Migration without interbreeding: evolutionary history of a highly selfing Mediterranean grass inferred from whole genomes

Christoph Stritt[1], Elena L. Gimmi[1], Michele Wyler[1], Abdelmonaim H. Bakali[2], Aleksandra Skalska[3], Robert Hasterok[3], Luis A.J. Mur[4], Nicola Pecchioni[5], Anne C. Roulin[1*]

***For correspondence:**
anne.roulin@botinst.uzh.ch
(telephone +41 44 63 48398)

[1]Institute for Plant and Microbial Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland; [2]National Institute of Agronomy, Regional Center of Errachidia, 529 Boutamine, Errachidia, Morroco; [3]Plant Cytogenetics and Molecular Biology Group, Institute of Biology, Biotechnology and Environmental Protection, Faculty of Natural Sciences, University of Silesia in Katowice, 28 Jagiellonska Street, 40-032 Katowice, Poland; [4]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Wales, UK; [5]Research Centre for Cereal and Industrial Crops, CREA – Council for Agricultural Research and Economics, S.S. 673 Km 25,200, 71122 Foggia, Italy

Abstract Whole genome sequences and coalescence theory allow the study of plant evolution in unprecedented detail. In this study we extend the genomic resources for the wild Mediterranean grass *Brachypodium distachyon* to investigate the scale of population structure and its underlying history at whole-genome resolution. The analysis of 196 accessions, spanning the Mediterranean from Iberia to Iraq, shows that the interplay of high selfing and seed dispersal rates has shaped genetic structure. At the continental scale, evolution in *B. distachyon* is characterized by the independent expansion of three lineages during the Upper Pleistocene. Today, these lineages may occur in sympatry yet do not interbreed. At the local scale, dispersal and selfing interact to maintain high genotypic diversity. Our study lays a foundation for the study of microevolution in *B. distachyon* and identifies adaptive phenotypic plasticity and frequency-dependent selection as key themes to be addressed with this model system.

Introduction

The Mediterranean Basin harbors 25,000 native plant species and is second only to the wet tropics in terms of plant species richness (***Myers et al., 2000***). Compared to more northern latitudes in America and Eurasia, where much of the present-day genetic variation in plants and animals can be explained by post-glaciation expansions (***Hewitt, 2004***), the combination of a mountainous topography and climatic stability allowed Mediterranean species to persist and diversify over prolonged periods of time (***Nieto Feliner, 2014***).

Because of this setting, the Mediterranean has since long been recognized as a great natural laboratory for the study of evolution. Numerous phylogeographic studies (reviewed in ***Thompson,***

2005; ***Médail and Diadema, 2009***) have revealed an impressive variation of evolutionary histories of Mediterranean plants throughout the Quaternary and the Pliocene, with a few lineages left over from more ancient times before the onset of the Mediterranean-type climate in the Miocene (***Rundel et al., 2016***). These studies, however, typically rely on single gene genealogies and have little power to infer evolutionary processes beyond the large-scale movement of pollen and seed: due to the high stochasticity of gene trees (***Hey and Machado, 2003***), evolutionary processes at ecological timescales, in particular demographic history and adaptation, are still poorly understood.

Whole genome sequences and coalescence theory offer new possibilities to study evolution in wild plant populations and overcome some of these limitations (***Weigel and Nordborg, 2015***). In their varying patterns of diversity and linkage, genomes contain records of past demographic and selective events. Recent methods allow to decipher this information and draw historical inferences from molecular patterns (reviewed in ***Pavlidis and Alachiotis, 2017***; ***Spence et al., 2018***). To take full advantage of these tools, however, a robust reference genome and additional resources such as annotations and linkage maps are required. For plants, genome-wide coalescence approaches have therefore been largely restricted to domesticated species (reviewed in ***Ross-Ibarra et al., 2007***; ***Gaut et al., 2018***) and *Arabidopsis thaliana* (***Alonso-Blanco et al., 2016***; ***Lee et al., 2017***; ***Fulgione et al., 2017***; ***Durvasula et al., 2017***). Clearly additional wild plant model species would be of great value to sample more evolutionary histories and broaden our understanding of how wild plants evolve and adapt to changing environments.

In this study, we extend the genomic resources for the Mediterranean grass *Brachypodium distachyon* to establish the species as a system to investigate plant microevolution. *B. distachyon* naturally occurs around the Mediterranean Basin and is a specialist of arid, oligotrophic habitats characterized by recurrent disturbance (***San Miguel, 2008***). It has a fully assembled diploid 272 Mb genome (***Vogel et al., 2010***) and is an established model to study various aspects of grass biology (***Vogel, 2016***; ***Scholthof et al., 2018***). Its allure as a model for plant evolution lies in its Mediterranean distribution, providing a window into an ecologically precious yet understudied region; its belonging to the Poaceae, a plant family dominating terrestrial ecosystems around the globe (***Gibson, 2009***); and, finally, its genome of intermediate complexity, where numerous gene copy number variants (***Gordon et al., 2017***) and transposable element polymorphisms (***Stritt et al., 2018***, ***2019***) contribute to molecular evolution.

An overview of population genetic studies with *B. distachyon* is provided in Supplementary file 1. A key question which has emerged from these studies concerns the role of seed dispersal in shaping genetic structure. A large-scale survey in Turkey with microsatellite markers reported high genotypic diversity within and little differentiation between populations (***Vogel et al., 2009***). This was explained through a combination of long-distance seed dispersal and extreme selfing rates, the former shuffling genotypes across the landscape, the latter preventing gene flow between diverged genotypes. Other studies, however, found little diversity within and high differentiation between populations in Turkey (***Dell'Acqua et al., 2014***) and Spain (***Marques et al., 2017***), in agreement with the increased genetic drift expected in selfing organisms (***Charlesworth, 2003***). Because these studies consider different populations and use different sampling schemes, it is difficult to tell whether their disagreement about the role of dispersal reflects different biological realities or study designs.

A second pattern highlighted by previous studies is the sympatry of highly diverged lineages differing in flowering phenology (***Tyler et al., 2016***; ***Gordon et al., 2017***). As in other widespread annual plants (e.g. ***Engelmann and Purugganan, 2006***), genetic variation for the control of flowering is present in *B. distachyon*, in particular in the vernalization pathway preventing the development of flowers in the cold season (***Ream et al., 2014***). It has been suggested that differences in flowering phenology could explain the lack of admixture between sympatric lineages (***Gordon et al., 2017***). So far, however, too few samples are available to understand how heritable variation in flowering

relates to geography and environmental differences.

In this study, we extend the genomic resources for *B. distachyon* with whole-genome sequences for 75 Italian and 11 French accessions sampled at local and regional scales. Using a total of 196 sequenced accessions spanning the Mediterranean from Iberia to Iraq, we describe the scale of population structure in this model species, infer the underlying history of population divergence, and discuss how the interplay of selfing and dispersal has shaped genetic structure across spatial scales. Furthermore, we conducted flowering time experiments including the newly sampled accessions and argue that previous studies have exaggerated the effect of flowering time differences on genetic structure in this species.

Results

Sampling, sequencing, and variant calling

We collected 719 *Brachypodium* plants across Italy, in the Languedoc region in Southern France and in the Atlas Mountains of Morocco (Supplementary file 2). Besides patching a conspicuous geographical gap in the *B. distachyon* collection, we sampled across different spatial scales in order to better understand the scale of population structure in this species. After using genetic markers to distinguish the morphologically similar sister species *B. distachyon*, *B. stacei* and their allotetraploid hybrid *B. hybridum* (***Catalán et al., 2012***), 86 *B. distachyon* accessions were selected for sequencing. In addition, we sequenced the reference accession Bd21 as a control and the *B. stacei* accession Cef2 as an outgroup for phylogenetic analyses and to estimate the ancestral states of variants. All 70 accessions sampled in Morocco were genotyped as *B. hybridum* and therefore not further analyzed in this study.

The 88 selected accessions were sequenced to a mean depth of 23 and joined with publicly available data for 52 accessions originating from Iberia and the eastern Mediterranean (***Gordon et al., 2017***), and with 57 recently sequenced accessions collected across Turkey (***Skalska et al., 2020***), resulting in a total of 196 accessions of *B. distachyon* and one of *B. stacei* (Supplementary file 2). The percentage of reads mapping to the *B. distachyon* reference genome confirms that – from a pragmatic genomic perspective – the sequenced accessions can be assigned to the same species: for the 196 accessions on average 97% (range 89–99%) of the reads mapped against Bd21, compared to 79% of the *B. stacei* reads (Supplementary file 2). From the reference-aligned reads 8,426,015 hard-filtered single nucleotide polymorphisms (SNPs) were identified in the 196 accessions.

Dealing with artificial heterozygosity due to structural variation

Identifying heterozygous SNPs in *B. distachyon* is complicated by a high number of gene copy number variants (***Gordon et al., 2017***). Among the SNPs passing the initial filtering thresholds, we observed many variants aggregating in islands of extended heterozygosity (IEH): regions of up to few thousand base pairs with extreme numbers of heterozygous sites (***Figure 1–Figure supplement 1***). IEH can be artifacts which arise when reads from paralogous stretches of DNA are collapsed onto a single region in the reference genome; but they can also represent real heterozygosity in the case of recent admixture between diverged lineages. To distinguish the two possibilities, we compared IEH among the initial hard-filtered SNPs and 5,774,928 SNPs in high-confidence regions, defined as regions where at least 90 percent of the accessions have a sequencing depth within 1.5 standard deviations of their mean depth. A median of 116 IEH were identified per accession in the hard-filtered set as compared to only two in the high-confidence set (***Figure 1***), indicating that most IEH are artifacts caused by structural variants. This finding is confirmed by considering a second high-quality genome assembly for the Turkish accession BdTR7a (***Stritt et al., 2019***). While 43,254 heterozygous sites are present after hard-filtering when mapping the BdTR7a reads against the

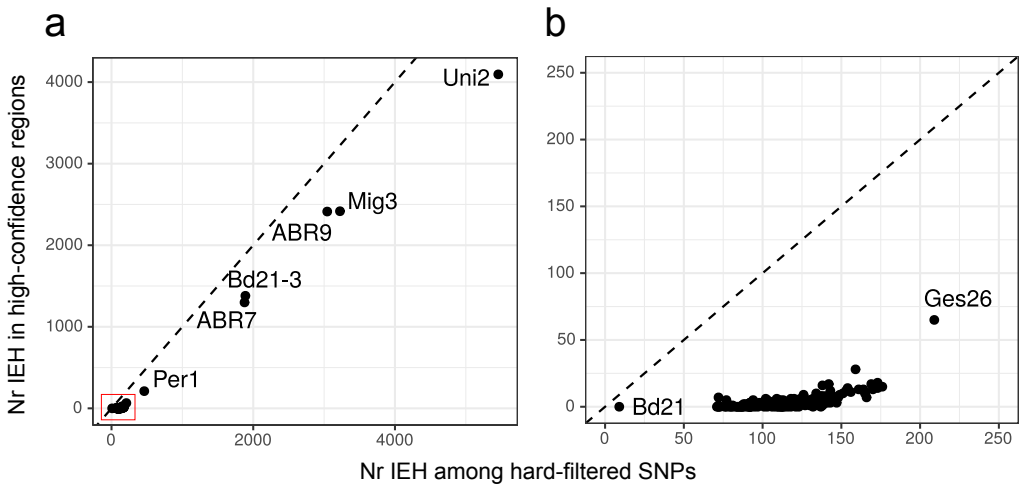

Figure 1. Islands of extended heterozygosity (IEH) among hard-filtered SNPs (x axis) and among SNPs in high confidence regions (y axis), defined as regions where at least 90 percent of the accessions have a sequencing depth within 1.5 standard deviations of their mean sequencing depth. **a)** Overview of all 196 accessions, highlighting the six outlier accessions. **b)** Zoom-in on the red square in panel a.

Figure 1–Figure supplement 1. Genome browser visualization of real and artificial IEH.

Figure 1–Figure supplement 2. Read coverage in IEH, allowing to distinguish real from artificial heterozygosity

reference accession Bd21, only 533 are obtained when mapping these reads against the BdTR7a assembly itself.

Six accessions (Uni2, Mig3, ABR9, Bd21-3, ABR7, Per1) deviate from the general pattern and have hundreds of IEH even in high confidence regions (***Figure 1a***). As the sequencing depth at these IEH is within the normal range and does not suggest the presence of structural variation (***Figure 1–Figure supplement 2***), we propose that this is real heterozygosity left over from recent interbreeding between diverged genotypes.

On average 46% of the IEH identified in any accession overlap with annotated genes, with a median of 50 genes per accession showing evidence of extended heterozygosity. As expected from the composition of the pan-genome (***Gordon et al., 2017***), the total set of 4,211 genes with IEH is enriched for the GO terms 'defense response', 'response to stress' and other functions associated with highly dynamic gene families.

Genetic structure: from major lineages ...

Principal component analysis (PCA), the ancestry estimation method sNMF and fineSTRUCTURE were used to characterize genetic structure in *B. distachyon* (***Figure 2***). All methods capture a clear subdivision of the 196 accessions into five major genetic clades, except for two highly admixed accessions from the Pyrenees which cannot be assigned (Arn1 and Mon3; ***Figure 1a***). This subdivision is particularly clear in the sNMF analysis (***Figure 2b***), where model fit improves in large steps until K=5, after which it declines in smaller steps as regional and local populations are recovered. At the base of the rooted *B. distachyon* phylogeny (***Figure 2c***), 15 Italian accessions form a clade (hereafter called C_Italia) consisting of nine plants from the Gargano peninsula in southeastern Italy and six collected across Italy, namely in Sicily, southern Campania and Liguria (***Figure 2a***). After the divergence of this basal lineage, the tree splits into two large lineages (***Figure 2c***): one comprising all the remaining plants from Italy (A_Italia) and plants primarily from the coastal regions of Turkey (A_East), the other comprising plants from the western Mediterranean (B_West) and plants primarily from the Anatolian plateau and south-eastern Turkey/Iraq (B_East). The same subdivision is evident in the PCA (***Figure 2d***). A geographic outlier are three accessions from southern Turkey

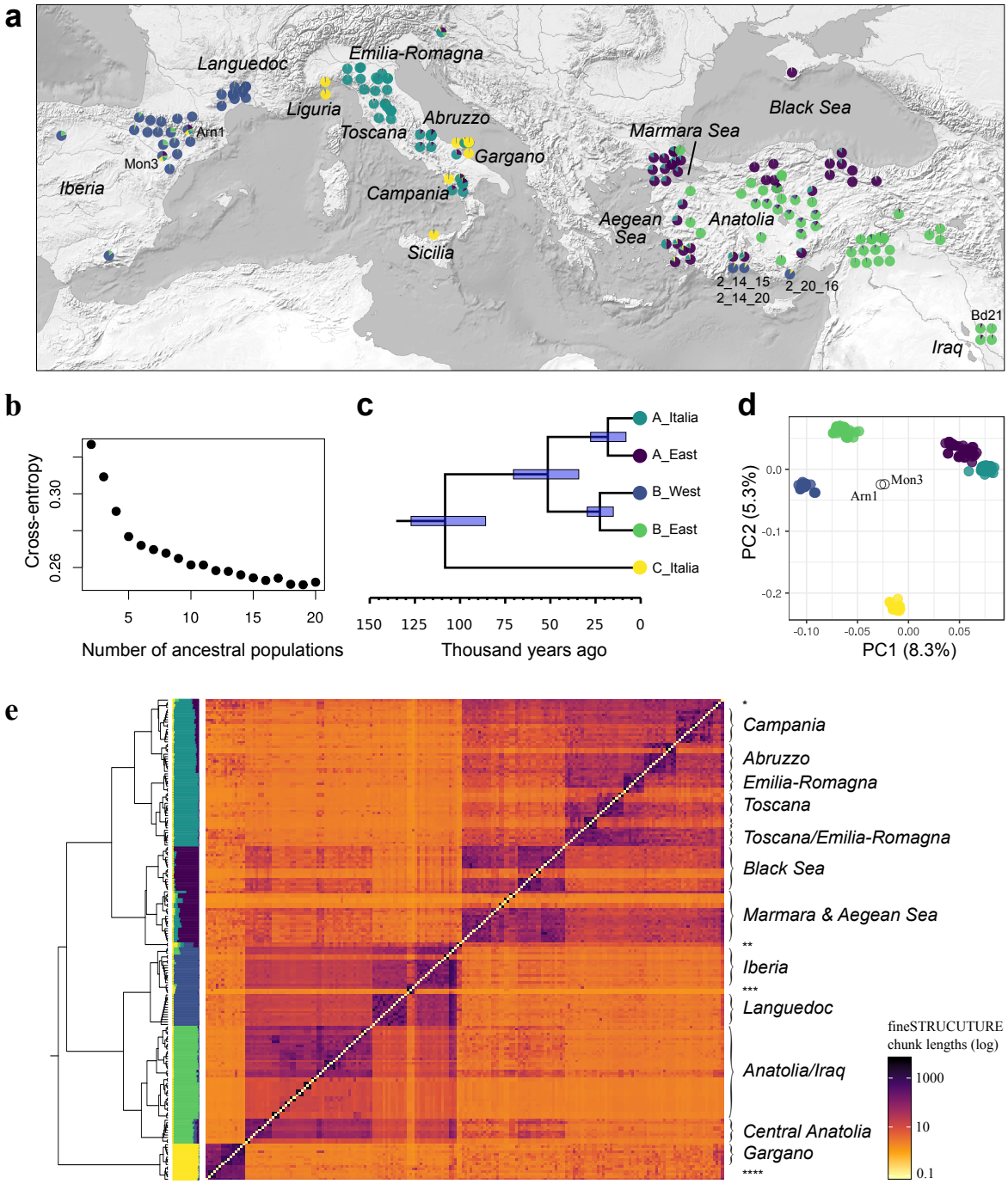

Figure 2. Geographic origin and genetic structure. **a)** Map and pie charts showing shared ancestry. Only a subset of accessions is displayed for the dense local samples in Italy and France. **b)** Cross-entropy for increasing K in the ancestry analysis with sNMF. **c)** Rooted phylogeny with divergence times, estimated under a multispecies coalescent model. **d)** Principal component analysis. **e)** Phylogenetic tree and coancestry matrix estimated with fineSTRUCTURE, as well as sNMF barplot for K=5. The fineSTRUCTURE chunk length is the recombination map distance donated to individuals in rows from individuals in columns (log scale). Names in italics highlight regional populations, asterisks indicate small clusters: * ABR9 collected in Albania and three accessions from the Gargano peninsula. ** Arn1 and Mon3, the two accessions which cannot be assigned to a cluster as they contain substantial ancestry from all five ancestral populations. *** Three Turkish accessions clustering with accessions from the western Mediterranean. **** Accessions from Sicily, Campania and Liguria.

Figure 2–Figure supplement 1. sNMF analysis with K from 2 to 20.

Figure 2–Figure supplement 2. Within-clade PCAs.

Figure 2–source data 1. Interactive phylogenetic tree/map on https://github.com/cstritt/bdis-phylogeo

(2_20_16, 2_14_15 and 2_14_20; ***Figure 2a***) which cluster with accessions from Spain and France. As these Turkish accessions were collected recently (Supplementary file 2) and not grown in contact with Iberian or French accessions, this finding can hardly be explained by mislabeling or seed contamination.

To estimate when the five major genetic clades diverged, we used the multispecies coalescent approach implemented in BPP (***Flouri et al., 2018***, ***Figure 2c***). Based on a sample of 200 random unlinked 1kb segments in 20 diverse accessions and an assumed mutation rate of $7x10^{-9}$ mutations per base pair per generation (***Ossowski et al., 2010***), the split between the basal C lineage and the others is dated to 107 thousand years ago (kya, 95% highest posterior density interval 86–128 ky) and the split between the A and B lineages to 52 kya (95% HPDI 35–72 ky). Within the A lineage, the separation between A_East and A_Italia is dated to 18 kya (95% HPDI 8–27 ky), while within the B lineage the separation between B_East and B_West is estimated at 22 kya (95% HPDI 15–30 ky).

At this coarse level of genetic structure, we thus observe an incongruence between genetic structure and geographic occurrence. Roughly half of the plants collected in Turkey and Iraq are more closely related to plants from the western Mediterranean than to other plants from Turkey. Similarly, in Italy the majority of the collected plants are more closely related to plants from Turkey than to the 15 Italian plants at the base of the phylogeny. Plotting ancestry components on the map and along the phylogenetic tree (***Figure 2a*** and ***Figure 2e***) shows that there is no evidence for gene flow between sympatric lineages, neither in Turkey nor in Italy. There is shared ancestry, however, within the A and B lineages. In the A lineage, for instance, plants from western Turkey (Marmara and Aegean Sea regions) share ancestry with Italian plants (***Figure 2a***). Similarly in the B lineage, plants from central Anatolia share ancestry with B_West accessions.

... to regional populations

Population subdivision in *B. distachyon* is hierarchical and does not end at the lineage level. Substructure within lineages is evident for increasing values of K in the sNMF analysis (***Figure 2–Figure supplement 1***), in within-lineage PCAs (***Figure 2–Figure supplement 2***), and in the fineSTRUCTURE analysis (***Figure 2e***, ***Figure 2–source data 1***). The latter is particularly informative: by modeling linkage disequilibrium in addition to mutation and coalescence (***Lawson et al., 2012***), this method reveals local and regional patterns not evident in clustering analyses based on unlinked SNPs. The fineSTRUCTURE analysis neatly recovers geographical populations in Italy (***Figure 2e***, ***Figure 2–source data 1***). In the A_Italia clade, it distinguishes populations from Campania, Abruzzo and Gargano in the South and from Toscana and Emilia-Romagna in the North, with the Apennines – the 1,200 km mountain chain spanning the peninsula in a NW-SE direction – forming a geographic barrier between regional populations. Also in the other lineage present in Italy, C_Italia, the 15 samples cluster according to their geographic origin. In the western Mediterranean, the Pyrenees separate an Iberian from a French population comprising all the newly collected accessions from the Languedoc region. Within the coastal A_East clade, fineSTRUCTURE divides the samples into a northern population with accessions from the Black Sea region, and a south-western population with samples from the Aegean and Marmara regions. Substructure within the B_East lineage is less obvious. While ten accessions from central Anatolia form a distinct group, the rest of the samples cover a large area form Baghdad to the Marmara Sea region and do not show an obvious mapping between phylogeny and geography.

In the following, to avoid confusion about the level of genetic structure under consideration, we will refer to the *lineages* A, B, C; the *geographic clades* A_Italia, A_East, B_East, B_West, C_Italia; and, finally, the *regional populations* Toscana, Black Sea, Languedoc etc, as described in ***Figure 2e***.

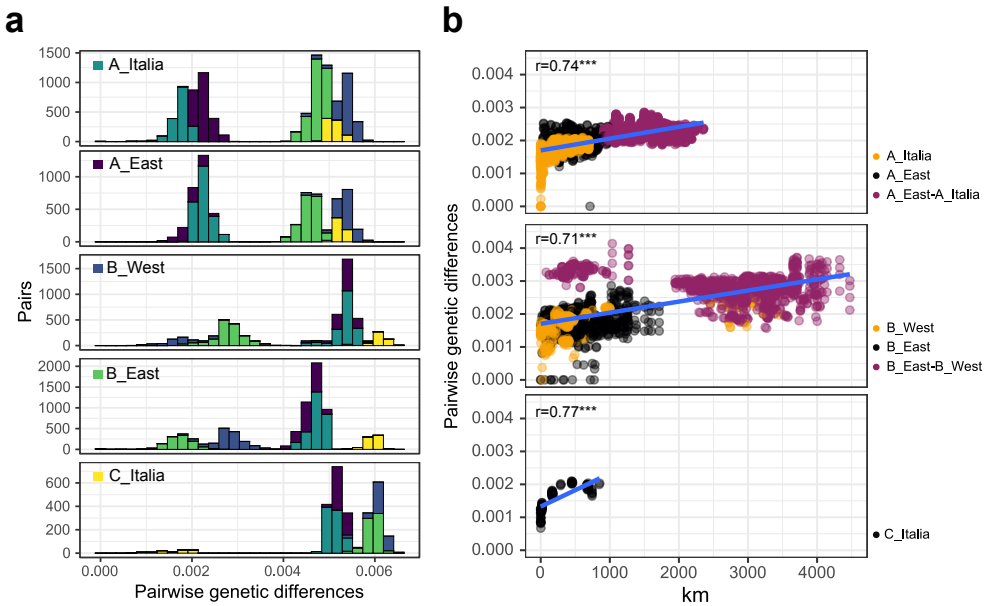

Figure 3. Genetic diversity and differentiation. **a)** Pairwise genetic differences (d_{XY}) within and between the five geographic clades. One panel is shown for each clade, the colors within the panel showing with which clade pairwise comparisons are made. **b)** Correlation between geographic and genetic distance within the three major genetic lineages A, B and C. Colors within the panels indicate between which geographic regions pairwise comparisons are made. Correlation coefficients from Mantel test and their significance are displayed.

High selfing rates throughout the species' range

Selfing rates often vary within species (***Barrett, 2003***; ***Griffin and Willi, 2014***), yet studies with *B. distachyon* so far unanimously reported a near absence of heterozygosity (***Vogel et al., 2009***; ***Marques et al., 2017***). To test whether this pattern holds across geographic regions, we estimated inbreeding coefficients and selfing rates for the 196 accessions.

Using SNPs in high-confidence regions, as defined above, we estimated a median inbreeding coefficient F of 0.996 among the 196 accessions (mad=0.003, Supplementary file 2), which translates to a median selfing rate of 0.998 using $s = \frac{2F}{1+F}$. When population structure is taken into account and F is estimated within geographic clades and regional populations, estimates are only marginally lower for most accessions, with a median of 0.991 at the clade and 0.989 at the population level (Supplementary file 2). Lower F values are estimated for the six accessions whose genomes contain numerous IEHs. At the clade level, significant differences in F are found between B_West and the two eastern clades ($P < 0.05$, Tukey's HSD), with lower values in B_West. This effect is also apparent at the population level, where significant differences are only found in comparisons with the Iberia population. Clearly this effect is driven by the outliers in the Iberia population: four out of the six accessions with numerous IEH and low F values belong to this population. Overall, our results thus show that extreme levels of homozygosity and high selfing rates predominate throughout the species' range and vary little between clades and populations.

Pairwise genetic differences between clades and across space

To understand how genetic diversity is partitioned across space and the different levels of genetic structure in *B. distachyon*, we took advantage of the coordinates available for each individual and estimated pairwise genetic differences (d_{XY}) at fourfold degenerate sites. Within the five geographic clades, median d_{XY} is 0.0018, that is, 1.8 SNPs per 1000 bp (***Figure 3a***). Values differ somewhat among clades: d_{XY} is higher in A_East (0.0019) than in the other lineages ($P < 2x10^{-16}$, Kruskal-Wallis),

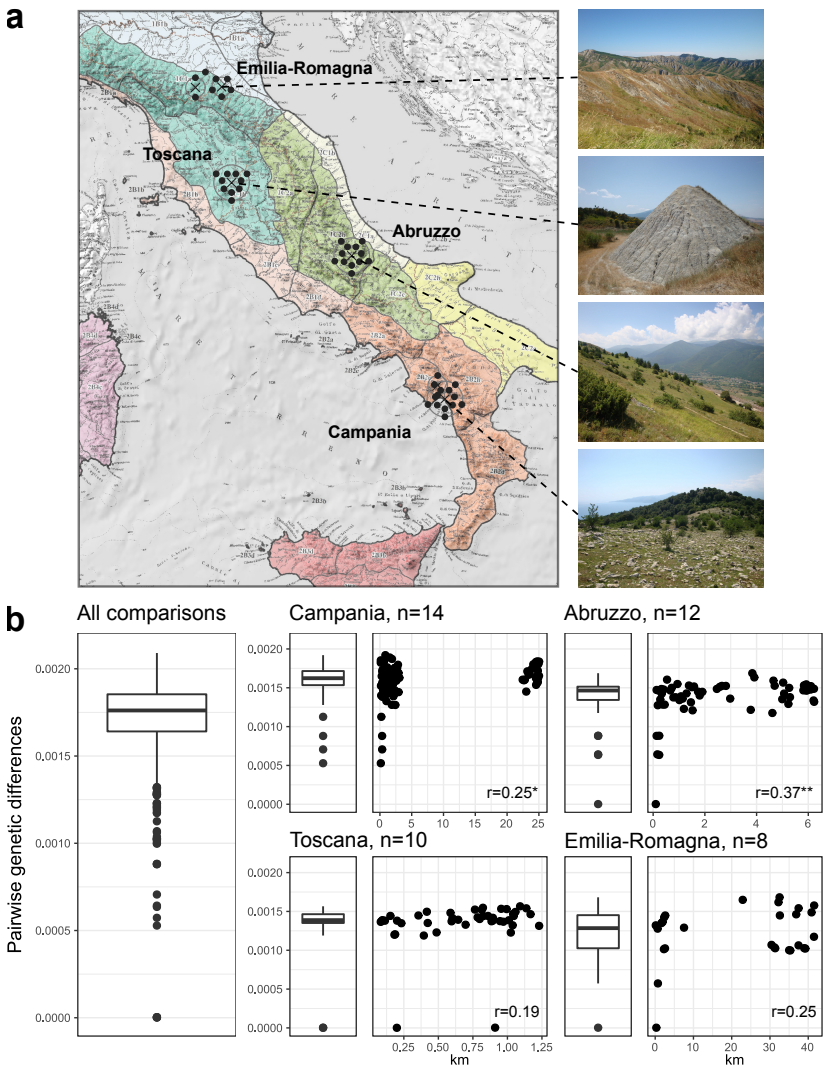

Figure 4. Genetic differentiation across ecoregions in Italy. **a)** Map showing the ecoregion classification of Blasi et al. (2014) and *B. distachyon* habitats in the four sampled regional populations. **b)** Genetic diversity within the A_Italia clade, for which the most extensive sampling is available. Boxplots show the distribution of d_{XY} across the whole clade and within regional populations, while dotplots show how d_{XY} relates to geographic distance.

and lower in C_Italia (0.0017) than in the others ($P < 0.03$). Levels of diversity in the B_East and B_West clades are similar (0.0017, $P = 0.37$), although significant differences appear when regional populations within these clades are compared: diversity is higher in the Languedoc population than in Iberia (0.0015 vs. 0.0013, $P = 0.01$), and higher in the large Anatolia/Iraq population (0.0016) than in the western Mediterranean ($P < 0.001$).

Within all three major lineages, d_{XY} increases with geographic distance, as shown by significant Mantel correlations (***Figure 3b***). While correlation coefficients are similar between lineages (r_A=0.75, r_B=0.71, r_C=0.77), some deviations are observed in the B lineage. On the one hand, there are genetically similar accessions with distant geographic origins, particularly in the B_East clade, in agreement with the previously observed poor mapping between geography and phylogeny in this clade. On the other hand, genetically distant individuals can be geographically close (upper left of the middle panel in ***Figure 3b***). These are comparisons of Turkish accessions with the three accessions collected in Turkey clustering genetically with B_West accessions (***Figure 2a***).

Genetic diversity and demographic history of *B. distachyon* in Italy

The local and regional sampling in Italy allows us to take a closer look at how genetic diversity is distributed across small spatial scales. In Italy, the Apennines separate *B. distachyon* into genetically differentiated regional populations (***Figure 4a***). This mountain chain is not only a barrier to gene flow, but also affects the levels of precipitation across the peninsula and largely defines the ecoregions of Italy (***Figure 4a***, ***Blasi et al., 2014***).

Figure 4b shows pairwise genetic differences in four regional populations for which samples were collected few dozen meters to few kilometers apart. Within populations, some more closely related individuals stand out (d_{XY}<0.001); these were typically collected within 1 km of each other. Most plants, however, even if growing close-by, are genetically quite distinct and can be more different from each other than from accessions kilometers away. Isolation by distance is weak at this scale and, while still significant in the two southern populations, is not significant in the Toscana and Emilia-Romagna populations. Genetic diversity declines step-wise from south to north, with a median d_{XY} of 0.0016 in Campania, 0.0015 in Abruzzo, 0.0014 in Toscana, and 0.0013 in Emilia-Romagna (***Figure 4b***).

To understand the demographic history underlying these patterns of diversity and divergence, we estimated divergence times and effective population sizes with fastsimcoal. A simple variable population size model (***Figure 5a***) was fitted to the derived allele frequency spectra estimated with fourfold degenerate (4D) sites (***Figure 5–Figure supplement 1***). The divergence between northern and southern Italian populations is dated to 18,462 kya (95% CI 6,041–37,264, ***Figure 5b***). In the South, the Campania and Abruzzo populations split 17,165 kya (CI 5,346–35,234), while in the North the Toscana and the Emilia-Romagna populations split 11,439 kya (CI 2,899–25,587). Population size estimates are consistently higher for the southern Campania and Abruzzo populations (***Figure 5b***), in agreement with the decline of genetic diversity from South to North. While the close correspondence of observed and predicted DAF suggests a good overall model fit, our best model systematically predicts less high-frequency shared alleles than observed (***Figure 5–Figure supplement 1***).

Flowering phenology in the greenhouse and outdoors

Previous studies described the co-occurrence of the A_East and B_East clades in Turkey and interpreted the lack of gene flow between these lineages as a result of variation in flowering time: accessions from the A_East clade require long vernalization and show a delay in flowering under greenhouse conditions (***Tyler et al., 2016***; ***Gordon et al., 2017***). Because much emphasis has been placed on flowering time differences as an explanation for genetic structure in the species, we included this aspect in our study and conducted two experiments.

Firstly, we tested how the newly sampled A_Italia and C_Italia accessions respond to vernalization treatments ranging from 0 to 12 weeks at 4°C under short-day conditions (see Material and Methods for the experimental set-up). Our experiment replicated the previously identified flowering behavior of the two B_East accessions Bd3-1 and BdTR3C and the two A_East accessions Bd1-1 and BdTR7a, which in this order required four, six, eight, and more than 12 weeks to satisfy the vernalization response (***Figure 6a***, ***Figure 6–source data 1***). With the exception of two C_Italia accessions which required six weeks of cold, all other 10 newly tested Italian accessions required long periods (>10 weeks) of cold followed by up to 40 days in warm conditions to flower (***Figure 6a***, ***Figure 6–source data 1***).

In the wild, plants all go through a prolonged vernalization process during winter which might produce different outcomes than the conditions tested in the greenhouse. To test for this possibility, we planted seeds of 105 accessions from all five geographic clades outdoors in autumn, with six replicates per accession. Flowering behavior in this experiment was quite different (***Figure 6b***, ***Figure 6–source data 1***). All plants emerged in autumn and flowered next spring within two weeks. However, clades still differ in their flowering behavior. C_Italia and B_East accessions tended to

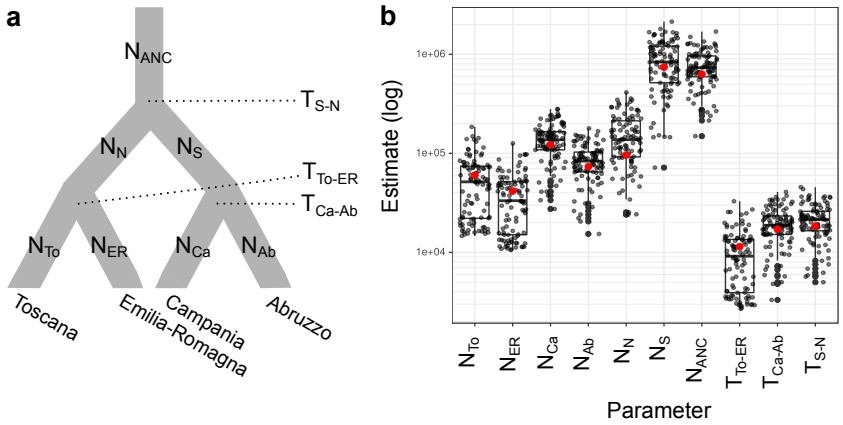

Figure 5. Demographic history of Italian *B. distachyon* estimated with fastsimcoal. **a)** Demographic model. **b)** Parameter estimates from 100 runs of the model. The red dots show the estimates from the run with the highest likelihood.

Figure 5–Figure supplement 1. Derived allele frequency spectra and residual errors of the model with the highest likelihood.

produce flowers more rapidly than A_East accessions, though the differences between these them are not statistically significant at the 5% level. A_Italia and B_West, on the other hand, take longer to flower than the other clades (LME, $P < 0.006$ for all comparisons between these two clades and the others), even though overall only two weeks separate the early from late flowering accessions.

Discussion

B. distachyon was initially developed as a model to study gene function in grasses, with an emphasis on applications in crop breeding. With this study we extend the genomic resources for this Mediterranean grass and lay a foundation for evolutionary studies by providing a historical explanation for genetic structure in the species. Population structure in *B. distachyon* is strongly hierarchical and reflects, as we discuss in the following, the combined effect of high selfing rates and effective seed dispersal. At the continental scale, dispersal and selfing help to explain why highly diverged lineages might be present in the same locality without any sign of admixture. At the local scale, selfing and dispersal interact to maintain high local genetic diversity, with important implications for how natural selection is expected to act in this species.

Extreme selfing and extended heterozygosity as a result of recent outcrossing

In plants, the wide occurrence of sexual reproduction and the elaborate means evolved to ensure it suggest that self-fertilization is disadvantageous under a wide range of conditions (***Barrett, 2010***). The best known exception are colonizing weeds (***Baker, 1955***; ***Stebbins, 1957***): here, the advantage of reproductive assurance during colonization can outweigh the costs of inbreeding. *B. distachyon* clearly is a colonizing weed: it thrives in disturbed areas like pastures or along paths and declines when stable conditions allow the establishment of perennial competitors (***San Miguel, 2008***).

Considering this life style, as well as the highly reduced, cleistogamous flowers (***Khan and Stace, 1999***) and the general pattern that wind-pollinated species tend to be either predominantly outcrossing or selfing (***Barrett and Eckert, 1990***), the highly selfing nature of *B. distachyon* comes at no surprise. Nevertheless the degree of homozygosity revealed by molecular markers is astonishing. In a large-scale survey of 187 accessions from Turkey with 43 microsatellite markers, only four heterozygous sites in four individuals were detected (***Vogel et al., 2009***). Similarly, ***Marques et al.*** (***2017***), looking at 137 plants in 14 Iberian populations with 10 markers, estimated inbreeding coefficients of 1 in all except one population.

Our study agrees with this finding and suggests that high selfing rates are prevalent throughout the species' range. It also corrects previous estimates of heterozygosity based on genome-wide SNPs, where we reported a median of 7.8 and 6.3% of heterozygous SNPs in the B_East and B_West clades (***Bourgeois et al., 2018***). These estimates were most likely too high as we did not take into account artificial heterozygosity due to structural variation. Indeed, distinguishing real from artificial heterozygosity proved to be challenging. As we have shown here, a method based on the clustering of heterozygous sites and sequencing depth can be used to identify artificial heterozygosity arising from deletions and duplications. As plant population genomics is extending to more complex and dynamic genomes, searching for islands of extended heterozygosity can be a simple first step to identify problems with heterozygous sites and to limit biases in downstream analyses.

While self-fertilization appears to be the norm in *B. distachyon*, we also find evidence for recent and historical outcrossing. A large part of the heterozygosity we observed occurred in only six accessions, in the form of extended islands of heterozygosity. Extended heterozygosity can result from the purging of deleterious homozygosity, as observed in inbred maize (***Brandenburg et al., 2017***), or from balancing selection (***Nordborg et al., 1996***). These processes, however, imply a selective advantage of heterozygosity and would hardly create hundreds of IEH in single accessions as observed here, but rather few IEH present at higher frequencies. Our results thus provide evidence that outcrossing does occur from time to time. As heterozygosity disappears within a few generations of selfing (***Hedrick, 2011***), the outcrossing events leading to the six heterozygous accessions must be recent and might have occurred during propagation of the accessions for research and seed distribution.

Historical admixture is evident in the two Iberian accessions Arn1 and Mon3 which share ancestry with all five geographic clades (***Figure 2***). Shared ancestry is also evident in other accessions. Here, however, descent from a common ancestor is the more parsimonious explanation than admixture. Ancestry is hardly ever shared between different clades occurring in the same region, but rather between closely related clades present in different regions. Admixture would thus require not only a rare outcrossing event, but also additional seed migration after an initial range expansion. A case for historical admixture has been made based on an incongruence between plastome and nuclear phylogenies (***Sancho et al., 2017***). Several B_East accessions have B_West plastome genotypes, but no admixture is evident at the nuclear level, which the authors explain through repeated backcrossing and loss of the B_West nuclear ancestry. Alternatively, incomplete lineage sorting of plastome haplotypes might have occurred, as for example described for *Hordeum* (***Jakob and Blattner, 2006***). This explanation requires no outcrossing and is thus more consistent with historically high selfing rates in *B. distachyon*.

Range expansions have led to the sympatry of reproductively isolated lineages

On a larger spatial scale, the dispersal ability of *B. distachyon* is evident in the widespread occurrence of three diverged genetic lineages. The A and B lineages have been described previously (***Gordon et al., 2017***), although with only seven accessions sampled the geographic spread of the A lineage was so far unknown. Here we show that the A lineage is present throughout Italy, and we report a new lineage of *B. distachyon*, C_Italia. This lineage diverged from the others 86–128 kya, forms the basal branch in the phylogeny, and occurs in sympatry with the A lineage present in Italy.

The geographic co-occurrence of genetically diverged lineages in *B. distachyon* is most easily explained by independent lineage expansions leading to overlapping ranges (***Avise, 1987***). In the Mediterranean region, east-west breaks are a common phylogeographic pattern (***Nieto Feliner, 2014***). A general trend of decreasing genetic diversity from east to west has been described in many animals and plants, including Poaceae species, suggesting that westward waves of colonization have been more common than eastward waves (***Conord et al., 2012***). A pattern of decreasing genetic diversity from east to west is present in the A as well as the B lineage (***Figure 3***), in agreement with

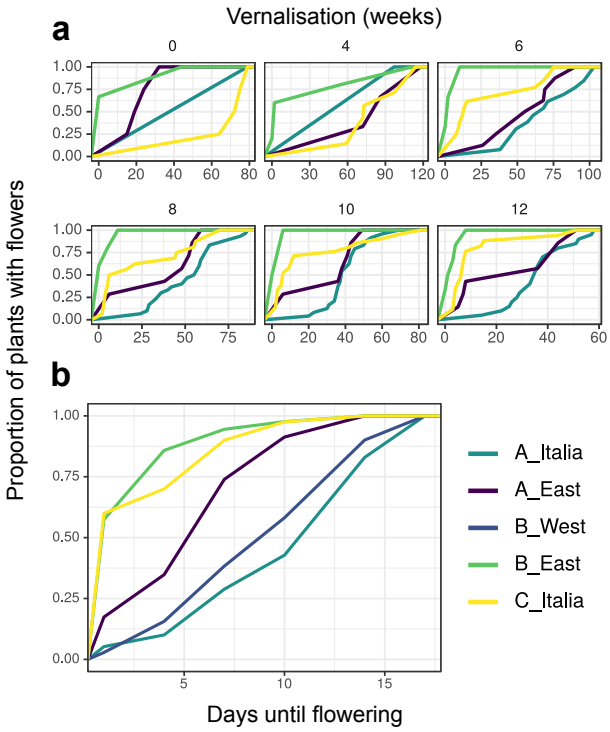

Figure 6. Flowering behavior in the greenhouse (a) and under semi-natural conditions (b). These cumulative distributions show, for each lineage, the proportion of plants with flowers through time, counted in days since the appearance of the first flower in any lineage. **a)** Greenhouse experiment with 16 accessions and 3 replicates per accession, testing six different vernalization periods (0 to 12 weeks). No plants from the B_West lineage were tested, which in other experiments were shown to have an early flowering phenotype like the B_East accessions. **b)** Experiment under 'semi-natural' conditions with 105 accessions and six replicates per accession.

Figure 6–source data 1. Tabulated results for the two flowering time experiments.

a general westwards expansion in *B. distachyon*. As we did not find *B. distachyon* in Morocco and the Balkans remained unsampled, more specific hypotheses about migration routes will have to await future studies.

A more detailed picture can be drawn for the A_Italia clade. Here a South-to-North gradient of declining diversity, shared ancestry and effective population size is observed, which suggests a northwards expansion of this clade. Divergence times between local populations in Italy indicate that this expansion occurred towards the end of the Last Glacial Period (LGP, c. 115,000–11,700 years ago). While the uncertainty in our molecular dating analysis, in particular the assumption of a specific mutation rate, advises caution when discussing events on such a small time-scale, a lineage expansion during the LGP does agree with the habitat preferences of *B. distachyon*. An environmental niche modeling study with *B. distachyon* showed that the range of the species likely increased during the LGP, as the more arid, cooler climate allowed it to colonize lower altitudes (***López-Alvarez et al., 2015***).

Previous studies have demoted geography to "a secondary factor in shaping population structure" (***Gordon et al., 2017***) and instead identified flowering time differences as "a main factor driving rapid intraspecific divergence" (***Sancho et al., 2017***). This view is questionable on both conceptual and empirical grounds. Flowering time differences are at best part of an explanation for genetic structure. In the scenario of subsequent lineage expansions we propose here, reproductive isolation might have evolved when the lineages were geographically isolated; and it might include other pre- and post-zygotic barriers in addition to flowering time, namely niche differentiation or genomic incompatibilities (***Baack et al., 2015***), and in particular high selfing rates (***Baker, 1959***; ***Hu,***

2015). Furthermore, differences observed in the greenhouse are most likely exaggerated through artificially short vernalization times. As our outdoors experiment shows, all accessions produced flowers within two weeks when they went through prolonged vernalization during winter.

While the role of flowering in reproductive isolation is less clear than assumed in previous studies, flowering remains an intriguing trait and appears to be under diversifying selection in *B. distachyon*. A geographic pattern of flowering time differences is observed in Turkey but not in Italy. In Turkey, the A lineage, which requires vernalization to flower, is present along the coast, while the B lineage, which requires little to no vernalization, is mainly found in inland Anatolia (***Figure 2***). ***Skalska et al.*** (***2020***) recently showed that the two lineages also differ in drought tolerance, indicating that the B_East accessions are adapted to the arid climate of central Anatolia. In Italy, no such geographic differences in vernalization requirement and flowering are observed, neither among lineages nor among regional populations. Possibly this reflects the stronger ecological contrasts in Turkey, where rain is scarce on the inland Anatolian plateau also in winter, while the coastal regions have the typical Mediterranean winter rainfalls (***Sensoy, 2008***).

Local consequences of selfing and dispersal

Within regional populations in Italy, isolation by distance is week, and sampling within few hundred meters can yield individuals as different from each other as individuals kilometers apart (***Figure 4b***). Clearly the plants collected locally did not descent from a recent colonization or founder event. Instead, our study supports the hypothesis that the interplay of high selfing and seed dispersal rates shapes local and regional genetic diversity, as proposed by ***Vogel et al.*** (***2009***) for populations in Turkey. With a selfing rate close to one, lineages can evolve independently of each other for a considerable period without the homogenizing effect of interbreeding, and lineage-specific mutations accumulate. Seed dispersal would then ensures that closely related genotypes do not occur in patches, as observed in *A. thaliana* (***Bomblies et al., 2010***), but are shuffled across the landscape.

Central to the selfing-dispersal hypothesis is that *B. distachyon* disperses its seeds easily and over long distances. Morphological adaptations to do so have been described: the species has barbed palea which facilitate zoochory (***Catalán et al., 2016***). Surprisingly, however, so far no study mentioned the animals themselves, sheep and goats. From our fieldwork and previous reports (***Vogel et al., 2009***) it is clear that the habitat in which *B. distachyon* thrives are arid, oligotrophic pastures grazed by sheep and goats. Not only do these animals keep away more competitive perennial plants, but they are also known to carry plant seeds over long distances and to be an important vector of long-distance seed dispersal (***Manzano and Malo, 2006***). In the Mediterranean landscape, which has been shaped by human activity through thousands of years (***Grove and Rackham, 2001***), the movement of domestic animals might thus be a crucial factor in the population dynamics of *B. distachyon* and similar plants.

The possibility that genetic diversity is heavily influenced by long-distance seed dispersal has important implications for the study of natural selection in *B. distachyon*. While selfing itself decreases the fixation time of adaptive mutations and might favor local adaptation (***Glémin and Ronfort, 2013***; ***Hartfield et al., 2017***), selfing combined with high rates of seed migration is expected to favor phenotypic plasticity over local adaptation (***Sultan and Spencer, 2002***). Furthermore, by maintaining high local genetic diversity, seed dispersal could make up for what is considered a main disadvantage of selfing: low diversity and failure to adapt to spatially and temporally heterogeneous environments (***Wright et al., 2013***). Indeed pathogens, an important and highly variable component of the biotic environment (***Karasov et al., 2014***), have been identified as a potentially strong selective pressure in *B. distachyon* (***Bourgeois et al., 2018***). As agents of frequency-dependent selection (***Brown and Tellier, 2011***), they could be an essential part of an explanation for why single lineages fail to rise to high local and regional frequencies.

Materials and Methods

Sampling & sequencing

719 dry *Brachypodium* plants were collected in 2016 across Italy, in the Languedoc region in Southern France, and in 2018 across the Atlas Mountains of Morocco. A representative set of 276 accessions were grown in the greenhouse from single seeds. Their genomic DNA was extracted from leaves with a CTAB-based method (***Stein et al., 2001***), quality-checked with Nanodrop (NanoDrop Technologies Inc), and quantified with Qubit (Invitrogen) in a realtime PCR.

As *B. distachyon* co-occurs with its morphologically similar sister species *B. stacei* and *B. hybridum* (***Catalán et al., 2016***), we first identified the species of the 276 accessions using the microsatellite marker ALB165 (***Giraldo et al., 2012***) and by sequencing the GIGANTEA gene (***López-Alvarez et al., 2012***) with ABI SOLiD. 143 individuals were identified as *B. distachyon*, 130 as *B. hybridum*, and three as *B. stacei* (Supplementary file 2). All 70 accessions collected in Morocco proved to be *B. hybridum* and were not included in this study.

86 *B. distachyon* accessions were selected for sequencing. In order to reduce the chance of sequencing close relatives, only accessions collected at least 50 meters apart were included. In addition, we also sequenced the reference accession Bd21 as a control and the *B. stacei* accession Cef2 as an outgroup. Paired-end libraries were prepared with the PCR-free KAPA Hyper Prep Kit (KAPA Biosystems). The 150 bp paired-end reads were sequenced with Illumina HiSeq2500 at the Genomics Facility Basel to a mean coverage of 23 (range 18–38, Supplementary file 2). The resulting sequencing data were joined with publicly available data for 52 accessions originating mainly from Iberia and Turkey (***Gordon et al., 2017***, downsampled to a mean coverage of 30), and with 57 recently sequenced Turkish accessions (***Skalska et al., 2020***, mean coverage 36). Paired-end reads for the former were downloaded from the JGI GOLD database (https://gold.jgi.doe.gov). To speed up computation and avoid potential bias due to unequal coverage, the publicly available reads were downsampled to a mean coverage of 30 using Picard tools v.1.97 (http://broadinstitute.github.io/picard).

Alignment, variant calling & filtering

Paired-end reads were aligned to version 3.0 of the *B. distachyon* reference genome, downloaded from Phytozome 12, with BWA MEM v. 0.7.12-r1039 (***Li, 2013***). Aligned reads were converted to the bam format and sorted with samtools v. 1.9 (***Li et al., 2009***), and read duplicates were removed with sambamba (***Tarasov et al., 2015***). Variants were then called with GATK v. 4.0.2.1. The following hard-filtering thresholds were applied to the combined variants: only SNPs were retained with a quality-by-depth > 20, a phred score < 20 for Fisher's exact test for strand bias, and a Z-score between -1 and 1 for Wilcoxon's rank sum test of Alt vs. Ref read mapping qualities. Additionally we removed SNPs in low complexity regions identified with dustmasker (***Morgulis et al., 2006***), resulting in a total of 8,426,015 hard-filtered SNPs.

Heterozygosity and masking of ambiguous regions

After hard filtering, an inspection of highly heterozygous regions with the Integrative Genomics Browser (***Thorvaldsdóttir et al., 2013***) suggested that many of the heterozygous SNPs remaining were artifacts due to structural variation. As wrongly identified heterozygosity might severely bias population genetic analyses, we first quantified this problem by scanning each accession for islands of extended heterozygosity (IEH): clusters of heterozygous sites as they arise when reads from structural variants are collapsed onto a single region in the reference genome. Defining thresholds through trial and error, an IEH was recorded if there were ten or more heterozygous SNPs along a distance of at least 300 bp. The program implementing this approach (detectIEH.py) is available on http://github.com/cstritt/popgen.

As hard filtering proved inadequate to deal with the abundant fake heterozygosity identified in the previous step, two additional treatments were applied to the initial set of 8,426,015 SNPs. The first was a masking approach with the goal to obtain high-confidence SNPs, including heterozygous sites, for analyses requiring comparatively few unlinked SNPs across the genome. For each accession a mask was created indicating the regions in the genome within 1.5 standard deviations of the accession-specific mean sequencing depth. In this way, an average of 85% of the genome was included per accession (range 0.76 (Bd1-1) to 0.88 (Bd21)). In a next step, the individual masks were combined with bedtools (***Quinlan and Hall, 2010***) to define high-confidence regions as regions with "normal" coverage in at least 90 percent of the accessions. Removing SNPs outside these regions resulted in a set of 5,774,928 SNPs across the 196 accessions. The second treatment consisted in recoding heterozygous genotypes among the 8,426,015 hard-filtered SNPs as missing. This SNP set was used for methods requiring densely linked SNPs across the genome. With this SNP set, accessions are effectively treated as haploid and no phasing is required.

Population structure

Genetic structure was analysed using principal component analysis, ancestry coefficient estimation, and fineSTRUCTURE, a coalescence-based approach which allows to estimate changing ancestry along the genome (***Lawson et al., 2012***). For the principal component analysis, the 5,774,928 SNPs in high-confidence regions were used and additionally filtered (minor allele frequency 0.05, LD threshold = 0.4) with the snpgdsLDpruning function of the R package SNPrelate (***Zheng et al., 2012***), resulting in 14,485 informative SNPs. The snpgdsPCA function of the same package was used to calculate the components. Ancestry coefficients were estimated with sNMF, as implemented in the R package LEA (***Frichot and François, 2015***), a method appropriate for high levels of inbreeding as it does not assume Hardy-Weinberg equilibrium. Using the same SNPs as for the PCA, sNMF was run for K values from 2 to 20, with 10 repetitions per K. Cross-entropy was plotted to assess model fit.

fineSTRUCTURE v. 4.0.1 (***Lawson et al., 2012***) was used to obtain a more detailed picture of population subdivision in *B. distachyon*. This method uses a hidden Markov model to infer, for each accession, the number and length of chunks along the genome which first coalesce with any other accession. Its output includes a square matrix which indicates the recombination map distance a 'recipient' genotype (rows) receives from the 'donor' genotypes (columns). As fineSTRUCTURE models linkage disequilibrium and performs best with dense haplotype data, the 8,426,015 hard-filtered SNPs were used after removing heterozygous sites and missing data. Recombination distances between loci were obtained from the linkage map of ***Huo et al.*** (***2011***). fineSTRUCTURE was run in its haploid mode, using default settings. To root the phylogenetic tree estimated by fineSTRUCTURE, we used SNPhylo (***Lee et al., 2014***) to estimate a rooted phylogeny from LD-pruned 4D sites, including the *B. stacei* accession Cef2 as an outgroup, and then re-rooted the fineSTRUCTURE tree accordingly.

Estimation of divergence times

To put a timescale on the divergence in *B. distachyon*, we used the multispecies coalescent approach implemented in BPP v.4.2.9 (***Rannala and Yang, 2003***; ***Flouri et al., 2018***). 20 diverse accessions were selected to represent the *B. distachyon* phylogeny, four per geographic clade. 200 random genomic regions of 1 kb length and at least 100 kb apart were chosen, as BPP assumes no recombination within and free recombination between loci. For each of the 20 accessions the 200 sequences were obtained by calling consensus sequences from the respective bam file. Inverse gamma priors were set to (3, 0.014) for the root age τ and to (3, 0.002) for the population size parameter θ, which corresponds to a mean theta of 0.001 and a mean root age of one million years, assuming a mutation rate of $7x10^{-9}$. The MCMC was run four times independently, each time with 408,000 iterations, including a burn-in of 8000 iterations and samples taken every second iteration. Convergence of the four runs was assessed with Tracer (***Rambaut et al., 2018***).

Genetic diversity and population differentiation

To estimate genetic diversity and differentiation, we estimated absolute genetic differentiation d_{xy} at 4D sites. The –includeNonVariantSites option of GATK v.3.4 was used to obtain variable and non-variable 4D sites. For each pairwise comparison between the 196 accessions, we then calculated the number of variable and the total number of sites. Heterozygous sites were ignored and only sites were considered for which both accessions in the comparison had a sequencing depth within 1.5 standard deviations of the accession-specific mean coverage. The script implementing this procedure (pairwise_nucdiff.py) is provided on http://github.com/cstritt/popgen. Pairwise geographic distances were calculated as Great Circle (WGS84 ellipsoid) distances between each pair, using the spDistsN1 function of the R package sp (***Pebesma and Bivand, 2005***). Mantel tests were used to test for isolation by distance, using the mantel.rtest function of the ade4 package with 100 permutations to assess the significance of the correlations (***Dray and Dufour, 2007***). To test if genetic diversity differs between regions, a Kruskal-Wallis test was performed followed by Dunn's test for multiple comparisons with a Benjamini-Hochberg correction.

Demographic inference

Fastsimcoal v.2.6 (***Excoffier et al., 2013***) was used to fit an explicit demographic model and obtain point estimates for population divergence times in the A_Italia clade. Two-dimensional derived allele frequency spectra were estimated with for fourfold degenerate (4D) sites, ignoring heterozygous sites and sampling one haplotype per individual. The ancestral state of 4D sites was estimated by parsimony using the *B. stacei* accession Cef2 as an outgroup, resulting in 90,707 variable and polarized sites included in the analysis. A simple model was fitted which allowed for different population sizes in each regional population and in the ancestral populations. 100 independent runs with 30 ECM cycles and 200,000 simulations per estimation step were performed. We report the point estimates from the run achieving the highest likelihood and a 95% confidence interval obtained from the distributions of parameter estimates of the 100 runs. A mutation rate of $7x10^{-9}$ mutations per bp per generation was assumed.

Flowering time experiments

For the greenhouse experiment, twelve Italian accessions with different geographic origins were analyzed (Cro24, Cm2, Leo2, Mca23, Mca18, Msa27, Ren22, San12, Sap49, Spm16, Sul5, Tar1). We further used four control accessions from Turkey and Iraq (Bd1-1, Bd3-1, BdTR3C, BdTR7a) previously investigated by other authors (***Ream et al., 2014***; ***Sharma et al., 2017***) in order to confirm the reproducibility of the flowering behavior and vernalization requirements. Plants from all 16 accessions were germinated for three weeks in the greenhouse. They were subsequently put in a cooling chamber for 0, 4, 6, 8, 10 or 12 weeks, respectively, and transferred to the greenhouse again. Three to five replicates were used per accession per time point. Time until flowering was counted from the day after the return to warm conditions (directly after the first three weeks of germination for the plants with 0 weeks vernalization). Flowering was defined as the emergence of the first spike. The cooling chamber was set to 4°C, 8 h light and 16 h dark. The greenhouse conditions were 16 h day at 20–21°C and 8 h dark at 18°C with a light intensity of 40–50 Klux and 60% relative humidity over the entire experiment.

For the outdoor experiment, conducted in the Botanical Garden of Zurich, Switzerland, six replicates for 37, 5, 25, 24 and 14 accessions of the A_Italia, A_East, B_West, B_East, and C_Italia clades were planted in November 2017. Only six plants did not germinate and 624 plants were used for phenotyping. To ensure no bias due to position effect, replicates were distributed randomly across trays. Mild heating was applied at the bottom of the compartment to prevent freezing when outdoor temperature fell below –10° C. From April 2018 on, plants were checked every three to four days for flowering. The first plant flowered the 21st of April 2018, which was used as a reference

(day 1). About 14% of the replicates from the A lineage desiccated before the end of the experiment and did not reach flowering. To test for differences in flowering time between clades a linear mixed effects (LME) model was used, as implemented in the R package MCMCglmm (***Hadfield, 2010***). To control for the six biological replicates, individual accessions were treated as random effects in the model.

Acknowledgments

The authors are especially thankful to the Swiss National Science Foundation and the University Research Priority Programs (URPP) *Evolution in Action* for funding this project. AS, LAJM and RH acknowledge the funding from the National Science Centre Poland (grant no. 2015/18/M/NZ2/00394). The authors would like to thank the Genetic Diversity Center at ETH Zurich for providing access on their High Performance Computing resources as well as Christian Beisal at BSSE-Basel, Gerhard Herren and Helen Zbinden for their support in the lab. The authors also would like to thank Marco Maccaferri, Saverio Sciandrello and Pietro Minissale for their help in sampling in Italy, Jean-Paul Pelletier for establishing connection between Switzerland and Morocco and Christian Parisod for his comments on an early draft of the manuscript.

Competing interests

The authors declare that no competing interests exist.

References

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KMM, Cao J, Chae E, Dezwaan TMM, Ding W, Ecker JRR, Exposito-Alonso M, Farlow A, Fitz J, Gan X, Grimm DGG, Hancock AMM, Henz SRR, Holm S, Horton M, et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016; 166(2):481–491. doi: 10.1016/j.cell.2016.05.063.

Avise JC. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annual Review of Ecology and Systematics. 1987; 18(1987):489–522. doi: 10.1146/annurev.ecolsys.18.1.489.

Baack E, Melo MC, Rieseberg LH, Ortiz-Barrientos D. The origins of reproductive isolation in plants. New Phytologist. 2015; 207(4):968–984. doi: 10.1111/nph.13424.

Baker HG. Self-compatibility and establishment after 'long-distance' dispersal. Evolution. 1955; 9(3):347. doi: 10.2307/2405656.

Baker HG. Reproductive methods as factors in speciation in flowering plants. Cold Spring Harbor Symposia on Quantitative Biology. 1959; 24:177–191. doi: 10.1101/SQB.1959.024.01.019.

Barrett SCH. Darwin's legacy: the forms, function and sexual diversity of flowers. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010; 365(1539):351–368. doi: 10.1098/rstb.2009.0212.

Barrett SCH. Mating strategies in flowering plants: the outcrossing–selfing paradigm and beyond. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2003; 358(1434):991–1004. doi: 10.1098/rstb.2003.1301.

Barrett SCH, Eckert CG. Variation and evolution of mating systems in seed plants. In: Kawano S, editor. *Biological Approaches and Evolutionary Trends in Plants* London, UK: Academic Press; 1990.p. 229–254.

Blasi C, Capotorti G, Copiz R, Guida D, Mollo B, Smiraglia D, Zavattero L. Classification and mapping of the ecoregions of Italy. Plant Biosystems. 2014; 148(6):1255–1345. doi: 10.1080/11263504.2014.985756.

Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. PLoS Genetics. 2010; 6(3):1–14. doi: 10.1371/journal.pgen.1000890.

Bourgeois Y, Stritt C, Walser JC, Gordon SP, Vogel JP, Roulin AC. Genome-wide scans of selection highlight the impact of biotic and abiotic constraints in natural populations of the model grass *Brachypodium distachyon*. The Plant Journal. 2018; 96(2):438–451. doi: 10.1111/tpj.14042.

Brandenburg JT, Mary-Huard T, Rigaill G, Hearne SJ, Corti H, Joets J, Vitte C, Charcosset A, Nicolas SD, Tenaillon MI. Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLoS Genetics. 2017; 13(3):1–30. doi: 10.1371/journal.pgen.1006666.

Brown JKM, Tellier A. Plant-parasite coevolution: bridging the gap between genetics and ecology. Annual Review of Phytopathology. 2011; 49(1):345–367. doi: 10.1146/annurev-phyto-072910-095301.

Catalán P, López-Alvarez D, Bellosta C, Villar L. Updated taxonomic descriptions, iconography, and habitat preferences of *Brachypodium distachyon*, *B. stacei*, and *B. hybridum* (Poaceae). Anales del Jardin Botanico de Madrid. 2016; 73(1):1–14. doi: 10.3989/ajbm.2428.

Catalán P, Müller J, Hasterok R, Jenkins G, Mur LaJ, Langdon T, Betekhtin A, Siwinska D, Pimentel M, López-Alvarez D. Evolution and taxonomic split of the model grass *Brachypodium distachyon*. Annals of Botany. 2012; 109(2):385–405. doi: 10.1093/aob/mcr294.

Charlesworth D. Effects of inbreeding on the genetic diversity of populations. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2003; 358(1434):1051–1070. doi: 10.1098/rstb.2003.1296.

Conord C, Gurevitch J, Fady B. Large-scale longitudinal gradients of genetic diversity: A meta-analysis across six phyla in the Mediterranean Basin. Ecology and Evolution. 2012; 2(10):2600–2614. doi: 10.1002/ece3.350.

Dell'Acqua M, Zuccolo A, Tuna M, Gianfranceschi L, Pè M. Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: a multi-faceted approach. BMC Genomics. 2014; 15(1):801. doi: 10.1186/1471-2164-15-801.

Dray S, Dufour AB. The ade4 Package: Implementing the duality diagram for ecologists. Journal of Statistical Software. 2007; 22(4). doi: 10.18637/jss.v022.i04.

Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Picó FX, Alonso-Blanco C, Hancock AM. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences. 2017; 114(20):5213–5218. doi: 10.1073/pnas.1616736114.

Engelmann K, Purugganan M. The molecular evolutionary ecology of plant development: flowering time in *Arabidopsis thaliana*. Advances in Botanical Research. 2006; 44(06):507–526. doi: 10.1016/S0065-2296(06)44013-1.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genetics. 2013; 9(10):e1003905. doi: 10.1371/journal.pgen.1003905.

Flouri T, Jiao X, Rannala B, Yang Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Molecular Biology and Evolution. 2018; 35(10):2585–2593. doi: 10.1093/molbev/msy147.

Frichot E, François O. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution. 2015; 6(8):925–929. doi: 10.1111/2041-210X.12382.

Fulgione A, Koornneef M, Roux F, Hermisson J, Hancock AM. Madeiran *Arabidopsis thaliana* reveals ancient long-range colonization and clarifies demography in Eurasia. Molecular Biology and Evolution. 2017; 35(3):564–574. doi: 10.1093/molbev/msx300.

Gaut BS, Seymour DK, Liu Q, Zhou Y. Demography and its effects on genomic variation in crop domestication. Nature Plants. 2018; 4(8):512–520. doi: 10.1038/s41477-018-0210-1.

Gibson DJ. Grasses & Grassland Ecology. Oxford University Press; 2009.

Giraldo P, Rodríguez-Quijano M, Vázquez JF, Carrillo JM, Benavente E. Validation of microsatellite markers for cytotype discrimination in the model grass *Brachypodium distachyon*. Genome. 2012; 55(7):523–527. doi: 10.1139/g2012-039.

Glémin S, Ronfort J. Adaptation and maladaptation in selfing and outcrossing species: New mutations versus standing variation. Evolution. 2013; 67(1):225–240. doi: 10.1111/j.1558-5646.2012.01778.x.

Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J, Lipzen A, Dochy N, Phillips J, Barry K, Geuten K, Budak H, Juenger TE, Amasino R, Caicedo AL, et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nature Communications. 2017; 8(1). doi: 10.1038/s41467-017-02292-8.

Griffin PC, Willi Y. Evolutionary shifts to self-fertilisation restricted to geographic range margins in North American *Arabidopsis lyrata*. Ecology Letters. 2014; 17(4):484–490. doi: 10.1111/ele.12248.

Grove A, Rackham O. The Nature of Mediterranean Europe - An Ecological History. Yale University Press; 2001.

Hadfield JD. MCMC methods for multi-respoinse generalized linear mixed models: The MCMCglmm R package. Journal of Statistical Software. 2010; 33(2):1–22. doi: 10.1002/ana.22635.

Hartfield M, Bataillon T, Glémin S. The evolutionary interplay between adaptation and self-fertilization. Trends in Genetics. 2017; 33(6):420–431. doi: 10.1016/j.tig.2017.04.002.

Hedrick P. Genetics of Populations. Jones & Bartlett Publishers; 2011.

Hewitt GM. Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2004; 359(1442):183–195. doi: 10.1098/rstb.2003.1388.

Hey J, Machado CA. The study of structured populations–new hope for a difficult and divided science. Nature Reviews Genetics. 2003; 4(7):535–43. doi: 10.1038/nrg1112.

Hu XS. Mating system as a barrier to gene flow. Evolution. 2015; 69(5):1158–1177. doi: 10.1111/evo.12660.

Huo N, Garvin DF, You FM, McMahon S, Luo MC, Gu YQ, Lazo GR, Vogel JP. Comparison of a high-density genetic linkage map to genome features in the model grass *Brachypodium distachyon*. Theoretical and Applied Genetics. 2011; 123(3):455–464. doi: 10.1007/s00122-011-1598-4.

Jakob SS, Blattner FR. A chloroplast genealogy of *Hordeum* (Poaceae): Long-term persisting haplotypes, incomplete lineage sorting, regional extinction, and the consequences for phylogenetic inference. Molecular Biology and Evolution. 2006; 23(8):1602–1612. doi: 10.1093/molbev/msl018.

Karasov TL, Horton MW, Bergelson J. Genomic variability as a driver of plant-pathogen coevolution? Current Opinion in Plant Biology. 2014; 18(1):24–30. doi: 10.1016/j.pbi.2013.12.003.

Khan MA, Stace Ca. Breeding relationships in the genus *Brachypodium* (Poaceae: Pooideae). Nordic Journal of Botany. 1999; 19(3):257–269. doi: 10.1111/j.1756-1051.1999.tb01108.x.

Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genetics. 2012; 8(1):11–17. doi: 10.1371/journal.pgen.1002453.

Lee Cr, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, Alonso-Blanco C, Weigel D, Nordborg M. On the post-glacial spread of human commensal *Arabidopsis thaliana*. Nature Communications. 2017; 8:14458. doi: 10.1038/ncomms14458.

Lee TH, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC genomics. 2014; 15(1):162. doi: 10.1186/1471-2164-15-162.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; doi: 10.1093/bioinformatics/btw748.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079. doi: 10.1093/bioinformatics/btp352.

López-Alvarez D, Manzaneda AJ, Rey PJ, Giraldo P, Benavente E, Allainguillaume J, Mur L, Caicedo AL, Hazen SP, Breiman A, Ezrati S, Catalán P. Environmental niche variation and evolutionary diversification of the *Brachypodium distachyon* grass complex species in their native circum-Mediterranean range. American Journal of Botany. 2015; 102(7):1073–1088. doi: 10.3732/ajb.1500128.

López-Alvarez D, López-Herranz ML, Betekhtin A, Catalán P. A DNA barcoding method to discriminate between the model plant *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). PLoS ONE. 2012; 7(12):e51058. doi: 10.1371/journal.pone.0051058.

Manzano P, Malo JE. Extreme long-distance seed dispersal via sheep. Frontiers in Ecology and the Environment. 2006; 4(5):244–248. doi: 10.1890/1540-9295(2006)004[0244:ELSDVS]2.0.CO;2.

Marques I, Shiposha V, López-Alvarez D, Manzaneda AJ, Hernandez P, Olonova M, Catalán P. Environmental isolation explains Iberian genetic diversity in the highly homozygous model grass *Brachypodium distachyon*. BMC Evolutionary Biology. 2017; 17(1):139. doi: 10.1186/s12862-017-0996-x.

Médail F, Diadema K. Glacial refugia influence plant diversity patterns in the Mediterranean Basin. Journal of Biogeography. 2009; 36(7):1333–1345. doi: 10.1111/j.1365-2699.2008.02051.x.

Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. Journal of Computational Biology. 2006; 13(5):1028–1040. doi: 10.1089/cmb.2006.13.1028.

Myers N, Mittermeier R, Mittermeier C, da Foncesca G, Kent J. Biodiversity hotspots for conservation priorities. Nature. 2000; 403: 853(February):853–858.

Nieto Feliner G. Patterns and processes in plant phylogeography in the Mediterranean Basin. A review. Perspectives in Plant Ecology, Evolution and Systematics. 2014; 16(5):265–278. doi: 10.1016/j.ppees.2014.07.002.

Nordborg M, Charlesworth B, Charlesworth D. Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proceedings of the Royal Society B: Biological Sciences. 1996; 263(1373):1033–1039. doi: 10.1098/rspb.1996.0152.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 2010; 327(5961):92–94. doi: 10.1126/science.1180677.

Pavlidis P, Alachiotis N. A survey of methods and tools to detect recent and strong positive selection. Journal of Biological Research-Thessaloniki. 2017; 24(1):7. doi: 10.1186/s40709-017-0064-0.

Pebesma EJ, Bivand RS. Classes and methods for spatial data in R. R News. 2005; 5(2):9–13. http://cran.r-project.org/doc/Rnews/.

Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. doi: 10.1093/bioinformatics/btq033.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic Biology. 2018; 67(5):901–904. doi: 10.1093/sysbio/syy032.

Rannala B, Yang Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 2003; 1656(August):1645–1656.

Ream TS, Woods DP, Schwartz CJ, Sanabria CP, Mahoy JA, Walters EM, Kaeppler HF, Amasino RM. Interaction of photoperiod and vernalization determines flowering time of *Brachypodium distachyon*. Plant Physiology. 2014; 164(2):694–709. doi: 10.1104/pp.113.232678.

Ross-Ibarra J, Morrell PL, Gaut BS. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceedings of the National Academy of Sciences. 2007; 104(Supplement 1):8641–8648. doi: 10.1073/pnas.0700643104.

Rundel PW, Arroyo MTK, Cowling RM, Keeley JE, Lamont BB, Vargas P. Mediterranean biomes: evolution of their vegetation, floras and climate. Annual Review of Ecology, Evolution, and Systematics. 2016; 47(September):383–407. doi: 10.1146/annurev-ecolsys-121415-032330.

San Miguel A. Management of Natura 2000 habitats. Pseudo-steppe with grasses and annuals of the Thero-Brachypodietea; 2008.

Sancho R, Cantalapiedra CP, López-Alvarez D, Gordon SP, Vogel JP, Catalán P, Contreras-Moreira B. Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time signatures, introgression and recombination in recently diverged ecotypes. New Phytologist. 2017; 218(4):1631–1644. doi: 10.1111/nph.14926.

Scholthof KBG, Irigoyen S, Catalán P, Mandadi KK. Brachypodium: A Monocot Grass Model Genus for Plant Biology. Plant Cell Advance Publication Published on. 2018; doi: 10.1105/tpc.18.00083.

Sensoy S. Climate of Turkey. Turkish State Meteorological Service; 2008.

Sharma N, Ruelens P, D'Hauw M, Maggen T, Dochy N, Torfs S, Kaufmann K, Rohde A, Geuten K. A flowering locus C homolog is a vernalization-regulated repressor in *Brachypodium* and is cold-regulated in Wheat. Plant Physiology. 2017; 173(2):1301–1315. doi: 10.1104/pp.16.01161.

Skalska A, Stritt C, Wyler M, Williams HW, Vickers M, Han J, Tuna M, Tuna GS, Susek K, Swain M, Rafa K. Genetic and methylome variation in Turkish *Brachypodium distachyon* accessions differentiate two geographically distinct populations. submitted. 2020; p. 1–17.

Spence JP, Steinrücken M, Terhorst J, Song YS. Inference of population history using coalescent HMMs: review and outlook. Current Opinion in Genetics and Development. 2018; 53:70–76. doi: 10.1016/j.gde.2018.07.002.

Stebbins GL. Self fertilization and population variability in the higher plants. The American Naturalist. 1957; 91(861):337–354. doi: 10.1086/281999.

Stein N, Herren G, Keller B. A new DNA extraction method for high-throughput marker analysis in a large-genome species such as *Triticum aestivum*. Plant Breeding. 2001; 120(4):354–356. doi: 10.1046/j.1439-0523.2001.00615.x.

Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass *Brachypodium distachyon*. Genome Biology and Evolution. 2018; 10(1):304–318. doi: 10.1093/gbe/evx276.

Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass *Brachypodium distachyon*. New Phytologist. 2019; 227(6):1736–1748. doi: 10.1111/nph.16308.

Sultan SE, Spencer HG. Metapopulation structure favors plasticity over local adaptation. The American Naturalist. 2002; 160(2):271–283. doi: 10.1086/341015.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics. 2015; 31(12):2032–2034. doi: 10.1093/bioinformatics/btv098.

Thompson JD. Plant Evolution in the Mediterranean. Oxford University Press; 2005. doi: 10.1086/523147.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013; 14(2):178–192. doi: 10.1093/bib/bbs017.

Tyler AL, Lee SJ, Young ND, Deiulio GA, Benavente E, Reagon M, Sysopha J, Baldini RM, Troìa A, Hazen SP, Caicedo AL. Population structure in the model grass *Brachypodium distachyon* is highly correlated with flowering differences across broad geographic areas. The Plant Genome. 2016; p. 1–55. doi: 10.3835/plantgenome2015.08.0074.

Vogel JP. Genetics and Genomics of Brachypodium. Springer International Publishing; 2016.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, Tice H, Grimwood J, McKenzie N, Huo N, Gu YQ, Lazo GR, Anderson OD, You FM, Luo MC, Dvorak J, et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010; 463(7282):763–768. doi: 10.1038/nature08747.

Vogel JP, Tuna M, Budak H, Huo N, Gu YQ, Steinwand Ma. Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. BMC Plant Biology. 2009; 9:88. doi: 10.1186/1471-2229-9-88.

Weigel D, Nordborg M. Population genomics for understanding adaptation in wild plant species. Annual review of genetics. 2015; 49(September):315–338. doi: 10.1146/annurev-genet-120213-092110.

Wright SI, Kalisz S, Slotte T. Evolutionary consequences of self-fertilization in plants. Proceedings of the Royal Society B: Biological Sciences. 2013; 280(1760):20130133. doi: 10.1098/rspb.2013.0133.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012; 28(24):3326–3328. doi: 10.1093/bioinformatics/bts606.

Supplementary files

Supplementary file 1: Overview of population genetic studies with *B. distachyon*

Supplementary file 2: Sample information, including coordinates, sequencing depth, inbreeding coefficients

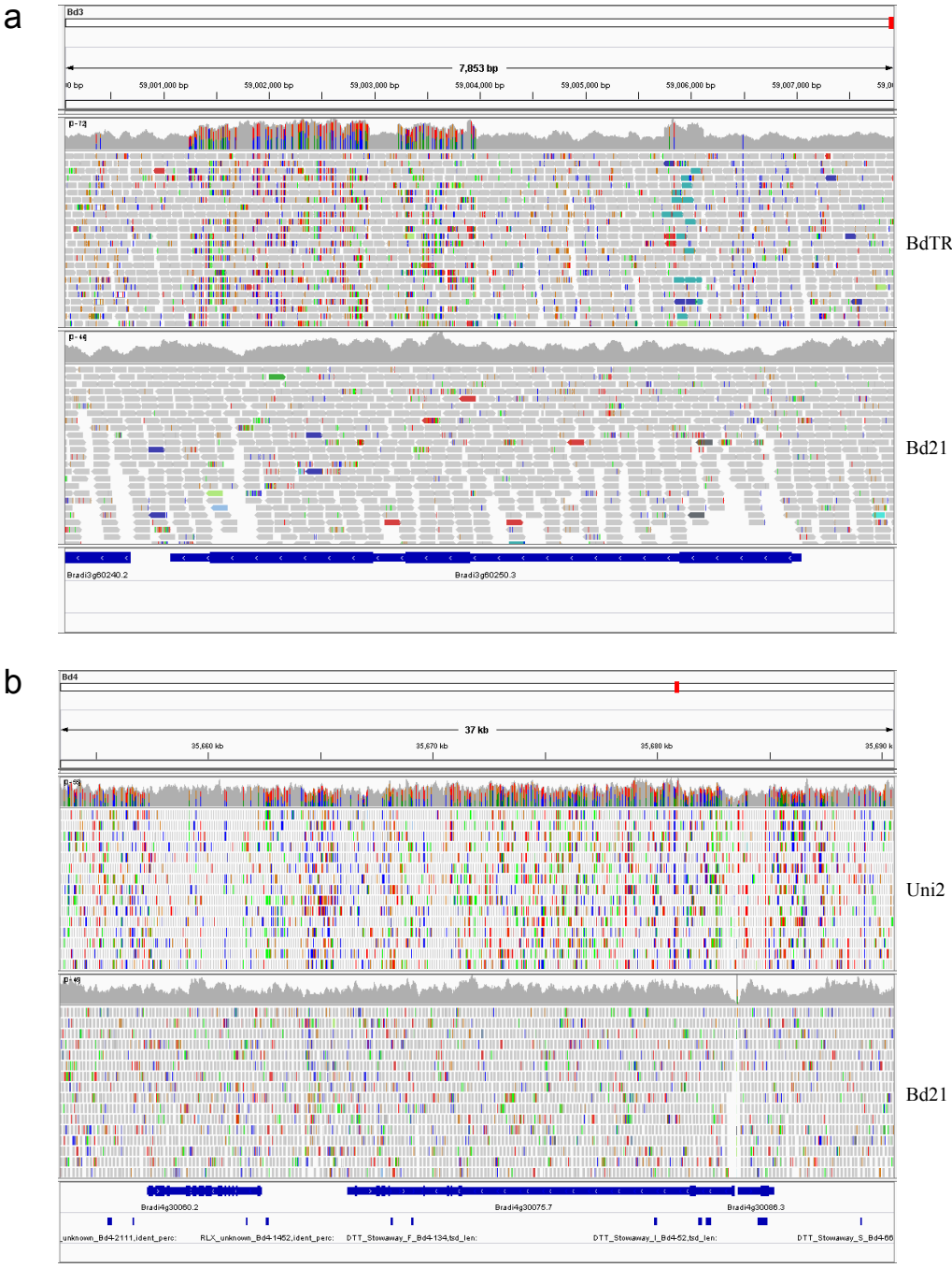

Figure 1–Figure supplement 1. Two islands of extended heterozygosity (IEH). Plots show paired-end reads aligned against the reference genome of two accessions: the reference itself (Bd21) and the two accessions BdTR7a and Uni2. **a)** An IEH cause by a gene copy number variant. Heterozygosity conincides with a doubling of sequencing depth. **b)** An IEH showing real heterozygosity due to recent admixture. Here heterozygosity stretches over long distances and the sequencing depth is within the normal range.

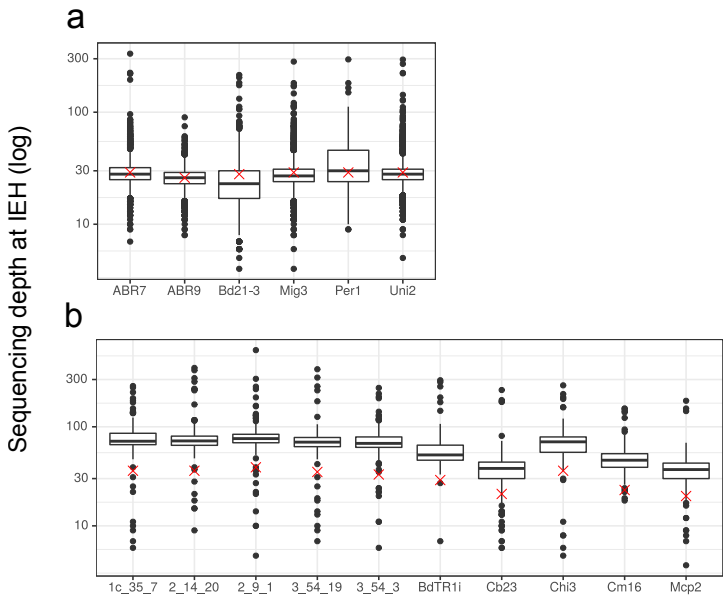

Figure 1–Figure supplement 2. Read coverage in islands of extended heterozygosity in **a)** the six outlier accessions and **b)** ten random accessions. Red crosses show the median read coverage across the genome for each accession, while the boxplots show read coverages in IEH.

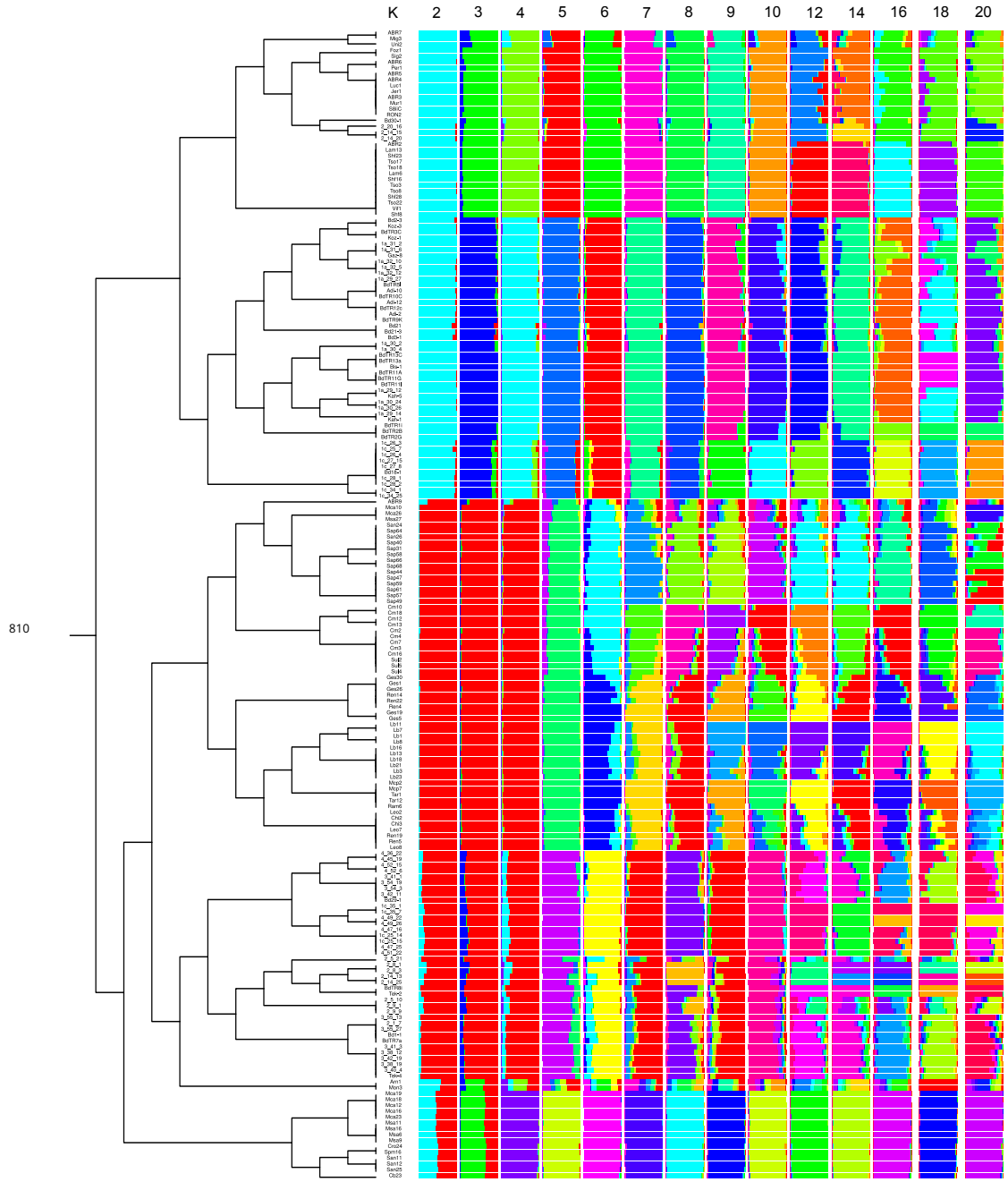

Figure 2–Figure supplement 1. sNMF analysis with K from 2 to 20. Barplots are ordered according to the fineSTRUCTURE phylogeny, depicted on the left.

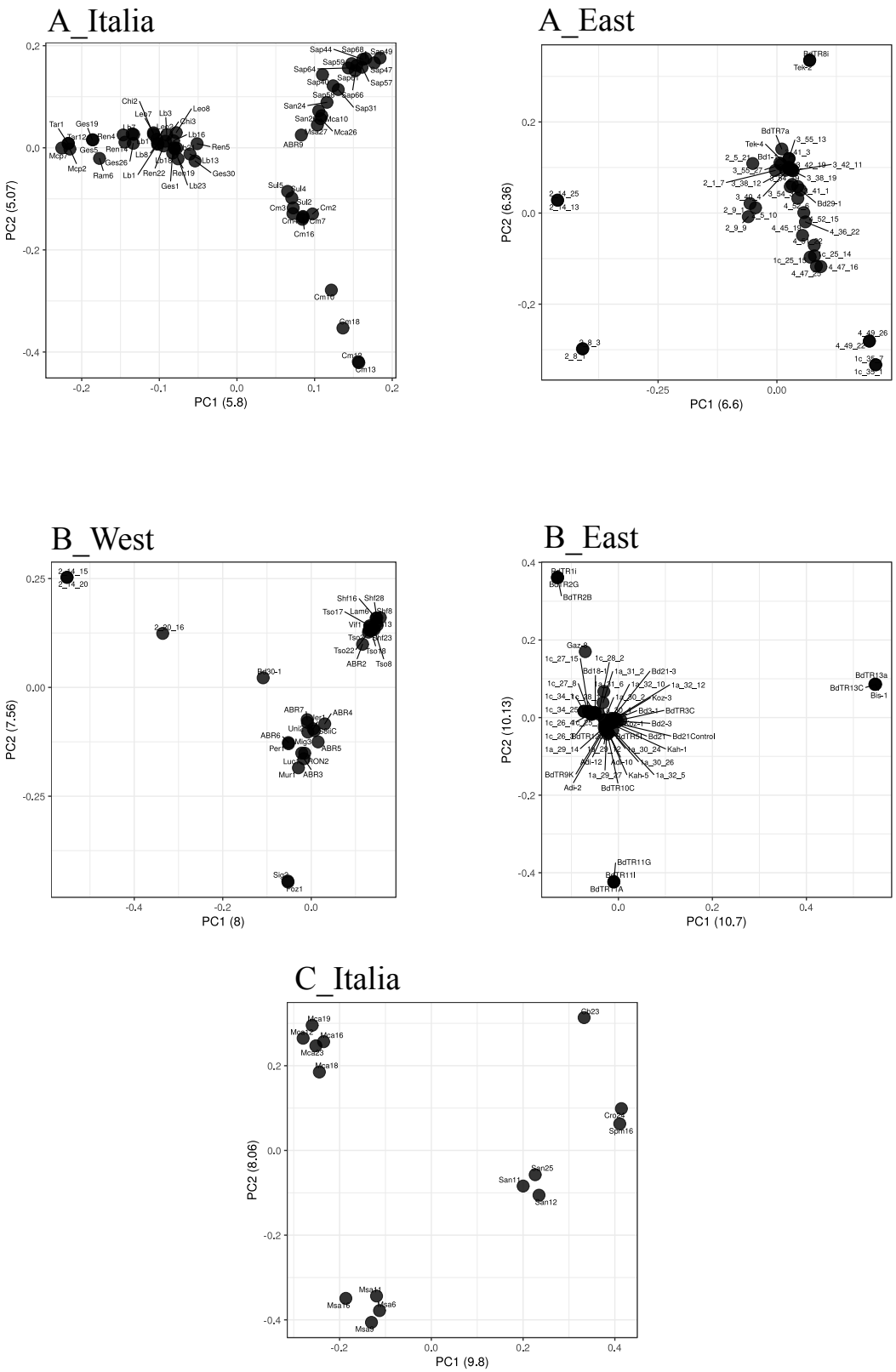

Figure 2–Figure supplement 2. Seperate PCAs for each of the five geographic clades.

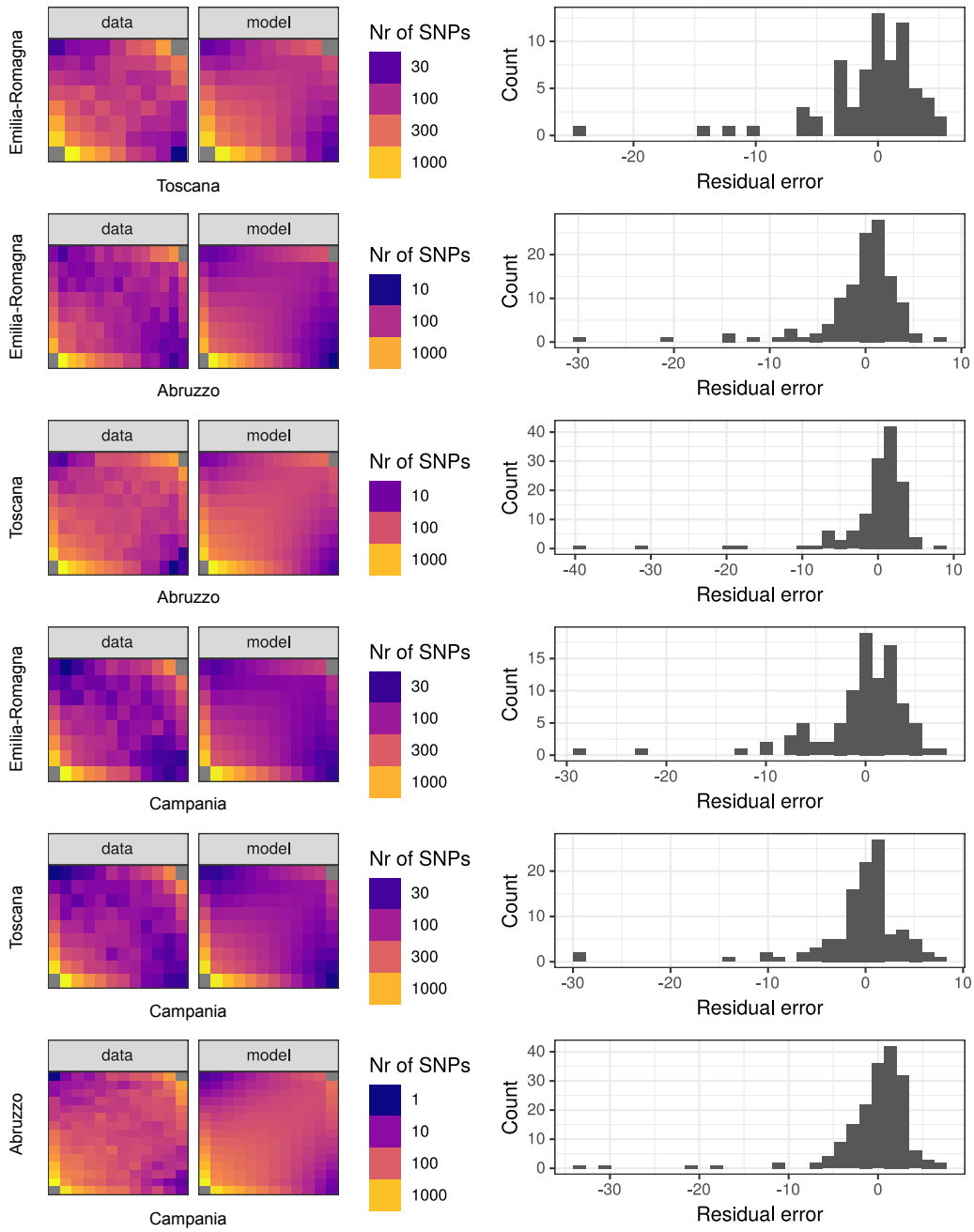

Figure 5–Figure supplement 1. Derived allele frequency spectra (DAF) and residual errors of the model with the highest likelihood. Each row shows a population comparison with observed and predicted DAF, the colors indicating the number of synonymous SNPs in each frequency class (log scale). The histograms show the Poisson residuals, calculated as $(model - data)/\sqrt{(model)}$. The largest deviations occur among high frequency shared alleles, of which the model predicts less than observed.