dines, 2002), allowing research that traces changes in species distributions over time, although inevitable biases in the record base must be accounted for (Isaac & Pocock, 2015).

In their fundamental treatise on the 'large genome constraint hypothesis', Knight, Molinari & Petrov (2005) highlight the need for a holistic approach to the study of genome size and its multiple correlations, suggesting that the integration of genome size estimates, trait data and species occurrence records would be of particular value in advancing our understanding of the role genome size has to play in ecology. Connecting the available information on plant genomes, traits and distributions outlined above would allow for this very approach and promises novel insights into the influences of genome size at landscape scales.

Aims and scope of the thesis

As highlighted above, a growing body of research, largely based on controlled experimental settings, points towards a role of genome size in plant ecology. The overarching aim of this thesis is to test the hypothesis that genome size, polyploidy and chromosome number have ecological ramifications that translate into effects at landscape scales within the study area of Britain and Ireland.

This aim necessitates the integration of extensive data on genetic characters, functional traits, species distributions and environmental parameters (including land cover). To achieve that need, Chapter 2 describes the generation of a flora-wide and taxonomically harmonised inventory of all vascular plant species in Britain and Ireland that underlies all subsequent chapters, containing a wealth of trait, genetic and descriptive information. Chapter 3 outlines the species distribution data that allows me to move to landscape scales.

Since such distribution data necessarily contains biases specific to biological records, the chapter also relays the methods of bias correction employed to achieve the highest level of reliability in the data used in the following chapters and presents some top-level findings regarding distribution trends of species within the flora.

In Chapter 4, I explore spatial patterns in the genome size and ploidy levels of angiosperm species across Britain and analyse the environmental factors driving them, including also a special focus on the impact of human activities. Considering changes in climate that have occurred over the past three decades, this chapter also determines the magnitude of range shifts along the North-South axis, and contextualises them with genome size. In Chapter 5, the dynamics reported on in earlier chapters are built upon through the application of genome size, along with functional traits and niche requirements, in predicting decreasing or increasing species trends. The final chapter offers a general synthesis of the findings across the preceding chapters and strives to reach some conclusions regarding the role of genome size in shaping the British and Irish flora through time and space, while also highlighting important focal points for future research.

Chapter 2 A snapshot of the vascular flora of Britain and Ireland

Publication information

The majority of this chapter was published in *Scientific Data* on 10 January 2022.

I am the lead author of the publication which forms the basis of all sections leading to the sub-chapters Results and Discussion. Thereafter the Results and Discussion provide some initial insights into the flora. The Methods section also contains additional information on the production of a flora-wide phylogeny, which occurred after publication of the paper. In order to integrate the publication into this thesis as a chapter, I have also changed the language from American English to British English.

Marie C Henniges, Ilia J Leitch and Andrew R Leitch developed the concept of the database presented here. Ilia J Leitch, Andrew R Leitch, Richard J Gornall, Max R Brown, Alex D Twyford, Peter M Hollingsworth, Kevin J Walker and Marie C Henniges planned the scope and practicality of the resource. Marie C Henniges extracted and compiled the datasets from a diversity of sources and carried out data validation. Clive A Stace made available his knowledge and allowed use of his published work. Maarten JM Christenhusz made available his knowledge on life forms. Sahr Mian and Robyn F Powell performed genome size measurements. Max R Brown compiled and calculated hybridisation scores. Laura Jones and Natasha de Vere contributed barcode information. Richard J Gornall made available his dataset of chromosome numbers and attributed numbers to the listed species, checked the species list and provided valuable guidance. Kevin J Walker contributed species status and distribution metrics. Alexandre Antonelli provided guidance on data compilation and R package development. Marie C Henniges coordinated the activities of all participants in the published paper. All authors contributed to the writing of the manuscript. Marie C Henniges provided a first draft. All authors approved the final version of the manuscript.

Henniges, M.C., Powell, R.F., Mian, S., Stace, C.A., Walker, K.J., Gornall, R.J., Christenhusz, M.J., Brown, M.R., Twyford, A.D., Hollingsworth, P.M., Jones, L., de Vere, N., Antonelli, A., Leitch, A.R. and Leitch, I.J., 2022. A taxonomic, genetic and ecological data resource for the vascular plants of Britain and Ireland. *Scientific Data*, **9**(1), 1-8.

Abstract

The vascular flora of Britain and Ireland is among the most extensively studied in the world, but the current knowledge base is fragmentary, with taxonomic, ecological and genetic information scattered across different resources. Here we present the first comprehensive data repository of native and non-native species optimised for fast and easy online access for ecological, evolutionary and conservation analyses. The inventory is based on the most recent reference flora of Britain and Ireland, with taxon names linked to unique Kew taxon identifiers and DNA barcode data. Our data resource for 3,227 species and 26 traits includes existing and unpublished genome sizes, chromosome numbers and life strategy and life-form assessments, along with existing data on functional traits, species distribution metrics, hybrid propensity, associated biomes, realised niche description, native status and geographic origin of non-native species. This resource will facilitate both fundamental and applied research and enhance our understanding of the flora's composition and temporal changes to inform conservation efforts in the face of ongoing climate change and biodiversity loss.

Introduction

There is a long history of botanical recording on the islands of Britain and Ireland, referred to here as 'BI', comprising England, Scotland, Wales, Northern Ireland, Republic of Ireland, Isle of Man and the Channel Islands (Fig. 2.1). The earliest systematic records date back to Revd John Ray in 1690. The Botanical Society of Britain and Ireland (BSBI) provides access to large-scale geographic distribution data based on more than 40 million occurrence records, allowing for unique research into changes within the flora, especially throughout the last century.

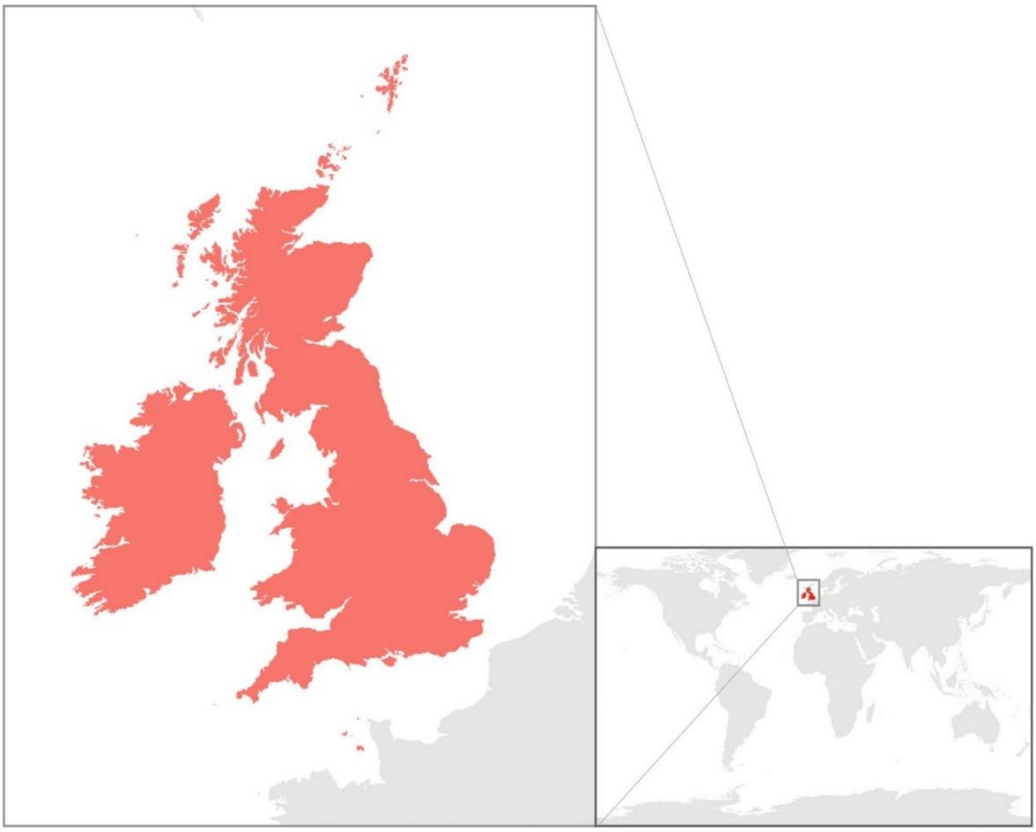


Fig. 2.1 Area covered by the database – Britain and Ireland. The area considered for our attribute database (red) comprises England, Scotland, Wales, Northern Ireland, the Republic of Ireland, the Isle of Man and Channel Islands.

In addition, a large community of researchers have contributed to a wide knowledge base for the BI flora, which includes large datasets on ecological traits, chromosome numbers and cytotype variation, population-level variation and genetic diversity, DNA barcoding resources, and many other traits (Fitter & Peat, 1994; Database for the Biological Flora of the British Isles; BSBI database search facility). The conservation status of species in the BI flora has been assessed, including via national red listing (BSBI, 2021). This diversity is protected *in situ* via a range of land management and habitat protection schemes and *ex situ* via large conservation collections and seed banking, with 72% of the UK's native and archaeophyte angiosperm species (see Table 2.1 for a glossary of terms used) currently conserved in seed banks (Clubbe *et al.,* 2020).

Table 2.1 Glossary of terms used within dataset.

Category	Term	Description
Native status	Native	Species which colonised the study region naturally since the last glaciation or that were present before that point
	Non-native/alien	Species which were most likely introduced by human activity, they are further subdivided into archaeophytes and neophytes
	Archaeophyte	Non-natives that were introduced by human activity before the year 1500, further subdivided into colonist, cultivated and denizen
	- colonist	Weedy species occurring on open ground
	- cultivated	Deliberately cultivated species
	- denizen	Species with near-native behaviour, able to compete with natives
	Neophyte	Non-natives that were introduced by human activity since the year 1500
	- casual	Not naturalised, persist only for a short time
	- naturalised	Established and self-perpetuating
	- survivor	Not naturalised, but able to persist for long times, often as relics in locations where they were planted
	Neonative	Species that arose from natural hybridisation between either a native and a non-native or between two non-native taxa, or that evolved from another neonative or non-native species within Britain & Ireland
Genome size	Genome size	The amount of DNA in an unreplicated nucleus as estimated by flow cytometry, given as 1 C (haploid nucleus) and 2 C (diploid nucleus), measured in picograms (pg) or mega base pairs (Mbp)

Realised niche	Ellenberg indicator values	Ordinal data for the preference of a species within an environmental gradient; data given for light, moisture, soil acidity, soil fertility, salt and temperature (each species is assigned a value (typically from 1 to 9) depending on its predicted preference within the environmental gradient); concept developed by Ellenberg (1974)
Life strategy	CSR strategy	Functional classification of each species' propensity for being a competitor (C), stress-tolerator (S) or ruderal (R); developed by Grime (1974)
Life-form *sensu* Raunkiær (1934)	Hydrophyte	Aquatic herb, buds are submerged in water or in soil underneath water, leaves may float or be submerged, flowering parts may emerge (= 'aquatics')
	Helophyte	Buds are fully submerged in water or within water-saturated soil, flowers and leaves emerge fully (= 'emergents')
	Geophyte	Above ground parts die outside the growing season, plant survives as a bulb, rhizome, tuber or root bud
	Hemicryptophyte	Herbaceous stems that tend to die back outside the growing season, buds survive on or just under the soil level, includes many biennial and perennial herbs
	Therophyte	Life cycle is completed within one growing season, surviving as a seed until the next growing season (= 'annuals')
	Chamaephyte	Herbaceous or woody stems, buds above soil, but not exceeding 50 cm (= 'shrubs')
	Phanerophyte	Persistent, woody stems, buds usually 3 m or more above ground, trees and larger shrubs (= 'trees')

BI also have a long history of agricultural development, beginning in prehistoric times (Fowler, 1983) and undergoing a series of changes towards high levels of intensification, especially during the last century (Green, 1990). Together these make the region a globally outstanding system for exploring the links between species richness, diverse ecological traits and genetic attributes, allowing for studies on the impacts of environmental and land use change on natural plant communities.

Despite these opportunities, large scale studies of the flora are challenging because of the current lack of a taxonomically harmonised repository of species present in the BI flora, optimised for comparative flora-wide assessments rather than information retrieval for individual species. The most recent version of a similar data source (Hill, Preston & Roy, 2004) dates back to 2004 and almost exclusively covers native species (Table 2.2). Another notable inventory, the *List of Vascular Plants of the British Isles* (Kent, 1992), including both native and non-native species, has served as the basis for subsequent checklists and keys (e.g. Hill, Preston & Roy, 2004; Stace, 2019). Since a large proportion (approx. 50% according to Stace & Crawley, 2015) of species present in BI today are not native, informed predictions of the species' future abundance and distribution require that attribute data are readily available for native and non-native plants alike. Trait-based approaches to species distribution modelling and community ecology are emerging to enable more informed forecasting of population level responses to changes in the abiotic environment, such as those driven by climate change (Schleuning *et al.,* 2020; Tikhonov *et al.,* 2020; Vesk *et al.,* 2021).

Here we present a comprehensive database and inventory of vascular plant species – both native and non-native – currently present in BI, together with diverse trait data. The species list is based on the most recent edition of the *New Flora of the British Isles* (Fourth Edition, Stace, 2019) (including name changes from the 2021 reprint), with each species name linked to its unique identification number according to the World Checklist of Vascular Plants (WCVP, 2020) to ensure taxonomic clarity and stability.

The repository encompasses 3,209 extant species and 18 extinct species (see Materials & Methods). Each entry includes associated intrinsic and functional traits, distribution and ecologically relevant data where available. In addition to information adapted from Stace (2019) such as taxonomic ranks, native or non-native status and origin (for non-native plants), we have collated other types of data from various sources (Table 2.2). These include data for several functional traits (e.g. Specific Leaf Area (SLA), and seed mass), realised niche descriptions (Ellenberg's indicator values (Ellenberg, 1974), Table 2.1), the life strategy of each species using the CSR strategy framework of Grime (1974) (Table 2.1), information on hybridisation propensity, genome sizes and chromosome numbers, along with DNA barcode sequences.

We consider that this comprehensive data repository will be crucial for enabling both fundamental and applied research to enhance our understanding of the biotic and abiotic factors influencing the distribution and composition of the vascular plant flora of BI. Such new insights will be invaluable for predicting how different species will respond to environmental challenges such as biodiversity loss, climate change, land use change and new pests and diseases and hence enable more informed decision making to ensure the long-term stewardship of the BI flora.

Table 2.2 Summary of the categories included in the database of vascular plants in BI.

Category	Percentage of species with data in the complete flora (percentage for natives/ non-natives given in brackets)	Databases and other reference sources of the data	Description
Taxonomy	100% *(100%/100%)*	Nomenclature and lower taxonomic ranks – Stace (2019, reprint 2021); World Checklist of Vascular Plants (WCVP), Higher taxonomic ranks (order, family) – NCBI via 'taxize', WCVP	Overview of species taxonomy, including kew_id, species binomials (Stace, 2019 (reprint 2021); WCVP), taxonomic rank (i.e. order, family, genus, subgenus, section, subsection, series, species, group, aggregate). Also provided are URLs to species pages on WCVP, POWO and IPNI.
Native status	(i) 98% *(-/-)*	(i) Stace (2019)	Description of level of nativity or establishment in Britain and Ireland ('Native', 'Archaeophyte denizen', 'Neophyte naturalised' etc., for full list and explanations see Table 2.1)
	(ii) 82% *(-/-)*	(ii) PLANTATT (Hill, Preston & Roy, 2004) and ALIENATT (pers. comm. Kevin J Walker)	
	(iii) 48% *(-/-)*	(iii) *Alien Plants* (Stace & Crawley, 2015); pers. comm. Kevin J Walker	
	Combined coverage: **99%**		
Functional traits	SLA: 56% *(69%/45%)*	Public data from the TRY database (Kattge *et al.*, 2020); for a list of specific publications see Table S2.2	Functional plant trait averages for (i) Specific Leaf Area (SLA, mm^2 mg^{-1}), (ii) Leaf Dry Matter Content (LDMC, g g^{-1}), (iii) Seed mass (mg), (iv) Leaf area (mm^2), and (v) Vegetative height (m). Also included is maximum vegetative height (m)
	LDMC: 47% *(65%/32%)*		
	Seed mass: 68% *(74%/63%)*		
	Leaf area: 51% *(66%/39%)*		
	Vegetative height: 75% *(88%/65%)*		
Realised niche description	Percentages given for each Ellenberg category, first the coverage derived from PLANTATT, then from Döring, 2017, then coverage for both sets combined:	(i) PLANTATT (Hill *et al*. 2004) (ii) Zeigerwerte von Pflanzen & Flechten in Mitteleuropa (Döring, 2017)	Ellenberg indicator values assigned to plant species as observed in Britain (data from PLANTATT) and in Central Europe (data from Döring, 2017). Listed Ellenberg categories are L (light), F (moisture, from German 'Feuchtigkeit'), R (reaction, soil acidity), N (nutrients, fertility), S (salt), T (temperature, only for European data). Numbers typically range across a scale of 1 to 9, with low numbers indicating an affinity to the lower end of the described environmental gradient. S and F have
	L: (i) 56% *(94%/23%)*		
	(ii) 60% *(94%/32%)*		
	61%		

	F: (i) 56% *(94%/23%)*		different scales with S spanning from 0 to 9 and F spanning from 1 to 12.
	(ii) 59% *(92%/31%)*		
	61%		
	R: (i) 56% *(94%/23%)*		
	(ii) 55% *(87%/29%)*		
	60%		
	N: (i) 56% *(94%/23%)*		
	(ii) 58% *(91%/30%)*		
	60%		
	S: (i) 56% *(94%/23%)*		
	(ii) 61% *(95%/32%)*		
	61%		
	T: (i) - *(-/-)*		
	(ii) 27% *(38%/17%)*		
	-		
Life strategy	(i) 14% *(27%/4%)*	(i) Electronic Comparative Plant Ecology (Hodgson *et al.*, 1995)	Life strategy of plants given as the CSR category established by Grime (1974). These can be either competitor (C), stress tolerator (S), ruderal (R), or a combination of these (e.g. CS, C/CSR)
	(ii) 45% *(63%/30%)*	(ii) Inferred from functional traits	
	Combined coverage: **45%**		
Growth form and succulence	(i) 86% *(89%/83%)* for growth form	Public data from the TRY database (Kattge *et al.*, 2020), for specific references see Table S2.2	(i) Plant growth form given as recorded by the TRY contributors Engemann and Günther. Categories used are aquatic, fern, graminoid, herb, shrub, and tree.
	(ii) 16 succulent species		(ii) Succulence was recorded when a species was mentioned as 'succulent' by any author in the growth form data from the TRY database (16 species).
Life-form	100% *(100%/100%)*	Pers. comm. Maarten JM Christenhusz	Life form categories as per Raunkiær (1934) (e.g. 'chamaephyte', 'hemicryptophyte', 'therophyte' or combinations thereof, see Table 1 for explanations)
Associated biome	48% *(86%/15%)*	Ecoflora database (Fitter & Peat, 1994)	Description of typical biome for the species (e.g. 'Mediterranean' or 'Boreo-Temperate')
Origin of non-native species	(i) 48% *(-/87%)*	Stace, 2019	(i) Description of country or region of origin (i.e. the most likely area plants were introduced from; not equal to complete foreign distribution) for non-native species.

	(ii) 46% (-/84%)		(ii) Information is also given as a TDWG level 1 code (Brummitt, 2001).
Species distributions	98% *(98%/97%)*	BSBI distribution database	Species occurrences within Britain and Ireland at hectad resolution for four time intervals: 1987 – 1999, post 2000, 2000 – 2009, 2010 – 2019. Data are given separately for Great Britain and the Isle of Man, Ireland and the Channel Islands.
Hybrid propensity	20% *(30%/11%)*	Stace *et al.*, 2015; pers. comm. Max R Brown	Hybrid propensity (*sensu* Whitney *et al.*, 2010), scaled hybrid propensity (weighted by the number of intragenic combinations within the genus)
DNA barcodes	44% *(87%/11%)* (with at least one record on BOLD), 935 species have sequence data for all three sequences (*rbcL, matK* and ITS2)	Pers. comm. L Jones & Natasha de Vere, de Vere *et al.*, 2012; Jones *et al.*, 2021	Hyperlinks to the Barcode of Life Data System (BOLD) record pages, which contains barcode sequences (*rbcL, matK* and ITS2), an image of the scanned herbarium specimen and details about sample collection
Genome size	66% *(77%/58%)* (with at least one measurement)	(i) Unpublished data from the Royal Botanic Gardens, Kew (RBG Kew) (ii) Šmarda *et al.,* 2019 (iii) Zonneveld, 2019 (iv) Plant DNA C-values database (Pellicer & Leitch, 2019)	Genome size measurements, given as 1C- and 2C-values in picograms (pg) and megabasepairs (Mbp)
	14% *(27%/4%)* (with at least one measurement from material sourced from the study region)		
Chromosome numbers	44% *(76%/17%)* (with at least one measurement from material sourced from the study region) 72% (*91%/57%*) (with chromosome numbers available from all sources combined)	Database curated at the University of Leicester by Richard J Gornall (i) Database curated at the University of Leicester by R.J.G. (ii) Šmarda *et al.,* 2019 (iii) Zonneveld, 2019 (iv) Plant DNA C-values database (Pellicer & Leitch, 2019)	Chromosome counts and estimates prepared from plant material from Britain and Ireland, an additional column adds further chromosome numbers from outside of the study area

Materials & Methods

The broad categories of data included in the repository are summarised in Table 2.2 and visualised in Fig. 2.2. Each category is explained in greater detail below, while full details together with accompanying notes are given in Table S2.1. Table 2.2 gives an overview of data coverage per category, both across all species and for native species separately. A complete list of data sources is available in Table S2.2.

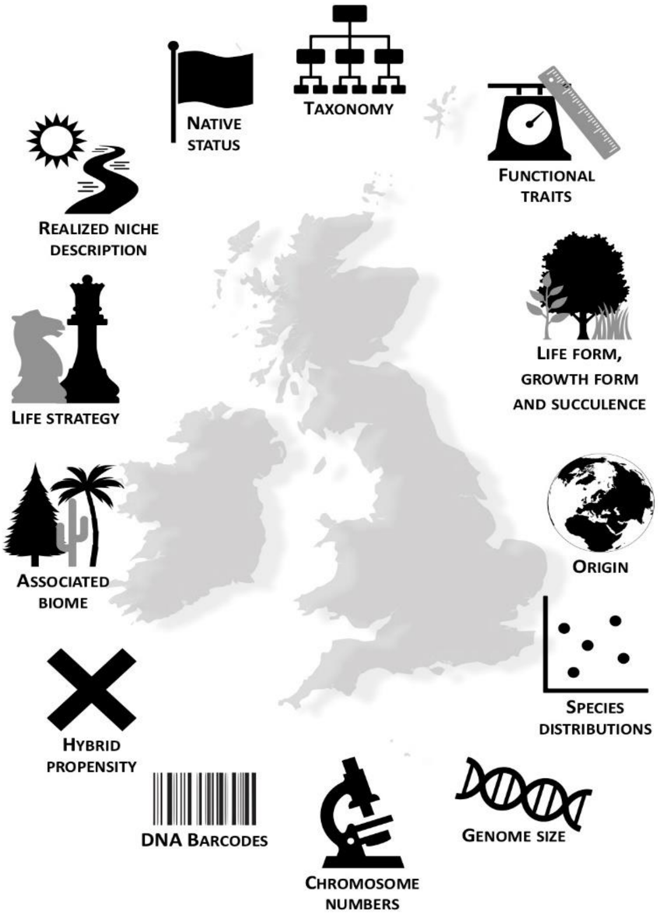


Fig. 2.2 Visualisation of the attributes presented in the database.

Generation of the species list

Taxon names listed in the most recent and widely accepted *New Flora of the British Isles'* index (Stace, 2019) were digitised via the Optical Character Recognition Software ReadirisTM 17 (IRIS). Results from the digitisation were transferred into a spreadsheet and obvious recognition errors were fixed. The resulting table contained 5,687 taxa and associated taxonomic authorities. A total of 360 unnamed hybrids were excluded, as well as species noted to have only questionable or unconfirmed records, leaving 5,038 species. Forty-one intergeneric hybrid species, 827 entries relating to (notho)subspecies, (notho)varieties, cultivars and forma were also removed along with 720 named hybrids. Species that were included by Stace (2019) but which he considered were not part of the flora (i.e. listed as 'other species' and 'other genera', e.g. genus *Tragus* or *Coreopsis verticillata*) were also excluded. Seven species that were labelled 'extinct' in the flora were included as there were indications that the species might be in the process of reintroduction (e.g. *Bromus interruptus*, *Bupleurum falcatum* and *Schoenoplectus pungens*). Extinct native and archaeophyte species without any signs of reintroduction (e.g. *Dryopteris remota*) are also listed but no additional data are provided and they are not included in calculations of completeness of data (Table 2.2). The final number of extant species listed here is therefore 3,209 (comprising 1,468 natives, 1,690 non-natives and 51 species with unknown status), plus 18 formally extinct species (natives and archaeophytes not seen in study region since 1999). Species names and taxonomic authorities were revised according to the 2021 reprint of the *New Flora of the British Isles*, communicated to us by Clive A Stace ahead of publication. Genera with less well-defined species – for example due to apomixis – contain additional information on subgenera, sections, and aggregates, as per Stace (2019). Since misidentifications are common in these groups, we include a column termed ‘unclear_species_marker’ that allows for these species to be quickly identified and excluded from analyses if appropriate. Such genera are often incompletely listed in our database since most microspecies are not sufficiently well defined.

Taxonomy

Nomenclature of the list was checked by Global Names Resolver in the R package ‘taxize’ (Chamberlain & Szöcs, 2013; Chamberlain *et al.,* 2020), using the *International Plant Name Index* (IPNI, 2020) as the data source, to remove any digitisation errors. Resolved names were used to determine accepted higher taxonomic hierarchy (family, order) again using taxize, with the National Center for Biotechnology Information (NCBI) database. Species that could not be resolved by the Global Names Resolver or did not yield matches in the NCBI database for their higher taxonomic ranks were manually checked for name matches in the *World Checklist of Vascular Plants* (WCVP, 2020). Species within the original species list that were found to be identical to a different spelling in WCVP were retained in the database. In such instances, and when slight spelling differences occurred, the columns ‘taxon_name‘ and ‘taxon_name_WCVP‘ differ. To improve clarity, each species is presented here with its unique identification number according to the WCVP (listed as ‘kew_id’) together with three additional columns (i.e. WCVP.URL, POWO.URL and IPNI.URL) which contain hyperlinks to the freely accessible taxon description websites of the *World Checklist of Vascular Plants* (WCVP, 2020), *Plants of the World Online* (POWO, 2020) and *International Plant Names Index* (IPNI, 2020), respectively. Thus, while the taxon names used in the database correspond to those used by Stace (2019), changes in the accepted species name since publication can be traced in columns ‘taxonomic_status’ and ‘accepted_kew_id’. The family classification of WCVP follows APG IV (2016) for angiosperms, Christenhusz *et al.* (2011) for gymnosperms and Christenhusz & Chase (2014) for ferns and lycopods.

Native status

We offer three different datasets which describe the status of a species as native or non-native, and its level of establishment in BI. The first is extracted from Stace (2019), the second contains the status codes used in PLANTATT (Hill, Preston & Roy, 2004) and the unpublished ALIENATT (pers. comm. author K.J.W.) datasets, and the third is extracted from *Alien Plants* (Stace & Crawley, 2015). The status from Stace (2019) and Stace & Crawley (2015) assigns a species to either native or non-native status, with non-natives subdivided into archaeophytes and neophytes at different levels of establishment (e.g. denizen, colonist etc., see Table 1). Status codes from the BSBI can be either AC (alien casual), AN (neophyte), AR (archaeophyte), N (native), NE (native endemic) or NA (native status doubtful).

Functional traits

Data for five ecologically relevant functional traits (i.e. seed mass, specific leaf area [SLA], leaf area, leaf dry matter content [LDMC] and vegetative height) were downloaded from public data available in the TRY database (Kattge *et al.,* 2020) (for specific authors see Table S2.1 and Table S2.2). Averages were calculated using the available measurements downloaded for each species, excluding rows where the measurement was zero. In addition, the maximum vegetative height for each species is given, where available.

Realised niche description

Realised niche descriptions based on assessments made on plants living in BI are given in the form of Ellenberg indicator values (Ellenberg, 1974), as published in PLANTATT (Hill, Preston & Roy, 2004). Ellenberg indicator values place each species along an environmental gradient (e.g. light or salinity) by assigning a number on an ordinal scale, depending on the species' preference for the specific gradient (Table 2.2). This information is often used to gain insights into environmental changes based on species occurrences (Hill, Mountford & Roy, 1999). For species listed under a previously accepted name in PLANTATT, the information was associated with the accepted synonym in Stace (2019). Due to the low coverage of PLANTATT for non-native species included in our list, we additionally include Ellenberg indicator values based on Central European assessments, as made available by Döring (2017). Each Ellenberg category is listed in a separate column, keeping the information from both data sources separate to avoid confounding of assessments based on two different regions (i.e. Britain and Ireland versus Central Europe).

Life strategy

To characterise the life strategy of a species, we used the CSR scheme developed by Grime (1974), which classifies each species as either a competitor (C), stress tolerator (S), ruderal (R) or a combination of these (e.g. CS, SR). CSR classifications were obtained from the *Electronic Comparative Plant Ecology* database (Hodgson *et al.,* 1995). Due to the low coverage of available CSR assessments for species in our database (i.e. data available for just 460 out of 3,209 extant species) we imputed CSR strategies for a further 981 species using available functional trait data, following the method proposed by Pierce *et al.* (2017). The functional leaf traits required for this method – i.e. specific leaf area, leaf area, leaf dry matter content – were obtained from the TRY database (Kattge *et al.*, 2020). Pre-existing (Hodgson, *et al., 1995*) and newly imputed CSR strategies are listed in separate columns.

Growth form, succulence and life-form

Plant growth form descriptions were obtained from the TRY database (Kattge *et al.,* 2020) and filtered for those entries given by specific contributors (Table 2.2) to maintain consistent use of growth form categories. Information on whether a species was considered to be a succulent was obtained by screening the entire growth form information obtained from the TRY database for the phrase 'succulence' or 'succulent'.

Species life-form categories according to Raunkiær (1934) were determined for each species in our dataset with regard to the typical life-form of the species as it grows in BI (pers. comm. Maarten JM Christenhusz).

Associated biome and origin

Information given in the Ecoflora database (Fitter & Peat, 1994) for the biome that each species is associated with was matched to the species names according to Stace (2019). The recognised biome categories follow Preston & Hill (2002) and are 'Arctic Montane', 'Boreal Montane', 'Boreo-Arctic Montane', 'Boreo-Temperate', 'Mediterranean', 'Mediterranean-Atlantic', 'Southern Temperate', 'Temperate', 'Wide Boreal' and 'Wide Temperate'.

For non-native species, the assumed origin (i.e. the region that plants were most likely to have been introduced to BI from, rather than the full non-BI distribution of a species) was adapted from Stace (2019) into a brief description of their country or region of origin. In addition, these descriptions were manually allocated to the TDWG level 1 regions listed in the World Geographical Scheme for Recording Plant Distributions (WGSRPD, TDWG, Brummitt, 2001).

Species distributions

Distribution metrics for each species are given as the number of 10 km square hectads in BI with records for the species in question within a specified time window (pre and post 2000, 1987-1999, 2000-2009 and 2010-2019). The data were derived from the BSBI Distribution Database and were extracted for each species, dividing the study region into Great Britain (incl. Isle of Man), Ireland and the Channel Islands, as previously partitioned for data available in PLANTATT (Hill, Preston & Roy, 2004). The database was queried using species and hectads for grouping, showing only records 'matching or within 2 km of county boundary' and excluding 'do-not-map-flagged' occurrences. The data were not corrected for sampling bias and should therefore only be used as an indication of trends.

Hybrid propensity

Data on hybridisation is provided for 641 species, obtained from the *Hybrid flora of the British Isles* (Stace, Preston & Pearman, 2015) which enumerates every hybrid reported in BI up until 2015 (pers. comm. Max R Brown). Each entry was transcribed manually, and then filtered to exclude (a) hybrids that have been recorded, but not formed in the

British Isles, (b) triple hybrids (mainly reported for the genus *Salix*), (c) doubtful records, (d) hybrids between subspecific ranks, and (e) hybrids where at least one parent is not native (only archaeophytes included). This left 821 hybrid combinations for data aggregation. The metric chosen here is hybrid propensity, which is a per-species metric of how many other species a focal species hybridises with (*sensu* Whitney *et al.,* 2010). A scaled hybrid propensity metric is also given which was calculated by weighting the hybrid propensity score by the number of intrageneric combinations for a given genus, to account for the greater opportunities of hybridisation in larger genera.

DNA barcodes

DNA barcode sequences for plant species present in BI are currently available for 1,413 species in our database. The information was derived from a dataset of *rbcL, matK* and ITS2 sequences compiled for the UK flora generated by the National Botanic Garden of Wales and the Royal Botanic Garden Edinburgh (de Vere *et al.,* 2012; Jones *et al.,* 2021; pers. comm. Laura Jones and Natasha de Vere). The data are given as a hyperlink to the record's page on the Barcode of Life Data Systems (BOLD, Ratnasingham & Hebert, 2007) which includes the DNA barcode sequences as well as scans of the herbarium specimen and information on the sample's collection. Most species have multiple record pages associated with them, due to the sampling of more than one individual. We include a maximum of three BOLD accessions per species; the full range of individuals sampled can be accessed via the original publications (de Vere *et al.,* 2012; Jones *et al.,* 2021). DNA barcodes are almost exclusively available for native species. Future releases of our database will increase the coverage of the non-native flora significantly. Where species in the BOLD database are attributed to a species name that is considered synonymous with another name in our list, the hyperlink is matched to the latest nomenclature (Stace, 2019). 1,421 species have at least one sequence associated with them, and 935 species have sequence data for all three sequences (*rbcL*, *matK* and ITS2).

Genome size and chromosome numbers

Genome size data for 2,117 specimens (at least one measurement per species) were obtained from various sources. Measurements for 467 species were newly estimated using plant material of known BI origin from the Millennium Seedbank of the Royal Botanic Gardens, Kew (Chapman, Miles & Trivedi, 2019). The measurements were made by flow cytometry using seeds or seedlings and following an established protocol (Pellicer, Powell & Leitch, 2020). Information on the extraction buffers and calibration standard species used are available in the file GS_Kew_BI.csv (https://catalogue.ceh.ac.uk/documents/9f097d82-7560-4ed2-af13-604a9110cf6d), along with peak CV values of the measurements as a quality control. Where more than one measurement is reported per species, the measurements were made on plant material from different populations or using different buffers. Previously published data for additional species were obtained from reports on the Czech flora (Šmarda *et al.,* 2019), and the Dutch flora (Zonneveld, 2019), and prime values listed in the Plant DNA C-values database (Leitch *et al.,* 2019; Pellicer & Leitch, 2019). Since significant intraspecific differences in genome size between plant material from different geographical origins have previously been described, predominantly due to cytotype diversity in ploidy level (Kolář *et al.,* 2017), genome size measurements from previously published sources were assessed with regard to the origin of the material. The column 'from_BI_material' (GS_BI.csv, BI_main.csv, see https://catalogue.ceh.ac.uk/

documents/9f097d82-7560-4ed2-af13-604a9110cf6d) allows users to filter for measurements made on material from BI to exclude a potential bias. The information was obtained from the original publication source of each measurement.

Chromosome numbers for 1,410 species (at least one chromosome number per species) determined exclusively from material collected in BI were obtained from an extensive dataset compiled by Richard J Gornall from various published studies, unpublished theses and personal communications from trusted sources. The counts were made between 1898 and 2017, with a large proportion stemming from efforts to achieve greater coverage of the flora by a team of cytologists based at the University of Leicester and headed by Richard J Gornall. Part of the dataset was previously incorporated into the BSBI's data catalogue but has since undergone revisions to incorporate new information and changes in taxonomy. The dataset contained many measurements at subspecies level which were allocated to the species level taxon in our list. This served to include as much of the often considerable infraspecific variation as possible. Since some species for which chromosome counts have been reported elsewhere are lacking chromosome counts from British or Irish material, they are absent from this dataset. To fill such gaps, we also present chromosome numbers from reports on the Czech flora (Šmarda *et al.,* 2019), the Dutch flora (Zonneveld, 2019), and the Plant DNA C-values database (Leitch *et al.,* 2019; Pellicer & Leitch, 2019).

Phylogeny construction

A phylogeny of the species in the BI flora was generated subsequent to the publication of Henniges *et al.* (2022). Many analyses of the BI flora database are likely to require information on the phylogenetic relatedness of species within it (Borges *et al.,* 2019).

To this end, a phylogeny was constructed based on pre-existing phylogenetic trees for seed-plants (Smith & Brown, 2018; Zanne *et al.,* 2014; synthesised by Qian & Jin, 2016), as contained within the R package 'V.PhyloMaker' (Jin & Qian, 2019). Species considered to be taxonomically unclear in the database were removed prior to pruning the megatree down to only include the BI-based species. Out of the remaining species, 1,993 could be matched perfectly to the backbone phylogeny. For those 659 species without a clear match, we used information from the WCVP to identify unambiguous synonyms, i.e. synonyms that are not associated with any other WCVP-accepted taxa. In 161 cases, where such a clear synonym could be found, species were matched to the backbone phylogeny via the synonymous taxon name. Finally, species that could neither be matched directly nor via a synonym were investigated further to find out if previous molecular studies had assigned these species a clear position within family- or genus-level phylogenies (for a detailed reference list for such information, see Table S2.3), giving information about their closest relatives within the megatree. Apart from these small-scale studies, I also used information from an unpublished phylogeny generated by Max R Brown and kindly shared with me by Max R Brown and Alex Twyford to guide these further attachment decisions. This additional reference phylogeny focused on species native to the UK and used separate plastid data and ITS alignment, as well as an APG IV (2016) tree to guide inference of family level relationships. 347 species were attached to the tree in this way, avoiding polytomies by respecting the dichotomous relationships found in the previous molecular studies. The resulting phylogeny contains 2,501 of the 3,227 species present in the database.

Software and visualisation

All data compilation and manipulation was carried out in R 3.5.3 – 4.1.3 (2022), with data management in Microsoft Excel (versions 2019-2022). 'Tidyverse' (Wickham *et al.,* 2019) packages were used for data manipulation and plotting of results.

All figures were generated in R 4.1.3 with post-processing in Microsoft PowerPoint and iWork Keynote, with line drawings of species generated in the raster graphics software Sketchbook. Maps were produced using the R packages 'sf' (Pebesma, 2018) and 'rnaturalearth' (South, 2017). The phylogeny was visualised in iTOL (Letunic & Bork, 2021).

R package and data set information

Data records

A static version of the data as of publication date is available from the NERC Environmental Information Data Centre (https://doi.org/10.5285/9f097d82-7560-4ed2-af13-604a9110cf6d). A metadata file (Database_structure.csv, see also Table S2.1) with explanations of the main dataset (BI_main.csv), additional datasets (GS_BI.csv, GS_Kew_BI.csv and chrom_num_BI.csv), and a complete list of all publications and sources used to compile the data (Detailed_sources.csv, see also Table S2.2) are included along with the data. The main database BI_main.csv lists all taxa included in this work along with their identification number (kew_id), associated taxonomic authorities, taxonomic ranks (order, family, genus, subgenus, section, subsection, series, species, group, aggregate), associated trait, distribution, and ecological data. The main database contains a summary of chromosome numbers and the smallest genome size measurement available per species.

Because more than one chromosome number and genome size measurement has been reported for many species – often reflecting considerable infraspecific variance – these additional chromosome number (chrom_num_BI.csv) and genome size (GS_BI.csv) data are published along with the main dataset as separate files. Detailed information about the newly generated genome size measurements from RBG Kew are summarised in GS_Kew_BI.csv, including information on the calibration standard species and extraction buffers used to estimate the genome size.

The data is also available as an R package on GitHub (https://github.com/RBGKew/BIFloraExplorer, Fig. 2.3), where we aim to provide new releases that will reflect new additions to the dataset as well as taxonomic changes.

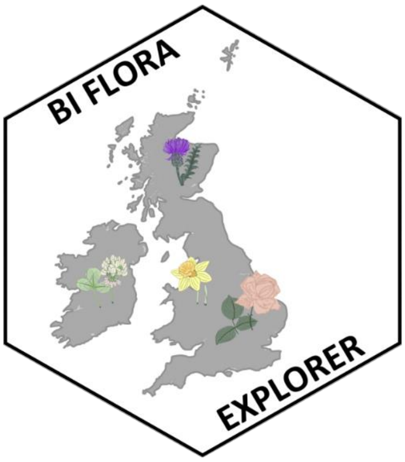


Fig. 2.3 Hex sticker for 'BIFloraExplorer' R package.

Technical validation

All data presented in the resource were compiled from a range of sources, the vast majority of which were from previously published field guides, atlases or peer reviewed articles. All such data are provided with full reference to their source (Table S2.1 and Table S2.2), allowing the user to validate particular pieces of information with ease. Any new unpublished data presented here were either determined experimentally, following best practice protocols (e.g. genome size data), calculated using peer reviewed methods (Pierce *et al.*, 2017), or supplied by one of the expert authors on this publication.

Where data were manually extracted from print sources, spot checks were conducted at various stages throughout the data collection to verify that mistakes had been kept to a minimum. When data were added from online or other digital resources, species binomial and – if available – taxonomic authority information were used to match data to the species in the list. This matching process was manually checked for each dataset.

Usage Notes

We present an easily accessible and downloadable database for the current vascular BI flora, comprising a full list of species with a range of associated ecological, genomic and distribution data. The data as of publication date are freely available for download from the EIDC (https://doi.org/10.5285/9f097d82-7560-4ed2-af13-604a9110cf6d). Species names are presented as published previously (Stace, 2019, with name changes from the 2021 reprint); changes in taxonomy are reflected in columns 'accepted_kew_id', 'accepted_name' and 'accepted_authors', as per WCVP and POWO. The development version of the dataset is available at https://github.com/RBGKew/BIFloraExplorer.

Results

Composition of the flora

There are 3,227 species that are considered part of the extant vascular flora in this database. These species fall into a total of 60 orders and 164 families, with half of all species falling into one of the five largest orders (Poales, Asterales, Rosales, Caryophyllales, Lamiales) and into one of the ten largest families (Rosaceae, Asteraceae, Poaceae, Fabaceae, Brassicaceae, Cyperaceae, Caryophyllaceae, Apiaceae, Lamiaceae and Plantaginaceae), Fig. 2.4.

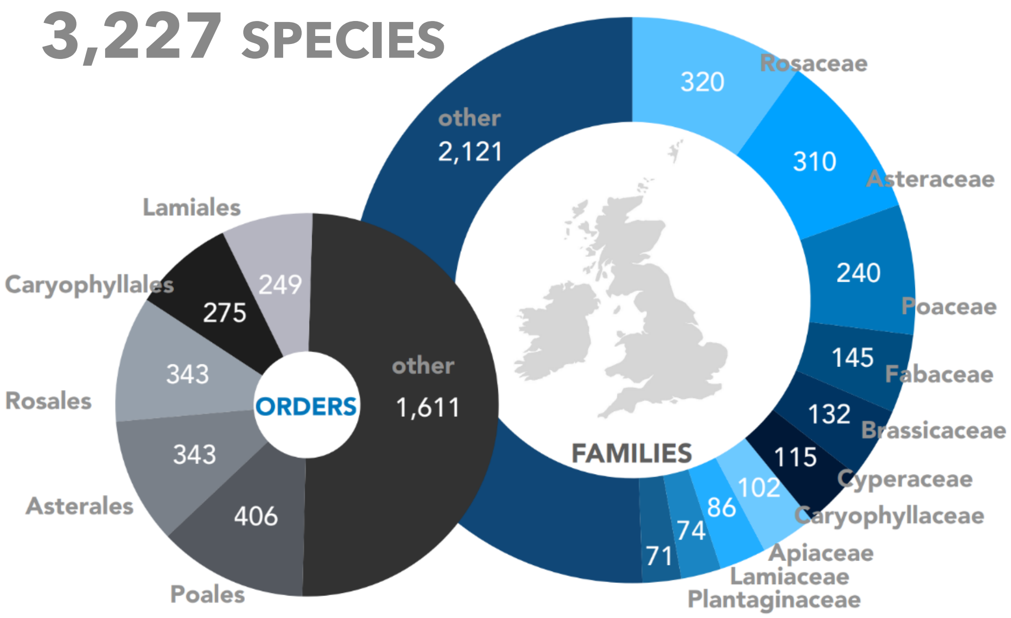


Fig. 2.4 Species composition of the BI flora at the family and order level. Donut charts represent the proportion of the flora within each clade. Total species numbers per clade are given within each segment of the charts.

While many previous checklists focused entirely or in part on representing the native flora, this database demonstrates that this reductive view misses more than half of the diversity currently present within the flora (Fig. 2.5), when considering status through the lens of the *New Flora of the British Isles*. Native species make up the largest single group within the flora (1,407 species), but the larger proportion is made up of non-native species (1,686 species). Of the latter, the comparatively small subgroup of archaeophytes (181 species) is dwarfed by the much more prominent subgroup of neophyte species (1,505 species), which are comparatively recent introductions to the flora (arrived within the last 500 years). The largest group within the neophytes is that of naturalised neophytes (936 species), species that have not only been introduced to BI but are also thriving. There are only four species that have the rarer status of neonatives, and 130 species are not assigned a clear status.

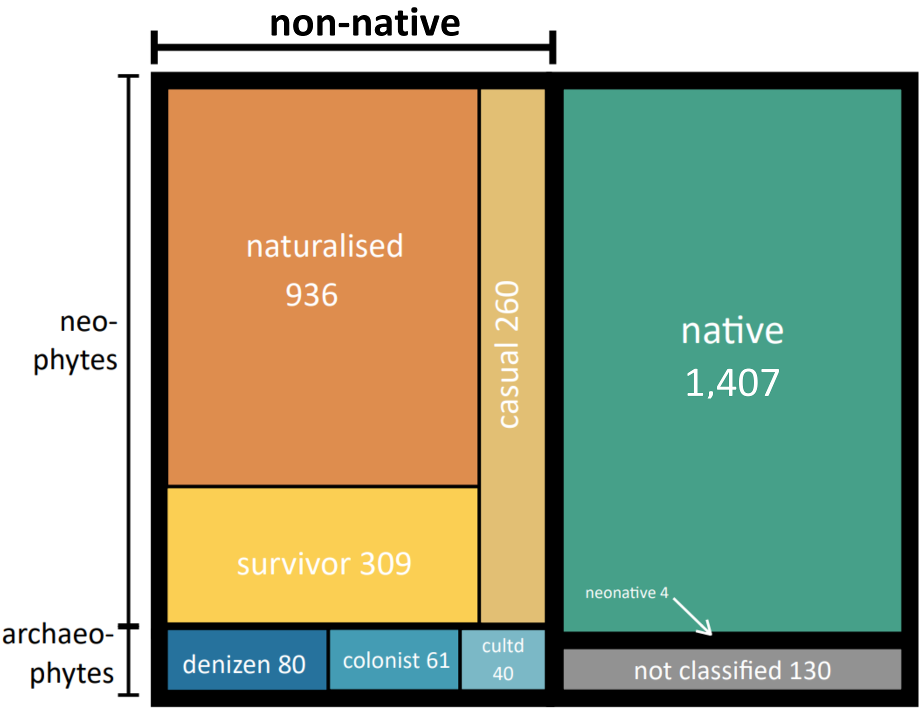


Fig. 2.5 Status of species within the BI flora. The treemap representation shows the hierarchical subdivision of the flora into native and non-native (alien) plants. Non-native plants are further subdivided into neophytes and archaeophytes, which in turn are split into naturalised, survivor and casual neophytes as well as denizen, colonist and cultivated (= cultd) archaeophytes (see Table 2.1). Four neonative species form too small a group to be discernible in this representation. While the largest single group is that of native plants, the higher level group of non-native plants encompasses more species overall.

Origin of introductions

Non-native species are introduced to Britain from all across the globe (Fig. 2.6), but the majority of them, 719 species, stem from other parts of Europe. Further common areas of origin are Temperate Asia (434 species), North America (259 species) and Africa (204 species). Smaller numbers of species were introduced from Southern America (119 species), Australasia (72 species) and Tropical Asia (28 species). Within those broader areas of origin, Southern Europe and the Mediterranean (184 species) stand out as common individual places from where many species have been introduced.

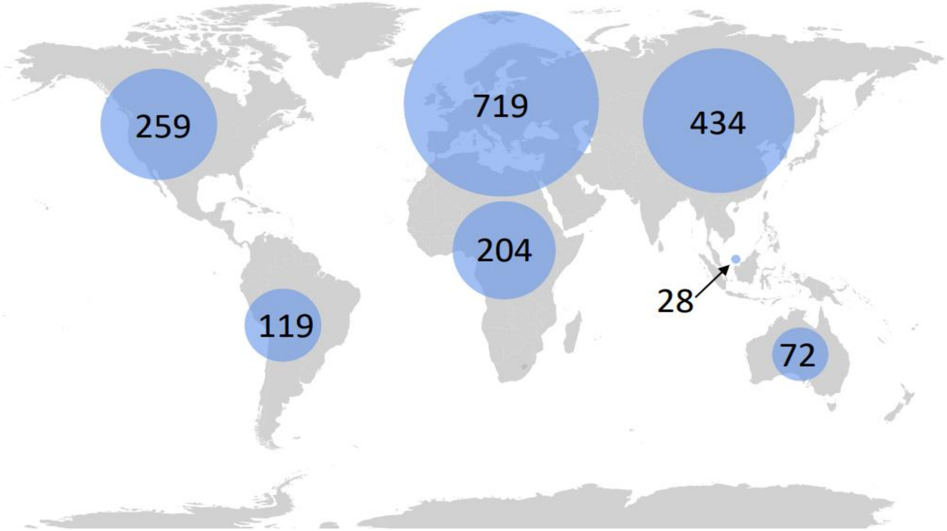


Fig. 2.6 Origin of 1,487 species that are not native to BI at TDWG Level 1. The bubble plot represents the number of species introduced to BI from each of the TDWG continental areas, with size and labelling of the bubbles proportional to the number of species from each location.

The biomes that species within Britain are commonly associated with range from Mediterranean to Arctic-Montane (Fig. 2.7), but the vast majority of BI's vascular plants prefers Temperate (537 species), Southern Temperate (269 species) and Boreo-Temperate biomes (228 species). The number of species typically found within warmer biomes (Mediterranean and Mediterranean-Atlantic) exceeds the number of species with a preference for colder conditions (Boreal Montane, Arctic Montane, Boreo-Arctic Montane and Wide Boreal), with 253 and 222 species respectively.

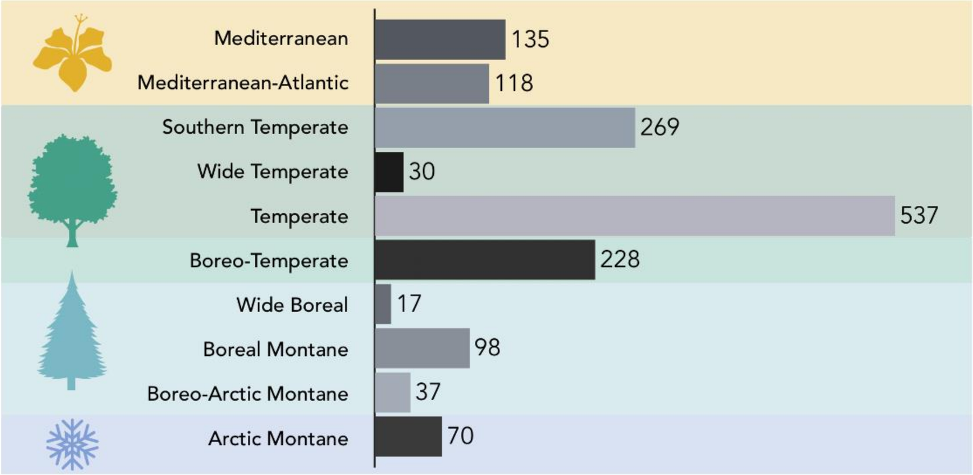


Fig. 2.7 Biomes associated with 1,531 species within the flora for which biome data were available.

Genome sizes

Genome sizes within the BI flora show the same characteristic skew towards small genomes that has been observed for all species (Fig. 2.8).

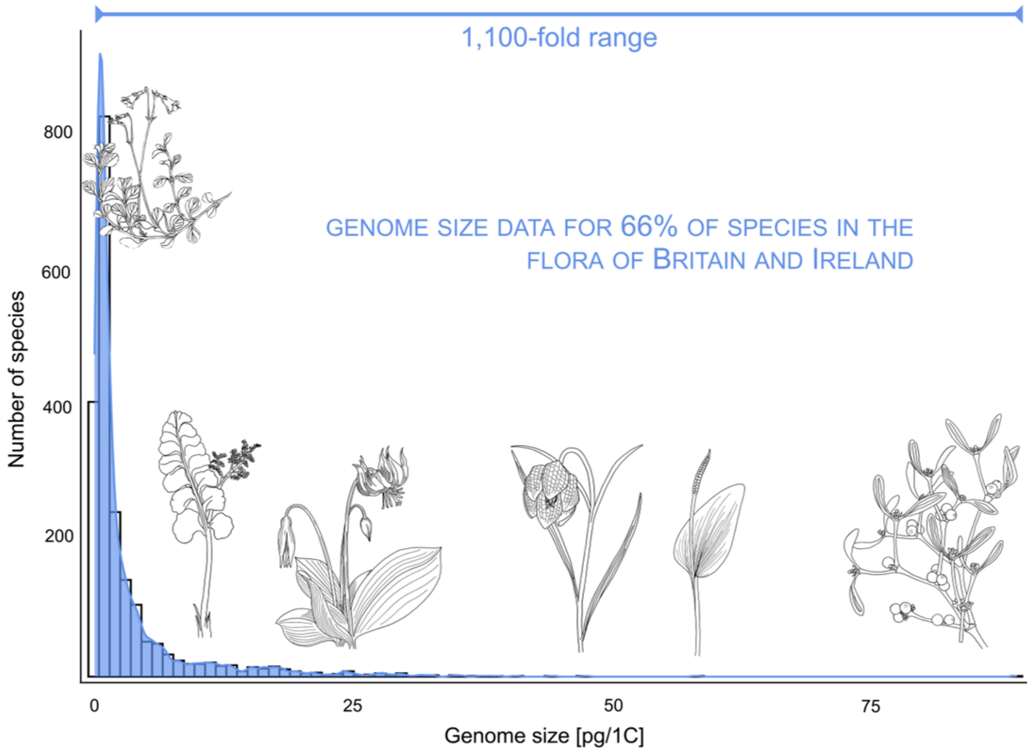


Fig. 2.8 Histogram and smoothed density of genome size data for vascular plants of BI. The number of species is plotted by genome size in picograms [pg] per haploid genome [1C] based on the 66% (2,117 species) of the native and non-native flora of BI for which data is currently available. The plants represented along the histogram are located close to their genome size and are, from left to right: *Linnaea borealis* L. (0.81 pg/1C), *Botrychium lunaria* (L.) Sw. (12.10 pg/1C), *Erythronium dens-canis* L. (24.99 pg/1C), *Fritillaria meleagris* L. (47.30 pg/1C), *Tulipa sylvestris* L. (58.00 pg/1C), *Viscum album* L. (88.90 pg/1C).

While the range of genome sizes from the smallest (*Selaginella selaginoides* (L.) P.Beauv., 0.08 pg/1C) to the largest (*Viscum album* L., 88.90 pg/1C) is remarkable, the vast majority (1,761 out of 2,117 species with data) of species have genome sizes that do not exceed 5 pg/1C.

Even though most genome sizes in the BI flora are small, some clades are characterised by a tendency towards larger genomes (Fig. 2.9, for a high resolution image see Fig. S2.1, the phylogenetic tree is available in Method S2.1). Notably, the far smaller groups of Lycophytes, Monilophytes and gymnosperms have larger genomes than angiosperms, with means of 4.56 pg/1C, 10.33 pg/1C and 17.56 pg/1C respectively compared to 2.88 pg/1C in angiosperms. Despite this, the largest genome present within the flora is that of an angiosperm, *Viscum album* L., while the smallest genome of *Selaginella selaginoides* (L.) P.Beauv. falls into the group of Lycophytes.

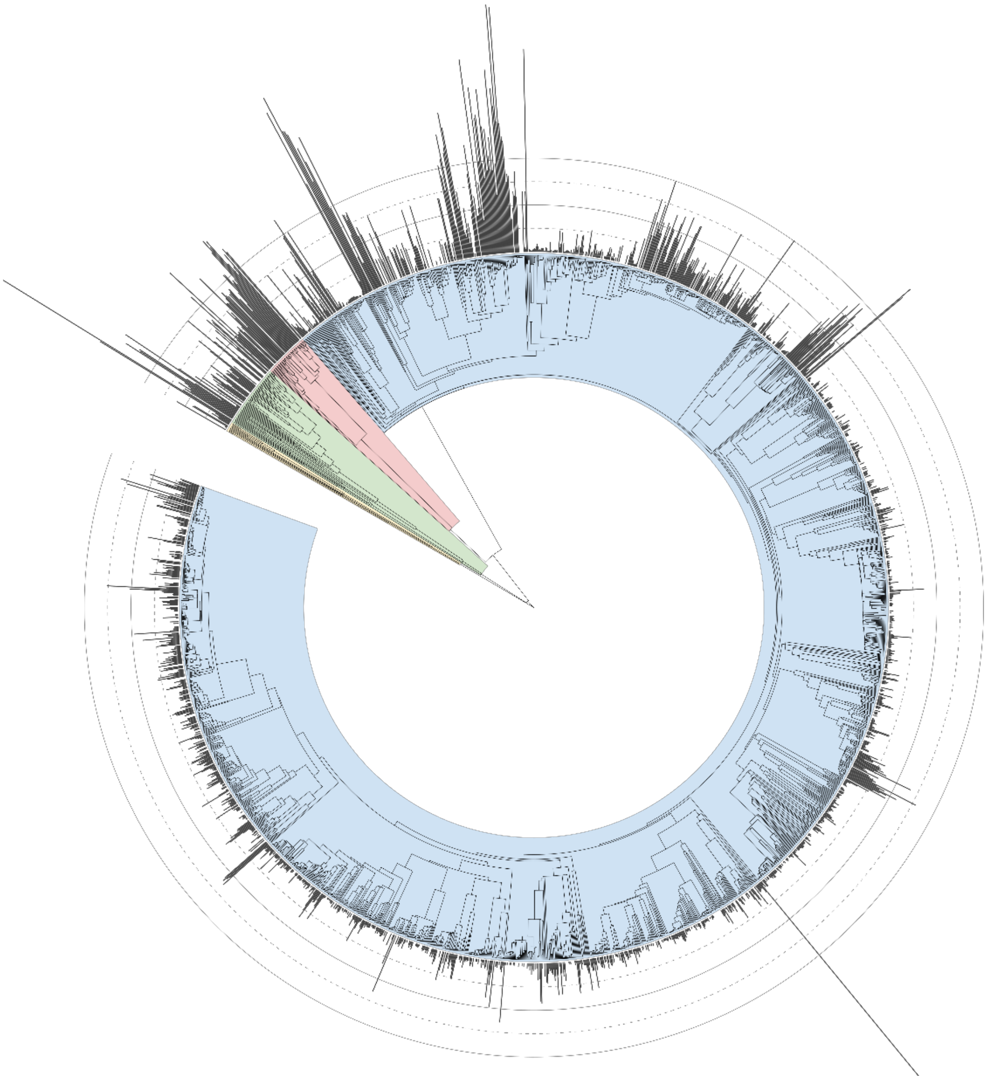


Fig. 2.9 Visualisation of the BI phylogeny and genome sizes. The circular representation of 2,501 species with phylogenetic information includes colour coding for the different clades, with Lycophytes in yellow, Monilophytes coded in green, gymnosperms in red and angiosperms overlaid in blue. The smallest known genome size for each species is plotted around the outside in pg/1C with gridlines at 5, 10, 15 and 20 pg for orientation. Lycophytes, Monilophytes and gymnosperms have larger genome sizes overall, but the overwhelmingly largest genome of the flora, that of *Viscum album* L., an angiosperm, is visible on the bottom right with a genome size of 88.90 pg/1C.

Genome sizes differ between plants of different status (Fig. 2.10) in the UK. Overall, neophytes have significantly larger genomes than both natives and archaeophyte species ($p < 0.001$), according to a pairwise T-test for multiple groups with a Bonferroni correction. Among archaeophytes, cultivated species have significantly larger genomes than both denizen and colonist species. Both naturalised and survivor type neophytes have larger genomes than casual neophytes.

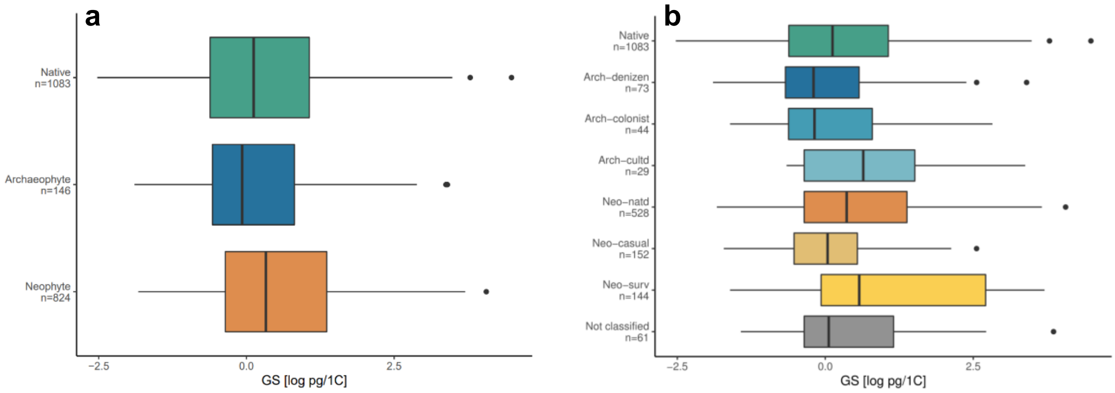


Fig. 2.10 Genome sizes by status. The boxplots show differences in genome size between the different status categories. **a** represents genome size for the three large categories (natives, archaeophytes, and neophytes). **b** splits the latter two groups into their constituent subgroups (denizen, colonist, cultivated (= cultd), naturalised (= natd), casual, and survivor (= surv) as well as adding a category for those species that had no categorisation for status. Group sizes are given with labels. Neonatives were omitted due to a low number of species (n = 3). Both native and archaeophyte species have significantly smaller genome sizes than neophytes ($p < 0.001$), but there is no significant different between the genome sizes of natives and archaeophytes.

Strikingly, genome sizes in the flora of BI appear to be linked with life strategy. Fig. 2.11 shows a ternary plot of Grime's Competitor – Stress-tolerator – Ruderal (CSR) classification for species in BI, where each species is assigned to a position between three poles representing the three life strategy characters of competitive, stress-tolerating or ruderal (i.e. weedy). Numeric CSR scores that were used to generate the figure are presented in Table S2.4. Most species pursue a mixed life strategy that incorporates varying levels of each of these three strategy characters. Centroids representing the average position of plants within the genome size quintiles, ranging from very small to very large genome sizes (very small: 0.15 - 0.53 pg/1C, small : 0.54 - 0.90 pg/1C, medium: 0.91 - 1.59 pg/1C, large: 1.60 - 4.18 pg/1C, very large: 4.3 - 47.3 pg/1C), reveal that plants with smaller genome sizes tend towards a ruderal strategy, while the largest genome size groups show increasing tendencies towards a more competitive and marginally more stress-tolerant lifestyle.

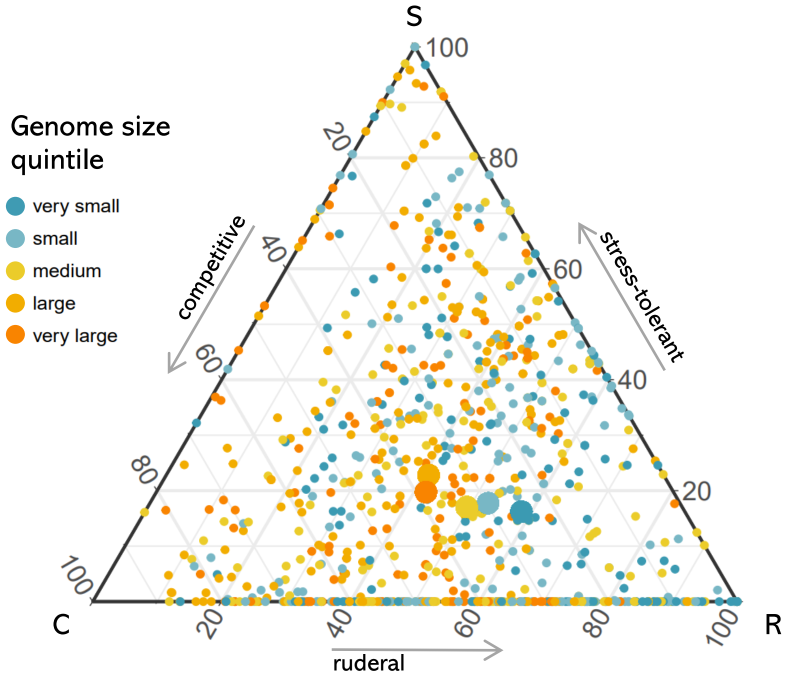


Fig. 2.11 Ternary plot of CSR strategy and genome size quintiles for 915 species. Strategies of different species are characterised by proximity to three poles: competitiveness, stress-tolerance and weediness (ruderal). Dots represent a plant's location with respect to all three poles with number from 0 to 100 along the outside indicating the score along each axis. E.g. species with exclusively ruderal life strategies are located at the far bottom left of the diagram with a score of 100 for ruderal and 0 for each of the other options. Colours indicate the quintile of genome size a species falls within (very small: 0.15 - 0.53 pg/1C, small: 0.54 - 0.9 pg/1C, medium: 0.91 - 1.59 pg/1C, large: 1.60 - 4.18 pg/1C, very large: 4.3 - 47.3 pg/1C). Larger dots represent the centroid of all species within each genome size quintile. With increasing genome size, species are less likely to be ruderals and more likely to follow a competitive life strategy. The large quintile shows slightly higher proclivity towards stress-tolerance.

Discussion

The results above are a glimpse of the extensive diversity encountered in the BI flora and highlight the value of having such an organised and comprehensive resource as the BIFloraExplorer dataset for a wide variety of analyses. Following in the footsteps of other flora-wide databases such as Pladias (Czech flora, Chytrý *et al.,* 2021) and AusTraits (Australian flora, Falster *et al.,* 2021), the dataset has the potential to boost research investigating the dynamics of the BI flora.

A flora of immigrants

BI's flora is an impoverished one with only 3,227 species, of which 1,407 are natives. This number is dwarfed by the extremely specious Australian flora (~28,900 native taxa) but also by continental European floras such as the Czech or German flora with around five thousand and seven thousand accepted taxa respectively (Wild *et al.,* 2019; Netzwerk Phytodiversität Deutschland & Bundesamt für Naturschutz, 2013). It is not unusual for floras in north-western Europe to have limited numbers of native species since repeated glaciation cycles have impacted the area and depleted its diversity (Ingrouille, 2012). This past is shared by BI, two thirds of which, with the exception of southern England, were covered by ice during the Last Glacial Maximum, 27,000 years ago (Ehlers & Gibbard, 2004; Clark *et al.,* 2012), the last remnants of which lasted until 11,300 years ago (Small & Fabel, 2016). While BI's soils and ecosystems are clearly still heavily impacted by this recent glaciation, the archipelago's sealocked nature presents another reason for the sparse species numbers. As the ice sheet retreated, sea levels rose, and after the Irish Sea first separated Ireland from Britain, the English Channel then separated the British Isles from the European mainland. Thus, BI were cut off from the continent approximately 8,500 ago (Preece, 1995) and as the rest of Europe was quickly repopulated by immigrant species, BI lagged behind (Ingrouille, 2012). The BI flora must therefore be considered a flora not only of immigrants, but of recent immigrants. For this reason, the inclusion of both native and non-native species within this database is of particular importance to gain some insight into dynamics within cohorts of plants that have arrived in BI at different times. While previous inventories of the BI flora have focused mostly or entirely on native species (e.g. Hill, Preston & Roy, 2004), it is clear from the results above that this approach leaves more than half of the flora unaccounted for. There is increasing interest in characterising the spread and movements of non-native species across BI. This tendency is reflected in excursion floras; in 1952, Clapham, Tutin & Warburg's flora was overwhelmingly focused on natives while Stace's *New Flora of the British Isles* lists the greater number of non-natives also present. This change in realisation of the importance of non-natives is also manifest in the increased reliability of non-native species records (both presence and absence) within the BSBI's distribution database since the 1980s (pers. comm. Kevin J Walker). As awareness of the dangers of plant invasions grows, so does the importance of understanding the non-native species in the flora (Kowarik & Lippe, 2008; Chytrý *et al.,* 2009; Pyšek *et al.,* 2022; Clements *et al.,* 2022). While a flora increasingly dominated by non-natives may sound like a change for the worse, research seems to indicate that with some exceptions (Manchester & Bullock, 2001), the new arrivals in the flora may actually be a welcome addition to an impoverished flora with little to no negative consequences for overall biodiversity (Maskell *et al.,* 2006; Thomas & Palmer, 2015).

In distinguishing natives and non-natives it must be stressed that in many cases, especially in BI following their tumultuous geological past, a native may simply be an immigrant species that arrived before any human record or observation existed to document their arrival, while a non-native may simply be anything that arrived subsequent to documentation (Webb, 1985).

Beyond a simple split into natives and non-natives, it can be seen that new arrivals into Britain are introduced from locations across the globe (Fig. 2.6), reflecting the role that globalisation and international trade have played in contributing to the composition of BI's flora (Hulme, 2009). Due to BI's characteristic humid temperate climate it is not surprising that the majority of plants in its flora favour temperate climes. Remarkably, while BI is considered to be lacking in habitat variation when compared to other parts of Europe which exhibit, for example, extremes in altitude (e.g. the Alps) and aridity (e.g. Mediterranean regions), there are, nevertheless, species within BI that favour the conditions present within both Mediterranean and Arctic biomes (Fig. 2.7, examples are *Arabis alpina* L., *Euphrasia frigida* Pugsley and *Veronica fruticans* Jacq. for Arctic Montane biomes and *Centranthus ruber* (L.) DC., *Datura stramonium* L. and *Fuchsia magellanica* Lam. as representatives of Mediterranean biomes), a phenomenon that will be further explored in the following chapters.

BI as a case study of genome sizes

Although the BI flora includes only a small fraction of the global plant biodiversity (approximately 308,312 vascular plants according to Christenhusz & Byng, 2016), species within BI with genome size data show they range nearly half (i.e. 1,100-fold) of the total ~2,400-fold range of genome sizes described for vascular plants as a whole (Leitch & Leitch, 2013). Genome size diversity in BI also mirrors the characteristic skew towards smaller genomes that has been observed at a global scale (Dodsworth, Leitch & Leitch, 2015). Visualisation of the genome size data on the phylogeny of the BI flora also indicates the presence of a strong phylogenetic signal in the genome size dataset. Such signal reveals the need to account for species phylogenetic structure in any analyses involving genome size data (Borges *et al.,* 2019).

Interestingly, the large group of neophyte species have significantly larger genomes than those species that are native or have existed in the UK for a longer period (archaeophytes). This means that species with larger genomes have been entering the flora. How this affects genome sizes across species assemblies in different regions of the study area is explored in Chapter 4.

Genome size has previously been shown to constrain life strategies. For example, Guignard *et al.* (2016, 2019), found within controlled field plots that high levels of nutrients favoured competitive species with higher ploidy levels and larger genome sizes. However, such a trend is complicated by observations that plants with extremely large genome sizes are more likely to be limited to stress-tolerant, slow-growing lifestyles (Bennett, 1972). Meanwhile species with smaller genomes have been associated with weediness and consequently greater invasion success (Suda *et al.,* 2015). Such results are mirrored in the findings for the BI flora (Fig. 2.11), where the quintile centroids of species with large and very large genome sizes lean towards competitive strategies whereas species within the small and very small quintiles are more ruderal (i.e. weedy) in their life strategy. As a potential meta-trait with constraining effects on a variety of plant traits and characters (e.g. Roddy *et al.,* 2020; Théroux-Rancourt *et al.,* 2021; Šímová & Herben, 2012; Bennett, 1971), genome size emerges as an interesting character which warrants further study in the context of the BI flora.

Chapter 3 Tackling sampling biases in the current knowledge of the British flora

Abstract

Although the record base for vascular plants in Britain and Ireland is extensive and well curated, it is fraught with biases that can skew findings obtained from it. Of particular note are spatial differences in recording effort linked with accessibility, changes in recording intensity over time, and the fact that inconspicuous and introduced species are often severely under-recorded.

Species distribution information at 10x10 km resolution for the most recent three date classes (1987-1999, 2000-2009 and 2010-2019) from the Botanical Society of Britain and Ireland's Distribution Database is used as the underlying information. Based on this data, I generate an updated dataset with the help of a frequency scaling method, accounting for biases from uneven sampling effort in time and space.

The resulting dataset conserves broad trends within the original data with regard to overall species numbers following a latitudinal diversity gradient, with most species in the South and species richness declining towards the North. I present bias-corrected diversity estimates for 3,136 plant species and illustrate the differences between raw and corrected estimations by focusing on three species of different status.

The approach's strength is particularly evident in the context of high human presence in the South of the study area where higher recording effort and the effect of garden and agricultural escapes would confound future analyses if not explicitly addressed. The resulting bias-corrected dataset presented here, although not perfect, allows for higher levels of confidence in any results derived from analyses of the British flora.

Introduction

A wealth of modern and historic species distribution data is available for the vascular flora of Britain and Ireland. The Botanical Society of Britain and Ireland (BSBI) Distribution Database (DDb) holds and curates this ever-expanding record base. Datasets are fed into this repository from the Vascular Plant Database (VPDb), from databases of county-wide recording, expert survey data and citizen science projects, leading to a total of over 40 million records to date (BSBI website, 2022; Amphlett, 2015; Walker *et al.*, 2010; Pescott *et al.*, 2018; Pescott *et al.,* 2019a). New data is added continuously, including not only present-day survey data but also information from historic datasets with records at sufficient spatial resolution (Walker *et al.,* 2010). The database allows ecologists, conservationists and landowners to make use of the wealth of organised and curated species occurrence data that has become available since the first *Atlas of the British Flora* (Perring & Walters, 1962), data that has drawn interest from the public, leading to increasing numbers of volunteer recorders (Preston, 2013). Despite being well-curated, the varied nature of species occurrence records means that the dataset is fraught with several biases that need to be accounted for (Isaac & Pocock, 2015; Dornelas *et al.,* 2013).

Data within the BSBI DDb is held in a variety of spatial resolution levels, each with its unique set of advantages and disadvantages (Amphlett, 2015; Pescott *et al.,* 2018). While much of the current recording effort is focused on monad (1 km x 1 km) or tetrad (2km x 2 km) level observations, projects looking to incorporate older records can benefit from using the spatial resolution of hectads (i.e. 10 km x 10 km grid squares). The greater reliability of hectad scale data is because hectad level recording has historically been the standard method employed for creating species lists and atlas maps (Pescott *et al.,* 2018). Monad and tetrad level records are fraught with a number of spatio-temporal biases, such as the tendency of monad and tetrad recording sites to be located in easily accessible and highly populated areas, the unevenness at which different counties are adapting to recording this finer-scale data and potential issues arising from inaccurate georeferencing. Although not devoid of biases, the use of hectad level records can lessen the impact of such distortions while also allowing comparisons across time. The remaining bias of uneven recording across space and time can be further reduced by using detection/non-detection data rather than abundance information and by grouping the data by date classes, instead of using yearly records. The date classes are designed to balance differences in sampling effort over time and are congruent with periods of recording for the *Atlas of the British and Irish Flora* (Preston, Pearman & Dines, 2002; Pescott *et al.*, 2018).

While recorders for the BSBI are instructed to follow unbiased sampling strategies (Groom *et al.*, 2011), inherent biases typical of biological records do exist within the data (Isaac & Pocock, 2015), not least due to the differences in historical and contemporary recording practices. While the choice of date class and hectad level data helps alleviate some of these biases, formal measures of bias correction are important to derive meaningful insights into trends, i.e. changes over time, from the datasets (Stroh *et al.,* 2014).

The Frescalo method (FREquency SCAling LOcal; Hill, 2012) was developed in the context of biological recording in Britain as a more sophisticated alternative to previously employed methods for bias correction available at the time, such as moving averages, extrapolations and the simple regression technique of the *Telfer* method (Telfer, Preston & Rothery, 2002; Isaac *et al.*, 2014; Rich & Karran, 2006; Groom, 2013a). While not the newest method for the correction of biases and calculation of species trends (Andermann *et al.,* 2022; Engemann *et al.,* 2015), Frescalo is often considered the tried and tested choice (Pescott, Powney & Roy, 2016; Groom, 2013a), and has been suggested to offer the most reliable results when used with data from the BSBI DDb by its maintainers (pers. comm. Kevin J Walker). Frescalo uses information about site similarities and proximities to select locally specific benchmark species (Hill, 2012). The presence or absence of the local benchmark species in a hectad is then used to calculate an estimation of sampling intensity as well as deriving a relative frequency of a given species at a given time (Hill, 2012). The approach allows inferences on time and site-specific sampling intensity as well as calculating location-specific likelihoods of occurrence and a trend for each species across the periods of time provided to the program.

This chapter outlines the acquisition and Frescalo-based bias correction of BSBI distribution data that all analyses in Chapters 4 and 5 are based on. I present observed species richness patterns and juxtapose them with those inferred by Frescalo. The clear advantage of using Frescalo-corrected data is demonstrated on three example species and potential drawbacks of the method are discussed.

Materials & Methods

Data acquisition

In February 2022, I downloaded hectad level detection/non-detection records for all 3,227 species listed within the 'BIFloraExplorer' (Henniges *et al.,* 2021 & 2022) during the most recent three date classes (1987-1999, 2000-2009, and 2010-2019) for the vice-counties VC1-113 (vice counties of Great Britain; Watson, 1883) and VCH1-40 (vice counties of Ireland; Webb, 1980), reflecting England, Scotland, Wales, Northern Ireland, the Republic of Ireland, the Isle of Man and the Channel Islands. The decision to restrict this analysis to only the three most recent date classes allows for a relatively high level of confidence in the sufficiency of records of non-native taxa (see glossary; Table 2.1) starting with the sampling period for the BSBI *New Atlas* (Preston, Pearman & Dines, 2002; Preston, 2002).

Since the BSBI DDb is not fully aligned with the nomenclature of Stace (2019), species names without a match in the database were checked manually to find any alternative spellings and mismatches. A total of 3,197 species had a clear, unambiguous match and were included in the query. Those species without a clear match were not used for this analysis since inconsistencies in their nomenclature and changes therein would make findings for those species not meaningful. That being said, the curation of the BSBI distribution database means that records from the most recent date classes with clear matches are less affected by biases due to nomenclature changes than would be expected in most comparable datasets (Dornelas *et al.*, 2013). Record grouping parameters were species and hectad. To exclude spurious records, I filtered out data points where grid-references did not align with vice-county boundaries (within 2 km) and such records that bore the 'do-not-map' label, indicating low levels of trust for an observation of a species in the wild, as opposed to a record of a cultivated species (e.g. in the context of a garden). All decisions were based on discussions with the maintainers and expert users of the database (pers. comm. Kevin J Walker, Oliver L Pescott, Tom A Humphrey).

Frescalo correction

I used a rendition of Hill's (2011) original program that was adapted for use within R (R version 4.1.3) in the package 'sparta' (August *et al.,* 2015), developed by the Biological Records Centre (https://www.brc.ac.uk/home). Frescalo was run on the records available for all species for the date classes 1987-1999, 2000-2009 and 2010-2019. Frescalo uses information on the spatial proximity and biological similarity of hectads to determine if observed differences in species occurrences are likely to represent reality or artifacts associated with differences in sampling effort. The result of this assessment is reflected in weights that are used to derive the bias-corrected outputs of Frescalo. I used the custom weights file ("LCUK") included in the 'sparta' package, based on remotely sensed land cover for the United Kingdom (including Northern Ireland, but excluding the Republic of Ireland), for this purpose. Hence the biological similarity is derived by the program based on the composition of a hectad's land cover. The alternative option of using vascular plant coverage was rejected since this may have introduced circularity into the inference.

The lack of land cover data for the Republic of Ireland from the same source means that this area is not included in the analysis presented below. I set the percentage of expected species within each hectad to be treated as benchmarks to 15% (alpha), following the settings used by Stroh *et al.* in their Red List for vascular plants in England (2014). This value is lower than the default suggested by Hill (2012), meaning that a shorter list of taxa are expected to be representative of a given hectad with their absence leading to the assumption that a hectad was under-recorded (i.e. a less strict definition of under-recording). I did not specify a list of species to be excluded as potential benchmarks, instead allowing the program to consider all species within the run as local benchmarks to avoid adding bias to the analysis. Although Hill (2012) notes that the setting of phi, the target frequency of frequency-weighted mean frequency, is not crucial to the successful application of Frescalo, I opted to follow best practice, running a trial iteration and raise phi according to the trial run's findings for the final run (phi = 0.80). The program was set to calculate decadal change with the default setting of arithmetic change. I inspected and compared the results of the runs with information presented for the 2014 Red List of vascular plants for England (Stroh *et al.,* 2014) as a 'sanity check'. For this purpose, I compared the spatial patterns observed in the overview maps (Fig. 3.1) from the Red List assessment's runs with my own. I also considered individual species in detail, comparing the results of Frescalo runs (occurrence maps and TFactor regressions) with information contained within the Red List assessments.

Data management, visualisations and statistics

Data management and manipulation was carried out in R (version 4.1.3) using the 'tidyverse' framework (Wickham *et al.,* 2019). Frescalo application was set up in R but computation was performed in Frescalo.exe (Hill, 2011). The initial reports from the Frescalo runs are presented here as outputted by the internal mapping script of the program with post-processing in Microsoft Powerpoint. Species line drawings were created in Sketchbook (raster graphics software).

Results

The Frescalo-based correction for sampling effort did not change the overall pattern of species richness observed across the United Kingdom. Higher levels of species richness were found in the South and fewer species in the North (Fig. 3.1). This pattern is reinforced by a similar trend towards lower sampling effort in the more remote areas of northern England and Scotland as well as Wales and the eastern coast of Northern Ireland, while most of England is comparatively well recorded (Fig. 3.1b). Consequently, the effect of the sampling effort multiplication (i.e. Frescalo correction) at a whole flora scale is more visible in the overall increase in assumed species numbers per hectad rather than in changes to the overall pattern of species richness across the study area (i.e. compare Fig. 3.1a & c). Note for Fig. 3.1c, the local scaling factors mapped in Fig. 3.1b are not a simple multiplication factor to be applied to species numbers in Fig. 3.1a. Instead the scaling factors are considered by the algorithm in determining the occurrence likelihood of each species within the hectad in question. Fig. 3.1c is therefore a representation of approximate species numbers after the scaling factor has been applied in the bias-corrected occurrence likelihood of each species.

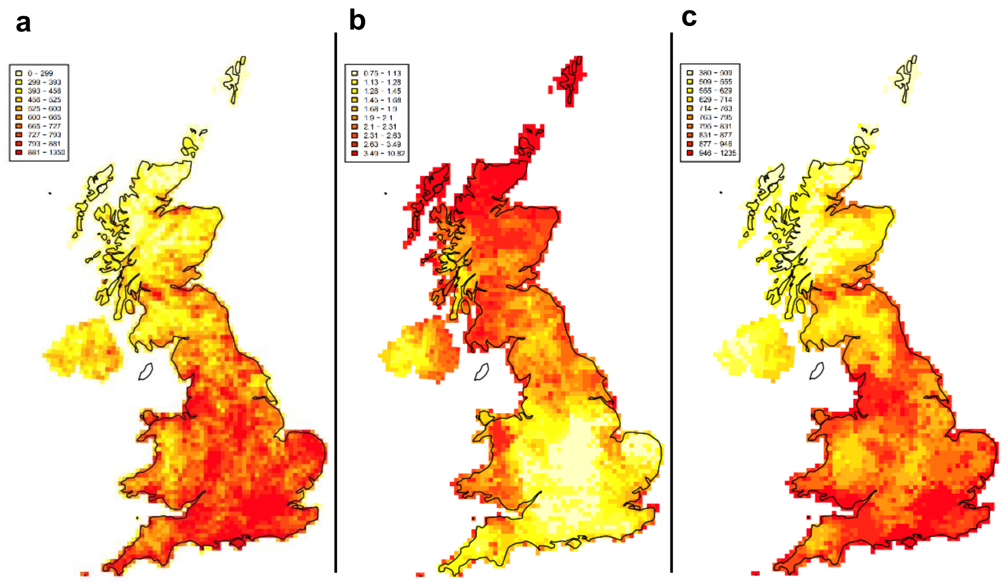


Fig. 3.1 Species richness and sampling effort output from Frescalo. a represents the recorded species numbers obtained from the BSBI. Increasingly red hues correspond to higher species numbers. **b** shows alpha, i.e. the local scaling factor to be applied to remediate the effects of uneven sampling. Yellow areas are comparatively well recorded while areas in red are affected by under-sampling. **c** shows the species numbers following Frescalo correction, with darker red areas again indicating locations with higher species richness. Each map represents a summary across the three date classes rather than reflecting on species numbers or scaling factor developments over time.

Frescalo estimates the adjusted likelihood of a species occurrence in a particular site at a particular time, given the estimated sampling intensity of the location as estimated via the occurrence of local benchmark species, which are represented as aggregations across the three date classes in Fig. 3.2. In addition to these occurrence likelihood profiles, Frescalo also calculates a 'Tfactor' (=time factor), for each date class, which reflects the detection probability of the focal species relative to the benchmark species it co-occurs with, averaged across all hectads. This contrasts with the local scaling factor alpha (Fig. 3.1b), which is calculated for a specific location rather than for a given species. A regression is then performed on the Tfactors with the resulting slope representing the magnitude of increase or decrease in relative abundance a species has undergone (Fig. 3.2). The estimate of trend used here and analysed in Chapter 5 was calculated as the arithmetic decadal change across all queried date classes. A full summary of Frescalo-corrected species distributions, hectad occurrence likelihoods and trends can be viewed in the appendix (Fig. S3.1; Tables S3.1 – S3.4), but three species are presented here to illustrate the effect of Frescalo corrections (Fig. 3.2). The three examples are chosen to reflect the different status categories present within the flora, a native species, an archaeophyte and a neophyte (glossary; Table 2.1).

Platanthera bifolia (L.) Rich. (Fig. 3.2a), a member of the Orchidaceae and native to the UK, is distributed across the study area, but appears to show a hotspot of occurrence in northern and western Scotland. This trend is evident in the occurrence data but is re-emphasised by the Frescalo correction, which bolsters the records made within similar and proximal hectads with low recording effort in northern Scotland, while reducing the likelihood of occurrence in hectads where records for the species exist, but the land cover within the hectad and the larger recorder effort suggest the potential for less characteristic records (i.e. records in areas where species observations may be indicative of frequent reintroduction due to escapes from gardens and high recorder effort rather than demonstrating the regular wild occurrence of the species). Across the three date classes within this analysis, *P. bifolia* shows a decreasing trend, meaning that its frequency across the UK – relative to the frequency of benchmark species typical of those hectads within which it occurs – has decreased steadily over time.

Borago officinalis L. (Fig. 3.2b), part of the Boraginaceae and cultivated archaeophyte with clear archaeological evidence placing it within Britain before the year 1500 (Preston, Pearman & Hall, 2004) and a strongly suspected Mediterranean background (Asadi-Samani, Bahmani & Rafieian-Kopaei, 2014), has been renowned as a source of courage in

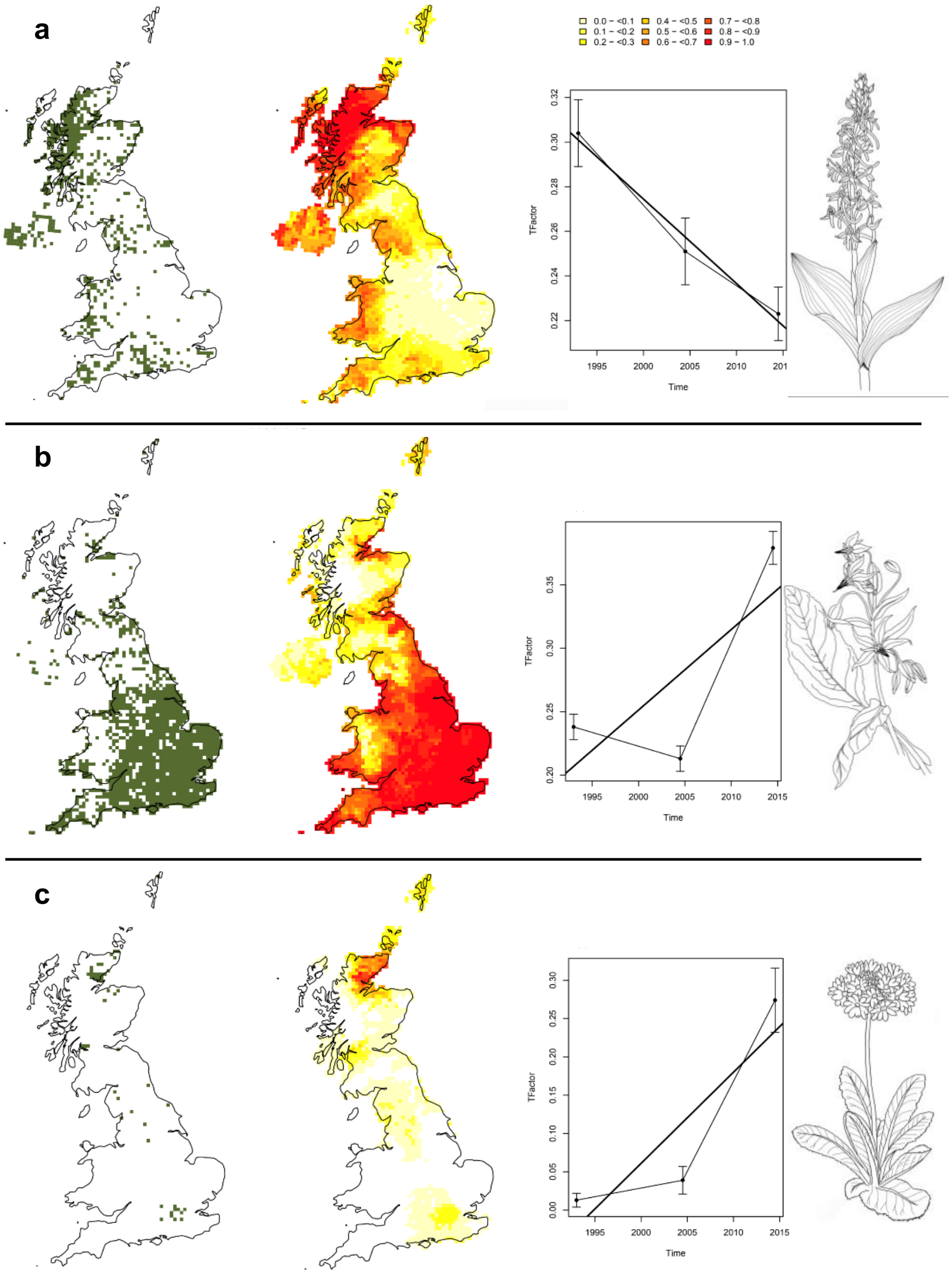


Fig. 3.2 Examples of Frescalo results for three species. Representative examples of the Frescalo output for **a**) the native species *Platanthera bifolia* (L.) Rich., **b**) the archaeophyte *Borago officinalis* L. and **c**) the neophyte *Primula denticulata* Sm. All maps display information across all three date classes. Maps in green on the left highlight all hectads where the species has ever been recorded since 1987. Red and yellow maps in the middle represent the Frescalo adjusted likelihood of an occurrence of the species in the given hectad across all three date classes. Plots on the right show the estimated relative frequency of the species (time factor = TFactor) during each date class. The trendline indicates the change across time as calculated by Frescalo. Line drawings show the habitus of each species.

ancient Rome (Fernie, 1890) and is still of value as a culinary and medicinal herb (Lozano-Baena *et al.*, 2016). Its early introduction to Britain is therefore not a surprise. Its occurrence is largely limited to southern England and the Midlands, with only a few occurrences further north and the only Scottish records limited mainly to the eastern coast. The Frescalo-derived trend of this species shows an initial decrease in relative frequency between the first (1987-1999) and second date class (2000-2009), but then a rapid increase in the most recent date class (2010-2019).

Primula denticulata Sm. (Fig. 3.3c), is a naturalised neophyte of the Primulaceae family. This species shows a sparser distribution than the other species above but has a strongly increasing trend across the three date classes. It has a recording hotspot in mountainous regions near Inverness with sporadic additional records scattered in the vicinity of various urban areas, most notably a cluster of occurrences in Greater London. Frescalo attributes more weight to the cluster of occurrences in northern Scotland than to the sporadic urban records.

Discussion

Even though the British flora is comparatively well described thanks to an organised and long recording history, the same flaws inherent in other biological recording around the world apply here (Isaac & Pocock, 2015; Zizka, Antonelli & Silvestro, 2021); for example (i) the further a sampling location is from any human settlements, the fewer visits and consequently records are available, and (ii) recorders may be more biased towards recording attractive and native species, leaving less ostentatious species as well as non-native plants under-recorded (Pescott, Humphrey & Walker, 2018). In addition, (iii) sampling effort is uneven through time, meaning that both temporal and spatial biases must be taken into account (Pescott, Humphrey & Walker, 2018).

In the bias correction here and for the following chapters, the choice of hectads for spatial, and date classes for temporal aggregation is conservative – the presence of a species, even an inconspicuous one, has a much higher likelihood of being spotted and hence recorded when coarsening spatio-temporal recording ranges (Isaac & Pocock, 2015) and would therefore increase the likelihood of the species being featured in the analysis. The drawback of this decision is that effects expected at smaller scales, i.e. short-term differences in occurrences within date classes or differences between the multitude of different habitats that are amalgamated within one hectad, will inevitably be invisible to the analyses presented in the following chapters. Another point to take into consideration is the use of detection/non-detection data. Reliable, comparable and comprehensive information on species compositions for the entirety of the flora of the UK is still missing, especially over the long term, with only relatively recent efforts designed to close this gap (Pescott *et al.*, 2019b&c). Consequently, the bias-corrected information presented here (and in the appendix, Fig. S3.1; Tables S3.1 – S3.4) and used for analyses in the following chapters should be seen as a top-level overview of the species within the flora, their distribution and changes therein over the past three decades that does not reflect habitat specific species compositions and their inner dynamics.

As previously described by Rich (2006) and is expected due to differences in the density in human settlements, Frescalo suggests higher levels of under-recording in the North, while most areas of England benefit from high levels of recording activity. The patterns of species richness revealed by Frescalo correction for sampling bias are congruent with previous findings presented by Stroh *et al.* (2014) in their Red List assessment for England, but here the analysis is expanded to Scotland, Wales and Northern Ireland.

It is important to note that species richness as it is derived from the Frescalo correction is simply the number of species recorded within a hectad, scaled by a local scaling factor to account for sampling effort. Hectads are large areas that are likely to contain a multitude of different habitats and – in the UK – will in almost all cases include human dwellings, particularly in the South, where species richness (as derived from distribution records and

Frescalo correction) is highest (Fig. 3.1a & c). It must therefore not be assumed that high species richness as it is used in this work is an indication of healthy, thriving ecosystems. Indeed, previous studies on global (Newbold *et al.,* 2015) and British (McClean *et al.,* 2011) species richness and other measures of biodiversity show steady declines in response to human actions (Hudson *et al.,* 2014); in the UK, this is particularly the case in and around arable lands (Sotherton, 1998).

Instead, the high species numbers in hectads in the South and particularly in urban areas (Fig. 3.1) are likely in part caused by greater numbers of garden and agricultural escapes, and the existence of a plurality of different fractured ecosystems in each hectad, as well as potential remnants of sampling bias towards highly populated areas where fewer plants are likely to go unnoticed. Particularly the impact of garden escapes that may become temporarily or permanently established and be recorded as wild occurrences is likely significant. The flora of urban domestic gardens in Britain has previously been shown to house more than 1,000 species, with 70% them being of non-native status; this set of species, especially those surviving sporadically due to human activities, were found to result in inflated estimates of species richness in areas close to human settlements (Loram *et al.,* 2008), concurring with the findings above. Despite such biases, an underlying latitudinal diversity gradient with remarkably higher levels of species richness in the South and fewer species in the North is visible (Fig. 3.1c). This gradient in species numbers has been well established at different scales across the globe and across multiple eukaryote groups (e.g. Hillebrand, 2004; Lamanna *et al.,* 2014).

The importance of the Frescalo correction, with its combination of proximity and similarity comparisons for hectads, is especially evident in the case of *Primula denticulata* patterns before and after the correction (Fig. 3.2c). The sporadic and potentially spurious records of *P. denticulata* around Greater London are very likely the result of garden escapes of this popular ornamental plant that may not persist in those locations for very long periods without human intervention. The impact of such records on subsequent analyses is reduced due to correction with Frescalo, although the taxon still remains present.

The measures outlined above mean that biases within the BSBI distribution data have been addressed using a well-tested and validated approach, particularly suitable for this data set, allowing a higher level of confidence in any findings stemming from them than would otherwise have been the case (see Chapters 4 and 5).

Chapter 4 Spatio-temporal analysis of the British flora reveals that land use changes are shifting the distribution of genome sizes, leading to an increased occurrence of species with larger genomes

This chapter is formatted for submission; however, the authors and the journal are still undecided.

Abstract

The abiotic environment in the United Kingdom has been impacted heavily by millennia of human presence. Plant genome sizes vary widely between species (at least 2,400-fold) and are believed to play a role in influencing a diversity of ecologically-relevant traits including a species' nutrient and water use efficiency. Assuming this is correct, species may respond differently to spatio-temporal differences in nutrient and water availability in the environment, depending on their genome size.

Using bias corrected distribution information on British angiosperms, climatic and nitrogen deposition data, I test the hypothesis that environmental factors influence spatial genome size distributions by plotting and modelling patterns of and hypothesised drivers behind weighted mean genome size per hectad as well as their change over the past three decades. Additionally, I explore the movement of distribution centroids of British plants since the late 1980s and explore the role that genome size and native status play in determining the magnitude of latitudinal range shifts.

Results show that hectad weighted mean genome sizes have increased by 5.5% over the past thirty years. Areas characterised by high levels of human disturbance and nitrogen pollution harbour species with larger genomes on average, but water availability correlates less strongly with the distribution of species with larger genome sizes across the generally wet temperate UK. While the majority (79.4%) of plants have shown northward range movements in the last three decades, species with larger genome sizes and especially those that are neophytes have expanded significantly further north than those with smaller genomes and natives.

These results extend previous findings from field experiments to landscape scales, demonstrating that nutrient pollution and effects of human activities can lift genome size-induced constraints on species distributions and significantly influence species movement and establishment.

Introduction

The landscapes of Britain have been shaped by human activities for millennia. Although arable farming was noted to be mostly absent from the area by Julius Caesar (Gerrish, 2022), there is ample evidence of prehistoric agricultural activities, both in the cultivation of varieties of grain and the keeping of livestock on dedicated grazing grounds (Curwen, 1927; Applebaum, 1958). Indeed, prior to the Industrial Revolution, the majority of the British population was employed in agricultural labour (Curwen, 1927). Throughout history, agricultural practices have been shaped by changes in climate (Applebaum, 1958) and have undergone steady intensification (Firbank *et al.,* 2000). Natural and semi-natural landscapes characteristic of Britain have thus developed in the presence of high levels of human influence over the past millennia.

Genome size is a fundamental plant character with significant repercussions on various aspects of plant physiology and is consequently expected to have a role in influencing the ecology of a species (Leitch & Bennett, 2007; Herben *et al.,* 2012). Previous research has shown genome size to be associated with a range of plant traits that are likely to constrain and shape ecological strategies and niche availabilities of plants (e.g. Bennett, 1971; Bennett, 1972; Masterson, 1994; Beaulieu *et al.*, 2007; Knight & Beaulieu, 2008; Veselý *et al.*, 2012; Sparrow & Miksche, 1960; Veselý *et al.,* 2013; Roddy *et al.,* 2020; Théroux-Rancourt *et al.,* 2021; see Chapter 5).

Various controlled experiments and field trials have shown that nutrient levels in the soil can shape the composition of plant communities, and that the nutrient demands of larger genomes play a key role in this dynamic (Guignard *et al.*, 2016; Šmarda e*t al.*, 2013; Walczyk *et al.,* 2019). This is because genomes are inherently costly with regard to nitrogen and phosphorus, acting as major sinks for these macronutrients (Hessen *et al.,* 2010). All else being equal, a plant that has to maintain a larger genome therefore potentially faces increased constraints under limiting nutrient resources compared with plants that have smaller genomes. Potentially therefore, differences between species' genome sizes in the British flora (1,100-fold range in the British and Irish flora, see Chapter 2) could be relevant in determining where species can grow and compete successfully depending on soil nutrient availability. In field experiments, the combined effect of nitrogen and phosphorus abundance has been shown to be associated with an increasing dominance, in terms of biomass production, of polyploid plants with larger genomes (Guignard *et al.*, 2016). If this observation applies across landscape scales, one might predict that in areas particularly exposed to high levels of atmospheric nitrogen deposition or the addition of NPK fertilisers, e.g. in the context of intensive agricultural use, this abundance in nutrients might lift genome size-imposed growth restrictions for species with large genomes and hence enable them to become established and thrive.

Genome size has also been shown to correlate positively with the size of stomatal guard cells as well as a variety of other leaf cells (Simonin & Roddy, 2018; Beaulieu *et al.*, 2008; Hodgson *et al.*, 2010; Théroux-Rancourt *et al.,* 2021; Wilson *et al.*, 2021) and is negatively associated with the density of stomata. Larger stomatal pores and intracellular spaces are often associated with lower water use efficiency (Faizullah *et al.*, 2021), and indeed, in pairwise comparisons, those species distributed in humid climates have larger genome sizes than their counterparts in arid conditions (Veselý *et al.*, 2020). However, the link between genome size and water use efficiency is complex, since the lower stomatal densities typically found in plants with larger genomes are associated with higher levels of water use efficiency. Nevertheless, there is a suggestion that levels of humidity as well as changes therein over the years may play a role in shaping patterns of species distributions, depending on genome size.

In this chapter, I examine spatial patterns and temporal trends in genome sizes and ploidy across the UK, as well as abiotic factors influencing them, drawing on species distribution, climatic and nitrogen deposition data for the past three decades. I demonstrate the drastic changes in land use across the last century and link these developments with genome size patterns associated with different land use types. Finally, I test whether genome size has played a role in influencing the extent of the northward movement of plants over the last few decades. Expanding on previous findings from tightly controlled field experiments (Guignard *et al.*, 2016; Šmarda *et al.*, 2013), I take the next step in testing for such correlations at the scale of landscapes and reveal that genome size may indeed contribute to influencing plant community composition across the UK in response to the environment.

Materials & Methods

Mapping weighted mean genome size and ploidy level

All analyses are based on the Frescalo-corrected species distributions outlined and presented in Chapter 3. Maps of mean genome size and ploidy per hectad were created for each of the three most recent BSBI (Botanical Society of Britain and Ireland) date classes (1987-1999, 2000-2009, 2010-2019). To reflect the sampling bias correction from Frescalo, I calculated hectad means for these genetic characters as the weighted mean ('smart' package; Martin, 2020). The weights used were the estimated probabilities of occurrence for a species in a hectad at a given time after rescaling relative to benchmark occurrences (available in Tables S3.2c, S3.3c and S3.4c).

There exists cytotype variability (genome size and ploidy level) amongst some plant species (Tate, Soltis & Soltis, 2005; Kolář *et al.*, 2017). This intraspecific variation had not been accounted for in the original release of the ‘BIFloraExplorer’ (Henniges *et al.,* 2021 & 2022), but was considered for the mapping of mean genome sizes and ploidies here. Each cytotype of a species with multiple cytotypes was assumed to account for an equal fraction of the overall local frequency derived from Frescalo. I compiled cytotype information for genome size, ploidy and chromosome numbers from the same sources (i.e. Šmarda *et al*., 2019; Zonneveld, 2019; Leitch *et al.*, 2019) that had supplied the genome size information within the ‘BIFloraExplorer’ dataset. Where a direct match of the species name used in the database and the individual source could not be obtained, I matched species via synonyms present within the World Checklist of Vascular Plants (WCVP, 2022). From the datasets, I manually assigned prime estimates, i.e. the most trusted genome size measurements and ploidy levels, based on the following:

1. Where multiple genome size values were available for a species, and where differences exceeded 30% of the smaller value, I assumed that the different estimates characterised different cytotypes.
2. Where differences in values were equal to or less than 30% of the smaller value, the estimates were ranked and only the most trusted value was chosen for the analysis. The most trusted measurements were assigned as follows:
    a. Values produced by RBG Kew (Kew) took precedence since this allowed me to use measurements made on known UK-sourced material that had been produced by the same team using the same equipment.
    b. If measurements were taken from publications by scientists outside Kew, then the most trusted measurement was chosen if a chromosome count and genome size estimate were published together, especially if the genome size and chromosome count had been estimated on the same plant. To prioritise the selection of genome size estimates and to keep as many values as possible from the same source, genome size estimates were

selected in the following order of priority: Šmarda *et al.* (2019), then Zonneveld (2019) and lastly the Plant DNA C-values Database (release 7.1, 2019).

c. Chromosome counts already present in the 'BIFloraExplorer' that had been confirmed by Richard Gornall were used to validate chromosome counts provided by the different datasets. Where two competing genome size measurements or chromosome counts were available and supported by equal amounts of evidence, the smaller count was chosen as the prime value.

d. Where genetic information was available at subspecies and variety level, these were also retained as prime values if there were suspected differences in ploidy.

e. If support for cytotype variation was sparse (i.e. very few or unreliable chromosome counts at different ploidy levels) in the chromosome counts supplied by Richard Gornall or by any entries in the Chromosome Counts Database (Rice *et al.,* 2015) then only the smallest genome size measurement from the prioritised source was retained.

f. Where a species had a genome size estimate but lacked information on ploidy and/ or chromosome numbers, and if sister taxa with chromosome count/ploidy data had similar (<30% different) genome sizes, then the species was assumed to have the same chromosome count and ploidy level as the sister taxa, and the assumption noted in Table S4.1. All underlying genomic information used for the compilation of this list is available in Table S4.2.

The extent of species examined in this chapter is restricted to herbaceous and graminoid, non-woody (i.e. excluding phanerophytes, see glossary; Table 2.1) angiosperms for ease of comparisons between species. This totals 1,585 species with sufficient information for Frescalo runs and with genetic information (1,698 when counting duplicates due to cytotype diversity).

Centre of mass

Using bias-corrected occurrence likelihood data (Tables S3.2c, S3.3c and S3.4c), I derived the distance and direction of movement of the centre of mass for each species' range between start (1987-1999) and end (2010-2019) date classes. Narrowly distributed species, i.e. those present in fewer than 5% of hectads (n = 150) in any date class, were excluded. The centre of mass for each species in each date class was calculated as the weighted mean latitude and longitude, with Frescalo occurrence likelihood serving as the weighing factor. I used the Haversine formula to calculate distances between centroids of the first and last date class, accounting for the curvature of the Earth, and the bearing using the 'geosphere' package (Hijmans, Williams & Vennes, 2020). The same was done to derive the distance travelled along the North-South axis only.

Environmental data acquisition and preparation

Information about three aspects of the abiotic environment was obtained to place genome size patterns and changes into a spatio-temporal context.

Climate data

I downloaded monthly mean temperature and total rainfall data from the Met Office via the CEDA Archive (https://archive.ceda.ac.uk/) for each year between 1987 and 2019

(Hollis *et al.*, 2022). Data for each month was extracted using the 'raster' package in R (Hijmans *et al.,* 2015). Mean monthly temperature and rainfall per hectad were then calculated using the 'terra' and 'raster' packages (Hijmans *et al.,* 2022), also averaging across the growing season, here defined as the period from April to July, and across all years to find the mean value per date class. The hectad shapefile used for this operation was based on the Ordnance Survey National Grids of 1936 (Ordnance Survey, 2015), made available by Roper (2015).

Nitrogen deposition data

I obtained wet and dry nitrogen deposition data from the dataset created by Tomlinson *et al.*, 2020 and 2021, downloaded from the UKCEH Environmental Information Data Centre (https://eidc.ac.uk/). Annual mean deposition values per hectad were extracted using the R package 'sp' (Pebesma *et al.,* 2012), for the four different deposition types available (NHx dry, NHx wet ('dry and wet deposition of reduced nitrogen'), NOy dry and NOy wet ('dry and wet deposition of oxidised nitrogen')) and then averaged across date classes. The subtypes of wet and dry deposition were added to form total wet and total dry nitrogen deposition values. Since the nitrogen deposition dataset only dates back to the year 1990, the means for the first date class (1987-1999) only incorporate information from 1990 onwards.

Land cover maps

The UK Centre for Ecology and Hydrology (UKCEH) have used satellite imagery to publish detailed land cover maps (LCMs) since 1990, with further releases at increasingly regular intervals. In order to reflect the three date classes, I downloaded land cover maps for the year 1990 (first date class), 2007 (second date class), 2017 (third date class) as well as an even more recent map for 2020 from the EDINA Digimap service (https://digimap.edina.ac.uk/). The maps were processed in QGIS (https://qgis.org/en/site/), where I used 'Zonal Statistics' to calculate the majority land cover type within each hectad. The land cover map of 2007 had two additional land cover categories ('montane habitats' and 'rough grassland') that were not part of the classification on earlier and later land cover maps. To avoid problems in making direct comparisons, I removed any hectads exhibiting these extra classes as the majority land cover (133 hectads removed).

Historic land use

For a look further into the past, I utilised ©Dudley Stamp's Land Utilisation Survey of Britain which had collated land use information in the 1930s. This was the first attempt of its kind in Britain (Stamp, 1931), aiming to document detailed changes in British land use for future generations. Remarkably, the survey was carried out by school children instructed by their teachers (Stamp, 1934). Again drawing on the EDINA Digimap service,

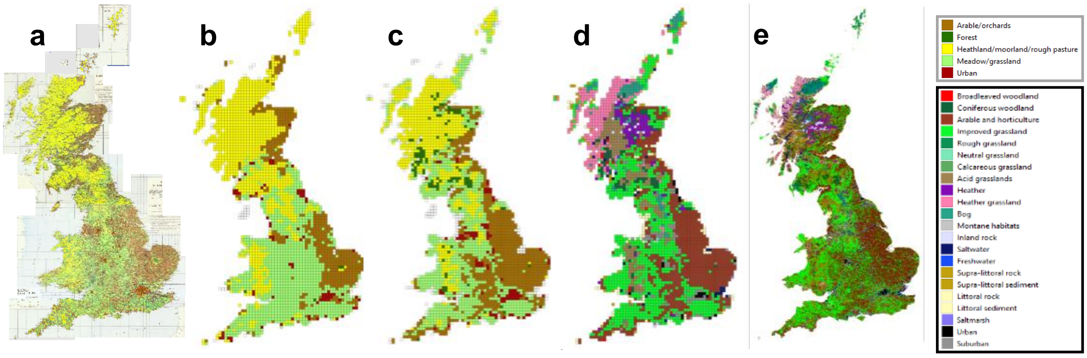


Fig. 4.1 Steps in the preparation of the hectad scale Dudley Stamp 1930s land utilisation map and modern land cover maps. a is the composite of original Dudley Stamp map material. **b** shows the digitised and hectad aggregated rendition of the Dudley Stamp map used below. **c** shows the UKCEH 2020 LCM re-classified to be comparable with Dudley Stamp's map. **d** is the hectad scale majority aggregation of the 2020 LCM and **e** shows the original 2020 LCM map. The legend explains the colour codes for each land cover type with the first set of categories (within the grey box) relating to the Dudley stamp classification and the second set (within the black box) relating to the UKCEH 2020 LCM categories.

who hold digitised copies of the original survey sheets, I loaded the material into ArcGIS 10.8 and georeferenced it (Fig. 4.1a). I used supervised classification in the 'Spatial Analyst' extension to extract information on land cover, making sure to build the training set with samples from different areas across the map to account for slight colouring differences from the scans of the original sheets. Finally, 'Zonal Statistics' in QGIS were used to find majority coverage of each hectad (Fig. 4.1b). This last step also alleviates to some extent the digitisation pitfalls highlighted by Zatelli *et al*. (2019), namely the misidentification of text on the map as a minor land cover type. The different steps of the process outlined above are visualised in Fig. 4.1. The hectad scale Dudley Stamp map is available as a shapefile (Method S4.1). Finally, since Dudley Stamp's classification and that of the later land cover maps are not identical, I made the decision to summarise categories to make the data more comparable. The reclassification is illustrated in Table 4.1.

Table 4.1 Reassignments of categories for comparisons between Dudley Stamp's Land Utilisation Survey data and the UKCEH's land cover maps (LCMs).

Comparison category	Dudley Stamp category	UKCEH LCM category
Arable/orchards	Arable land Orchards and nursery gardens	Arable and horticulture
Forest and woodland	Forest and woodland	Broadleaved woodland Coniferous woodland
Heathland/moorland/rough pasture	Heathland, moorland and rough pasture	Acid grassland Rough grassland Bog Heather Heather grassland Inland rock Saltmarsh
Meadow/grassland	Meadowland and permanent grassland	Calcareous grassland Improved grassland Neutral grassland
Urban	Chief urban areas	Urban Suburban
Not applicable	Not applicable	Freshwater Saltwater Littoral rock Littoral sediment Supra-littoral rock Supra-littoral sediment

Data management, visualisations and statistics

Microsoft Excel was used for data management and data manipulation. Analyses relied on packages of the 'tidyverse' (Wickham *et al.,* 2019) in R (R version 4.1.3). All maps, coefficient tables and plots were created in QGIS Desktop 3.24.2 'Tisler' and R (packages 'sf', 'spdep', 'tmap', 'maptools', 'kableExtra', 'ggplot2' and 'ggalluvial' (Pebesma, 2018; Bivand *et al.,* 2015; Tennekes *et al.,* 2022; Bivand *et al.,* 2022; Zhu, 2019; Wickham, Chang & Wickham, 2016; Brunson, 2018)), with post-processing in Microsoft PowerPoint. The full data frame used for modelling (including environmental data and hectad weighted genome size and ploidy) is available in Table S4.3.

The change of genome sizes over time was assessed using a Wilcoxon Signed Rank Test to account for the non-independence of repeated data for the same set of hectads. Variable selection for spatial models of hectad weighted mean genome size and ploidy per hectad in the final date class and of change in genome size per hectad was based on Pearson correlation assessed in the 'corrplot' package (Wei *et al.,* 2017) and iterative dropping of each model term to minimise AIC. Predictors for hectad weighted mean genome size in the final date class were environmental variables and species richness (the estimated number of species present in any hectad, following Frescalo correction, as per Chapter 3). In modelling change in weighted mean genome size per hectad over the course of the three date classes, I used changes within the predictors over the same time span.

The relative importance of variables in non-spatial linear models was assessed using the 'relaimpo' package (Grömping & Matthias, 2021). I chose the 'lmg' metric (proposed by Lindemann, Merenda & Gold, 1980), which decomposes R^2 into a set of non-negative contributions, summing automatically to the total R^2. This approach has been shown to be robust to the pitfalls of collinearity since the metric averages across different orderings for the predictors (Grömping, 2007).

Spatial signal (spatial autocorrelation) was assessed by calculating Moran's I of outcome and predictors. Residuals of the non-spatial model were also plotted and inspected for spatial signal. Lagrange multiplier diagnostics (Anselin *et al.,* 1996) for spatial dependence were used to identify the nature of the spatial dependence present within the data (spatial lag and spatial error dependence). Due to strong evidence of both spatial lag and spatial error dependence, I followed guidance by Anselin, Le Gallo & Jayet (2008) and corrected for the dependence with the largest test statistic, in this instance the spatial error dependence. The 'spatialreg' package (Bivand *et al.,* 2019) was used to run the final spatial model.

Differences in hectad weighted mean genome size profiles of different land use categories were assessed using ANOVA with Tukey post-hoc tests for multiple comparisons on those land use classes that are the majority cover in more than 15 hectads.

I tested for the presence of phylogenetic signal (Blomberg's K and Pagel's λ) in the genome size and magnitude of change along the North-South axis data using the 'phytools' package (Revell, 2012) with 10,000 randomisations. The association of genome size and northward movement was then tested using Phylogenetic Generalised Least Squares (PGLS, Symonds & Blomberg, 2014) regression as implemented in the packages 'ape' and 'nlme' (Paradis *et al.,* 2019; Pinheiro *et al.,* 2017), based on the flora-wide phylogeny described in Chapter 2. To account for cytotype variation, I attached each cytotype to the base species within the phylogeny, resulting in an expansion of the phylogeny from 2,501 leaves to 2,742, which was ultimately used to account for phylogenetic signal here (the resulting phylogeny is available in Method S4.2). I tested model fit based on Brownian, Blomberg and Pagel correlation structures and chose Pagel's due to it yielding the lowest AIC value. Genome size data was log transformed and the magnitude of northward movement was sqrt-transformed. Further, I performed quantile regression (Koenker & Bassett, 1978), as implemented in the 'quantreg' package (Koenker *et al.,* 2018), on the same data to find if the association between genome size and northward movement differed at different quantiles of northward movement. A Bonferroni correction was applied to the quantile regression results to account for multiple comparisons. A phylogenetically corrected ANOVA ('phytools' package) was chosen to test for differences in the northwards movement of plants of different status (i.e. native, archaeophyte, neophyte; see glossary; Table 2.1).

Results

Spatio-temporal patterns of genome size and ploidy

The patterns of genome size and ploidy show two very different trends.

Hectad weighted mean ploidy level across the UK follows a clear latitudinal and altitudinal gradient (Fig. 4.2a) that stays consistent across the three date classes; the South is characterised by lower hectad weighted mean ploidy levels while the North and especially the Scottish Highlands exhibit higher hectad weighted mean ploidy levels on average. The changes in hectad weighted mean ploidy levels are negligible across the three date classes, with changes never exceeding +/-0.09.

The pattern of hectad weighted mean genome size is strikingly different, with distinct areas characterised by smaller and others by larger hectad weighted mean genome sizes (Fig. 4.2b). An overall trend of smaller hectad weighted mean genome sizes in the North (especially the North West) and larger hectad weighted mean genome sizes in the South (especially the South East) is visible, but in addition to this trend, there are clear hot spots (large genome sizes), such as in urban areas (Greater London in particular), and cold spots (small genome sizes) e.g. in western Scotland and northern Wales. There are also clear trends in hectad weighted mean genome size profiles between the three date classes (Fig.

4.3a). Most (2,842) hectads have experienced weighted mean genome size increases from the first to the most recent date class, with only 159 showing a decreasing weighted mean.

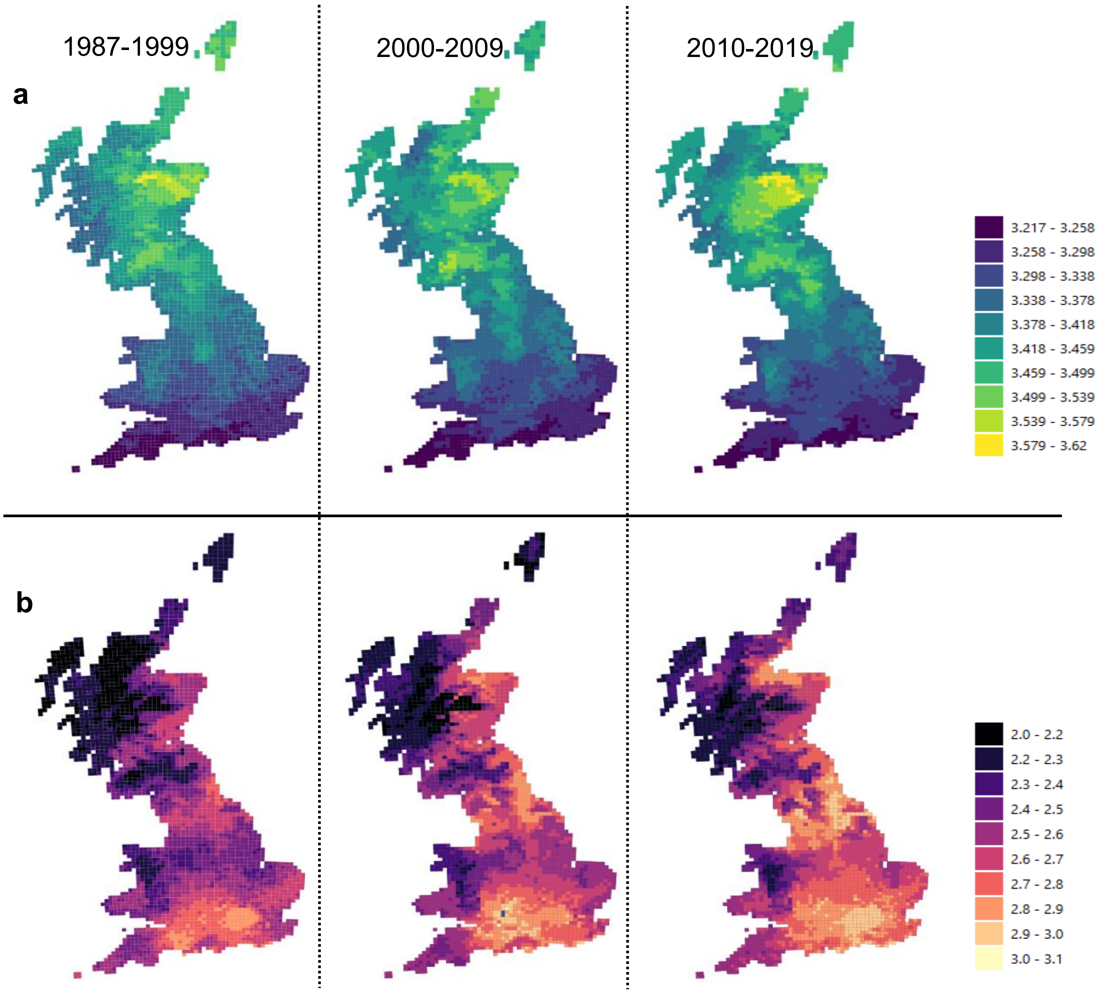


Fig. 4.2 Patterns of ploidy and genome size in space and time. Series **a** and **b** show the patterns of weighted mean ploidy and genome size respectively throughout the three date classes. The legend for ploidy level is given in the number of chromosome sets in the nucleus (x), while the legend for genome size is given in pg/1C. A clear gradient from North to South and very few changes through time in the ploidy graphs are juxtaposed with locally distinctive patterns and a gradual change towards larger genomes in more recent years.

The greatest increases in hectad weighted mean genome sizes are localised in England and northern Scotland. The Wilcoxon Signed Rank Test with continuity correction revealed a significant increase in hectad weighted mean genome sizes across the whole study area, both overall and from one date class to the next (all $p < 0.0001$, Fig. 4.3b). The total mean increase in hectad weighted mean genome sizes across the study area between the first and last date class amounts to 5.5% (from a mean of 2.4 pg/1C in the first date class to a mean of 2.6 pg/1C in the last date class).

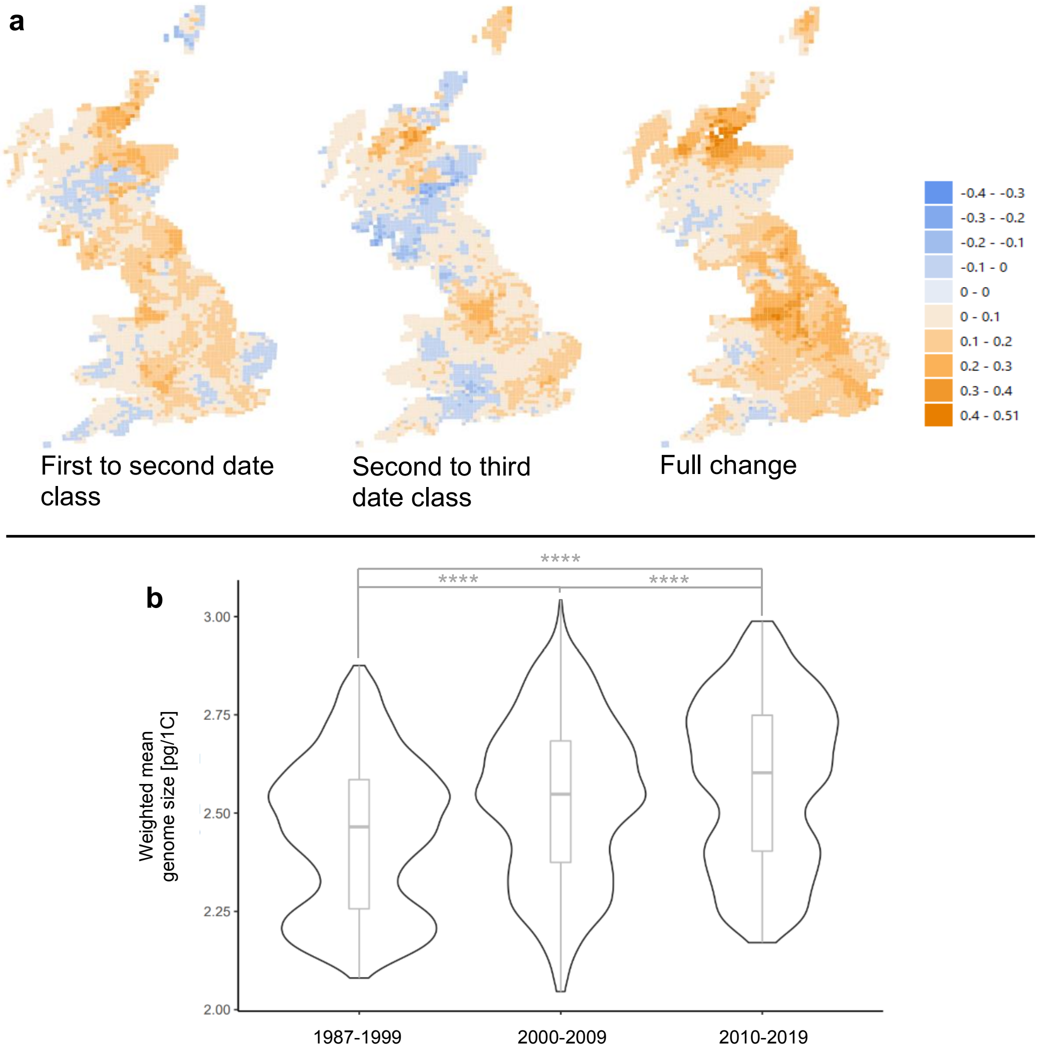


Fig. 4.3 Changes in genome size. a shows the change in hectad weighted mean genome size of each hectad, between the first and second, second and third and first and third date class (full change). Orange hues indicate increases while areas with decreasing genome sizes are coded in blue. The violin plots and integrated boxplots in **b** illustrate the gradual increase in genome size across the study area. Significant differences were found between all groups ($p < 0.0001$), as indicated by asterisks. All hectad weighted mean genome sizes are in pg/1C for 1,698 species and cytotypes with available data.

Land use changes in the long- and short-term

Fig. 4.4 shows the change in land cover present in the study area from the 1930s to 2020. The expansion of agricultural land, particularly in the West of England, and the slightly increasing space occupied by urban areas are visible, but also the increasing reforestation in Scotland can be made out with more hectads exhibiting mostly forest cover. The alluvial plot (Fig. 4.4a) demonstrates the fate of hectad majority cover; each hectad is represented here as a line with colours indicative of the majority cover in 2020. The strata at each time point represent the proportions of the different land cover types within them. A more detailed plot of land cover changes can be seen in S4.1.

Roughly half (362 hectads, 46.8%) of current agricultural land (773 hectads) was converted from areas that were previously meadow or grassland areas between the 1930s and 1990. While only three hectads were mostly covered in forest in the 1930s, there now is a substantial group of such hectads (114 hectads). The overwhelming majority of them (104 hectads) were previously classed as heathland, moorland or rough pasture.

Predictors of hectad weighted mean genome size and ploidy

Numeric predictors

All numeric predictors (species richness, rainfall, temperature, wet and dry nitrogen deposition) as well as hectad weighted mean genome size and ploidy were found to be spatially autocorrelated (Moran's I 0.91, 0.91, 0.94, 0.87, 0.93, 0.91 and 0.96, all $p < 0.0001$). The same is true for the changes in predictors, hectad weighted mean genome size and ploidy across the three date classes (Moran's I 0.82, 0.83, 0.95, 0.83, 0.89, 0.83 and 0.78, all $p < 0.0001$). Maps of each predictor are available in Fig. S4.2.

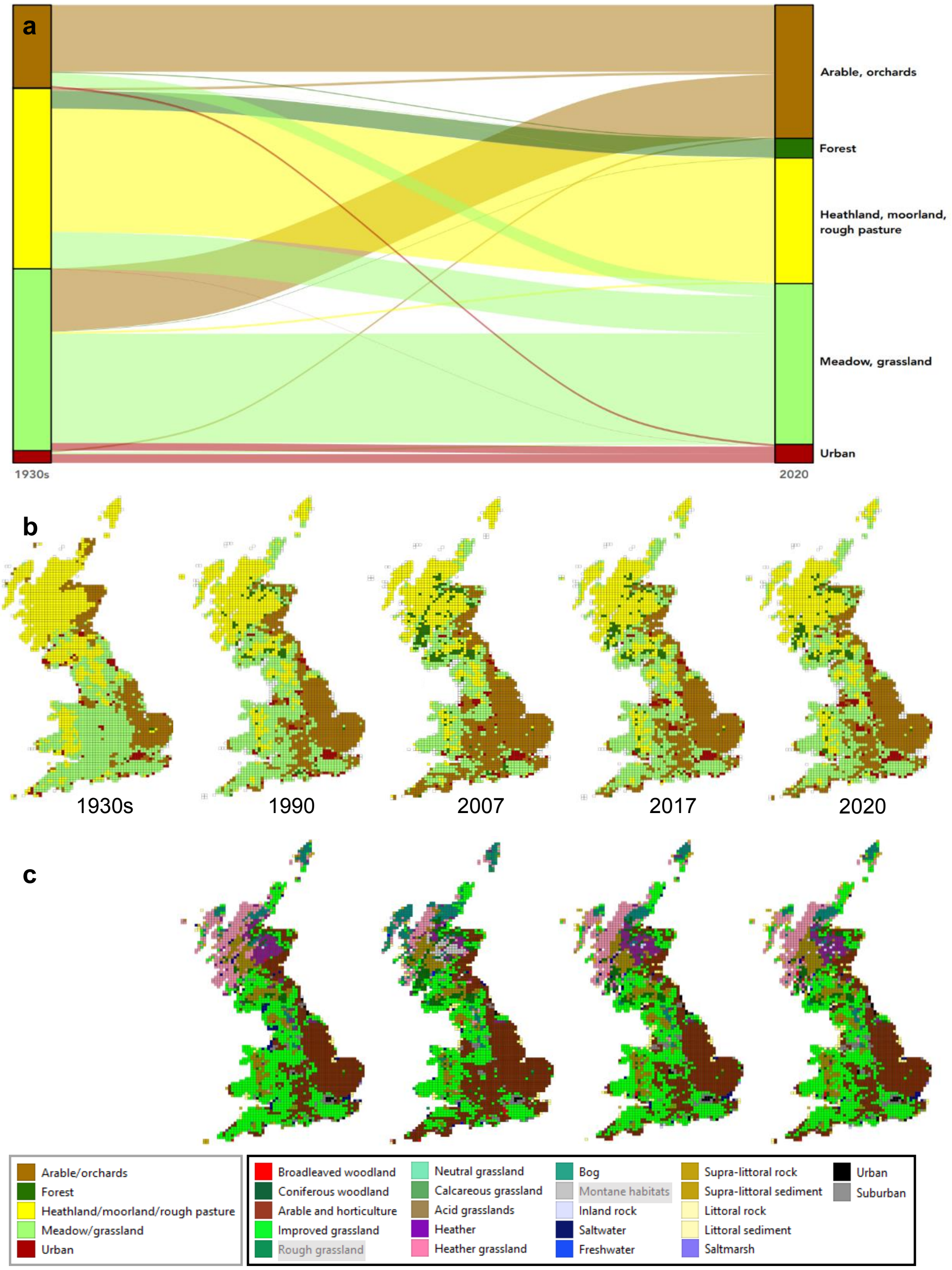


Fig. 4.4 Land cover changes. a is an alluvial plot that visualises the change in majority land cover for the 2,655 hectads (represented as individual lines) for which this information is available at each of the time points considered here (1930s, 1990, 2007, 2017 and 2020). **b** and **c** map the land cover by hectad categorised according to the Dudley Stamp 1930s map (**b**) and UKCEH land cover maps (**c**). 2007 represents a special case, since the UKCEH LCM's categories for this period are not perfectly aligned with those used in the preceding and following years, making direct comparisons more challenging. Hectads with majority cover for one of those land cover types that were not assigned in all time periods. The land cover types only present in the 2007 LCM are highlighted in grey in the legend. Legends for maps are given inside the grey box for Dudley Stamp categories and inside the black box for UKCEH categories). Colours in **a** indicate the majority cover the hectad falls into in the final date class and correspond to the legend and map in **b**.

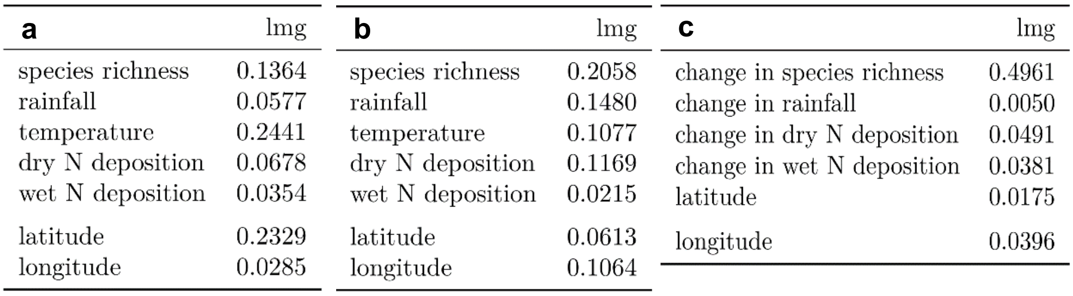
Linear models for individual predictors of hectad weighted mean genome size and ploidy are presented in Fig. S4.3 and S4.4 and show opposing responses for hectad weighted mean genome size and ploidy to all tested predictors. While hectad weighted mean genome sizes decrease with increasing latitudes, hectad weighted mean ploidy level increases. Hectad weighted mean ploidy level also increases with rainfall per growing season. Conversely, hectad weighted mean genome size increases with increasing species numbers, temperature per growing season and both wet and dry nitrogen deposition, while hectad weighted mean ploidy level decreases in response to these predictors.

Changes in hectad weighted mean genome size over time showed less clear relationships with changes occurring in the different predictors, although hectad weighted mean genome size increases over time were correlated with rising species numbers, increasing temperatures and wet nitrogen deposition (Fig. S4.5).

In preparation for multivariate modelling I inspected correlations between predictor variables to diagnose collinearity that would necessitate exclusions of variable combinations. I found that Pearson correlations (Fig. S4.6) among potential predictors of changes in hectad weighted mean genome size over time were not high enough to preclude any combinations of variables from multivariate analyses, with all correlations well below +/-0.4, with the exception of the correlation between the change in temperature and the change in dry nitrogen deposition over time, which was -0.56. Conversely, most of the predictor variables for prediction of mean genome size within hectads of the last date class showed stronger correlations, once again indicating the need for a spatial modelling approach. Single term deletions on the multivariate linear models revealed a benefit in dropping the temperature component from the change models for both hectad weighted mean genome size and ploidy.

The relative importance of variables was derived from linear models that did not account for spatial correlations, but which did include latitude and longitude as predictors. The resulting lmg metrics (Lindemann, Merenda & Gold, 1980) of variable importance for all predictors of hectad weighted mean genome size and ploidy in the last date class as well as the change in hectad weighted mean genome size and ploidy over time are summarised in Table 4.2. Unsurprisingly given the clear gradient observed when mapping hectad weighted mean ploidy (Fig. 4.2a), the overwhelmingly most effective predictors for it are temperature and latitude (Table 4.2a), with increasing latitude and decreasing temperature associated with larger ploidy levels (Fig. S4.4). Changes in hectad weighted mean ploidy across the three date classes were negligible and were therefore not considered in the following analyses (data not shown). The best predictor of hectad weighted mean genome size is species richness, with rainfall a distant second (Table 4.2b). The importance of species richness becomes even more apparent in the variable importance for the model of change in hectad weighted mean genome size. Here, the change in species richness outcompetes the other predictors by an order of magnitude (Table 4.2c).

Table 4.2 Relative importance of predictors in linear models. **a** hectad weighted mean ploidy level in the last date class, **b** hectad weighted mean genome size in the last date class and **c** the change in hectad weighted mean genome size from the first to the last date class. Lmg is the metric of variable importance used (Lindemann, Merenda & Gold, 1980) and describes the variance explained by each predictor, summing to the total R^2 of each model (R^2 = 0.77, 0.80 and 0.64, respectively).



a	lmg
species richness	0.1364
rainfall	0.0577
temperature	0.2441
dry N deposition	0.0678
wet N deposition	0.0354
latitude	0.2329
longitude	0.0285

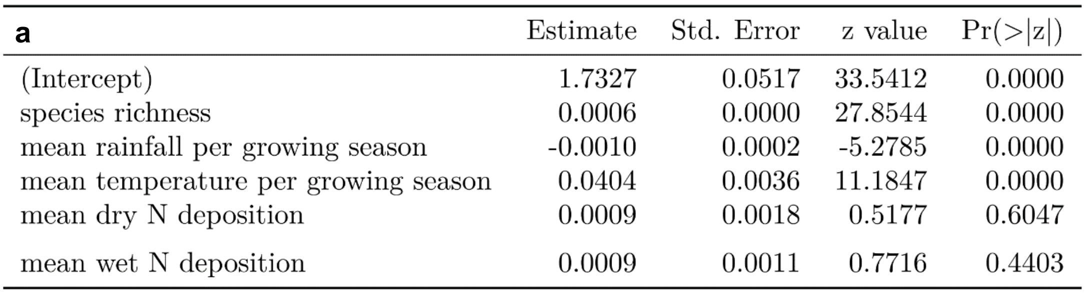
b	lmg
species richness	0.2058
rainfall	0.1480
temperature	0.1077
dry N deposition	0.1169
wet N deposition	0.0215
latitude	0.0613
longitude	0.1064

c	lmg
change in species richness	0.4961
change in rainfall	0.0050
change in dry N deposition	0.0491
change in wet N deposition	0.0381
latitude	0.0175
longitude	0.0396

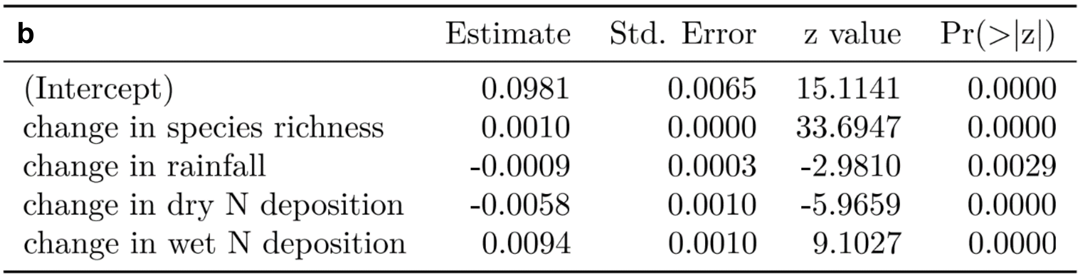
When grouping the hectads by the land use (as categorised by UKCEH LCMs; Fig. 4.4c) they fall into in the last date class, change in species richness remains the main factor influencing changes in hectad weighted mean genome size across the majority of land cover types. However, there are some interesting emergent predictors when interrogating certain land cover categories in this way. In 'suburban' and 'urban' hectads, change in dry nitrogen deposition emerges as a substantial secondary predictor of change in hectad weighted mean genome size. In areas where the majority cover is 'heather', changes in wet nitrogen deposition and rainfall also add considerable explanatory power to the models. Finally, the change in wet nitrogen deposition becomes an important predictor in addition to changes in species richness in 'acid grassland', 'bog' and 'coniferous woodland' land cover types. Indeed, in 'acid grassland' hectads, change in wet nitrogen deposition is the most helpful predictor, ahead of species richness, with a considerable effect of change in dry nitrogen deposition as well (Table S4.4).

The residuals within the models for hectad weighted mean genome size in the final date class and for the change in hectad weighted mean genome size also showed spatial patterning (Moran's I 0.65, $p < 0.0001$). This, on top of the spatial non-independence within predictors and outcome does suggest the importance of accounting for spatial dependence structures within the models themselves to avoid chronic under- or overestimation of the regression in proximate areas. Lagrange multiplier diagnostics for spatial dependence showed that both spatial error and spatial lag dependence were present and significant within the models ($p < 0.0001$ in all cases). The test statistics for the spatial error dependence were higher in both models (RLMerr = 3,976.0 and RLMerr = 1,296.2, RLMlag = 21.7 and RLMlag = 96.6), suggesting the greater importance of correcting for the non-independence in the error structure (Aneselin, Gallo & Jayet, 2008). The final, spatial linear regression model showed only slight, but significant effects of species richness, mean rainfall and mean temperature per growing season on hectad weighted mean genome size in the last date class (Table 4.3a). Species richness and temperature were positively associated with hectad weighted mean genome size, while an increase in rainfall had a negative effect on hectad weighted mean genome size. Neither wet nor dry nitrogen deposition showed significant effects.

Table 4.3 Summary of spatial models. a hectad weighted mean genome size in the final date class and **b** of change in hectad weighted mean genome size between the first and last date class.



a	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.7327	0.0517	33.5412	0.0000
species richness	0.0006	0.0000	27.8544	0.0000
mean rainfall per growing season	-0.0010	0.0002	-5.2785	0.0000
mean temperature per growing season	0.0404	0.0036	11.1847	0.0000
mean dry N deposition	0.0009	0.0018	0.5177	0.6047
mean wet N deposition	0.0009	0.0011	0.7716	0.4403



b	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0981	0.0065	15.1141	0.0000
change in species richness	0.0010	0.0000	33.6947	0.0000
change in rainfall	-0.0009	0.0003	-2.9810	0.0029
change in dry N deposition	-0.0058	0.0010	-5.9659	0.0000
change in wet N deposition	0.0094	0.0010	9.1027	0.0000

The model for change in hectad weighted genome size over time based on changes in the predictors across the three date classes revealed significant associations for all predictors retained in the model (Table 4.3b). A positive change in species richness was once again associated with an increase in hectad weighted mean genome size, while an increase in rainfall concurred with a decrease in hectad weighted mean genome size. Changes in wet and dry nitrogen deposition have relatively strong and opposing effects in this model, with increases in wet nitrogen deposition associated with an increase in genome size while increases in dry nitrogen deposition correlate with genome size decreases.

Land cover

Having already observed the differences in variable importance associated with different land cover types, I wanted to find out how hectad weighted mean genome size and its change over time differs by land use.

Weighted mean genome sizes per hectad showed clear differences across the different land cover types (Fig. 4.5). Notably, hectad weighted mean genome sizes in 'urban' and 'suburban', 'arable and horticulture' as well as 'improved grassland' and 'littoral sediment' hectads were all significantly larger than in any of the other land cover types tested (Tukey HSD p < 0.0001, except for the comparison between 'improved grassland' and 'littoral sediment' with 'saltwater' (p = 0.0241 and p = 0.0041), for a full list of comparisons see Table S4.5.

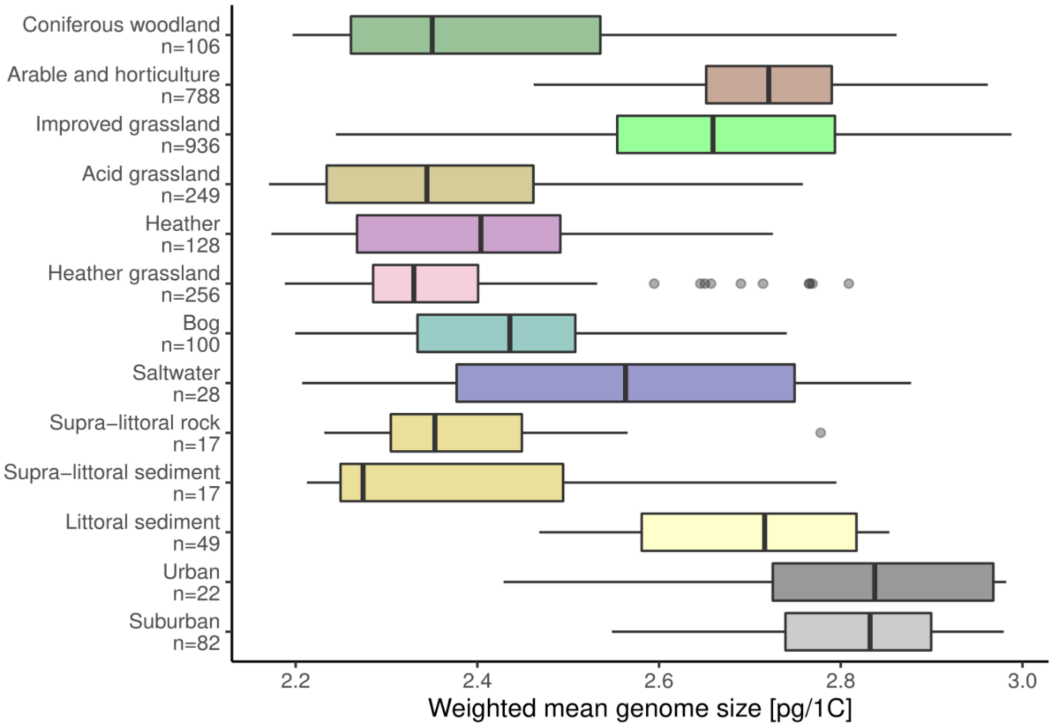


Fig. 4.5 Hectad weighted mean genome sizes in different land cover categories for 2,778 hectads. Boxplot representation of weighted mean genome size profiles in hectads associated with different land cover types in the final date class. Only categories represented by more than 15 hectads are shown. Colours correspond to default UKCEH land cover colour code. Land cover types 'urban and 'suburban', but also 'arable and horticulture', 'improved grassland' and 'littoral sediment' stand out from all others as harbouring plants with significantly larger mean genome sizes. The number of hectads falling into each group is given along the y-axis.

While across all land cover types, the hectad weighted mean genome size has been increasing steadily across the three date classes, some land cover types stand out. Compared to the 5.5% increase in mean genome size between the first and last date class when analysing data from all land cover types together (Fig. 4.3), weighted mean genome sizes in 'acid grassland' hectads only increase by 2.5% with a stagnation of the increase between the second and third date class. In contrast, particularly large increases in weighted mean genome size can be observed in 'bog' hectads (8.1% increase) and in 'suburban' hectads (7.3% increase), where the rise was steady across date classes.

Centre of mass

The vast majority (79.4%) of species exhibited a northward shift from the 1987-1999 to the 2010-2019 date class, with 933 out of the total 1,175 species with sufficient information moving North. Fig. 4.6 shows the distances and direction of movement for those species with available data for status. Plants with different status in the study area also showed

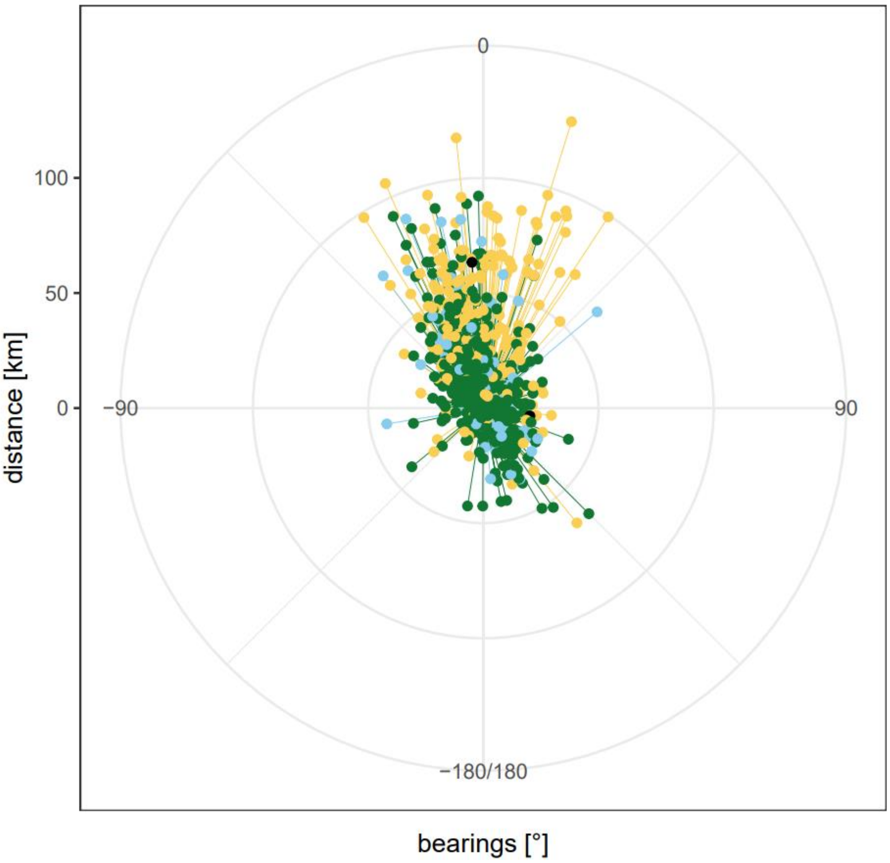


Fig. 4.6 Shifts in centre of mass between the first and last date class. The shifts for the centre of mass of 1,163 species. 825 native (green), 111 archaeophyte (blue), 224 neophyte (yellow) and 3 neonative (black) species are represented with respect to the distance as well as the direction of the movement. The vast majority (i.e. 79.4%) and especially neophyte are moving towards the North.

different potential for movements towards the North, with archaeophytes and especially neophytes moving significantly larger distances compared to natives, who only performed marginal shifts of usually less than 25 km towards the North (Fig. 4.7). The differences were significant (phylogenetic ANOVA, all $p < 0.01$), even when accounting for the pronounced phylogenetic signal within the northward movement data ($K = 0.0338039$, $\lambda = 0.77175$, $p < 0.0001$).

Evocative of this trend, the strongest shifts northward were shown by *Cupressus macrocarpa* Hartw. ex Gordon (Cupressaceae, 130 km), *Jacobaea maritima* (L.) Pelser & Meijden (Asteraceae, 118 km) and *Lemna minuta* Kunth (Araceae, 106 km), representing a neophyte survivor and two naturalised neophytes respectively. The species moving furthest South are two natives, *Callitriche platycarpa* Kütz. (Plantaginaceae, 65 km) and *Catapodium marinum* (L.) C.E.Hubb. (Poaceae, 53 km), and *Cedrus libani* A.Rich. (Pinaceae, 64 km), another neophyte survivor. Meanwhile among some of the species with an almost entirely static centre of mass are the native plants *Trifolium repens* L. (Fabaceae), *Plantago lanceolata* L. (Plantaginaceae) and *Juncus effusus* L. (Juncaceae).

Beyond the signal already found within northward movement data, phylogenetic signal was also significant and substantial in the genome size data ($K = 0.259017$, $\lambda = 0.999934$, $p < 0.0001$), suggesting the importance of a phylogenetic correction. The magnitude of northward shift in the flora was found to be significantly and positively associated with genome size (Fig. 4.8a) when tested using PGLS regression ($p < 0.0001$). Quantile regression revealed that this positive association is especially driven by those species that move the furthest distances (0.9 and 0.75 quantiles), where the positive association is the steepest (Fig. 4.8b).

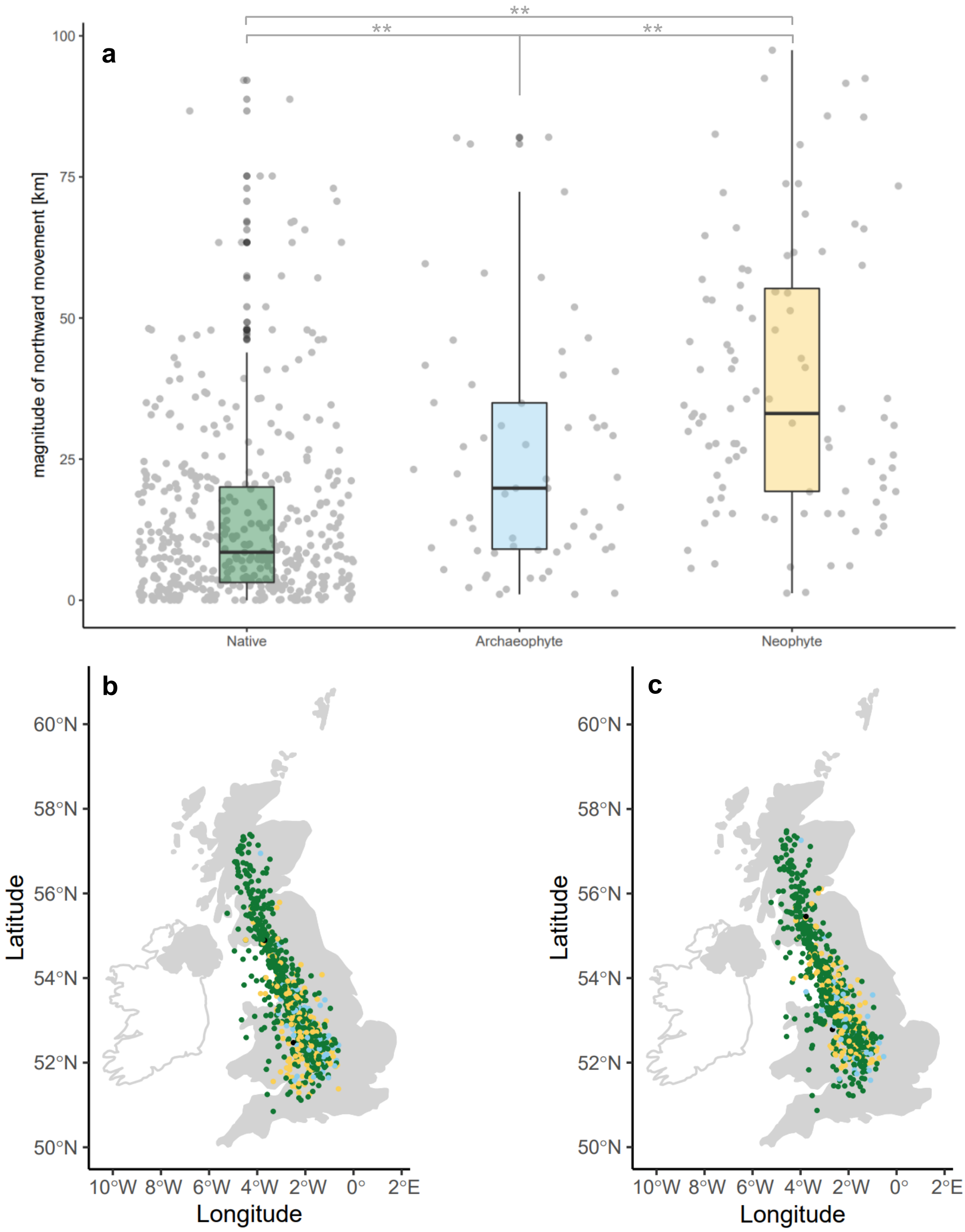


Fig. 4.7 The northward movements of natives and non-natives. The different movement profiles of plants of different status towards the North are represented as a box- and scatterplot (**a**). All comparisons were significant after accounting for phylogenetic signal, as indicated by asterisks. The data represents 443 natives, 62 archaeophytes and 99 neophytes with available phylogenetic information. Neonatives were not tested due to a scarcity of records (n = 3). **b** & **c** show the location of centre of masses in the first date class (1987-1999, **b**) and the last date class (2010-2019, **c**). Neophyte plants that showed the strongest northward movement typically have centres of mass in the far South in the first date class, while natives are spread out across the whole length of the UK. In the last date class neophyte centres of mass had shifted further north leaving fewer centres of mass in the far South. The location plots encompass information on all 825 native (green), 111 archaeophyte (blue), 224 neophyte (yellow) and 3 neonative (black) species for which centre of mass could be calculated.

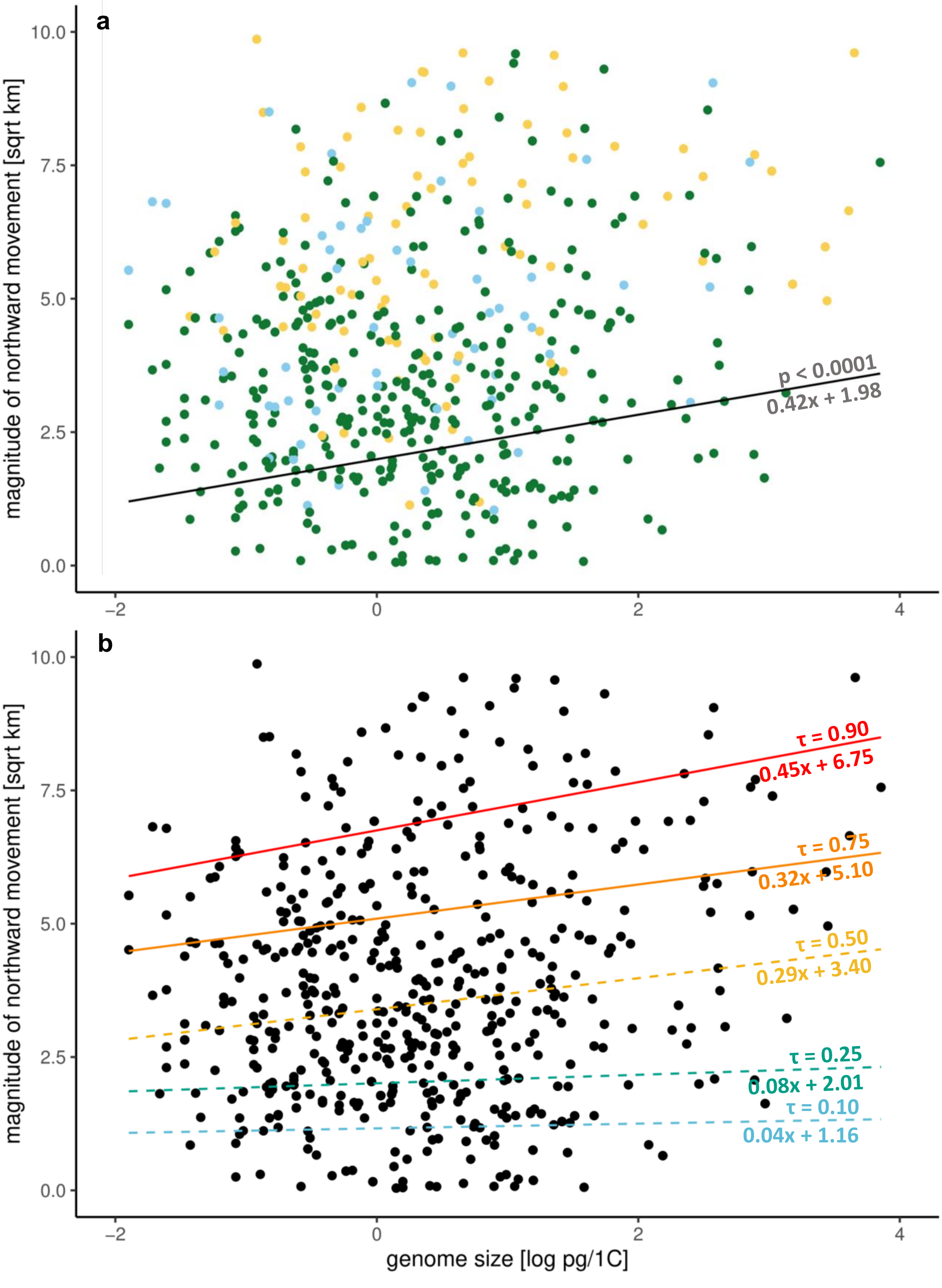


Fig. 4.8 The association of northward shifts and genome size for 604 species. The square root of the magnitude of northward shifts of each species is plotted against its log transformed genome size along with PGLS regression (**a**) and quantile regression lines (**b**). Colours in **a** correspond with status of species (green stands for natives, blue for archaeophytes, and yellow for neophytes). Larger northward shifts are associated with larger genome sizes. Notably, it is especially neophytes who are performing large movements northwards The PGLS fit is highly significant (p < 0.0001). The quantile regression was performed on the conditional quantiles τ = 0.9, 0.75, 0.5, 0.25 and 0.1. Only regressions for τ = 0.9 and 0.75 are significant (p < 0.05). Dashed lines indicate non-significance. Lines are labelled with the corresponding equation (format mx + c, where m = slope and c = intercept).

Discussion

Genome size patterns and their predictors

Weighted mean genome size and ploidy levels per hectad show inverse patterns across the UK. The clear increase in hectad weighted mean ploidy levels towards the North concurs with findings of Rice *et al.* (2019) who found that polyploid frequency increases towards the poles. Recent findings from an analysis of the global distribution of genome sizes showed a similar pattern of increasing genome sizes towards both poles, although in the far North (above latitudes of c. 50-60N), particularly in regions with recent glaciation histories, the relationship was reversed, with further increases towards the North characterised by increasingly smaller genomes (Bureš *et al.,* 2022 (in press)). This latter finding is corroborated by the patterns of weighted mean genome size in the very recently glaciated UK (Clark *et al.*, 2012), presented here.

Meanwhile, the comparatively small size of the study area means that the factors at play on a global level may not be apparent in this study. While both global genome size and ploidy distributions are likely to be linked to climate and soil properties (e.g. nutrient poorer soils in the tropics), it appears that the smaller geographic scales and extremely high levels of human disturbance characterising the UK might lead to different dynamics. In the analysis here, hectad weighted mean ploidy level on the one hand is predicted mostly by temperature and latitude in simple linear models without spatial considerations beyond the inclusion of coordinate data. This is in agreement with the findings of Rice *et al.* (2019), who found temperature to be the most relevant factor in predicting polyploid frequencies. Such a clear picture does not emerge in the more spatially distinctive patterns of hectad weighted mean genome size, where non-spatial models identify species richness as the major predictor of hectad weighted mean genome size and especially of changes in hectad weighted mean genome size over time. In this study, weighted mean genome size per hectad increases with temperature but decreases with rainfall, both in assessments of the effect of individual factors and in the final spatial models. In contrast, the model exploring changes in hectad weighted mean genome size over time performs better with the exclusion of temperature altogether, while local increases in rainfall since the 1980s and 1990s are associated with decreases in genome size.

The correlation between genome size and water use efficiency is complex. Species with larger genome sizes and hence stomatal guard cell sizes are thought to lose less water than species with smaller guard cells for the same total stomatal pore area, which might suggest increased water use efficiency in species with larger genomes. However, those species may also open and close their stomata more slowly in response to changing weather, which may have the opposite effect on water use efficiency (Faizullah *et al.,* 2021). While an increase in mean air temperature for Central England has been reported (Watts *et al.,* 2015), with summers now between 1 and 6°C warmer and in some regions up to 60% drier than in 1990 (Met Office, 2022), the UK is still a comparatively wet and cool area with relatively few areas affected by droughts on a regular basis, although the effects of climate change are already felt in increased frequencies of extreme weather events from droughts to storms (Kendon *et al.*, 2022). The relatively limited range in temperatures and aridity across the UK may not be sufficient for strong trends to emerge. However, this situation may well change in the future, since unmitigated climate change is expected to cause unprecedented increases in temperature and decreases in rainfall with the potential to overturn landscape-level community assemblages (Ritchie *et al.,* 2019).

Wet and dry nitrogen deposition is used in the models above as a proxy for overall eutrophication, a known driver of declines in species richness (Payne *et al.*, 2017; Stevens *et al.,* 2004) and a hypothesised enabling factor in the increased dominance of plants with larger genomes (Guignard *et al.*, 2016; Šmarda *et al.,* 2013; Peng *et al.,* 2022). In the current study, there is some uncertainty regarding the role of nitrogen deposition on hectad weighted mean genome size. Neither wet nor dry nitrogen deposition are significantly associated with hectad weighted mean genome sizes in the spatial multiple regression analysis of the last date class. Nevertheless, changes in each deposition type have opposing and significant correlation with changes in hectad weighted mean genome size over time, with increases in dry nitrogen deposition associated with decreases in the temporal change in hectad weighted mean genome size but increases in wet nitrogen deposition leading to increases in hectad weighted mean genome size. In individual regressions of these predictors, this opposing effect on change in hectad weighted mean genome size is also seen, while individual regressions of the association of hectad weighted mean genome size in the last date class and nitrogen deposition show positive correlations with both deposition types. One of the most unexpected results, that areas with increasing dry nitrogen deposition show decreasing hectad weighted mean genome sizes, might be explained by the nature of dry deposition which is expected to be highest near urban centres, along motorways and near agricultural sites, where the effects of other human impacts may be more prominent and perhaps obscure the expected effects of additional nitrogen. Wet nitrogen deposition on the other hand travels further away from emission sources and is deposited more evenly (Tomlinson *et al.*, 2021). The fact that this deposition type does seem to correlate with increases in hectad weighted mean genome size offers support for the hypothesis that high levels of nutrient availability in soils are expected to lift the constraint of highly nutrient-demanding species with large genomes, allowing them to colonise more widely. Meanwhile, the contradicting nature of the findings regarding the role of nitrogen deposition on genome size patterns may also mean that the hypothesised role of nutrients on shaping plant communities via genome size simply does not emerge at the scales tested. It must be noted that the data on atmospheric nitrogen deposition used here is likely not the ideal measure to test for the effect of increased nutrient availability, especially when considering that the full effects of nutrient limitation on genome size are likely to arise from the combined effect of nitrogen and phosphorus (Guignard *et al.,* 2016).

In the setting of steady increases in the intensity of agriculture and consequently eutrophication from agricultural run-off across the UK in the last century (Smart *et al.,* 2003; Firbank *et al.,* 2000), nutrient pollution from agricultural lands continues to be a major and poorly controlled (Sharma, 2020) threat to soil health, even as more effective policies have caused steady declines in atmospheric nitrogen deposition (Tomlinson *et al.,* 2021). While information on fertiliser application for England in 2015 is available (UKCEH Land Cover® plus) and was used in preliminary models, the restricted extent of this dataset excluded many interesting geographical areas of hectad weighted mean genome size distributions and led to inconclusive findings. What is clear, however, is that areas characterised by human activity can be expected to be more drastically affected by nutrient pollution, either in the form of atmospheric nutrient deposition (especially 'urban' and 'suburban' land cover categories) or nutrient pollution from fertiliser application and livestock manure ('arable and horticulture' as well as 'improved grassland' land cover types, the latter of which is most commonly used as productive grazing land (NatureScot, 2018)). This is confirmed by higher levels of nitrogen, phosphorus and potassium application in hectads with mostly agricultural use ('arable and horticulture', to a lesser extent 'improved grasslands', see Fig. S4.7), suggesting that land use might offer further insights into the shaping effects behind the genome size distribution patterns observed. In particular, while mere nitrogen deposition cannot account for the combined effects of nitrogen and phosphorus pollution, the coupled application of NPK fertilisers in agricultural environments suggests that larger mean genome sizes within agricultural land use types might reflect the synergistic effect of both nutrients.

Human impact favours larger genomes in the UK

Hectads most impacted by humans, i.e. those with a majority of 'urban', 'suburban', 'arable and horticulture' or 'improved grassland' cover clearly harboured the largest mean genomes (Fig. 4.5). These are the very hectads most likely to suffer from high levels of nutrient pollution either from atmospheric deposition or from agricultural practices, offering support for the hypothesis that abundant nutrient supply removes constraints on species with larger genome sizes, thus driving the average genome size of species occupying such hectads upwards. The specific associations between land use, nutrient pollution and other effects of human activities with genome size patterns may benefit from structural equation modelling to help untangle some of these interrelated effects.

Beyond the effects of nutrient levels alone, human disturbance in itself might be a factor in driving genome size differences and change across Britain. Lim *et al.* (2014) noted that the strongly felt presence of humans is a major driver of plant invasions, suggesting that the level of human disruption present within small, industrialised nations such as the UK might fundamentally alter the way threats to biodiversity must be contextualised and countered. The finding that higher levels of species richness are associated with larger hectad weighted mean genome size may also be related to human actions, especially due to higher levels of species richness near metropolitan areas. As discussed in detail in Chapter 3, the pattern of species richness (Fig. 3.1) considered here is most likely less reflective of thriving and diverse natural communities, but is likely instead influenced by recurrent introductions of species from agriculture and domestic gardens, recognised as a major route for plant introductions worldwide (Guo *et al.,* 2019). Notably, neophyte species, i.e. recent additions to the flora, have larger genomes than native species (Fig. 2.10), and are likely to be more frequently beneficiaries of frequent re-introductions, especially in hectads with high levels of human disturbance (e.g. as garden escapes).

Beyond the effect of garden escapes near clusters of human settlements, some species, specifically generalists with wider ranges, have previously been found to benefit from the novel niches created by human disturbance, leading to an overall loss in distinctiveness across such disturbed assemblages (Newbold *et al.,* 2018). Given the findings presented here, plants with larger genomes appear to be another group of such beneficiaries of increasing disturbance.

While land use changes in the past thirty years have been subtle, the last century has seen significant levels of agricultural intensification across the UK, especially in the wake of World War II (Robinson & Sutherland, 2002; Smart *et al.,* 2003), which is reflected in the vast expansion of 'arable and horticultural' land documented by the Dudley Stamp and subsequent land cover maps (Fig. 4.4). Given the strong association of larger hectad weighted mean genome size with land cover types characterised with high levels of human activity, it is conceivable that the drastic changes in land use have positively influenced the establishment of plants with large genome sizes across the UK. While a look back in time to the genetic composition of areas about to undergo change towards more intensive agricultural use in and before the 1930s is challenging due to increasingly severe biases within biological records (Isaac & Pocock, 2015), the association of hectad weighted mean genome size with wet nitrogen deposition and changes within it certainly suggests a role of nutrient pollution and hence agricultural practises in driving community genome sizes. Ritchie *et al.* (2019), suggest that ongoing climate change will likely bring about an overhaul of land use across Britain, with warmer temperatures and higher CO_2 levels predicted to lead to westward expansions of arable lands, but also to potential needs for extensive artificial irrigation to maintain productivity especially in the South East. The expansion of intensively managed agricultural lands and uncertainty about the effectiveness of legislation on nutrient pollution in the future (Sharma, 2020; DEFRA, 2022) will pose risks to nutrient poorer habitats in particular, whose species richness has been shown frequently to decrease in response to nitrogen deposition (Maskell *et al.,* 2010; Stevens *et al*., 2004). Plants with larger genomes may well emerge as winners of this trend and their potential to become dominant and 'crowd out' biodiversity in the presence of an abundance of nutrients as suggested in field experiments (Guignard *et al.*, 2016) and under higher CO_2 (Ritchie *et al.,* 2019) as hypothesised by Faizullah *et al.,* (2021) might then become a threat for ecosystems in Britain. Whether this advantage due to nutrient pollution will be sufficient to outweigh the increasing aridity expected to result from climate change, or whether this might become the limiting factor influencing the distribution of species in the UK, especially those with larger genome sizes, will be an important development to watch.

Northward movements

Genome size also appears to correlate with range shifts of the British flora. It is clear from the results shown above that the vast majority of species in the UK are on a northward trajectory (Fig. 4.6). Northward shifts in animal and more rarely in plant distributions have been recorded and are often interpreted in the context of climate change (e.g. see Hickling *et al.* (2006) for a variety of animal groups and Lenoir *et al.* (2008) for plants). Groom (2013b) undertook to test for such movements in the UK's vascular plants between 1978 and 2011. Here I have expanded this temporal range to the year 2019 and my approach differs from Groom's in several ways (namely his use of kriging instead of Frescalo for smoothing the effects of recording bias, different timescales and the use of native species only), but produces similar results of a tendency towards northward movements.

The differences in range shifts of native and non-native plants are notable. Particularly neophytes are showing strong northward movement, while the majority of natives are almost static. This likely indicates that many of the relatively new arrivals in the flora are still in the process of expanding their range to their full potential within the UK, with continued human-driven dispersal of useful and charismatic plants a crucial factor, while many native species have likely reached the limit of northward expansion that is feasible to them (Pearce-Higgins *et al.,* 2017). The strong northward shifts of neophytes therefore may therefore be a reflection of the joint impact of gradual warming of the study area opening new habitats for plants with higher temperature requirements, along with the movement of newly introduced species from areas of initial introduction and cultivation towards more sparsely populated areas in the North (Groom, 2013b). The comparative unreliability in non-native records which was noted to make inferences about neophyte distribution shifts challenging by Braithwaite (2010), was here addressed by only using date classes that occurred after notable changes in thinking made the recording of neophytes more mainstream and expansive, but traces of it are likely to have an impact on the findings. Meanwhile, Hill & Preston (2015) demonstrated on plants native to the UK that boreal species were disappearing from the South of Britain, and, by comparing changes in the frequencies of boreal plants with similar species with warmer preferences, found that climate change appeared to be an emerging driver of vascular plant declines in Britain, suggesting that at least part of the northward movement demonstrated here is likely due to the gradual warming of the UK's climate.

The fact that plants with larger genomes are migrating further northward (Fig. 4.8) might be considered surprising given that plants with smaller genomes are often believed to have greater trait flexibility, enabling them to inhabit a broader range of environmental niches compared to species with larger genomes that are more constrained in the ecological options available to them (i.e. 'the large genome constraint hypothesis', Knight *et al*., 2005; see also Suda *et al.,* 2015 and Faizullah *et al.,* 2021). Vinogradov (2003) in fact noted that threatened plant species tended to be characterised by possessing larger genomes than species with lower levels of vulnerability to extinction, suggesting that some of the DNA sequences, such as repetitive DNA, which dominate large genomes, may constrain the ecological and evolutionary potential of such species as they act as a burden. Such findings are supported by more recent studies showing the dynamics of repetitive DNA turnover are more constrained in species with large genomes, reducing their ability to generate genetic diversity upon which selection can act (Novák *et al.,* 2020). Field experiments (Guignard *et al.,* 2016; Šmarda *et al.,* 2013) have previously shown, however, that in locations where limiting factors (nutrient limitation in particular) are removed, plants with larger genome sizes may find themselves at a competitive advantage and become dominant. It is possible then, that the wet and nutrient-rich environments of the UK are ideal locations for plants with larger genome sizes to swiftly move into new environments, although there are likely upper limits since the very largest genome sizes are typically associated with long minimum generation times and large diaspore sizes making their expansion into new niches more challenging (Cavalier-Smith, 2005).

As indicated in Fig. 4.7 and Fig. 4.8a, it is especially neophytes that are moving far distances and that thus dominate the upper quantiles of northward movement where the positive association with genome size was steepest (Fig. 4.8). It appears therefore that the newcomers in the flora are a strong influence on the changes in genome size patterns observed here, with neophyte species characterised by larger genomes than those of native and archaeophyte species, on average. Out of the 10 species with the largest genome sizes in this analysis, seven were neophytes. Lim *et al.* (2014) also note that successful invasive species in Britain are often characterised by high moisture and nutrient requirements, traits that would be shared with plants with larger genomes in areas where plants can take advantage of higher levels of human impact (e.g. high levels of nutrient additions) as well as the wetter conditions of northern parts of Britain.

The faster movement in the ranges of neophytes and plants with larger genomes in general demonstrated here might place some of them at a competitive advantage as they may well be more able to keep up with the increasingly fast-paced changes in land use and climate across the UK (Sandel *et al.,* 2011), to the potential detriment of native species with smaller genomes. Although it remains to be explored to what extent it is the intrinsic property of a larger genome that is apparent here, as opposed to a mere correlational tendency of neophyte species to have larger genomes, the role of genome size as a potential predictor of species success in the face of ongoing anthropogenic change should be considered in more detail.

The aim of this chapter was the elucidation of how spatial changes and patterns within the flora influenced the distribution of hectad weighted mean genome size and ploidy levels. Weighted mean genome sizes of species in hectads across the UK show uneven patterns, with the largest values found in areas of high human impacts. There also is a correlation between genome size and the trajectories of plants within the UK, particularly for species which have been introduced into the flora more recently (i.e. neophytes, Fig. 4.8a) and hence may not have established their full potential range. The results suggest that genome size may be a helpful addition to models that aim to determine species at risk of disappearance from the British flora as a whole and/or locally or at risk of becoming dominant and hence potentially affecting the survival of native species. This could be because genome size covaries with various functional traits (e.g. Bennett, 1971; Bennett, 1972: Masterson, 1994; Beaulieu et al., 2007; Knight and Beaulieu, 2008; Veselý et al., 2012; Sparrow and Miksche, 1960; Veselý et al., 2013; Roddy et al., 2020; Théroux-Rancourt et al., 2021). Unlike these other traits however, which can vary with development, age and environment, genome size is an inherent character that is relatively easy to obtain.

Chapter 5 Genome size informs predictions of species at risk of decline mediated through functional traits

This chapter is formatted for submission; however, the authors and the journal are still undecided.

Abstract

The identification of species in decline is of vital importance in a time of unprecedented anthropogenic changes that require targeted conservation actions. Traits and plant characters shape the functional and environmental niche that plants are able to occupy and can consequently help us to distinguish species predisposed to decline in response to environmental change. Genome size sets limits on and correlates with a multitude of plant traits, and may offer additional information to models seeking to identify species at risk. However, its putative value for such analyses remains underexplored.

Based on species records from the flora of the United Kingdom, I use the Frescalo method to calculate decreasing and increasing species trends (based on regressions of relative frequencies) over the past thirty years and characterise 'winners' and 'losers' with regard to status, biome associations and genetic characters. Using a random forest classification algorithm built on functional traits, Ellenberg indicator values and genome size, I predict species trends and determine if genome size can be an informative addition for such predictions. Path analysis is used to explore how genome size might be linked with trend via interactions with traits and niche requirements.

My findings indicate that species showing increasing trends are typically non-natives from Mediterranean biomes with larger genomes. Random forest derived predictions of trend categories correctly identify species with declining trends in 77% of cases with an overall model accuracy of 70%. Genome size emerges as a helpful feature for pinpointing species at risk, and appears to exert its role indirectly via impacts on functional traits.

These findings suggest that genome size can help us improve trait-based models for the identification of species at risk from environmental change. Although the extent to which this applies to species outside the UK remains to be explored, trait-based models including genome size promise to be highly beneficial for informing targeted conservation, especially in areas where distribution data is sparse.

Introduction

In the context of global biodiversity in decline and anthropogenic threats to it mounting further (Bellard, Marino & Courchamp, 2022), methods that aid our ability to identify early those species heading for extinction are crucial. Vulnerability assessments such as those undertaken by the IUCN Red List (2022) are the most widely adopted approach for the identification of species at risk, but distribution data at sufficient temporal and spatial resolution to support Red List assessments are not always available. Consequently, any information that can be linked with increased risk of decline and extinction is crucial to allow policymakers to prioritise focal species, thus maximising conservation effects (Pearce-Higgins *et al.,* 2017).

As basic descriptors of plant function, functional traits have long received attention as predictors of species' responses to environmental gradients and their ability to adapt to change. Indeed, Alexander von Humboldt pioneered the exploration of plant trait patterns and their link with the environment as far back as the early 19th century (Päßler & Ette, 2020). Since there is a vast array of plant traits and the ease with which they can be sampled differs greatly between trait types, much research has focused on recognising major dimensions in plant function and determining which traits are the most suited to capturing this fundamental diversity. For example, Grime's CSR (competitor – stress-tolerator – ruderal; 1974 & 1977) scheme is often used as a concise conceptualisation of strategy information inherent in functional traits. While CSR assessments are certainly useful, their use is often limited to comparisons within local floras in which they are calibrated, with only recent advances towards CSR classifications that are built upon globally comparable trade-offs between frequently measured traits (Pierce *et al.,* 2017). The leaf-height-seed scheme was proposed by Westoby (1998) as an alternative to the complex strategy descriptors that are Grime's CSR axes, suggesting instead the use of specific leaf area, canopy height and seed mass as easily measured plant traits that cover the major dimensions of the functional diversity exhibited by plants, a concept that has received support over the CSR scheme (Pierce *et al.,* 2014). Broadly, plant height is typically considered a reflection of the plant's proclivity to dominate vegetation cover and capture light, seed mass reflects dispersal and establishment ability and leaf traits (specific leaf area in particular) characterise the dynamics and trade-offs of plant growth and resource efficiency (higher specific leaf area is often found in fast-growing species) (Violle *et al.,* 2009; Thomson *et al.,* 2011; Tamme *et al.,* 2014, Carboni *et al.,* 2016). The notion that height and leaf economics are major axes of plant function concurs with the findings of Díaz *et al.* (2016) in their analysis of the entire global plant trait space. Leaf traits, height and seed mass have been used in the assessment of species responses to their biotic and abiotic environment (e.g. Lake & Leishmann, 2004; Pollock, Morris & Vesk, 2012; Carboni *et al.,* 2018) and the need for sophisticated models that capture the variable effects of functional trait combinations was highlighted by Vesk (2013) and Vesk *et al.* (2021).

Similar to plant traits, Ellenberg's indicator values (Ellenberg, 1974; Ellenberg *et al.,* 1991, see glossary; Table 2.1) offer fundamental information about a species, reflecting their ecological requirements. The broad axes of the indicator values represent moisture (Ellenberg F), nutrients (Ellenberg N), light (Ellenberg L), soil reactivity (Ellenberg R) and salinity (Ellenberg S). Based on quantitative observations of realised niche conditions in the field, their true ability to describe the abiotic environment in the way Ellenberg intended has been disputed (Schaffers & Sýkora, 2000). However, they do represent strikingly informative characterisations of niche requirements – a type of data that would otherwise require extensive environmental sampling (Diekmann, 2003) and careful integration of a variety of factors (e.g. for Ellenberg F: soil moisture content, precipitation, ground water level etc. (Schaffers & Sýkora, 2000)). Unsurprisingly, Ellenberg values have become popular metrics in attempts to predict plant performance based on niche preferences (Pyšek, Prach & Smilauer, 1995; Thompson & McCarthy, 2008; Lim *et al.*, 2014; Powney *et al.*, 2014a).

While traits and Ellenberg values are frequently employed as predictors of species success, the role that genome size, ploidy and chromosome numbers might play has received little attention, although multiple studies point towards an influence of these genetic characters on trait space (Faizullah *et al.,* 2021), community composition (Guignard *et al.,* 2016; Šmarda *et al.,* 2013; Peng *et al.,* 2022) and responses to environmental pressures (Vinogradov, 2003; Hegarty & Hiscock, 2008; Pandit, White & Pocock, 2014; Chapter 4). Genome size is at the very base of plant physiology, setting hard biophysical limits on minimum cell size, packing densities and cell-division speed (Van't Hof & Sparrow, 1963; Francis, Davies & Barlow, 2008; Beaulieu *et al.,* 2008; Símová & Herben, 2012; Roddy *et al.,* 2020; Bennett, 1971 & 1972). Unsurprisingly, these limits mean that genome size has been found to correlate strongly with a multitude of traits and characters (including stomatal size, pollen size, UV-sensitivity and life strategy (Masterson, 1994; Beaulieu *et al.,* 2008; Knight *et al.,* 2010; Knight & Beaulieu, 2008; Sparrow & Miksche, 1960; Veselý *et al.,* 2012; Veselý *et al.,* 2013)).

Genome size has also been shown to correlate with the functional traits included in the leaf-height-seed hypothesis. A clear positive relationship between genome size and seed mass, assumed to stem from the constraint of genome size on minimum cell size, has long been established (Knight, Molinari & Petrov, 2005; Beaulieu *et al.,* 2007). However, the effect of genome size on plant height and specific leaf area is less clear and varies depending on the clade in question. Trees typically have smaller genomes, leading to an overall negative association between genome size and height (Knight & Beaulieu, 2008), but non-woody species appear to show the opposite trend, i.e. a positive correlation between genome size and plant height (Rios, Kenworthy & Munoz, 2015; Herben *et al.,* 2012). Specific leaf area, too, has been shown to be positively or negatively associated with genome size, depending on the taxonomic context (Kang *et al.,* 2014; Herben *et al.,* 2012).

Studies of the associations between Ellenberg values and genome size have been sparse (Bureš *et al.,* 2004; Chrtek *et al.,* 2009; Kubešová *et al.,* 2010) and correlations have not been observed consistently between genome size and the indicator values. Meanwhile there is theoretical support for a potential link between genome size and the Ellenberg values for nutrients (N) and moisture (F). The hypothesised altered water use efficiency in plants differing in genome size (Faizullah *et al.,* 2021), supported by the finding that plants occurring in humid conditions tended towards larger genomes than those from arid environments (Veselý *et al*., 2020), suggests that the maintenance of larger genomes might lead to a preference for higher moisture levels and consequently a higher Ellenberg F score in those species. Much more support is available for a link between nutrient levels in the soil and species with larger genome sizes, which would suggest the existence of a positive link between Ellenberg N and genome size. The costly nature of nucleic acids with regard to nitrogen and phosphorus was proposed by Hessen *et al.* (2010), and Šmarda e*t al.* (2013), Guignard *et al*. (2016) and Peng *et al.* (2022) have all demonstrated that nitrogen and phosphorus enrichment favours species with increased genome size, leading to changes in species community composition in field experiments.

In addition to genome size, the conjunction of ploidy and chromosome number is a further component of the genetic make-up of a species. Although linked by the history of genome duplications in a species' ancestry, a trend towards genome downsizing following polyploidy means that plants with higher ploidy levels may not necessarily have larger genomes (Renny-Byfield & Wendel, 2014). Indeed, in certain circumstances, the expected effects of large genomes and high ploidy levels contradict one another, as in the case of invasiveness, where species with larger genomes are less likely to be invasive, but those with higher ploidy level (and chromosome number) are more likely to be invasive

(Vinogradov, 2003; Pandit *et al.,* 2014). This points to the potential benefits of including all genetic characters together in analyses of trait driven species success to capture all potentially important information contained across them.

The complex roles that genome size, ploidy and chromosome number are expected to play in the context of ecology suggest that their inclusion in models of species success may be important. Indeed, Herben *et al.* (2012), demonstrated that even when accounting for functional traits, genome size offered additional predictive value in models of regional plant abundance.

In this chapter, I construct a random forest classifier to distinguish between plants with increasing and decreasing species trends (based on regressions of relative frequencies, see Chapter 3, Fig. 3.2) across the UK in the past thirty years, based on Ellenberg values, functional traits and genetic characters. Following an assessment of variable importance from the random forest models, I then conduct a phylogenetic path analysis to gain insights into the way in which predictors tie in with one another to exert their effect on trend outcomes, focusing in particular on the way that genome size might factor into the equation.

Materials & Methods

Estimation of species trends

The trend information derived from the Frescalo bias correction on plant detection/ non-detection data from the Botanical Society of Britain and Ireland's (BSBI) distribution database (DDb) outlined in Chapter 3 was used here to derive insights into the success of individual species across the three most recent complete date classes of data (1987-1999, 2000-2009 and 2010-2019). The trend estimate (visualised in Fig. 3.2) is based on a regression across the relative frequencies of a species in each of the date classes, as compared with benchmark species. Species showing a positive regression coefficient are described as showing an increasing species trend (‘winners’), while those with a negative coefficient represent species with a declining trend (‘losers’). All results from Frescalo runs, including the trend regression results, are available in Appendix 2.

To test the reliability of the Frescalo-estimated species trends, I also (i) calculated alpha hulls (Edelsbrunner *et al.,* 1983; Burgman & Fox, 2003) as a measure of Extent of Occurrence (Joppa *et al.,* 2016) using the ‘ConR’ package (Dauby *et al.,* 2017), with alpha = 0.2, (ii) derived the decadal change in alpha hull size and (iii) compared this decadal change with the Frescalo trend estimates. Both measures were clearly positively associated and consequently the Frescalo-based trend estimate was used in the following analyses since it had the added benefit of intrinsic correction for sampling bias.

Dataset compilation

I assembled a dataset of potential predictors of species trend – as defined above – from the ‘BIFloraExplorer’s’ (Henniges *et al.,* 2021) functional trait data, niche descriptors and information on genetic characters. The dataset for all presented analyses following the Frescalo correction was restricted to angiosperms associated with a graminoid or herbaceous growth form (i.e. not woody), explicitly filtering out species associated with phanerophytic life forms (see glossary; Table 2.1).

The functional traits specific leaf area, leaf area, leaf dry matter content, mean canopy height and seed mass were used in analyses. Use of Grime’s CSR values within the models was considered, but since the scores along the three axes had been determined by utilising Pierce *et al.*’s (2017) suggested method, using the trade-offs between specific leaf area, leaf area and leaf dry matter content, this meant that use of the CSR scores would have necessitated exclusion of all leaf traits, making interpretation in the context of the leaf-height-seed scheme impossible.

Ellenberg indicator values describing the realised niche of plants with respect to nutrients (N), water (F), soil reactivity (R), salinity (S) and light (L) *sensu* Hill, Preston & Roy, 2004 were used as presented in the 'BIFloraExplorer' (Henniges *et al.,* 2021). Assessments by Döring (2017) were not used to fill gaps in our knowledge of realised niche descriptions, since the mixture of subjective estimates from two different sources in the same quantitative analyses posed the risk of confounding results. Additional information on Ellenberg values is presented in Chapter 2, Table 2.2.

The genetic characters chromosome number, ploidy level and genome size were also included in the dataset of potential predictors of species trend. I used the dataset of genetic information previously described in Chapter 4, taking into account cytotype variation by treating each cytotype as an additional 'species' that shares the same traits, Ellenberg values and trend, but differs with regard to chromosome number, ploidy and genome size.

In addition to the dataset on predictors for subsequent analysis, I also created a dataset to characterise and give an overview of the species with increasing and decreasing species trends with regard to their status (*sensu* Stace, 2019, i.e. native, archaeophyte, neophyte etc., see glossary; Table 2.1), their biome association (i.e. Temperate, Mediterranean, Boreal etc.), and CSR strategy (all from the 'BIFloraExplorer' (Henniges *et al.,* 2021), see Chapter 2 and glossary).

The phylogeny described in Chapter 2 (with attachments of cytotypes as in Chapter 4) allowed me to test for phylogenetic signal in all predictor variables (Pagel's λ with 10,000 randomisations) and the trend data ('D-statistic'; Fritz & Purvis, 2010), using the 'phytools' package and the 'caper' package (Revell, 2012; Orme *et al.,* 2013).

Data management, visualisations and formal analyses

All data management, manipulation and analyses were carried out in R (version 4.1.3), relying on 'tidyverse' packages (Wickham *et al*., 2019). Plots were generated using 'ggplot2' (Wickham, Chang & Wickham, 2016), 'treemap' (Tennekes & Ellis, 2017) and 'ggtern' (Hamilton & Ferry, 2018). Differences in each predictor between trend categories were assessed using phylogenetically corrected ANOVAs as implemented in the 'phytools' package with 1,000 randomisations. All steps of random forest runs were executed in the 'tidymodels' framework (Kuhn & Wickham, 2020).

Random Forest

Due to the expected highly complex role of genetic characters in affecting species trends, I chose the random forest algorithm to build a predictive model for binary species trends. The random forest, first proposed by Breiman (2001) is a machine learning algorithm founded on the basic principle of the decision tree, where successive splits in the dataset based on predictor variables are used to arrive at highly accurate predictions of an outcome variable. Improving the predictive power of this very simplistic algorithm, the random forest uses the concept of the wisdom of the crowds, constructing an entire 'forest' of decision trees, each based on a random subset of data, and then averaging across their individual predictions (Liu, 2014; Biau & Scornet, 2016). In addition to achieving high predictive accuracy on non-linear problems, being remarkably robust to outliers and making no assumptions about interdependencies in the data, the random forest is also more interpretable with regard to variable importance (Auret & Aldrich, 2012).

Feature selection for the random forest algorithm was based on multiple steps. First, I calculated and inspected Pearson correlations between all predictors, noting that none of the correlations warranted exclusion of variables (all well below 0.5, excepting the correlation between ploidy and chromosome number which was 0.62). Secondly, I applied two variable selection steps contained within the 'tidymodels' framework: a near-zero variance filter to exclude sparse and unbalanced variables and a filter that removes variables that are linear combinations of one another which would make model fitting and later inferences about variable importance challenging. Neither filter suggested the removal of any of the predictors. Thirdly, I performed Boruta feature selection (Kursa, Jankowski & Rudnicki, 2010) as implemented in the R package 'Boruta' (Kursa & Rudnicki, 2010) on ten random subsamples. This process aims to remove 'unhelpful' variables and to this end generates 'shadow features', an alternative version of a specific variable, where observations are randomly shuffled against the outcome to generate an arbitrary and consequently non-predictive version of the original feature. The shadow features are then included in a set of random forest iterations. Only those features that are persistently found by the algorithm to be more helpful than the most informative of the randomised shadow features are retained for the final model (Kumar *et al.,* 2017). Boruta feature selection was run with 100 iterations on each of ten random subsets of the original dataset. Based on the findings of the Boruta feature selection, both ploidy and chromosome number were excluded since they were not found to improve models. All functional traits, Ellenberg values and genome size were retained since they all provided helpful information for models to reach their final predictions. Since species with decreasing trends were more common than those with increasing trends, I applied the synthetic minority over-sampling technique (Chawla *et al.,* 2002) implemented in the 'themis' package (Hvitfeldt, 2020).

For use in the subsequent models, I experimented with using the outcome (trend) either as a numeric variable (directly derived from Frescalo assessments), coded as a categorical variable (decreasing, neutral, increasing) with a threshold of +/- 10% change to distinguish the categories, or as a binary variable (decreasing and increasing). When using the numeric or threshold representation of trend in the random forest models, I achieved very low levels of accuracy with the set of predictors outlined above. Hence, the outcome variable is treated as binary in this chapter and the random forest classifier algorithm is used to predict whether a species falls into the decreasing or increasing trend category based on functional traits, Ellenberg values and genome size.

I generated ten random subsets of the data and within them performed ¾ training/testing splits, leaving ¼ for model testing. From each training set, I created a random 10-fold cross-validation object for tuning of the hyperparameters 'mtry' (the number of predictors available to the algorithm at each split), 'trees' (number of trees to be built for the random forest) and 'min_n' (the number of observations at which a node must be split further), which were tuned via grid-based tuning, aiming to maximise accuracy.

The final models with tuned hyperparameters were run in the ranger engine on the testing sets to assess the final performance of the models. The total number of trees generated across all random forests was 8,000. Estimates given below summarise the average performance across the ten runs on random subsamples of the dataset. Variable importance as estimated by the random forest models is presented as averages across all ten independent runs, giving an indication of the variation in importance observed on different subsets of the original data. Instead of the default mean decrease in impurity importance metric (Breiman, 2001), I used the more elaborate permutation-based variable importance metric which considers features to be important if they improve the prediction accuracy of the overall model (Cutler, Cutler & Stevens, 2012). Since Pearson correlations between the used predictors are low, the potential pitfall of permutation-based variable importance was avoided (Cutler, Cutler & Stevens, 2012).

Phylogenetic path analysis

Since random forests cannot give clear insights into the interconnectedness of variables that allow them to exert their predictive potential (Auret & Aldrich, 2012), I conducted a confirmatory path analysis in an attempt to untangle the predictors' relationships, with a particular focus on the way genome size might be linked with the other predictors and how it might relate to trend outcomes. Based on multiple regression, this approach represents variables within a network of interdependencies, comparing the feasibility of a range of causal models in the context of the data presented to it. Due to the levels of phylogenetic signal present within each of the predictors, I used a phylogenetic path analysis as implemented in the package 'phylopath' (van der Bijl, 2018 & 2022). Developed by Hardenberg & Gonzalez-Voyer (2013) by integrating the concept of Phylogenetic Generalised Least Squares (PGLS; Symonds & Blomberg, 2014) with Shipley's (2000) 'd-separation' method, phylogenetic path analysis allows users to gain insights into the unresolved causal structures at the root of correlations (Shipley, 2016; Gonzalez-Voyer & Hardenberg, 2014) in the presence of phylogenetic signal.

Since an abundance of variables within path analyses can lead to confusing and unintelligible causal structures (Streiner, 2005), I simplified the model to include only a subset of predictors (genome size, the functional traits specific leaf area, seed mass and vegetative height as well as the Ellenberg values for moisture (F) and nutrients (N)), with binary trend as the outcome. The selection of these Ellenberg values was based on the identity of Ellenberg F and N as the most powerful predictors identified by the random forest runs. The three functional traits were chosen since they are most frequently used to represent the three dimensions of the leaf-height-seed strategy scheme (Westoby, 1998).

To meet assumptions of the regressions within the phylogenetic path analysis, all functional traits and genome size were log-transformed. All variables were also standardised. Phylogenetic path analysis based on the same dataset used for random forest runs, but restricted to the above traits, was run at default settings, with phylogenetic correction based on the phylogeny constructed and described in Chapter 2, including the addition of cytotypes described in Chapter 4.

I tested a set of 16 hypothesised causal models, summarised as acyclic path diagrams in Fig. S5.1. All models are based on exploratory PGLS analysis conducted on functional traits, Ellenberg values and genome size. Since bidirectional arrows would cause difficulties in the interpretation of path analysis results (Streiner, 2005), arrows between the two Ellenberg values and among all the functional traits were assumed to be unidirectional. The competing models were ranked by their C-statistic information criterion (CICc) to assess support of each hypothesised causal structure. The structure with the best support was used to build the final directed acyclic diagram with 500 bootstrap iterations. Strength of paths is expressed as the standardised regression coefficients and their significance was derived via 95% confidence intervals.

Only complete observations (i.e. species that had full coverage across all functional traits, niche descriptors (Ellenberg values) and genetic data) were included in random forest models, leaving 960 observations including cytotypes that were counted as separate species (98 species had more than one cytotype, leading to 104 observations being added due to additional cytotypes). The removal of species without phylogenetic information meant that this number was restricted to 784 for the phylogenetic path analysis. The visualisation of 'winners' and 'losers' with regard to distribution trend is based on all available data since completeness across all traits was not required for that analysis. Sample sizes are given with the figures.

Results

Characterising 'winners' and 'losers'

Trend information (based on regression across relative occurrence frequencies, Chapter 3, Fig. 3.2) is available for 2,249 species, with 1,353 showing decreasing and 896 species showing increasing trends. Of these, 1,195 are native plants, 160 are archaeophytes, 885 are neophytes, and 3 are neonatives.

Of those species with genome size and phylogenetic data available (i.e. 1,743 species, including cytotypes), the results show that genome size stands out in the context of native

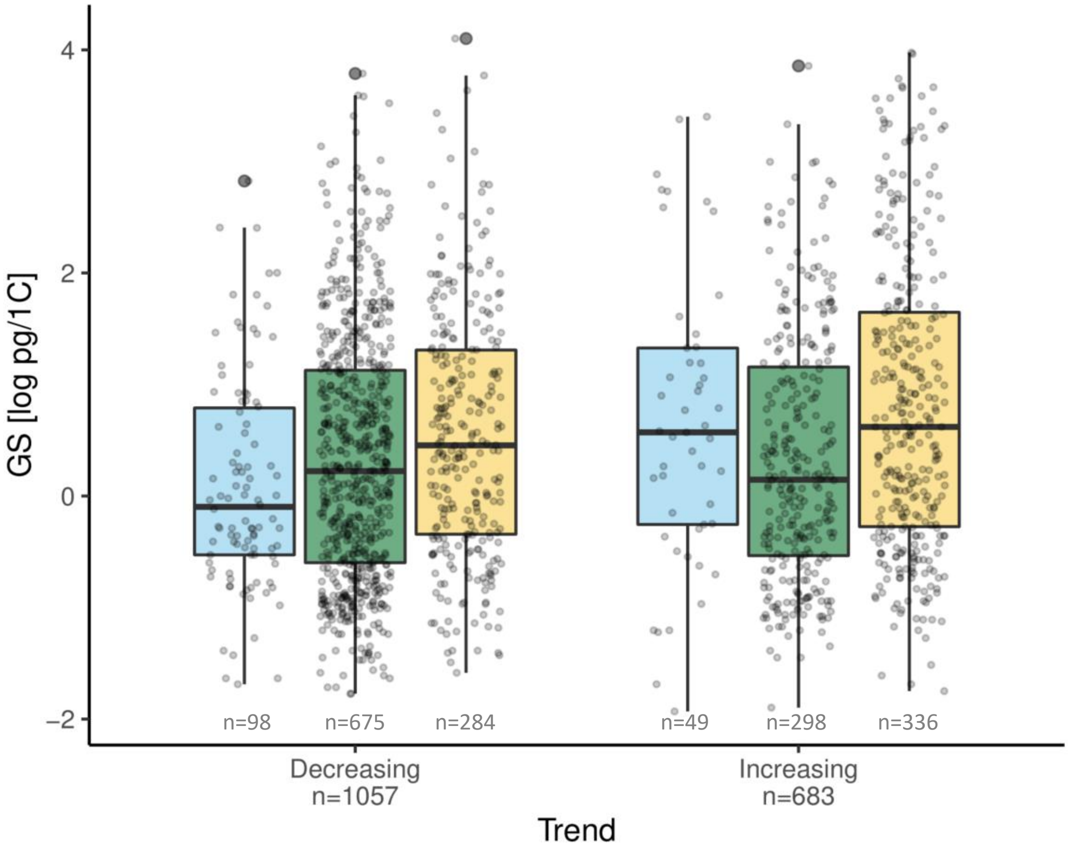


Fig. 5.1 Trend, status and genome size of 1,743 species. Boxplots show the log transformed genome size profiles of species with different status within the decreasing and increasing groups. Green stands for natives (973 species), blue for archaeophytes (147 species), and yellow for neophytes (620 species). The increasing group is dominated by neophytes (49.2%) and is characterised by overall larger genomes. The decreasing category contains mostly native species (63.9%). The data shown here is reduced to those species with phylogenetic information, concurring with data used for the phylogenetic ANOVA and does not include neonatives (n=3).

and non-native plants (Fig. 5.1). Species with an increasing trend (684 species) have significantly larger genome sizes (phylogenetic ANOVA F = 12.74, $p < 0.05$; mean genome size 4.21 pg/1C compared with a mean genome size of 2.86 pg/1C among the 1,059 species showing a decreasing trend) and neophytes are clearly more prominent within the increasing group (298 natives, 49 archaeophytes, 336 neophytes and 1 neonative) compared with species following a decreasing trend (675 natives, 98 archaeophytes, 284 neophytes and 2 neonatives).

Fig. 5.2 represents the make-up of the group of decreasing ('losers') and increasing ('winners') species with regard to status (Fig. 5.2a & b), biomes (Fig. 5.2c & d) and genetic characters (genome size and ploidy, Fig. 5.2e & f). The majority (58.9%) of species showing increasing trends is made up of non-natives (Fig. 5.2a), especially naturalised neophytes, who account for 36.8% of all species in this group, with casual and survivor neophytes accounting for 11.4 and 4.8% respectively. 41.0% of the increasing group are native plants and only 5.9% are archaeophytes. Conversely, when looking at the group of species with decreasing trends, natives are in the majority with 61.4%, neophytes make up a significantly smaller proportion with especially the proportion of naturalised neophytes and survivors only half (18.8% and 1.9%) of their proportion among 'winners'. Archaeophytes are more common among the 'losing' species with 7.9% (Fig. 5.2b).

With regard to the biomes that species are associated with, the most striking difference is the proportion of Mediterranean (both Mediterranean and Mediterranean-Atlantic) species, which are far more prevalent among 'winners' (24.2%) than among 'losers' (11.0%, Fig. 5.2c & d). The most prominent biome association in both groups, however, is a variety of Temperate biome types, which collectively account for 65.9% of 'winners' and 72.1% of 'losers'. While fewer Temperate and Boreo-Temperate species fall within the 'winning' category, species associated with Southern Temperate and Wide Temperate biomes make up larger proportions among the 'winners' (22.1% and 4.4%) than among the 'losers' (16.7%

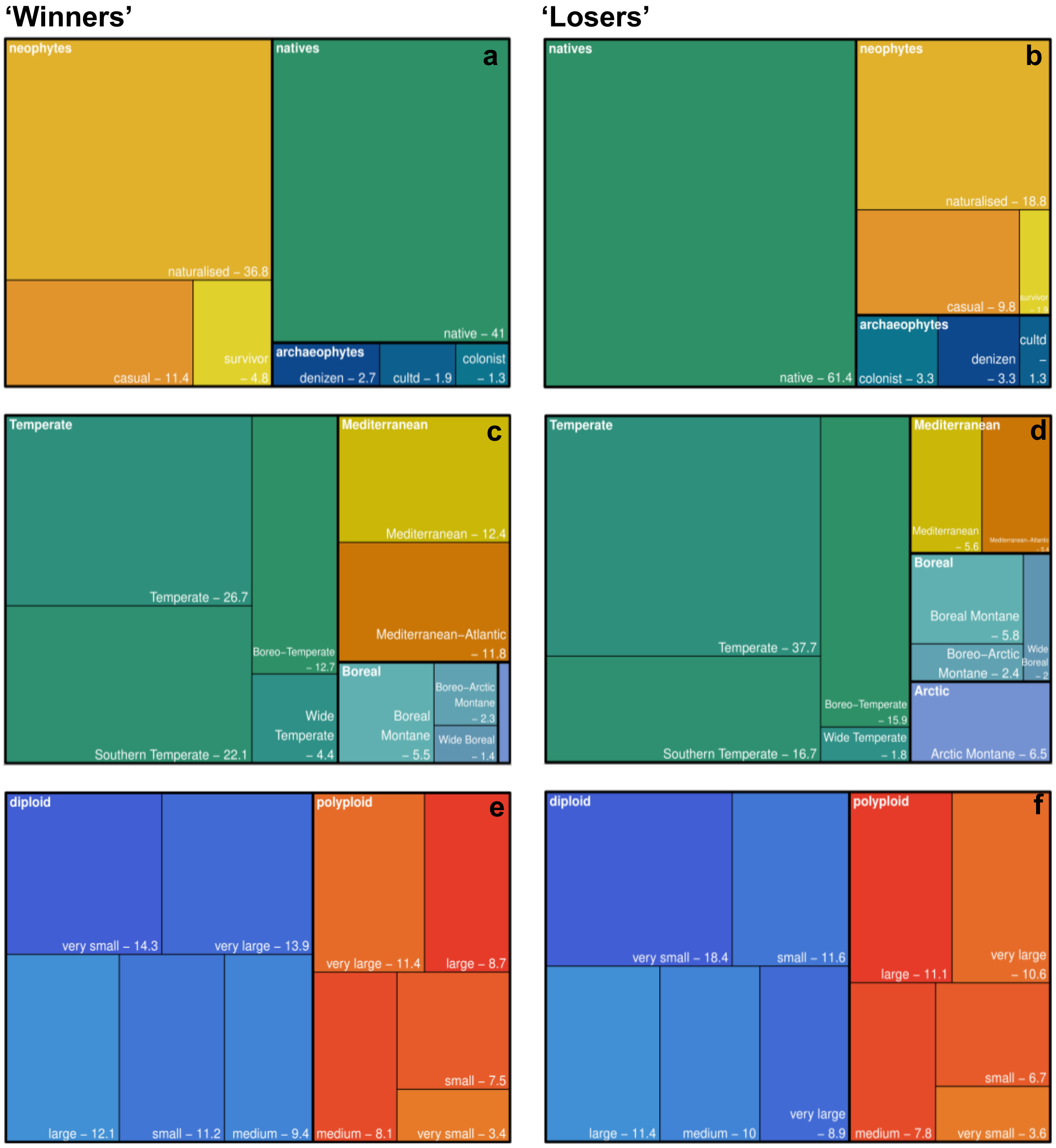


Fig. 5.2 Treemaps representing the composition of the increasing and decreasing groups of species. a and **b** represent status (native, neophytes, archaeophytes; see Table 2.1), **c** and **d** show the biome (Temperate, Mediterranean, Boreal, Arctic etc.) are associated with and **e** and **f** illustrate genetic categories (ploidy and genome size). **a**, **c** and **e** represent the group of 'winners' (i.e. increasing trends), **b**, **d** and **f** show the composition of the decreasing group. Subgroups within **e** and **f** correspond to genome size quintiles (very small: 0.15 - 0.53 pg/1C, small: 0.54 - 0.90 pg/1C, medium: 0.91 - 1.59 pg/1C, large: 1.60 - 4.18 pg/1C, very large: 4.3 - 47.3 pg/1C). Numbers after group names show the percentage that this category takes up within the overarching groups (plants with increasing or decreasing species trend). Treemaps are based on available data for each descriptor (a – 895 species, b – 1,347 species, c – 434 species, d – 941 species, e – 678 species, f – 1,055 species).

and 1.8%). Boreal species representation is relatively balanced between the two groups, with 9.2% among 'winners' and 10.2% among 'losers'. This is in stark contrast to plants with a preference for Arctic conditions (Arctic Montane biomes), who are hardly represented within the increasing group (0.7%), but contribute 6.5% to the 'losing' group, showing that almost all species associated with this biome are suffering declines.

Polyploid plants are just as prevalent among 'losers' (39.6%) as among 'winners' (39.1%, 5.2e & f). Genome sizes, here given in quintiles (very small: 0.15 - 0.53 pg/1C, small : 0.54 - 0.90 pg/1C, medium: 0.91 - 1.59 pg/1C, large: 1.60 - 4.18 pg/1C, very large: 4.3 - 47.3 pg/1C), show some subtle shifts between groups. Plants with very small genomes are the most prevalent group among diploids across both the 'winners' and 'losers' (14.3% and 18.4%, respectively). The second largest group among 'winners' are diploid plants with the largest genome sizes (13.9%). This group is the least frequent diploid quintile among 'losing' species (8.9%). The quintile representing large genome sizes is slightly more common among 'winners' (12.1%) than among 'losers' (11.4%). Among the polyploids, the largest categories are those with large or very large genome size, but while the large genome size quintile accounts for 8.7% of 'winning' species, it accounts for 11.1% among the 'losers', with the very large quintile overtaking the large quintile as the group with the highest proportion among 'winning' species.

Finally, I characterised species with decreasing and increasing trends with regard to their CSR-strategy (Grime, 1974; Fig. 5.3). Species with increasing trends show a tendency towards a competitive lifestyle and are less likely to be ruderals or stress-tolerators compared with species showing a decreasing trend which more often lean towards adopting a ruderal life strategy. This tendency is not specific to natives or non-natives, since both groups show the same pattern when visualised independently (Fig. 5.3b & c). However, while natives overall and particularly those with increasing trends show a stronger proclivity towards stress-tolerance, the opposite is true for non-natives, which are generally less likely to be stress-tolerators, with the mean S-score for 'winning' non-natives being close to zero.

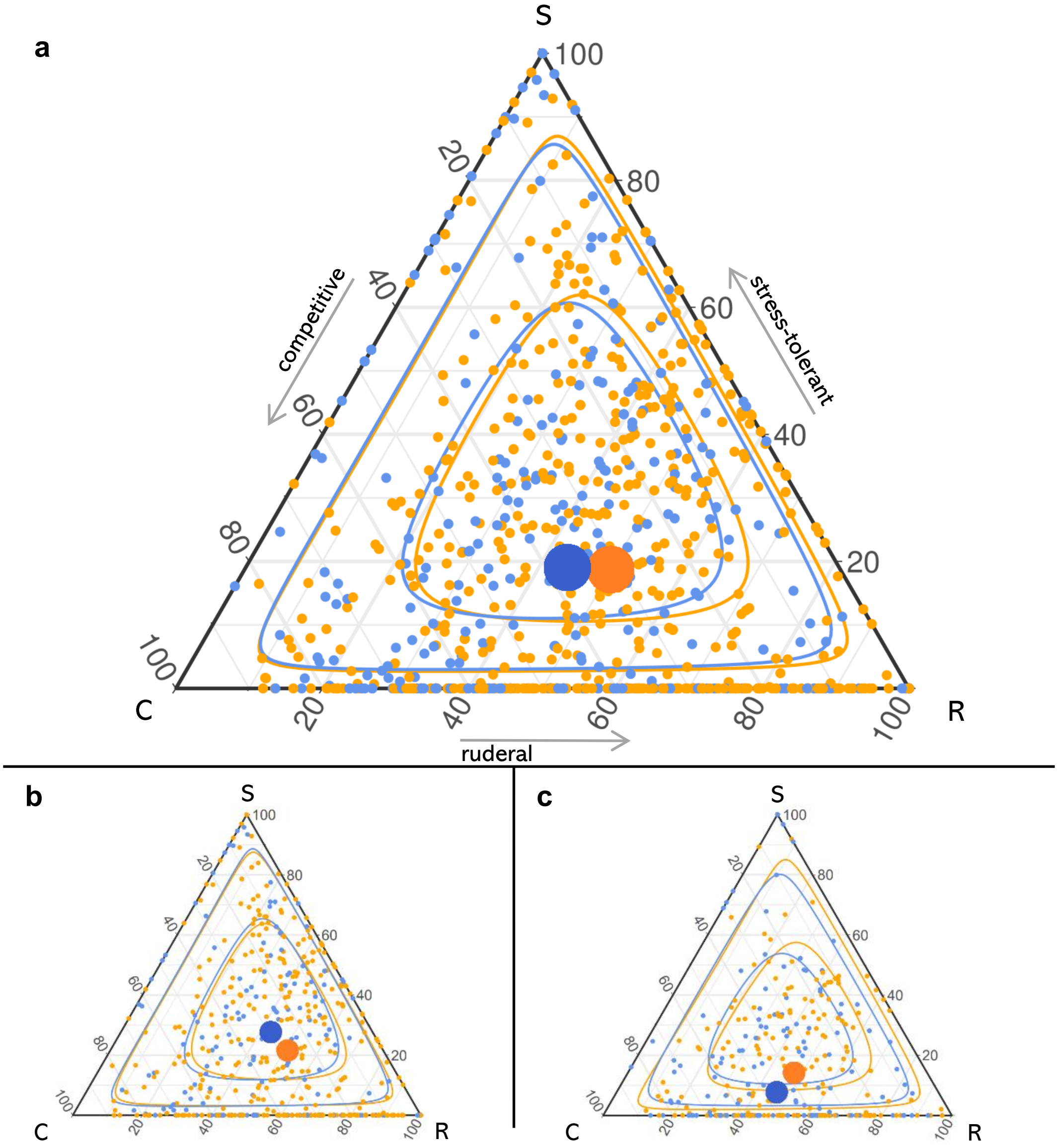


Fig. 5.3 Ternary plots of CSR strategy and trend. Life strategies of species are represented as their location within the space between the three poles competitor (C), stress-tolerator (S) and ruderal (R), in accordance with Grime's CSR scheme. Colours indicate the group each species falls into, with orange denoting decreasing and blue denoting increasing trend. The centroid of each group is represented by the large, darker dots. 50 and 95% confidence interval as calculated using Mahalanobis distance (Hamilton, 2015). **a** represents all 915 species with available data for CSR scores and trend, **b** and **c** show the available data for natives (571 species) and non-natives (342 species) respectively. Species showing an increasing trend are tending more towards adopting a competitive life strategy and are less likely to be ruderals or stress-tolerators. Stress-tolerance is more typically observed among natives generally and those natives showing an increasing trend in particular, while non-natives with increasing trends are rarely stress-tolerators.

Predictions of species trends

I used a random forest classifier to find out if I could accurately predict if a species was showing an overall increasing or decreasing species trend based on traits, niche preferences (Ellenberg values) and genetic characters (although only genome size was included in final models following Boruta feature selection).

Out of the 960 species with complete data across all predictors, 651 were decreasing and 309 had an increasing trend. This skew was addressed with the themis package's synthetic minority over-sampling technique (SMOTE, Chawla *et al.,* 2002). The removal of all species that did not have full coverage across all the functional trait, Ellenberg value and genetic character data meant that native species dominate the dataset used for further analysis (739 natives, 113 archaeophytes, 106 neophytes, and 2 neonatives).

There was no strong phylogenetic signal within the binary trend (increase or decrease) data (D = 0.7417643, D-statistic of zero corresponds to Brownian motion evolution, a value of one corresponds to complete absence of signal). Consequently I did not include phylogenetic information in the random forest runs. There was, however, significant phylogenetic signal within the genetic characters (genome size λ = 0.907082, ploidy level λ = 0.223224, and chromosome number λ = 0.379103, all $p < 0.0001$), the functional traits (specific leaf area λ = 0.846305, leaf dry matter content λ = 0.767634, leaf area λ = 0.0803706, mean vegetative height λ = 0.894468, all $p < 0.0001$; and seed mass λ = 0.027421, $p < 0.05$), and the Ellenberg values (F λ = 0.958421, N λ = 0.83935, R λ = 0.698443, L λ = 0.793788, S λ = 0.945842, all $p < 0.0001$). Nevertheless, the phylogenetic signal present within the predictor variables is not problematic for random forest predictions since the non-parametric nature of random forests means they make far fewer assumptions about variable independence and normality than parametric tests (Dankowski & Ziegler, 2016).

All parameters except ploidy and chromosome number were found to be helpful to the random forest models in making predictions by the Boruta feature selection step.

The final accuracy in predicting binary trend achieved across runs was 69.8% with a ROC AUC of 73.1% (perfect distinction between both categories would correspond to a score of 100% on both metrics). Out of bag (OOB) prediction error rates averaged 14.2%. The algorithm correctly identified decreasing species in 76.9% of cases, and species with increasing trends in 52.8% of cases. A summary of the accuracies, error rates and confusion matrices of the runs on each subset is presented in Table S5.1.

Out of the predictors used within the random forests, the Ellenberg values for nutrients (N) and moisture (F) were the most informative, followed by leaf area, Ellenberg R (reactivity) and specific leaf area (Fig. 5.4). Mean vegetative height, leaf dry matter content

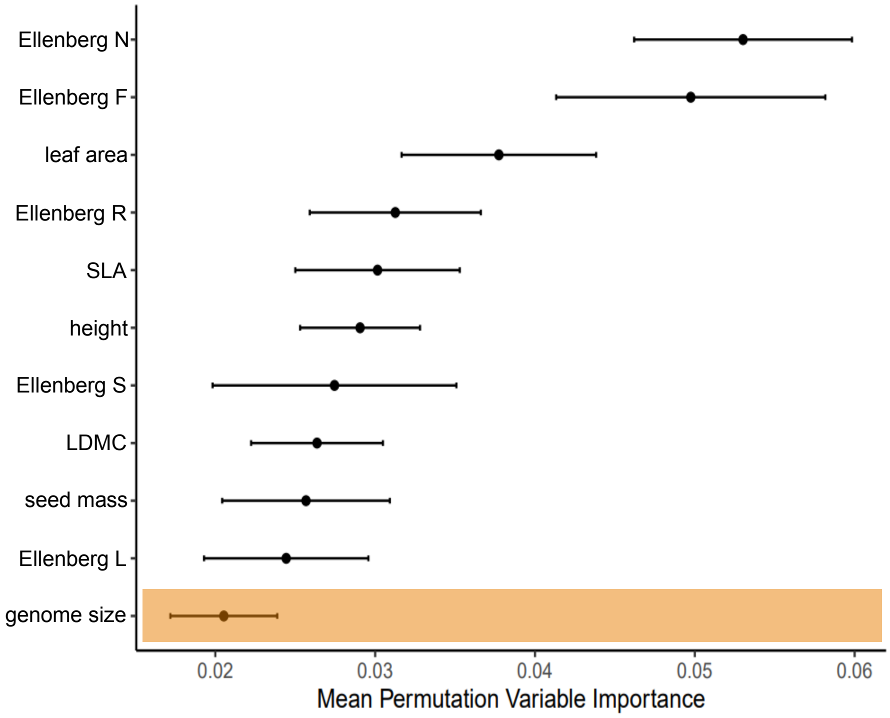


Fig. 5.4 Variable importance of random forest. The permutation-based importance of each variable in informing the final random forest ensemble for ten random subsets of the original data (8,000 trees). Mean and standard errors are derived from the ten independent runs. The focal variable genome size is highlighted in orange and contributes to the model to a lesser degree than the functional traits and Ellenberg values. The most important variables for the random forest are Ellenberg N and Ellenberg F. (SLA = specific leaf area, LDMC = leaf dry matter content).

and seed mass were less crucial, similar to Ellenberg S (salinity) and Ellenberg L (light). Finally, genome size was the least important of the eleven informative predictors.

In the dataset used for random forest runs, species showing a decreasing trend were associated with lower nutrient level (Ellenberg N, phylogenetic ANOVA F = 20.85, $p < 0.002$) and soil reactivity scores, (Ellenberg R, phylogenetic ANOVA F = 5.55, p = 0.07) but higher moisture preferences (phylogenetic ANOVA F = 6.18, p = 0.063). Their leaf areas were smaller (phylogenetic ANOVA F = 8.44, $p < 0.05$), but their specific leaf area bigger (phylogenetic ANOVA F = 2.04, p = 0.29). Genome sizes of species with decreasing trends tended to be smaller, although not significantly so in this subset with complete records across all predictors (phylogenetic ANOVA F = 3.43, p = 0.174). Results for all phylogenetic ANOVAs and group means of predictors can be found in Table S5.2.

Untangling causalities behind the predictions

Having found that inclusion of genome size is helpful (albeit the least informative of eleven informative variables) for the prediction of species trends (Fig. 5.4), I conducted a phylogenetic path analysis to find out how genome size might impact the other predictors in influencing the success of a species (here defined as showing an increasing trend). The model with by far the most support across the 16 models tested, although even this model provided a relatively poor fit with the data (p = 0.004), included genome size as an exogenous variable which directly influences seed mass, specific leaf area and vegetative height, represented as a directed acyclic graph in Fig. 5.5. A summary report for all models is available in Table S5.3. This association is positive for seed mass and height but negative for specific leaf area. Links among the functional traits vary, with a negative correlation of specific leaf area and seed mass and a positive link between height and seed mass. All functional traits were positively associated with the species' realised niche, as encapsulated in the Ellenberg N and F values, with the exception of seed mass, which showed a negative association with Ellenberg F. The effect of Ellenberg N on trend is positive and strong, suggesting that plants preferring nutrient-rich environments are more likely to exhibit increasing trends. On the other hand, the association between moisture preference (Ellenberg F) and trend was negative and non-significant. Meanwhile, an indirect pathway for the effect of moisture requirements on trend via Ellenberg N is relatively weak but significant.

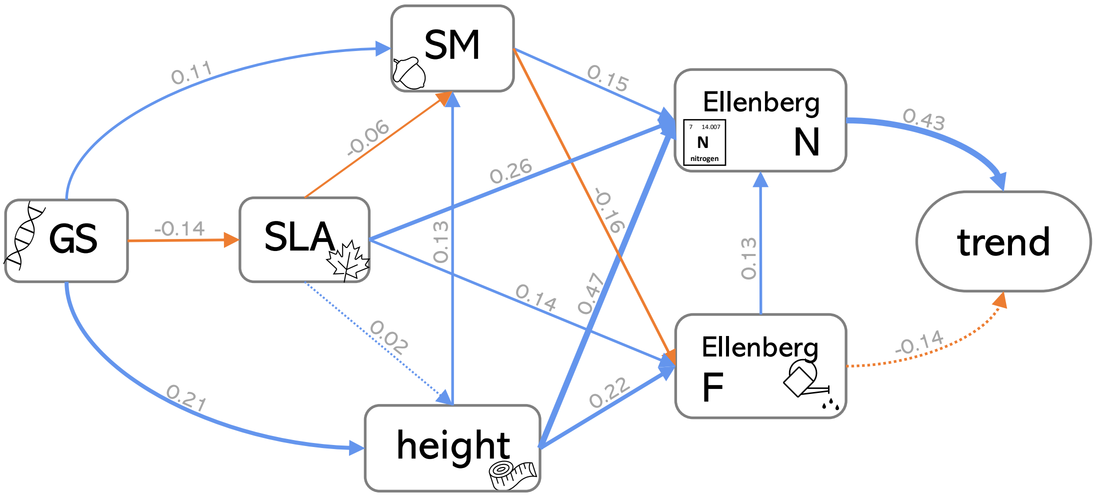


Fig. 5.5 The directed acyclic graph representing the most supported model of the relationships of predictors and species trend. The result of the phylogenetic path analysis show the hypothesised model of causal effects on species trend most supported by the data. Genome size (GS) is the only exogenous variable and directly influences the functional traits (seed mass = SM, specific leaf area = SLA and vegetative height = height), which in turn influence Ellenberg values (for nutrients (N) and moisture (F)) and finally trend. Numbers along the paths and path thickness correspond to the standardised path coefficients derived from path analysis. Orange paths indicate negative effects, while blue paths correspond to positive effects. Dashed lines indicate non-significant relationships (where the 95% interval includes zero), and solid lines indicate statistically significant relationships.

Discussion

The comparatively well-sampled flora of the UK, where species trends can be assessed over relatively long time scales with some confidence, offers itself for an exploratory study such as this. My aim in this chapter was to construct a model capable of predicting species trends based on easily measured plant characters and traits, and to explore whether genome size and other genetic traits might be able to contribute to the accuracy of such a model. I show how the model is able to predict increasing and decreasing range trends over the past three decades. The study demonstrates the value of predictive tools based on traits and characters in assessing species dynamics, as previously shown by Powney *et al.* (2014a), which might also be applicable to biodiversity hotspots worldwide, even when they do not exhibit the same density of records (Meyer, Weigelt & Kreft, 2016; Paton *et al.,* 2020). I find a role of genome size in the predictive framework, and I explore the way in which it might exert its effect in the context of the other predictors. Although currently limited to the UK, the greatest potential value of this approach of trait-based predictive models of species success with inclusion of genome size lies in its application to areas with much sparser levels of historic and current botanical activity, where species declines are more challenging to assess. Such transferability for a trait-based model of species success was previously demonstrated by Powney *et al.* (2014b). While further research is needed to demonstrate the validity of these findings in such different contexts, they could well contribute to more accurate tools to guide targeted conservation approaches.

The challenge of capturing species trends

My results show that a greater number of species within the UK is showing decreasing trends in their relative frequencies than increasing trends. The metric of species trend used here corresponds to the change in relative frequency derived with the Frescalo method (Hill, 2012) and is presented in binary format. While the *Atlas of the British and Irish Flora* (Preston, Pearman & Dines, 2002) has previously aimed to capture change (in range size) within the plants it covered via the Change Index, based on the Telfer method (Telfer, Preston & Rothery, 2002), more elaborate characterisations are required to capture meaningful measures of change from data biased by spatio-temporal recording differences

(Hodgson, 2003). Although the Frescalo-based bias correction used here follows the best practise approach currently available which is recommended for this dataset, its use does not alter the fact that the determination of accurate change metrics for plants in the flora from flawed distribution records is fraught with difficulties, which are outlined and discussed in detail in Chapter 3.

The magnitude of frequency change that a species can exhibit is directly dependent on the frequency of records available for that species at the start of the sampling period. Non-native species, especially neophytes, that are new arrivals to the flora and that are only present in very few locations or under-recorded would show explosive expansions of their range simply because they started from a comparatively low point. Such a rapid expansion would not necessarily be caused by any intrinsic trait or ability of the species to thrive, but might reflect another facet of sampling bias that cannot be entirely removed by Frescalo. Using binary trends, a decision partially dictated by low prediction accuracies of regression type random forests achieved with the dataset, levels the playing field to some extent, removing skews in the data from such spuriously extreme changes. Nevertheless, the binary nature of the data comes at the cost of showing differences between species with increasing, decreasing and stable trends. When the data were split into these three categories using a 10% decadal change threshold, as suggested in a comparable study of population changes in moths (Coulthard *et al.,* 2019), low accuracies were achieved once again and the approach was abandoned.

Species trends – who are the 'winners' and 'losers'?

The findings presented here indicate that those species with an overall increasing trend are more often neophytes, and are more often species that are associated with Mediterranean and Southern Temperate biomes. Chapter 4 had already demonstrated that neophyte species have greater northward shifts than native and archaeophyte species (Fig. 4.7), which would explain why those species are more prevalent among the group of species with increasing trends than among those that are decreasing in frequency. The human impacts described and discussed in Chapters 3 & 4 are likely to play a major role in making neophytes and plants with Mediterranean and Southern biome preferences so successful.

Guo *et al.* (2019) demonstrated that 94% of species that had become naturalised as neophytes anywhere in the world were cultivated in domestic gardens, suggesting that the flora of the UK would be strongly impacted by introductions of garden escapes. The appeal of attractive and exotic Mediterranean species (RHS Gardening, 2022) will likely have contributed to favouring those plants associated with Southern Temperate and Mediterranean biomes in their expansion across Britain. The concurrent tendency of plants associated with Arctic-Montane conditions to decline (Fig. 5.2c & d) is consistent with the previous findings by Hill & Preston (2015), who demonstrated losses of boreal plants in southern Britain. The decline of Arctic-Montane species is likely caused by climate change, which is reducing habitat availability for cold-adapted species. Climate change probably also explains part of why the Mediterranean species are increasing in frequency. Pearce-Higgins *et al.* (2015 & 2017) identified that upland species (plants and animals) were typically at risk from the effects of habitat loss due to climate change and urged towards their protection, offering support for the above finding of the striking absence of Arctic-Boreal plants from the 'winners'. The repeated introductions of attractive garden species and the increasing attention paid to non-natives in the wild likely also plays a role in the trends observed here.

Increasing species (both natives and non-natives) showed tendencies towards competitive life strategies (Fig. 5.3). At a global scale, the size of the naturalised range of herbaceous plants has previously been found to be positively associated with competitive tendencies, while native range sizes were positively associated with ruderal tendencies, and all range size estimates (native and naturalised) were negatively related to stress-tolerance (Liao *et al.,* 2021). On the much smaller scale of the UK, natives and non-natives mostly differ in the extent to which stress-tolerance contributes to their life strategy (Fig. 5.3b & c). Non-natives score extremely low on the stress-tolerance axis, especially those showing increasing species trends, while native plants have, on average, higher stress-tolerator scores, with increasing species more likely to be stress-tolerant than decreasing species. Many native species will have adapted to a very specific niche over long periods of time, including specific stress adaptations, while successful naturalisation appears to be facilitated mostly by higher ruderality and competitiveness (Guo *et al.,* 2018). The quick growth, high levels of flower and seed production and highly nutritious tissues typically found in competitive and ruderal plants, predispose them to higher invasion success while simultaneously making them more attractive for introductions by humans for horticulture and agriculture (Guo *et al.,* 2018; Pysek & Richardson, 2008; van Kleunen *et al.,* 2010; van Kleunen *et al.,* 2018), explaining the strong tendency of non-natives towards high ruderal and competitor scores. Stress-tolerance on the other hand, which is associated with slow growth cycles, low seed production and highly specialised adaptation to particular stressors (Grime, 1979; Grime & Pierce, 2012; Alexander *et al.,* 2011) is not helpful for naturalisation but appears to be a strategy associated with successful natives in the UK, perhaps suggesting an advantage of stress adaptations when niches are faced with ongoing environmental changes.

Finally, species with increasing trends are more often associated with larger genome sizes than those with decreasing trends (Fig. 5.1). This is consistent with the findings regarding the northward movement of species presented in Chapter 4, where species with larger genomes moved further northwards than those with smaller genomes (Fig. 4.8). The pattern observed in this chapter is driven by neophyte plants with larger genomes (Fig. 5.2 & 5.3), likely at an advantage from their human cultivation in gardens, high levels of garden fertiliser application as well as more widespread eutrophication across the UK (Smart *et al.,* 2003; Firbank *et al.,* 2000), as well as the effect of climate change making the UK's climate increasingly favourable to neophyte species, especially those from Mediterranean areas. The tendency of increasing species (among both natives and non-natives) towards more competitive life strategies may also be related to genome size; Chapter 2 showed that plants with larger genome sizes also tended to be those adopting a more competitive strategy (Fig. 2.11). Newbold *et al.* (2018) had previously found on a global level that species with broad distribution ranges were often positively impacted by human disturbances, while more narrow-ranged habitat specialists suffered. Typically, broad range sizes and a lack of specialisation are associated with ruderal lifestyles (Guo *et al.,* 2019) which are themselves associated with smaller genome sizes (Fig. 2.11; Suda *et al.,* 2015). Given this, my finding that increasing species in the UK are characterised by significantly larger genomes overall might be considered surprising. My putative interpretation is that the severe level and specific forms of human impact across the UK, namely high levels of eutrophication and human-driven repeated introductions of neophytes, might give competitive plants with larger genome sizes an advantage that translates into the overall increasing species trends observed here for non-native competitive species with larger genome sizes. However, it is also possible that the observed patterns are influenced mostly by a tendency of neophytes to fall into the increasing category, paired with the strong correlation between genome size and neophytic status. In this case, the expected effect of genome size may simply not hold true in this dataset.

The role of functional traits and niche preferences in predicting species trends

The vital role of functional traits in influencing species establishment, distribution and response to changes has been shown for many different groups of organisms (Newbold *et al.,* 2013; Coulthard *et al.,* 2019, Pollock, Morris & Vesk, 2012; Vesk *et al.,* 2021). Pollock, Morris & Vesk (2012); Vesk (2013) highlight that different values of individual traits can fundamentally modulate the correlations between the environment, plant success and other traits. They therefore suggest flexible modelling approaches to capture the full predictive potential of functional traits. For this chapter, I decided to use a random forest algorithm to tackle the complex relationships and modulations expected to be present within the dataset.

The advantages of this approach are manifold; random forests are well known to perform well with highly complex, non-linear tasks since the successive splits and combination of multiple trees built on different data are able to capture information regardless of the shape of the input data. Further, the algorithm is able to handle a multitude of data in very different formats and – to some extent – imbalances in the data, allowing me to incorporate Ellenberg values, traits and genome size together without elaborate transformations that might compromise interpretability. Drawbacks of random forests are their inability to extrapolate and their opacity when compared to conventional statistical models. Although the latter is less pronounced in random forests than in other machine learning algorithms, which otherwise outperform random forests with regard to prediction accuracies and flexibility (Ghannam & Techtmann, 2021), it still necessitates additional analyses (in this case a path analysis) to explain the success of individual contributions detected by the algorithm.

Although imperfect, the relatively high level of accuracy (~70%) achieved with the random forest models and especially the high success in pinpointing declining species (~77%) suggest that the random forests obtained sufficient information from the chosen traits and characteristics to reliably classify species into decreasing and increasing trend categories.

The leaf-height-seed hypothesis has been proposed as an alternative to Grime's CSR-scheme. It suggests that the easily measurable traits of specific leaf area, height at maturity and seed mass can adequately capture the ecology and overall strategy of a plant (Westoby, 1998). These functional traits were used here, with the added leaf traits of leaf area and leaf dry matter content in an attempt to capture further aspects of leaf traits potentially at play (Díaz *et al.,* 2016), to generate models of species trend. Of these traits, leaf area, specific leaf area and height were found to be important predictors of species trends, while seed mass was of much lower importance to the final models (Fig. 5.4). The lesser importance of seed mass in the models probably suggests that positive and negative developments of species frequencies in the UK do not hinge massively on their ability to disperse far distances and become established (Westoby, 1998). More important characters appear to be instead associated with the ability of species to become competitive quickly (i.e. height, leaf area and specific leaf area; Dayrell *et al.,* 2018; Westoby, 1998; and perhaps genome size too, see below). It is also likely that across the UK, there are enormous effects of human intervention, with human-mediated dispersal of species making seed mass a much less relevant trait than it may be in larger and less disturbed locations.

Carboni *et al.* (2018) found that specific leaf area and plant height were important predictors of invasion success along gradients of human disturbance and environmental parameters in France, with a non-significant role of seed mass. The same study showed – similar to the results presented above – that larger plants were more successful invasives, but also that successful invasives had higher specific leaf areas (a finding that is corroborated by Lake & Leishman, 2004), which is counter to my finding regarding the marginally bigger specific leaf area of species with decreasing trends. Carboni *et al.* (2018) do find, however, that the success of plants with smaller specific leaf areas is increased by higher levels of human disturbance offering a potential explanation for the results presented above.

Of even greater use to the random forest classification models than these functional traits are the Ellenberg values for nutrients and moisture. Ellenberg values are not used as commonly in modern ecological settings due to their subjective nature (Ellenberg *et al.,* 1991), the lack of this classification for most floras worldwide and inherent difficulties in determining Ellenberg values for new species (Chytrý *et al.,* 2018). Additionally, caution is advised with regard to the actual meaning of each indicator. Schaffers & Sýkora (2000) showed how Ellenberg values compared with measurable conditions within niches and found that while Ellenberg F correlated satisfactorily with moisture, when characterising the soil's sensitivity to drought during the driest months in particular, Ellenberg N values only correlated weakly with measurable soil parameters. Instead, a multitude of studies point towards Ellenberg N being a reflection of overall productivity rather than nutrient content (Hill & Carey, 2009; Ertsen, Alkemade & Wassen, 1998), suggesting that high values of Ellenberg N integrate a multitude of environmental factors that allow high productivity, such as soil consistency, moisture retention and pH or the presence of disturbances (Schaffers & Sýkora, 2000). The strong predictive power of the Ellenberg values within the model presented here (Fig. 5.4) highlights that despite difficulties with Ellenberg values, they do offer a highly informative description of niche preferences.

Therefore continued interest in these values, as evidenced by recent efforts to expand the system to new floras (Berg, Welk & Jäger, 2017; Chytrý *et al.,* 2018), should be encouraged.

The identity of the two most important predictors (Ellenberg N and F) in this model is relevant to the effects of eutrophication over the last century (Smart *et al.,* 2003; Firbank *et al.,* 2000) and to the expected aridification of the UK projected to result from continuing climate change (Ritchie *et al.,* 2019). Already, I find that species with decreasing trends have tendencies towards lower Ellenberg N values, suggesting a preference for niches with lower productivity, but higher Ellenberg F values, indicating a preference for moist environments. Thus it appears that eutrophication and climate change are already involved in driving species declines within the UK.

Genome size is a helpful addition when predicting species trends

The fact that genome size is expected to have a multitude of complex and sometimes opposing correlations with a multitude of traits and characters makes its incorporation into conventional models challenging, especially in the context of already noisy ecological data. The framework of the nonparametric and flexible random forest algorithm used here allowed me to explore whether genome size might be helpful in predicting plant success as hypothesised and previously hinted at by Herben *et al.* (2012). Although genome size is found by the model to be the least important predictor, its survival of the Boruta feature selection step shows that there is valuable information to be gleaned from the genome size of a species. Indeed, Herben *et al.* (2012) suggest that the effect of genome size on cell size and division rates and their physiological repercussions are difficult to capture in any trait other than genome size.

The coding of cytotypes as two or more separate 'species' with exactly the same traits in the models above (i.e. 104 out of the 960 'species' exist as duplicates or triplicates of the

98 species with cytotype diversity) means that the role of genome size was very likely downplayed. Although an increasing number of studies aims to characterise morphological and distributional differences between cytotypes (e.g. Halverson *et al.,* 2008; Richardson & Hanks, 2011; Pegoraro *et al.,* 2016; reviewed by Kolář *et al.,* 2017), trait information on separate cytotypes of the same species is often not available. However, different cytotypes of the same species are likely to display different trait values and distributions, indeed often competitively excluding each other when competing in the field (Collins, Naderi & Mueller-Schaerer, 2011; Laport *et al.,* 2013; Walczyk & Hersch-Green, 2019; Pegoraro *et al.,* 2019). It is therefore not unreasonable to assume that – had detailed trait information been available on separate cytotypes – genome size might have emerged as an even more influential predictor.

Additionally, the findings presented in Chapter 4 point towards a spatially heterogeneous distribution pattern of genome size across the UK, potentially influenced in part by nutrient availability, with limitations due to a lack of nutrients lifted in some areas, leading to the higher prevalence of larger genomes in such areas. While the outcome variable of decreasing and increasing species trends here looks at the flora as a whole and does not take into account these spatial differences, incorporating spatial patterns into future models may also elevate the importance of genome size as a predictor of species success in some areas more than in others.

The Boruta feature selection step led to the exclusion of ploidy and chromosome number in random forest runs. Pandit *et al.* (2014) had previously found that the inclusion of genome size and chromosome number together in models aiming to predict the likelihood of a species becoming invasive increased the overall explanatory power of each, but also that genome size and ploidy had contradicting effects. They hypothesised that with increasing ploidy level the effect of genome size would change. The fact that our model did not consider chromosome number or ploidy level as being helpful in its predictions is therefore surprising. It is possible that the signal one might expect from the hybrid vigour associated with polyploidy (Soltis & Soltis, 2000; Birchler, 2015; Te Beest *et al.,* 2012) is obscured by the use of duplicated cytotype data in this analysis. Therefore, further analyses along the same lines as presented here would benefit from the investigation of trait variation between cytotypes. Potentially, however, the predictive capacity of ploidy is simply too limited in this context to emerge – especially compared with the contributions of Ellenberg values and functional traits – even if the data availability allowed the integration of cytotype variation.

The fact that the random forest analysis was almost entirely restricted to native species due to data availability means that the predictive potential of genome size among non-native species could not be sufficiently explored here. Differences in the way that genome size impacts species trends for established species and new arrivals may well be substantial and should be considered by future research, as data availability for non-native species increases. Additionally, any insights into the causalities behind the tendency of neophytes to have larger genomes in the UK would be valuable for further assessments of the helpfulness of genome size in capturing predictive information.

Genome size is likely to be indirectly linked with trend

A phylogenetic path analysis helped me gain insights into potential causality structures among the predictors used in the random forest models and how they exert their effect on trend. The 16 hypothetical models put forward three alternative potential causal flows; firstly independence of effects of the three groups of predictors (genome size, functional traits and Ellenberg values), secondly a setting where traits influence niche requirements which in turn act directly on trend, and thirdly the reverse setting where belonging within a specific realised niche leads to different trait values and then to trends. Within those three causal flows, the position of genome size was shifted to determine which causality structure for the integration of genome size was most consistent with the data (visualised in Fig. S5.1). The structure most supported (Fig. 5.5) was consistent with traits conditioning niche requirements (i.e. Ellenberg values) and genome size exerting its role as an exogenous variable with direct links with all three functional traits, but indirect effects on Ellenberg values and trend. The fact that all models, including the most supported model presented here, had low p-values suggested that further links and nodes are required to more accurately represent the causality structure within the network. Importantly, the 16 competing hypothetical structures that were preformulated here are themselves based on assumptions made about likely causality structures – with a special focus on elucidating the position of genome size – leading to an inherent level of subjectivity in the path analysis. In the most supported model genome size has positive links with seed mass and height, and a negative effect on specific leaf area. While the finding of a positive correlation between genome size and seed mass agrees with several previous smaller studies reviewed by Knight, Molinari & Petrov (2005) and the larger study of Beaulieu *et al.* (2007), genome size was previously shown to correlate negatively with plant height (Knight & Beaulieu, 2008), although the effect was not significant under phylogenetic correction. This latter trend is likely driven by trees, which typically have smaller genomes (Knight & Beaulieu, 2008) and the opposing trend, concurrent with my findings, was shown in graminoid (Rios, Kenworthy & Munoz, 2015) and, including a phylogenetic correction, in herbaceous plants (Herben *et al.,* 2012), which are also studied here. The negative correlation of genome size with specific leaf area is also supported by findings from Herben *et al.* (2012), but does not appear to be universally applicable across groups of plants (Kang *et al.,* 2014). Lower specific leaf areas are typically found in highly nutrient and water use efficient species with lower metabolic rates, suggestive of lower competitive performance, while greater plant height is a hallmark of greater dominance and competitive success (Violle *et* *al.,* 2009; Carboni *et al.,* 2018; Peng *et al.,* 2022), hinting once again at the complex entanglement of genome size, its immediate physiological correlations and the consequential interactions between the traits influenced and their ecological ramifications.

The directed acyclic graph (Fig. 5.5) most supported by the underlying data makes clear that the strongest associations of larger genome size eventually translate into a tendency towards higher Ellenberg N values, and consequently into a preference for productive environments. The suggested role of moisture in defining the productivity described by Ellenberg N (Schaffers & Sýkora, 2000) led to the assumption of an additional unidirectional pathway from Ellenberg F (Streiner, 2005), the second most important predictor in the random forest presented above, to Ellenberg N, the most important predictor. Contrary to the high importance of Ellenberg F postulated by the random forest, the phylogenetic path analysis suggests that the direct link between Ellenberg F and trend is actually not significant and instead suggests that moisture requirements might influence their role indirectly via Ellenberg N. Far from undermining the role of moisture requirements for influencing species trends (and hence in part the role of climate change), this finding simply indicates that there are likely additional links missing from the model that, when added, would show a closer approximation of the true pathway via which moisture requirements might impact trend.

The findings are consistent with an indirect role of genome size on species trend, mediated through the sometimes opposing and usually complex links that genome size has with a multitude of plant traits. The traits chosen to present in this context are far from the only ones that genome size has been shown to influence (Van't Hof & Sparrow, 1963; Francis, Davies & Barlow, 2008; Beaulieu *et al.,* 2008; Šímová & Herben, 2012; Roddy *et al.,* 2020; Bennett, 1971 & 1972), but the inclusion of further traits would have created difficulties in the interpretation of the analysis due to complexity. An analysis integrating other processes likely impacted by genome size, such as photosynthetic rates or water use efficiency (Roddy *et al.,* 2020; Beaulieu *et al.,* 2008) may well offer further insights into the links of genome size with species trend and overall success, and might help further in elucidating the link with moisture requirements in particular. Sadly, the limited availability of such data for UK species means that opportunities to extend these studies are currently constrained. What is clear from this analysis and falls in with the findings of previous field experiments (Guignard *et al.,* 2016; Šmarda *et al.,* 2013; Peng *et al.,* 2022) as well as with Chapter 4 is the fact that genome size appears to correlate with various aspects of plant ecology in a fundamental way and may do so, at least in part, via an association with nutrient availability in the environment.

Whether the role of genome size differs from what is presented here and offers the same level of predictive power in a global context should be a focus of future research, offering the potential to improve predictive models of species success in areas where species records are sparser and such methods could be crucial for impactful conservation approaches.

Chapter 6 General discussion

Summary of findings

This thesis aims to elucidate the role that genome size might play at landscape scales, and determine the extent to which it can be found to correlate with distribution, movement and success of plants in response to the environment. To facilitate this approach, I created a comprehensive repository of the vascular plants of Britain and Ireland, both native and non-native (see glossary; Table 2.1), bringing together the vast knowledge of Britain and Ireland's expert botanists. Use of this repository, in conjunction with the UK's unique recording history, allowed me to characterise dynamic changes within the flora, and explore the patterns and influence of genome size.

The three major findings with regard to genome size can be summarised as follows: (i) Distinct spatial distributions of genome size and ploidy (as weighted mean per hectad) are visible, with genome size distribution in particular showing a pattern that coincides with land cover types most impacted by humans, suggesting nutrient pollution as a potential influence. (ii) Plants with larger genomes are also shown to move further distances along the north-south axis of the UK as species ranges respond to anthropogenic pressures over time. (iii) Finally, this thesis trials the use of genome size as a predictive variable in random forest models to identify species with decreasing species trends across the UK and points towards an indirect role of genome size that adds value to trait-based models of species success.

This thesis shows that more than half of the British and Irish flora is currently made up of non-native species. It is also increasingly impacted by climate change; pervasive movements towards the north, especially of neophyte species, and a dominance of Mediterranean introductions amongst species with increasing species trends are reported here and illustrate the growing impact of temperature rises and decreases in rainfall.

Tying in with previous research from field experiments, these novel findings show that knowledge of genome size can help us to identify priorities for conservation, a finding of particular importance against the backdrop of mounting human pressures on ecosystems worldwide. To my knowledge, this thesis represents the first holistic, flora-wide assessment of the role that genome size might play in influencing plant distributions, going beyond individual correlation studies and tightly controlled experimental settings, as called for by Knight, Molinari & Petrov (2005).

A flora in *flux*

Despite its impressive history of organised botanical recording dating back to the late 17th century (Ray, 1690), comprehensive studies of the flora have long been held back by a lack of consistency between the multitude of repositories, checklists and distribution records available for the British and Irish flora, particularly with regard to taxonomy and the treatment of plants with different status (native, non-native). The first major output of this thesis, a comprehensive and taxonomically harmonised repository of all vascular plants within the flora, developed in collaboration with leading experts, is intended as a starting point for a more integrative and holistic approach to characterising the flora and changes occurring within it (Henniges *et al.,* 2021 & 2022).

In this thesis, I demonstrate that the British and Irish flora is undergoing dynamic changes. Following Stace (2019), the data includes all species that 'the plant-hunter might reasonably be able to find […] in any one year' and thus offers a comprehensive look at species currently characterising ecosystems in the study area. Out of these species, more than half are non-native, suggesting that the inclusion of such species is crucial should one wish to characterise the true species composition present within the area; a focus on native species only, often for good reasons (such as a lack of trust in the reliability of occurrence records for the non-native flora, e.g. Groom, 2013b), will therefore necessarily miss a big part of the picture. Pyšek *et al.* (2004) lament the incongruous treatment of non-natives (invasives in particular), their assumed status (neophyte, archaeophyte; see glossary; Table 2.1) and naturalisation level (naturalised, casual etc.) in many floras and highlight the value of flora-wide analyses that incorporate non-natives.

The analyses presented in this thesis (Chapter 4, Chapter 5) clearly highlight the strong and often contradicting effects of non-native species within this dataset, even though particular care was taken to incorporate them. On the one hand, both chapters highlight how differently natives and non-natives appear to have developed in the UK in the past three decades; the northward movement of non-natives is far more pronounced than that of natives (Fig. 4.7 & Fig. 4.8) and non-natives are also more prevalent among species with increasing species trends (i.e. species with positive slopes on regressions of their relative frequency over time, Fig. 5.1, Fig. 5.2 a&b). Further, tendencies in adapted life strategies differ between successful (i.e. increasing species trend) natives and non-natives, with a tendency towards stress tolerance in natives but not in non-natives (Fig. 5.3 b&c). On the other hand, due to the historic under-representation of non-native species in local floras, data such as the Ellenberg indicator values were not available for non-native species, meaning that not all analytic avenues allowed the integration of both native and non-native species (such as the random forest modelling in Chapter 5). The importance of including all components of a flora, irrespective of status, when generating new datasets must therefore be stressed to avoid such bottlenecks and limitations in the future.

The impact of humans on compositional changes in the flora of a small, industrialised nation such as the UK is considerable (Lim *et al.,* 2014). Introductions of ornamental garden plants in particular and their naturalisation beyond the garden are not often considered by ecological studies (Pergl *et al.,* 2016), but have tremendous impact on possible biases in analyses such as the ones presented in this thesis (as highlighted in

Chapter 3) and are also likely a strong influence on the compositional changes observed here (Chapters 4 & 5). Their full inclusion in floras, species checklists and flora-wide analysis is therefore required to fully characterise changes and threats therein.

Beyond the success of non-native species, changes in the flora are also visible when considering the biome types that its species are associated with. Those taxa with increasing species trends were more often linked with Mediterranean and Southern Temperate biomes, while I found that cold-adapted species with associations to an Arctic Montane biome were overwhelmingly among those species with decreasing trends (Fig. 5.2 c&d). Species adapted to specific biomes show clear patterns of occurrence across the UK, with Arctic Montane species typically centring their occurrence in the mountainous, northern parts of Scotland (Fig. 6.1). Thus existing at the northern boundary of the UK, and with further retreat options limited at high altitudes of the Highlands, such species are at severe risk as climate change alters their habitat and allows species from the South to infringe on it.

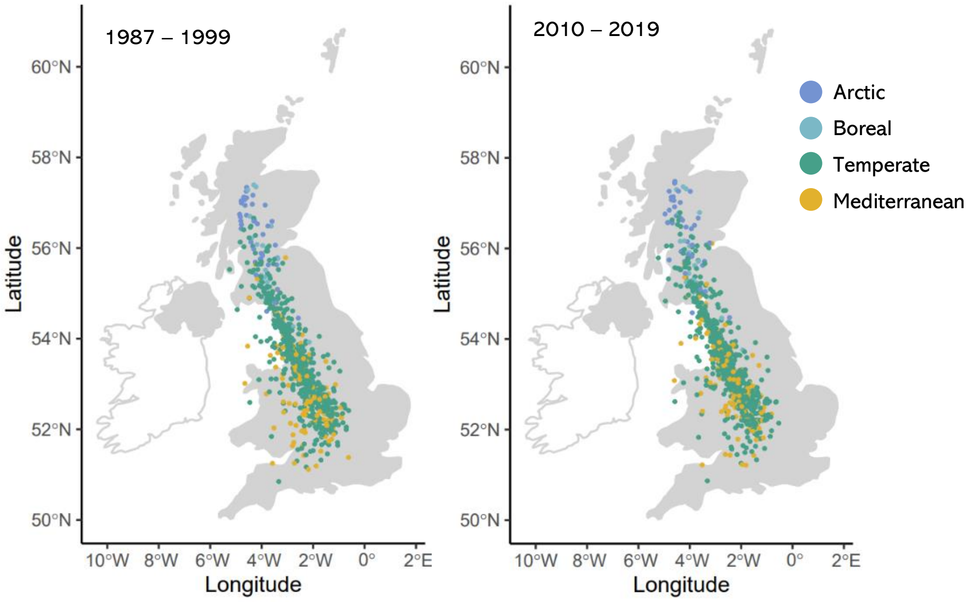


Fig. 6.1 Distribution centroid and biome associations for 964 species. a & **b** show the location of centre of masses in the first date class (1987-1999) and the last date class (2010-2019), respectively. Colours indicate if the centroid belongs to a species associated with an Arctic, Boreal, Temperate or Mediterranean biome (see legend). The centre of Arctic and Boreal species' distributions is typically located in the far North, Temperate species centre their mass throughout the length of the UK and Mediterranean species tend to centre in the South. All species show subtle advances northward over the past three decades.

These differences in success of plants from typically hot and typically cold biomes (Fig. 5.2c & d), paired with the demonstrated further movement of neophyte species towards the North (Fig. 4.7 & 4.8) suggests that impacts of climate change are already affecting the flora's composition, with further and more severe changes to be expected in the near future (Richie *et al.,* 2019). Emerging from this work and in agreement with previous studies by Hill & Preston (2015) as well as Pearce-Higgins *et al.* (2015 & 2017), cold-adapted species and those occurring in upland habitats are at particular risk from current and future impacts of climate change.

Is a large genome always a burden? – a question of circumstances

This thesis presents evidence that mean genome sizes (at hectad scale) have steadily increased in nearly all areas of the UK in the past thirty years (Fig. 4.3). This finding is surprising given the fact that large genomes have often been associated with physiological disadvantages; large genomes necessitate slower growth cycles (Bennett, 1971; Bennett, 1972; Bennett, 1987), are more restricting with regard to the levels of soil nutrients and pollution plants with such genomes can endure and thrive in (e.g. Guignard *et al.,* 2016; Vidic *et al.,* 2009; Sparrow & Miksche, 1961) and have been suggested to impact negatively water use efficiency and photosynthetic rates (Faizullah *et al.,* 2021). All this suggests that having a larger genome constrains the trait space and adaptation flexibility of plants (Faizullah *et al.,* 2021). Concurrent with these physiological drawbacks, larger genomes have been shown to be at greater risk of extinction globally (Vinogradov, 2003) and are rarely found to be invasive plants (Suda *et al.,* 2015).

Given these findings and notions, large genomes are typically thought of as burdensome ('large genome constraint hypothesis', Knight, Molinari & Petrov, 2005), causing concerns for plants that maintain large genomes, often with the help of specific adaptations (Veselý, Bureš & Šmarda, 2013), in the face of changing environments. In this thesis, I find that large genomes may not always be as restricting as they are often thought to be. Beyond the steady increase in weighted mean genome size across the past thirty years, I also find that plants with increasing trends have larger genomes than those with decreasing trends (Fig. 5.1) and poleward movement is more pronounced in plants with larger genomes (Fig. 4.8). While this could be an artifact caused by the success of introduced species that happen to have larger genomes rather than a direct consequence of large genomes, it appears as though the specific context of the highly disturbed UK might compensate for some of the typical drawbacks associated with large genomes. For example, ongoing eutrophication may lead to a similar release from nutrient limitation and consequent dominance as has previously been observed in field experiments (Guignard *et al.,* 2016; Šmarda *et al.,* 2013; Peng *et al.,* 2022), and human-driven introductions of species in new environments via gardens and agriculture as well as the disturbance of intact ecosystems may make large propagule sizes and longer generational times less prohibitive for successful dispersal and establishment.

On the other hand, the fact that plants with larger genomes move further distances might also indicate that they are more successful in colder regions where their physiology may be better adapted, necessitating such movement in the first place. Plants with the largest genomes are often highly specialised, having adaptations that allow them to be competitive under limiting environmental nutrient availability, such as the storage tissues of geophytes (Veselý, Bureš & Šmarda, 2013). Indeed, after the parasitic plant *Viscum album* L. (88.9 pg/1C) with by far the largest genome in the study area, the three plants with the next largest genomes, *Tulipa sylvestris* L. (58.0 pg/1C), *Fritillaria meleagris* L. (47.3 pg/1C) and *Paris quadrifolia* L. (44.2 pg/1C) all have storage tissues typical for geophytic lifestyles. This particular adaptation affords plants with large genomes the ability to store nutrients slowly accumulated and to pre-divide cells prior to the next growing season, when hydraulic expansion, driven by water accumulation in vacuoles, causes rapid growth (Bennett, 1972). This strategy is not disadvantaged by cold temperatures, unlike growth by cell division. Thus with a warming climate, the advantage of growth by cell division that underpins growth in species with small genomes provides an ever increasing competitive advantage, potentially pushing species with large genomes northward. There is also evidence of higher frost resistance in species with larger genomes (MacGillivray & Grimes, 1995) – an advantage that would become increasingly irrelevant under increasing temperatures, which could eventually impact species with large genomes negatively, as average temperatures across the UK continue to increase.

Since the Frescalo-derived species data operate at hectad scale, the analyses presented in this thesis do not allow for insights into changes in community composition. Further research is needed to elucidate if the positive species trends and further northward shifts associated with larger genomes translate into greater dominance of the species in the field or instead simply means plants with larger genomes shift their range northward, where they are most competitive.

From individual correlations to landscapes

Research on genome size has long been dominated by individual correlation studies (e.g. as reviewed in Knight, Molinari & Petrov, 2005) and has recently benefited from insights derived from controlled field experiments (e.g. Guignard *et al.,* 2016). Given the multitude of suspected effects of genome size at all levels of plant physiology, ecology and evolution, the notion that it might improve our understanding of plant biogeography suggests itself.

Increasingly, studies of genetic characters exist at global scales (Rice *et al.,* 2019 and Bureš *et al.,* 2022 (in press)) and highlight gradients of both ploidy and genome size across biomes. Meanwhile the intermediate step – between field experiments and global patterns – promises to offer insights into the local drivers that interact with genome size in shaping plant distributions. To my knowledge, this is the first attempt to consider in detail both the external correlates of genome size patterns as well as the predictive power of genome size on species trends to gain a greater understanding of the role it might play in shaping the distribution dynamics within a national flora.

This thesis demonstrates distinctive spatial patterns of genome size across the UK that can be linked with human presence (Chapter 4); areas close to human settlements and land used for agriculture are characterised by larger mean genome size (considered at hectad scale, Fig. 4.5). The analysis presented here could not conclusively identify either atmospheric nitrogen deposition (in part due to contrary findings for wet and dry deposition types) or climatic factors (temperature, rainfall) as major correlates behind these patterns, but the location of genome size hotspots (Fig. 4.2b) suggests that human actions are a major factor. Whether this is due to repeated introduction and spread of plants with larger genomes, due to nutrient pollution that lifts nutritional constraints on genome size or some other underlying effect remains to be confirmed in future research. Particularly the role of joint nitrogen and phosphorus availability could not be fully elucidated here. Previous findings from field experiments (Guignard *et al.,* 2016; Berendse *et al.,* 2021) have found that atmospheric nitrogen deposition alone does not have the same levels of impact on species composition (and the mediation via genome size) as the effect of an industrial fertiliser containing nitrogen and phosphorus jointly. Different locations worldwide have differing levels of nutrient limitation (Du *et al.,* 2020), with the far North more impacted by nitrogen limitation and the tropics more by phosphorus limitation. Anthropogenic nitrogen deposition appears to be shifting the limitation away from nitrogen and towards phosphorus (Peñuelas *et al.,* 2013; Crowley *et al.,* 2012), potentially explaining why nitrogen deposition alone did not help to explain the patterns and change in genome sizes across the UK conclusively.

Regardless of the exact pathway by which human action shapes the landscape of genome size hotspots in the UK, it is clear that at least in this case study, the presence of humans appears to influence the genome size profiles of the flora around them. Genome size information can thus be considered a helpful link in understanding how human-made environmental disturbances will impact species compositions.

Genome size can help inform conservation efforts

Beyond characterising patterns of genome size at landscape scales, this thesis furthers the concept that the limitations inherent in genomes of different sizes could influence which environmental conditions a plant can occur in and how it responds to changing environments. I demonstrate that knowledge of genome size can help predict species success (as expressed in the form of changes in relative frequencies of species) in the UK (see Chapter 5), suggesting its value in informing targeted conservation efforts.

Regardless of the external influences that condition how genome size might exert its effects, a growing body of literature, including the work presented in this thesis, suggests a role for genome size in predicting ecological trajectories of plant species. Herben *et al.* (2012) concluded that information on cell size and division rates contained within genome size data offered helpful data for models of regional species abundances, while Schmidt & Drake (2011) and Pandit, White & Pocock (2014) successfully used genetic characters (genome size and chromosome number) in predicting invasive success.

The role of genome size in predisposing species to become invasive has received some attention over the years (reviewed in Suda *et al.,* 2015). Species with small genomes are considered to have higher potential to become invasive due to their ability to have short generation times, small propagule size and high resource efficiency, among others (Pandit, White & Pocock, 2014; Suda *et al.,* 2015). Indeed, the species most frequently reported as invasives in the UK with high perceived impact scores according to Dehnen-Schmutz *et al.* (2022) all have small genomes; these species are *Impatiens glandulifera* Royle (0.83 pg/1C) *Reynoutria japonica* Houtt. (0.78 pg/1C), *Rhododendron ponticum* L. (0.81 pg/1C), *Crassula helmsii* (Kirk) Cockayne (0.32 pg/1C) and *Heracleum mantegazzianum* Sommier & Levier (1.83 pg/1C).

The fact that many non-natives (especially naturalised neophytes) are showing positive trends in their relative frequencies (Chapter 5) and underwent greater northward movement in the study area than native species (Chapter 4) chimes with the definition of invasive species, i.e. that they are non-natives that have become naturalised, forming self-replacing populations and that they have the potential to spread over long distances and cause harm to native species and environments (Richardson *et al.,* 2000). However, the finding that successful and far-moving species identified here tend to have larger genomes suggests that it may not be invasives *sensu strictu* that are affecting the patterns observed. Species with larger genomes have increasingly strong tendencies towards a competitive life strategy (Fig. 2.11), which is also associated with increasing species frequency trends in both native and non-native species (Fig. 5.3) and is the same group that Guignard *et al.* (2016) found to respond most positively to the combined application of nitrogen and phosphorus in field trials.

While species with a small genome size may be predisposed to exhibit invasive tendencies such as efficient spread and rapid growth (Suda *et al.,* 2015), those with a large genome size might have a competitive advantage in highly disturbed areas where – given the right environmental conditions – they might become dominant, negatively impacting native biodiversity. Consequently, it appears that different environmental contexts may well favour plants with different genome sizes. As outlined above, species with large genomes are not always at risk of extinction, but can – under the right environmental conditions – become a potential threat to the biodiversity around them. Knowledge of the genetic make-up of plants (e.g. genome size and whether or not they are polyploid), combined with information about local conditions in the environment, thus promises to aid in the identification of species at risk of invasive or overly competitive behaviour, and conversely, extinction.

The rapid changes of the Anthropocene, from climate change to pollution and habitat conversion, are endangering biodiversity across the globe (Tilman *et al.,* 2017). Improving our understanding of the factors that modulate the way in which different species respond to those changes is crucial to better use limited conservation resources and to develop biodiversity bonds, which will become increasingly important in the context of a growing appreciation that monetising biodiversity will drive its stewardship (Dasgupta, 2021). Work in this thesis demonstrates that genome size might constitute a useful and easily obtained source of information that is currently omitted from trait-based assessments of species at risk from declines and extinction (e.g. Pollock, Morris & Vesk, 2012; Vesk *et al.,* 2021).

Bellard, Marino & Courchamp (2022) highlight that while different sources of anthropogenic environmental threats are often ranked by their expected level of impact for different regions, each ecosystem will be faced with a set of diverse threats that can impact species assemblages in a variety of ways. Agricultural habitat conversion, invasive species and pollution have previously been suggested to be of particular importance for future species extinctions in Europe (Harfoot *et al.,* 2021), but Ritchie *et al.* (2019) also points to the impacts of climate change, especially with regard to changes in growth conditions for the vegetation in Britain, as a major threat. As outlined in Chapters 1, 4 and 5 in more detail, genome size has been hypothesised to be linked with water use efficiencies by constraining cell size ranges (Faizullah *et al.,* 2021), suggesting that knowledge of genome size can help identify species at risk from changes in climatic conditions. Meanwhile the heightened sensitivity of plants with large genomes to pollution (Vidic *et al.,* 2009; Temsch *et al.,* 2010; Sparrow & Miksche, 1961; Einset & Collins, 2018), including increased biomass accumulation with high levels of nutrients (Guignard *et al.,* 2016; Šmarda *et al.,* 2013; Peng *et al.,* 2022) could lead to different responses of plants, mediated in part by genome size, to different kinds of pollution.

The correlation of land use in the UK with spatial genome size patterns (Fig. 4.5) is also of importance here. Leclère *et al.* (2020) analysed a set of scenarios regarding the future of land use globally and stressed the importance of land use change in the current biodiversity crisis, a sentiment shared by Lanz, Dietz & Swanson (2018) and Elmqvist, Zipperer & Güneralp (2015) in considering the detrimental role of agricultural and urban expansion on biodiversity. My analyses suggest that knowledge of genome size might help us predict which species could profit from unmitigated habitat conversions and which might suffer, allowing a better understanding of what future communities under different scenarios will look like.

Avenues for future research

Further research on the British and Irish flora

The coarse resolution of distribution records at hectad and date class-scales used here allowed me to quantify change over relatively long periods, but also necessarily poses difficulties. Particularly the wide variety of habitats amalgamated into the synthetic unit of hectads constrains the findings to a top-level view of dynamics within the flora. While it is unlikely that the inherent difficulties of variable sampling biases and lacking comparability between modern and historic data will allow long term analyses of higher resolution species distribution trends with the methods currently available (see e.g. Pescott, Humphrey & Walker, 2018; Pescott *et al.,* 2019a), emerging datasets such as the one generated by the National Plant Monitoring Scheme (Pescott *et al.,* 2019b&c) may allow for more detailed insights into the role of genome size in shaping natural plant assemblies. Such datasets are also available for grassland settings in France, China and the United States, making comparisons between floras feasible in the future (Violle *et al.,* 2015; Li *et al.,* 2015; Pearson *et al.,* 2016).

Helpful information on the effect of nitrogen and phosphorus may become available from DEFRA's 'June survey of agriculture and horticulture' datasets (https://www.gov.uk/agricultural-survey) or from the comprehensive models of environmental change carried out by the UK-SCAPE project (https://uk-scape.ceh.ac.uk/).

An exciting new avenue for spatially explicit trait-based distribution models lies in Hierarchical Modelling of Species Communities (HMSC, Ovaskainen *et al.,* 2017; Tikhonov *et al.,* 2020), which integrate traits, distribution and environmental information as well as phylogenetic data to predict local species success and biodiversity development in response to environmental change. Use of the data generated for this thesis and especially the exploration of genome size as an addition to such models would provide greater resolution and hence allow a more differentiated understanding of the potential drivers behind distributional changes and biodiversity patterns across the UK.

Does the role of genome size differ across the globe?

While patterns of genome size in the UK appear to be particularly correlated with human presence, characteristics such as soil types, climate and competition can be expected to be even more relevant elsewhere. This thesis highlights the value of genome size in predicting species trends and occurrence patterns. While such trends can be quantified with relative ease and confidence in the well-recorded UK (if existing sampling biases are appropriately accounted for), many locations around the globe and in particular some biodiversity hotspots did not benefit from the same levels of historical biological recording (Meyer, Weigelt & Kreft, 2016; Paton *et al.,* 2020). The potential of trait-based predictive models incorporating genome size will be of particular value in such under-recorded areas, provided the predictive power of genome size is found to apply in such different environments and the mechanisms by which genome size exerts its role in different contexts is taken into consideration (Powney *et al.,* 2014b). Consequently, it will be important to expand this analysis to different floras, to find out just how context-dependent the role of genome size might be in shaping plant distributions and determining local species success.

An analysis of the role of genetic characters in countries with lower levels of human impact and greater diversity with regard to climatic conditions and soil types is required to understand the success of plants with large genomes. Potentially the results here are driven by the combination of a mild climate and high levels of human impacts that is found in the UK, even though the effects of climate change are already emerging. In areas where water is more limiting and the driving force of artificial nutrient pollution and repeated species introductions less pronounced, subtle changes in climate might turn out to impact the success of plants with large genomes to a greater or lesser extent than could be observed in this work. This is especially true since nutrient availability to plants is impaired by drought conditions (Sardans & Peñuelas, 2012), which would place plants with large genomes at a potential additional disadvantage.

Moving the analyses presented here to different floras and a global viewpoint promises to further our understanding of the role that ploidy, in addition to genome size, might have to play. While the patterns in ploidy across the UK mirrored previous findings by Rice *et* *al.* (2019), neither ploidy nor chromosome number were identified as informative trait in random forest models of species trend. Since spatial patterns of polyploid frequencies were found to be mostly predicted by temperature (Rice *et al.,* 2019), but temperature ranges across the UK are comparatively limited, with temperature fluctuations being still ameliorated by the effects of the gulf stream, it is possible that the predictive potential of ploidy might only emerge in spatial contexts with greater climatic variability. Indeed, Pandit, White & Pocock (2014) found that when modelling invasiveness, both chromosome number/ploidy and genome size contributed important information. Whilst such genetic characters are to some degree correlated, they still may individually have a role to play in shaping plant communities and responses to the environment and do need to be considered together in future modelling approaches even if the significance of each genetic character is likely context-dependent.

Concluding remarks

I embarked on this project attempting to determine if the previously established correlations of genome size with many aspects of physiology might translate into a role of genome size in shaping species distribution and success in the spatial context of Britain's landscapes. I demonstrate that genome size can be a helpful addition to models of species success and movement, adding, on top of its entanglements with many functional traits, information that may not be available if genome size is omitted from such models. In the context of the intensive human footprint across the UK, disadvantages of large genomes appear to be compensated, leading to overall increases in mean genome size across the past decades. While further research is needed to explore in what ways genome size exerts its ecological effects and how the role of genome size varies across different environmental contexts, this thesis makes clear that genome size should have a role to play in scientific efforts that aim to understand and predict the response of individual species and species assemblages to the unprecedented changes of the Anthropocene.

References

Alexander, J.M., Kueffer, C., Daehler, C.C., Edwards, P.J., Pauchard, A., Seipel, T. and Miren Consortium, 2011. Assembly of nonnative floras along elevational gradients explained by directional ecological filtering. *Proceedings of the National Academy of Sciences*, **108**(2), pp.656-661.

Amphlett, A., 2015. Using the BSBI Distribution Database. https://database.bsbi.org/static/Using_the_BSBI_Distribution_Database.pdf

Andermann, T., Antonelli, A., Barrett, R.L. and Silvestro, D., 2022. Estimating alpha, beta, and gamma diversity through deep learning. *Frontiers in Plant Science,* **13,** pp.839407.

Anselin, L., Bera, A.K., Florax, R. and Yoon, M.J., 1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, **26**(1), pp.77-104.

Anselin, L., Gallo, J.L. and Jayet, H., 2008. Spatial panel econometrics. In *The Econometrics of Panel Data*, pp. 625-660. Springer, Berlin, Heidelberg.

Applebaum, S., 1958. Agriculture in Roman Britain. *The Agricultural History Review*, **6**(2), pp.66-86.

Asadi-Samani, M., Bahmani, M. and Rafieian-Kopaei, M., 2014. The chemical composition, botanical characteristic and biological activities of *Borago officinalis*: a review. *Asian Pacific Journal of Tropical Medicine*, **7**, pp.22-28.

August, T., Powney, G., Harrower, C., Hill, M. and Isaac, N.J.B., 2015. sparta: Trend analysis for unstructured data (v0.1.30). [R package].

Auret, L. and Aldrich, C., 2012. Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, **35**, pp.27-42.

Beaulieu, J.M., Leitch, I.J., Patel, S., Pendharkar, A. and Knight, C.A., 2008. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist*, **179**(4), pp.975-986.

Beaulieu, J.M., Moles, A.T., Leitch, I.J., Bennett, M.D., Dickie, J.B. and Knight, C.A., 2007. Correlated evolution of genome size and seed mass. *New Phytologist*, **173**(2), pp.422.

Bellard, C., Marino, C. and Courchamp, F., 2022. Ranking threats to biodiversity and why it doesn't matter. *Nature Communications*, **13**(1), pp.1-4.

Bennett, M.D., 1971. The duration of meiosis. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, **178**(1052), pp.277-299.

Bennett, M.D., 1972. Nuclear DNA content and minimum generation time in herbaceous plants. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, **181**(1063), pp.109-135.

Bennett, M.D., 1987. Variation in genomic form in plants and its ecological implications. *New Phytologist*, **106**, pp.177-200.

Bennett, M.D. and Leitch, I.J., 2001. Plant DNA C-values Database (Release 1.0) https://cvalues.science.kew.org/

Bennett, M.D. and Leitch, I.J., 2005. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Annals of Botany*, **95**(1), pp.45-90.

Bennett, M. D., and Smith, J. B., 1976. Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences*, **274**(933), pp.227-274.

Berendse, F., Geerts, R.H., Elberse, W.T., Bezemer, T.M., Goedhart, P.W., Xue, W., Noordijk, E., ter Braak, C.J. and Korevaar, H., 2021. A matter of time: Recovery of plant species diversity in wild plant communities at declining nitrogen deposition. *Diversity and Distributions*, **27**(7), pp.1180-1193.

Berg, C., Welk, E. and Jäger, E.J., 2017. Revising Ellenberg's indicator values for continentality based on global vascular plant species distribution. *Applied Vegetation Science*, **20**(3), pp.482-493.

Biau, G. and Scornet, E., 2016. A random forest guided tour. *Test*, **25**(2), pp.197-227.

Biological Records Centre, 2021. Sparta: Trend Analysis for Unstructured Data (v0.2.19). [R package]. https://github.com/BiologicalRecordsCentre/sparta/blob/master/R/frescalo.R

Birchler, J.A., 2015. Heterosis: The genetic basis of hybrid vigour. *Nature Plants*, **1**(3), pp.1-2.

Bivand, R. *et al.*, 2015. Package 'spdep'. *The Comprehensive R Archive Network*. [R package].

Bivand, R. *et al.*, 2022. Package 'maptools'. [R package].

Bivand, R. and Piras, G., 2019. spatialreg: Spatial Regression Analysis. R package version 1.1-8. [R package].

Borges, R., Machado, J.P., Gomes, C., Rocha, A.P. and Antunes, A., 2019. Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics*, **35**(11), pp.1862-1869.

Botanical Society of Britain & Ireland. https://bsbi.org/maps-and-data [accessed October 2022]

Botanical Society of Britain and Ireland, 2021. GB Red List of Vascular Plants. https://bsbi.org/taxon-lists [accessed January 2022]

Braithwaite, M.E., 2010. How well has BSBI chronicled the spread of neophytes? *Watsonia*, **28**(1), pp.21-31.

BRC Vascular Plant Database (VPDb). CEH Wallingford.

Breiman, L., 2001. Random forests. *Machine Learning*, **45**(1), pp.5-32.

Brummitt R.K., 2001. *World Geographical Scheme for Recording Plant Distributions*, Edition 2. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards [accessed September 2022]

Brunson, J.C., 2018. Package 'ggalluvial'. *The Comprehensive R Archive Network (CRAN)*. [R package]

BSBI Distribution Database. CEH Wallingford. https://database.bsbi.org/ [accessed October 2019]

BSBI database search facility (taxon, literature and cytology database). https://websites.rbge.org.uk/BSBI/intro.php [accessed January 2022]

Buggs, R.J.A., 2021. The origin of Darwin's "abominable mystery". *American Journal of Botany*, **108**(1), pp.22-36.

Bureš, P., Elliott, T.L., Veselý, P., Šmarda, P., Forest, F., Leitch, I.J., Nic Lughadha, E., Soto Gomez, M., Pironon, S., Brown, M.J.M., Šmerda, J., Zedek, F. The global biogeography of angiosperm genome size is shaped by climate and range size. Submitted to *New Phytologist* (Dec 2022).

Bureš, P., Wang, Y.F., Horová, L. and Suda, J., 2004. Genome size variation in Central European species of Cirsium (Compositae) and their natural hybrids. *Annals of Botany*, **94**(3), pp.353-363.

Burgman, M.A. and Fox, J.C., 2003. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. In *Animal Conservation Forum* **6**, pp. 19-28. Cambridge University Press.

© Dudley Stamp, L./Geographical Publications Ltd, Audrey N. Clark, Environment Agency/DEFRA and Great Britain Historical GIS

Carboni, M., Calderon-Sanou, I., Pollock, L., Violle, C., DivGrass Consortium and Thuiller, W., 2018. Functional traits modulate the response of alien plants along abiotic and biotic gradients. *Global Ecology and Biogeography*, **27**(10), pp.1173-1185.

Carboni, M., Münkemüller, T., Lavergne, S., Choler, P., Borgy, B., Violle, C., Essl, F., Roquet, C., Munoz, F., DivGrass Consortium and Thuiller, W., 2016. What it takes to invade grassland ecosystems: Traits, introduction history and filtering processes. *Ecology Letters*, **19**(3), pp.219-229.

Carpenter, S.R. *et al.*, 2009. Science for managing ecosystem services: Beyond the Millennium Ecosystem Assessment. *Proceedings of the National Academy of Sciences*, **106**(5), pp.1305-1312.

Cavalier-Smith, T., 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Annals of Botany*, **95**(1), pp.147-175.

Chadwick, O.A., Derry, L.A., Vitousek, P.M., Huebert, B.J. and Hedin, L.O., 1999. Changing sources of nutrients during four million years of ecosystem development. *Nature*, **397**(6719), pp.491-497.

Chamberlain, S.A. and Szöcs, E., 2013. taxize: taxonomic search and retrieval in R. *F1000Research*, **2**, pp.191.

Chamberlain, S., Szöcs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., O'Donnell, J., Oksanen, J. and Tzovaras, B.G., 2020. taxize: Taxonomic information from around the web. *R package version 0.9*, pp.*92*.

Chapman, T., Miles, S. and Trivedi, C., 2019. Capturing, protecting and restoring plant diversity in the UK: RBG Kew and the Millennium Seed Bank. *Plant Diversity*, **41**(2), pp.124-131.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, pp.321-357.

Chénais, B., Caruso, A., Hiard, S. and Casse, N., 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene*, **509**(1), pp.7-15.

Christenhusz, M.J. and Byng, J.W., 2016. The number of known plants species in the world and its annual increase. *Phytotaxa*, **261**(3), pp.201-217.

Christenhusz, M.J. and Chase, M.W., 2014. Trends and concepts in fern classification. *Annals of Botany*, **113**(4), pp.571-594.

Christenhusz, M.J., Reveal, J.L., Farjon, A., Gardner, M.F., Mill, R.R. and Chase, M.W., 2011. A new classification and linear sequence of extant gymnosperms. *Phytotaxa*, **19**, pp.55-70.

Chrtek Jr, J., Zahradníček, J., Krak, K. and Fehrer, J., 2009. Genome size in Hieracium subgenus Hieracium (Asteraceae) is strongly correlated with major phylogenetic groups. *Annals of Botany*, **104**(1), pp.161-178.

Chytrý, M. *et al.*, 2021. Pladias Database of the Czech Flora and Vegetation. *Preslia*, 93, pp.1-87

Chytrý, M., Pyšek, P., Wild, J., Pino, J., Maskell, L.C. and Vilà, M., 2009. European map of alien plant invasions based on the quantitative assessment across habitats. *Diversity and Distributions*, **15**(1), pp.98-107.

Chytrý, M., Tichý, L., Dřevojan, P., Sádlo, J. and Zelený, D., 2018. Ellenberg-type indicator values for the Czech flora. *Preslia*, **90**(2), pp.83-103.

Clapham, A.R., Tutin, T.G. and Moore, D.M., 1990. *Flora of the British Isles*. Cambridge University Press.

Clapham, A.R., Tutin, T.G. and Warburg, E.F., 1952. *Flora of the British Isles*. Cambridge University Press.

Clark, C.D., Hughes, A.L., Greenwood, S.L., Jordan, C. and Sejrup, H.P., 2012. Pattern and timing of retreat of the last British-Irish Ice Sheet. *Quaternary Science Reviews*, **44**, pp.112-146.

Clements, D.R., Upadhyaya, M.K., Joshi, S. and Shrestha, A., 2022. Global Plant Invasions on the Rise. In *Global Plant Invasions* pp.1-28. Springer.

Clubbe, C., Ainsworth, A.M., Bárrios, S., Bensusan, K., Brodie, J., Cannon, P., Chapman, T., Copeland, A.I., Corcoran, M., Dani Sanchez, M. and David, J.C., 2020. Current knowledge, status, and future for plant and fungal diversity in Great Britain and the UK Overseas Territories. *Plants, People, Planet,* **2**(5), pp.557-579.

Collins, A.R., Naderi, R. and Mueller-Schaerer, H., 2011. Competition between cytotypes changes across a longitudinal gradient in *Centaurea stoebe* (Asteraceae). *American Journal of Botany*, **98**(12), pp.1935-1942.

Coulthard, E., Norrey, J., Shortall, C. and Harris, W.E., 2019. Ecological traits predict population changes in moths. *Biological Conservation*, **233**, pp.213-219.

Crowley, K.F. *et al.*, 2012. Do nutrient limitation patterns shift from nitrogen toward phosphorus with increasing nitrogen deposition across the northeastern United States? *Ecosystems*, **15**(6), pp.940-957.

Curwen, E.C., 1927. Prehistoric agriculture in Britain. *Antiquity*, **1**(3), pp.261-289.

Cutler, A., Cutler, D.R. and Stevens, J.R., 2012. Random forests. In *Ensemble Machine Learning*, pp. 157-175. Springer, Boston, MA.

Dankowski, T. and Ziegler, A., 2016. Calibrating random forests for probability estimation. *Statistics in Medicine*, **35**(22), pp.3949-3960.

Darwin, C., 1903. *More letters of Charles Darwin: a record of his work in a series of hitherto unpublished letters* (Vol. **2**). D. Appleton.

Dasgupta, P., 2021. *The economics of biodiversity: The Dasgupta review.* HM Treasury.

Database for the Biological Flora of the British Isles. British Ecological Society. https://www.britishecologicalsociety.org/publications/journals/journal-of-ecology/biological-flora-database/ [accessed June 2022]

Dauby, G., Stévart, T., Droissart, V., Cosiaux, A., Deblauwe, V., Simo-Droissart, M., Sosef, M.S., Lowry, P.P., Schatz, G.E., Gereau, R.E. and Couvreur, T.L., 2017. *ConR*: An R package to assist large-scale multispecies preliminary conservation assessments using distribution data. *Ecology and Evolution*, **7**(24), pp.11292-11303.

Davies, T.J., Barraclough, T.G., Chase, M.W., Soltis, P.S., Soltis, D.E. and Savolainen, V., 2004. Darwin's abominable mystery: insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences*, **101**(7), pp.1904-1909.

Dayrell, R.L., Arruda, A.J., Pierce, S., Negreiros, D., Meyer, P.B., Lambers, H. and Silveira, F.A., 2018. Ontogenetic shifts in plant ecological strategies. *Functional Ecology*, **32**(12), pp.2730-2741.

Dehnen-Schmutz, K., Pescott, O.L., Booy, O. and Walker, K.J., 2022. Integrating expert knowledge at regional and national scales improves impact assessments of non-native species. *NeoBiota*, **77**, pp.79-100.

Deng, M., Liu, L., Jiang, L., Liu, W., Wang, X., Li, S., Yang, S. and Wang, B., 2018. Ecosystem scale trade-off in nitrogen acquisition pathways. *Nature Ecology & Evolution*, **2**(11), pp.1724-1734.

Department for Environment, Food & Rural Affairs, 2022. Nutrient pollution: reducing the impact on protected sites. https://www.gov.uk/government/publications/nutrient-pollution-reducing-the-impact-on-protected-sites/nutrient-pollution-reducing-the-impact-on-protected-sites [accessed November 2022]

de Vere, N., Rich, T.C., Ford, C.R., Trinder, S.A., Long, C., Moore, C.W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K. and Wilkinson, M.J., 2012. DNA barcoding the native flowering plants and conifers of Wales. *PloS One*, **7**(6), e37945.

Díaz, S. *et al.*, 2016. The global spectrum of plant form and function. *Nature*, **529**(7585), pp.167-171.

Diekmann, M., 2003. Species indicator values as an important tool in applied plant ecology–a review. *Basic and Applied Ecology*, **4**(6), pp.493-506.

Dodsworth, S., Chase, M.W. and Leitch, A.R., 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society*, **180**(1), pp.1-5.

Dodsworth, S., Leitch, A.R. and Leitch, I.J., 2015. Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development*, **35**, pp.73-78.

Doležel, J., Bartoš, J., Voglmayr, H., Greilhuber, J., 2003. Nuclear DNA content and genome size of trout and human. *Cytometry Part A,* **51**, pp.127-128.

Döring, M., 2017. Zeigerwerte von Pflanzen & Flechten in Mitteleuropa. GBIF Secretariat. Checklist dataset https://doi.org/10.15468/tpngma [accessed via GBIF.org on 21 December 2020]

Dornelas, M. *et al.*, 2013. Quantifying temporal change in biodiversity: challenges and opportunities. *Proceedings of the Royal Society. Series B. Biological Sciences*, **280**(1750), pp.20121931.

Doyle, J.J. and Coate, J.E., 2019. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences*, **180**(1), pp.1-52.

Du, E., Terrer, C., Pellegrini, A.F., Ahlström, A., van Lissa, C.J., Zhao, X., Xia, N., Wu, X. and Jackson, R.B., 2020. Global patterns of terrestrial nitrogen and phosphorus limitation. *Nature Geoscience*, **13**(3), pp.221-226.

Edelsbrunner, H., Kirkpatrick, D. and Seidel, R., 1983. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, **29**(4), pp.551-559.

EDINA Digimap https://digimap.edina.ac.uk/ [accessed August 2022]

Ehlers, J. and Gibbard, P.L., 2004. *Quaternary glaciations-extent and chronology: part I: Europe*. Elsevier.

Ellenberg, H., 1974. Zeigerwerte der Gefässpflanzen Mitteleuropas. *Scripta Geobotanica*, **9**, pp.9-166.

Ellenberg, H., Weber, H.E., Düll, R., Wirth V., Werner, W., Paulißen, D., 1991. Zeigerwerte von Pflanzen in Mitteleuropa. *Scripta Geobotanica* **18**, Goltze Verlag, Göttingen

Einset, J. and Collins, A.R., 2018. Genome size and sensitivity to DNA damage by X-rays—plant comets tell the story. *Mutagenesis*, **33**(1), pp.49-51.

Elliott, T.A. and Gregory, T.R., 2015. Do larger genomes contain more diverse transposable elements? *BMC Evolutionary Biology*, ***15***(1), pp.1-10.

Elmqvist, T., Zipperer, W.C. and Güneralp, B., 2015. Urbanization, habitat loss and biodiversity decline: solution pathways to break the cycle. In *The Routledge Handbook of Urbanization and Global Environmental Change*, pp. 163-175. Routledge.

Elser, J.J., Bracken, M.E., Cleland, E.E., Gruner, D.S., Harpole, W.S., Hillebrand, H., Ngai, J.T., Seabloom, E.W., Shurin, J.B. and Smith, J.E., 2007. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecology Letters*, **10**(12), pp.1135-1142.

Engemann, K., Enquist, B.J., Sandel, B., Boyle, B., Jørgensen, P.M., Morueta-Holme, N., Peet, R.K., Violle, C. and Svenning, J.C., 2015. Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution*, **5**(3), pp.807-820.

Ertsen, A.C.D., Alkemade, J.R.M. and Wassen, M.J., 1998. Calibrating Ellenberg indicator values for moisture, acidity, nutrient availability and salinity in the Netherlands. *Plant Ecology*, **135**(1), pp.113-124.

Escudero, M. and Wendel, J.F., 2020. The grand sweep of chromosomal evolution in angiosperms. *New Phytologist*, **228**(3), pp.805-808.

Faizullah, L., Morton, J.A., Hersch-Green, E.I., Walczyk, A.M., Leitch, A.R. and Leitch, I.J., 2021. Exploring environmental selection on genome size in angiosperms. *Trends in Plant Science*, **26**(10), pp.1039-1049.

Falster, D. *et al.*, 2021. AusTraits, a curated plant trait database for the Australian flora. *Scientific Data*, ***8***(1), pp.1-20.

Fernie, W.T., 1890. The History and Capabilities of Herbal Simples: IX.—Borage. *The Hospital*, **8**(190), pp.90.

Firbank, L.G., Smart, S.M., van de Poll, H.M., Bunce, R.G.H., Hill, M.O., Howard, D.C., Watkins, J.W., Stark, G.J., 2000. Causes of Change in British Vegetation, ECOFACT, vol. **3**. Department of the Environment, Transport and the Regions, London, pp.1-99.

Fitter, A.H. and Peat, H.J., 1994. The ecological flora database. *Journal of Ecology*, **82**(2), 415-425. http://www.ecoflora.co.uk [accessed 14 April 2020]

Fleischmann, A., Michael, T.P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E.M., Greilhuber, J., Müller, K.F. and Heubl, G., 2014. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany*, **114**(8), pp.1651-1663.

Fowler, P.J., 1983. *The farming of prehistoric Britain*. Cambridge University Press Archive.

Fox, D.T., Soltis, D.E., Soltis, P.S., Ashman, T.L. and Van de Peer, Y., 2020. Polyploidy: a biological force from cells to ecosystems. *Trends in Cell Biology*, **30**(9), pp.688-694.

Francis, D., Davies, M.S. and Barlow, P.W., 2008. A strong nucleotypic effect on the cell cycle regardless of ploidy level. *Annals of Botany*, **101**(6), pp.747-757.

Franks, P.J. and Farquhar, G.D., 2001. The effect of exogenous abscisic acid on stomatal development, stomatal mechanics, and leaf gas exchange in *Tradescantia virginiana*. *Plant Physiology*, **125**(2), pp.935-942.

Fritz, S.A. and Purvis, A., 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, **24**(4), pp.1042-1051.

Gerrish, J., 2022. *Julius Caesar: The Gallic War Books V-VI*. Liverpool University Press.

Ghannam, R.B. and Techtmann, S.M., 2021. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal,* **19**, pp.1092-1107.

Gonzalez-Voyer, A. and Hardenberg, A.V., 2014. An introduction to phylogenetic path analysis. In *Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology*, pp. 201-229. Springer, Berlin, Heidelberg.

Green, B.H., 1990. Agricultural intensification and the loss of habitat, species and amenity in British grasslands: a review of historical change and assessment of future prospects. *Grass and Forage Science*, **45**(4), pp.365-372

Greilhuber, J., Doležel, J., Lysák, M.A. and Bennett, M.D., 2005. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. *Annals of Botany*, **95**(1), pp.255-260.

Greilhuber, J. and Leitch, I.J., 2013. Genome size and the phenotype. In *Plant Genome Diversity Volume* **2**, pp. 323-344. Springer, Vienna.

Greilhuber, J.P., Rudall, P.J., Cribb, P.J., Cutler, D.F., Humphries, C.J., 1995. Chromosomes of the monocotyledons (general aspects). Monocotyledons: systematics and evolution. Royal Botanic Garden Kew, pp. 379-414.

Grime, J.P., 1974. Vegetation classification by reference to strategies. *Nature*, **250**(5461), pp.26-31.

Grime, J.P., 1977. Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. *The American Naturalist*, **111**(982), pp.1169-1194.

Grime, J.P., 1979. *Plant Strategies and Vegetation Processes Vegetation Processes*. John Wiley & Sons, Limited.

Grime, J.P., 1983. Prediction of weed and crop response to climate based upon measurements of nuclear DNA content. *Aspects of Applied Biology*, **4**, pp. 87-98.

Grime, J.P. and Mowforth, M.A., 1982. Variation in genome size—an ecological interpretation. *Nature*, **299**(5879), pp.151-153.

Grime, J.P. and Pierce, S., 2012. *The Evolutionary Strategies that Shape Ecosystems*. John Wiley & Sons.

Grömping, U., 2007. Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, **17**, pp.1-27.

Grömping, U. and Matthias, L., 2021. Package 'relaimpo'. *Relative Importance of Regressors in Linear Models; R Fundation for Statstical Computing: Vienna, Austria*. [R package]

Groom, Q.J., 2013a. Estimation of vascular plant occupancy and its change using kriging. *New Journal of Botany*, **3**(1), pp.33-46.

Groom, Q.J., 2013b. Some poleward movement of British native vascular plants is occurring, but the fingerprint of climate change is not evident. *PeerJ*, **1**, e77.

Groom, Q., Northumberland, S., Walker, K., Yorkshire, N.W., Mcintosh, J., 2011. BSBI recording the British and Irish flora 2010–2020 Annex 1: guidance on sampling approaches. BSBI Recording the British and Irish Flora 2010–2020, pp.1–9.

Guignard, M.S., Crawley, M.J., Kovalenko, D., Nichols, R.A., Trimmer, M., Leitch, A.R. and Leitch, I.J., 2019. Interactions between plant genome size, nutrients and herbivory by rabbits, molluscs and insects on a temperate grassland. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, **286**(1899), pp.20182619.

Guignard, M.S., Nichols, R.A., Knell, R.J., Macdonald, A., Romila, C.A., Trimmer, M., Leitch, I.J. and Leitch, A.R., 2016. Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. *New Phytologist*, **210**(4), pp.1195-1206.

Guo, W.Y., van Kleunen, M., Pierce, S., Dawson, W., Essl, F., Kreft, H., Maurel, N., Pergl, J., Seebens, H., Weigelt, P. and Pyšek, P., 2019. Domestic gardens play a dominant role in selecting alien species with adaptive strategies that facilitate naturalization. *Global Ecology and Biogeography*, **28**(5), pp.628-639.

Guo, W.Y., van Kleunen, M., Winter, M., Weigelt, P., Stein, A., Pierce, S., Pergl, J., Moser, D., Maurel, N., Lenzner, B., Kreft, H., Essl, F., Dawson, W. and Pyšek, P., 2018. The role of adaptive strategies in plant naturalization. *Ecology Letters*, **21**(9), pp.1380-1389.

Halverson, K., Heard, S.B., Nason, J.D. and Stireman III, J.O., 2008. Origins, distribution, and local co-occurrence of polyploid cytotypes in *Solidago altissima* (Asteraceae). *American Journal of Botany*, **95**(1), pp.50-58.

Hamilton, N.E., 2015. ggtern: An extension to 'ggplot2', for the creation of ternary diagrams. [R package version 1.0.6.0]. Retrieved from http:// www.ggtern.com/

Hamilton, N.E. and Ferry, M., 2018. ggtern: Ternary diagrams using ggplot2. *Journal of Statistical Software*, **87**, pp.1-17.

Hardenberg, A.V. and Gonzalez-Voyer, A., 2013. Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution: International Journal of Organic Evolution*, **67**(2), pp.378-387.

Harfoot, M.B., Johnston, A., Balmford, A., Burgess, N.D., Butchart, S.H., Dias, M.P., Hazin, C., Hilton-Taylor, C., Hoffmann, M., Isaac, N.J. and Iversen, L.L., Outhwaite, C.L., Visconti, P. and Geldmann, J., 2021. Using the IUCN Red List to map threats to terrestrial vertebrates at global scale. *Nature Ecology & Evolution*, **5**(11), pp.1510-1519.

Hegarty, M.J. and Hiscock, S.J., 2008. Genomic clues to the evolutionary success of polyploid plants. *Current Biology*, **18**(10), pp.R435-R444.

Henniges, M.C., Powell, R.F., Mian, S., Stace, C.A., Walker, K.J., Gornall, R.J., Christenhusz, M.J., Brown, M.R., Twyford, A.D., Hollingsworth, P.M., Jones, L., de Vere, N., Antonelli, A., Leitch, A.R. and Leitch, I.J., 2021. A taxonomic, genetic and ecological data resource for the vascular plants of Britain and Ireland. NERC Environmental Information Data Centre (2021). (Dataset). https://doi.org/10.5285/9f097d82-7560-4ed2-af13-604a9110cf6d [accessed December 2022]

Henniges, M.C., Powell, R.F., Mian, S., Stace, C.A., Walker, K.J., Gornall, R.J., Christenhusz, M.J., Brown, M.R., Twyford, A.D., Hollingsworth, P.M., Jones, L., de Vere, N., Antonelli, A., Leitch, A.R. and Leitch, I.J., 2021. BIFloraExplorer: A taxonomic, genetic and ecological data resource for the vascular plants of Britain and Ireland (v0.1.0). [R package]. https://github.com/RBGKew/BIFloraExplorer.

Henniges, M.C., Powell, R.F., Mian, S., Stace, C.A., Walker, K.J., Gornall, R.J., Christenhusz, M.J., Brown, M.R., Twyford, A.D., Hollingsworth, P.M., Jones, L., de Vere, N., Antonelli, A., Leitch, A.R. and Leitch, I.J. 2022. A taxonomic, genetic and ecological data resource for the vascular plants of Britain and Ireland. *Scientific Data*, **9**(1), pp.1-8.

Herben, T., Suda, J., Klimešová, J., Mihulka, S., Říha, P. and Šímová, I., 2012. Ecological effects of cell-level processes: genome size, functional traits and regional abundance of herbaceous plant species. *Annals of Botany*, **110**(7), pp.1357-1367.

Hessen, D.O., Elser, J.J., Sterner, R.W. and Urabe, J., 2013. Ecological stoichiometry: an elementary approach using basic principles. *Limnology and Oceanography*, **58**(6), pp.2219-2236.

Hessen, D.O., Jeyasingh, P.D., Neiman, M. and Weider, L.J., 2010. Genome streamlining and the elemental costs of growth. *Trends in Ecology & Evolution*, **25**(2), pp.75-80.

Hibberd, J.M. and Jeschke, D.W., 2001. Solute flux into parasitic plants. *Journal of Experimental Botany*, **52**(363), pp.2043-2049.

Hickling, R., Roy, D.B., Hill, J.K., Fox, R. and Thomas, C.D., 2006. The distributions of a wide range of taxonomic groups are expanding polewards. *Global Change Biology*, **12**(3), pp.450-455.

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R. and Leitch, I.J., 2017. Is there an upper limit to genome size? *Trends in Plant Science*, ***22***(7), pp.567-573.

Hijmans, R.J., Bivand, R., Forner, K., Ooms, J., Pebesma, E. and Sumner, M.D., 2022. Package 'terra'. [R package]

Hijmans, R.J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J.A., Lamigueiro, O.P., Bevan, A., Racine, E.B., Shortridge, A. and Hijmans, M.R.J., 2015. Package 'raster'. [R package], *734*.

Hijmans, R.J., Williams, E. and Vennes, C., 2020. Geosphere: Spherical Trigonometry (v1.5-10) [R package]

Hill, M.O., 2011. Frescalo - a computer program to analyse your biological records. https://www.brc.ac.uk/biblio/frescalo-computer-program-analyse-your-biological-records [downloaded March 2019]

Hill, M.O., 2012. Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in Ecology and Evolution*, **3**(1), pp.195-205.

Hill, M.O. and Carey, P.D., 2009. Prediction of yield in the Rothamsted Park Grass Experiment by Ellenberg indicator values. *Journal of Vegetation Science,* **8**(4), pp.579-586.

Hill, M.O., Mountford, J.O., Roy, D.B. and Bunce, R.G.H., 1999. *Ellenberg's indicator values for British plants. ECOFACT Volume 2 Technical Annex*. Institute of Terrestrial Ecology.

Hill, M.O. and Preston, C.D., 2015. Disappearance of boreal plants in southern Britain: habitat loss or climate change?. *Biological Journal of the Linnean Society*, **115**(3), pp.598-610.

Hill, M.O., Preston, C.D. and Roy, D.B., 2004. *PLANTATT-attributes of British and Irish plants: status, size, life history, geography and habitats*. Centre for Ecology & Hydrology.

Hillebrand, H., 2004. On the generality of the latitudinal diversity gradient. *The American Naturalist,* **163**(2), pp.192-211.

Hodgson, J.G., 2003. Change the change index? *BSBI News* **93**, pp.44–47

Hodgson, J.G., Grime, J.P., Hunt, R. and Thompson, K., 1995. *The electronic comparative plant ecology*. London: Chapman & Hall.

Hodgson, J.G. *et al.*, 2010. Stomatal vs. genome size in angiosperms: the somatic tail wagging the genomic dog?. *Annals of Botany*, **105**(4), pp.573-584.

Hollis, D., McCarthy, M., Kendon, M., Legg, T. 2022. HadUK-Grid Climate Observations by UK countries, v1.1.0.0 (1836-2021). NERC EDS Centre for Environmental Data Analysis, *26 May 2022*. doi:10.5285/59a7cd0dcd474f5f906ead4073a9be8b [accessed June 2022]

Houlton, B.Z., Wang, Y.P., Vitousek, P.M. and Field, C.B., 2008. A unifying framework for dinitrogen fixation in the terrestrial biosphere. *Nature*, **454**(7202), pp.327-330.

Hudson, L.N. *et al.*, 2014. The PREDICTS database: a global database of how local terrestrial biodiversity responds to human impacts. *Ecology and Evolution*, **4**(24), pp.4701-4735.

Hulme, P.E., 2009. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, **46**(1), pp.10-18.

Hvitfeldt, E., 2020. themis: Extra Recipe Steps for Dealing with Unbalanced Data. https://CRAN.R-project.org/package=themis. [R package]

Ingrouille, M., 2012. *Historical ecology of the British flora*. Springer Science & Business Media.

IPNI. International Plant Names Index, 2020. Published on the Internet http://www.ipni.org, The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Botanic Gardens [retrieved 10 December 2020]

IUCN, 2022. The IUCN Red List of Threatened Species. Version 2022-1. https://www.iucnredlist.org. [accessed November 2022]

Isaac, N.J. and Pocock, M.J., 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**(3), pp.522-531.

Isaac, N.J., van Strien, A.J., August, T.A., de Zeeuw, M.P. and Roy, D.B., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, **5**(10), pp.1052-1060.

Jenczewski, E., Mercier, R., Macaisne, N. and Mézard, C., 2013. Meiosis: Recombination and the control of cell division. In *Plant Genome Diversity Volume **2**,* pp.121-136. Springer, Vienna.

Jenkins, D.G. *et al.,* 2007. Does size matter for dispersal distance? *Global Ecology and Biogeography*, ***16***(4), pp.415-425.

Jiao, Y. *et al.,* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**(7345), pp.97-100.

Jin, Y. and Qian, H., 2019. V. PhyloMaker: an R package that can generate very large phylogenies for vascular plants. *Ecography*, **42**(8), pp.1353-1359.

Jones, L., Twyford, A.D., Ford, C.R., Rich, T.C., Davies, H., Forrest, L.L., Hart, M.L., McHaffie, H., Brown, M.R., Hollingsworth, P.M. and de Vere, N., 2021. Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom. *Molecular Ecology Resources*, **21**(6), pp.2050-2062.

Joppa, L.N., Butchart, S.H., Hoffmann, M., Bachman, S.P., Akçakaya, H.R., Moat, J.F., Böhm, M., Holland, R.A., Newton, A., Polidoro, B. and Hughes, A., 2016. Impact of alternative metrics on estimates of extent of occurrence for extinction risk assessment. *Conservation Biology*, **30**(2), pp.362-370.

Kang, M., Tao, J., Wang, J., Ren, C., Qi, Q., Xiang, Q.Y. and Huang, H., 2014. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. *New Phytologist*, **202**(4), pp.1371-1381.

Kattge, J. *et al.,* 2020. TRY plant trait database–enhanced coverage and open access. *Global Change Biology*, **26**(1), pp.119-188. https://doi.org/10.1111/gcb.14904 [accessed September 2021]

Kendon, M., McCarthy, M., Jevrejeva, S., Matthews, A., Sparks, T., Garforth, J. and Kennedy, J., 2022. State of the UK Climate 2021. *International Journal of Climatology*, **42**, pp.1-80.

Kent, D.H., 1992. *List of vascular plants of the British Isles.* Botanical Society of the British Isles.

Knight, C.A. and Ackerly, D.D., 2002. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*, **5**(1), pp.66-76.

Knight, C.A. and Beaulieu, J.M., 2008. Genome size scaling through phenotype space. *Annals of Botany*, **101**(6), pp.759-766.

Knight, C.A., Clancy, R.B., Götzenberger, L., Dann, L. and Beaulieu, J.M., 2010. On the relationship between pollen size and genome size. *Journal of Botany*, 2010, pp.7.

Knight, C.A., Molinari, N.A. and Petrov, D.A., 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*, **95**(1), pp.177-190.

Koenker, R. and Bassett, G., 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society,* **46**, pp. 33–50.

Koenker, R. *et al.*, 2018. Package 'quantreg'. Cran R-project.org [R package]

Kolář, F., Čertner, M., Suda, J., Schönswetter, P. and Husband, B.C., 2017. Mixed-ploidy species: progress and opportunities in polyploid research. *Trends in Plant Science*, **22**(12), pp.1041-1055.

Kowarik, I. and Lippe, M.V.D., 2008. Pathways in plant invasions. In *Biological invasions* pp.29-47. Springer, Berlin, Heidelberg.

Kubešová, M., Moravcova, L., Suda, J., Jarošík, V. and Pyšek, P., 2010. Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia*, **82**(1), pp.81-96.

Kuhn, M. and Wickham, H., 2020. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. *Boston, MA, USA*. [accessed on 10 December 2020]

Kumar, S.S. and Shaikh, T., 2017. Empirical evaluation of the performance of feature selection approaches on random forest. In *2017 international conference on computer and applications (ICCA)* pp. 227-231. IEEE.

Kursa, M.B., Jankowski, A. and Rudnicki, W.R., 2010. Boruta – a system for feature selection. *Fundamenta Informaticae*, **101**(4), pp.271-285.

Kursa, M.B. and Rudnicki, W.R., 2010. Feature selection with the Boruta package. *Journal of Statistical Software*, **36**, pp.1-13.

Lake, J.C. and Leishman, M.R., 2004. Invasion success of exotic plants in natural ecosystems: the role of disturbance, plant attributes and freedom from herbivores. *Biological Conservation*, **117**(2), pp.215-226.

Lamanna, C. *et al.*, 2014. Functional trait space and the latitudinal diversity gradient. *Proceedings of the National Academy of Sciences*, **111**(38), pp.13745-13750.

Lambers, H., Brundrett, M.C., Raven, J.A. and Hopper, S.D., 2011. Plant mineral nutrition in ancient landscapes: high plant species diversity on infertile soils is linked to functional diversity for nutritional strategies. *Plant and Soil*, **348**(1), pp.7-27.

Land Cover map of Great Britain, 1990. [TIFF geospatial data], Scale 1:250000, Tiles: GB, Updated: 1 December 1990, CEH, Using: EDINA Environment Digimap Service, <https://digimap.edina.ac.uk> [downloaded 2022-08-12 16:35:56.113]

Land Cover Map, 2007. [TIFF geospatial data], Scale 1:250000, Tiles: GB, Updated: 18 July 2008, CEH, Using: EDINA Environment Digimap Service, <https://digimap.edina.ac.uk> [downloaded 2022-08-12 16:35:56.134]

Land Cover map of Great Britain, 2017. [TIFF geospatial data], Scale 1:250000, Tiles: GB, Updated: 30 June 2020, CEH, Using: EDINA Environment Digimap Service, <https://digimap.edina.ac.uk> [downloaded 2022-08-12 16:35:56.143]

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C. and Soltis, P.S., 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany*, **105**(3), pp.348-363.

Lanz, B., Dietz, S. and Swanson, T., 2018. The expansion of modern agriculture and global biodiversity decline: an integrated assessment. *Ecological Economics*, **144**, pp.260-277.

Laport, R.G., Hatem, L., Minckley, R.L. and Ramsey, J., 2013. Ecological niche modeling implicates climatic adaptation, competitive exclusion, and niche conservatism among *Larrea tridentata* cytotypes in North American deserts, 2. *The Journal of the Torrey Botanical Society*, **140**(3), pp.349-363.

Lawson, T. and Vialet-Chabrand, S., 2019. Speedy stomata, photosynthesis and plant water use efficiency. *New Phytologist*, **221**(1), pp.93-98.

Leclère, D. *et al.,* 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. *Nature*, **585**(7826), pp.551-556.

Leitch, I.J. and Bennett, M.D., 2004. Genome downsizing in polyploid plants. *Biological journal of the Linnean Society*, **82**(4), pp.651-663.

Leitch I.J. and Bennett M.D., 2007. Genome size and its uses: the impact of flow cytometry. In: Doležel J, Greilhuber J, Suda J eds. *Flow Cytometry with Plant Cells*: Weinheim: Wiley-VCH, pp.153-176.

Leitch, I.J., Johnston, E., Pellicer, J., Hidalgo, O. & Bennett, M.D., 2019. Plant DNA C-values Database (release 7.1, April 2019). https://cvalues.science.kew.org/ [accessed September 2022]

Leitch, I.J. and Leitch, A.R., 2013. Genome size diversity and evolution in land plants. In *Plant Genome Diversity Volume* **2**, pp.307-322. Springer, Vienna.

Lenoir, J., Gégout, J.C., Marquet, P.A., de Ruffray, P. and Brisse, H., 2008. A significant upward shift in plant species optimum elevation during the 20th century. *Science*, **320**(5884), pp.1768-1771.

Letunic, I. and Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, **49**(W1), pp.293-296.

Li, F.W. and Harkess, A., 2018. A guide to sequence your favorite plant genomes. *Applications in Plant Sciences*, **6**(3), e1030.

Li, Z., Ma, W., Liang, C., Liu, Z., Wang, W. and Wang, L., 2015. Long-term vegetation dynamics driven by climatic variations in the Inner Mongolia grassland: findings from 30-year monitoring. *Landscape Ecology*, **30**(9), pp.1701-1711.

Liao, H., Li, D., Zhou, T., Huang, B., Zhang, H., Chen, B. and Peng, S., 2021. The role of functional strategies in global plant distribution. *Ecography*, **44**(4), pp.493-503.

Lim, J., Crawley, M.J., de Vere, N., Rich, T. and Savolainen, V., 2014. A phylogenetic analysis of the British flora sheds light on the evolutionary and ecological factors driving plant invasions. *Ecology and Evolution*, **4**(22), pp.4258-4269.

Lindeman RH, Merenda PF, Gold RZ (1980). *Introduction to Bivariate and Multivariate Analysis.* Scott, Foresman, Glenview, IL.

Liu, Y., 2014. Random forest algorithm in big data environment. *Computer Modelling & New Technologies*, **18**(12A), pp.147-151.

Loram, A., Thompson, K., Warren, P.H. and Gaston, K.J., 2008. Urban domestic gardens (XII): the richness and composition of the flora in five UK cities. *Journal of Vegetation Science*, **19**(3), pp.321-330.

Lozano-Baena, M.D., Tasset, I., Muñoz-Serrano, A., Alonso-Moraga, Á. and de Haro-Bailón, A., 2016. Cancer prevention and health benefices of traditionally consumed Borago officinalis plants. *Nutrients*, **8**(1), pp.48.

MacGillivray, C.W. and Grime, J.P., 1995. Genome size predicts frost resistance in British herbaceous plants: implications for rates of vegetation response to global warming. *Functional Ecology*, **9**(2), pp.320-325.

Manchester, S.J. and Bullock, J.M., 2001. The impacts of non-native species on UK biodiversity and the effectiveness of control. *Journal of Applied Ecology*, **37**(5), pp.845-864.

Martin, S. 2020. Smart: Steve Martin's Assorted R Toolbox (v0.1.3). [R package].

Maskell, L.C., Firbank, L.G., Thompson, K., Bullock, J.M. and Smart, S.M., 2006. Interactions between non-native plant species and the floristic composition of common habitats. *Journal of Ecology*, pp.1052-1060.

Maskell, L.C., Smart, S.M., Bullock, J.M., Thompson, K. and Stevens, C.J., 2010. Nitrogen deposition causes widespread loss of species richness in British habitats. *Global Change Biology*, **16**(2), pp.671-679.

Masterson, J., 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, **264**(5157), pp.421-424.

Met Office, 2022. Climate change in the UK. https://www.metoffice.gov.uk/weather/climate-change/climate-change-in-the-uk [accessed October 2022]

McClean, C.J., Van den Berg, L.J., Ashmore, M.R. and Preston, C.D., 2011. Atmospheric nitrogen deposition explains patterns of plant species loss. *Global Change Biology*, **17**(9), pp.2882-2892.

Menge, D.N., Batterman, S.A., Hedin, L.O., Liao, W., Pacala, S.W. and Taylor, B.N., 2017. Why are nitrogen-fixing trees rare at higher compared to lower latitudes? *Ecology*, **98**(12), pp.3127-3140.

Meyer, C., Weigelt, P. and Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, ***19***(8), pp.992-1006.

NatureScot, 2018. Improved grassland (UK BAP Broad Habitat). https://www.nature.scot/sites/default/files/2018-02/Broad%20Habitat%20-%20Improved%20Grassland.pdf [accessed October 2022]

Netzwerk Phytodiversität Deutschland & Bundesamt für Naturschutz (Hrsg.), 2013. Verbreitungsatlas der Farn- und Blütenpflanzen Deutschlands. *Münster (Landwirtschaftsverlag)*, pp.912.

Newbold, T., Hudson, L.N., Contu, S., Hill, S.L., Beck, J., Liu, Y., Meyer, C., Phillips, H.R., Scharlemann, J.P. and Purvis, A., 2018. Widespread winners and narrow-ranged losers: Land use homogenizes biodiversity in local assemblages worldwide. *PLoS Biology*, **16**(12), e2006841.

Newbold, T. *et al.*, 2015. Global effects of land use on local terrestrial biodiversity. *Nature*, **520**(7545), pp.45-50.

Newbold, T., Scharlemann, J.P., Butchart, S.H., Şekercioğlu, Ç.H., Alkemade, R., Booth, H. and Purves, D.W., 2013. Ecological traits affect the response of tropical forest bird species to land-use intensity. *Proceedings of the Royal Society. Series B. Biological Sciences*, **280**(1750), p.20122131.

Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblížková, A., Dodsworth, S., Kovařík, A., Pellicer, J., Wang, W., Macas, J., Leitch, I.J. and Leitch, A.R., 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature Plants*, **6**(11), pp.1325-1329.

Ordnance Survey, 2015. A guide to coordinate systems in Great Britain. An introduction to mapping coordinate systems and the use of GPS datasets with Ordnance Survey mapping. https://www.bnhs.co.uk/2019/technology/grabagridref/OSGB.pdf [accessed October 2022]

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N. and Pearse, W., 2013. The caper package: comparative analysis of phylogenetics and evolution in R. [R package]

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F.G., Duan, L., Dunson, D., Roslin, T. and Abrego, N., 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, **20**(5), pp.561-576.

Pandit, M.K., White, S.M. and Pocock, M.J., 2014. The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. *New Phytologist*, **203**(2), pp.697-703.

Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H.S., Desper, R. and Didier, G., 2019. Package 'ape' [R package]. *Analyses of Phylogenetics and Evolution*, **2**(4), pp.47.

Päßler, U. and Ette, O., 2020. *Alexander von Humboldt: Geographie der Pflanzen. Unveröffentlichte Schriften aus dem Nachlass*. J.B. Metzler

Paterson, A.H., Wang, X., Li, J. and Tang, H., 2012. Ancient and recent polyploidy in monocots. In *Polyploidy and Genome Evolution,* pp.93-108. Springer, Berlin, Heidelberg.

Paton, A. *et al.*, 2020. Plant and fungal collections: Current status, future perspectives. *Plants, People, Planet*, **2**(5), pp.499-514.

Payne, R.J., Dise, N.B., Field, C.D., Dore, A.J., Caporn, S.J. and Stevens, C.J., 2017. Nitrogen deposition and plant biodiversity: past, present, and future. *Frontiers in Ecology and the Environment*, **15**(8), pp.431-436.

Pearce-Higgins, J.W., Ausden, M.A., Beale, C.M., Oliver, T.H. and Crick, H.Q.P., 2015. Research on the assessment of risks & opportunities for species in England as a result of climate change. Natural England, UK

Pearce-Higgins, J.W. *et al.,* 2017. A national-scale assessment of climate change impacts on species: Assessing the balance of risks and opportunities for multiple taxa. *Biological Conservation*, **213**, pp.124-134.

Pearson, D.E., Ortega, Y.K., Eren, Ö. and Hierro, J.L., 2016. Quantifying "apparent" impact and distinguishing impact from invasiveness in multispecies plant invasions. *Ecological Applications*, **26**(1), pp.162-173.

Pebesma, E.J., 2018. Simple features for R: standardized support for spatial vector data. *R J.*, ***10***(1), pp.439. [R package]

Pebesma, E., Bivand, R., Pebesma, M.E., RColorBrewer, S. and Collate, A.A.A., 2012. Package 'sp' (v1.5-0). [R package]. *The Comprehensive R Archive Network*. https://doi.org/10.5285/9b203324-6b37-4e91-b028-e073b197fb9f [retrieved March 2022]

Pegoraro, L., Cafasso, D., Rinaldi, R., Cozzolino, S. and Scopece, G., 2016. Habitat preference and flowering-time variation contribute to reproductive isolation between diploid and autotetraploid *Anacamptis pyramidalis*. *Journal of Evolutionary Biology*, **29**(10), pp.2070-2082.

Pegoraro, L., de Vos, J.M., Cozzolino, S. and Scopece, G., 2019. Shift in flowering time allows diploid and autotetraploid *Anacamptis pyramidalis* (Orchidaceae) to coexist by reducing competition for pollinators. *Botanical Journal of the Linnean Society*, **191**(2), pp.274-284.

Pellicer, J., Fay, M.F. and Leitch, I.J., 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, **164**(1), pp.10-15.

Pellicer, J., Hidalgo, O., Dodsworth, S. and Leitch, I.J., 2018. Genome size diversity and its impact on the evolution of land plants. *Genes*, **9**(2), pp.88.

Pellicer, J. and Leitch, I.J., 2019. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*, **226**(2), pp.301-305.

Pellicer, J., Powell, R.F. and Leitch, I.J., 2021. The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants. In *Molecular Plant Taxonomy*, pp.325-361. Humana, New York, NY.

Peng, Y., Yang, J., Leitch, I.J., Guignard, M.S., Seabloom, E.W., Cao, D., Zhao, F., Li, H., Han, X., Yong, J., Leitch, A.R., Wei, C., 2022. Plant genome size modulates grassland community responses to multi-nutrient additions. *New Phytologist,* **236**, pp.2091-2102.

Peñuelas, J., Poulter, B., Sardans, J., Ciais, P., Van Der Velde, M., Bopp, L., Boucher, O., Godderis, Y., Hinsinger, P., Llusia, J., Nardin, E., Vicca, S., Obersteiner, M. and Janssens, I.A., 2013. Human-induced nitrogen–phosphorus imbalances alter natural and managed ecosystems across the globe. *Nature Communications*, **4**(1), pp.1-10.

Pergl, J., Sádlo, J., Petřík, P., Danihelka, J., Chrtek Jr, J., Hejda, M., Moravcová, L., Perglová, I., Štajerová, K. and Pyšek, P., 2016. Dark side of the fence: ornamental plants as a source of wild-growing flora in the Czech Republic. *Preslia*, **88**(2), pp.163-184.

Perring, F.H. and Walters, S.M., eds. 1962. *Atlas of the British Flora.* Thomas Nelson & Sons, London.

Pescott, O.L., Humphrey, T.A., Stroh, P.A. and Walker, K.J., 2019a. Temporal changes in distributions and the species atlas: How can British and Irish plant data shoulder the inferential burden? *British & Irish Botany*, **1**(4), pp.250-282.

Pescott, O.L., Humphrey, T.A. and Walker, K.J., 2018. A short guide to using British and Irish plant occurrence data for research. Wallingford, NERC/Centre for Ecology & Hydrology, pp.9. (Unpublished)

Pescott, O.L., Powney, G.D. and Roy, D.B., 2016. Approaches to Bayesian occupancy modelling for habitat quality assessment. Wallingford, NERC/Centre for Ecology & Hydrology, pp.23.

Pescott, O.L., Walker, K.J., Harris, F., New, H., Cheffings, C.M., Newton, N., Jitlal, M., Redhead, J., Smart, S.M. and Roy, D.B., 2019b. The design, launch and assessment of a new volunteer-based plant monitoring scheme for the United Kingdom. *PloS One*, **14**(4), e0215891.

Pescott, O.L., Walker, K.J., Jitlal, M., Smart, S.M., Maskell, L., Schmucki, R., Day, J., Amos, C., Peck, K., Robinson, A. & Roy, D.B., 2019c. The National Plant Monitoring Scheme: A Technical Review. JNCC Report No. 622, JNCC, Peterborough, ISSN 0963-8091.

Pierce, S., Bottinelli, A., Bassani, I., Ceriani, R.M. and Cerabolini, B.E., 2014. How well do seed production traits correlate with leaf traits, whole-plant traits and plant ecological strategies? *Plant Ecology*, **215**(11), pp.1351-1359.

Pierce, S. *et al.*, 2017. A global method for calculating plant CSR ecological strategies applied across biomes world-wide. *Functional Ecology*, **31**(2), pp.444-457.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B. and Maintainer, R., 2017. Package 'nlme'. *Linear and nonlinear mixed effects models, version*, **3**(1). [R package]

Plants of the World Online, 2020. RBG Kew. http://www.plantsoftheworldonline.org/ [viewed 24 February 2020]

Pollock, L.J., Morris, W.K. and Vesk, P.A., 2012. The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, **35**(8), pp.716-725.

Powney, G.D., Preston, C.D., Purvis, A., Van Landuyt, W. and Roy, D.B., 2014b. Can trait-based analyses of changes in species distribution be transferred to new geographic areas? *Global Ecology and Biogeography*, **23**(9), pp.1009-1018.

Powney, G.D., Rapacciuolo, G., Preston, C.D., Purvis, A. and Roy, D.B., 2014a. A phylogenetically-informed trait-based analysis of range change in the vascular plant flora of Britain. *Biodiversity and Conservation*, **23**, pp.171-185.

Preece, R.C., 1995. Introduction-Island Britain: a Quaternary perspective. *Geological Society, London, Special Publications*, **96**(1), pp.1-2.

Preston, C.D., 2002. Approaches to native and alien species. *Transactions of the Suffolk Naturalists' Society*, **38**, pp.37-48.

Preston, C.D., 2013. Following the BSBI's lead: the influence of the Atlas of the British flora, 1962–2012. *New Journal of Botany*, **3**(1), pp.2-14.

Preston, C.D. and Hill, M.O., 2002. The geographical relationships of the British and Irish flora: a comparison of pteridophytes, flowering plants, liverworts and mosses. *Journal of Biogeography*, **26**(3), pp.629-642.

Preston, C.D., Pearman, D. and Dines, T.D., eds. 2002. *New Atlas of the British & Irish Flora*. Oxford University Press.

Preston, C.D., Pearman, D.A. and Hall, A.R., 2004. Archaeophytes in Britain. *Botanical Journal of the Linnean Society*, **145**(3), pp.257-294.

Pyšek, P., Pergl, J., Dawson, W., Essl, F., Kreft, H., Weigelt, P., Winter, M. and van Kleunen, M., 2022. European Plant Invasions. In *Global Plant Invasions* pp.151-165. Springer.

Pyšek, P., Prach, K. and Smilauer, P., 1995. Relating invasion success to plant traits: an analysis of the Czech alien flora. *Plant Invasions: General Aspects and Special Problems*, pp.39-60.

Pyšek, P. and Richardson, D.M., 2008. Traits associated with invasiveness in alien plants: where do we stand? In *Biological Invasions*, pp. 97-125. Springer, Berlin, Heidelberg.

Pyšek, P., Richardson, D.M., Rejmánek, M., Webster, G.L., Williamson, M. and Kirschner, J., 2004. Alien plants in checklists and floras: towards better communication between taxonomists and ecologists. *Taxon*, **53**(1), pp.131-143.

QGIS Development Team, 2022. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org [downloaded August 2022]

Qian, H. and Jin, Y., 2016. An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *Journal of Plant Ecology*, **9**(2), pp.233-239.

R Core Team, 2022. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. *http://www. R-project. org/*.

Ratnasingham, S. and Hebert, P.D., 2007. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, **7**(3), pp.355-364.

Raunkiær, C., 1934. *The life forms of plants and statistical plant geography; being the collected papers of C. Raunkiær.* Oxford at the Clarendon Press.

Ray, J., 1690. *Synopsis methodica stirpium Britannicarum, in qua tum notae generum characteristicae traduntur, tum species singulae breviter describuntur.* S. Smith (Londini).

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H. and Qi, J., 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant*, **11**(3), pp.414-428.

Renny-Byfield, S. and Wendel, J.F., 2014. Doubling down on genomes: polyploidy and crop plants. *American Journal of Botany*, **101**(10), pp.1711-1725.

Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, (**2**), pp.217-223. [R package]

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N.M., Salman-Minkov, A., Mayzel, J., Chay, O. and Mayrose, I., 2015. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytologist*, **206**(1), pp.19-26.

Rice, A., Šmarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., Meiri, S., Belmaker, J. and Mayrose, I., 2019. The global biogeography of polyploid plants. *Nature Ecology & Evolution*, **3**(2), pp.265-273.

Rich, T.C., 2006. Floristic changes in vascular plants in the British Isles: geographical and temporal variation in botanical activity 1836–1988. *Botanical Journal of the Linnean Society*, **152**(3), pp.303-330.

Rich, T.C. and Karran, A.B., 2006. Floristic changes in the British Isles: comparison of techniques for assessing changes in frequency of plants with time. *Botanical Journal of the Linnean Society*, **152**(3), pp.279-301.

Richardson, M.L. and Hanks, L.M., 2011. Differences in spatial distribution, morphology, and communities of herbivorous insects among three cytotypes of *Solidago altissima* (Asteraceae). *American Journal of Botany*, **98**(10), pp.1595-1601.

Richardson, D.M., Pyšek, P., Rejmanek, M., Barbour, M.G., Panetta, F.D. and West, C.J., 2000. Naturalization and invasion of alien plants: concepts and definitions. *Diversity and Distributions*, **6**(2), pp.93-107.

Rios, E.F., Kenworthy, K.E. and Munoz, P.R., 2015. Association of phenotypic traits with ploidy and genome size in annual ryegrass. *Crop Science*, **55**(5), pp.2078-2090.

Ritchie, P.D., Harper, A.B., Smith, G.S., Kahana, R., Kendon, E.J., Lewis, H., Fezzi, C., Halleck-Vega, S., Boulton, C.A., Bateman, I.J. and Lenton, T.M., 2019. Large changes in Great Britain's vegetation and agricultural land-use predicted under unmitigated climate change. *Environmental Research Letters*, **14**(11), pp.114012.

Robinson, R.A. and Sutherland, W.J., 2002. Post-war changes in arable farming and biodiversity in Great Britain. *Journal of Applied Ecology*, **39**(1), pp.157-176.

Roddy, A.B. *et al.,* 2020. The scaling of genome size and cell size limits maximum rates of photosynthesis with implications for ecological strategies. *International Journal of Plant Sciences*, **181**(1), pp.75-87.

Rohr, R.P., Saavedra, S., Peralta, G., Frost, C.M., Bersier, L.F., Bascompte, J. and Tylianakis, J.M., 2016. Persist or produce: a community trade-off tuned by species evenness. *The American Naturalist*, **188**(4), pp.411-422.

Roper, C. 2015. OSGB Grids - A collection of Ordance Survey National Grids in Shapefile (.shp) [OSGB 1936] and GeoJSON (.geojson) [WGS84] formats. https://github.com/charlesroper/OSGB_Grids [accessed March 2019]

Royal Horticultural Society Gardening, 2022. Mediterranean garden plants. https://www.rhs.org.uk/garden-design/design-with-plants/mediterranean-garden-plants [accessed in November 2022]

Sandel, B., Arge, L., Dalsgaard, B., Davies, R.G., Gaston, K.J., Sutherland, W.J. and Svenning, J.C., 2011. The influence of Late Quaternary climate-change velocity on species endemism. *Science*, **334**(6056), pp.660-664.

Sandve, S.R., Rohlfs, R.V. and Hvidsten, T.R., 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics*, **50**(7), pp.908-909.

Sardans, J. and Peñuelas, J., 2012. The role of plants in the effects of global change on nutrient availability and stoichiometry in the plant-soil system. *Plant Physiology*, **160**(4), pp.1741-1761.

Sattler, M.C., Carvalho, C.R. and Clarindo, W.R., 2016. The polyploidy and its key role in plant breeding. *Planta*, **243**(2), pp.281-296.

Schaffers, A.P. and Sýkora, K.V., 2000. Reliability of Ellenberg indicator values for moisture, nitrogen and soil reaction: a comparison with field measurements. *Journal of Vegetation Science*, **11**(2), pp.225-244.

Schleuning, M., Neuschulz, E.L., Albrecht, J., Bender, I.M., Bowler, D.E., Dehling, D.M., Fritz, S.A., Hof, C., Mueller, T., Nowak, L., Sorensen, M.C., Böhning-Gaese and K., Kissling, W.D., 2020. Trait-based assessments of climate-change impacts on interacting species. *Trends in Ecology & Evolution*, **35**(4), pp.319-328.

Schley, R.J., Pellicer, J., Ge, X.J., Barrett, C., Bellot, S., Guignard, M.S., Novák, P., Suda, J., Fraser, D., Baker, W.J., Dodsworth, S., Macas, J., Leitch, A.R., Leitch, I.J., 2022. The ecology of palm genomes: repeat-associated genome size expansion is constrained by aridity. *New Phytologist*, **236**, pp.433-446.

Schmidt, J.P. and Drake, J.M., 2011. Time since introduction, seed mass, and genome size predict successful invaders among the cultivated vascular plants of Hawaii. *PLoS One*, **6**(3), e17391.

Schubert, I. and Vu, G.T., 2016. Genome stability and evolution: attempting a holistic view. *Trends in Plant Science*, **21**(9), pp.749-757.

Sharma, A., 2020. The wicked problem of diffuse nutrient pollution from agriculture. *Journal of Environmental Law*, **32**(3), pp.471-502.

Shipley, B., 2016. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R*. Cambridge University Press.

Shipley, B., 2000. A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling, **7**(2),* pp.206-218.

Simonin, K.A. and Roddy, A.B., 2018. Genome downsizing, physiological novelty, and the global dominance of flowering plants. *PLoS Biology*, **16**(1), e2003706.

Šímová, I. and Herben, T., 2012. Geometrical constraints in the scaling relationships between genome size, cell size and cell cycle length in herbaceous plants. *Proceedings of the Royal Society. Series B. Biological Sciences*, **279**(1730), pp.867-875.

Small, D. and Fabel, D., 2016. Was Scotland deglaciated during the younger Dryas? *Quaternary Science Reviews*, **145**, pp.259-263.

Šmarda, P., Hejcman, M., Březinová, A., Horova, L., Steigerova, H., Zedek, F., Bureš, P., Hejcmanova, P. and Schellberg, J., 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 Phytologist*, **200**(3), pp.911-921.

Šmarda, P., Knápek, O., Březinová, A., Horová, L., Grulich, V., Danihelka, J., Veselý, P., Šmerda, J., Rotreklová, O. and Bureš, P., 2019. Genome sizes and genomic guanine+ cytosine (GC) contents of the Czech vascular flora with new estimates for 1700 species. *Preslia*, **91**(2), pp.117-142.

Smart, S.M., Clarke, R.T., van de Poll, H.M., Robertson, E.J., Shield, E.R., Bunce, R.G.H. and Maskell, L.C., 2003. National-scale vegetation change across Britain; an analysis of sample-based surveillance data from the Countryside Surveys of 1990 and 1998. *Journal of Environmental Management*, **67**(3), pp.239-254.

Smith, S.A. and Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, **105**(3), pp.302-314.

Soltis, P.S. and Soltis, D.E., 2000. The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences*, **97**(13), pp.7051-7057.

Sotherton, N.W., 1998. Land use changes and the decline of farmland wildlife: an appraisal of the set-aside approach. *Biological Conservation,* **83**(3), pp.259-268.

South, A., 2017. rnaturalearth: world map data from Natural Earth. [R package version 0.1.0.] *The R Foundation. https://CRAN. R-project. org/package= rnaturalearth*.

Sparrow, A.H. and Miksche, J.P., 1960. The relationship between chromosomal, nuclear, or cellular size, and the radiosensitivity of different plant taxa. *Radiation Research* 1960 **12**, pp.474.

Sparrow, A.H. and Miksche, J.P., 1961. Correlation of nuclear volume and DNA content with higher plant tolerance to chronic radiation. *Science*, **134**(3474), pp.282-283.

Stace, C., 2019. *New Flora of the British Isles – Fourth Edition*. C&M Floristics.

Stace, C.A., Preston, C.D. and Pearman, D.A., 2015. *Hybrid flora of the British Isles*. Botanical Society of Britain and Ireland.

Stace, C. A. & Crawley, M. J., 2015. *Alien plants.* Collins New Naturalist Library, Book **129**. HarperCollins UK.

Stamp, L.D., 1931. The land utilisation survey of Britain. *Scottish Geographical Magazine,* **47**(3), pp.144-150.

Stamp, L.D., 1934. Land Utilisation Survey as a School and College Exercise. *Journal of Geography*, **33**(4), pp.121-130.

Sterner, R.W. and Elser, J.J., 2002. *Ecological stoichiometry: the biology of elements from molecules to the biosphere.* Princeton University Press

Stevens, C.J., Dise, N.B., Mountford, J.O. and Gowing, D.J., 2004. Impact of nitrogen deposition on the species richness of grasslands. *Science*, **303**(5665), pp.1876-1879.

Stevens, C.J., Payne, R.J., Kimberley, A. and Smart, S.M., 2016. How will the semi-natural vegetation of the UK have changed by 2030 given likely changes in nitrogen deposition? *Environmental Pollution*, **208**, pp.879-889.

Streiner, D.L., 2005. Finding our way: an introduction to path analysis. *The Canadian Journal of Psychiatry*, **50**(2), pp.115-122.

Stroh, P., Leach, S.J., August, T.A., Walker, K.J., Pearman, D.A., Rumsey, F.J., Harrower, C.A., Fay, M.F., Martin, J.P., Pankhurst, T. and Preston, C.D., 2014. A vascular plant red list for England. *BSBI News*, **9**, pp.1-193.

Suda, J., Meyerson, L.A., Leitch, I.J. and Pyšek, P., 2015. The hidden side of plant invasions: the role of genome size. *New Phytologist*, **205**(3), pp.994-1007.

Swift, H., 1950. The constancy of desoxyribose nucleic acid in plant nuclei. *Proceedings of the National Academy of Sciences*, **36**(11), pp.643-654.

Symonds, M.R. and Blomberg, S.P., 2014. A primer on phylogenetic generalised least squares. In *Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology*, pp. 105-130. Springer, Berlin, Heidelberg.

Tamme, R., Götzenberger, L., Zobel, M., Bullock, J.M., Hooftman, D.A., Kaasik, A. and Pärtel, M., 2014. Predicting species' maximum dispersal distances from simple plant traits. *Ecology*, **95**(2), pp.505-513.

Tank, D.C., Eastman, J.M., Pennell, M.W., Soltis, P.S., Soltis, D.E., Hinchliff, C.E., Brown, J.W., Sessa, E.B. and Harmon, L.J., 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist*, **207**(2), pp.454-467.

Tate, J.A., Soltis, D.E. and Soltis, P.S., 2005. Polyploidy in plants. In *The Evolution of the Genome*, pp. 371-426. Academic Press.

te Beest, M., Le Roux, J.J., Richardson, D.M., Brysting, A.K., Suda, J., Kubešová, M. and Pyšek, P., 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany*, ***109***(1), pp.19-45.

Telfer, M.G., Preston, C.D. and Rothery, P., 2002. A general method for measuring relative change in range size from biological atlas data. *Biological Conservation*, **107**(1), pp.99-109.

Temsch, E.M., Temsch, W., Ehrendorfer-Schratt, L. and Greilhuber, J., 2010. Heavy metal mollution, selection, and genome size: The species of the Žerjav study revisited with flow cytometry. *Journal of Botany*, **2010**, pp.11.

Tennekes, M. and Ellis, P., 2017. Package 'treemap'. *R Package Version*, pp.2-4. [R package]

Tennekes, M., Nowosad, J., Gombin, J., Jeworutzki, S., Russell, K., Zijdeman, R., Clouse, J., Lovelace, R. and Muenchow, J., 2022. Package 'tmap'. [R package]

Théroux-Rancourt, G., Roddy, A.B., Earles, J.M., Gilbert, M.E., Zwieniecki, M.A., Boyce, C.K., Tholen, D., McElrone, A.J., Simonin, K.A. and Brodersen, C.R., 2021. Maximum CO_2 diffusion inside leaves is limited by the scaling of cell size and genome size. *Proceedings of the Royal Society. Series B. Biological Sciences*, **288**(1945), pp.20203145.

The Angiosperm Phylogeny Group, 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**(1), pp.1-20.

Thomas, C.D. and Palmer, G., 2015. Non-native plants add to the British flora without negative consequences for native diversity. *Proceedings of the National Academy of Sciences*, **112**(14), pp.4387-4392.

Thompson, K. and McCarthy, M.A., 2008. Traits of British alien and native urban plants. *Journal of Ecology*, **96**(5), pp.853-859.

Thomson, F.J., Moles, A.T., Auld, T.D. and Kingsford, R.T., 2011. Seed dispersal distance is more strongly correlated with plant height than with seed mass. *Journal of Ecology*, **99**(6), pp.1299-1307.

Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehikoinen, A., de Jonge, M.M., Oksanen, J. and Ovaskainen, O., 2020. Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, **11**(3), pp.442-447.

Tilman, D., Clark, M., Williams, D.R., Kimmel, K., Polasky, S. and Packer, C., 2017. Future threats to biodiversity and pathways to their prevention. *Nature*, **546**(7656), pp.73-81.

Tomlinson, S.J., Carnell, E.J., Dore, A.J., Dragosits, U., 2020. Nitrogen deposition in the UK at 1km resolution, 1990-2017 NERC Environmental Information Data Centre. https://catalogue.ceh.ac.uk/documents/9b203324-6b37-4e91-b028-e073b197fb9f [downloaded June 2022]

Tomlinson, S.J., Carnell, E.J., Dore, A.J. and Dragosits, U., 2021. Nitrogen deposition in the UK at 1 km resolution from 1990 to 2017. *Earth System Science Data*, **13**(10), pp.4677-4692.

van de Peer, Y., Maere, S. and Meyer, A., 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, **10**(10), pp.725-732.

van de Peer, Y., Mizrachi, E. and Marchal, K., 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics*, **18**(7), pp.411-424.

van der Bijl, W., 2018. phylopath: Easy phylogenetic path analysis in R. *PeerJ*, **6**, e4718.

van der Bijl, W., 2022. phylopath: Perform phylogenetic path analysis. https://cran.r-project.org/web/packages/phylopath/phylopath.pdf [R package]

van Kleunen, M. *et al.*, 2018. The changing role of ornamental horticulture in alien plant invasions. *Biological Reviews*, **93**(3), pp.1421-1437.

van Kleunen, M., Weber, E. and Fischer, M., 2010. A meta-analysis of trait differences between invasive and non-invasive plant species. *Ecology Letters*, **13**(2), pp.235-245.

Van't Hof, J. and Sparrow, A.H., 1963. A relationship between DNA content, nuclear volume, and minimum mitotic cycle time. *Proceedings of the National Academy of Sciences*, **49**(6), pp.897-902.

Veselý, P., Bureš, P. and Šmarda, P., 2013. Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives. *Annals of Botany*, **112**(6), pp.1193-1200.

Veselý, P., Bureš, P., Šmarda, P. and Pavlíček, T., 2012. Genome size and DNA base composition of geophytes: the mirror of phenology and ecology? *Annals of Botany*, **109**(1), pp.65-75.

Veselý, P., Šmarda, P., Bureš, P., Stirton, C., Muasya, A.M., Mucina, L., Horová, L., Veselá, K., Šilerová, A., Šmerda, J. and Knápek, O., 2020. Environmental pressures on stomatal size may drive plant genome size evolution: evidence from a natural experiment with Cape geophytes. *Annals of Botany*, **126**(2), pp.323-330.

Vesk, P.A., 2013. How traits determine species responses to environmental gradients. *Journal of Vegetation Science*, **24**(6), pp.977-978.

Vesk, P.A., Morris, W.K., Neal, W.C., Mokany, K. and Pollock, L.J., 2021. Transferability of trait-based species distribution models. *Ecography*, **44**(1), pp.134-147.

Vidic, T., Greilhuber, J., Vilhar, B. and Dermastia, M., 2009. Selective significance of genome size in a plant community with heavy metal pollution. *Ecological Applications*, **19**(6), pp.1515-1521.

Vinogradov, A.E., 2003. Selfish DNA is maladaptive: evidence from the plant Red List. *Trends in Genetics*, **19**(11), pp.609-614.

Violle, C. *et al.*, 2015. Vegetation ecology meets ecosystem science: Permanent grasslands as a functional biogeography case study. *Science of the Total Environment*, **534**, pp.43-51.

Violle, C., Garnier, E., Lecoeur, J., Roumet, C., Podeur, C., Blanchard, A. and Navas, M.L., 2009. Competition, traits and resource depletion in plant communities. *Oecologia*, **160**(4), pp.747-755.

Vitousek, P.M., Porder, S., Houlton, B.Z. and Chadwick, O.A., 2010. Terrestrial phosphorus limitation: mechanisms, implications, and nitrogen–phosphorus interactions. *Ecological Applications*, **20**(1), pp.5-15.

Walczyk, A.M. and Hersch-Green, E.I., 2019. Impacts of soil nitrogen and phosphorus levels on cytotype performance of the circumboreal herb *Chamerion angustifolium*: implications for polyploid establishment. *American Journal of Botany*, **106**(7), pp.906-921.

Walker, K., Pearman, D. Ellis B., McIntosh J. and Lockton, A., 2010. Recording the British and Irish flora 2010-2020. *Botanical Society of the British Isles, London.*

Wang, X., Morton, J.A., Pellicer, J., Leitch, I.J. and Leitch, A.R., 2021. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *The Plant Journal*, **107**(4), pp.1003-1015.

Watson, H.C., 1883. *Topographical botany: being local and personal records towards shewing the distribution of British plants traced through the 112 counties and vice-counties of England, Wales, and Scotland*. B. Quaritch.

Watts, G. *et al.*, 2015. Climate change and water in the UK–past changes and future prospects. *Progress in Physical Geography*, **39**(1), pp.6-28.

WCVP. World Checklist of Vascular Plants, version 2.0, 2022. Facilitated by the Royal Botanic Gardens, Kew. http://wcvp.science.kew.org/ [retrieved 10 June 2021]

Webb, D.A., 1980. The biological vice-counties of Ireland. *Proceedings of the Royal Irish Academy. Section B. Biological, Geological, and Chemical Science*, pp. 179-196. Royal Irish Academy.

Webb, D.A., 1985. What are the criteria for presuming native status? *Watsonia*, **15**(3), pp.231-236.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y. and Zemla, J., 2017. Package 'corrplot'. *Statistician*, **56**(316), e24. [R package]

Wendel, J.F., 2015. The wondrous cycles of polyploidy in plants. *American Journal of Botany* **102**(11), pp.1753-1756.

Westoby, M., 1998. A leaf-height-seed (LHS) plant ecology strategy scheme. *Plant and Soil*, **199**(2), pp.213-227.

Whitney, K.D., Ahern, J.R., Campbell, L.G., Albert, L.P. and King, M.S., 2010. Patterns of hybridization in plants. *Perspectives in Plant Ecology, Evolution and Systematics*, **12**(3), pp.175-182.

Wickham, H. *et al.*, 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, **4**(43), pp.1686.

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H. and Dunnington, D., 2016. Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version, **2**(1), pp.1-189. [R package]

Wild, J. et al., 2019. Plant distribution data for the Czech Republic integrated in the Pladias database. *Preslia*, **91**(1), pp.1-24.

Wilson, M.J., Fradera-Soler, M., Summers, R., Sturrock, C.J. and Fleming, A.J., 2021. Ploidy influences wheat mesophyll cell geometry, packing and leaf function. *Plant Direct*, **5**(4), e00314.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B. and Rieseberg, L.H., 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, **106**(33), pp.13875-13879.

Zanne, A.E. *et al.*, 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature*, **506**(7486), pp.89-92.

Zatelli, P., Gobbi, S., Tattoni, C., La Porta, N. and Ciolli, M., 2019. Object-based image analysis for historic maps classification. *International Archives Of The Photogrammetry, Remote Sensing And Spatial Information Sciences*, **XLII-4/W14**, pp.247-254.

Zenil-Ferguson, R., Ponciano, J.M. and Burleigh, J.G., 2016. Evaluating the role of genome downsizing and size thresholds from genome size distributions in angiosperms. *American Journal of Botany*, **103**(7), pp.1175-1186.

Zhu, H., 2019. *KableExtra*: Construct complex table with 'kable' and pipe syntax. [R package version], **1**(0).

Zizka, A., Antonelli, A. and Silvestro, D., 2021. *Sampbias*, a method for quantifying geographic sampling biases in species distribution data. *Ecography*, **44**(1), pp.25-32.

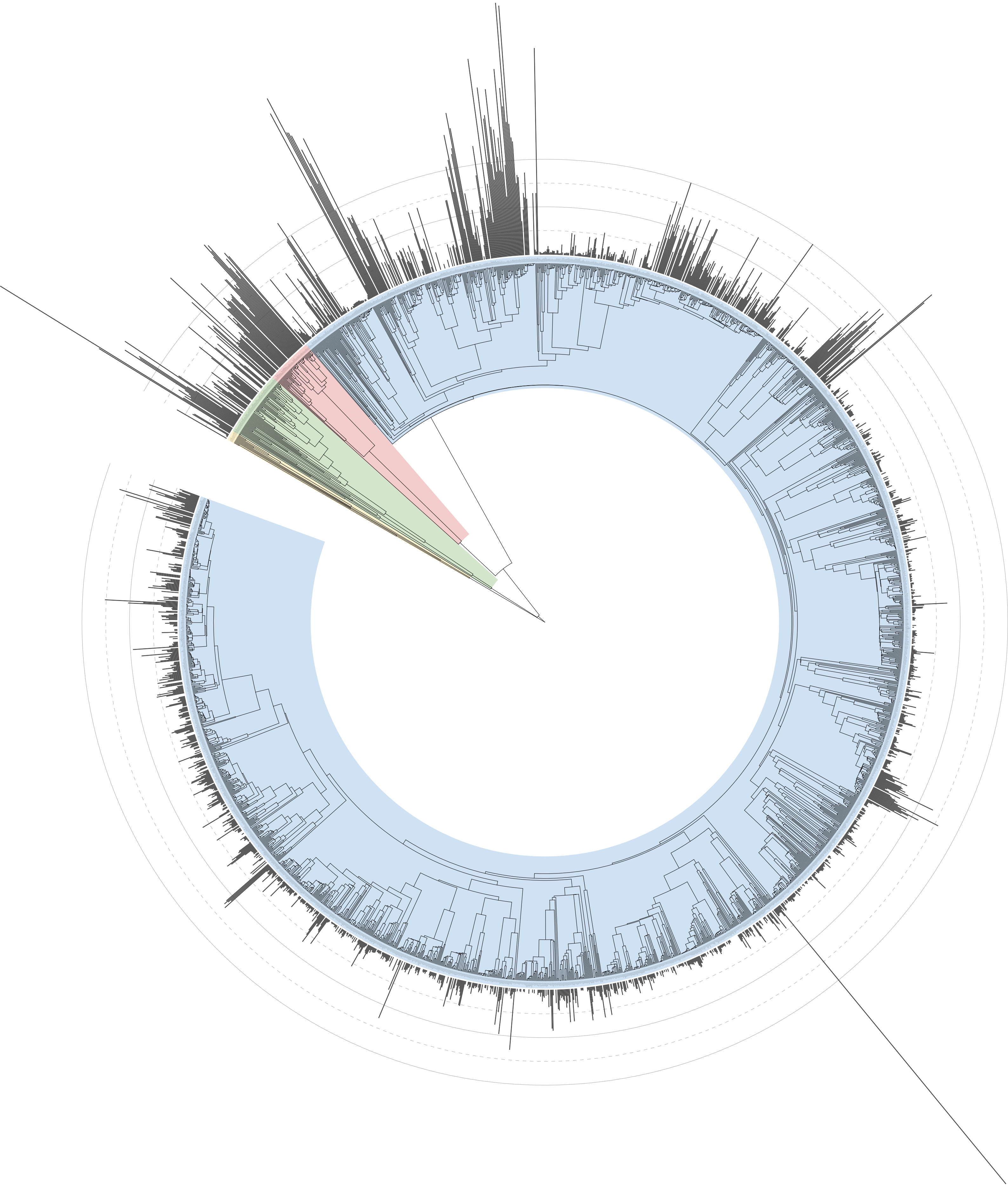
Zonneveld, B.J., 2019. The DNA weights per nucleus (genome size) of more than 2350 species of the Flora of The Netherlands, of which 1370 are new to science, including the pattern of their DNA peaks. In *Forum Geobotanicum* **8**, pp.24-78.

Appendix 1 Supporting information for Chapter 2

Supporting Tables

Supporting Figure

Supporting Method



Supporting Figure S2.1 High resolution phylogeny of the British and Irish vascular flora. The circular representation of 2,501 species with phylogenetic information includes colour coding for the different clades, with Lycophytes in yellow, Monilophytes coded in green, gymnosperms in red and angiosperms overlaid in blue. The smallest known genome size for each species is plotted around the outside in pg/1C with gridlines at 5, 10, 15 and 20 pg for orientation. Lycophytes, Monilophytes and gymnosperms have larger genome sizes overall, but the overwhelmingly largest genome of the flora, that of *Viscum album* L., an angiosperm, is visible on the bottom right with a genome size of 88.90 pg/1C.

Appendix 2 Supporting information for Chapter 3

All Supporting Tables and Figures in Appendix 2 are available online at https://github.com/mariehenniges/BI_flora_thesis_appendices.

Supporting Tables

Supporting Table S3.1 Dataframe containing the results of linear regressions on time factors across three date classes (1987-1999, 2000-2009, 2010-2019).

Supporting Table S3.2a Dataframe listing time factors calculated for each species (1987-1999).

Supporting Table S3.2b Location report for each hectad (1987-1999).

Supporting Table S3.2c Listing of rescaled species frequencies per hectad (1987-1999).

Supporting Table S3.3a Dataframe listing time factors calculated for each species (2000-2009).

Supporting Table S3.3b Location report for each hectad (2000-2009).

Supporting Table S3.3c Listing of rescaled species frequencies per hectad (2000-2009).

Supporting Table S3.4a Dataframe listing time factors calculated for each species (2010-2019).

Supporting Table S3.4b Location report for each hectad (2010-2019).

Supporting Table S3.4c Listing of rescaled species frequencies per hectad (2010-2019).

Supporting Figure

Supporting Figure S3.1 Maps and regression graphs of Frescalo outputs across three date classes (1987-1999, 2000-2009, 2010-2019).

Appendix 3 Supporting information for Chapter 4

Supporting Tables

Supporting Figures

Supporting Methods

Supporting Table S4.4 Relative importance of predictors in linear models of change in hectad weighted mean genome size by land use in the last date class. Lmg is the metric of variable importance used (Lindemann, Merenda & Gold, 1980) and describes the variance explained by each predictor, summing to the total R² of each model.

Acid grassland	lmg
change in species richness	0.1427
change in rainfall	0.0053
change in dry N deposition	0.1128
change in wet N deposition	0.1341
latitude	0.0541
longitude	0.1069

Bog	lmg
change in species richness	0.4003
change in rainfall	0.0113
change in dry N deposition	0.0321
change in wet N deposition	0.0708
latitude	0.0561
longitude	0.0060

Heather	lmg
change in species richness	0.3332
change in rainfall	0.0875
change in dry N deposition	0.0313
change in wet N deposition	0.0697
latitude	0.0445
longitude	0.0450

Arable and horticulture	lmg
change in species richness	0.5154
change in rainfall	0.0044
change in dry N deposition	0.0284
change in wet N deposition	0.0115
latitude	0.0107
longitude	0.0275

Coniferous woodland	lmg
change in species richness	0.5066
change in rainfall	0.0139
change in dry N deposition	0.0251
change in wet N deposition	0.0963
latitude	0.1498
longitude	0.0407

Heather grassland	lmg
change in species richness	0.5150
change in rainfall	0.0157
change in dry N deposition	0.0399
change in wet N deposition	0.0219
latitude	0.0848
longitude	0.0312

Improved grassland	lmg
change in species richness	0.5765
change in rainfall	0.0100
change in dry N deposition	0.0565
change in wet N deposition	0.0105
latitude	0.0196
longitude	0.0586

Saltwater	lmg
change in species richness	0.3693
change in rainfall	0.0308
change in dry N deposition	0.0256
change in wet N deposition	0.0789
latitude	0.1822
longitude	0.2395

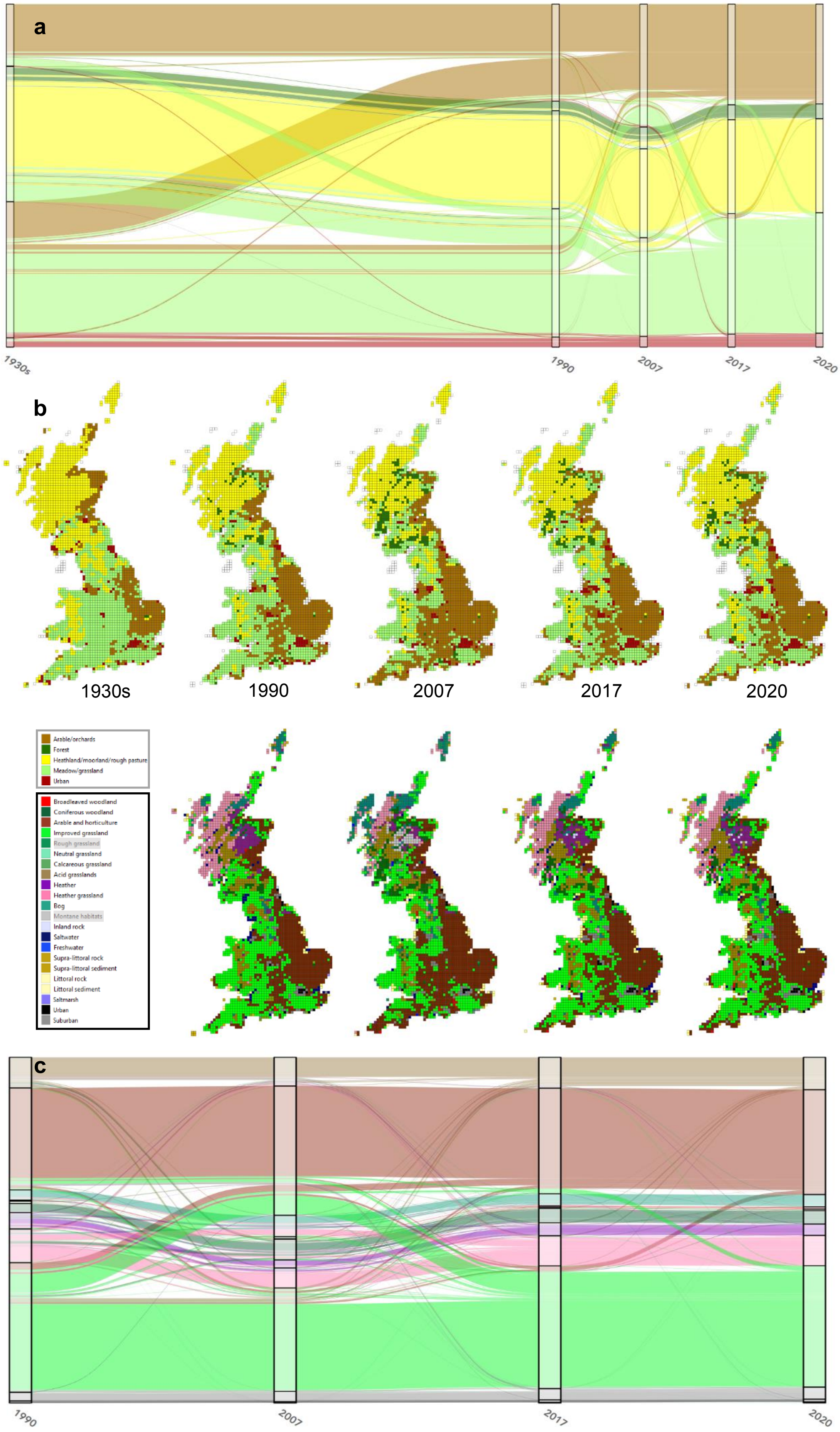
Suburban	lmg
change in species richness	0.6287
change in rainfall	0.0192
change in dry N deposition	0.1308
change in wet N deposition	0.0057
latitude	0.0235
longitude	0.0042

Littoral sediment	lmg
change in species richness	0.2603
change in rainfall	0.0149
change in dry N deposition	0.0260
change in wet N deposition	0.0167
latitude	0.0098
longitude	0.0084

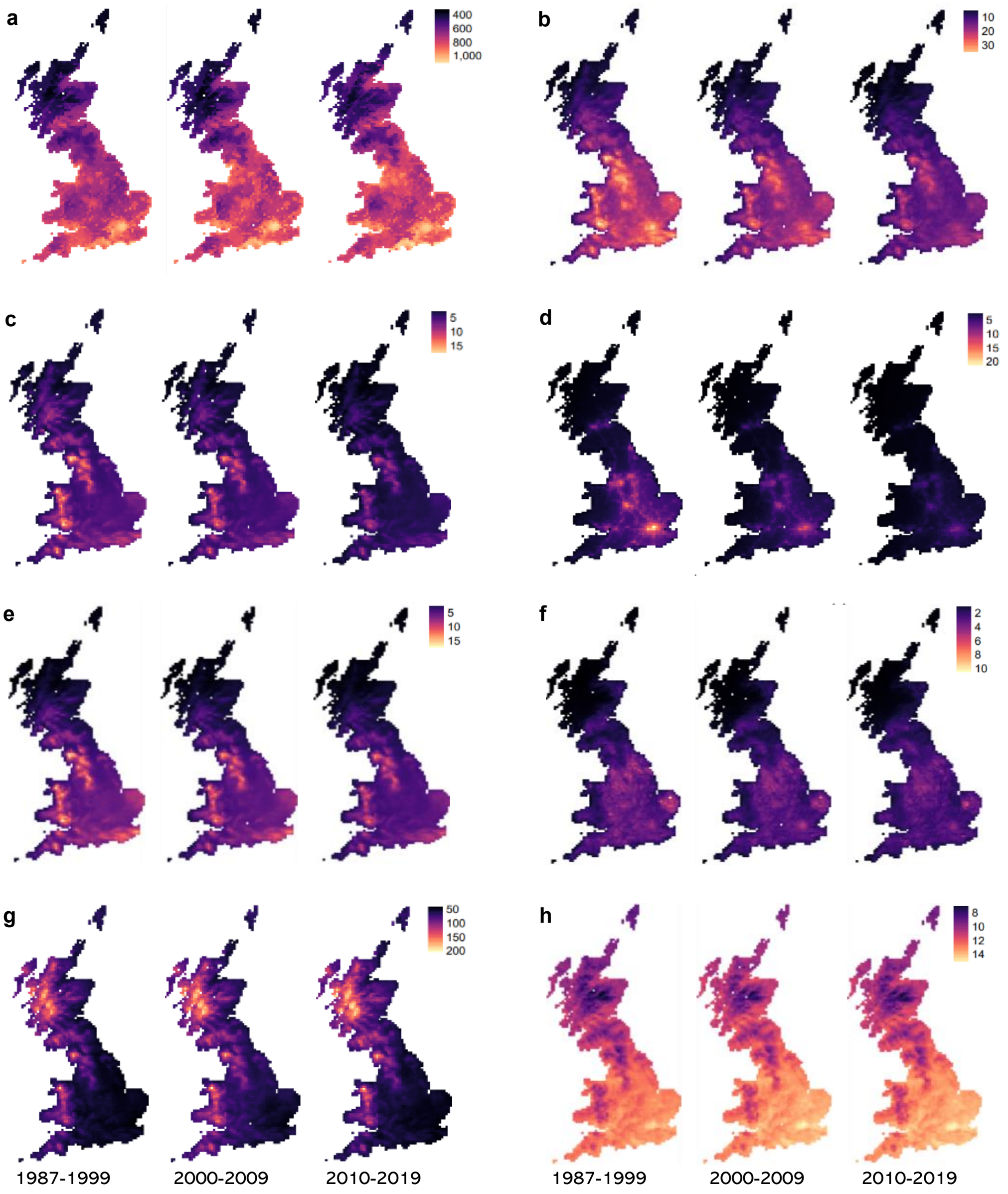
Urban	lmg
change in species richness	0.4622
change in rainfall	0.0665
change in dry N deposition	0.1052
change in wet N deposition	0.0351
latitude	0.0254
longitude	0.0313

Supporting Table S4.5 Tukey post-hoc test results for comparison of mean weighted genome sizes per hectad between the different land cover types.

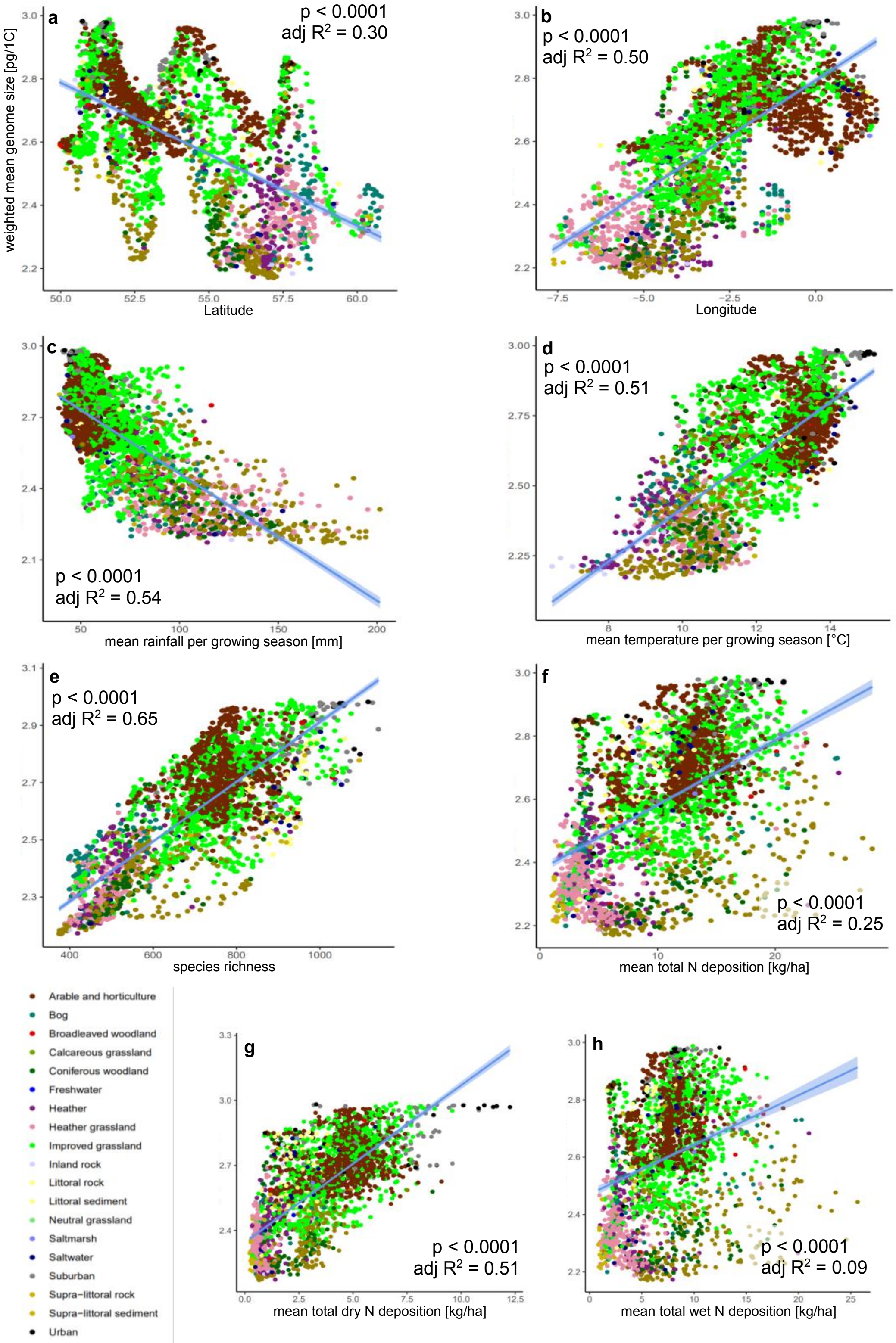
group1	group2	null.value	estimate	conf.low	conf.high	p.adj	p.adj.signif
Acid grassland	Arable and horticulture	0	0.3638	0.3305	0.3971	0.0000	****
Acid grassland	Bog	0	0.0688	0.0145	0.1231	0.0019	**
Acid grassland	Coniferous woodland	0	0.0537	0.0005	0.1068	0.0457	*
Acid grassland	Heather	0	0.0346	-0.0153	0.0845	0.5170	ns
Acid grassland	Heather grassland	0	-0.0131	-0.0539	0.0277	0.9980	ns
Acid grassland	Improved grassland	0	0.2945	0.2618	0.3272	0.0000	****
Acid grassland	Littoral sediment	0	0.3321	0.2605	0.4038	0.0000	****
Acid grassland	Saltwater	0	0.2006	0.1091	0.2920	0.0000	****
Acid grassland	Suburban	0	0.4507	0.3923	0.5091	0.0000	****
Acid grassland	Supra-littoral rock	0	0.0284	-0.0866	0.1433	1.0000	ns
Acid grassland	Supra-littoral sediment	0	0.0231	-0.0919	0.1381	1.0000	ns
Acid grassland	Urban	0	0.4607	0.3587	0.5627	0.0000	****
Arable and horticulture	Bog	0	-0.2950	-0.3437	-0.2463	0.0000	****
Arable and horticulture	Coniferous woodland	0	-0.3101	-0.3576	-0.2627	0.0000	****
Arable and horticulture	Heather	0	-0.3292	-0.3729	-0.2855	0.0000	****
Arable and horticulture	Heather grassland	0	-0.3769	-0.4099	-0.3439	0.0000	****
Arable and horticulture	Improved grassland	0	-0.0693	-0.0915	-0.0471	0.0000	****
Arable and horticulture	Littoral sediment	0	-0.0317	-0.0992	0.0359	0.9430	ns
Arable and horticulture	Saltwater	0	-0.1632	-0.2514	-0.0750	0.0000	****
Arable and horticulture	Suburban	0	0.0869	0.0337	0.1401	0.0000	****
Arable and horticulture	Supra-littoral rock	0	-0.3354	-0.4479	-0.2230	0.0000	****
Arable and horticulture	Supra-littoral sediment	0	-0.3407	-0.4531	-0.2282	0.0000	****
Arable and horticulture	Urban	0	0.0969	-0.0022	0.1961	0.0626	ns
Bog	Coniferous woodland	0	-0.0152	-0.0791	0.0488	1.0000	ns
Bog	Heather	0	-0.0342	-0.0954	0.0270	0.8220	ns
Bog	Heather grassland	0	-0.0819	-0.1360	-0.0279	0.0000	****
Bog	Improved grassland	0	0.2257	0.1774	0.2739	0.0000	****
Bog	Littoral sediment	0	0.2633	0.1833	0.3433	0.0000	****
Bog	Saltwater	0	0.1317	0.0337	0.2298	0.0006	***
Bog	Suburban	0	0.3819	0.3135	0.4502	0.0000	****
Bog	Supra-littoral rock	0	-0.0405	-0.1608	0.0799	0.9960	ns
Bog	Supra-littoral sediment	0	-0.0457	-0.1660	0.0746	0.9890	ns
Bog	Urban	0	0.3919	0.2839	0.4999	0.0000	****
Coniferous woodland	Heather	0	-0.0190	-0.0793	0.0412	0.9980	ns
Coniferous woodland	Heather grassland	0	-0.0668	-0.1198	-0.0138	0.0021	**
Coniferous woodland	Improved grassland	0	0.2408	0.1938	0.2879	0.0000	****
Coniferous woodland	Littoral sediment	0	0.2785	0.1992	0.3577	0.0000	****
Coniferous woodland	Saltwater	0	0.1469	0.0495	0.2444	0.0000	****
Coniferous woodland	Suburban	0	0.3970	0.3296	0.4645	0.0000	****
Coniferous woodland	Supra-littoral rock	0	-0.0253	-0.1451	0.0945	1.0000	ns
Coniferous woodland	Supra-littoral sediment	0	-0.0305	-0.1504	0.0893	1.0000	ns
Coniferous woodland	Urban	0	0.4071	0.2996	0.5145	0.0000	****
Heather	Heather grassland	0	-0.0477	-0.0974	0.0019	0.0735	ns
Heather	Improved grassland	0	0.2599	0.2167	0.3031	0.0000	****
Heather	Littoral sediment	0	0.2975	0.2205	0.3746	0.0000	****
Heather	Saltwater	0	0.1660	0.0703	0.2617	0.0000	****
Heather	Suburban	0	0.4161	0.3512	0.4810	0.0000	****
Heather	Supra-littoral rock	0	-0.0063	-0.1247	0.1122	1.0000	ns
Heather	Supra-littoral sediment	0	-0.0115	-0.1299	0.1069	1.0000	ns
Heather	Urban	0	0.4261	0.3203	0.5320	0.0000	****
Heather grassland	Improved grassland	0	0.3076	0.2753	0.3400	0.0000	****
Heather grassland	Littoral sediment	0	0.3453	0.2737	0.4168	0.0000	****
Heather grassland	Saltwater	0	0.2137	0.1224	0.3050	0.0000	****
Heather grassland	Suburban	0	0.4638	0.4056	0.5220	0.0000	****
Heather grassland	Supra-littoral rock	0	0.0415	-0.0734	0.1564	0.9930	ns
Heather grassland	Supra-littoral sediment	0	0.0362	-0.0786	0.1511	0.9980	ns
Heather grassland	Urban	0	0.4739	0.3720	0.5758	0.0000	****
Improved grassland	Littoral sediment	0	0.0376	-0.0296	0.1049	0.8200	ns
Improved grassland	Saltwater	0	-0.0939	-0.1819	-0.0060	0.0241	*
Improved grassland	Suburban	0	0.1562	0.1034	0.2090	0.0000	****
Improved grassland	Supra-littoral rock	0	-0.2661	-0.3784	-0.1539	0.0000	****
Improved grassland	Supra-littoral sediment	0	-0.2714	-0.3836	-0.1591	0.0000	****
Improved grassland	Urban	0	0.1662	0.0673	0.2652	0.0000	****
Littoral sediment	Saltwater	0	-0.1316	-0.2402	-0.0229	0.0041	**
Littoral sediment	Suburban	0	0.1186	0.0357	0.2014	0.0002	***
Littoral sediment	Supra-littoral rock	0	-0.3038	-0.4329	-0.1747	0.0000	****
Littoral sediment	Supra-littoral sediment	0	-0.3090	-0.4381	-0.1799	0.0000	****
Littoral sediment	Urban	0	0.1286	0.0109	0.2463	0.0181	*
Saltwater	Suburban	0	0.2501	0.1497	0.3505	0.0000	****
Saltwater	Supra-littoral rock	0	-0.1722	-0.3132	-0.0312	0.0036	**
Saltwater	Supra-littoral sediment	0	-0.1775	-0.3185	-0.0364	0.0022	**
Saltwater	Urban	0	0.2602	0.1295	0.3908	0.0000	****
Suburban	Supra-littoral rock	0	-0.4223	-0.5446	-0.3001	0.0000	****
Suburban	Supra-littoral sediment	0	-0.4276	-0.5498	-0.3053	0.0000	****
Suburban	Urban	0	0.0100	-0.1001	0.1202	1.0000	ns
Supra-littoral rock	Supra-littoral sediment	0	-0.0052	-0.1626	0.1521	1.0000	ns
Supra-littoral rock	Urban	0	0.4324	0.2843	0.5805	0.0000	****
Supra-littoral sediment	Urban	0	0.4376	0.2895	0.5857	0.0000	****


a
1930s
1990
2007
2017
2020
b
1930s
1990
2007
2017
2020
Arable/orchards
Forest
Heathland/moorland/rough pasture
Meadow/grassland
Urban
Broadleaved woodland
Coniferous woodland
Arable and horticulture
Improved grassland
Rough grassland
Neutral grassland
Calcareous grassland
Acid grasslands
Heather
Heather grassland
Bog
Montane habitats
Inland rock
Saltwater
Freshwater
Supra-littoral rock
Supra-littoral sediment
Littoral rock
Littoral sediment
Saltmarsh
Urban
Suburban
c
1990
2007
2017
2020

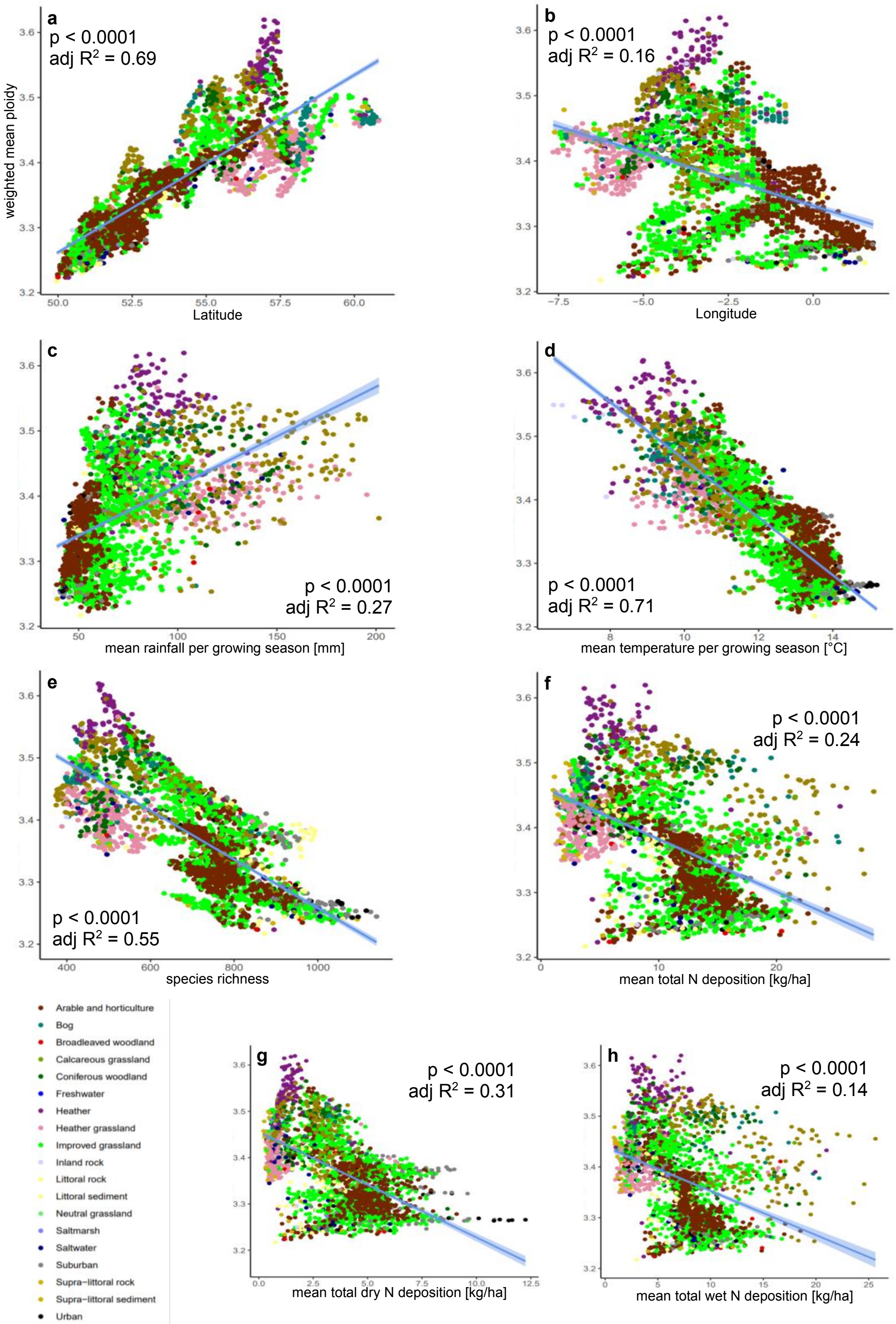
Supporting Figure S4.1 Detailed overview of land cover changes. a and c are more detailed alluvial plots that visualise the fate of each hectad with land cover data at each of the time points considered here (1930s, 1990, 2007, 2017 and 2020). While a shows broader categories of land cover, the plot in c shows far more detailed land cover information, but does not allow comparison with the 1930s. b represents the land cover by hectad categorised according to the Dudley Stamp 1930s map (top) and UKCEH land cover maps (bottom). 2007 represents a special case, since the UKCEH land cover map's categories for this period are not perfectly aligned with those used in the preceding and following years, making direct comparisons more challenging. Hectads with majority cover for one of those land cover types that were not assigned in all time periods are not included in alluvial plots. The land cover types only present in the 2007 LCM are highlighted in grey in the legend. Legends (inside the grey box for Dudley Stamp categories, inside the black box for UKCEH categories) are valid for maps and alluvial plots. Colours in alluvial plots indicate the majority cover the hectad falls into in the final date class.



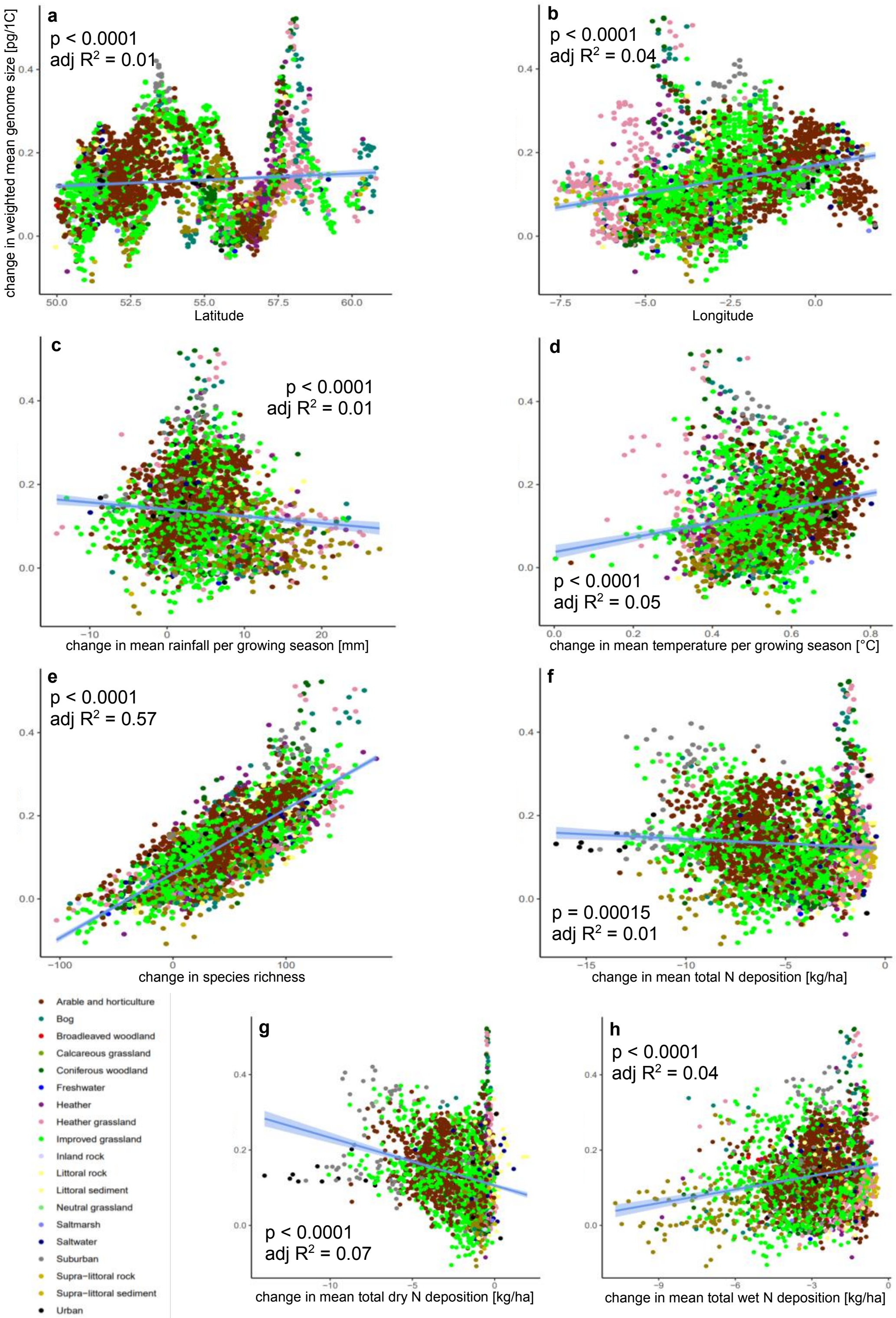
Supporting Figure S4.2 Map representations of different predictor variables used in spatial models. Species richness following Frescalo correction (a). Mean total nitrogen deposition [kg/ha] (b). Mean NOy wet deposition [kg/ha] (c). Mean NOy dry deposition [kg/ha] (d). Mean NHx wet deposition [kg/ha] (e). Mean NHx dry deposition [kg/ha] (f). Mean annual rainfall per growing season [mm] (g). Mean annual temperature per growing season [°C] (h).



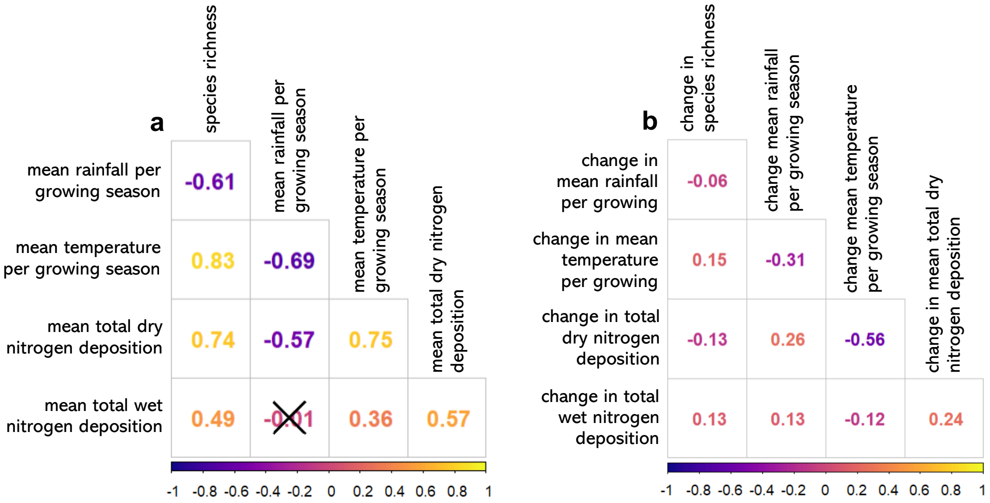
Supporting Figure S4.3 Linear models of weighted mean genome size per hectad by different spatial parameters for the final date class (2010-2019). Each plot illustrates the relationship between weighted mean genome size [pg/1C] by hectad and one spatial parameter (latitude (a), longitude (b), mean rainfall per growing season [mm] (c), mean temperature per growing season [°C] (d), species number per hectad after Frescalo correction (e), mean total nitrogen deposition [kg/ha] (f), mean total dry and wet nitrogen deposition [kg/ha] (g and h)). p-values and adjusted R² are given with each plot. Each dot represents a hectad with colours indicating the majority land cover type present there in 2017 (see legend).



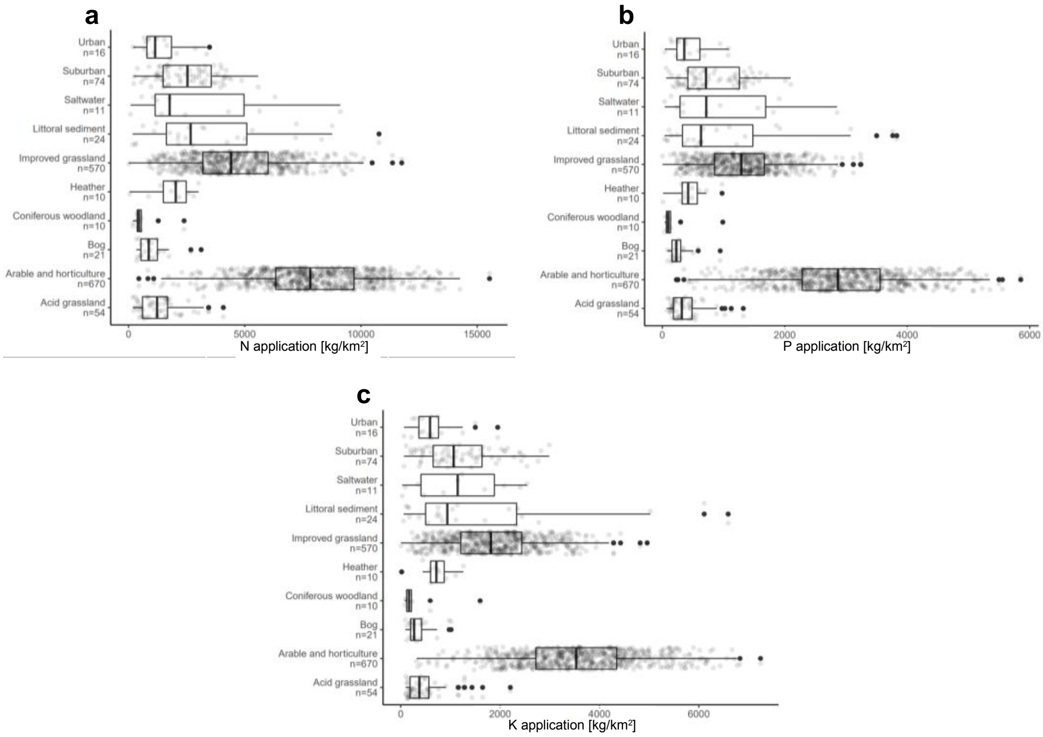
Supporting Figure S4.4 Linear models of weighted mean ploidy per hectad by different spatial parameters for the final date class (2010-2019). Each plot illustrates the relationship between weighted mean ploidy by hectad and one spatial parameter (latitude (a), longitude (b), mean rainfall per growing season [mm] (c), mean temperature per growing season [°C] (d), species number per hectad after Frescalo correction (e), mean total nitrogen deposition [kg/ha] (f), mean total dry and wet nitrogen deposition [kg/ha] (g and h)). p-values and adjusted R² are given with each plot. Each dot represents a hectad with colours indicating the majority land cover type present there in 2017 (see legend).



Supporting Figure S4.5 Linear models of change in weighted mean genome size per hectad by changes in different spatial parameters between 1987-1999 and 2010-2019. Each plot illustrates the relationship between change in weighted mean genome size [pg/1C] by hectad and one spatial parameter (latitude (a), longitude (b), change in mean rainfall per growing season [mm] (c), change in mean temperature per growing season [°C] (d), change in species number per hectad after Frescalo correction (e), change in mean total nitrogen deposition [kg/ha] (f), change in mean total dry and wet nitrogen deposition [kg/ha] (g and h)). p-values and adjusted R² are given with each plot. Each dot represents a hectad with colours indicating the majority land cover type present there in 2017 (see legend).



Supporting Figure S4.6 Pearson correlation metrics for predictors used in linear models. a shows correlations between predictors in the model for weighted mean genome size per hectad in the last date class (1987-1999), b shows correlations between predictors for the model for change in weighted mean genome size per hectad. Numbers and colours indicate the strength and direction of the correlation. Crossed out numbers are non-significant.

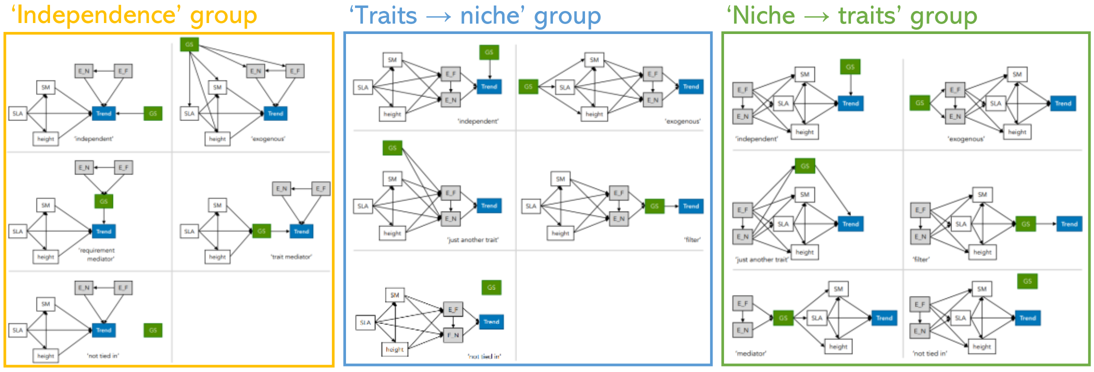


Supporting Figure S4.7 Representation of the amount of N (nitrogen, a), P (phosphorous, b) and K (potassium, c) applied to hectads of different land cover types. Information was derived from the datasets made available on https://www.ceh.ac.uk/data/ukceh-land-cover-plus-fertilisers-and-pesticides.

Appendix 4 Supporting information for Chapter 5

Supporting Figure

Supporting Tables



Supporting Figure S5.1 Hypothesised causal structures tested in path analysis represented as directed acyclic graphs. Based on previous PGLS analysis conducted on all traits, the hypotheses of causal interaction between functional traits (SM = seed mass, SLA = specific leaf area, height = canopy height), Ellenberg values (Ellenberg N = E_N and Ellenberg F = E_F) and genome size (GS) and fall into three broad categories. The first category of proposed models assumes that genome size, functional traits and Ellenberg values each have separate effects on trend ('independence'), the second category ('traits > niche) postulates that the functional traits influence the realised niche requirements (Ellenberg values) of a plant and the third category ('niche > traits') assumes that the characterisation of the niche a plant occurs in effects functional traits which in turn has impacts on trend. Within each category, the exact way in which genome size ties into the proposed network is changed to gain insights into the way it might affect trend. Names given with each path diagram describe the role that genome size would be expected to play if this model received support from phylogenetic path analysis.

Supporting Table S5.1 Detailed results from ten independent random forest runs on ten random subsets of data. For each run, the table details the hyperparameters derived from grid-based tuning, measures of the success of the algorithm overall (accuracy, ROC AUC and the OOB (out of bag) prediction error as well as confusion matrices and derived from them the percentage of correctly identified decreasing and increasing species.

Run	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Hyper-parameters	mtry=4 trees=500 min_n=6	mtry=2 trees=1000 min_n=2	mtry=2 trees=1000 min_n=2	mtry=2 trees=2000 min_n=2	mtry=6 trees=500 min_n=2	mtry=2 trees=500 min_n=2	mtry=6 trees=1000 min_n=10	mtry=2 trees=500 min_n=2	mtry=4 trees=500 min_n=2	mtry=6 trees=500 min_n=6
Accuracy	0.696	0.717	0.708	0.742	0.658	0.708	0.692	0.654	0.675	0.729
ROC AUC	0.702	0.723	0.776	0.793	0.721	0.731	0.694	0.690	0.751	0.733
OOB prediction error	0.1399774	0.1422186	0.1401969	0.1495144	0.1314461	0.1437683	0.1435094	0.1327389	0.1412994	0.1505671
Confusion matrix	D I D 133 46 I 27 34	D I D 140 40 I 28 32	D I D 129 45 I 25 41	D I D 134 29 I 33 44	D I D 124 45 I 37 34	D I D 129 35 I 35 41	D I D 132 42 I 32 34	D I D 117 47 I 36 40	D I D 120 34 I 44 42	D I D 146 29 I 36 29
%correct for decr	74.30 %	77.77 %	74.14 %	82.21 %	73.37 %	78.66 %	75.86 %	71.34 %	77.92 %	83.42 %
%correct for incr	55.74%	53.33 %	62.12 %	57.14 %	47.89 %	53.95 %	51.52 %	52.63 %	48.84 %	44.62 %

Supporting Table S5.2 Summary of group means and phylANOVA results for RF predictors on subset used for RF runs. Units of traits are in order of appearance in table (mm^2, mm/mg^{-1}, g/g^{-1}, mg, m, pg/1C). Ellenberg values do not have a unit.

Predictor	**Group mean**		**F**	**p**
	Decreasing	**Increasing**		
Leaf area	3483	3485	8.44	0.023
SLA	26.7	25.5	2.04	0.292
LDMC	0.205	0.201	0.90	0.483
Seed mass	3.23	10.3	5.88	0.062
Mean vegetative height	0.445	0.486	5.62	0.068
Genome size	2.64	3.81	3.43	0.174
Ellenberg F	5.65	5.29	6.18	0.063
Ellenberg N	4.58	5.23	20.82	0.002
Ellenberg R	6.29	6.52	5.55	0.073
Ellenberg L	7.09	7.16	0.50	0.599
Ellenberg S	0.18	0.55	28.26	0.001

Supporting Table S5.3 Summary report of phylopath path analysis. Model name is given with k, the number of independence claims inherent in the model, q, the number of parameters estimated, the C-statistic and its p-value, CICc and delta_CICc scores, i.e. the C-statistic information criterion corrected for small sample sizes (which converges to CIC for large sample sizes) and the difference in CICc with the best model, as well as the relative likelihoods (l) and CICc weights (w). Model names correspond to the acyclic diagrams in Supporting Figure S5.1, with t standing for traits and n standing for niche.

Model	**k**	**q**	**C**	**p**	**CICc**	**delta_CICc**	**l**	**w**
TN_exogenous	6	22	29.3	3.57e-03	74.6	0.0	1.00e+00	1.00e+00
NT_filter	7	21	49.8	6.65e-03	93.0	18.4	1.03e-04	1.03e-04
TN_just_another_trait	7	21	61.3	6.96e-08	104.5	29.9	3.25e-07	3.25e-07
TN_not_tied_in	9	19	77.6	2.28e-09	116.6	41.9	7.81e-10	7.81e-10
TN_independent	8	20	76.3	7.60e-10	117.4	42.8	5.07e-10	5.07e-10
NT_exogenous	6	22	74.5	4.58e-11	119.8	45.2	1.53e-10	1.53e-10
NT_just_another_trait	5	23	75.8	3.34e-12	123.2	48.6	2.77e-11	2.77e-11
NT_not_tied_in	8	20	84.5	2.54e-11	125.6	51.0	8.50e-12	8.50e-12
NT_independent	7	21	86.9	1.44e-12	130.1	55.5	8.78e-13	8.78e-13
TN_filter	8	20	117.3	0.00e+00	158.4	83.8	6.41e-19	6.41e-19
NT_mediator	9	19	233.0	0.00e+00	272.0	197.4	1.37e-43	1.37e-43
I_exogenous	7	21	229.3	0.00e+00	272.6	197.9	1.04e-43	1.04e-43
I_trait_mediator	11	17	249.9	0.00e+00	284.7	210.0	2.46e-46	2.46e-46
I_not_tied_in	12	16	285.2	0.00e+00	317.9	243.3	1.50e-53	1.50e-53
I_independent	11	17	284.3	0.00e+00	319.1	244.5	8.06e-54	8.06e-54
I_niche_mediator	11	17	299.2	0.00e+00	334.0	259.3	4.87e-57	4.87e-57