On climate change and genetic evolution in *Arabidopsis thaliana*

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Moises Exposito-Alonso
aus Alacant, Spanien

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 30.10.2018

Dekan	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter	Prof. Dr. Detlef Weigel
2. Berichterstatter	Prof. Dr. Oliver Bossdorf
3. Berichterstatter	Prof. Dr. Jon Ågren

A mi familia.

To my friends.

To our beautiful Planet Earth.

Table of content

Summary

Global climate change is already impacting Earth's biodiversity, but we are still struggling to understand which species will perish and which will thrive. As many species will not tolerate a rapidly-changing climate nor migrate fast enough to escape it, survival will depend on whether populations are able to genetically adapt. Some species, however, seem to rapidly adapt and spread in the new *status quo* of human-dominated ecosystems. We are just beginning to understand the genomic footprints of past adaptation to climates and how this has prepared populations for future rapid adaptation, but many questions still need to be answered. Furthermore, evolution and adaptation knowledge is rarely integrated into predictive biodiversity models, even though that would increase the accuracy of predictions and help design better conservation strategies. Here I aim to tackle those challenges using the mustard-related plant *Arabidopsis thaliana*, for which there are public genomic sequences, geographic information, and seed collections of thousands of individuals.

Chapter One was my first approach to understand how populations of the same species might respond to climate change. I examined survival of 220 natural *Arabidopsis thaliana* lines whose genomes are known to a simulated extreme drought in the greenhouse. Severe droughts are being forecast as some of the most drastic threats for plant communities as a consequence of global change. Extending the use of environmental niche models in combination with genome-wide association techniques, I found the hotspots of adaptive variants are primarily at the North and South margins of the species' distribution range. The populations at those areas, that live in more extreme environments, will perhaps become reservoirs of adaptive variation under future, more hostile climates.

In Chapter Two, I carried out a large-scale field experiment to directly quantify climate-driven selection in natural conditions. We planted a global panel of 517 natural *A. thaliana* lines in rainfall-manipulated common gardens both in a region with a moderate climate, in Central Europe, and in a region with a more extreme environment, the Mediterranean. Using image analysis to estimate reproductive success, I generated close to

25,000 fitness measurements. Combining fitness and genomic data, I could infer massive changes in genome-wide allele frequencies within one generation, especially under hot temperatures and reduced precipitation where many Central European genotypes died. Integrating the theory of local adaptation with machine learning tools, I showed that a significant portion of natural selection is predictable from the climate at the geographic areas where genetic variants are found. Following a decrease in rainfall in the future, I then predicted that the intensity of natural selection will increase the most in transition areas from the Mediterranean to Central Europe, putting populations at evolutionary risk. This is in stark contrast to the generally accepted notion that marginal "warming" populations are at higher risk of extinction than populations at the center of the geographic distribution.

Chapter Three, in contrast to the previous chapters that studied the adaptive value of pre-existent variants to future climate change, focuses on how novel mutations could directly contribute to adaptation. Using herbarium samples as genetic snapshots in time, I studied a 400-year-old lineage of *A. thaliana* that was isolated in North America. I was able to identify over 5,000 new mutations, some of which generated novel morphological differences likely related to adaptation to the newly colonized continent. I concluded that even large organisms such as plants might evolve and adapt from new mutations in contemporary timescales.

This work advances our knowledge on how and whether different populations of a species will genetically adapt to the changing climate. Some of the insights generated here include (1) that adaptation to climate occurs thanks to hundreds of genetic variants (polygenic adaptation), (2) that new mutations occur often enough that they could contribute to rapid adaptation in colonizing populations, and (3) that statistical models that learn the relationship between current climates and genetic variants can be used to predict whether populations will have the appropriate genetic makeup to adapt to climate change or whether they will be at evolutionary risk. All in all, these studies move us one step closer to address ecological challenges using the genetic theory of evolution.

Zusammenfassung

Der globale Klimawandel beeinflusst schon jetzt die Biodiversität der Erde. Dennoch kämpfen wir weiterhin damit, zu verstehen, wie Arten reagieren werden. Weil viele Arten ein sich schnell veränderndes Klima nicht tolerieren werden, oder nicht schnell genug migrieren können, um eben jenem zu entrinnen, wird ein Überleben davon abhängen, ob Populationen in der Lage sind, sich genetisch anzupassen. Andere Arten aber scheinen sich rasch anzupassen und zu verbreiten im neuen *Status Quo* unseres vom Menschen dominierten Ökosystems. Wir beginnen gerade erst damit, die genomischen Fußabdrücke vergangener Anpassungen ans Klima zu verstehen, und wie diese Populationen für zukünftige Anpassungen vorbereitet haben. Gleichzeitig sind viele Fragen nach wie vor unbeantwortet. Zudem wird das Wissen um Evolution und Anpassung kaum in Modelle integriert, die Biodiversität voraussagen, obwohl eine solche Integration die Genauigkeit der Vorhersagen steigern sowie helfen würde, bessere Konservations-Strategien zu entwerfen. In der vorliegenden Doktorarbeit möchte ich diese Herausforderungen angehen.

Kapitel Eins war mein erster Ansatz, zu verstehen, wie Populationen der gleichen Art auf den Klimawandel reagieren könnten. Ich untersuchte das Überleben von 220 natürlichen *Arabidopsis thaliana* (Ackerschmalwand) Linien unter simulierter extremer Dürre im Gewächshaus. Strenge Dürren, eine Konsequenz des globalen Wandels, werden als eine der drastischsten Bedrohungen für Pflanzengemeinschaften vorhergesagt. Indem ich die Anwendung von Ökologischen Nischenmodellen mit Genomweiten Assoziationstechniken erweiterte, fand ich, dass eine Reihe adaptiver genetischer Varianten primär an den Rändern des Verbreitungsgebiets von Ackerschmalwand präsent war. Möglicherweise werden diese Populationen unter zukünftigen, “feindlichen” Klimabedingungen durch ihr momentanes Dasein in extremeren Umgebungen zu Quellen adaptiver Vielfalt.

In Kapitel Zwei realisierte ich einen großangelegten Feldversuch, um direkt Klima-gesteuerte Selektion unter natürlichen Bedingungen zu quantifizieren. Wir pflanzten eine globale Auswahl von 517 natürlichen Ackerschmalwand-Linien in einem Regenfall-manipulierten

Common Garden Experiment sowohl in einer Region mit moderatem Klima in Mitteleuropa, als auch in einer Region mit extremeren Klimabedingungen, im Mittelmeerraum. Mittels Bildanalysen zum Abschätzen von Samenproduktion/reproduktivem Erfolg erzeugte ich nahezu 25.000 Fitness-Messungen. Mit diesen Daten konnte ich massive Veränderungen in genomweiten Allelfrequenzen innerhalb einer Generation ableiten, besonders bei hohen Temperaturen und verringertem Regenfall, Bedingungen, unter welchen viele mitteleuropäische Genotypen vertrockneten. Indem ich Theorien zu lokaler Anpassung mit maschinellem Lernen verknüpfte, zeigte ich, dass ein signifikanter Anteil der vom Klima gelenkten natürlichen Selektion vorhersagbar ist. Damit, kombiniert mit dem Wissen um lokale genetische Vielfalt, mache ich die Vorhersage, dass Populationen in den Übergangsgebieten zwischen Mittelmeer und Mitteleuropa das höchste evolutionäre Risiko tragen, wenn Regenfall in der Zukunft plötzlich abnimmt. Dies steht in großem Kontrast zur generell akzeptierten Idee, dass marginell "wärmere" Populationen einem höheren Risiko Auszusterben ausgesetzt sind als solche Populationen, die sich im Zentrum der geographischen Ausdehnung einer Art befinden.

Kapitel Drei widmet sich, im Gegensatz zu den vorherigen Kapiteln, deren Fokus auf dem adaptiven Wert bereits existierender genetischer Varianten im Angesicht des Klimawandels lag, neuen Mutationen, und wie diese zur Anpassung beitragen könnten. Ich nutzte Herbariumproben als "genetische Schnappschüsse" durch die Zeit, um eine 400 Jahre alte *A. thaliana* Linie zu untersuchen, die sich isoliert in Nordamerika befand. Es gelang mir, über 5.000 neue Mutationen zu identifizieren, von denen einige neue morphologische Unterschiede verursachen, die wahrscheinlich mit Anpassung in Verbindung stehen. Ich zog den Schluss, dass selbst große Organismen wie Pflanzen bereits in verhältnismäßig zeitnahen Zeitrahmen nur auf der Basis neuer Mutationen evolvieren und sich anpassen könnten.

Die vorliegende Arbeit treibt unser Wissen darüber voran, wie und ob sich verschiedene Populationen derselben Art genetisch an das sich wandelnde Klima anpassen werden. Die hervorgebrachten Einsichten beinhalten unter anderem, dass (1) Anpassung an das Klima mittels hunderter genetischer Varianten erfolgt (polygenetische Anpassung), (2) neue Mutationen oft genug auftreten, um zu rascher Anpassung sich neu ansiedlender

Populationen beitragen zu können, und (3) dass wir statistische Modelle, die den Zusammenhang zwischen gegenwärtigen Klimabedingungen und genetischen Varianten lernen, dazu benutzen können um vorherzusagen, ob Populationen die entsprechende genetische Zusammensetzung für eine Anpassung an den Klimawandel mitbringen, oder ob für sie ein evolutionäres Risiko vorliegt. *Summa summarum* bringen uns diese Studien dem Ziel, ökologische Herausforderungen mittels der genetischen Evolutionstheorie anzugehen, einen Schritt näher.

Publications

Published

Exposito-Alonso, M., Becker, C., Schuenemann, V.J., Reiter, E., Setzer, C., Slovak, R., Brachi, B., Hagmann, J., Grimm, D.G., Jiahui, C., Busch, W., Bergelson, J., Ness, R.W., Krause, J., Burbano, H.A., Weigel, D., (2018). **The rate and effect of new mutations in a colonizing plant lineage.** PLOS Genetics, https://doi.org/10.1371/journal.pgen.1007155 | (2016) bioRxiv, https://doi.org/10.1101/050203.

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H.A., Weigel, D., (2018). **Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana***. Nature Ecology & Evolution 2, 352–358. https://doi.org/10.1038/s41559-017-0423-0 | (2017) bioRxiv, https://doi.org/10.1101/118067.

Under review

Exposito-Alonso, M., 500 Genomes Field Experiment Team, Burbano, H. A., Bossdorf, O., Nielsen, R., Weigel, D. (2018) **A map of climate change-driven selection in *Arabidopsis thaliana***. bioRxiv, https://doi.org/10.1101/321133.

Exposito-Alonso, M., Rodríguez, R.G., Barragán, C., Capovilla, G., Chae, E., Devos, J., Dogan, E.S., Friedemann, C., Gross, C., Lang, P., Lundberg, D., Middendorf, V., Kageyama, J., Karasov, T., Kersten, S., Petersen, S., Rabbani, L., Regalado, J., Reinelt, L., Rowan, B., Seymour, D.K., Symeonidi, E., Schwab, R., Tran, D.T.N., Venkataramani, K., Van de Weyer, A.-L., Vasseur, F., Wang, G., Wedegärtner, R., Weiss, F., Wu, R., Xi, W., Zaidem, M., Zhu, W., García-Arenal, F., Burbano, H.A., Bossdorf, O., Weigel, D., (2017) **A rainfall-manipulation experiment with 517 *Arabidopsis thaliana* accessions**. bioRxiv, https://doi.org/10.1101/186767.

Published manuscripts not included in this dissertation

Exposito-Alonso, M., Brennan, A., Alonso-Blanco, C., Picó, F.X., (2018). **Spatio-temporal variation in fitness responses to contrasting environments in *Arabidopsis thaliana***. Evolution, http://doi.org/10.1111/evo.13508.

Vasseur, F., Exposito-Alonso, M., Ayala-Garay, O., Wang, G., Enquist, B.J., Violle, C., Ville, D., Weigel, D., (2018). **Adaptive diversification of growth allometry in the plant *Arabidopsis thaliana***. Proceedings of the National Academy of Sciences U.S.A, https://doi.org/10.1073/pnas.1709141115.

Lee, C.-R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., Alonso-Blanco, C., Weigel, D., Nordborg, M., (2017). **On the post-glacial spread of human commensal *Arabidopsis thaliana*.** Nature Communications. 8, 14458.

https://doi.org/10.1038/ncomms14458.

Iakovidis, M., Teixeira, P.J.P.L., Exposito-Alonso, M., Cowper, M.G., Law, T.F., Liu, Q., Vu, M.C., Dang, T.M., Corwin, J.A., Weigel, D., Dangl, J.L., Grant, S.R., (2016). **Effector-triggered immune response in *Arabidopsis thaliana* is a quantitative trait**. Genetics 204, 337–353. https://doi.org/10.1534/genetics.116.190678.

1001 Genomes Consortium, (2016). **1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana***. Cell 166, 481–491. https://doi.org/10.1016/j.cell.2016.05.063.

INTRODUCTION

Humans are dramatically impacting the Earth — its surface, atmosphere and oceans —, and seriously threaten many of the life forms inhabiting it (Steffen et al. 2015, Newbold et al. 2016). Not only there are reasons to preserve biodiversity because of its intrinsic value and beauty, but also because it provides the natural resources and ecosystem services to sustain all human populations (Pearson 2016). In order to preserve biodiversity and ecosystems, we first need to improve our understanding of how species adapt to the environment, and make robust predictions of extinction risk. In this thesis, I use the model plant *Arabidopsis thaliana* to ask (1) what are the genetic determinants of climate adaptation in plants, (2) where geographically can we find the most adaptive variation in a species, and (3) how can we make geographically-explicit predictions of evolutionary risk — the risk of failing to adapt. To put my work into context, I below introduce some necessary concepts from different research fields and describe several major breakthroughs in these fields from a chronological perspective. Section One introduces the main problem of biodiversity loss driven by climate change, and explains the first ecological approaches used to tackle this problem. As researchers over time have realized how important it is to account for evolution in biodiversity projections, I then move in Section Two to describe the foundation of the field of evolutionary genetics. I specifically review the contributions of this field towards understanding how species adapt to new environments. In Section Three I discuss two general areas of interaction between ecology and genetic evolution that I will explore with my study system, which I describe in Section Four. Finally n Section Five I present the specific questions pursued in this thesis.

1. Species fate in the Anthropocene

Replacement of natural ecosystems by human systems is the most important factor leading to the loss of species on our planet (Vitousek et al. 1997, Urban 2015, Newbold et al. 2016). This reduction of worldwide species is estimated to be 1000 times faster than the "background" species loss (i.e. without humans (Pimm et al. 2014)) since the 1900s. Losses

of native species exceed 40% in many regions of the world (Newbold et al. 2016). This led Max Planck scientist and Nobel Laureate P. K. Crutzen and colleagues, to name the present geological era as the Anthropocene (Crutzen 2002), and the current ongoing extinction as the Anthropocene mass extinction — the 6^{th} mass extinction in Earth's life history. On the other hand, counterintuitively, when looking at reduced local geographic scales, the number of species present is currently increasing in many regions (Newbold et al. 2016). This is because a minority of species has benefited from hitchhiking along human trading and traveling routes to spread over many new territories (Tatem 2009, Seebens et al. 2017). Therefore, the observed local increase of biodiversity is due to the arrival of foreign species, in some areas thousands of them (van Kleunen et al. 2015a, Seebens et al. 2017). When these species rapidly spread and increase in abundance, or even harm native ecosystems, they are called invasive species (Dukes and Mooney 1999, Mooney and Cleland 2001, van Kleunen et al. 2015b, Seebens et al. 2017). These two human-driven biodiversity alterations, the extinction of many species and the rapid spread of a human-commensal minority of species, are a longstanding concern. That is what sparked in 1948 the creation of the International Union for Conservation of Nature (IUCN). This organization created the red list of species (www.iucnredlist.org) to monitor biodiversity changes and to provide policy guidelines to conserve present-day biodiversity and ecosystems.

There are many other human-driven impacts on biodiversity that are more subtle than those derived from direct habitat destruction. These are caused by the progressive alteration of Earth's conditions, prominently, the increase in CO_2 emissions and the consequential global warming (Rosenzweig et al. 2008) (others, which I am not discussing here, are pollution and landscape fragmentation, see (Fahrig 2003, Zalasiewicz et al. 2017)). In response to rising concern over accumulating evidence for soot and CO_2 increase in the atmosphere, researchers in the 1970s started to develop models to project the potential change in climate. Understandably, the first predictions resulted in heated debates and controversy over whether there will be a warming trend (via CO_2) or a cooling trend (via air-suspended particles) (Edwards 2011). This confusion was of such magnitude that the U.S. National Academy of Sciences stated already over 40 years ago (National Academy of Sciences 1975): "we do not have a good quantitative understanding of our climate [... so] it does not seem possible to predict climate". This of course has rapidly changed since the

1980s, because of better data acquisition and monitoring of climate, new mathematical models, and the commitment of many countries in the formation of the Intergovernmental Panel For Climate Change in 1988 (IPCC, www.ipcc.ch). These efforts resulted in complex 3D models of the Earth's Ocean-Atmosphere systems that already in the 1980s predicted with reasonable accuracy the climate change we are currently experiencing (Intergovernmental Panel on Climate Change 2014, Hausfather 2017).

A similar trend is occurring nowadays in the fields of ecology and conservation biology. There are now very consistent and well-documented responses of many species specifically to climate change (i.e. not due to habitat destruction). Generally, species are losing populations at the warm (typically southern) margin of their geographic distribution (Pimm et al. 2014), while their populations in the cold (typically northern or high altitude) range margin are migrating northwards (or upwards) and invading new territories (Parmesan and Yohe 2003, Seebens et al. 2015). Despite this, researchers have warned that we are not yet able to accurately assess and predict abundance and distribution changes of species (Pearson and Dawson 2003, 2004, Fordham et al. 2012), what is the necessary first step to devise effective conservation policies (Dawson et al. 2011). The original attempts to make predictions of species responses to climate change came from modeling climate tolerance limits and migration rates of species, but, as described below, it has become increasingly clear that we must, among others, also take into account the genetic diversity and evolutionary potential of species.

1.1. Forecasting biodiversity changes

The typical forecast of biodiversity changes driven by climate change is based on Hutchinson's classic concept of the ecological niche (1957). The set of environments where a species successfully survives and reproduces represent the species' environmental niche, which is limited due to certain physiological constraints. The so-called "Environmental Niche Models" (ENMs) (Guisan and Thuiller 2005), infer the environmental niche limits of a species by overlapping climate maps with maps of sightings of a species (they can also include land use or other features of the landscape). Both types of data are very abundant nowadays. For example, worldclim.org provides maps at 1 km^2 resolution of monthly climate averages from

1960-1990 (Hijmans et al. 2005). The Global Biodiversity Information Facility, GBIF.org, contains over 1 billion sightings of over 1 million species. If the sightings-based geographic limits of a species coincide with the environmental niche (Guisan and Thuiller 2005), ENM would correctly estimate the tolerance limits of a species, i.e. the environmental ranges within which it can survive. Once the environmental limits of a species are known, researchers can explore how the geographic limits of the species distribution would shrink or move given the IPCC climate map projections to the future (2050—2100, also publicly available also at worldclim.org). There are two implicit population processes involved in ENM-based predictions: either migration, where populations expand into previously unoccupied areas where the new environments come to fit the species' niche, or local extinction, where populations experience new environments outside the species' niche ultimately leading to geographic distribution shrinkage.

Comprehensive forecasting studies using ENMs with many species suggest that >20% of vertebrates and >40% of plants will see their distribution shrink in the 21st century (Warren et al. 2018); putting about 15% of all known species in peril of extinction (Thomas et al. 2004, Urban 2015). ENMs have also been applied — though not as comprehensively — to forecast which areas outside a species' geographic range could be under risk of invasions (Trethowan et al. 2011, Suárez-Mota et al. 2016, Qiao et al. 2017, Barbet-Massin et al. 2018). Comprehensive projections of possible invasions would be timely to device preventive strategies, given that the number of invasive species is increasing exponentially across the world and they might be facilitated by global change (Dukes and Mooney 1999, Whitney and Gabler 2008, Seebens et al. 2017).

1.2. Evolving in response to climate change

Apart from their coarse-grained and highly variable predictions (Araújo et al. 2005), ENMs have been criticized for their many simplistic assumptions (Sinclair et al. 2010). Among the most unnatural assumptions is the consideration of species as single entities that are neither evolving nor diverse. In evolutionary genetics, species are described as dynamic ensembles of populations, with various degrees of genetic diversity, population connectivity and population-specific adaptations to different environments across the species distribution

(Leimu and Fischer 2008, Hereford 2009). Geographic and climate heterogeneity, together with past climate changes of alternate glacial and interglacial periods, have generated such a legacy of diverse, locally adapted populations in many species (Hewitt 2000, 2004, Davis and Shaw 2001). This has two intuitive implications: First, that different populations of a species will have different abilities to evolve and adapt to the exact same environmental change — although obviously, the magnitude of climate change varies regionally. Second, the more environments the populations of a species have experienced in the past, the more adaptive variation the species will have overall, and the more likely it will be that at least some populations survive future environments (Jump et al. 2008).

It has become clear that many species are indeed evolving and adapting at time scales that we can directly observe (Gibbs and Grant 1987, Reznick and Ghalambor 2001, Hairston et al. 2005, Franks et al. 2007, Merilä and Hendry 2014, Bergland et al. 2014, Messer et al. 2016, Bosse et al. 2017, Nosil et al. 2018). However, building geographic predictive models that include the process of genetic adaptation involving hundreds to millions of individuals in a species has proven to be as complex as the challenge that climate scientists faced in the 1970s (Araújo and Rahbek 2006, Hoffmann and Sgrò 2011, Thuiller et al. 2013, Fordham et al. 2014, Catullo et al. 2015, Bay et al. 2017, Rudman et al. 2018). Encouragingly, we already know that genetic predictions can, in some cases, be useful and accurate. In artificial selection during plant and animal breeding — where a specific selection pressure is deliberately applied to a population where all individuals are genetically sequenced or pedigreed — quantitative genetic theory has been used very successfully to predict performance gains (Falconer and Mackay 1996). The lack of generality of some quantitative genetic models, together with the scarcity of data and more dynamic structures (Grant and Grant 2002, Nosil et al. 2018), have lead to a contradicting projections in some wild populations (Merilä et al. 2001, Walsh and Blows 2009, Hoffmann et al. 2017, Pujol et al. 2018). I propose that we first must gather and generate more comprehensive genomic and fitness datasets of wild populations. We can then use these to validate new theoretical genetic models that better describe wild populations, and finally predict biodiversity responses to climate change while accounting for the evolutionary process.

2. The genetics of evolution and adaptation

2.1. One century of theory

Quantitative and population genetics theory describes mathematically how natural selection, mutations, and demographic drift, act over genetic variation of populations, ultimately leading to adaptation. These theories were developed by R. A. Fisher, S. Wright and J. B. S. Haldane in the 1930s (Fisher 1930, Wright 1931, Haldane 1932), and were the foundation of the Modern Synthesis of Evolution. The genetic theory of evolution (arguably) started exactly 100 years ago, with a seminal paper by R.A. Fisher (1918). Fisher demonstrated that not only discrete traits such as flower color can be explained by the laws of Mendelian inheritance, but also continuous traits such as fitness (survival and offspring production). The difference was that the latter were determined by many genetic variants that individually conformed to Mendel's law (i.e. polygenic architecture). This groundbreaking work implied that potentially all the diversity of forms and species seen in nature could be explained by genetics.

A quintessential example of the application of this theory to real populations was the case of industrial adaptation by the peppered moth *Biston betularia.* In 1848 in Manchester, England, 99% of individuals were whitish, and only 1% were blackish. The frequency of the dark morph increased to about 98% in 1898, as the Birch trees around factories turned black due to coal burning and the dark morph was less conspicuous to predators on the surface of blackened trees. Haldane calculated that in order for the black morph to increase in frequency so rapidly, it need to have a natural selection advantage (selection coefficient) of 30% (Haldane 1924). That is, for every new offspring produced by a white morph moth, there were three black morph ones. Predictably, the selection coefficient should reverse with decreasing coal burning, and so indeed the white-colored morph increased again in the second half 20^{th} century (Clarke et al. 1985, Van't Hof et al. 2013). In another example for the application of population genetics to understand wild populations, Wright calculated the migration rate between *Drosophila pseudoobscura* populations of North America, as mutation and selection alone could not explain the frequency of lethal recessive alleles present in those populations (such alleles were "seen" in the lab based on their phenotypic

effects, as DNA sequencing did not exist!) (Dobzhansky and Wright 1941, Wright 1943). Although the two examples mentioned above are very simple, they showcase the predictive power that genetic models can provide to ecological studies.

2.2. Genome sequencing, standing variation and *de novo* mutations

Almost a century after the emergence of the field of population genetics, the technical revolution of genome sequencing allowed researchers to empirically study some of the classic concepts of population genetics. Apart from the very fact that we now have complete genome sequences for many species (Lewin et al. 2018), Genome-Wide Association (GWA) studies have been especially impactful in expanding our knowledge of how genetics influences organisms (Yu et al. 2006, Yang et al. 2010, Gibson 2011, Burghardt et al. 2017, Boyle et al. 2017). They have enabled us to verify that the vast majority of both continuous and discrete traits are heritable, and even to pinpoint genetic variants across the genome that control them. From height or diabetes in humans to flowering timing in plants, or burrow digging in mice (Steiner et al. 2007, Yang et al. 2010, 1001 Genomes Consortium 2016). In simple words, GWA provides a measure of whether individuals carrying a specific genetic variant of interest, are particularly prone to be on one side of the spectrum of a trait. The overall variation in the trait explained by all genetic effects is called heritability of a trait (Falconer and Mackay 1996). An exciting application of GWA in the context of adaptation to climate change is the identification of genetic variants underlying ecologically relevant traits, which can help to determine the fitness of organisms across environments (Bergelson and Roux 2010). In GWA, we study pre-existing (or standing) genetic variation of a population, which has been accrued from new mutations along its history. Ideally, if we identify the standing variants that dictate the survival and reproduction of the individuals of a species in any given environment, we could assess the evolutionary risk of its each population under climate change — the risk of failing to adapt and becoming extinct — and perhaps even design conservation policies to bring such important genetic variants to local populations to promote adaptation (Sexton et al. 2011, Aitken and Whitlock 2013).

The process of adaptation discussed above relies on a population having pre-existing variation, but one could also ask how easily populations can acquire additional mutations

that will allow them to survive. Our knowledge of new mutations derives mostly from laboratory experiments, specifically from mutation accumulation lines or long-term evolution experiments (Halligan and Keightley 2009). Such experiments start with a population of identical individuals or clones that are propagated over generations. By exposing populations to different environments, laboratory evolution experiments have shown that fast-growing organisms, such as *Escherichia coli*, *Saccharomyces cerevisiae*, *Chlamydomonas reinhardtii*, or *Drosophila melanogaster*, can adapt rapidly from new mutations, as they have a short generation times (Papadopoulos et al. 1999, Dunham et al. 2002, Burke et al. 2010, Lagator et al. 2014). In nature, this process has also been observed for viruses, which have even faster generation times. For example, HIV or influenza virus adapt rapidly to the selective pressures of the immune system within a single (human) host thanks to their high rate of mutations (Feder et al. 2016, Hadfield et al. 2017). A powerful approach to understanding the evolution of organisms over time is by reconstructing their genealogies. When the individuals sampled separately through time, as is the case of sampled viruses in patients, these genealogies can be used to calculate the mutation rate of the species, date when the different lineages of the species separated in time, and potentially identify which and when adaptive mutations arose (Drummond et al. 2003). Applying this methodology to viruses such as SIV/HIV, researchers showed that some branch lineages of the virus acquired a set of mutations before they became pandemic (Rambaut et al. 2004). Because of their slow generation time, new mutations in multicellular organisms are not thought to be an effective source for rapid adaptation (Barrett and Schluter 2008).

2.3. The genomic signatures of adaptation

The interplay between mutation rate, standing variation, the size of the population and the strength and number of variants under natural selection, are all factors that affect the dynamics of adaptation (Charlesworth and Charlesworth 2010). These dynamics, in turn, leave signatures in genetic diversity across the genome (Ellegren and Galtier 2016). Generally, the dynamics of selection over genetic variants has been classified into three types: “hard sweeps”, “soft sweeps”, and “polygenic adaptation” (also called incomplete or partial sweeps) (Pritchard et al. 2010, Booker et al. 2017). Hard sweeps occur when very strong selection favors a specific mutation in the population, increasing its frequency to

100% over a few generations (Colosimo et al. 2005). During this process, mutations that were linked to the positively selected mutation get dragged along, or swept, to high frequency, leaving a conspicuous valley of low genetic diversity of the population in the selected region in the genome. The fewer outcrossing and recombination events (both of which are also a function of time), the wider this valley (Neher 2013, Corbett-Detig et al. 2015). Soft sweeps occur in a similar fashion as hard sweeps but, when mutation rate is high, similar favorable mutations might arise in multiple genome backgrounds, so when they increase in frequency they drag along multiple background mutations, leaving a much less appreciable decrease in local diversity. In contrast to hard sweeps, adaptation in the case of soft sweeps is not limited by mutations. Because of the large population sizes and fast mutation rate, soft sweeps might be most common in bacteria (Barrick et al. 2009), viruses (Feder et al. 2016), or insects with short generation times (Karasov et al. 2010). Polygenic adaptation occurs when selection is not very strong or is acting over many genetic variants — this is particularly likely when there is plentiful standing variation in the population and fitness-related traits are complex. Individual advantageous variants rarely reach 100% frequency. Instead, they smoothly increase in frequency, a subtle footprint that can only be detected when pooling the signal across many variants and comparing multiple populations that might have experienced opposite selection (Berg and Coop 2014). One can speculate that adaptation in plants and animals with long generation times, small population sizes, more complex genomes and fitness architecture (Boyle et al. 2017) might generally follow such a polygenic adaptation or partial sweep model (Hernandez et al. 2011, Fournier-Level et al. 2011, Thurman and Barrett 2016), although direct and comprehensive evidence is mostly lacking.

3. Evolutionary genetics at the service of ecology

3.1 Evolution at the edge

The realization that evolution can take place in short timescales, previously thought to be dominated by ecological or demographic processes, has lead to the emergent field of eco-evolutionary dynamics. In this field, ecological processes such as biotic or climate changes are studied from the angle of evolutionary adaptation of populations. And vice versa, evolutionary outcomes such as lethal mutations can lead to changes in ecological

conditions such as the extinction of an important pollinator species. As the Earth is warming, populations at the more equatorial margin of a species' distribution are expected to be the first to manifest a shrinkage. The questions are, therefore: what are the eco-evolutionary processes that lead to the formation of geographic distribution limits in the past? And can we extrapolate such knowledge to understand the current distribution shifts in response to climate change? (Sexton et al. 2009). A first intuition from the ecological niche concept suggests that geographic limits are formed by physiological limits, but why then are invasive species expanding their distributions so dramatically? Or why can edge populations not adapt to the new environment if species have adapted to different climates in the past? To answer these questions, one probably needs a deeper understanding of the genetic processes that lead to adaptation and extinction in time and space.

Although simplistic in a number of ways, the explanation of distribution range shifts by the ecological niche is likely a major contributor to range limits. We can, however, rebrand this concept with evolutionary genetics thinking. In evolutionary genetics, species have genetically variable individuals, and natural selection favors the fittest genotype in each of the environments experienced across the distribution. That is, each genotype has its own optimal niche that might overlap to some degree with others. Areas that are closest to the species' niche limits would only let the genotypes adapted to extreme conditions survive. This natural selection force would tend to generate highly locally adapted populations at the edges, so most standing adaptive genetic variants would be found at the margins of the distribution (Jump and Penuelas 2005, Kawecki 2008). Because natural selection is a "filtering force", populations at the edges might become small, what could ultimately increase drift, putting a limit to adaptation (Willi et al. 2006, Bridle and Vines 2007, Kawecki 2008). In such a scenario, losing warm edge populations as a result of climate change could be a major loss of important diversity for the species (Hampe and Petit 2005).

The second explanation derives from population genetics theory and explains the distribution range limits in terms of migration and drift, without the need for environmental-driven natural selection. Let us imagine a newly formed species that starts growing and expanding in space. As populations disperse randomly in space from the origin, small founder effects, inbreeding, and drift generate a concentric pattern of gradual genetic

differentiation and decreasing diversity. This process also generates a so-called "isolation-by-distance" pattern (Wright 1943): the genetic distance between two individuals increases with geographic distance. This scenario of an expanding population already implies that one reason that geography limits exist is because individuals did not have enough time to migrate further apart yet. Therefore, all else being equal, the lower the dispersal ability of a species, the narrower its geographic limits. The second reason comes from the fact that drift, which increases towards the edges, decreases the efficiency of purifying selection to remove random deleterious mutations, such that edge populations accumulate more of these mutations (Lynch et al. 1995, Henry et al. 2015). When the number of detrimental mutations increases too much, the survival and reproduction of the edge populations decline below the replacement rate, and a new geographic limit is formed (Henry et al. 2015). These processes are a serious danger for threatened endemic species because of their very reduced geographic range (Lynch and Gabriel 1990, Lynch et al. 1995). In summary, the two reasons geographic range limits exist are: (1) limited dispersal ability, and (2) small founder effects occurring during the migration/expansion process, that increase drift towards the periphery. The associated consequences lead to local extinction due to mutational meltdown (Lynch and Gabriel 1990, Lynch et al. 1995).

The two explanations of range limits and local extinctions discussed above, the "Selectionist" vs "Neutralist" explanations, are not mutually exclusive (in fact there are others that shared principles with both (Bridle and Vines 2007)). In order to test the assumptions and emerging patterns of both hypothesis, comprehensive genomic catalogs of the populations within a species are needed (Sexton et al. 2009).

3.2 The genetic paradox of invasion

An extreme case of geographic distribution expansions is when species migrate over very long distances and even colonize other continents. As discussed above, population genetics principles tell us that because migrations occur via a limited number of founder individuals, there will be a population bottleneck that decreases genetic diversity and increases drift. Studying the differences in allelic richness and diversity of plants and animals, Dlugosch and Parker (2008) showed that such a decline in diversity is significant; although perhaps not as

dramatic as complete bottlenecks. Nevertheless, that invasions are accompanied by adaption is not rare (Lee 2002), raising the question of how populations can adapt to new environments despite their low diversity as a consequence of a founder effect. This conundrum was first noticed by the founders of the invasion genetics field, H.G. Baker & G.L. Sebbins (1965), and was later coined as the "genetic paradox of invasion" (Estoup et al. 2016), and seems to challenge both aforementioned Neutralist and Selectionist hypotheses of geographic distribution limits. The proposed solutions to this paradox can generally be separated into scenarios in which adaptation still occurs from standing variation and scenarios in which adaptation occurs from new mutations. In the first case, there is only limited depletion of standing variation during colonization, for example, if the founding population was unusually diverse, if there were recurrent migrations, or if diversity was increased through introgressions from local relatives (Dlugosch et al. 2015, Whitney et al. 2015). In the second case, the bottleneck is strong or complete, and genetic adaptation can only occur through new mutations (Colautti et al. 2017). While there is plenty of examples where adaptation occurred rapidly from standing variation in multicellular and macroscopic organisms such as animals and plants (Barrett and Schluter 2008), the adaptation from *de novo* mutations remains largely undocumented.

4. The plant of 1001 Genomes

My interest in plants derives from their important role on Earth. Living plants make up the majority of Earth's biomass, over 80% of the total ~550 Gt of carbon (Bar-On et al. 2018). Ironically, the combustion of such plant biomass stored in different forms by humans is a major driver of climate change. Plants are the primary producers of ecosystems, fixing approximately 50-60 Gt of carbon per year (Melillo et al. 1993, Woodward 2007). Therefore, any change of such an integral part of ecosystems is likely to cause a cascade of impacts on all other dependent organisms (Tylianakis et al. 2008). In fact, global patterns of primary production have already changed since the 1980s as a response to climate change (Nemani et al. 2003). The world's plant biodiversity risk assessment (Kew 2016) concluded that one fifth of the almost 400,000 vascular plant species is threatened with extinction. This is likely an underestimate given that many more species are likely to be already at risk because they

have narrow ranges and are therefore less likely to have been identified and formally described (Pimm and Raven 2017).

The major impacts of climate change on plant species are expected to be increased variability in precipitation (Schwalm et al. 2017) and droughts (Dai 2012). Plants use multiple mechanisms to cope with droughts: tolerating dehydration/hydration (as mosses do), avoiding water loss or absorbing more water (as Mediterranean shrubs do), or escaping drought periods by timing germination and reproduction to wetter seasons (as annual plants do) (Ludlow 1989). The utilization of different mechanisms varies across species and within species, as has been shown in the model plant *Arabidopsis thaliana* (Franks 2011, Juenger 2013) and its relatives (Bouzid et al. 2018). Variation in magnitude and overall drought resistance strategy across populations could enable an evolutionary response of the species to global change, as some of these strategies can pre-adaptat plants to a drier and hotter future climate (Vasseur et al. 2018).

Arabidopsis thaliana is perhaps the plant with the most comprehensive genomic catalog to date for a wild species, with over 1,135 genomes sequenced individuals from worldwide populations (1001 Genomes Consortium 2016). It was first adopted as a model organism for genetics and molecular biology during the mid-20th century (Laibach 1943) and became an established model in the late 1980s (Meyerowitz 2001). In 2000, it became the first plant with a complete reference genome (Arabidopsis Genome Initiative 2000). Soon after it came to be appreciated as a useful model for ecology (Pigliucci 2003, Tonsor et al. 2005). Its native geographic range is large, ranging from forests in Scandinavia to drylands of North Africa (Krämer 2015) — although in the South populations they seem to be sparse and their persistence limited by a minimum rainfall per year (Brennan et al. 2014). In addition, *A. thaliana* presents many of the traits of the "perfect weed", as defined by H. G. Baker (1965), being an annual herb, self-fertilizing and highly reproductive. Invasive species tend to share at least some of these traits (Razanajatovo et al. 2016), and *A. thaliana* has also colonized multiple continents in historic times (Platt et al. 2010, 1001 Genomes Consortium 2016).

The availability of thousands of genome sequences a broad geographic distribution, high colonization ability, and variable survival under climate extremes, make *A. thaliana* a

fantastic system to study the evolutionary genetics of adaptation to climate and to develop new environmental models for ecological forecasting (Hancock et al. 2011, Savolainen et al. 2013, Krämer 2015, Weigel and Nordborg 2015).

5. Objectives

Since its inception, the genetic theory of evolution has helped to explain how populations adapt to their environment. In this thesis, I have applied concepts of evolutionary genetics to the study of ecological challenges such as adaptation to climate change and invasion biology. For adaptation to occur, populations must have some standing genetic variation and/or rapidly accrue new mutations that provide a fitness advantage under the new environment.

In my first dissertation project, I asked whether *A. thaliana* harbors standing genetic variation that supports differential survival under an extreme climatic event. And if so, where are adaptive variants present across the geographic range of the species? To address this question, I used the 1001 Genomes resource of *Arabidopsis thaliana* and exposed over 200 natural lines in the greenhouse to a simulated drought. This approach led to the identification of a large number of adaptive variants that increase survival under severe drought stress. I applied the concept of environmental niches to genetic variants, to test whether adaptive variants were more commonly present at the core or at the edge of the geographic distribution (Exposito-Alonso et al. 2018d). I found that the edges — perhaps because of their more extreme environments — are hotspots of such adaptive variation.

Next, I determined the strength of selection that future environments will exert over genetic variants. I also asked what areas across the geographic distribution will suffer the strongest selection pressures in the future. To approach these questions, I designed replicated rainfall-controlled field experiments growing over 500 natural lines at two contrasting locations, one at the core and another at the edge of the geographic distribution of *A. thaliana*. This yielded substantial insights into how natural selection is distributed across the genome, and how much allele frequencies would change as a response to a climate pressure. Subsequently, I developed new environmental models to project selection intensity across the geographic range of the species. I concluded that populations in the

transition from Southern to Central Europe might be under the highest evolutionary risk, given that they do not have much standing genetic variation that might be adaptive under the climates they are expected to encounter in the future (Exposito-Alonso et al. 2017, 2018c).

Finally, because adaptation to new environments could occur via new mutations, I asked whether one can find new mutations of potential adaptive relevance in a new colonizing population. I addressed this by reconstructing the genealogy of a recent migration of *A. thaliana* to the New World with genome sequences from herbarium samples and live plants. Because I found evidence that selection had acted on new mutations, I concluded that we should not underestimate evolution from *de novo* mutations in contemporary plant invasions (Exposito-Alonso et al. 2018a).

In the discussion, I describe the impact of these studies in the areas of ecology, evolution, genetics and conservation biology, and propose future directions and technologies that will be central to advancing the genetic theory of adaptation and improving eco-evolutionary forecasting.

CHAPTER ONE

Genomic basis and evolutionary potential for extreme drought adaptation

The content of this chapter has been published as:

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H.A., Weigel, D., (2018). *Nature Ecology & Evolution*, https://doi.org/10.1038/s41559-017-0423-0.

see Thesis Appendix I

Abstract

Because the earth is currently experiencing dramatic climate change, it is of critical interest to understand how species will respond to it. The chance of a species to withstand climate change will likely depend on the diversity within the species and, particularly, whether there are subpopulations that are already adapted to extreme environments. However, most predictive studies ignore that species comprise genetically diverse individuals. We have identified genetic variants in Arabidopsis thaliana that are associated with the survival of an extreme drought event, a major consequence of global warming. Subsequently, we determined how these variants are distributed across the native range of the species. Genetic alleles conferring higher drought survival showed signatures of polygenic adaptation and were more frequently found in Mediterranean and Scandinavian regions. Using geo-environmental models, we predicted that Central European, but not Mediterranean, populations might lag behind in adaptation by the end of the 21st century. Further analyses showed that a population decline could nevertheless be compensated by natural selection acting efficiently over standing variation or by migration of adapted individuals from populations at the margins of the species' distribution. These findings highlight the importance of within-species genetic heterogeneity in facilitating an evolutionary response to a changing climate.

Contributions

Conceived and designed the project: MEA. Advised and contributed to image phenotyping:

GW and FV. Provided additional phenotype data: FV. Performed chromopainter analyses: MEA and WD. Performed drought experiments, processed image data, carried out statistical analyses: MEA. Advised and oversaw the project: DW and HAB. Wrote the first draft: MEA. Wrote the final manuscript: HAB and DW. Commented manuscript: HAB, DW, FV, GW.

CHAPTER TWO

A map of climate change-driven natural selection

The content of this chapter was published as two preprints. The second is a supplemental appendix of the first.

Exposito-Alonso, M., 500 Genomes Field Experiment Team, Burbano, H. A., Bossdorf, O., Nielsen, R., Weigel, D. (2018). *bioRxiv*, https://doi.org/10.1101/321133.

Exposito-Alonso, M., Rodríguez, R.G., Barragán, C., Capovilla, G., Chae, E., Devos, J., Dogan, E.S., Friedemann, C., Gross, C., Lang, P., Lundberg, D., Middendorf, V., Kageyama, J., Karasov, T., Kersten, S., Petersen, S., Rabbani, L., Regalado, J., Reinelt, L., Rowan, B., Seymour, D.K., Symeonidi, E., Schwab, R., Tran, D.T.N., Venkataramani, K., Van de Weyer, A.-L., Vasseur, F., Wang, G., Wedegärtner, R., Weiss, F., Wu, R., Xi, W., Zaidem, M., Zhu, W., García-Arenal, F., Burbano, H.A., Bossdorf, O., Weigel, D., (2017). *bioRxiv*, https://doi.org/10.1101/186767.

See Thesis Appendix II

Abstract

Through the lens of evolution, climate change is an agent of natural selection that forces populations to change and adapt, or face extinction. Current assessments of the risks associated with climate change, however, do not typically take into account that natural selection can dramatically impact the genetic makeup of populations. We made use of extensive genome information in *Arabidopsis thaliana* and measured how rainfall-manipulation affected the fitness of 517 natural lines grown in Spain and Germany. This allowed us to directly infer selection at the genetic level. Natural selection was particularly strong in the hot-dry Spanish location, killing 63% of lines and significantly changing the frequency of ~5% of all genome-wide variants. A significant proportion of this selection over variants could be predicted from the climate (mis)match between experimental sites and the geographic areas where variants are found (R2=29-52%). Field-validated predictions across the species range indicated that Mediterranean and Western Siberia populations — at the edges of the species' environmental limits — currently experience the strongest climate-driven selection, and Central Europeans the weakest. With rapidly increasing droughts and rising temperatures in Europe, we forecast a wave of

directional selection moving North, putting many native *A. thaliana* populations at evolutionary risk.

Contributions

Conceived the project outline: MEA, HAB, and DW. Designed, implemented and coordinated the project: MEA. Seed Bulking: MEA. Seed aliquoting: MEA, RGR, RW. Field setup: MEA, RGR, RW. Pictures of plants: MEA, RGR, RW, FW, PL, ES. Sowing Spain: MEA, RGR, HAB. Sowing Germany: MEA, FV, RW, DL, DS, BR, PL, JK, RW, WX, KV, SK. Thinning seedlings Spain: RGR.Thinning seedlings Germany: MEA, PL, GC, ES, VM, AVdW, JD, DTNT. Flower monitoring Spain: RGR. Flower monitoring Germany: MEA, LR, VM, RW, CG. Fruit images Spain: RGR. Fruit images Germany: MEA, LR, VM, RW, CG. Image processing: MEA. Data curation & analysis: MEA. Figures: MEA. Statistical analyses: MEA. Statistical advice: RN. Supervision and discussion of analysis interpretation: RN. HAB, OB, RN, and DW. Writing first draft: MEA. Writing final manuscript: MEA, HAB, OB, RN, and DW. Commenting: all authors.

CHAPTER THREE

The rate and effect of *de novo* mutation in a colonizing lineage

The content of this chapter has been published as:

Exposito-Alonso, M., Becker, C., Schuenemann, V.J., Reitter, E., Setzer, C., Slovak, R., Brachi, B., Hagmann, J., Grimm, D.G., Jiahui, C., Busch, W., Bergelson, J., Ness, R.W., Krause, J., Burbano, H.A., Weigel, D., (2018). *PLOS Genetics*, https://doi.org/10.1371/journal.pgen.1007155.

see Thesis Appendix III

Abstract

By following the evolution of populations that are initially genetically homogeneous, much can be learned about core biological principles. For example, it allows for detailed studies of the rate of emergence of de novo mutations and their change in frequency due to drift and selection. Unfortunately, in multicellular organisms with generation times of months or years, it is difficult to set up and carry out such experiments over many generations. An alternative is provided by "natural evolution experiments" that started from colonizations or invasions of new habitats by selfing lineages. With limited or missing gene flow from other lineages, new mutations and their effects can be easily detected. North America has been colonized in historic times by the plant *Arabidopsis thaliana*, and although multiple intercrossing lineages are found today, many of the individuals belong to a single lineage, HPG1. To determine in this lineage the rate of substitutions – the subset of mutations that survived natural selection and drift –, we have sequenced genomes from plants collected between 1863 and 2006. We identified 73 modern and 27 herbarium specimens that belonged to HPG1. Using the estimated substitution rate, we infer that the last common HPG1 ancestor lived in the early 17th century, when it was most likely introduced by chance from Europe. Mutations in coding regions are depleted in frequency compared to those in other portions of the genome, consistent with purifying selection. Nevertheless, a handful of mutations is found at high frequency in present-day populations. We link these to detectable phenotypic variance in traits of known ecological importance, life history, and growth, which

could reflect their adaptive value. Our work showcases how, by applying genomics methods to a combination of modern and historic samples from colonizing lineages, we can directly study new mutations and their potential evolutionary relevance.

Contributions

Conceptualization: MEA, CB, JB, JK, HAB, DW. Data curation: MEA, CB.Formal analysis: MEA, CB, JH, DGG, RWN, HAB. Funding acquisition: WB, JB, JK, HAB, DW. Investigation: MEA, CB, VJS, ER, CS, RS, BB, JH, DGG, JC, RWN, HAB. Methodology: MEA, CB, VJS, ER, CS, RS, BB, JH, DGG, JC, RWN, HAB. Supervision: WB, JB, RWN, JK, HAB, DW. Validation: MEA, CB. Visualization: MEA, CB. Writing original draft: MEA. Review & editing: MEA, CB, JB, RWN, JK, HAB, DW.

DISCUSSION

This dissertation was motivated by my curiosity of how this world, full of diverse life forms, came to be through evolution, and whether living beings will have the resilience and adaptability to overcome some of Humanity's most pernicious actions. Here I leveraged the extensive information on the geo-climatic distribution of the model plant *A. thaliana*, its vast genomic resources, and the ability to experimentally quantify fitness of a large number of individuals, to gain a number of ecological and evolutionary insights. These include knowledge of the evolutionary risk caused by climate change and on the adaptive potential of different populations from both standing genetic variation and new mutations. Below, I will discuss the broader implications of my doctoral research and future directions.

1. Adaptation at the edges, risk at the center

The extent and consequences of climate change will rarely be identical across the distribution of a species, rather, they will vary regionally or locally (Giorgi and Lionello 2008, Dai 2012). Additionally, many populations of a species have probably experienced different climates and migrations in the past, which endows them with a different legacy of genetic variants and thus a different adaptive toolset for facing climate change. Evolutionary ecologists expect the differences in the adaptability of populations to be most dramatic between core and edge populations, although current theories focus either on drift or natural selection as the leading drivers of such differences (Kawecki 2008, Henry et al. 2015). Understanding all the above will have important consequences for how one performs risk assessments under climate change and what conservation strategies one will recommend. Below I describe the knowledge gleaned from studying geographic patterns of genetic diversity in *A. thaliana* and climate-driven natural selection in multiple field stations across the species' distribution.

1.1. The genomic legacy of past climate changes

To test the emergent genomic patterns predicted by the Selectionist and Neutralist hypotheses of distribution rage limits, we first need to define the geographic distribution of *A. thaliana* and its center and marginal areas (Krämer 2015). We can do this qualitatively using the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/species/3052436), which has over one billion digitalized sightings covering virtually all ecosystems of the world, even in remote regions of the Sahara desert. *Arabidopsis thaliana* seems to be most common in Central Europe, where the distribution is rather continuous with sightings not further apart from each other than a couple of hundred kilometers. Subsetting the almost 100,000 GBIF geo-referred records to one per 0.1 latitude/longitude degree to avoid sampling effort bias, I calculated that the median geographic point of *A. thaliana*'s distribution is in Central Europe. Specifically, between Cologne and Frankfurt, Germany (50°16'N, 8°00'E), what I then defined as the current center of the geographic distribution. Human-influenced landscapes with ample, disturbed spaces (which are considered as conducive to *A. thaliana* growth), and moderate climates with abundant rainfall might explain the high density of the species in Central Europe (1001 Genomes Consortium 2016, Lee et al. 2017). The northernmost sightings are in northern Scandinavia, North of the Arctic circle (>66°N). The southernmost sightings within the native distribution are the Cape Verde Islands (adjacent to the coast of Senegal), and mountain tops of Ethiopia and Kenya. However, the geographic distribution of sightings is disjoint from the Mediterranean coasts of North Africa and southwards. We can then label North Africa and the Mediterranean as the warm edge of the distribution, where the populations are small, sparse, and isolated (Brennan et al. 2014, Durvasula et al. 2017, Exposito-Alonso et al. 2018b). In the West, the geographic distribution is truncated by the Atlantic coast of European. Although *A. thaliana* is nowadays also found in North America, this area does not belong to the native geographic range but results from a historical introduction (Platt et al. 2010, Exposito-Alonso et al. 2018a). Towards the East, sighting records begin to be intermittent from Ukraine onwards, although they are locally present near Moscow, the Caucasus, the northern plains of India, and the Yangtze River in China (Yin et al. 2010, Zou et al. 2017).

By studying the 1001 Genomes catalog of *A. thaliana* (1001 Genomes Consortium 2016), I tested genomic patterns predicted by the Neutralist hypothesis. In summary,

calculating genome-wide distances between all 1,135 individuals, my colleagues and I found that there is a significant isolation by distance pattern (1001 Genomes Consortium 2016, Lee et al. 2017). We also found that individuals at the edge of continuous range in Europe (Sweden, West Siberia, Spain and the Mediterranean) were the most divergent to all others, while individuals at the center (Central Europe, United Kingdom, and East Europe) were the least differentiated, both among each other and compared to all others (1001 Genomes Consortium 2016, Exposito-Alonso et al. 2018d). These two patterns support the notion that migration and gene flow are geographically limited throughout the distribution and that population drift increases towards the edges. However, local genetic diversity was the highest in the Mediterranean (1001 Genomes Consortium 2016, Exposito-Alonso et al. 2018d), contradicting the Neutralist hypothesis, which would predict Central Europe to be the most diverse area.

Understanding the climate history of the Quaternary glacial periods in Europe, we can explain the aforementioned, allegedly puzzling, diversity pattern. During the last glacial maximum, many areas of Europe and North America were covered by ice. During the harshest glacial extremes, some population managed to survive in Mediterranean refugia. These are “relict populations”, and similar populations been identified in other species (Hampe and Jump 2011). During interglacial periods, recolonization of the North probably occurred from one or more of these refugia (Hewitt 1999). In *A. thaliana*, based on genetic sharing, we have evidence that more than one major colonization occurred (Lee et al. 2017, Fulgione and Hancock 2018) and that many relict populations survive around the Mediterranean, where they typically live in habitats largely unaffected by human interference (1001 Genomes Consortium 2016). Because relict populations are old and had the chance to accumulate mutations, genetic diversity in the Mediterranean, in general, is higher than elsewhere. In contrast, because European populations originated from migrations that carried only a fraction of the species diversity (i.e. suffered a population bottleneck), they have a lower genetic diversity than their Mediterranean counterparts. It appears that multiple recolonizations, from Iberia, Italy, or the Balkans, are common in many species, and that Central Europe became a contact zone in which different lineages admixed (Petit et al. 2003, Eckert et al. 2008). Such a mixing and homogenization of genotypes could

also explain that Central European population of *A. thaliana* are not particularly divergent from any other group, but rather a blend of different groups (Exposito-Alonso et al. 2018d).

Since *A. thaliana* is a widespread species that re-colonized Europe in post-glacial times and migrated to America and Australia in historic times, it does not seem that its geographic limits are dictated by its dispersal ability. The Neutralist hypothesis also puts forward drift and the consequential accumulation of deleterious mutations as an important factor shaping the geographic limits of species. I therefore studied the geographic distribution of nonsynonymous mutations, i.e. mutations that cause amino acid changes in the encoded proteins and the majority of which is deleterious (Eyre-Walker and Keightley 2007). I found that populations from the warm edge carried more of these mutations (Exposito-Alonso et al. 2018c). One could mistakenly interpret this result as low efficiency of purifying selection, resulting in accumulation of deleterious mutations in the marginal and isolated relict populations (1001 Genomes Consortium 2016). However, the total number of mutations might also be directly related to the old age of populations. In order to appropriately compare populations with different levels of diversity, I investigated the ratio of nonsynonymous to synonymous mutations (K_n/K_s) The ratio correlated with latitude, i.e. the lower the latitude, the lower the proportion of nonsynonymous mutations (Exposito-Alonso et al. 2018c). This indicated that the warm edge populations actually have experienced highly efficient selection, and conversely, that Central and North European populations experienced less efficient selection. As the efficiency of natural selection does not seem to coincide with the limits of the geographic distribution in North Africa, West Siberia, and northern Sweden, we therefore must reject the Neutralist hypothesis.

Although the arguments for extinction because of genetic drift are sound (Lynch et al. 1995, Frankham 2005), and such scenarios might be very important in mammals (Abascal et al. 2016), they do not seem to explain the geographic limits of *A. thaliana* (Exposito-Alonso et al. 2018a, 2018c). The concepts of genetic drift and bottlenecks are still useful to interpret diversity patterns in the context of relict and non-relict populations of *A. thaliana* and have important applications in conservation biology beyond this species. Analogous to the Out-of-Africa theory in humans (Excoffier et al. 2008), Central and North European populations in *A. thaliana* suffered a bottleneck during the post-glacial recolonization, which

might have limited the efficiency of selection and led to a high proportion of deleterious mutations in their genome. On the other hand, relict populations from the warm edge tend to harbor the most genetic diversity in *A. thaliana* and deleterious mutations seem to have been purged more efficiently from their genomes. Based on this, it would be recommendable that conservation policies should focus on edge populations with high genetic value, particularly those at glacial refugia (Hampe and Petit 2005, Jump et al. 2008), rather than those populations from areas of high abundance of the species (Araújo and Williams 2000, Galetti et al. 2009, Tédonzong Dongmo et al. 2018).

1.2. The landscape of standing variation and natural selection

If none of the Neutralist ideas hold, does natural selection limit and shape the geographic distribution of *A. thaliana*? To directly quantify natural selection in the wild, common garden experiments constitute one of the most powerful and widely accepted approaches. This gold standard was set by the pioneering work of J. Clausen, D. D. Keck and W. M. Hiesey (1941), who carried out field experiments across a climate gradient from the Californian coast to the Sierra Nevada. With a similar spirit, I designed two rainfall-manipulated field experiments with *A. thaliana* in Spain and German field stations. I concluded that natural selection was the strongest in the field station at the warm edge, while it was very weak at the core of the distribution (Exposito-Alonso et al. 2017, 2018c). Developing a new type of field-validated environmental models (GWES), I extrapolated my insights into the strength of natural selection at the genetic level, to all known European populations of *A. thaliana* using present climate databases (Hijmans et al. 2005). Expectedly, selection intensity was highest in hot and dry regions. Taking the median climate of the geographic *A. thaliana* distribution as the species' climatic niche center, I calculated Euclidean distances in climate space (defined by 98 climate variables) from this center to all the studied populations (Exposito-Alonso et al. 2018c). Correlating the strength of selection and the environmental distance to the species' niche center, I confirmed that climate-driven selection increased towards the niche periphery – which geographically corresponds to the Mediterranean, Western Siberia and Scandinavia (Spearman's r=0.42, $P<10^{-16}$). In addition, the local genomes at those areas had signatures of highly efficient selection (the aforementioned low of K_n/K_s ratio) (Exposito-Alonso et al. 2018c). This is perhaps not so surprising as the efficiency of selection

depends both on population drift but also on the strength of selection. This suggests that selection driven by past climates could have had a role in removing deleterious *A. thaliana*'s genomic diversity.

Altogether, I found multiple lines of support for natural selection limiting the geographic distribution of *A. thaliana*, in favor of the evolution-rebranded ecological niche concept (Exposito-Alonso et al. 2018c). Although we do not have comprehensive genetic data for other species to study genetic selection coefficients across the distribution, the Selectionist hypothesis seems to be supported in 77% of the species studied in meta-analyses of common garden experiments (Lee-Yaw et al. 2016). These results come from experiments of multiple species grown at the core and at the edge of their distribution. The experiments repeatedly found that individuals' survival or fecundity significantly dropped in common gardens outside the species' geographic distribution limits (Lee-Yaw et al. 2016).

1.3. Mind the dry edge

My ultimate aim of studying eco-evolutionary processes across the distribution of a species was to predict risks that climate change will impose on the survival of species and use the resulting insights to devise potential conservation policies. Given that I identified climate variables as the natural selective pressures limiting geographic distributions, the obvious conclusion is that if natural selection shifts in the future, it could put populations at risk of local extinction. Specifically, droughts and high temperatures typical of South Europe (Seager et al. 2007, Giorgi and Lionello 2008) are expected to move northwards (Intergovernmental Panel on Climate Change 2014). Using the GWES environmental models fitted with present data, with 2050 climate maps (Intergovernmental Panel on Climate Change 2014), I predicted an increase in the strength of natural selection moving towards Central Europe (Exposito-Alonso et al. 2018c).

Whether populations will be able to adapt to this new wave of natural selection will depend on the local standing genetic variation, i.e. the presence of potentially adaptive alleles. Growing diverse genotypes under simulated drought conditions in the greenhouse

and conducting GWA, I found a clear pattern in the geographic distribution of survival-related genetic variants (Exposito-Alonso et al. 2018d). The identified adaptive variants were mostly present at the latitudinal edges of the species' range, Scandinavia and the Mediterranean regions (Exposito-Alonso et al. 2018d). I believe that the fact that populations at the Northern edge are also adapted to dry conditions is the result of cross-stress tolerance between cold and dry environments (Thomashow 1999, Swindell 2006, Exposito-Alonso et al. 2018d). Together, the knowledge of local genetic variation and increasing selective forces indicate that populations with the highest evolutionary risk are living in areas in the transition between the Mediterranean and European regions. Although much emphasis has been put on how rising temperatures might threaten warm edge populations of species (Southern Europe in our case), our results rather point to drying areas of Central Europe to have the highest evolutionary risk — a risk that might be more common in plant than animal communities (Thuiller et al. 2005).

By identifying the locations where specific adaptive variants are currently present and the areas that might suffer most strongly from climate change-driven selection pressures, one can propose more effective conservation policies. One example is the use of assisted gene flow, where one aims to improve or diversify the local gene pool to aid adaptation to climate change (Sexton et al. 2011, Aitken and Whitlock 2013, Supple et al. 2018). While *A. thaliana* might not become a globally threatened species, the evolutionary scenario depicted here might be shared by many other temperate plant species (Thuiller et al. 2005), especially because many have southern relict populations (Hampe and Jump 2011) and perhaps even north edge populations also display a cross-adaptation from cold to drought stresses. In such a scenario, a conservation strategy of transplanting seeds between different warm edge populations, or from the cold and warm edges to the center, could help preserve relict as well as Central European populations.

2. Rapid evolution from *de novo* mutations

2.1. Ancient DNA to study mutational processes in real time

The most extensive recent migration and expansion of *A. thaliana* has probably occurred in North America, where an extreme founder effect manifests itself as a massive drop of

diversity compared to the populations in the native area (Platt et al. 2010). A similarity of N. American and European climates and the opportunity to hitchhike along human traveling routes probably facilitated the rapid spread of *A. thaliana* across thousands of kilometers in a matter of 400 years (Exposito-Alonso et al. 2018a). I dated the origin of the colonization by one of the most common lineages in N. America based on genealogies reconstructed with over 5,000 mutations that the individuals from this colonizing lineage accumulated over time. I also found evidence that negative purifying selection acted at this timescale, as there were fewer mutations in coding regions than expected, and those that remained were at a lower frequency than expected. On the other hand, my results also suggested that positive selection might have acted on some of the mutations that rose to high frequency and were responsible for variation in root length and gravitropism, which in turn were correlated with precipitation at the collection sites. The obvious implication of this research is that invasive species could evolve and adapt to new environments even when they have to rely only on *de novo* mutations, and even in short periods of time (Colautti et al. 2017).

A key feature of our North American *A. thaliana* study was the use of ancient DNA (aDNA) to sample mutation accumulation over time. If the series of samples goes sufficiently back in time, it allows directly calculating substitution rates from the complete genome sequences. From herbarium specimens, such samples can be obtained as far back as 500 years (Lang et al. 2018), while from archeological remains it can be in the order of thousands of years (Gutaker et al. 2017, Swarts et al. 2017, Di Donato et al. 2018). The knowledge of mutation rates in plants could also be used to understand the origins and reservoirs of noxious species as well as to assess risks that they evolve to bypass alien species control such as herbicides or biological agents (Kreiner et al. 2017). For example, in a reasonably well-sized patch of 10,000 plants of *A. thaliana*, every generation there would be over 8 thousand new mutations (i.e. the product of the per base mutation rate, the genome size, and the total number of individuals in a population). Most of these mutations would be lost if they are neutral, but if one provides a 10% in fitness advantage, theory says that the probability that all individuals of the populations will have the advantageous mutations in a few generations is approximately 20% (under a number of assumptions that oversimplify the calculation) (Patwa and Wahl 2008). The practical consequence is that the evolved population's growth rate would be 10% faster than the original one.

2.2. The impact of polygenicity on invasion genetics

Our findings of multiple new mutations associated with ecologically relevant traits point to a scenario of polygenic adaptation in North American *A. thaliana*. This result has important consequences. Imagine that the distribution of fitness effects of new mutations in an environment is exponential (Thurman and Barrett 2016, Exposito-Alonso et al. 2018c), with probability mass function: $f(x) = \lambda e^{-\lambda x}$. This distribution implicitly asserts that there are more mutations with small effects and fewer with strong effects. The flatter the exponential distribution is (or the "more polygenic" the architecture; $\lambda \downarrow$), the less biased is the abundance of mutations with very small effect compared to intermediate or strong effect ones. This is in opposition to a very steep exponential distribution (or the "more monogenic" the architecture; $\lambda \uparrow$), where the majority of mutations has virtually zero effects, and only a minuscule number of mutations has strong effects. An interesting mathematical property of polygenic-like compared to monogenic-like distributions is that the average effect is overall higher, as the mean of the exponential distribution is equal to λ^{-1}. As a consequence, if the fitness architecture in a new environment is polygenic, a random mutation would have at least some effect on average, and natural selection can act upon it. Consequentially in polygenic adaptation, populations would not need to wait so much time until advantageous mutations appear, compared to the case of a mono(/oligo)genic mode of adaptation. If the above theoretical hypothesis holds true, it could further support the solution of the paradox of invasion that says that new mutations contribute to rapid adaptation during invasions (Dlugosch et al. 2015, Colautti et al. 2017).

3. Towards a multigenic theory of adaptation

3.1. From Mendel to GWAs

One of the fundamental questions in genetics is which and how genetic factors contribute to phenotypic variation in species. Mendel focused on traits whose inheritance was simple in peas but in this dissertation all three GWA studies — whether on root development traits, survival to extreme drought, or seed production in outdoor conditions — showed that traits

were polygenic (Exposito-Alonso et al. 2018a, 2018d, 2018c). This is not coincidental. There are now thousands of GWA studies for all kind of traits in plants and animals (including humans), and in summary these indicate that the majority of traits have multiple causative genetic variants (Rasamivelona et al. 1995, Gravois and Bernhardt 2000, Hirschhorn et al. 2002, Marwede et al. 2004, Goddard and Hayes 2009, Atwell et al. 2010, Vink et al. 2014, Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014, Escott-Price et al. 2015, Loh et al. 2015, Fan and Song 2016, 1001 Genomes Consortium 2016, Field et al. 2016, Bartoli and Roux 2017, Boyle et al. 2017, Bosse et al. 2017, Martin et al. 2017, Kita et al. 2017). Of course, many conspicuous examples in the literature present discoveries where adaptation is seemingly monogenic. Many of these studies, however, only focus on the first candidates discovered in a GWA (Jones et al. 2012, Bay et al. 2018), or investigate very charismatic traits such as coloration of mammals or birds, which are particularly prone to be controlled by one or a few genes (Steiner et al. 2007, Uy et al. 2016, Bourgeois et al. 2017, Jones et al. 2018) — including the classic peppered moth (Van't Hof et al. 2013). Furthermore, although specific traits might be controlled by a few genetic variants with strong effect, the architecture of fitness might still be polygenic, as it depends on many other traits. Studies of fitness of wild plants and crops, which depend on physiological, resource allocation, or metabolic traits, indicate it is indeed a quantitative trait (Holland 2007, Ingvarsson and Street 2011, Fournier-Level et al. 2011, Anderson et al. 2014, Price et al. 2018). In our rainfall-manipulated experiments, we measured the fitness of multiple genotypes and empirically quantified contributions of genetic variants to fitness (i.e. selection coefficients). This showed strong selection affecting thousands of variants that changed in frequency in different degrees (Exposito-Alonso et al. 2018c). In such a scenario, despite selection was strong, I predict the dynamics of adaptation to follow a polygenic adaptation model instead of a selective sweep model. Indeed the per-allele selection coefficients quantified in the field correlated with smooth geographic allele frequency gradients across the sampled *A. thaliana* native populations (Berg and Coop 2014, Exposito-Alonso et al. 2018c), rather than with genetic diversity valleys in the genome characteristic of adaptation via hard selective sweeps (Nielsen et al. 2005). Our results support the notion that polygenic adaptation plays a prominent role in plant adaptation, which will have consequences in how we predict the demographic dynamics and evolutionary responses of species to climate change.

3.2. The classic monogenic vs infinitesimal population models of adaptation

The classic mathematical models of population genetics are almost exclusively developed for the demographics of (hard) selective sweeps, i.e. fitness depends only on one mutation that is positively selected. This legacy has continued to dominate in recently-developed models of genetic adaptation to a new environment called evolutionary rescue models (Bell 2017). The models begin by assuming that upon environmental change, a population declines with rate r, as the mean absolute fitness of individuals is $1-r$. If a new mutation provides a fitness advantage to an individual so that its absolute fitness is: $(1-r)(1+s)$, one can calculate the frequency increase of the new mutation in every generation, and the probability that the population would recover through adaptation: $P \approx 2Nu\ (s-r)/r$. Intuitively, this probability depends on the number of new advantageous mutations $2Nu$ and the relative advantage that the new mutation provides compared to the disadvantage of the old one, $(s-r)/r$ (Gomulkiewicz and Holt 1995, Orr and Unckless 2014, Bell 2017). The resulting demographics are characterized by a U shape, i.e. population decreases with the change in environment and later it recovers thanks to the new advantageous mutation.

A more realistic modeling of fitness and adaptation derives from Fisher's "infinitesimal model" (also loosely called polygenic model). The model assumes that if effectively an infinite number of variants determine a trait, each with infinitesimally small effects on the trait, it would be sufficient to model the relationship between the total number of genetic differences between individuals (or genetic variance σ_g) and the total phenotypic differences between individuals (phenotypic variance σ_T). The ratio of these two is the heritability of a trait: $h^2 = \sigma_g/\sigma_T$ (Falconer and Mackay 1996). The elegance of this model is that it deals with traits that are under selection and inherited in some proportion, circumventing the modeling of many specific genetic variants (Fisher 1918). The infinitesimal model was extended to adaptation to a new environment by Lynch & Lande (1993), who defined fitness as the instantaneous rate of population increase, r, which can be written as: $r = r_\theta - \frac{(z-\theta)^2+\sigma_e+\sigma_g}{2\sigma_\theta}$, where r_θ is the fitness when the mean phenotype is at the optimum for an environment, θ. The mean population phenotype is z, and the

phenotypic variance is divided into additive genetic variance, σ_g, which captures deviations from the mean of different genotypes, and environmental noise, σ_e, which generates random deviations of each individual's phenotype. The degree of adaptation of the population is defined by the phenotypic distance from the optimum, $z - \theta$, and the strength of selection is captured by the steepness of fitness decay from the optimum, σ_θ. The faster the environment changes, the larger will be the lag of adaptation. There is also a maximum environmental change to which the population can adapt, and it can be calculated from the strength of selection and the heritability of the trait. To my knowledge, Lynch & Lande's (1993) was the first theoretical model that attempted to predict species responses to an environmental change assuming non-monogenic natural selection.

Although the infinitesimal model revolutionized quantitative genetics thanks to its ability to model continuous traits, it makes several artificial assumptions for the sake of mathematical simplicity. These that are unlikely to be met in wild populations (Morrissey et al. 2010) and they could actually prevent us from accurately predicting evolutionary responses of populations to climate change (Hoffmann et al. 2017, Pujol et al. 2018). These assumptions are: (1) Perfect random mating of populations, although it is known that gene flow is limited in many wild species; particularly those that self-fertilize like *A. thaliana* and many other plant species. (2) All genetic variants in the genome are in linkage equilibrium, although all species show some extent of linkage *dis*equilibrium in the genome (Corbett-Detig et al. 2015). Notably, this is not specific to self-fertilizing species, where recombination between different genotypes less often occurs, but it also applies to species that obligately outcross (such as humans) (Reich et al. 2001). (3) All fitness effects of genetic variants are additive and of identical magnitude, although genetic variants have dramatically different (sometimes very large) effects on fitness (Thurman and Barrett 2016), and different types of epistasis between variants probably contribute to adaptation (Carlborg et al. 2006, Le Rouzic 2014, Sohail et al. 2017, Hoffmann et al. 2017). Because most species violate these assumptions to different degrees, leading to substantially different population dynamics (Neher 2013), I believe there is a need to rethink population genetic models of adaptation that allow for polygenic architecture, yet at the same time are less constrained by the classic infinitesimal assumptions.

3.3. A multigenic rethinking of adaptation

Because of the above limitations, I propose that to accurately make predictions of population dynamics, we move towards a "multigenic theory of adaptation" (Kirkpatrick et al. 2002). I use the term multigenic to refer to those cases where a finite, yet large, number of genetic variants with different effect sizes determines the fitness of an individual (i.e. polygenic in the broad sense). This is in order to distinguish it from the Fisherian polygenic model (in the strict sense, synonymous with the infinitesimal model) that is heavily loaded with the assumptions described above. Most of the extensive GWA literature cited throughout section 3 arguably falls into this category of multigenic trait architecture. In a multigenic model of adaptation, selection is described with a vector of selection coefficients for all p genetic variants involved in fitness. The distribution of selection coefficients can follow an exponential distribution, allowing some genetic variants to have stronger effects than others (Exposito-Alonso et al. 2018c). The linkage disequilibrium among variants needs to be taken into account as well, as co-existence of multiple beneficial mutations in the same background would generate highly fit genotypes. This can be described by a $n \times p$ genome matrix X of the different genotypes in a population where alternative and reference alleles at every polymorphic site are coded as -1 and 1. The fitness of a genotype w_j would be expressed as $\prod_{i=1}^{p}(1 + s_i x_{ji})^e$, for a multiplicaticative multigenic selection model (Exposito-Alonso & Nielsen, *unpublished*). This model overcomes the assumption of additive effects on fitness, in favor of multiplicative effects (Wade et al. 2001), and also allows for a non-geometric increase in fitness with adaptive mutations (the e term). Selection coefficients and other necessary parameters can be estimated from sequencing individuals of a population and measuring their fitness (Anderson et al. 2014, Price et al. 2018, Exposito-Alonso et al. 2018c), or from genome-wide allele frequency changes by re-sequencing population that are evolving over multiple generations (Iranmehr et al. 2017). Because of the above necessary complexities, it is not possible to directly derive equations of the persistence of a population as before. However, differently from 20th-century geneticists, we can use computers to simulate populations based on the above realistic multigenic equations under a variety of conditions, in order to study future trajectories (Messer 2013, Thornton 2014). Knowing that new mutations can arise in historic times and generate trait variation, simulations can include new mutations at a certain rate drawn from

a certain probability distribution of fitness effects (Martin and Lenormand 2006, Exposito-Alonso et al. 2018c). Another direct advantage of computer simulations is that they can naturally incorporate stochastic processes such as demographic drift by sampling from statistical distributions (e.g. github.com/MoisesExpositoAlonso/popgensim). I foresee that the development of such a multigenic theory and robust and flexible platforms for population genetic simulations will be a requirement to integrate evolutionary processes to predict species responses under climate change (Fordham et al. 2014, Urban et al. 2016).

4. Conclusion: The future of eco-evolutionary forecast

My work showcases how evolutionary genetics can help to address ecological challenges that are becoming increasingly urgent in the 21st century. Using over 1,000 genomes and producing the largest *A. thaliana* fitness data resource in field experiments to date, I have been able to describe general patterns of natural selection along the genome and their interaction with the environment (Exposito-Alonso et al. 2018d, 2018c); contributing to increase the knowledge on the genomic basis of local adaptation (Hancock et al. 2011, Fournier-Level et al. 2011). My thesis also substantially contributes to solving the challenge of incorporating evolution into ecological forecasting under climate change. I moved from the typical presence/absence modeling of a species' geographic distributions (Guisan and Thuiller 2005) to model the geographic distribution of genetic mutations within a species (Fitzpatrick and Keller 2015, Exposito-Alonso et al. 2018d), and modeling the selective pressure imposed by climate over local genetic variants across the species' range (Exposito-Alonso et al. 2018c). Apart from the discussed multigenic theory of adaptation, I think there are four central technological advancements that will enable robust eco-evolutionary predictions in the future: (1) Use of stochastic computer simulations to further integrate non-genetic biological mechanisms of population dynamics. (2) Leverage the predictive power of machine learning. (3) Acquire large genetic and demographic data set for many species. (4) Use hierarchical and network approaches to extend species predictions to ecosystemic responses.

The power of stochastic computer simulations is in the details. Differently from analytical solutions, simulations can increase in complexity with essentially no further cost on computational resources. For example, they can easily incorporate migration from other populations that can catalyze local adaptation, given that we have information about dispersal of the species (Broquet and Petit 2009, Aguilée et al. 2015, Al-Asadi et al. 2018), or even human trade routes and socioeconomic parameters such as import/export of goods between countries (Seebens et al. 2015) (freely-available at www.worldbank.org). One can also define realistic species-specific demographic models that are largely ignored in analytical population genetics, for example, dormant seed banks (Salguero-Gómez et al. 2015)(Charlesworth 1973) that remain in the soil and buffer genetic changes over time, or different reproductive systems such as self-fertilizing that decreases the efficiency of selection (Neher 2013). While this eco-evolutionary simulation framework could be very powerful in ecological forecasting, it is not yet commonly used (Guillaume and Rougemont 2006, Bocedi et al. 2014, Brown et al. 2016, Rudman et al. 2018).

When large amounts of data are available, machine learning excels at data integration and prediction in comparison to classic probabilistic statistics (Bzdok et al. 2018). This can be particularly useful to define the starting parameters of population simulations. In my thesis, I used present associations between climate variables and selected features from a few study locations, such as presence/absence of genetic variants or relative fitness of genotypes, and extrapolated these properties on a spatial grid; and into the future. Machine learning has generated breakthroughs in the technology industry (Taigman et al. 2014, Silver et al. 2016), but perhaps due to the small size of datasets, it has generally being underapplied in ecology and evolution. The evolution of populations is sometimes considered stochastic and/or chaotic, ultimately unpredictable (Erwin 2006). But some recent inspiring applications of machine learning and big data tells us differently (Lässig et al. 2017, Reznick and Travis 2018). These include predicting species interactions in communities (Desjardins-Proulx et al. 2016), predicting population growth and decline from genetic diversity data (Schrider and Kern 2018), or predicting evolutionary trajectories (Neher et al. 2014, Nosil et al. 2018, Exposito-Alonso et al. 2018c).

I think that four exponentially improving data acquisition technologies will be central in gathering sufficient information on species presence and genetics in order to validate predictive models. First, crowd-sourced data through phone apps allow free, citizen-based, worldwide "watches" of species that amount to millions of sightings per year (iNaturalist.org, iSpotnature.org, eBird.org). This could allow instant validation of models for short-term predictions of extinction of populations or invasion risk. Short-term predictions might aid immediate actions in conservation biology (Dietze et al. 2018, White et al. 2018). Second, remote sensor-based technologies continue to make breakthroughs such as the digital reconstructions of the vegetation in ecosystems and even their health status (Asner et al. 2004, 2009), or the monitoring of large numbers of animals (Kranstauber et al. 2011). Third, new portable sequencing technology (such as Oxford Nanopore, nanoporetech.com) can contribute to the genetic monitoring of populations and species (Parker et al. 2017). This approach has successfully applied to track outbreaks of Ebola, Flu, or Zika almost in real time (nextstrain.org). Hopefully, this will also become widespread for animal and plant populations. Forth, ancient DNA technology that enables sequencing of past populations (Shapiro and Hofreiter 2014, Orlando et al. 2015), allows us to sample longer time series. Combining backward-in-time species predictions based on climate records with aDNA, one can validate the hindcast predictability of eco-evolutionary models to then apply them to long-term forward-in-time predictions (i.e. forecasting). Comprehensive climate datasets was a key to the success of the Intergovernmental Panel for Climate Change. Biodiversity researchers are now hoping for a boost in data production following the foundation of the analog Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services in 2012 (IPBES http://www.ipbes.net).

Ultimately, we will need to forecast not a single species, but entire ecosystems. Predicting the intricate meshes of a myriad of individuals and species interacting is also much more complex than predicting the overall ecosystem services and functions such as primary production (Nemani et al. 2003, Campbell et al. 2017, Bar-On et al. 2018). A hierarchical approach based on our knowledge of ecological networks could make a very positive contribution. No matter how much databases grow, is unlikely that we can monitor every population of every species. By building networks of all species in an ecosystem, predictions could focus on keystone or indicator species, which will drive the majority of the

network connection (Lewinsohn et al. 2006, Bascompte et al. 2006, Tylianakis et al. 2008). Once predictions are developed for heavily connected species or indicator species, their predicted presence can serve as predictors for other species. Examples of these species could be major ecosystem hub species such as the most predominant tree species in a forest (Iverson and Prasad 1998, Gedney and Valdes 2000, Laliberté and Tylianakis 2010), but also species flagged as endangered in the IUCN red list, as those are undergoing the earliest impacts and might be good sensors (Dufrene and Legendre 1997).

Because the rate of climate change, species extinction, and invasive species spread are not slowing down (Seebens et al. 2017, Brown and Caldeira 2017, Warren et al. 2018), we will have to become much better at forecasting adaptation and extinction of species to face present and future global environmental and ecological challenges. And we will. For the good of both humankind and our planet Earth.

Glossary

aDNA technology — A set of molecular techniques to retrieve, process, and analyze DNA sequences from museum specimens, archaeological findings, fossil remains, and other unusual sources of DNA such as sediments.

Admixture — The process of outcrossing of different genotypes from two or more distinct populations.

Architecture, Trait — Also genetic architecture. It refers to the number of genes or alleles involved in determining the trait, the distribution of effects or contributions to the trait, and the relationships of additivity, dominance and/or epistasis among the involved variants.

Polygenic, Trait — A trait, typically continuous, that is determined by small effects of many genetic variants throughout the genome. Also called complex or quantitative trait. A Fisherian polygenic trait assumes an infinite number of genetic variants underlying the trait. Contrast: monogenic (Mendelian) or oligogenic trait, that is determined by a one or a few genetic variants, respectively.

Bottleneck — The sudden reduction of a population size, which has an associated increase in genetic drift. When related to a migration, it is called founder effect.

Diversity, Genetic — The number of genetic variants in a population. In its strict sense, the average probability that a site differs in two randomly sampled genomes.

Diversity, Neutral — The number of genetic variants in a population that are not under natural selection. Synonym mutations, which do not generate protein changes, are typically considered approximately neutral.

Eco-evolutionary dynamics — A term that generally refers to the interplay between ecology and evolution. For example, a change in environment or ecological context can cause a change in the population mean of a trait and consequently the frequency of the underlying causal alleles. On the other hand, the rise of new mutations, can change ecological relationships, e.g. when a population of commensal insects in a plant becomes parasitic.

ENM — Environmental Niche Model. An environmental niche is an n-dimensional space that a species occupies. It is typically related to abiotic environments such as annual precipitation, but potentially could be extended to biotic environments, e.g. the niche would be all those spaces where another species already exists. The niches are typically modeled using decision trees.

Evapotranspiration — The process by which water is transferred from Earth's land to the atmosphere, both through transpiration of plants and direct evaporation. Potential evapotranspiration is high when the temperature is high; actual evapotranspiration has an upper limit imposed by water availability. When the potential is higher than the actual evapotranspiration, the soil dries out.

Evolution, Darwinian — New species originate by descent with variability and by natural selection, which is the process whereby some of the offspring are better adapted to their environment and tend to survive and produce, in turn, more offspring (Darwin 1859). Also referred to Darwinism.

Evolution, Modern Synthesis of — Integration of Darwinism and Mendel's laws of inheritance lead by E. Mayr, G. L. Stebbins and T. Dobzhansky in the second half of 20^{th} century. It is based on the mathematical breakthroughs of quantitative and population genetics in the early 20^{th} century by R. A. Fisher, S. Wright, and J. B. S. Haldane.

Fixation — Said of a mutation when it reaches 100% of frequency in the population.

Genealogy A representation of ancestral connections between two or more genetic sequences. Typically called phylogeny when representing species connections. Also called tree in the broad sense.

Genetic drift A process whereby a population experiences changes in allele frequencies just due to random sampling error during reproduction, which increases with smaller numbers. For example, if a population consists only of two hermaphrodite individuals that produce only two offspring, there is a 0.56 chance that a mutation present as one copy in one of the parents is lost. If they produce 100 offspring, there is only a 10^{-13} chance that none of the progeny will inherit that mutation.

GWA Genome-Wide Association. Most commonly, a linear (fixed or mixed) model is used to estimate the effect that a variant has on a trait of a population of genotyped individuals.

Heritability The proportion or percentage of differences in a trait between the individuals of a population that can be explained by their genetic differences.

Isolation by distance An emerging geographic and genetic pattern of populations that migrate and progressively become differentiated from each other in their genetic makeup. The more geographically separated, the more genetically distinct are two randomly picked individuals.

K_n/K_S ratio Also known as K_a/K_S or d_N/d_S. Metric of a genome or a region within a genome such as a gene. The ratio between the nonsynonymous to synonymous substitutions, i.e. those that generate a protein change and that might be subject to natural selection, and those that do not and are considered quasi-neutral. When K_n/K_S equals 1, sequences evolve only by the influence of genetic drift. When >1 only occurs when positive selection is very strong and adaptation is not mutation limited. Most studied species have K_n/K_S <1; on average nonsynonymous mutations are under purifying selection.

LD Linkage Disequilibrium. LD is the nonrandom association of alleles at two or more loci. In other words, two mutations are in LD when they are both more likely to be found together in some individuals, and vice versa, both missing in other individuals. It can be measured as the correlation between the presence of two mutations in the genome in a population. Contrast: Linkage Equilibrium.

N_e Effective population size. In population genetics, most models assume an idealized population where individuals randomly mate, population size is constant, and generations are non-overlapping. Sometimes also called a Wright-Fisher or Hardy-Weinberg population. N_e represents the size of the population in such an idealized population whose drift- and selection-driven allele frequency dynamics are equivalent as the real population, which violates some of the above assumptions and has often a very different census size. For example, the world census size of *Arabidopsis thaliana* would be much larger than the effective population size, because the species has suffered population bottlenecks and expansions, because it only 2% of the matings are between different individuals, and because the soil seed banks generate overlapping populations.

Population, Relict Populations of a species that survived last glacial era (115,000 – 11,700 years ago).

Population, Edge Also called marginal populations. Populations of a species at the periphery of its geographic distribution. Contrast: populations at the center or core populations.

Rate, Mutation	The rate of replication error of the DNA of an organism found in the gametes that give rise to the offspring.
Rate, Substitution	The observed rate of mutation after drift and selection forces have acted upon mutations. In its strict sense, is the rate of mutations that become fixed in a species.
Selection, Natural	The process whereby the individuals best adapted to an environment survive and reproduce the most, so that their genetic variants are passed on in a larger proportion than those non-adapted individuals. Those genetic variants that are favored are said to be under positive selection, those that are disfavored are under negative, also called purifying, selection. The latter is thought to be acting continuously in many of the spontaneous new mutations are detrimental.
Selection, Coefficient	The quantification of natural selection over a genetic variant. Expressed as a fraction or percentage of relative fitness advantage or disadvantage with respect to a reference. The reference has unit one and is taken as either the fittest individual or the average individual in a population.
SNP	Single Nucleotide Polymorphism. Genetic polymorphism or variant at a specific position in the genome. Many times biallelic (particularly for GWA), in which case one allele is considered the reference and the other the alternative.

References

1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. Cell 166:481–491.

Abascal, F., A. Corvelo, F. Cruz, J. L. Villanueva-Cañas, A. Vlasova, M. Marcet-Houben, B. Martínez-Cruz, J. Y. Cheng, P. Prieto, V. Quesada, J. Quilez, G. Li, F. García, M. Rubio-Camarillo, L. Frias, P. Ribeca, S. Capella-Gutiérrez, J. M. Rodríguez, F. Câmara, E. Lowy, L. Cozzuto, I. Erb, M. L. Tress, J. L. Rodriguez-Ales, J. Ruiz-Orera, F. Reverter, M. Casas-Marce, L. Soriano, J. R. Arango, S. Derdak, B. Galán, J. Blanc, M. Gut, B. Lorente-Galdos, M. Andrés-Nieto, C. López-Otín, A. Valencia, I. Gut, J. L. García, R. Guigó, W. J. Murphy, A. Ruiz-Herrera, T. Marques-Bonet, G. Roma, C. Notredame, T. Mailund, M. M. Albà, T. Gabaldón, T. Alioto, and J. A. Godoy. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biology 17:251.

Aguilée, R., P. de Villemereuil, and J.-M. Guillon. 2015. Dispersal evolution and resource matching in a spatially and temporally variable environment. Journal of Theoretical Biology 370:184–196.

Aitken, S. N., and M. C. Whitlock. 2013. Assisted Gene Flow to Facilitate Local Adaptation to Climate Change. Annual Review of Ecology, Evolution, and Systematics 44:367–388.

Al-Asadi, H., D. Petkova, M. Stephens, and J. Novembre. 2018, July 9. Estimating recent migration and population size surfaces.

Anderson, J. T., C.-R. Lee, and T. Mitchell-Olds. 2014. Strong selection genome-wide enhances fitness trade-offs across environments and episodes of selection. Evolution 68:16–31.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Araújo, M. B., R. G. Pearson, W. Thuiller, and M. Erhard. 2005. Validation of species–climate impact models under climate change. Global Change Biology 11:1504–1513.

Araújo, M. B., and C. Rahbek. 2006. How does climate change affect biodiversity? Science 313:1396–1397.

Araújo, M. B., and P. H. Williams. 2000. Selecting areas for species persistence using occurrence data. Biological Conservation 96:331–345.

Asner, G. P., D. E. Knapp, A. Balaji, and G. Paez-Acosta. 2009. Automated mapping of tropical deforestation and forest degradation: CLASlite. Journal of Applied Remote Sensing 3:033543.

Asner, G. P., D. Nepstad, G. Cardinot, and D. Ray. 2004. Drought stress and carbon uptake in an Amazon forest measured with spaceborne imaging spectroscopy. Proceedings of the National Academy of Sciences of the United States of America 101:6039–6044.

Atwell, S., Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, R. Jiang, N. W. Muliyati, X. Zhang, M. A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. R. Ecker, N. Faure, J. M. Kniskern, J. D. G. Jones, T. Michael, A. Nemri, F. Roux, D. E. Salt, C. Tang, M. Todesco, M. B. Traw, D. Weigel, P. Marjoram, J. O. Borevitz, J. Bergelson, and M. Nordborg. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631.

Baker, H. G. 1965. Characteristics and modes of origin of weeds. Pages 147–172 *in* H. G. Baker and G. L. Stebbins, editors. The Genetics of Colonizing Species. Academic Press Inc., NY.

Baker, H. G., and G. L. Stebbins, editors. 1965. The Genetics of Colonizing Species: Proceedings of the First International Union of Biological Sciences Symposia on General Biology. First Edition.

Academic Press Inc.
Barbet-Massin, M., Q. Rome, C. Villemant, and F. Courchamp. 2018. Can species distribution models really predict the expansion of invasive species? PloS One 13:e0193085.
Bar-On, Y. M., R. Phillips, and R. Milo. 2018. The biomass distribution on Earth. Proceedings of the National Academy of Sciences of the United States of America:201711842.
Barrett, R. D. H., and D. Schluter. 2008. Adaptation from standing genetic variation. Trends in Ecology & Evolution 23:38–44.
Barrick, J. E., D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461:1243–1247.
Bartoli, C., and F. Roux. 2017. Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. Frontiers in Plant Science 8:763.
Bascompte, J., P. Jordano, and J. M. Olesen. 2006. Asymmetric coevolutionary networks facilitate biodiversity maintenance. Science 312:431–433.
Bay, R. A., R. J. Harrigan, V. Le Underwood, H. Lisle Gibbs, T. B. Smith, and K. Ruegg. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science 359:83–86.
Bay, R. A., N. Rose, R. Barrett, L. Bernatchez, C. K. Ghalambor, J. R. Lasky, R. B. Brem, S. R. Palumbi, and P. Ralph. 2017. Predicting Responses to Contemporary Environmental Change Using Evolutionary Response Architectures. The American Naturalist 189:463–473.
Bell, G. 2017. Evolutionary Rescue. Annual Review of Ecology, Evolution, and Systematics 48:605–627.
Bergelson, J., and F. Roux. 2010. Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nature reviews. Genetics 11:867–879.
Berg, J. J., and G. Coop. 2014. A Population Genetic Signal of Polygenic Adaptation. PLoS Genetics 10:e1004412–e1004412.
Bergland, A. O., E. L. Behrman, K. R. O'Brien, P. S. Schmidt, and D. A. Petrov. 2014. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS Genetics 10:e1004775.
Bocedi, G., S. C. F. Palmer, G. Pe'er, R. K. Heikkinen, Y. G. Matsinos, K. Watts, and J. M. J. Travis. 2014. RangeShifter: a platform for modeling spatial eco-evolutionary dynamics and species' responses to environmental changes. Methods in Ecology and Evolution / British Ecological Society 5:388–396.
Booker, T. R., B. C. Jackson, and P. D. Keightley. 2017. Detecting positive selection in the genome. BMC Biology 15:98.
Bosse, M., L. G. Spurgin, V. N. Laine, E. F. Cole, J. A. Firth, P. Gienapp, A. G. Gosler, K. McMahon, J. Poissant, I. Verhagen, M. A. M. Groenen, K. van Oers, B. C. Sheldon, M. E. Visser, and J. Slate. 2017. Recent natural selection causes adaptive evolution of an avian polygenic trait. Science 358:365–368.
Bourgeois, Y. X. C., B. Delahaie, M. Gautier, E. Lhuillier, P.-J. G. Malé, J. A. M. Bertrand, J. Cornuault, K. Wakamatsu, O. Bouchez, C. Mould, J. Bruxaux, H. Holota, B. Milá, and C. Thébaud. 2017. A novel locus on chromosome 1 underlies the evolution of a melanic plumage polymorphism in a wild songbird. Royal Society Open Science 4:160805.
Bouzid, M., F. He, G. Schmitz, R. Haeusler, A. Weber, T. Mettler, and J. de Meaux. 2018, June 8. Arabidopsis species deploy distinct strategies to cope with drought stress.
Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An Expanded View of Complex Traits: From Polygenic to

Omnigenic. Cell 169:1177–1186.
Brennan, A. C., B. Méndez-Vigo, A. Haddioui, J. M. Martínez-Zapater, F. X. Picó, and C. Alonso-Blanco. 2014. The genetic structure of *Arabidopsis thaliana* in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. BMC Plant Biology 14:17.
Bridle, J. R., and T. H. Vines. 2007. Limits to evolution at range margins: when and why does adaptation fail? Trends in Ecology & Evolution 22:140–147.
Broquet, T., and E. J. Petit. 2009. Molecular Estimation of Dispersal for Ecology and Population Genetics. Annual Review of Ecology, Evolution, and Systematics 40:193–216.
Brown, J. L., J. J. Weber, D. F. Alvarado-Serrano, M. J. Hickerson, S. J. Franks, and A. C. Carnaval. 2016. Predicting the genetic consequences of future climate change: The power of coupling spatial demography, the coalescent, and historical landscape changes. American Journal of Botany 103:153–163.
Brown, P. T., and K. Caldeira. 2017. Greater future global warming inferred from Earth's recent energy budget. Nature 552:45–50.
Burghardt, L. T., N. D. Young, and P. Tiffin. 2017. A Guide to Genome-Wide Association Mapping in Plants. Current Protocols in Plant Biology:22–38.
Burke, M. K., J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose, and A. D. Long. 2010. Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467:587–590.
Bzdok, D., N. Altman, and M. Krzywinski. 2018. Statistics versus machine learning. Nature Methods 15:233.
Campbell, J. E., J. A. Berry, U. Seibt, S. J. Smith, S. A. Montzka, T. Launois, S. Belviso, L. Bopp, and M. Laine. 2017. Large historical growth in global terrestrial gross primary production. Nature 544:84–87.
Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel, and L. Andersson. 2006. Epistasis and the release of genetic variation during long-term selection. Nature Genetics 38:418–420.
Catullo, R. A., S. Ferrier, and A. A. Hoffmann. 2015. Extending spatial modeling of climate change responses beyond the realized niche: estimating, and accommodating, physiological limits and adaptive evolution. Global Ecology and Biogeography: a Journal of Macroecology 24:1192–1202.
Charlesworth, B. 1973. Selection in Populations with Overlapping Generations. V. Natural Selection and Life Histories. The American Naturalist 107:303–311.
Charlesworth, B., and D. Charlesworth. 2010. Elements of Evolutionary Genetics. Roberts and Company Publishers.
Clarke, C. A., G. S. Mani, and G. Wynne. 1985. Evolution in reverse: clean air and the peppered moth. Biological Journal of the Linnean Society. Linnean Society of London 26:189–199.
Clausen, J., D. D. Keck, and W. M. Hiesey. 1941. Regional Differentiation in Plant Species. The American Naturalist 75:231–250.
Colautti, R. I., J. M. Alexander, K. M. Dlugosch, S. R. Keller, and S. E. Sultan. 2017. Invasions and extinctions through the looking glass of evolutionary ecology. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 372:20160031.
Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal Jr, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, and D. M. Kingsley. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science 307:1928–1933.
Corbett-Detig, R. B., D. L. Hartl, and T. B. Sackton. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biology 13:e1002112.
Crutzen, P. J. 2002. Geology of mankind. Nature 415:23.

Dai, A. 2012. Increasing drought under global warming in observations and models. Nature Climate Change 3:52–58.

Davis, M. B., and R. G. Shaw. 2001. Range shifts and adaptive responses to Quaternary climate change. Science 292:673–679.

Dawson, T. P., S. T. Jackson, J. I. House, and I. C. Prentice. 2011. Beyond predictions: biodiversity conservation in a changing climate. Science 332 (6025): 53–58

Desjardins-Proulx, P., I. Laigle, T. Poisot, and D. Gravel. 2016. Ecological Interactions and the Netflix Problem. bioRxiv. https://doi.org/10.1101/089771

Di Donato, A., E. Filippone, M. R. Ercolano, and L. Frusciante. 2018. Genome Sequencing of Ancient Plant Remains: Findings, Uses and Potential Applications for the Study and Improvement of Modern Crops. Frontiers in Plant Science 9:441.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, M. A. Kenney, C. M. Laney, L. G. Larsen, H. W. Loescher, C. K. Lunch, B. C. Pijanowski, J. T. Randerson, E. K. Read, A. T. Tredennick, R. Vargas, K. C. Weathers, and E. P. White. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proceedings of the National Academy of Sciences of the United States of America 115:1424–1432.

Dlugosch, K. M., S. R. Anderson, J. Braasch, F. A. Cang, and H. D. Gillette. 2015. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. Molecular Ecology 24:2095–2111.

Dlugosch, K. M., and I. M. Parker. 2008. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Molecular Ecology 17:431–449.

Dobzhansky, T., and S. Wright. 1941. Genetics of Natural Populations. V. Relations between Mutation Rate and Accumulation of Lethals in Populations of *Drosophila pseudoobscura*. Genetics 26:23–51.

Drummond, A., O. G. Pybus, and A. Rambaut. 2003. Inference of viral evolutionary rates from molecular sequences. Advances in Parasitology 54:331–358.

Dufrene, M., and P. Legendre. 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

Dukes, J. S., and H. A. Mooney. 1999. Does global change increase the success of biological invaders? Trends in Ecology & Evolution 14:135–139.

Dunham, M. J., H. Badrane, T. Ferea, J. Adams, P. O. Brown, F. Rosenzweig, and D. Botstein. 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. Proceedings of the National Academy of Sciences of the United States of America 99:16144–16149.

Durvasula, A., A. Fulgione, R. M. Gutaker, S. I. Alacakaptan, P. J. Flood, C. Neto, T. Tsuchimatsu, H. A. Burbano, F. X. Picó, C. Alonso-Blanco, and A. M. Hancock. 2017. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences.

Eckert, C. G., K. E. Samis, and S. C. Lougheed. 2008. Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. Molecular Ecology doi: 10.1111/j.1365-294X.2007.03659.x

Edwards, P. N. 2011. History of climate modeling: History of climate modeling. Wiley Interdisciplinary Reviews: Climate Change 2:128–139.

Ellegren, H., and N. Galtier. 2016. Determinants of genetic diversity. Nature reviews. Genetics 17:422–433.

Erwin, D. H. 2006. Evolutionary contingency. Current Biology: CB 16:R825–6.

Escott-Price, V., R. Sims, C. Bannister, D. Harold, M. Vronskaya, E. Majounie, N. Badarinarayan, GERAD/PERADES, IGAP consortia, K. Morgan, P. Passmore, C. Holmes, J. Powell, C. Brayne, M. Gill, S. Mead, A. Goate, C. Cruchaga, J.-C. Lambert, C. van Duijn, W. Maier, A. Ramirez, P. Holmans, L. Jones, J. Hardy, S. Seshadri, G. D. Schellenberg, P. Amouyel, and J. Williams. 2015. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain: a Journal of Neurology 138:3673–3684.

Estoup, A., V. Ravigné, R. Hufbauer, R. Vitalis, M. Gautier, and B. Facon. 2016. Is There a Genetic Paradox of Biological Invasion? Annual Review of Ecology, Evolution, and Systematics 47:51–72.

Excoffier, L., M. Foll, and R. J. Petit. 2008. Genetic Consequences of Range Expansions. Annual Review of Ecology, Evolution, and Systematics 40:481–501.

Exposito-Alonso, M., C. Becker, V. J. Schuenemann, E. Reiter, C. Setzer, R. Slovak, B. Brachi, J. Hagmann, D. G. Grimm, J. Chen, W. Busch, J. Bergelson, R. W. Ness, J. Krause, H. A. Burbano, and D. Weigel. 2018a. The rate and potential relevance of new mutations in a colonizing plant lineage. PLoS Genetics 14:e1007155.

Exposito-Alonso, M., A. C. Brennan, C. Alonso-Blanco, and F. X. Picó. 2018b. Spatio-temporal variation in fitness responses to contrasting environments in *Arabidopsis thaliana*. Evolution 71:550.

Exposito-Alonso, M., H. A. Burbano, O. Bossdorf, R. Nielsen, and D. Weigel. 2018c. A map of climate change-driven natural selection in *Arabidopsis thaliana*. bioRxiv. https://doi.org/10.1101/321133

Exposito-Alonso, M., R. G. Rodríguez, C. Barragán, G. Capovilla, E. Chae, J. Devos, E. S. Dogan, C. Friedemann, C. Gross, P. Lang, D. Lundberg, V. Middendorf, J. Kageyama, T. Karasov, S. Kersten, S. Petersen, L. Rabbani, J. Regalado, L. Reinelt, B. Rowan, D. K. Seymour, E. Symeonidi, R. Schwab, D. T. N. Tran, K. Venkataramani, A.-L. Van de Weyer, F. Vasseur, G. Wang, R. Wedegärtner, F. Weiss, R. Wu, W. Xi, M. Zaidem, W. Zhu, F. García-Arenal, H. A. Burbano, O. Bossdorf, and D. Weigel. 2017. A rainfall-manipulation experiment with 517 *Arabidopsis thaliana* accessions. bioRxiv. https://doi.org/10.1101/186767

Exposito-Alonso, M., F. Vasseur, W. Ding, G. Wang, H. A. Burbano, and D. Weigel. 2018d. Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*. Nature Ecology & Evolution 2:352–358.

Eyre-Walker, A., and P. D. Keightley. 2007. The distribution of fitness effects of new mutations. Nature Reviews. Genetics 8:610–618.

Fahrig, L. 2003. Effects of Habitat Fragmentation on Biodiversity. Annual Review of Ecology, Evolution, and Systematics 34:487–515.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. Longman.

Fan, Y., and Y.-Q. Song. 2016. Finding the Missing Heritability of Genome-wide Association Study Using Genotype Imputation. Matters 2 (5): e201604000013.

Feder, A. F., S.-Y. Rhee, S. P. Holmes, R. W. Shafer, D. A. Petrov, and P. S. Pennings. 2016. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. eLife 5. https://doi.org/10.7554/eLife.10670.

Field, Y., E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton, D. Golan, L. Yengo, G. Rocheleau, P. Froguel, M. I. McCarthy, and J. K. Pritchard. 2016. Detection of human adaptation during the past 2,000 years. Science 354 (6313): 760–64.

Fisher, R. A. 1918. The correlation among relatives on the supposition of Mendelian inheriatnce. Australian Journal of Agricultural Research 14:742–757.

Fisher, S. R. A. 1930. The genetical theory of natural selection. The Clarendon Press.

Fitzpatrick, M. C., and S. R. Keller. 2015. Ecological genomics meets community-level modeling of

biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecology Letters 18:1–16.
Fordham, D. A., B. W. Brook, C. Moritz, and D. Nogués-Bravo. 2014. Better forecasts of range dynamics using genetic data. Trends in Ecology & Evolution 29:436–443.
Fordham, D. A., H. Resit Akçakaya, M. B. Araújo, J. Elith, D. A. Keith, R. Pearson, T. D. Auld, C. Mellin, J. W. Morgan, T. J. Regan, M. Tozer, M. J. Watts, M. White, B. A. Wintle, C. Yates, and B. W. Brook. 2012. Plant extinction risk under climate change: are forecast range shifts alone a good indicator of species vulnerability to global warming? Global Change Biology 18:1357–1371.
Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. 2011. A map of local adaptation in *Arabidopsis thaliana*. Science 334:86–89.
Frankham, R. 2005. Genetics and extinction. Biological Conservation 126:131–140.
Franks, S. J. 2011. Plasticity and evolution in drought avoidance and escape in the annual plant *Brassica rapa*. The New phytologist 190:249–257.
Franks, S. J., S. Sim, and A. E. Weis. 2007. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. Proceedings of the National Academy of Sciences of the United States of America 104:1278–1282.
Fulgione, A., and A. M. Hancock. 2018. Archaic lineages broaden our view on the history of *Arabidopsis thaliana*. The New Phytologist. 219 (4): 1194–98.
Galetti, M., H. C. Giacomini, R. S. Bueno, C. S. S. Bernardo, R. M. Marques, R. S. Bovendorp, C. E. Steffler, P. Rubim, S. K. Gobbo, C. I. Donatti, R. A. Begotti, F. Meirelles, R. de A. Nobre, A. G. Chiarello, and C. A. Peres. 2009. Priority areas for the conservation of Atlantic forest large mammals. Biological Conservation 142:1229–1241.
Gedney, N., and P. J. Valdes. 2000. The effect of Amazonian deforestation on the northern hemisphere circulation and climate. Geophysical Research Letters 27:3053–3056.
Gibbs, H. L., and P. R. Grant. 1987. Oscillating selection on Darwin's finches. Nature 327:511–513.
Gibson, G. 2011. Rare and common variants: twenty arguments. Nature Reviews. Genetics 13:135–145.
Giorgi, F., and P. Lionello. 2008. Climate change projections for the Mediterranean region. Global and Planetary Change 63:90–104.
Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews. Genetics 10:381–391.
Gomulkiewicz, R., and R. D. Holt. 1995. When does Evolution by Natural Selection Prevent Extinction? Evolution 49:201–207.
Grant, P. R., and B. R. Grant. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. Science 296:707–711.
Gravois, K. A., and J. L. Bernhardt. 2000. Heritability and Genotype × Environment Interactions for Discolored Rice Kernels. Crop Science 40:314.
Guillaume, F., and J. Rougemont. 2006. Nemo: an evolutionary and population genetics programming framework. Bioinformatics 22:2556–2557.
Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.
Gutaker, R. M., E. Reiter, A. Furtwängler, V. J. Schuenemann, and H. A. Burbano. 2017. Extraction of ultrashort DNA molecules from herbarium specimens. BioTechniques 62:76–79.
Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. 2017. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics , May, 224048.

Hairston, N. G., Jr, S. P. Ellner, M. A. Geber, T. Yoshida, and J. A. Fox. 2005. Rapid evolution and the convergence of ecological and evolutionary time. Ecology Letters 8:1114–1127.
Haldane, J. B. 1924. A mathematical theory of natural and artificial selection. Part I. Trans. Camb. Phil. Soc 23:19–41.
Haldane, J. B. 1932. The Causes of Evolution. Princeton University Press.
Halligan, D. L., and P. D. Keightley. 2009. Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. Annual Review of Ecology, Evolution, and Systematics 40:151–172.
Hampe, A., and A. S. Jump. 2011. Climate Relicts: Past, Present, Future. Annual Review of Ecology, Evolution, and Systematics 42:313–333.
Hampe, A., and R. J. Petit. 2005. Conserving biodiversity under climate change: the rear edge matters. Ecology Letters 8:461–467.
Hancock, A. M., B. Brachi, N. Faure, M. W. Horton, L. B. Jarymowycz, F. G. Sperone, C. Toomajian, F. Roux, and J. Bergelson. 2011. Adaptation to climate across the *Arabidopsis thaliana* genome. Science 334:83–86.
Hausfather, Z. 2017, October 5. How well have climate models projected global warming? www.carbonbrief.org/analysis-how-well-have-climate-models-projected-global-warming.
Henry, R. C., K. A. Bartoń, and J. M. J. Travis. 2015. Mutation accumulation and the formation of range limits. Biology Letters 11:20140871.
Hereford, J. 2009. A quantitative survey of local adaptation and fitness trade-offs. The American Naturalist 173:579–588.
Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. Cord Melton, A. Auton, G. McVean, 1000 Genomes Project, G. Sella, and M. Przeworski. 2011. Classic Selective Sweeps Were Rare in Recent Human Evolution. Science 331:920–924.
Hewitt, G. M. 1999. Post-glacial re-colonization of European biota. Biological journal of the Linnean Society. Linnean Society of London 68:87–112.
Hewitt, G. M. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–913.
Hewitt, G. M. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 359:183–95; discussion 195.
Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978.
Hirschhorn, J. N., K. Lohmueller, E. Byrne, and K. Hirschhorn. 2002. A comprehensive review of genetic association studies. Genetics in medicine: official journal of the American College of Medical Genetics 4:45–61.
Hoffmann, A. A., and C. M. Sgrò. 2011. Climate change and evolutionary adaptation. Nature 470:479–485.
Hoffmann, A. A., C. M. Sgrò, and T. N. Kristensen. 2017. Revisiting Adaptive Potential, Population Size, and Conservation. Trends in Ecology & Evolution 32:506–517.
Holland, J. B. 2007. Genetic architecture of complex traits in plants. Current Opinion in Plant Biology 10:156–161.
Hutchinson, G. E. 1957. Concluding remarks Cold Spring Harbor Symposia on Quantitative Biology, 22: 415--427. GS SEARCH.
Ingvarsson, P. K., and N. R. Street. 2011. Association genetics of complex traits in plants. The New Phytologist 189:909–922.
Intergovernmental Panel on Climate Change. 2014. Climate Change 2013 - The Physical Science Basis:

Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Iranmehr, A., A. Akbari, C. Schlötterer, and V. Bafna. 2017. CLEAR: Composition of Likelihoods for Evolve And Resequence Experiments. Genetics.
Iverson, L. R., and A. M. Prasad. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs 68:465–485.
Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team, E. S. Lander, F. Di Palma, K. Lindblad-Toh, and D. M. Kingsley. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.
Jones, M. R., L. S. Mills, P. C. Alves, C. M. Callahan, J. M. Alves, D. J. R. Lafferty, F. M. Jiggins, J. D. Jensen, J. Melo-Ferreira, and J. M. Good. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science 360:1355–1358.
Juenger, T. E. 2013. Natural variation and genetic constraints on drought tolerance. Current Opinion in Plant Biology 16:274–281.
Jump, A. S., R. Marchant, and J. Peñuelas. 2008. Environmental change and the option value of genetic diversity. Trends in plant science 14:51–58.
Jump, A. S., and J. Penuelas. 2005. Running to stand still: adaptation and the response of plants to rapid climate change. Ecology Letters 8:1010–1020.
Karasov, T., P. W. Messer, and D. A. Petrov. 2010. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS genetics 6:e1000924.
Kawecki, T. J. 2008. Adaptation to Marginal Habitats. Annual Review of Ecology, Evolution, and Systematics 39:321–342.
Kew, R. B. G. 2016. The state of the world's plants report--2016. Kew bulletin / Royal Botanic Gardens.
Kirkpatrick, M., T. Johnson, and N. Barton. 2002. General models of multilocus evolution. Genetics 161:1727–1750.
Kita, R., S. Venkataram, Y. Zhou, and H. B. Fraser. 2017. High-resolution mapping of cis-regulatory variation in budding yeast. Proceedings of the National Academy of Sciences of the United States of America 114:E10736–E10744.
van Kleunen, M., W. Dawson, F. Essl, J. Pergl, M. Winter, E. Weber, H. Kreft, P. Weigelt, J. Kartesz, M. Nishino, L. A. Antonova, J. F. Barcelona, F. J. Cabezas, D. Cárdenas, J. Cárdenas-Toro, N. Castaño, E. Chacón, C. Chatelain, A. L. Ebel, E. Figueiredo, N. Fuentes, Q. J. Groom, L. Henderson, Inderjit, A. Kupriyanov, S. Masciadri, J. Meerman, O. Morozova, D. Moser, D. L. Nickrent, A. Patzelt, P. B. Pelser, M. P. Baptiste, M. Poopath, M. Schulze, H. Seebens, W.-S. Shu, J. Thomas, M. Velayos, J. J. Wieringa, and P. Pyšek. 2015a. Global exchange and accumulation of non-native plants. Nature 525:100–103.
van Kleunen, M., M. Röckle, and M. Stift. 2015b. Admixture between native and invasive populations may increase invasiveness of Mimulus guttatus. Proc. R. Soc. B 282:20151487.
Krämer, U. 2015. Planting molecular functions in an ecological context with *Arabidopsis thaliana*. eLife 4. https://doi.org/10.7554/eLife.06100.
Kranstauber, B., A. Cameron, R. Weinzerl, T. Fountain, S. Tilak, M. Wikelski, and R. Kays. 2011. The Movebank data model for animal tracking. Environmental Modelling & Software 26:834–835.
Kreiner, J. M., J. R. Stinchcombe, and S. I. Wright. 2017. Population Genomics of Herbicide Resistance: Adaptation via Evolutionary Rescue. Annual Review of Plant Biology. 69 (April): 611–35.

Lagator, M., N. Colegrave, and P. Neve. 2014. Selection history and epistatic interactions impact dynamics of adaptation to novel environmental stresses. Proceedings. Biological sciences / The Royal Society 281:20141679.

Laibach, F. 1943. *Arabidopsis thaliana* (L.) Heynh. als Objekt für genetische und entwicklungsphysiologische Untersuchungen. Bot. Archiv 44:439–455.

Laliberté, E., and J. M. Tylianakis. 2010. Deforestation homogenizes tropical parasitoid–host networks. Ecology 91:1740–1747.

Lang, P. L. M., F. M. Willems, J. F. Scheepens, H. A. Burbano, and O. Bossdorf. 2018. Using herbaria to study global environmental change. PeerJ Preprints.

Lässig, M., V. Mustonen, and A. M. Walczak. 2017. Predicting evolution. Nature Ecology & Evolution 1:0077.

Lee, C. E. 2002. Evolutionary genetics of invasive species. Trends in Ecology & Evolution 17:386–391.

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

Lee-Yaw, J. A., H. M. Kharouba, M. Bontrager, C. Mahony, A. M. Csergő, A. M. E. Noreen, Q. Li, R. Schuster, and A. L. Angert. 2016. A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. Ecology letters. 19 (6): 710–22.

Leimu, R., and M. Fischer. 2008. A meta-analysis of local adaptation in plants. PloS One 3:e4010.

Le Rouzic, A. 2014. Estimating directional epistasis. Frontiers in Genetics 5:198.

Lewin, H. A., G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, and G. Zhang. 2018. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences of the United States of America 115:4325–4333.

Lewinsohn, T. M., P. Inácio Prado, P. Jordano, J. Bascompte, and J. M. Olesen. 2006. Structure in plant–animal interaction assemblages. Oikos 113:174–184.

Loh, P.-R., G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, Schizophrenia Working Group of Psychiatric Genomics Consortium, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler, M. C. O'Donovan, B. M. Neale, N. Patterson, and A. L. Price. 2015. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature Genetics 47:1385–1392.

Ludlow, M. M. 1989. Strategies of response to water stress. Pages 269–281 *in* K. H. Kreeb, H. Richter, and T. M. Minckley, editors. Structural and functional responses to environmental stress. The Hague, the Netherlands: SPB Academic.

Lynch, M., J. Conery, and R. Burger. 1995. Mutation Accumulation and the Extinction of Small Populations. The American Naturalist 146:489–518.

Lynch, M., and W. Gabriel. 1990. Mutation load and the survival of small populations. Evolution 44:1725–1737.

Lynch, M., and R. Lande. 1993. Evolution and extinction in response to environmental change. Biotic Interactions and Global Change:234–250.

Martin, A. R., M. Lin, J. M. Granka, J. W. Myrick, X. Liu, A. Sockell, E. G. Atkinson, C. J. Werely, M. Möller, M. S. Sandhu, D. M. Kingsley, E. G. Hoal, X. Liu, M. J. Daly, M. W. Feldman, C. R. Gignoux, C. D. Bustamante, and B. M. Henn. 2017. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell 171:1340–1353.e14.

Martin, G., and T. Lenormand. 2006. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60:893–907.
Marwede, V., A. Schierholt, C. Möllers, and H. C. Becker. 2004. Genotype × Environment Interactions and Heritability of Tocopherol Contents in Canola. Crop Science 44:728.
Melillo, J. M., A. D. McGuire, D. W. Kicklighter, B. Moore, C. J. Vorosmarty, and A. L. Schloss. 1993. Global climate change and terrestrial net primary production. Nature 363:234.
Merilä, J., and A. P. Hendry. 2014. Climate change, adaptation, and phenotypic plasticity: The problem and the evidence. Evolutionary Applications 7:1–14.
Merilä, J., B. C. Sheldon, and L. E. B. Kruuk. 2001. Explaining stasis: Microevolutionary studies in natural populations. Pages 199–222 *in* A. P. Hendry and M. T. Kinnison, editors. Microevolution Rate, Pattern, Process. Springer Netherlands, Dordrecht.
Messer, P. W. 2013. SLiM: simulating evolution with selection and linkage. Genetics 194:1037–1039.
Messer, P. W., S. P. Ellner, and N. G. Hairston Jr. 2016. Can Population Genetics Adapt to Rapid Evolution? Trends in Genetics: TIG. 32 (7): 408–18.
Meyerowitz, E. M. 2001. Prehistory and history of Arabidopsis research. Plant physiology 125:15–19.
Mooney, H. A., and E. E. Cleland. 2001. The evolutionary impact of invasive species. Proceedings of the National Academy of Sciences of the United States of America 98:5446–5451.
Morrissey, M. B., L. E. B. Kruuk, and A. J. Wilson. 2010. The danger of applying the breeder's equation in observational studies of natural populations. Journal of evolutionary biology 23:2277–2288.
National Academy of Sciences. 1975. Understanding Climatic Change: A program for action. National Academy of Sciences.
Neher, R. A. 2013. Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation. Annual Review of Ecology, Evolution, and Systematics 44:195–215.
Neher, R. A., C. A. Russell, and B. I. Shraiman. 2014. Predicting evolution from the shape of genealogical trees. eLife 3. https://doi.org/10.7554/eLife.03568.
Nemani, R. R., C. D. Keeling, H. Hashimoto, W. M. Jolly, S. C. Piper, C. J. Tucker, R. B. Myneni, and S. W. Running. 2003. Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science 300:1560–1563.
Newbold, T., L. N. Hudson, A. P. Arnell, S. Contu, A. De Palma, S. Ferrier, S. L. L. Hill, A. J. Hoskins, I. Lysenko, H. R. P. Phillips, V. J. Burton, C. W. T. Chng, S. Emerson, D. Gao, G. Pask-Hale, J. Hutton, M. Jung, K. Sanchez-Ortiz, B. I. Simmons, S. Whitmee, H. Zhang, J. P. W. Scharlemann, and A. Purvis. 2016. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. Science 353:288–291.
Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. Genome Research 15:1566–1575.
Nosil, P., R. Villoutreix, C. F. de Carvalho, T. E. Farkas, V. Soria-Carrasco, J. L. Feder, B. J. Crespi, and Z. Gompert. 2018. Natural selection and the predictability of evolution in Timema stick insects. Science 359:765–770.
Orlando, L., M. T. P. Gilbert, and E. Willerslev. 2015. Reconstructing ancient genomes and epigenomes. Nature Reviews. Genetics 16:395–408.
Orr, H. A., and R. L. Unckless. 2014. The population genetics of evolutionary rescue. PLoS genetics 10:e1004551.
Papadopoulos, D., D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. 1999. Genomic evolution during a 10,000-generation experiment with bacteria. Proceedings of the National Academy of Sciences of the United States of America 96:3807–3812.
Parker, J., A. J. Helmstetter, D. Devey, T. Wilkinson, and A. S. T. Papadopulos. 2017. Field-based species

identification of closely-related plants using real-time nanopore sequencing. Scientific reports 7:8345.

Parmesan, C., and G. Yohe. 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature 421:37–42.

Patwa, Z., and L. M. Wahl. 2008. The fixation probability of beneficial mutations. Journal of the Royal Society, Interface / the Royal Society 5:1279–1289.

Pearson, R. G. 2016. Reasons to Conserve Nature. Trends in Ecology & Evolution 31:366–371.

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography: a Journal of Macroecology 12:361–371.

Pearson, R. G., and T. P. Dawson. 2004. Bioclimate envelope models: what they detect and what they hide - response to Hampe (2004): Correspondence. Global Ecology and Biogeography: a Journal of Macroecology 13:471–473.

Petit, R. J., I. Aguinagalde, J.-L. de Beaulieu, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, S. Fineschi, D. Grivet, M. Lascoux, A. Mohanty, G. Müller-Starck, B. Demesure-Musch, A. Palmé, J. P. Martín, S. Rendell, and G. G. Vendramin. 2003. Glacial Refugia: Hotspots But Not Melting Pots of Genetic Diversity. Science 300:1563–1565.

Pigliucci, M. 2003. Selection in a Model System: Ecological Genetics of Flowering Time in *Arabidopsis thaliana*. Ecology 84:1700–1712.

Pimm, S. L., C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, and J. O. Sexton. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science 344:1246752.

Pimm, S. L., and P. H. Raven. 2017. The Fate of the World's Plants. Trends in Ecology & Evolution 32:317–320.

Platt, A., M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio, N. W. Mulyati, J. Agren, O. Bossdorf, D. Byers, K. Donohue, M. Dunning, E. B. Holub, A. Hudson, V. Le Corre, O. Loudet, F. Roux, N. Warthmann, D. Weigel, L. Rivero, R. Scholl, M. Nordborg, J. Bergelson, and J. O. Borevitz. 2010. The scale of population structure in *Arabidopsis thaliana*. PLoS Genetics 6:e1000843.

Price, N., B. T. Moyers, L. Lopez, J. R. Lasky, J. Grey Monroe, J. L. Mullen, C. G. Oakley, J. Lin, J. Ågren, D. R. Schrider, A. D. Kern, and J. K. McKay. 2018. Combining population genomics and fitness QTLs to identify the genetics of local adaptation in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America:201719998.

Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current biology: CB 20:R208–15.

Pujol, B., S. Blanchet, A. Charmantier, E. Danchin, B. Facon, P. Marrot, F. Roux, I. Scotti, C. Teplitsky, C. E. Thomson, and I. Winney. 2018. The Missing Response to Selection in the Wild. Trends in Ecology & Evolution 33:337–346.

Qiao, H., L. E. Escobar, and A. T. Peterson. 2017. Accessible areas in ecological niche comparisons of invasive species: Recognized but still overlooked. Scientific Rreports 7:1213.

Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes. 2004. The causes and consequences of HIV evolution. Nature reviews. Genetics 5:52–61.

Rasamivelona, A., K. A. Gravois, and R. H. Dilday. 1995. Heritability and Genotype × Environment Interactions for Straighthead in Rice. Crop Science 35:1365.

Razanajatovo, M., N. Maurel, W. Dawson, F. Essl, H. Kreft, J. Pergl, P. Pyšek, P. Weigelt, M. Winter, and M. van Kleunen. 2016. Plants capable of selfing are more likely to become naturalized. Nature Communications 7:13313.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Reznick, D. N., and C. K. Ghalambor. 2001. The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. Genetica 112-113:183–198.

Reznick, D., and J. Travis. 2018. Is evolution predictable? Science 359:738–739.

Rosenzweig, C., D. Karoly, M. Vicarelli, P. Neofotis, Q. Wu, G. Casassa, A. Menzel, T. L. Root, N. Estrella, B. Seguin, P. Tryjanowski, C. Liu, S. Rawlins, and A. Imeson. 2008. Attributing physical and biological impacts to anthropogenic climate change. Nature 453:353–357.

Rudman, S. M., M. A. Barbour, K. Csilléry, P. Gienapp, F. Guillaume, N. G. Hairston Jr, A. P. Hendry, J. R. Lasky, M. Rafajlović, K. Räsänen, P. S. Schmidt, O. Seehausen, N. O. Therkildsen, M. M. Turcotte, and J. M. Levine. 2018. What genomic data can reveal about eco-evolutionary dynamics. Nature Ecology & Evolution 2:9–15.

Salguero-Gómez, R., O. R. Jones, C. R. Archer, Y. M. Buckley, J. Che-Castaldo, H. Caswell, D. Hodgson, A. Scheuerlein, D. A. Conde, E. Brinks, H. de Buhr, C. Farack, F. Gottschalk, A. Hartmann, A. Henning, G. Hoppe, G. Römer, J. Runge, T. Ruoff, J. Wille, S. Zeh, R. Davison, D. Vieregg, A. Baudisch, R. Altwegg, F. Colchero, M. Dong, H. de Kroon, J.-D. Lebreton, C. J. E. Metcalf, M. M. Neel, I. M. Parker, T. Takada, T. Valverde, L. A. Vélez-Espino, G. M. Wardle, M. Franco, and J. W. Vaupel. 2015. The compadre Plant Matrix Database: an open online repository for plant demography. The Journal of Ecology 103:202–218.

Savolainen, O., M. Lascoux, and J. Merilä. 2013. Ecological genomics of local adaptation. Nature Reviews. Genetics 14:807–820.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511:421–427.

Schrider, D. R., and A. D. Kern. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in genetics: TIG. 34 (4): 301–12.

Schwalm, C. R., W. R. L. Anderegg, A. M. Michalak, J. B. Fisher, F. Biondi, G. Koch, M. Litvak, K. Ogle, J. D. Shaw, A. Wolf, D. N. Huntzinger, K. Schaefer, R. Cook, Y. Wei, Y. Fang, D. Hayes, M. Huang, A. Jain, and H. Tian. 2017. Global patterns of drought recovery. Nature 548:202–205.

Seager, R., M. Ting, I. Held, Y. Kushnir, J. Lu, G. Vecchi, H. -P. Huang, N. Harnik, A. Leetmaa, N. -C. Lau, C. Li, J. Velez, and N. Naik. 2007. Model Projections of an Imminent Transition to a More Arid Climate in Southwestern North America. Science 316:1181–1184.

Seebens, H., T. M. Blackburn, E. E. Dyer, P. Genovesi, P. E. Hulme, J. M. Jeschke, S. Pagad, P. Pyšek, M. Winter, M. Arianoutsou, S. Bacher, B. Blasius, G. Brundu, C. Capinha, L. Celesti-Grapow, W. Dawson, S. Dullinger, N. Fuentes, H. Jäger, J. Kartesz, M. Kenis, H. Kreft, I. Kühn, B. Lenzner, A. Liebhold, A. Mosena, D. Moser, M. Nishino, D. Pearman, J. Pergl, W. Rabitsch, J. Rojas-Sandoval, A. Roques, S. Rorke, S. Rossinelli, H. E. Roy, R. Scalera, S. Schindler, K. Štajerová, B. Tokarska-Guzik, M. van Kleunen, K. Walker, P. Weigelt, T. Yamanaka, and F. Essl. 2017. No saturation in the accumulation of alien species worldwide. Nature Communications 8:14435.

Seebens, H., F. Essl, W. Dawson, N. Fuentes, D. Moser, J. Pergl, P. Pyšek, M. van Kleunen, E. Weber, M. Winter, and B. Blasius. 2015. Global trade will accelerate plant invasions in emerging economies under climate change. Global Change Biology 21:4128–4140.

Sexton, J. P., P. J. McIntyre, A. L. Angert, and K. J. Rice. 2009. Evolution and Ecology of Species Range Limits. Annual Review of Ecology, Evolution, and Systematics 40:415–436.

Sexton, J. P., S. Y. Strauss, and K. J. Rice. 2011. Gene flow increases fitness at the warm edge of a

species' range. Proceedings of the National Academy of Sciences of the United States of America 108:11704–11709.

Shapiro, B., and M. Hofreiter. 2014. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. Science 343:1236573.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489.

Sinclair, S., M. White, and G. Newell. 2010. How useful are species distribution models for managing biodiversity under future climates? Ecology and Society 15.

Sohail, M., O. A. Vakhrusheva, J. H. Sul, S. L. Pulit, L. C. Francioli, Genome of the Netherlands Consortium, Alzheimer's Disease Neuroimaging Initiative, L. H. van den Berg, J. H. Veldink, P. I. W. de Bakker, G. A. Bazykin, A. S. Kondrashov, and S. R. Sunyaev. 2017. Negative selection in humans and fruit flies involves synergistic epistasis. Science 356:539–542.

Steffen, W., K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, and S. Sörlin. 2015. Planetary boundaries: guiding human development on a changing planet. Science 347:1259855.

Steiner, C. C., J. N. Weber, and H. E. Hoekstra. 2007. Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biology 5:e219.

Suárez-Mota, M. E., E. Ortiz, J. L. Villaseñor, and F. J. Espinosa-García. 2016. Ecological Niche Modeling of Invasive Plant Species According to Invasion Status and Management Needs: The Case of *Chromolaena odorata* (Asteraceae) in South Africa. Polish Journal of Ecology 64:369–383.

Supple, M. A., J. G. Bragg, L. M. Broadhurst, A. B. Nicotra, M. Byrne, R. L. Andrew, A. Widdup, N. C. Aitken, and J. O. Borevitz. 2018. Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change. eLife 7. https://doi.org/10.7554/eLife.31835.

Swarts, K., R. M. Gutaker, B. Benz, M. Blake, R. Bukowski, J. Holland, M. Kruse-Peeples, N. Lepak, L. Prim, M. Cinta Romay, J. Ross-Ibarra, J. de Jesus Sanchez-Gonzalez, C. Schmidt, V. J. Schuenemann, J. Krause, R. G. Matson, D. Weigel, E. S. Buckler, and H. A. Burbano. 2017. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357:512–515.

Swindell, W. R. 2006. The association among gene expression responses to nine abiotic stress treatments in *Arabidopsis thaliana*. Genetics 174:1811–1824.

Taigman, Y., M. Yang, M. 'aurelio Ranzato, and L. Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Pages 1701–1708 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington, DC, USA.

Tatem, A. J. 2009. The worldwide airline network and the dispersal of exotic species: 2007-2010. Ecography 32:94–102.

Tédonzong Dongmo, L. R., J. Willie, A. M. P. Keuko, J. K. Kuenbou, G. Njotah, M. N. Tchamba, N. Tagg, and L. Lens. 2018. Using abundance and habitat variables to identify high conservation value areas for threatened mammals. Biodiversity and Conservation 27:1115–1137.

Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. F. de Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. van Jaarsveld, G. F. Midgley, L. Miles, M. A. Ortega-Huerta, A. Townsend Peterson, O. L. Phillips, and S. E. Williams.

2004. Extinction risk from climate change. Nature 427:145.
Thomashow, M. F. 1999. PLANT COLD ACCLIMATION: Freezing Tolerance Genes and Regulatory Mechanisms. Annual review of plant physiology and plant molecular biology 50:571–599.
Thornton, K. R. 2014. A C++ Template Library for Efficient Forward-Time Population Genetic Simulation of Large Populations. Genetics 198:157–166.
Thuiller, W., S. Lavorel, M. B. Araújo, M. T. Sykes, and I. C. Prentice. 2005. Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Sciences of the United States of America 102:8245–8250.
Thuiller, W., T. Münkemüller, S. Lavergne, D. Mouillot, N. Mouquet, K. Schiffers, and D. Gravel. 2013. A road map for integrating eco-evolutionary processes into biodiversity models. Ecology letters 16 Suppl 1:94–105.
Thurman, T. J., and R. D. H. Barrett. 2016. The genetic consequences of selection in natural populations. Molecular Ecology 25:1429–1448.
Tonsor, S. J., C. Alonso-Blanco, and M. Koornneef. 2005. Gene function beyond the single trait: natural variation, gene effects, and evolutionary ecology in *Arabidopsis thaliana*. Plant, cell & environment 28:2–20.
Trethowan, P. D., M. P. Robertson, and A. J. McConnachie. 2011. Ecological niche modelling of an invasive alien plant and its potential biological control agents. S. Afr. J. Bot. 77:137–146.
Tylianakis, J. M., R. K. Didham, J. Bascompte, and D. A. Wardle. 2008. Global change and species interactions in terrestrial ecosystems. Ecology Letters 11:1351–1363.
Urban, M. C. 2015. Accelerating extinction risk from climate change. Science 348:571–573.
Urban, M. C., G. Bocedi, A. P. Hendry, J.-B. Mihoub, G. Pe'er, A. Singer, J. R. Bridle, L. G. Crozier, L. De Meester, W. Godsoe, A. Gonzalez, J. J. Hellmann, R. D. Holt, A. Huth, K. Johst, C. B. Krug, P. W. Leadley, S. C. F. Palmer, J. H. Pantel, A. Schmitz, P. A. Zollner, and J. M. J. Travis. 2016. Improving the forecast for biodiversity under climate change. Science 353.
Uy, J. A. C., E. A. Cooper, S. Cutie, M. R. Concannon, J. W. Poelstra, R. G. Moyle, and C. E. Filardi. 2016. Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. Proceedings. Biological sciences / The Royal Society 283.
Van't Hof, A. E., P. Nguyen, M. Dalíková, N. Edmonds, F. Marec, and I. J. Saccheri. 2013. Linkage map of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): a model of industrial melanism. Heredity 110:283–295.
Vasseur, F., M. Exposito-Alonso, O. J. Ayala-Garay, G. Wang, B. J. Enquist, D. Vile, C. Violle, and D. Weigel. 2018. Adaptive diversification of growth allometry in the plant *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America:201709141.
Vink, J. M., J. J. Hottenga, E. J. C. de Geus, G. Willemsen, M. C. Neale, H. Furberg, and D. I. Boomsma. 2014. Polygenic risk scores for smoking: predictors for alcohol and cannabis use? Addiction 109:1141–1151.
Vitousek, P. M., H. A. Mooney, and J. Lubchenco. 1997. Human domination of Earth's ecosystems. Science. 277 (5325): 494–99.
Wade, M. J., R. G. Winther, A. F. Agrawal, and C. J. Goodnight. 2001. Alternative definitions of epistasis: dependence and interaction. Trends in Ecology & Evolution 16:498–504.
Walsh, B., and M. W. Blows. 2009. Abundant Genetic Variation + Strong Selection = Multivariate Genetic Constraints: A Geometric View of Adaptation. Annual Review of Ecology, Evolution, and Systematics 40:41–59.
Warren, R., J. Price, E. Graham, N. Forstenhaeusler, and J. VanDerWal. 2018. The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C. Science

360:791–795.
Weigel, D., and M. Nordborg. 2015. Population Genomics for Understanding Adaptation in Wild Plant Species. Annual Review of Genetics 49:315–338.
White, E. P., G. M. Yenni, S. D. Taylor, E. M. Christensen, E. K. Bledsoe, J. L. Simonis, and S. K. Morgan Ernest. 2018. Developing an automated iterative near-term forecasting system for an ecological study. bioRxiv. https://doi.org/10.1101/268623.
Whitney, K. D., K. W. Broman, N. C. Kane, S. M. Hovick, R. A. Randell, and L. H. Rieseberg. 2015. Quantitative trait locus mapping identifies candidate alleles involved in adaptive introgression and range expansion in a wild sunflower. Molecular Ecology 24:2194–2211.
Whitney, K. D., and C. A. Gabler. 2008. Rapid evolution in introduced species,"invasive traits" and recipient communities: challenges for predicting invasive potential. Diversity and Distributions 14:569–580.
Willi, Y., J. Van Buskirk, and A. A. Hoffmann. 2006. Limits to the Adaptive Potential of Small Populations. Annual Review of Ecology, Evolution, and Systematics 37:433–458.
Woodward, F. I. 2007. Global primary production. Current Biology: CB 17:R269–73.
Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97–159.
Wright, S. 1943. Isolation by Distance. Genetics 28:114–138.
Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature genetics 42:565–569.
Yin, P., J. Kang, F. He, L.-J. Qu, and H. Gu. 2010. The origin of populations of *Arabidopsis thaliana* in China, based on the chloroplast DNA sequences. BMC Plant Biology 10:22.
Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38:203–208.
Zalasiewicz, J., C. N. Waters, C. P. Summerhayes, A. P. Wolfe, A. D. Barnosky, A. Cearreta, P. Crutzen, E. Ellis, I. J. Fairchild, A. Gałuszka, P. Haff, I. Hajdas, M. J. Head, J. A. Ivar do Sul, C. Jeandel, R. Leinfelder, J. R. McNeill, C. Neal, E. Odada, N. Oreskes, W. Steffen, J. Syvitski, D. Vidas, M. Wagreich, and M. Williams. 2017. The Working Group on the Anthropocene: Summary of evidence and interim recommendations. Anthropocene 19:55–60.
Zou, Y.-P., X.-H. Hou, Q. Wu, J.-F. Chen, Z.-W. Li, T.-S. Han, X.-M. Niu, L. Yang, Y.-C. Xu, J. Zhang, F.-M. Zhang, D. Tan, Z. Tian, H. Gu, and Y.-L. Guo. 2017. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. Genome Biology 18:239.

Acknowledgments

I am very grateful to all the people that accompanied me during the process of this PhD and during my life. Below I will mention some names and sadly overlook many. If you are reading this, thank you!

I would like to thank Detlef. He is an inspiration, scientifically and personally. He supported my ideas from the beginning — perhaps betting against the odds — and provided me with a virtually complete scientific freedom. Thanks for teaching me to think big(ger), to explain complex concepts with elegance and simplicity, to put people in front of scientific product, to take the high road.

I would like to thank Hernán who has, from the start, helped me in the good and the bad days, in the scientific and in the personal. He started as a Skype supervisor and became a friend. I admire Hernán. He is an example of patience, dedication, care, and good heart, and I wish to keep applying his lessons throughout my life.

I thank Oliver for accepting me in his community, for sharing, for always helping with a smile, and for Heuberger Tor. Thanks for emanating sympathy and humbleness to everybody around you.

I am grateful to Rasmus for spending so many hours making my brain hurt (particularly on Wednesdays), for welcoming me in Berkeley since the first email and always, for helping me build a future. I never took all this for granted.

I feel very proud and lucky of having such four great mentors and friends. They made possible that my PhD thesis (and my self) could grow in so many conceptual directions.

Thanks to my teachers and undergrad supervisors, as they propelled me to keep growing scientifically within and beyond Spain: Ana, Jordi and, specially, Xavi.

Thanks to Weigel Lab. It is a wonderful, inclusive, and happy place. I particularly want to thank Rebecca, the hero that makes a seemingly utopian (weigel)world, possible. I will remember Weigel Lab as a very large family.

Thanks everybody from Weigel Lab and in different institutes of Madrid for taking the dust out of their boots, and go out to the field with me in a wonderful adventure (see the full list in the author contributions and acknowledgments of the bioRxiv paper doi.org/10.1101/186767).

Thanks to François and Niek for sailing with me towards the crazy project of GrENE-net. It has been (and will continue to be) an extreme fun.

I am grateful to Hernán, Patti, and Talia, for commenting the thesis. As always, t improved a lot!

Thanks to all my friends, both the “sciency” friends and “de toda la vida” friends.

I would like to start with the first group of people that received me in Tübingen and that sadly had to leave (way) too soon: Dan, Danelle, Claude, Nacho, Beth, and George (and Hernán, but he is already above). Thanks for all the scientific and personal discussions with beers. I miss you!

Thanks to Talia and Mike. You have helped me a lot in the last years (thanks Mike for your pedagogy expertise; thanks Talia for being the best officemate I could ever imagine). You are wonderful people and I have learned a lot from you. I am just mad at you because you came to Tübingen 2 years late!

Thanks to my cohort of PhD students and the Perla Latina, that are always up for a beer or a coffee, and that have supported me in the scientific and in the social parts of my life. Thanks Sergio, Julian, Cristina, Effie, Clemens, Patti, and many others. Special thanks to Patti for improving my grammar (even outside work), for teaching me German, and for listening to too many hours of my rantings!

Gracias a mis amigos en España. Mis amigos de Sevilla, que me ofrecieron la oportunidad de reinventarme y convivieron conmigo en una gran comunidad de un gran hotel y una pensión en triana y más allá: Álvaro, Jesús, Alberto, Miguel, Gemma, Noa, y Ester. Gracias a mis “amigos de toda la vida”. Hay tantos bebés que debería enumerar que da vergüenza, así que sentíos todos incluidos! Gracias en particular a las chicas que han estado conmigo desde los 4 años: Marina, María Z, María R, Lara y (por ende) Rafa. Chicas, sin haber crecido con vosotras, no sería la misma persona. Gracias Irene por descubrir conmigo la naturaleza de la Terreta y los alrededores.

A mi familia, mi yaya Honorata, mamá Adela, papá Antonio, y mi hermano David. Sin ellos, no estaría aquí. Han sido los pilares inamovibles de mi vida que me han sostenido y ayudado a desarrollarme. Para que mi familia lo pueda leer, traduzco aquí el sumario de la tesis a Español:

El cambio climático global está impactando la biodiversidad de la Tierra, pero todavía no somos capaces de predecir qué especies sobrevivirán y cuales se beneficiarán. Puesto que muchas especies no podrán tolerar un tal cambio climático no tendrán la habilidad de migrar

rápidamente a latitudes más altas, la supervivencia dependerá de si se pueden adaptar genéticamente, es decir, si pueden evolucionar. Otras especies parecen adaptarse rápidamente en el nuevo status quo en el cual los humanos dominamos los ecosistemas. Estamos empezando a entender las huellas que adaptaciones a cambios climáticos pasados han dejado en los genomas de las poblaciones y cómo esto las ha podido preparar para posibles futuras adaptaciones rápidas, pero todavía quedan muchas preguntas por resolver. Además, el conocimiento actual sobre evolución y adaptación raramente se incluye en modelos predictivos de biodiversidad, aunque obviamente ayudarían a mejorar las predicciones y a diseñar estrategias de conservación más efectivas. Aquí abordo estos retos usando la planta *Arabidopsis thaliana*, de la familia de la mostaza, sobre la cual tenemos información genética, geográfica y morfológica, de miles de individuos.

En el capítulo uno, estudié cómo poblaciones de la misma especie podrían responder de forma más o menos efectivamente al mismo cambio climático. Examiné la supervivencia de 220 líneas naturales (o variedades) de la planta *Arabidopsis thaliana* a condiciones de sequía extrema simuladas en un invernadero. Las sequías severas, consecuencia del cambio climático, se espera sean uno de las amenazas más grandes para las comunidades vegetales. Usando la técnica de modelos de nicho ambiental en combinación con asociaciones genómicas, pude determinar una serie de variantes genéticas adaptativas, y que precisamente se encontraban en los márgenes de la distribución geográfica de la especie. Quizá al haber vivido en ambientes más extremos, las poblaciones en los bordes de la distribución podrían ser un reservorio de variación adaptativa en un futuro clima más hostil.

En el capítulo dos, hice experimentos de campo a gran escala para cuantificar la selección natural dirigida por el clima en condiciones naturales. Plantamos un panel de 517 líneas naturales de *A. thaliana* en experimentos de jardines comunes con precipitación controlada en una región con clima moderado, en Europa Central, y una región con clima más extremo, el Mediterráneo. Usando análisis de imágenes, estimé el éxito reproductivo de las plantas y generé cerca de 25,000 medidas de fitness. Con estos datos, pude inferir cambios masivos en frecuencia alélica a lo largo del genoma en una sola generación, siendo más extremo en altas temperaturas y precipitación reducida, ya que muchos genotipos centroeuropeos murieron. Integrando teorías de adaptación local y técnicas de machine learning, demostré que una parte significativa de la selección natural es predecible. Con esto, y en combinación con el conocimiento de la composición genética de poblaciones, hice predicciones que indican que las áreas entre el Mediterráneo y Europa Central sufrirán el riesgo evolutivo más alto debido a una reducción repentina de precipitación en el futuro. Estos resultados contrastan con la visión generalmente aceptada que las poblaciones en el borde ecuatorial, más cálido, están en mayor riesgo de extinción que las poblaciones en el centro de la distribución geográfica.

En el capítulo tres, estudié el valor adaptativo de nuevas mutaciones, en lugar de adaptaciones ya existentes como en los capítulos previos. Usando muestras de herbario

como fotografías en el tiempo, estudié un linaje de *A. thaliana* que se originó hace 400 años tras aislarse durante una migración a Norte América. En este linage, identifiqué 5,000 nuevas mutaciones, algunas de las cuales producían diversidad morfológica en las plantas que podría estar relacionado con la adaptación al continente recientemente colonizado. Con esto pude concluir que incluso organismos de gran tamaño como plantas pueden evolucionar en cortos periodos de tiempo usando sólo nuevas mutaciones.

En general esta tesis doctoral ha avanzado nuestro conocimiento en cómo y si es posible que diferentes poblaciones de una especie se adapten genéticamente al cambio climático. Algunos de los descubrimientos más importantes son: (1) que la adaptación al clima sucede por la acumulaciones de cientos de variaciones genéticas de forma concertada (adaptación poligénica), (2) que nuevas mutaciones aparecen de forma relativamente rápida con lo que pueden contribuir en la adaptación en tiempo real, por ejemplo en plantas invasoras, y (3) que modelos que aprenden de la asociación entre climas actuales y variantes genéticas pueden ser usados para predecir si poblaciones estarán en riesgo evolutivo por el cambio climático en un futuro. En conjunto, todos estos estudios nos sitúan en una nueva etapa para entender y solucionar retos ecológicos usando la teoría de la evolución genética.

Thesis Appendix I

“Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*”

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H.A., Weigel, D., (2018). *Nature Ecology & Evolution*, https://doi.org/10.1038/s41559-017-0423-0.

Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*

Moises Exposito-Alonso[1], François Vasseur[1§], Wei Ding[1], George Wang[1], Hernán A. Burbano[1,2], Detlef Weigel[1]*.

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[2]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

*Corresponding author. Email: weigel@weigelworld.org

[§]Current address: CNRS, UMR5175, Centre d'Ecologie Fonctionelle et Evolutive, F-34000 Montpellier, France (FV). [ς]Current address: Computomics, Davis, California, United States (GW).

Keywords: climate change, polygenic adaptation, GWA, environmental niche models, random forest, drought, *Arabidopsis thaliana*, image processing.

One-sentence summary: "Future genetic changes in *A. thaliana* populations could be forecast by combining climate change models with genomic predictions based on experimental phenotypic data."

Short title: "Genetic adaptation to extreme drought in *A. thaliana*"

Because earth is currently experiencing dramatic climate change, it is of critical interest to understand how species will respond to it. The chance of a species to withstand climate change will likely depend on the diversity within the species and, particularly, whether there are subpopulations that are already adapted to extreme environments. However, most predictive studies ignore that species comprise genetically diverse individuals. We have identified genetic variants in *Arabidopsis thaliana* that are associated with survival of an extreme drought event, a major consequence of global warming. Subsequently, we determined how these variants are distributed across the native range of the species. Genetic alleles conferring higher drought survival showed signatures of polygenic adaptation, and were more frequently found in Mediterranean and Scandinavian regions. Using geo-environmental models, we predicted that Central European, but not Mediterranean, populations might lag behind in adaptation by the end of the 21st century. Further analyses showed that a population decline could nevertheless be compensated by natural selection acting efficiently over standing variation or by migration of adapted individuals from populations at the margins of the species' distribution. These findings highlight the importance of within-species genetic heterogeneity in facilitating an evolutionary response to a changing climate.

Ongoing climate change has already shifted latitudinal and altitudinal distributions of many plant species[1]. Future changes in distributions by local extinctions and migrations are most commonly inferred from niche models that are based on current climate across species ranges[2,3]. Such approaches, however, ignore that an adaptive response can occur also *in situ* if there is sufficient variation in genes responsible for local adaptation[4–6]. The plant *Arabidopsis thaliana* is found under a wide range of contrasting environments, making it distinctively suited for studying evolutionary adaptation to a changing climate[7–9]. For the next 50 to 100 years, extreme drought events, potentially one of the strongest climate change-related selective pressures[10], are predicted to become pervasive across the Eurasian range of *A. thaliana[2,11]*. An attractive hypothesis is that populations from the Southern edge of the species' range[12] provide a reservoir of genetic variants that can make individuals resistant to future, more extreme, climate conditions[12,13]. To investigate the potential of *A. thaliana* to adapt to extreme drought events, we first linked genetic variation to survival under an experimental extreme-drought treatment. By combining genome-wide association (GWA) techniques that capture signals of local and/or polygenic adaptation[14] with environmental niche models[8,15], we then predicted genetic changes of populations under future climate change scenarios. An unexpected result of our predictions is that populations at both the Northern and Southern margins of the species' range will likely more easily adapt to increased extreme drought events, due to these populations carrying a greater spectrum of drought survival alleles.

RESULTS AND DISCUSSION

Differential survival to an extreme drought event. We began by exposing a high-quality subset of 211 geo-referenced natural inbred *A. thaliana* accessions[16] to an experimental extreme drought event during the vegetative phase, which killed the plants before they could reproduce (Table S1). After two weeks of normal growth, plants were challenged by a terminal severe drought for over six weeks and imaged every 2-4 days (Fig. 1A) (see Supplementary Methods section 2). To quantify the rate of leaf senescence, a polynomial linear mixed model was fit to the time series of green pixels per pot (Fig. 1B-D, Video S1). The average genotype deviations from the mean quadratic-term in the model provided the best estimate of this survivorship trait in late stages of drought (Supplementary Fig. 3, see details in Supplementary Methods), ranging from -5 to +5 x 10^{-4} green pixels/day^2. The most sensitive genotypes survived only about 32 days, while the most resilient plants survived about 15 days longer. Genotype-dependent survival probably reflects both constitutive as well as induced drought responses, i.e., both environment-dependent and -independent behaviors of the tested accessions. Additional environments need to be examined in order to disentangle these two types of responses.

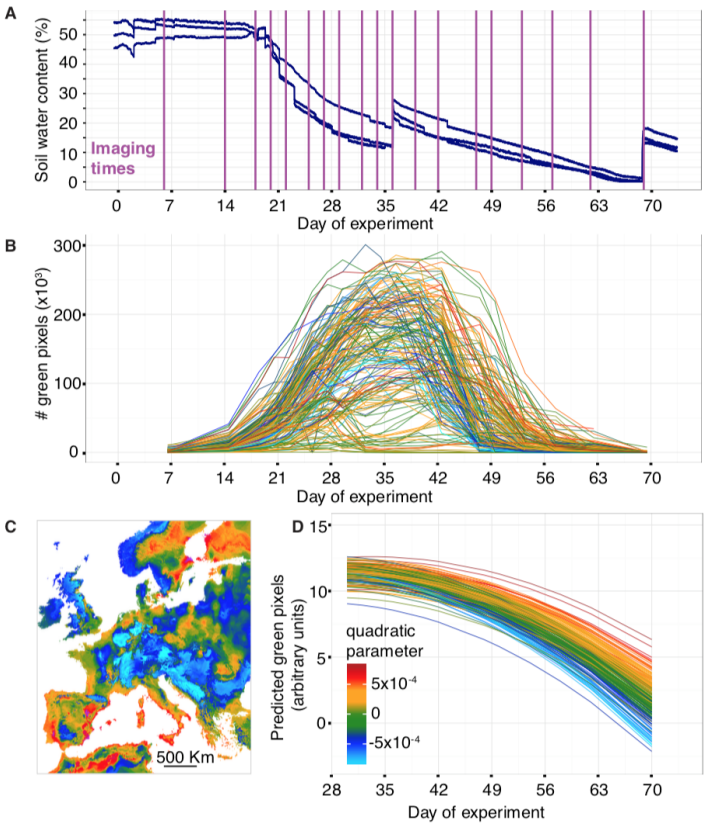


Figure 1. Terminal drought treatment and phenotyping of 211 accessions.

(A) Soil water content as measured by sensors in three well spaced experimental trays. Vertical lines indicate

dates of image acquisition. **(B)** Trajectories of total rosette area of 200 randomly chosen pots (see Video S1). Color index according to quadratic parameter in (D). (**C**) Map projection of the environmental niche model prediction of the quadratic parameter (the drought-survival index) in (D). **(D)** Decay trajectory modeled with a polynomial regression, with genotypes as random factors, from the day of maximum number of green pixels until the end of the experiment. Each line corresponds to one genotype.

The amount of water available during our drought experiment translates to only about 30-40 mm of monthly rainfall, and as expected, accessions with higher survival come from regions with low precipitation during the warmest season (correlation with climate variable bio18 [www.worldclim.org, ref. [17]]: Pearson correlation, r=-0.19, p=0.005), and specifically with low precipitation during May and June (r≤-0.19, p≤0.005) (see Fig. 2A). To further exploit current climatic data, we used 19 bioclimatic variables and random forest models[18] for environmental niche modeling (ENM) to predict the geographic distribution of the drought-survival index across Europe (Fig. 1C). Surprisingly, we found that individuals with higher drought survival were not only likely to be present around the Mediterranean, but also at the opposite end of the species' range in Sweden[19] (Fig. 1C, ENM cross-validation accuracy=89%, Table S10). In contrast to the warm-dry Mediterranean climate, Scandinavian dry periods occur on average at freezing temperatures (Supplementary Fig. 12). Consequently, precipitation might occur as snow and soil water content is frozen, thus water is not accessible to plants, producing a physiological drought response[20].

Survival across geographically structured population lineages. We then studied whether the different genetic lineages of *A. thaliana* are locally adapted[6] to low precipitation regimes via increased drought-survival. Using an extended panel of 762 *A. thaliana* accessions (Table S1) we carried out genetic clustering[21] and studied population size trajectories[22] (Fig. 2). This corroborated the existence of a so-called Mediterranean 'relict' group[12] and ten other derived groups of relictual (e.g. Spanish groups) or other (e.g. Central Europe) origin, as an apparent result of complex migration and admixture processes[23]. A generalized linear model indicated that genetic group membership explained a significant amount of drought-survival variance (GLM: R^2=12.8%; p=4 x 10^{-5}), with the North (N) Swedish and Northeastern (NE) Spanish groups each having on average higher survival than the other groups (t-test p≤0.01). A population graph estimated by Treemix[24] suggested a gene flow edge between the Mediterranean and Scandinavian drought-resistant genetic groups, potentially indicative of historical sharing of drought survival alleles (Fig. 2D). Finally, an ENM of the genetic group membership with climatic variables from the accession's geographic origin confirmed that the most important predictive variable of genetic structure was precipitation during the warmest quarter (bio18), followed by mean temperature of the driest quarter (bio9), and minimum temperature of the coldest month (bio6) (ENM accuracy > 95%. Supplementary Fig. 8 and

Table S10). As our results indicate that the deepest genetic split parallels contrasts in local precipitation regimes and ability to survive drought, we expect that decline in rainfall could lead to a future loss of certain genetic groups and/or to turnover of genetic diversity[11] (see Fig.12 Supplementary Fig. 8).

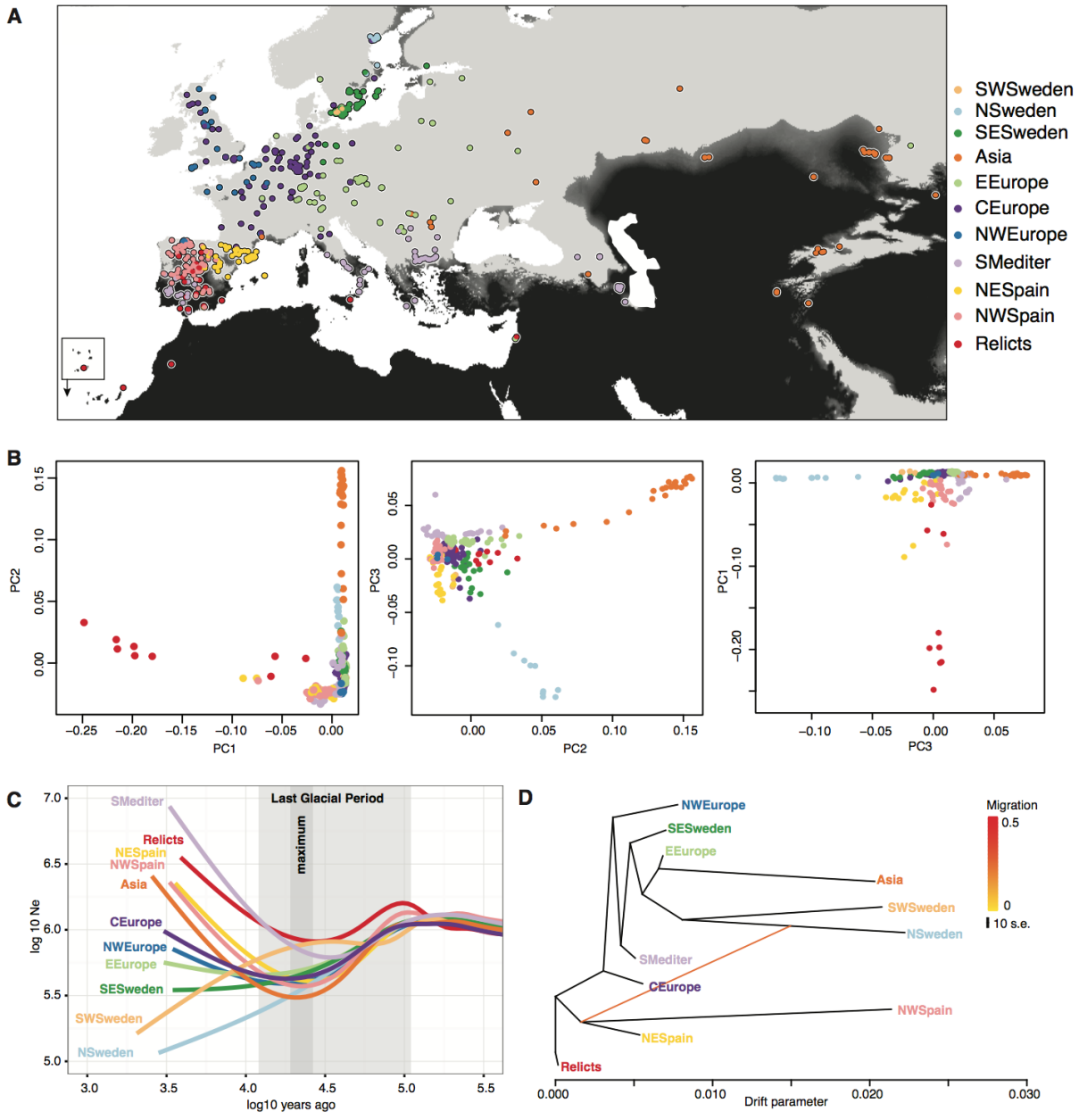


Figure 2. Population structure and history of 762 high-quality genomes.

(A) Geographic locations and 11 genetic clusters estimated by ADMIXTURE (k=11 having the lowest cross-validation error). Black indicates less than 40 mm of June rainfall (1960 to 1990 average), which corresponds to the amount of water provided in our drought experiment (Fig. 1). Note areas of very low June rainfall in the Mediterranean basin and along the coast in Scandinavia (partially obscured by colored circles). Cape Verde Islands are shown as inset. **(B)** Principal Component Analysis of genome-wide SNPs. **(C)** Effective

population sizes in time estimated from MSMC. **(D)** Population ancestral graph and the first migration trajectory from Treemix.

The genomic basis of survival. Because the potential of populations to adapt to drought will ultimately depend on specific genetic variants and the selected trait architecture, we identified drought-associated loci with EMMAX[25], a genome-wide association (GWA) method. Although genotype-associated variance[25] h^2 was relatively high, 50%, no individual SNP was significantly associated with drought survival (minimum $p \sim 10^{-7}$, after FDR or Bonferroni corrections $p>0.05$) (Supplementary Fig. 5, Table S3). Significant associations in multiple phenotypes have been detected in similarly powered *A. thaliana* experiments[26]. While multiple testing adjustment can over-correct p-values and obscure true associations, the absence of significant associations may also be due to (i) polygenic trait architecture, with many small-effect loci[27], and/or (ii) confounding by strong population structure, consistent with the association of drought survival with genetic group membership.

Polygenic signal of adaptation. To test for polygenic adaptation, we repeated the GWA analyses with a model that specifically handles both oligo- and polygenic architectures, BSLMM[28]. BSLMM estimates, among other parameters, the probability that each SNP comes from a group of major-effect loci. Around half of the top non-significant EMMAX SNPs were found to have over 99% probability of belonging to such a major-effect group (Fisher's exact test of overlap, $p=3 \times 10^{-7}$; see Supplementary Methods 3.3). We further tested the polygenic hypothesis using the population genetic approach of Berg & Coop[14]. The test is based on the principle that if populations diverge in a specific trait such as drought-survival that is due to many loci, there should be an orchestrated shift in their allele frequencies. After testing some 60 groups of EMMAX SNP hits of variable size and at different ranks, we detected the most significant signal of polygenic adaptation with the group that included the 151 top SNPs (Table S9). The signal was lost for ranks below the top 300-400 EMMAX SNPs (Table S9). We then compared summary statistics of the top 151 SNPs with background SNPs matched in frequency to avoid GWA discovery biases. The top 151 SNPs showed high F_{st} values, consistent with allele frequency differentiation between populations (Supplementary Fig. 5). Tajima's D values were positive (U Mann-Whitney $p<0.05$), indicating intermediate allele frequencies at the GWA loci (Supplementary Fig. 5), which could be a result of selection favoring alternative alleles in different ecological niches of the species[29]. The genomic regions containing the top SNPs did not show any evidence for precipitous reductions of haplotypic diversity, as would be expected for hard selective sweeps[30] (Supplementary Fig. 5). Together these patterns fit the expectations of local

adaptation from a polygenic trait controlled by some hundred loci[31] — a scenario that should enable a fast response to new environmental shifts.

Ancestry associations suggest a Mediterranean origin of survival alleles. During local adaptation, the relevant loci diverge due to natural selection across populations, which generates a statistical correlation with population groups[32]. In this situation, the default correction of population structure applied in GWA might obscure some of the true associations. There are cases where F_{st} scans can be useful to identify overly divergent loci that could be involved in local adaptation. However, in cases of strong population structure, the mean genome-wide F_{st} is high[32], complicating outlier detection (Supplementary Fig. 4). One can recover relevant variants that are deeply divergent across populations and therefore invisible to conventional GWA by first assigning ancestry to each SNP. Using ChromoPainter[33], which relies on linkage disequilibrium information, we segmented each genome in question into its different population ancestries (here 11 groups). The first outcome of this analysis was that individuals from NW and NE Spain and, to a lesser extent, the Southern Mediterranean (Fig. 2A), have inherited many DNA segments from relictual individuals (Supplementary Fig. 7). In a generalized linear model framework, we then tested whether the ancestries of individuals at a SNP coincided with the observed phenotypic differences in drought-survival. Performing this "ancestry" genome-wide association (aGWA) and using a permutation correction of p-values (see Supplementary Methods 3.6), we detected 8 distinct peaks ($p<0.001$, Fig. 3A) including over 1,000 significant SNPs (70 SNPs after linkage disequilibrium pruning) (Table S4). The most prominent peak was located on chromosome 5 and explained over 20% of the variance in drought survival (Table S4). There was no overlap in top SNPs between GWA and aGWA because they search for different association signals. Our aGWA resembles other admixture mapping techniques[34], and might be most useful for associations in scenarios of adaptive introgression and local adaptation. Although we do not know yet whether our observations can be generalized, our work demonstrates the power of using alternative GWA approaches in situations where adaptive variation is expected to be tightly linked to population history and structure.

To understand the origin of aGWA-identified SNPs, we constructed trees for all concatenated aGWA SNPs and for genome-wide background SNPs. Although the individuals from both the warm (Iberia and relicts) and cold (Scandinavia) edges of the species distribution are far apart in genome-wide SNPs, they are closely related in drought-associated SNPs (Fig. 3B). Overall, this is consistent with a common Mediterranean origin of drought-adaptive genetic variants of both Northern and Southern individuals (Fig. 2D, Fig. 3B), and highlights the relevance of populations at the latitudinal extremes of the species range as a possible genetic reservoir for future climate change adaptation[12].

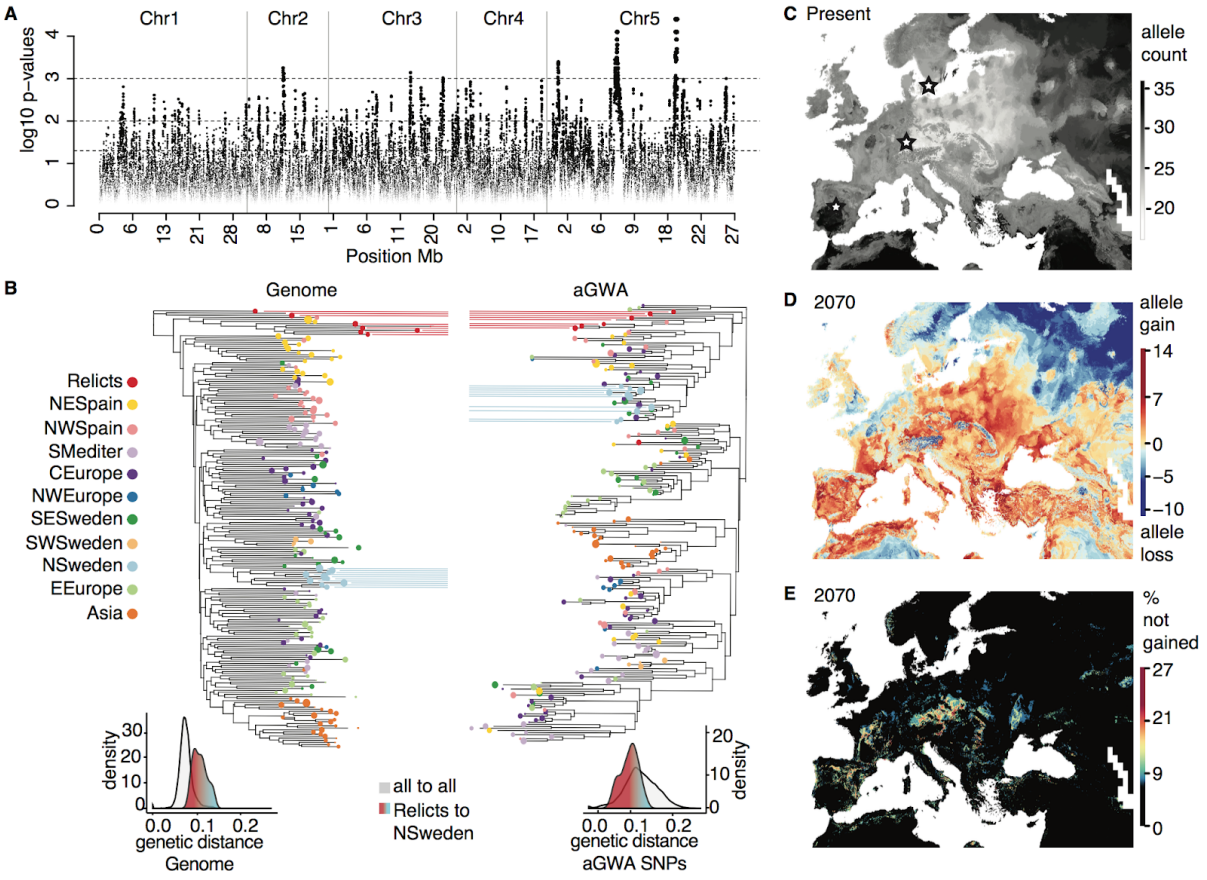


Figure 3. Ancestry GWA of drought survival and environmental predictions.

(**A**) Manhattan plot of SNPs from ancestry GWA (aGWA) after permutation correction of p-values. Dashed lines indicate significant thresholds at p<0.05, 0.01, and 0.001. (**B**) Top, Neighbour-joining phylogeny of 1,000 concatenated genome-wide SNPs compared with a phylogeny of all significant aGWA SNPs (ca. 1,000). Colors indicate population clusters (Fig. 2). Relicts and N. Swedish groups are highlighted. Bottom, genetic distances for genome-background SNPs or aGWA SNPs. (**C**) Environmental niche models of 70 top aGWA SNPs (after LD pruning), trained with climate averages from 1960-1990, and then (**D**) used to forecast gain or loss of alleles in 2070 under free migration. (**E**) Discrepancy of alleles that can be gained by 2070 between the geographically constrained (PCA control) model and the free migration model.

Drought survival is a resilience trait independent on phenology. Drought adaptation can be accomplished by diverse mechanisms, with cross-stress resistance being pervasive[35]. An annual life history enables drought survival through an escape strategy based on the acceleration of the life cycle from germination to flowering and seed production. An alternative strategy, the avoidance strategy, is employed by many xeric perennials with increased water efficiency[36]. Previous drought experiments with *A. thaliana* have shown that both strategies exist, although early flowering, which is associated with an escape strategy, was more favourable under water-limiting conditions[37,38]. In

our experiment, drought-survival was not negatively correlated with flowering time in unstressed conditions[39] (Pearson correlation, r=0.07, p=0.12). Although a correlation was not significant at the individual ecotype level, the GWA effect sizes of drought-survival for the top 151 SNPs were positively correlated with the ability of the same SNPs to delay flowering (Pearson correlation, r=0.51, $p=1x10^{-11}$, see Supplementary Methods 3.4). Given the described trade-off between escape by flowering and water use efficiency in *A. thaliana[37,40,41]*, our drought-survival index might be related to the avoidance strategy, although this needs to be tested with specific physiological experiments (Supplementary Fig. 11, Table S6). Gene enrichment analysis revealed a weak signal for membrane transport (see Supplementary Methods 3.7). Adjustment of osmotic balance through cell membrane transport is a drought avoidance mechanism[42] that might also confer cross-tolerance to other abiotic stresses[43]. Therefore, it might be of relevance for Scandinavian *A. thaliana* accessions or other populations in extreme environments (Supplementary Fig. 12)[19].

Forecast of genetic changes to global warming reveals regional differences in evolutionary potential. It is expected that populations with increased survival to severe abiotic stresses should have an evolutionary advantage in face of the predicted increase in drought frequency and intensity both around the Mediterranean and in Europe, which will constitute a critical hazard for many plants[2,11], including *A. thaliana*. Surprisingly, environmental niche models (ENM) of species distributions, which have been used to predict future changes of species' ranges [2,3], do not usually include information of within-species diversity that can lead to adaptation from standing variation[44–46]. This could in turn lead to overestimates of extinction rates[47–49]. By fitting ENMs of current climate with SNP data, using a similar rationale as for the "climate GWA" of Hancock and colleagues[7], we attempted to forecast the most likely genetic makeup under current and future climate conditions. We trained one ENM for each of the 151 GWA and 70 aGWA drought-associated SNPs to predict which allele, either the high or the low survival one, is more likely, given a set of environmental variables (all ENM 5CV accuracy >92%; Table S3-4, Supplementary Fig. 13-16). Consequently, from each model, we geographically mapped the potential distribution of the high survival allele using available environmental datasets (www.worldclim.org, ref. [17]). Finally, concatenating the resulting 221 maps, we inferred the most likely individual genotype at each location. At present, individuals from both northern and southern edges of the species' Eurasian and N. African range are predicted to harbor more drought-survival alleles than those located in between (Fig. 3C, Supplementary Fig. 15-16, with the quadratic term in a regression of allele count on latitude being positive at $p=10^{-3}$), corroborating our previous observations. Using the trained ENM, we also forecast the distribution of the 221 drought-survival alleles in 2070 (rpc 8.5, IPCC, www.ipcc.ch, ref. [17]). While it was expected that populations in the Mediterranean Basin need to become more

drought resistant[11], our predictions anticipate a greater increase in the total number of drought-survival alleles for Central Europe (Fig. 3, Supplementary Fig. 14-15). This is because by 2070 rainfall in Central Europe will likely become more similar to that in the Mediterranean[2,11] (Supplementary Fig. 12).

Because some drought-survival alleles are currently not present in Central Europe, we speculated that gene migration might be necessary to facilitate adaptation to future conditions[50]. An underlying assumption of the ENM is that alleles will be present wherever required by the environment, but this assumption of "universal migration" may not be realistic for future predictions if the presence of alleles is currently geographically restricted. We therefore included two geographic boundary conditions in the ENM to generate alternative models that were either more or less "migration-limited" (see Supplementary Methods 4.2). After fitting all possible models and predicting allele distributions with future climate, we calculated the difference of predicted allele presence per map grid cell between the naïve, free migration ENM and the two geographically constrained ones (Fig. 3D-E). If an allele has currently a narrow distribution or is specific to a certain genetic background, its future presence in an area might not be predicted by the constrained models, even though the climate variables coincide with the SNP's environmental range. Such a scenario seems to apply to Central Europe, as the deficit in drought-survival alleles predicted by the free over the constrained models was 8-30% (18-66 out of 221) (Fig. 3E; with the quadratic term in a regression of the allele count difference on latitude being negative at $p<10^{-10}$). Central European populations may therefore be under threat of lagging adaptation by the end of the 21st century.

In the end, for a population to persist, not only must drought-survival alleles be present locally, but they also need to increase in frequency[51]. The chance of this occurring will depend on current local allele frequencies and the strength of natural selection favouring the drought-survival alleles. Therefore, we studied current allele frequencies at three representative locations with the highest sampling density in our dataset (40 samples within a 50 kilometer area): Madrid (Spain), Tübingen (Germany) and Malmö (Sweden), which are near the southern edge, center and northern edge of the Eurasian and N. African range, respectively. Based on ENM predictions, we calculated allele frequency changes from present to 2070. Frequencies are predicted to increase significantly only in the Tübingen population (Student's t test, $p<10^{-16}$, Table S11), but not in Madrid and Malmö, indicating that these two populations might already be adapted to the future local climate. Although not all drought-associated alleles are found in Tübingen (32 of 70 aGWA SNPs and 136 of 151 GWA SNPs), increasing the number of the alleles in single genotypes should be feasible, since there are already single genotypes that have 24 (aGWA) and 123 (GWA) of these alleles (see Supplementary Methods 4.2). Running 50-generations simulations starting at the present Tübingen frequency of each of the drought-survival alleles and assuming a range of selection coefficients, we estimated that

a 1-3% of fitness advantage on average would be necessary to increase frequencies to match those of the adapted Madrid and Malmö populations (Supplementary Fig. 17, see Supplementary Methods 4.2). Such selection could take place efficiently when populations are large, as is typical for highly-proliferative weeds[51,52].

Conclusion

Leveraging the genetic resources available for *A. thaliana*, we have begun to address the question of how climate change will affect biodiversity. We provide evidence for the possibility of adaptive genetic variation to extreme drought events from standing variation. Specifically, we found that drought survival in *A. thaliana* has a polygenic basis and that favorable alleles are more abundant toward the edges of the species' distribution range. Extreme adaptation at range edges might thus be critical for a species' persistence under climate change. Although many aspects of future adaptation are not considered here, namely non-drought related or seasonal climate change[51], biotic interactions, phenotypic plasticity, or novel adaptive mutations, our spatially explicit analyses emphasize the potential of adaptive evolution from standing variation to mitigate climate change's detrimental effects.

METHODS

Study populations. 211 natural inbred lines from the 1001 Genomes project[16] were grown in a terminal drought experiment, and 762 lines were analyzed for genetic structure and genome-environment models. These two subsets were selected based on sequence quality and homogeneity of geographic distribution (see Supplementary Methods 1.1). We retrieved the genomes corresponding to the above natural lines from http://1001genomes.org/data/GMI-MPI/releases/v3.1/ and extracted the biallelic SNPs with >95% calling rate. This resulted in keeping ~4M SNP.

Genetic structure. To understand the genetic structure of *Arabidopsis thaliana* we ran, on the 762 samples, the software ADMIXTURE v1.2 (ref. [21]) assuming two to 20 groups and using a 5-fold cross-validation procedure. The number of groups with the smallest cross-validation error was 11 (Fig. 2, Fig. Supplementary Table 1, Supplementary Fig. 8). We computed a genomic PCA using PLINK v1.9 (ref. [53]). The three first PC axes explained 33.5% of the genomic variance (see Supplementary Methods 3).

We used genomes with probability >0.9 of assignment to one of the 11 ADMIXTURE groups to run MSMC v.3 (ref. [22]). This was done in quartets of genomes, i.e. four genomes for

within-population coalescent mode, and two genomes of each of two populations for the cross-coalescent mode (Fig. 2, Supplementary Fig. 5). Using the 11 genetic groups as population lineages, we run Treemix assuming zero to five migration edges[24] (Fig. 2, Supplementary Fig. 5)

Terminal drought experiment. Stratified seeds from the selected 211 natural lines were sown in greenhouse pots and abundantly watered every three days during two weeks. Thereafter watering only occurred every three weeks, which dramatically reduced soil water content (Fig. 1, Supplementary Methods 1.2). Top-view photographs of the potting trays were done at 20 timepoints during the whole experiment with a high resolution Panasonic DMC-TZ61 digital camera mounted in a closed black box setting to ensure image consistency (Supplementary Methods 2). Using customized Python scripts and the module Open Computer Vision, we segmented the green plant-leave pixels from the brown soil background to monitor plant area over time (Supplementary Video). Starting from the day with the largest rosettes areas, until the end of the experiment, we modeled the decay of green area (i.e. # pixels) using a polynomial generalized linear mixed model with Poisson link as described in the MCMCglmm R package v.2.25 (see Supplementary Methods 2). The random genotype effects captured the average deviation of each genotype from a general intercept, slope and quadratic curvature. After calculating the heritability of each of the three coefficient deviations and their correlation with the genotype's climate variables of origin, we understood that it was the quadratic curvature that was the most suitable to use as index of survival (Supplementary Methods 2).

Genome-Wide Association (GWA). Using the index of survival per genotype as the trait and the SNPs with a minimum allele frequency > 5% as predictors (n=879,654 SNPs), we carried out associations using the linear mixed model implemented in EMMAX software[25] to find SNPs that excessively contributed to the prediction of survival of genotypes (Supplementary Table 3) (see Supplementary Methods 3.3). To corroborate the identified top SNPs we also performed a Bayesian Sparse Linear Mixed Model (BSLMM) with GEMMA software[28]. EMMAX fist a model as: $Y = X_i\beta + Zu + \epsilon$, where Y is the vector of trait values, X is the alternative allele dosage at SNP i and β the allelic effect of SNP i on the trait. Population structure is corrected with a random genotype term (of 211 levels) represented by u, which follows a Multivariate Normal distribution $\mathcal{N}(0, A\sigma^2_G)$, where A is the relationship matrix between all individual genotypes built from SNP information and σ^2_G is the genotype-associated variance. Different from EMMAX, the BSLMM model of GEMMA fits a multilocus model such as: $Y = X\beta + \epsilon$, where all SNPs are fitted at once but there is a strong prior distribution of the β coefficients. These are constrained to follow a mixture of

two distributions, one that expects many small effects and another that generates few strong effects. Because all SNPs are included in the model, the population structure is implicitly accounted for.

To determine whether the top SNPs identified in the GWA might have been subject to polygenic adaptation, we used the method from Berg & Coop[14]. We did this for several groupings of top SNPs and reported the group that yielded the strongest signal (see all results in Supplementary Table 9).

Using painted chromosomes generated using ChromoPainter v. 2.0.7 (ref. [33]), we carried out another set of associations between the survival trait and the local ancestry category (11 groups) of a chunk of the genome. We used a linear model, $Y = \mu + X\beta + \epsilon$, and reported the positions in the genome with the least mean square error (i.e. highest R^2) (Supplementary Table 4). To compute p-values, we took an empirical p-value distribution approach based on 1,000 random permutation runs (see Supplementary Methods 3.6). To understand the ancestry of the associated genomic positions, we concatenated the SNP genotypes of the top-associated positions, computed genetic distances between natural lines and generated a Neighbour Joining tree. This tree was compared with a tree built from an equal number of randomly-picked background SNPs.

Genome-wide diversity and selection summary statistics. We calculated genome-wide F_{st} among the ADMIXTURE-defined groups and Tajima's D with PLINK v1.9 (ref. [53]) and likelihood of a selective sweep with SweeD (ref. [30]). We investigated the enrichment of the top SNPs in the upper tail of the distributions of those statistics by calculating a right-tailed t-test in contrast with genome-background SNPs with the same frequency values (Supplementary Fig. 4, Supplementary Table 3, rank columns).

Environmental Niche Models. We used classification and regression Random Forest models implemented in the *randomForest* R package, available environmental databases www.worldclim.com v.1.4 (ref. [17], 19 bioclimatic variables at 2.5 arc-minutes resolution), and geographic locations of GWA-identified alleles, to fit environmental niche models (ENM). To evaluate model's predictive ability for each allele, we used a 5-fold cross-validation procedure in which ⅘ parts of the data were used to train the model and ⅕ was used to test it. This enabled us to assign a percentage of successful assignment of an allele given the environmental variables at a location (Supplementary Tables 3-4). The fitted Random Forest model was used to generate potential geographic distributions of survival-associated alleles which, all overlapped, provided a geographic map of density of survival alleles. Using existing predictions of the same 19 bioclimatic variables to 2050 and 2070 under both low (2.6 rcp) and high (8.5 rcp) CO_2 accumulation scenarios, we re-predicted the distribution of alleles to the different future scenarios using the previously fitted

Random Forest models. Because of the implicit assumption of free movement of alleles, we generated two additional models per SNP: (1) ENM including the latitude and longitude variables in the Random Forest models and (2) ENM including the three first PC axes geographically modeled with present day climate (see below). By repeating predictions with future climate data, but keeping the latitude, longitude and PC components constant, some alleles would not be predicted in areas where the appropriate environment exists but which are outside of the current geographic distribution (1) or current local genomic background (2) (see Supplementary Methods 4, Supplementary Fig. 13-16).

Apart from the potential distribution of putatively adaptive alleles, we also modeled the geographic distribution of continuous traits, namely the aforementioned PCA components of population structure or the index of survival under drought itself. In those cases the Random Forest was of the regression type and the predictive ability was computed for the test data calculating the squared Pearson's correlation coefficient between predicted and true values (see Supplementary Methods 4).

To complement observations of presence and absence of alleles from ENM predictions, we carried out Wright-Fisher simulations of single biallelic SNPs (for details see Supplementary Methods 4.2.4). We ran simulations for 50 discrete generations. The population size was assumed of 300,000 plants, as inferred from diversity data, and was constant over time. Fitness was only determined by the selection coefficient of the drought alleles, which varied from 0 to 20% in an array of simulation runs. The starting frequency of the allele was set equal to the present day frequency of all natural lines sampled in a given geographic area (e.g., Tübingen). These simulations could be extended in the future to incorporate joint fitness effects from multiple adaptive mutations and complex environment-driven demographic processes (Supplementary Methods 4.2.4).

Code availability. Code for the image analysis pipeline available at http://github.com/MoisesExpositoAlonso/hippo with DOI: https://doi.org/10.5281/zenodo.1039888, code for ancestryGWA is available at https://github.com/MoisesExpositoAlonso/aGWA with DOI: https://doi.org/10.5281/zenodo.1039882, code for Wright-Fisher population simulations at http://github.com/MoisesExpositoAlonso/popgensim with DOI: https://doi.org/10.5281/zenodo.1039886.

Data availability. Phenotypic datasets available in the Supplementary Dataset at https://www.nature.com/articles/s41559-017-0423-0. Processed genome matrices are available at http://1001genomes.org/data/GMI-MPI/releases/v3.1/. Raw reads are stored in the www.ncbi.nlm.nih.gov/sra archive under the ID number: SRP056687.

Additional information

Supplementary information is available for this paper at: https://www.nature.com/articles/s41559-017-0423-0

Acknowledgements

We thank R. Wedegärtner for assistance with the greenhouse drought experiment, I. Henderson for the recombination map, the Petrov, Coop, Ross-Ibarra, Gaut and Schmitt labs for discussions. We thank J. Lasky, X. Picó, A. Hancock, H. Thomassen, T. Mitchell-Olds, J. Mujica, P. Lang, and D. Seymour for comments and the Weigel and Burbano labs for discussion. This work was supported by the President's Fund of the Max Planck Society, project "Darwin" to HAB and by central Max Planck Society funds and the ERC (AdG IMMUNEMESIS) to DW.

Author contributions

MEA conceived and designed the project. GW and FV helped and advised on image phenotyping and FV provided additional phenotypes. MEA and WD performed chromosome painter analyses. MEA performed the drought experiment, processed the image data, and designed and carried out the statistical analyses. DW and HAB advised and oversaw the project. MEA wrote the first draft and together with HAB and DW wrote the final manuscript with input from all authors.

The authors declare no competing financial interest.

REFERENCES

1. Parmesan C, Yohe G. A globally coherent fingerprint of climate change impacts across natural systems. Nature. 2003;421: 37–42. doi:10.1038/nature01286

2. Thuiller W, Lavorel S, Araújo MB, Sykes MT, Prentice IC. Climate change threats to plant diversity in Europe. Proc Natl Acad Sci U S A. National Acad Sciences; 2005;102: 8245–8250. doi:10.1073/pnas.0409902102

3. Jezkova T, Wiens JJ. Rates of change in climatic niches in plant and animal populations are much slower than projected climate change. Proc R Soc B. The Royal Society; 2016;283: 20162104. doi:10.1098/rspb.2016.2104

4. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends Ecol Evol. 2008;23: 38–44. doi:10.1016/j.tree.2007.09.008

5. Hereford J. A quantitative survey of local adaptation and fitness trade-offs. Am Nat. 2009;173: 579–588. doi:10.1086/597611

6. Turesson G. The species and the variety as ecological units. Hereditas. 1922;3: 100–113.

7. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, et al. Adaptation to climate across the Arabidopsis thaliana genome. Science. 2011;334: 83–86. doi:10.1126/science.1209244

8. Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. A map of local adaptation in Arabidopsis thaliana. Science. 2011;334: 86–89. doi:10.1126/science.1209271

9. Lasky JR, Des Marais DL, McKay JK, Richards JH, Juenger TE, Keitt TH. Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate. Mol Ecol. 2012;21: 5512–5529. doi:10.1111/j.1365-294X.2012.05709.x

10. Siepielski AM, Morrissey MB, Buoro M, Carlson SM, Caruso CM, Clegg SM, et al. Precipitation drives global variation in natural selection. Science. American Association for the Advancement of Science; 2017;355: 959–962. doi:10.1126/science.aag2773

11. Dai A. Increasing drought under global warming in observations and models. Nat Clim Chang. Nature Research; 2012;3: 52–58. doi:10.1038/nclimate1633

12. Hampe A, Petit RJ. Conserving biodiversity under climate change: the rear edge matters. Ecol Lett. Wiley Online Library; 2005;8: 461–467. doi:10.1111/j.1461-0248.2005.00739.x

13. Lee-Yaw JA, Kharouba HM, Bontrager M, Mahony C, Csergő AM, Noreen AME, et al. A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. Ecol Lett. 2016; doi:10.1111/ele.12604

14. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10: e1004412–e1004412. doi:10.1371/journal.pgen.1004412

15. Dormann CF, Schymanski SJ, Cabral J, Chuine I, Graham C, Hartig F, et al. Correlation and process in species distribution models: bridging a dichotomy. J Biogeogr. Blackwell Publishing Ltd; 2012;39: 2119–2131. doi:10.1111/j.1365-2699.2011.02659.x

16. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. Elsevier; 2016;166: 481–491. doi:10.1016/j.cell.2016.05.063

17. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. John Wiley & Sons, Ltd.; 2005;25: 1965–1978. doi:10.1002/joc.1276

18. Breiman L. Random Forests. Mach Learn. Kluwer Academic Publishers; 2001;45: 5–32. doi:10.1023/A:1010933404324

19. Mojica JP, Mullen J, Lovell JT, Monroe JG, Paul JR, Oakley CG, et al. Genetics of water use physiology in locally adapted Arabidopsis thaliana. Plant Sci. 2016;251: 12–22. doi:10.1016/j.plantsci.2016.03.015

20. Ingram J, Bartels D. The molecular basis of dehydration tolerance in plants. Annu Rev Plant Physiol Plant Mol Biol. annualreviews.org; 1996;47: 377–403. doi:10.1146/annurev.arplant.47.1.377

21. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

22. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. Nature Publishing Group; 2014;46: 919–925. doi:10.1038/ng.3015

23. Lee C-R, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, et al. On the post-glacial spread of human commensal Arabidopsis thaliana. Nat Commun. 2017;8: 14458. doi:10.1038/ncomms14458

24. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. Public Library of Science; 2012;8: e1002967. doi:10.1371/journal.pgen.1002967

25. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42: 348–354. doi:10.1038/ng.548

26. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465: 627–631. doi:10.1038/nature08800

27. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. Nature Publishing Group; 2011;13: 135–145. doi:10.1038/nrg3118

28. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013;9: e1003264. doi:10.1371/journal.pgen.1003264

29. Hedrick PW. Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. Annu Rev Ecol Evol Syst. 2006;37: 67–93. doi:10.1146/annurev.ecolsys.37.091305.110132

30. Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol. 2013;30: 2224–2234. doi:10.1093/molbev/mst112

31. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. Elsevier Ltd; 2010;20: R208–15. doi:10.1016/j.cub.2009.11.055

32. Josephs EB, Stinchcombe JR, Wright SI. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? New Phytol. 2017; doi:10.1111/nph.14410

33. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8: e1002453. doi:10.1371/journal.pgen.1002453

34. Shriner D, Adeyemo A, Ramos E, Chen G, Rotimi CN. Mapping of disease-associated variants in admixed populations. Genome Biol. 2011;12: 223. doi:10.1186/gb-2011-12-5-223

35. Tardieu F. Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. J Exp Bot. 2012;63: 25–31. doi:10.1093/jxb/err269

36. Ludlow MM. Strategies of response to water stress. In: Kreeb KH, Richter H, Minckley TM, editors. Structural and functional responses to environmental stress. The Hague, the Netherlands: SPB Academic.; 1989. pp. 269–281.

37. Kenney AM, McKay JK, Richards JH, Juenger TE. Direct and indirect selection on flowering time,

water-use efficiency (WUE, δ13C), and WUE plasticity to drought in Arabidopsis thaliana. Ecol Evol. Wiley Online Library; 2014;4: 4505–4521. Available: http://onlinelibrary.wiley.com/doi/10.1002/ece3.1270/full

38. Bac-Molenaar JA, Granier C, Keurentjes JJB, Vreugdenhil D. Genome-wide association mapping of time-dependent growth responses to moderate drought stress in Arabidopsis. Plant Cell Environ. Wiley Online Library; 2016;39: 88–102. doi:10.1111/pce.12595

39. Vasseur F, Wang G, Bresson J, Schwab R, Weigel D. Image-based methods for phenotyping growth dynamics and fitness in large plant populations [Internet]. bioRxiv. 2017. p. 208512. doi:10.1101/208512

40. Juenger TE, Mckay JK, Hausmann N, Keurentjes JJB, Sen S, Stowe KA, et al. Identification and characterization of QTL underlying whole-plant physiology in Arabidopsis thaliana: δ13C, stomatal conductance and transpiration efficiency. Plant Cell Environ. Blackwell Science Ltd; 2005;28: 697–708. doi:10.1111/j.1365-3040.2004.01313.x

41. McKay JK, Richards JH, Mitchell-Olds T. Genetics of drought adaptation in Arabidopsis thaliana: I. Pleiotropy contributes to genetic correlations among ecological traits. Mol Ecol. Wiley Online Library; 2003;12: 1137–1151. Available: https://www.ncbi.nlm.nih.gov/pubmed/12694278

42. Jarzyniak KM, Jasiński M. Membrane transporters and drought resistance - a complex issue. Front Plant Sci. 2014;5: 687. doi:10.3389/fpls.2014.00687

43. Swindell WR. The association among gene expression responses to nine abiotic stress treatments in Arabidopsis thaliana. Genetics. 2006;174: 1811–1824. doi:10.1534/genetics.106.061374

44. Pauls SU, Nowak C, Bálint M, Pfenninger M. The impact of global climate change on genetic diversity within populations and species. Mol Ecol. 2013;22: 925–946. doi:10.1111/mec.12152

45. Brown JL, Weber JJ, Alvarado-Serrano DF, Hickerson MJ, Franks SJ, Carnaval AC. Predicting the genetic consequences of future climate change: The power of coupling spatial demography, the coalescent, and historical landscape changes. Am J Bot. 2016;103: 153–163. doi:10.3732/ajb.1500117

46. Fitzpatrick MC, Keller SR. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecol Lett. 2015;18: 1–16. doi:10.1111/ele.12376

47. Catullo RA, Ferrier S, Hoffmann AA. Extending spatial modelling of climate change responses beyond the realized niche: estimating, and accommodating, physiological limits and adaptive evolution. Glob Ecol Biogeogr. 2015;24: 1192–1202. doi:10.1111/geb.12344

48. Moritz C, Agudo R. The future of species under climate change: resilience or decline? Science. 2013;341: 504–508. doi:10.1126/science.1237190

49. Hoffmann AA, Sgrò CM. Climate change and evolutionary adaptation. Nature. 2011;470: 479–485. doi:10.1038/nature09670

50. Aitken SN, Whitlock MC. Assisted gene flow to facilitate local adaptation to climate change. Annu Rev Ecol Evol Syst. Annual Reviews; 2013;44: 367–388. doi:10.1146/annurev-ecolsys-110512-135747

51. Fournier-Level A, Perry EO, Wang JA, Braun PT, Migneault A, Cooper MD, et al. Predicting the evolutionary dynamics of seasonal adaptation to novel climates in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2016;113: E2812–21. doi:10.1073/pnas.1517456113

52. Roux F, Giancola S, Durand S, Reboud X. Building of an experimental cline with Arabidopsis thaliana to estimate herbicide fitness cost. Genetics. 2006;173: 1023–1031. doi:10.1534/genetics.104.036541

53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

Supplementary Information Guide for:

Exposito-Alonso et al.**: Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana***

SUPPLEMENTARY VIDEO in Supplementary_Video.gif **71**

SUPPLEMENTARY METHODS

1. Experimental design and biological material

1.1 Choice of accessions from the 1001 Genomes resource

The 1001 Genomes project has released resequencing data for 1,135 natural inbred lines, also called accessions or ecotypes (http://1001genomes.org). We applied several filters to select the most informative, least biased accessions for our experiment. (i) The first filter removed 176 accessions with low quality genome information, < 10X genome coverage and < 90% congruence of SNPs called from Max Planck Institute and Gregor Mendel Institute pipelines[1]. (ii) The second filter removed 244 nearly-identical accessions, many from N. America. For this, we calculated pairwise genome-wide identity-by-state differences using PLINK v1.9 (ref. [2]). When pairs differed in less than < 0.01 changes per polymorphic site, we randomly removed one member of the pair. The overlap between (i) and (ii) was 762 accessions (Supplementary Fig. 1, 2, Supplementary Table 1). For geographic analyses in the native Eurasian and N. African range (e.g. environmental niche models), we used the 729 accessions that were within 50°W to 100°E longitude (see section 4.2.1). For the terminal drought experiment, we used 211 of these 729 accessions. The seeds were progeny of 1001 Genomes collection seed stocks obtained from the Arabidopsis Biological Resource Center (CS78942).

1.2 Greenhouse terminal drought experiment

The 211 accessions included both vernalization-requiring, slow-flowering and vernalization-independent, fast-flowering ones. Because of the difficulties associated with disentangling drought-induced mortality and reproduction-associated senescence at the end of the plant life cycle, our study focused on lethal drought stress during the vegetative stage, i.e., before flowering. We did not apply a vernalization treatment to reduce flowering time variance. (Note that onset of flowering, or flowering time, was not a confounding factor. See section 2.2.)

Seeds were aliquoted in Eppendorf tubes, suspended in 1% agar solution, stratified in a 4ºC dark room for 5 days to promote germination, and then pipetted into pots filled with sifted soil (CL-P, Einheitserde Werkverband e.V., Deutschland). When multiple seeds germinated per pot, all but one were removed at random. We sowed 8 replicates per genotype in 49 trays of 8 x 5 cells (5.5 x 5.5 x 10 cm) using a randomized incomplete block design. We excluded corner cells, where edge effects are

strongest.

During the first two weeks after sowing, from the time point defined as day 0, trays were watered close to soil saturation once every 3 days, with temperature maxima from 20 to 25ºC under 16 hours natural and supplemental light (the experiment ran during the months of July and August). The photosynthetic active radiation (PAR, wavelengths from 400 to 700 nm) was on average 5.73 mole m^{-2} day^{-1}, and ranged from 0.1 mole m^{-2} day^{-1} during the night to a maximum of 15 to 66 mole m^{-2} day^{-1} during the day, depending on insolation. After this period, seedlings were challenged with a terminal drought, with "recovery waterings" after 3 and 6 weeks (see small peaks in Fig. 1), in order to increase the variance in survival. The overall watering during the drought period (4 l in each tray of 40 x 60 cm), correspond to approximately 33 mm of rainfall (4,000 + 4,000 cm^3 water/ 2,400 cm^2 surface = 3.3 cm). We monitored water content using moisture sensors (Parrot SA, Paris, France) (see water content graph in Fig. 1A). We monitored rosette green area by imaging at 20 time points (Fig. 1A) using a customized system (see below).

1.2.1 Advantages and disadvantages of the experimental system

Drought experiments are known to be difficult in that different plants might not experience the same stress depending on their developmental stage and size because they consume the available water at different rates[3,4]. Therefore, it would be ideal to measure each pot's water content and adjust the watering of each pot every day[3,4]. In our experimental design we used over 1,600 pots, thus pot-specific watering adjustment was not feasible[5]. Automatized setups[6] can be powerful systems for such accurate watering and physiological studies, but they come with a limitation in the number of pots used (around 500, ref. [6]) and are difficult to implement in field conditions. In our design, we compromised pot-specific monitoring and water adjustment for high-throughput and scalability to field conditions.

Because we were concerned that the aforementioned effect of plant size could dramatically affect the results, we took several precautions. First, we used pots filled with about 245 cm^3 of soil (6 x 6 cm surface area x 10 cm depth), which is about twice as much than the more typical 125 cm^3 soil (5 x 5 x 5cm) often used in *A. thaliana* experiments. By using larger pots, we aimed to reduce the impact of plant differential water consumption relative to the impact of drought treatment. Second, we designed our experiment as a terminal drought to make sure all plants quickly experienced a water availability that was below their individual physiological limits. We argue that this approach should be less sensitive to differential water consumption of plants than if we had only imposed a moderate drought[5], which may be of more interest when trying to simulate an agricultural setting, rather than an extreme event in natural settings without supplementing water. Third, if plant size was

a major contributor to our drought survival index, there ought to be a relationship between plant size and drought survival. We did not find such a relationship (see section 3.4).

1.3 Validation field experiment

In order to validate the observations of local adaptation to low rainfall regimes from the greenhouse extreme drought experiment, we carried out a follow-up field experiment utilizing the same experimental design as for the previous greenhouse experiment. We sowed the same genotypes in Madrid (Spain) and in Tübingen (Germany)[7]. We grew plants in semi-natural conditions, using the same industrial soil and trays in both locations, under PVC foil tunnels with openings to the outside. The plants experienced temperatures and photoperiods very close to natural conditions, while at the same time we could artificially control the rainfall amounts. We simulated in both locations Spanish and German average precipitation based on real-time monitoring of rainfall with sensors next to the foil tunnels. We had in total four treatment combinations of two photo-temperature regimes and two rainfall regimes. Out of 7 pots per genotype, we counted how many survived until reproduction (i.e., produced fruits). The drought-survival index measured in the greenhouse correlated with this field survival variable in the low-precipitation treatment in Madrid (Spearman's rho=0.17, n=211, p=0.01) but did not in the high-precipitation treatment in Tübingen (p=0.99), as expected (see Supplementary Table 1 for genotype values).

We used the same imaging system in the field as for the greenhouse experiment (section 2.1), and thus had exceptional power to validate the robustness of the imaging system. We used images of the same trays that had been photographed twice on the same day to measure replicability of the image pipeline (these pairs of pictures included multiple trays at 11 timepoints). Spearman's rank correlation of green pixels per pot retrieved from two such images was 0.97 (n=1,508, $p<10^{-16}$). This high correlation indicates a high replicability of the illumination and segmentation procedures. The small error would become even more negligible when averaging all replicates per genotype and modeling trajectories per genotypes.

1.4 Experiment under optimal growing conditions

In a first experiment, we grew the same 211 genotypes under optimal watering and nutrient conditions and monitored vegetative growth by image analysis[8] (see Supplementary Table 5 for a description of the 24 traits extracted from the images). This set of traits was used to investigate whether variation of drought-survival index was correlated with growth under optimal conditions.

2. Drought phenotyping

2.1 Image analysis pipeline

Plants were imaged using a Panasonic DMC-TZ61 digital camera and a customized closed black box at a distance of 40 cm from the tray. This produced very consistent images in terms of illumination (only from in-camera flash) and focal distance to the plants. After benchmarking different camera settings, we set relative exposure to -⅔ and ISO to 100. White balance was set for flashlight illumination, which was consistent thanks to the customized black box.

We extracted leaf area per plant over time using the imaging module Open Computer Vision in Python (ref. [9]) (Video S1), with these steps: (i) 5 pixel mean denoise of the whole-tray image. (ii) Fixed Hue Saturation Value (HSV) segmentation of "green" values. The threshold values were determined manually by selecting pixels from plants at five timepoints through the temporal series to capture the different green values that plants display at different stages of development. (iii) Cropping of each pot to extract individual plant images. (iv) Counting of green pixels (code available at http://github.com/MoisesExpositoAlonso/hippo with DOI: https://doi.org/10.5281/zenodo.1039888). Pots with green pixels but without plants were excluded after careful visual inspection of all images.

2.2. Drought survival index

After determining the peak of green area for the majority of pots, we modeled the daily number of green pixels per pot. Several different models, including up to third order polynomial models, and several error correction factors, either raw or genotype averages, were tested. All models were ranked based on parameter convergence in an Monte Carlo Markov Chain (MCMC) walk and AIC values. The final model was a generalized linear mixed model with Poisson link of the form:

$$y = i + s\,t + q\,t^2 + \epsilon_{gi} + \epsilon_{gs} + \epsilon_{gq} + \epsilon_{tray} + \epsilon_{pos} + \epsilon$$

, where green area, *y,* was the response variable, and an intercept (*i*), slope (*s*), and quadratic coefficients (*q*) with time (*t*) were fitted as fixed effects. Genotypes were treated as random factors that were allowed to deviate from the main trends, following a normal distribution (0, $I\sigma_g$). Tray block and position within the tray grid were fitted as random factors also following a normal distribution. To estimate these parameters, we performed 10,000 iterations in a MCMC and 1,000 burn-in using the glmmMCMC R package (ref. [10]).

The variance from all genotype-dependent components relative to the total phenotypic variance was ~10%. Genotype values of the three parameters of interest (intercept, slope, and

quadratic coefficient) were used for Genome Wide Association (GWA) and downstream analyses. Additive genetic variance was estimated from linear mixed models using a kinship matrix (see GWA section (3.3)). The intercept, slope and quadratic deviations had narrow-sense heritabilities (h^2, or kinship-associated variance) of 0%, 0%, and 49%; respectively. We chose the latter as the drought-survival index. This parameter informed about survival during the late stage of the experiment, as can be observed from a high correlation between the drought survival index and the raw green pixels in the final monitored days (Supplementary Fig. 3).

Because the drought-survival index could depend on the developmental stage, size and subsequent water consumption of the plants when the drought treatment started, we computed the pixel decay polynomial model with and without a covariate of flowering time, rosette area and rosette dry mass under optimal conditions (indicative of overall developmental speed; see source of the phenotype in section 1.3 and phenotypic correlations in section 3.4.1). The model described above was run three times, each with one of the mentioned phenotypes as random factors to get the variance explained. The variances explained by each of the test factors (V_{factor}/V_{total}) were 1.15 x 10^{-4} (95% Highest Posterior Probability: 6.74 x 10^{-9}, 1.07 x 10^{-2}) for flowering time, 1.86 x 10^{-8} (9.38 x 10^{-11}, 1.60 x 10^{-5}) for rosette dry mass, and 2.67 x 10^{-7} (1.84 x 10^{-11}, 4.53 x 10^{-2}) for rosette area.

To provide an intuitive understanding of the drought survival index, we looked at the relationship between the index and the last day on which a plant was clearly alive, defined as the last day with at least 5,000 green pixels left. The relationship between the drought-survival index and the last living day was highly significant ($p < 10^{-16}$). The most sensitive plants survived for 32 days, and the most resistant ones were alive for 15 days longer.

3. Population and quantitative genetics

3.1 Population structure

From the vcf file with SNP calls of the 1001 Genomes project (http://1001genomes.org/data/GMI-MPI/releases/v3.1/), we identified biallelic SNPs with a genotype calling rate >95%, which resulted in a genome matrix of ~4 million SNPs x 762 accessions (section 1.1). We defined genetic clusters with ADMIXTURE v1.2 (ref. [11]) (Supplementary Table 2). As a model-free alternative to ADMIXTURE, we used PCA implemented in PLINK v1.9 (ref. [2]). The first three axes explained 16.0%, 9.6% and 7.9% of the total genetic variance. ADMIXTURE clustering and PCA were used to understand population structure and to relate it to phenotype variables. We assessed population splits and migrations with a population ancestral graph using TreeMix v. 1.12 (ref. [12]), a tree based on genome-wide allele frequency differences across populations.

Additionally, we calculated a proxy of local genetic diversity[13] at each of the 762 locations sampled by computing the genome-wide number of polymorphic sites between such a focal sample and the geographically closest other sample.

3.1.1 Association of genetic group membership with drought

Using the ADMIXTURE membership probabilities of each genome, we carried out univariate linear regressions with drought survival index as phenotype. The groups that yielded positive relationships were NE. Spanish ($p<0.05$), Mediterranean ($p=0.06$), and the N. Swedish groups ($p<0.001$). The groups negatively associated were Central Europe ($p=0.06$), Asia ($p<0.001$), and E. Europe ($p<0.001$). This broadly coincided with the map of drought-survival prediction (Fig. 1D, S11). We also carried out regressions between the drought survival index and the first three axes from a genomic PCA and found that only PC3 was significantly associated with the drought survival index (GLM $R^2=0.076$; $p=5.15x10^{-5}$). The N. Swedish and NE. Spanish groups showed particularly low values in PC3 compared to the rest (Supplementary Fig. 8).

3.2 Coalescent rates over time

Only the accessions with ≥90% of membership probability in one of the genetic groups were used. Using MSMC v.3 (ref. [14], http://github.com/stschiff/msmc), we performed within-group coalescent with four genomes and cross-genetic group coalescent with two genomes for each group. In total, 333 runs were performed, with each genetic group being tested at least 3 times. The results were summarized using a smoothing generalized additive model in R (Fig. 2C).

3.3 Genome wide association (GWA) analyses

3.3.1 Linear Mixed Models (LMMs) with EMMAX

We used 879,654 biallelic SNPs with a minimum allele frequency (MAF) of 5% for genome wide association (GWA) using EMMAX (ref. [15]). We carried out GWA for all climatic variables and 11 phenotypes (Supplementary Table 5). The GWA is based on linear mixed models that test, one by one, each of the SNPs, and correct the results by population structure using a random factor with a variance/covariance kinship matrix built from genome-wide SNPs. In *A. thaliana*, which is a selfing species with geographically confined genetic lineages, this method can correct for coancestry[16].

To rule out the possibility that drought survival measurements were dependent on the developmental stage of the plant during the experiment, we carried out the GWA with and without a covariate of flowering time that had been scored in controlled conditions (flowering time being a proxy of developmental speed; section 1.3). The top SNP hits were the same with or without this

covariate, and we only show results without the covariate. To account for familywise error in GWA we used Bonferroni correction (p value x number of SNPs) and the Benjamini-Hochberg false discovery rate (FDR) correction[17]. The kinship-associated variance of drought-survival — an approximation of narrow sense heritability, h^2, was 49%. When we fit a kinship calculated from only the 151 top polygenic GWA SNPs (see section 3.5.2), the estimate of h^2 was 52%. This is probably a better estimate than that from the genome-wide-based kinship matrix, as the putatively causal SNPs are better "tagged" in the 151 SNP kinship matrix.

3.3.2 Bayesian Sparse Linear Mixed Models (BSMLMMs) with GEMMA

The Bayesian Sparse Linear Mixed model (BSLMM) implemented in GEMMA (ref. [18]) accommodates both poly- and oligogenic architectures in a GWA framework. It models two effect hyperparameters, a basal effect, *alpha*, that captures the fact that many SNPs contribute to the phenotype, and an extra effect, *beta*, that captures the stronger effect of only a subset of SNPs. The parameter measuring the probability of having another extra effect, *gamma*, can be used to prioritize SNPs (see Reference Manual of GEMMA, http://www.xzlab.org/software.html). Over 40% of the top 151 SNPs from EMMAX were found to have over 99% percentile of the gamma inclusion probability in GEMMA (Fisher's exact test odds ratio =17.21, $p=3x10^{-7}$). The estimate of realized heritability with BSLMM was 50%, which is in agreement with the EMMAX analyses. The 95% highest posterior density (95%HPD) from 1,000 MCMC steps ranged from 25-85%.

3.4 Multivariate analyses of phenotypes and GWA summary statistics

For a description and sources of all variables used, see Supplementary Table 5.

3.4.1 Pairwise correlations

We computed all-against-all Pearson product-moment correlation coefficients among accession line means (n=211 accessions) of phenotypic and climate variables (Supplementary Table 5, 6). To study genetic correlations, we performed the same analyses with SNP effect sizes (n=151 drought-associated SNPs) estimated from multiple GWA (Supplementary Table 7).

The phenotype correlations (Supplementary Table 6) showed that the drought survival index was negatively correlated with reproductive allocation and number of seeds ($r<-0.16$, $p<0.02$), suggesting a fitness trade-off between stressful and optimal growth environments. Drought-survival was not correlated with flowering time ($r=0.07$, $p=0.12$) nor plant size (rosette area $r=0.11$, $p=0.1$; and rosette dry mass, $r=0.12$, $p=0.07$) (Supplementary Table 6).

Drought survival SNP effects negatively correlated with the SNP effect sizes of most individual

precipitation variables ($r < -0.4$; $p < 10^{-8}$, Supplementary Table 7), indicating that alleles that increased drought survival were found in more arid geographic regions, i.e. regions with high temperatures and lower precipitation at different times of the year. Drought survival SNP effects were also positively correlated with SNP effects of rosette area, dry mass, and flowering time (Supplementary Table 7). These analyses have two-fold interests: (1) GWA-estimated effects have been corrected by population structure, thus correlations should not reflect phenotypic differences caused by drift of populations. (2) SNPs can have pleiotropic effects and this can limit adaptation due to genetic constraints[19] (see section 4.2.4.3).

3.4.2 Canonical Correlation Analysis (CCA)

We further utilized Canonical Correlation Analysis (CCA) using the CCA R package[20] to decompose environment-phenotype associations of SNP effects. This was done for all genome-wide SNPs (n=~800,000) and for the 151 drought-associated SNPs (Supplementary Table 8).

CCA of genome-wide SNPs revealed the first canonical correlation axis (CC1) to be driven by lower flowering time (T_repro, loading=-0.77), lower rosette dry mass (loading=-0.76) and higher annual temperature (bio1, loading=0.5). CC2 indicates that lower plant photosystem stress (FvFm, loading=0.60) is related to higher mean temperature of the wettest quarter and higher precipitation seasonality (bio8, bio15, loadings>0.25). CC3 shows that lower drought survival (loading=-0.58) effects are related to higher precipitation in the driest (bio17, loading=0.44) and warmest quarters (bio18, loading=0.35).

CCA of the 151 top GWA SNPs yielded a first canonical correlation coefficient of 0.99, with a phenotype canonical variate driven by lower drought survival, higher rosette area and dry mass (loadings >0.75), and a climatic canonical variate dominated by higher precipitation during the wettest month (bio13) and wettest quarter of year (bio16) (loadings >0.75).

3.5 Polygenic adaptation signal

3.5.1 Classic Q_{st}-F_{st} comparison

Q_{st}/F_{st} ratios have experimental evidence as appropriate indicators of local adaptation in *A. thaliana[21]* and are widely popular in evolutionary ecology studies[22]. Genome-wide F_{st} across the eleven populations was computed from 211 genomes using vcftools v0.1.12b (ref. [23]). We estimated the mean and confidence intervals based on the standard error of the mean, obtaining a mean F_{st}=0.042 (95% cumulative distribution =0.360). We calculated Q_{st} for the drought-survival index as the between-genetic group variance divided by the total variance. We used the MCMCglmm package v.2.25 in R (ref. [10]) with a 10,000 step chain, 1,000 burn-in steps, and fitted the genetic

group as random effect. This resulted in a global Q_{st} = 0.143 (90%HPD=0.052 - 0.338). We also calculated F_{st} and Q_{st} between each pair of subpopulations (Supplementary Table 12) using the same methods. This revealed that while many median Q_{st} values were above the median F_{st}, none were above the 95% percentile of F_{st} due to the tail of the distribution of background F_{st} being long.

When the variance across populations was calculated using the NE. Spanish and the N. Sweden population groups (the two groups with highest values of drought survival and thus putatively locally adapted) against the rest, we obtained Q_{st}= 0.377 (0.047 - 0.987). We thus concluded that a significant $Q_{st} > F_{st}$ signal is only observable at the individual level when the hypothetical populations that underwent local adaptation were used in the calculation of the variance.

3.5.2 Berg & Coop methodology

We tested for a polygenic adaptation signature following Berg & Coop[24], an extension of the Q_{st}/F_{st} ratio test based on SNP frequency per population and effect sizes as estimated from a GWA analysis. We used different groups and numbers of ranked SNPs after pruning linked SNPs ($r^2 > 0.6$), to learn about the robustness of this test and the apparent number of SNPs that contribute to the signal (Supplementary Table 9). Since this test does not use direct phenotypes but calculates the average phenotype per population based on allele frequencies of GWA SNPs, we could perform the test with 762 high quality accessions. Since results did not vary between analyses with 762 and 211 samples, we only report the analyses with 762 genomes (Supplementary Table 9). The median linkage disequilibrium of the final 151 top SNP set was r^2=0.26 (1s^t quartile: 0.22; 3rd quartile: 0.33) and the median distance between SNPs within the same chromosome was ~119 kb (15 kb, 2.6 Mb).

3.6 ChromoPainter and ancestry GWA

We ran ChromoPainter version 2.0.7 (available at http://paintmychromosomes.com; ref. [25]) on the 762 genomes dataset, after imputing missing genotypes with Beagle version 3.3.2 (ref. [26]) using default parameters. ChromoPainter analyses require a "training" run to estimate several hyperparameters. We ran 10 expectation maximization iterations on chromosome 2 (the smallest chromosome). We informed ChromoPainter with a published recombination map of *Arabidopsis thaliana[27]* that we reshaped to our SNP dataset. We used the command:

ChromoPainterv2 -i 10 -in -iM -j -g haplotypefile -r recombinationfile -a 0 0 -t labelfile

We used the output hyperparameters to run ChromoPainter on all chromosomes in an unsupervised all-to-all genomes mode, with the command:

ChromoPainterv2 -n 4.737068 -M 0.000421 -j -g haplotypefile -r recombinationfile -a 0 0 -t labelfile

3.6.1 Global proportion of ancestral chromosome segments

To study the ancestry relationships of each of the genetic groups, we counted the number of chromosomal segments (termed "chunks" in the original ChromoPainter paper[25]) that each genome "received" from all other genomes. The segment varied in size depending on local recombination rates and between genomes, but *a posteriori* analyses indicated that the median size was in the order of magnitude of kilo and megabases. To make the counts more informative, we show boxplots per ADMIXTURE group rather than counts per individual (Supplementary Fig. 7 A-K). This showed, for example, that NW. Spain, NE. Spain and S. Mediterranean (the latter to a lower degree), were "painted" mostly by relict DNA segments. Next, we tried to infer how well the drought survival of an individual correlated with the number of segments inherited from a certain ancestry. This indicated that only N. Sweden and relicts passed DNA segments that were correlated with the drought-survival index of the receiving individual (Supplementary Fig. 7L). The Pearson correlation coefficient was calculated excluding the individuals from the same admixture group as the predictor.

3.6.2 aGWA for admixture mapping

If populations are locally adapted, F_{st} outlier scans can be used to identify genetic variants under divergent selection[28,29]. However, when populations become isolated and diverge genetically, as is the case in *Arabidopsis thaliana*, F_{st} values are shifted to high values across the entire genome even when subsequent admixture happens, making the identification of outliers difficult[28] (Supplementary Fig. 4). Thus, we must rely on LD and identity by descent to find DNA segments characteristic of the different populations. If subsequent but incomplete admixture occurred between the locally adapted populations, it is expected that the individuals that retained the DNA segments responsible for local adaptation, would show the largest phenotypic differences with those that did not retain or never had the DNA segment. This is the principle of admixture mapping[30].

With the above rationale, we developed an admixture mapping technique[30] repurposing the output of ChromoPainter. The "painted" genome matrix produced by ChromoPainter has 762 states (one per individual in the analysis) and we repainted it into a genome matrix of 11 states (the genetic groups from ADMIXTURE analysis, which are geographically and environmentally separated). We then computed a regression of the drought-survival phenotype on the population group specific to a SNP as: $Y = \mu + X\beta + \epsilon$; where *Y* is a vector of *i=1…211* individual's phenotypes, μ is the mean phenotype, *A* is the 211 x 11 design sparse matrix of the ancestry states, *b* is a 1 x 11 vector of effects that each ancestry has in the mean phenotype, and ε is the uncorrelated random residuals

assumed to be normally distributed. This model was repeated for each SNP in our dataset (~2 million imputed and 'painted' SNPs, see section 3.6). We report R^2 and p-value of each SNP model (Supplementary Table 4). Since we already knew that the phenotype is associated with the membership assigned per individual, we expected that the ADMIXTURE membership of any random SNP would be on average also associated, because of linkage resulting from common ancestry. Therefore, we implemented an empirical p-value distribution correction to only detect those SNPs whose ancestry explained an even larger proportion of variance than the whole-genome ancestry. The permutation was done within each individual genome, shuffling the SNP states at a distance of 1,000 to 10,000 SNP positions — defined from analysing the typical size of "homogeneously painted DNA segments" (code is available at https://github.com/MoisesExpositoAlonso/aGWA with DOI :https://doi.org/10.5281/zenodo.1039882). We permuted the dataset 1,000 times and repeated this "aGWA" analysis to build p-value distributions. Since the nature of the associations is very different from that of a standard GWA analysis, we did not expect and did not find any overlap of top aGWA SNPs with the top SNPs from conventional GWA. The closest was a conventional GWA SNP that was 8 kb away from an aGWA SNP. The closest gene to both encodes a defensin-like protein; a family of proteins with broad anti-fungal and anti-bacterial activity[31].

Our approach is conceptually related to admixture mapping in humans, which has focused on local enrichment of Neandertal- and Denisovan-like variants, and which has led to the identification of a TLR immunity gene[32] as adaptive. It has also helped to increase the power for detection of background-dependent disease risk in humans with mixed ancestries, e.g. African-American individuals[33], or other more complex mixtures[34]. Such approaches constitute a powerful tool for understanding the genetic basis of local adaptation when complex demographic scenarios of admixture exist.

Phylogeny of aGWA SNPs

To learn about the distribution and shared ancestries of the drought-related alleles, we computed a neighbour joining phylogeny of all concatenated SNP hits from aGWA ($p<0.001$) and compared it with a genome background phylogeny of 1,000 randomly chosen genome-wide SNPs (Fig 3B). This revealed that while for genome-wide SNPs the distance between accessions from N. Swedish and Mediterranean relicts is higher than the average between any random pair of accessions (Student's t test, $p<2x10^{-16}$). However, when the same distances were calculated based on aGWA SNPs, the N. Swedish and Mediterranean relicts were much closer than the average pair of accessions (Student's t test, $p<2x10^{-16}$) (Fig. 3B). The same analyses showed also higher affinity of N. Swedish and NE. Spanish populations (Student's t test, $p<10^{-10}$).

3.7 Test for annotation enrichment

Using the TAIR10 gene annotation of *Arabidopsis thaliana* (available at arabidopsis.org/portals/genAnnotation/functional_annotation/), we tested whether a specific annotation class was enriched in our GWA and aGWA hits. Among genes overlapping with the 151 top GWA hits were the nitrate transporter gene *NRT1.8*, which among other functions mediates cadmium tolerance and is related to ABA transport[35–37], the *CATION/CARNITINE TRANSPORTER 4* (*OCT4*), which mediates homeostasis of metabolites and promotes lateral root formation[38], and the sugar transporter gene *SWEET8*, which is upregulated during salt stress[39]. The strongest peak unique to the aGWA hits fell inside the *CATION EXCHANGER 9*, a gene that is important for K^+, Na^+ and Mn^{++} homeostasis and which confers salt tolerance when introduced into yeast[40]. An empirical distribution test based on random draws of genes showed, however, only marginal enrichment. The 30 genes defined by the 151 top GWA SNPs were weakly enriched for gene annotation enrichment of cell membrane transport (6/30; p=0.01), and the 23 genes defined by the 70 top aGWA SNPs were marginally enriched for membrane transport (7/23; p=0.06). Testing for overrepresentation with PANTHER (www.pantherdb.org) and including genes adjacent to the GWA and aGWA SNPs revealed weak enrichment of aGWA genes for ferredoxin metabolic processing (p=0.03) and vesicle-mediated transport (p=0.05), and of GWA genes for growth-related functions (p=0.0007) and metabolite biosynthetic processes (p=0.0002). It is difficult to know what to conclude from this, but the most noteworthy finding is probably that there was no link to flowering time, in contrast to previous QTL and GWA studies of *A. thaliana* response to drought[41–43].

4. Environmental and forecasting analyses

4.1 Environmental data

The environmental data comprised the Last Glacial Period (LPG, ~22,000 years ago), recent averages from 1960-1990, and two 2070 climate projections of contrasting socio-economic scenarios, the 2.6 and 8.5 CO_2 representative concentration pathways[44,45] (rcp). The data were retrieved from www.worldclim.com v.1.4 (ref. [46]). They consist of 19 bioclimatic variables at 2.5 arc-minutes geographic resolution (CCSM4).

4.2 Environmental Niche Models (ENM)

We carried out ENMs with a number of response variables (for summary statistics see Supplementary Table 10), namely the drought-survival phenotype, flowering time, the genomic principal component axes, the discrete population groups, the local genetic diversity, and the SNPs

identified in GWA and aGWA analyses.

4.2.1 Geographic areas covered and niche limits

To train ENMs, we removed accessions from Japan and from N. America, as they are considered recent introductions[47] and their genetics might not reflect long-term climatic adaptation. The remaining sampled locations used to train the models were within 15 to 63º N latitude and 23°W to 88ºE longitude, but we only predicted from 34 to 63º N and 10.5°W to 35ºE, to avoid extrapolation of data. Predictions for the last glacial maxima were masked in those areas that were likely tundra or covered by ice sheets at the time (<5ºC and <0ºC annual temperature, respectively), as predictions for such areas would be irrelevant.

Because the sampling in the 1001 Genomes project was not even across the species range, predictions for underrepresented regions such as N. Africa, the Middle East, or Russia must be taken with caution. In order to be explicit about for which areas we could make the most robust predictions, we show the sampling density per 1°x1° latitude x longitude grid, which varies from 1 to 60 individuals (Supplementary Fig. 8D), and plot trends of predicted values against other variables, such as latitude or climate variables (e.g. Supplementary Fig. 9-11), only at those grid cells where there is at least one sample.

Finally, it is worth noting that even for the most pessimistic climate change scenario (rcp 8.5), the values of annual precipitation (bio1) and the precipitation during the warmest season (bio18) were always above the present minimum precipitation values where *A. thaliana* is currently found (see Supplementary Fig. 12). Therefore, we expect that transgressive phenotypes are not required to survive future climates.

4.2.2 Random Forest models

After trying different methods, including generalized linear models, MaxEnt, and linear discriminant models, we opted for random forest models because they are nonparametric, nonlinear, allow both continuous and discrete response variables, and are computationally efficient[48]. Additionally, the implementation of an "importance" parameter of each predictor variable available in the *randomForest* R package v.4.6 (ref. [49]) makes ranking of variables straightforward. To mitigate the overfitting problem typical of machine learning methods, a 5-fold cross validation procedure was used. We randomly divided the dataset into five parts, used four parts as training dataset and one part as testing dataset, and repeated this five times. Reported accuracy from cross validation was the R^2 of a linear model between observed and predicted values for continuous variables, and the rate of successful assignment of categories relative to the total number of observations for discrete

variables. To build the final forest, a total of 50 classification or regression trees per cross-validation set were used, and six variables were tested for each classification split.

4.2.3 ENM of genetic groups

We modeled the presence of population structure as a discrete response variable in ENM; either using eleven genetic groups as states, or the two relict and non-relict states.

In order to formally quantify the relevance of genetic group membership, we calculated the percentage of map grid cells that each genetic group occupies. For this we only considered areas where at least one genome per 1°x1° latitude x longitude was observed (Supplementary Fig. 8A) and where tundra or ice sheet are not expected (important for LGM comparisons).

When we used the present-data trained relict/non-relict ENM with past climate data from the last glacial maxima, we found that relicts likely occupied almost a quarter of the non-glaciated areas, compared to less than 2% today (Supplementary Fig. 8D), in agreement with genomic inferences of higher effective population size in the past (Fig. 2C). The reason that the relicts' environmental niche is predicted with 100% accuracy under 5-fold cross-validation (5CV) is likely that the local number of relict individuals is low, 26 accessions out of 762, and because their niche is very restricted.

Under a future high CO_2 increase socio-economic scenario, the ENM with 11 genetic groups predicts that the S. Mediterranean group will expand most dramatically into Central-European areas, replacing groups currently occupying these areas (Supplementary Fig. 8 C, F). Although these models are not mechanistic, they illustrate that genetic groups from the Mediterranean and from temperate areas have contrasting environmental niches and thus might replace each other under future climate warming.

4.2.4 Genome Environment Models (GEMs)

All 151 GWA and 70 aGWA SNPs were modeled as a bivariate discrete variable (drought-sensitive and drought-survival allele) in a random forest. The prediction accuracy and the most important predictor for each model is shown per SNP in Supplementary Table 3 and 4.

After modeling presence/absence of each drought-survival SNP, we projected the present inferred allele distributions in a map and then summarized all maps by intersecting them. In this way, we generated a continuous map surface of the total number of drought-resistant alleles in a given location. Ancestral GWA and conventional GWA models showed overall similar patterns (Supplementary Fig. 13-14), but the latter seemed to point to drought alleles being more

concentrated not only in Southern and Northern areas but also Western areas of Europe (e.g. UK, France). While this might be the case, it could also be due to conventional GWA SNPs suffering from high-frequency bias, making them more likely to be present in geographic areas with more samples (Supplementary Fig. 8). After we had trained the models with present data, we used them to predict allele distributions in 2070 under low and high CO_2 increase scenarios. While patterns were similar in both scenarios, for further analyses we used the most "pessimistic" high CO_2 increase scenario to be able to show main trends more clearly.

Migration assumptions

For each SNP we trained three models in order to overcome the "universal (or free) migration" assumption, implicit when using a current climate-trained ENMs with future climate data (e.g. ref. [50]). Although typically the free-migration model may not be entirely appropriate for predictions, it might be more realistic for cosmopolitan species with continental-scale migrations in the recent past, as is the case for *A. thaliana[47]*. Nevertheless, we designed two additional models to account for limited migration. The free model includes only the 19 bioclimatic variables as predictors of the drought-survival alleles. The first geographically-controlled model includes in addition the first three PC genomic axes as predictors (Supplementary Fig. 8G-J), in an attempt to limit prediction of allele presence to geographic areas where the genomic background that they reside on is present today. The second geographically-controlled model, which is even more restrictive, includes latitude and longitude together with the 19 bioclimatic variables. For all models we not only show the predicted maps (Supplementary Fig. 13-14) but also provide residuals of predicted *vs* observed (empirical) number of alleles in the locations where we have a sample. We also show their relationship with latitude (Supplementary Fig. 15-16).

Allele frequency change predicted by GEMs

We took 40 individuals approximately within 50 km of each other at three locations with the highest density of samples in our dataset: Madrid (Spain), Tübingen (Germany) and Malmö (Sweden) (Fig. 3C, Supplementary Fig. 8). We tested overall future allele frequency changes of all SNPs per population as well as SNP-specific allele frequency changes.

First, we calculated the mean allele frequency differences between future (rcp 8.5, 2070) and present predictions. This proved to be significant in most locations and models (Supplementary Table 11), although the direction of change was different between the two edge populations, Madrid and Malmö, and the Tübingen population from the center of the range. The former showed a decrease or a steady state in allele frequency, and the latter showed a highly significant increase in all models and

SNP subsets (Supplementary Table 11).

Second, we calculated the differences in frequency (*F*) between present (*pres*) and future (*2070*) populations per SNP and tested the difference using a Student's t test and a pooled standard error (*se*) of the frequency measurements: $t = \frac{F_{SNP\ 2070} - F_{SNP\ pres}}{\sqrt[2]{se^2_{SNP\ 2070} + se^2_{SNP\ pres}}}$. This not only revealed the main trend in frequency change, but also the distribution of differences in alleles (see histograms in Supplementary Fig. 13-14). It corroborated the general trend observed for all SNPs (Supplementary Table 11) and in addition showed that the global distribution of allele frequency changes in Tübingen is skewed to the right for some SNPs (increase of drought allele frequency).

Possible genetic trade offs of drought survival and flowering time

Contrary to our expectations, there were areas in the Mediterranean that were predicted to lose drought-survival alleles under climate change (Supplementary Fig. 9-10). These are areas that already today suffer from low precipitation (reached zero precipitation in summer, Supplementary Fig. 12) and will probably not become much drier in future summers. On the other hand, temperatures will keep increasing, which likely will demand an acceleration of flowering time (for which there is a trade-off with drought avoidance). Predictions at the phenotypic level (Supplementary Fig. 9-10) showed this trend: drought-survival will increase only in the transition areas between Mediterranean and more temperate regions (Supplementary Fig. 9) and might decrease in areas that are already dry (Supplementary Fig. 11). On the other hand, flowering time was predicted to decrease in the Mediterranean (Supplementary Fig. 10-11). We note that the SNP effects on drought survival and flowering time were positively correlated, as disclosed by Canonical Correlation Analyses (section 3.4).

Population genetics simulations

The prediction of an allele in 2070 does not directly inform about the actual possibility of adaptation. This depends on (**i**) the frequency of the alleles and haplotypes in the population, (**ii**) the recombination rate, and (**iii**) the strength and efficiency of selection. Indeed, geographic predictions of alleles are probably not good indicators of future allele frequency because random forest models tend to predict either presence or absence of alleles (and not co-existence, i.e. intermediate frequency). That is why we do not compare empirically obtained present allele frequencies with frequencies calculated from future predicted presence of alleles in the different locations of Tübingen, Madrid and Malmö.

To obtain further insights into population dynamics required for adaptation, we simulated

allele frequency changes in a Wright-Fisher population under a mutation-selection balance with inbreeding, as *Arabidopsis thaliana* is a predominant selfer (code available at http://github.com/MoisesExpositoAlonso/popgensim with DOI: https://doi.org/10.5281/zenodo.1039886). We started the simulations with the present frequencies of drought-related alleles of the 221 aGWA/GWA SNPs, with (codominant) selection coefficients (s) ranging from 0.01 to 20% fitness advantage. We considered SNPs as independent, that is, we did not include linkage disequilibrium information nor a recombination rate (see next section).

We carried out forward-in-time simulations for 50 generations, the approximate number of generations of natural populations of *A. thaliana* from today until 2070 at an average generation time of around 1.3-1.8 years[51]. We assumed a mutation rate (μ) calculated from laboratory mutation accumulation lines[47,52], although its effect in a few generations might be negligible. We used a selfing coefficient (ψ) of 98%, a conservative lower bound estimate from heterozygosity in individuals collected from nature[53,54]. The population size (N) was estimated from the genomic diversity in our dataset: The 40 genomes within the Tübingen area had a genome-wide nucleotide diversity of 0.004 (section 3.1). With the equation: $4Ne \times \mu = \pi$, we solved for effective population size (N_e) and transformed it into population size following the relationship[55]: $N_e = \frac{1}{1+F} \times N = \frac{2-\psi}{\psi} \times N$. This yielded N = 300,000 plants, which might be reasonable given that we consider an area of 50 km around Tübigen, and stands with hundreds or thousands of plants are not uncommon.

After running the simulations, we asked what selection coefficient would be needed to reach quasi-fixation frequencies of each allele (frequency >0.9) or to match the drought-allele frequencies in Madrid or Malmö (assuming that those populations are better adapted in comparison with Tübingen). When the frequency of a specific allele was higher in Tübingen than in Madrid or Malmö, we assumed selection would not be necessary and the coefficient was assumed to be zero. The results indicated that selection coefficients should be strong (but see ref. [56]) for alleles to become fixed (Supplementary Fig. 17). However, the distribution of selection coefficients was centered around 1-3% fitness advantage for Tübingen allele frequencies to match Malmö or Madrid (Supplementary Fig. 17) (but see next section 4.2.4.5). We did not simulate different degrees of drift in our analyses. We reasoned that when the inequality: $N_e s > 1$, holds, the weight of drift relative to selection is thought to be imperceptible[13].

Considerations regarding recombination

As stated above, assessing whether a population can adapt depends on the frequencies of drought-resistant individuals and drought-resistance alleles in the population, the rate of realized recombination, for crossing and reassortment of advantageous alleles, and the strength of selection.

Simulations that take LD between advantageous SNPs into account could inform about processes such as the Hill-Robertson effect, hitchhiking, or background selection[57].

In our simulations we did not work with haplotypes of SNPs in LD as they are found in individuals, but considered SNPs to be independent. This can be seen *a priori* as a strong assumption. Of relevance is that even in Tübingen there are already some individuals that have many of the 151 GWA drought-resistant alleles, with one exceptional individual having 123/151 drought-survival alleles. The three next best individuals have 107, 105, and 99 alleles. Fifteen of the 28 drought-resistance alleles are not present in the Tübingen population and will have to be imported by migration. Therefore, to produce a hypothetical "fully adapted haplotype" with 136 alleles from the current standing variation (123 alleles are 90% of all 136 present alleles), only 13 drought-resistant alleles would have to be recombined and introgressed into the already advantageous haplotype. Such introgression events might not be limited by low frequencies of the advantageous alleles, as some were found at intermediate or at as high as 90-95% frequency. Furthermore, in a scenario with a haplotype in the population with 123 alleles already present, simple individual-based simulations show that already with selection coefficients in the order of 0.5% advantage per allele, the 123 alleles haplotype will become completely fixed in the population within 50 generations. Results of aGWA indicate similar patterns of exceptional individuals with many drought alleles (24 alleles are 63% of all 32 present alleles) but more alleles are missing in the Tübingen population, as their frequency is lower and geographic distributions are narrower than for conventional GWA alleles.

We also used a series of approximate calculations to ascertain how many recombination events are required to generate a hypothetical "fully adapted haplotype". In *A. thaliana*, there are on average 1.4 meiotic crossovers for each of the five chromosomes[58]. Together with independent segregation of the five chromosomes, the parental haplotypes are rearranged at around 12 positions. A population of ~300,000 individuals (*N*) with a lower bound outcrossing rate of 1% (=1-F) over 50 generations (g), could thus undergo around ~2 million recombination events, $N \times (1 - F) \times r \times g$, or about 1 event per 50 bp. Note that this could be a conservative estimate as outcrossing events amount up to 14% of reproduction events in geographically close plants (ref. [54]). This result suggests that recombination might be less limiting than expected *a priori*.

GEM limitations

There is a long list of factors that we did not take into account and that will influence future plant response to climate change. We briefly enumerate them here:

A. We only focus on adaptation to drought, but other environmental stresses could have similar

detrimental effects such as extremely high temperatures or ecosystem destruction. In addition, fluctuation in selection gradients and seasonal environmental variation are other possible consequences of climate change[59,60]. The distribution of pathogens will almost certainly change under climate change as well, and biotic interactions can also play a relevant role in population dynamics, which we ignored[61–63].

B. We can only explain ca. 50% of the drought survival variance with 221 SNPs.

C. We evaluated drought survival in a controlled greenhouse experiment, but the extrapolation to natural conditions may be difficult. This would require field experiments assessing fitness *in situ* to confirm that the identified SNPs actually report a fitness advantage[64].

D. Because the high narrow sense heritability suggests a mostly-additive genetic architecture, we carried out predictions with allele counts. However, we acknowledge that there is variation of the magnitude of the SNP effects (Supplementary Table 3-4), and epistatic effects might exist.

E. Although long-term evolution should be driven by genetic adaptation, it is expected that phenotypic plasticity will partially buffer the detrimental consequences of environmental change[65].

F. The existence of a seed bank in *A. thaliana[51,66]* would cause longer generation times and overlapping generations, and both alter the speed and dynamics of allele frequencies[67].

G. Although our rough calculations suggest that recombination would not be a limitation for future adaptation in *A. thaliana* populations, we have not incorporated such processes in our modeling, as doing so is not a trivial matter[57,68]. This ignores phenomena such as background selection or hitchhiking effects that could arise from phenotypic trade-offs and the currently realized composition of haplotypes in the population.

H. Finally, on the positive side, our predictions are based on existing diversity, but *de novo* mutations are likely to make a contribution as well, especially in species with high reproduction rate, short generation time, and large population sizes[47,69,70]. In this sense our approach would be rather conservative.

SUPPLEMENTARY REFERENCES

1. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell. Elsevier; 2016;166: 481–491. doi:10.1016/j.cell.2016.05.063

2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

3. Verslues PE, Agarwal M, Katiyar-Agarwal S, Zhu J, Zhu J-K. Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. Plant J.

2006;45: 523–539. doi:10.1111/j.1365-313X.2005.02593.x

4. Jones HG. Monitoring plant and soil water status: established and novel methods revisited and their relevance to studies of drought tolerance. J Exp Bot. 2007;58: 119–130. doi:10.1093/jxb/erl118

5. Rymaszewski W, Vile D, Bédiée A, Dauzat M, Luchaire N, Kamrowska D, et al. Stress-response gene expression reflects morpho-physiological responses to water deficit. Plant Physiol. 2017; doi:10.1104/pp.17.00318

6. Granier C, Aguirrezabal L, Chenu K, Cookson SJ, Dauzat M, Hamard P, et al. PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit. New Phytol. 2005;169: 623–635. doi:10.1111/j.1469-8137.2005.01609.x

7. Exposito-Alonso M, Rodríguez RG, Barragán C, Capovilla G, Chae E, Devos J, et al. A rainfall-manipulation experiment with 517 Arabidopsis thaliana accessions [Internet]. bioRxiv. 2017. doi:10.1101/186767

8. Vasseur F, Wang G, Bresson J, Schwab R, Weigel D. Image-based methods for phenotyping growth dynamics and fitness in large plant populations [Internet]. bioRxiv. 2017. p. 208512. doi:10.1101/208512

9. Bradski G. The opencv library. Doctor Dobbs Journal. M AND T PUBLISHING INC; 2000;25: 120–126. Available: http://elibrary.ru/item.asp?id=4934581

10. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw. 2010;33: 1–22. Available: http://mirror.dcc.online.pt/CRAN/web/packages/MCMCglmm/vignettes/Overview.pdf

11. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

12. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. Public Library of Science; 2012;8: e1002967. doi:10.1371/journal.pgen.1002967

13. Charlesworth B, Charlesworth D. Elements of Evolutionary Genetics [Internet]. Roberts and Company Publishers; 2010. Available: https://books.google.de/books?id=dgNFAQAAIAAJ

14. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. Nature Publishing Group; 2014;46: 919–925. doi:10.1038/ng.3015

15. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42: 348–354. doi:10.1038/ng.548

16. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465: 627–631. doi:10.1038/nature08800

17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995; 289–300.

18. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013;9: e1003264. doi:10.1371/journal.pgen.1003264

19. Mitchell-Olds T. Pleiotropy causes long-term genetic constraints on life-history evolution in Brassica rapa. Evolution. [Society for the Study of Evolution, Wiley]; 1996;50: 1849–1858. doi:10.2307/2410742

20. González I, Déjean S, Martin P, Baccini A. CCA: An R Package to Extend Canonical Correlation Analysis. Journal of Statistical Software, Articles. 2008;23: 1–14. doi:10.18637/jss.v023.i12

21. Porcher E, Giraud T, Goldringer I, Lavigne C. Experimental demonstration of a causal relationship between heterogeneity of selection and genetic differentiation in quantitative traits. Evolution. Wiley Online Library; 2004;58: 1434–1445. Available: http://www.ncbi.nlm.nih.gov/pubmed/15341147

22. Leinonen T, Scott McCairns RJ, O'Hara RB, Merilä J. QST–FST comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat Rev Genet. Nature Publishing Group; 2013;14: 179–190. doi:10.1038/nrg3395

23. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

24. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10: e1004412–e1004412. doi:10.1371/journal.pgen.1004412

25. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8: e1002453. doi:10.1371/journal.pgen.1002453

26. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98: 116–126. doi:10.1016/j.ajhg.2015.11.020

27. Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, et al. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat Genet. 2013;45: 1327–1336. doi:10.1038/ng.2766

28. Josephs EB, Stinchcombe JR, Wright SI. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? New Phytol. 2017; doi:10.1111/nph.14410

29. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. Nature Research; 2012;484: 55–61. doi:10.1038/nature10944

30. Shriner D, Adeyemo A, Ramos E, Chen G, Rotimi CN. Mapping of disease-associated variants in admixed populations. Genome Biol. 2011;12: 223. doi:10.1186/gb-2011-12-5-223

31. Tesfaye M, Silverstein KAT, Nallu S, Wang L, Botanga CJ, Karen Gomez S, et al. Spatio-Temporal Expression Patterns of Arabidopsis thaliana and Medicago truncatula Defensin-Like Genes. PLoS One. Public Library of Science; 2013;8: e58992. doi:10.1371/journal.pone.0058992

32. Dannemann M, Andrés AM, Kelso J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. Am J Hum Genet. 2016;98: 22–33. doi:10.1016/j.ajhg.2015.11.015

33. Rife DC. Populations of hybrid origin as source material for the detection of linkage. Am J Hum

Genet. ncbi.nlm.nih.gov; 1954;6: 26–33. Available: http://www.ncbi.nlm.nih.gov/pubmed/13138567

34. Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, et al. Interrogating local population structure for fine mapping in genome-wide association studies. Bioinformatics. 2010;26: 2961–2968. doi:10.1093/bioinformatics/btq560

35. Li J-Y, Fu Y-L, Pike SM, Bao J, Tian W, Zhang Y, et al. The Arabidopsis nitrate transporter NRT1.8 functions in nitrate removal from the xylem sap and mediates cadmium tolerance. Plant Cell. 2010;22: 1633–1646. doi:10.1105/tpc.110.075242

36. Gojon A, Gaymard F. Keeping nitrate in the roots: an unexpected requirement for cadmium tolerance in plants. J Mol Cell Biol. 2010;2: 299–301. doi:10.1093/jmcb/mjq019

37. Jarzyniak KM, Jasiński M. Membrane transporters and drought resistance - a complex issue. Front Plant Sci. 2014;5: 687. doi:10.3389/fpls.2014.00687

38. Lelandais-Brière C, Jovanovic M, Torres GAM, Perrin Y, Lemoine R, Corre-Menguy F, et al. Disruption of AtOCT1, an organic cation transporter gene, affects root development and carnitine-related responses in Arabidopsis. Plant J. 2007;51: 154–164. doi:10.1111/j.1365-313X.2007.03131.x

39. Ma S, Gong Q, Bohnert HJ. Dissecting salt stress pathways. J Exp Bot. 2006;57: 1097–1107. doi:10.1093/jxb/erj098

40. Apse MP, Sottosanto JB, Blumwald E. Vacuolar cation/H+ exchange, ion homeostasis, and leaf development are altered in a T-DNA insertional mutant of AtNHX1, the Arabidopsis vacuolar Na+/H+ antiporter. Plant J. 2003;36: 229–239. doi:10.1046/j.1365-313X.2003.01871.x

41. Lovell JT, Juenger TE, Michaels SD, Lasky JR, Platt A, Richards JH, et al. Pleiotropy of FRIGIDA enhances the potential for multivariate adaptation. Proceedings of the Royal Society of London B: Biological Sciences. The Royal Society; 2013;280: 20131043. doi:10.1098/rspb.2013.1043

42. Bac-Molenaar JA, Granier C, Keurentjes JJB, Vreugdenhil D. Genome-wide association mapping of time-dependent growth responses to moderate drought stress in Arabidopsis. Plant Cell Environ. Wiley Online Library; 2016;39: 88–102. doi:10.1111/pce.12595

43. Vasseur F, Bontpart T, Dauzat M, Granier C, Vile D. Multivariate genetic analysis of plant responses to water deficit and high temperature revealed contrasting adaptive strategies. J Exp Bot. 2014;65: 6457–6469. doi:10.1093/jxb/eru364

44. Moss RH, Edmonds JA, Hibbard KA, Manning MR, Rose SK, van Vuuren DP, et al. The next generation of scenarios for climate change research and assessment. Nature. 2010;463: 747–756. doi:10.1038/nature08823

45. Guiot J, Cramer W. Climate change: The 2015 Paris Agreement thresholds and Mediterranean basin ecosystems. Science. 2016;354: 465–468. doi:10.1126/science.aah5015

46. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. John Wiley & Sons, Ltd.; 2005;25: 1965–1978. doi:10.1002/joc.1276

47. Exposito-Alonso M, Becker C, Schuenemann VJ, Reitter E, Setzer C, Slovak R, et al. The rate and effect of de novo mutations in natural populations of Arabidopsis thaliana [Internet]. bioRxiv.

2016. p. 050203. doi:10.1101/050203

48. Breiman L. Random Forests. Mach Learn. Kluwer Academic Publishers; 2001;45: 5–32. doi:10.1023/A:1010933404324

49. Liaw A, Wiener M. Classification and Regression by randomForest [Internet]. R News. 2002. pp. 18–22. Available: http://CRAN.R-project.org/doc/Rnews/

50. Thuiller W, Lavorel S, Araújo MB, Sykes MT, Prentice IC. Climate change threats to plant diversity in Europe. Proc Natl Acad Sci U S A. National Acad Sciences; 2005;102: 8245–8250. doi:10.1073/pnas.0409902102

51. Falahati-Anbaran M, Lundemo S, Stenøien HK. Seed dispersal in time can counteract the effect of gene flow between natural populations of Arabidopsis thaliana. New Phytol. 2014;202: 1043–1054. doi:10.1111/nph.12702

52. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010;327: 92–94. doi:10.1126/science.1180677

53. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The scale of population structure in Arabidopsis thaliana. PLoS Genet. 2010;6: e1000843. doi:10.1371/journal.pgen.1000843

54. Bomblies K, Yant L, Laitinen R a., Kim S-T, Hollister JD, Warthmann N, et al. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet. 2010;6: e1000890–e1000890. doi:10.1371/journal.pgen.1000890

55. Nordborg M, Donelly P. The Coalescent Process with selfing. Genetics. 1997;146: 1185–1195.

56. Roux F, Giancola S, Durand S, Reboud X. Building of an experimental cline with Arabidopsis thaliana to estimate herbicide fitness cost. Genetics. 2006;173: 1023–1031. doi:10.1534/genetics.104.036541

57. Haller BC, Messer PW. SLiM 2: Flexible, Interactive Forward Genetic Simulations. Mol Biol Evol. 2017;34: 230–240. doi:10.1093/molbev/msw211

58. Salomé PA, Bomblies K, Fitz J, Laitinen RAE, Warthmann N, Yant L, et al. The recombination landscape in Arabidopsis thaliana F2 populations. Heredity . 2012;108: 447–455. doi:10.1038/hdy.2011.95

59. Fournier-Level A, Perry EO, Wang JA, Braun PT, Migneault A, Cooper MD, et al. Predicting the evolutionary dynamics of seasonal adaptation to novel climates in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2016;113: E2812–21. doi:10.1073/pnas.1517456113

60. Chevin L-M, Visser ME, Tufto J. Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. Evolution. 2015;69: 2319–2332. doi:10.1111/evo.12741

61. Karasov TL, Kniskern JM, Gao L, DeYoung BJ, Ding J, Dubiella U, et al. The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature. Nature Publishing Group; 2014;512: 436–440. doi:10.1038/nature13439

62. Jakob K, Goss EM, Araki H, Van T, Kreitman M, Bergelson J. Pseudomonas viridiflava and P. syringae—natural pathogens of Arabidopsis thaliana. Mol Plant Microbe Interact. 2002;15:

1195–1203. doi:10.1094/MPMI.2002.15.12.1195

63. Araújo MB, Luoto M. The importance of biotic interactions for modelling species distributions under climate change. Glob Ecol Biogeogr. Wiley Online Library; 2007;16: 743–753. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1466-8238.2007.00359.x/full

64. de Villemereuil P, Gaggiotti OE, Mouterde M, Till-Bottraud I. Common garden experiments in the genomic era: new perspectives and opportunities. Heredity. nature.com; 2016;116: 249–254. doi:10.1038/hdy.2015.93

65. Chevin LM, Lande R. When do adaptive plasticity and genetic evolution prevent extinction of a density-regulated population? Evolution. 2010;64: 1143–1150. doi:10.1111/j.1558-5646.2009.00875.x

66. Lundemo S, Falahati-Anbaran M, Stenøien HK. Seed banks cause elevated generation times and effective population sizes of Arabidopsis thaliana in northern Europe. Mol Ecol. 2009;18: 2798–2811. doi:10.1111/j.1365-294X.2009.04236.x

67. Charlesworth B. Selection in populations with overlapping generations. VI. Rates of change of gene frequency and population growth rate. Theor Popul Biol. Elsevier; 1974;6: 108–133. Available: https://www.ncbi.nlm.nih.gov/pubmed/4418568

68. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. Genetics Soc America; 2003;165: 2213–2233. Available: http://www.ncbi.nlm.nih.gov/pubmed/14704198

69. Colautti RI, Lau JA. Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. Mol Ecol. 2015;24: 1999–2017. doi:10.1111/mec.13162

70. Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. Mol Ecol. 2015;24: 2095–2111. doi:10.1111/mec.13183

71. Berg JJ, Coop G. A coalescent model for a sweep of a unique standing variant. Genetics. 2015;201: 707–725. doi:10.1534/genetics.115.178962

SUPPLEMENTARY TABLES

Supplementary tables are available as a .xlsx file at: https://static-content.springer.com/esm/art%3A10.1038%2Fs41559-017-0423-0/MediaObjects/41559_2017_423_MOESM3_ESM.xlsx

Supplementary Table 1. Accession information.
1001 Genomes IDs, common names, countries of origin, and geographical and environmental information.

Supplementary Table 2. ADMIXTURE cross-validation for K= 2 to 20.

Supplementary Table 3. Diversity statistics and annotations of top GWA hits.

Supplementary Table 4. Annotation of top aGWA SNP hits.

Supplementary Table 5. Information on phenotypic and climate traits.

Supplementary Table 6. Climate and phenotype correlations (per accession).
Pearson product-moment correlation coefficients between all phenotype and climate variables of Supplementary Table 5. Lower triangle shows p-values, upper triangle correlation coefficients. The drought index parameter of choice (m1d_polqua) negatively correlates with the precipitation in the driest month and quarter, bio14 and bio18, respectively.

Supplementary Table 7. Climate and phenotype correlations (per SNP).
Pearson product-moment correlation coefficients of GWA effects of a large subset of all phenotype and climate variables of Supplementary Table 5. Correlations are done only with the top 151 SNPs identified in the drought survival GWA and tested for polygenic selection.

Supplementary Table 8. Canonical Correlation Analysis (CCA).
CCA between GWA effects on different phenotypes and the SNP associations with climate variables.

Supplementary Table 9. Polygenic model for different top SNP groups.
We applied the Berg & Coop model[71] of polygenic adaptation to different groups of top SNPs and report the value of Q_x statistics.

Supplementary Table 10. Importance of variables in Random Forest analyses.

For each random forest model, the importance of bioclimatic variables is reported. For classification random forest, importance is reported as the mean decreased accuracy (MDA) and for regression random forest, importance is reported as the mean square error (MSE). MDA is the number of misclassified observations when removing a variable and MSE is the increase of mean square error produced by removing a variable.

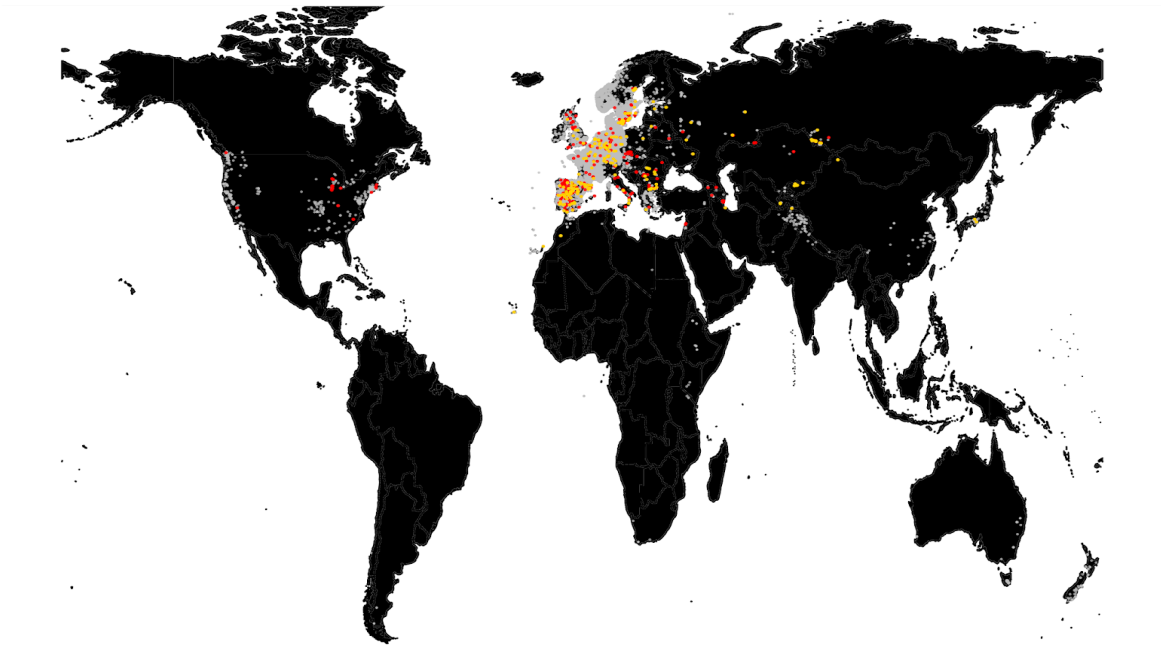
Supplementary Table 11. Allele frequency change.

Student's t-test results of allele frequency changes in the locations of Madrid, Tübingen and Malmö under the three forecasting Genome Environment Models: free migration, principal components control, and geography control.

Supplementary Table 12. Q_{st} F_{st} pairwise comparisons.

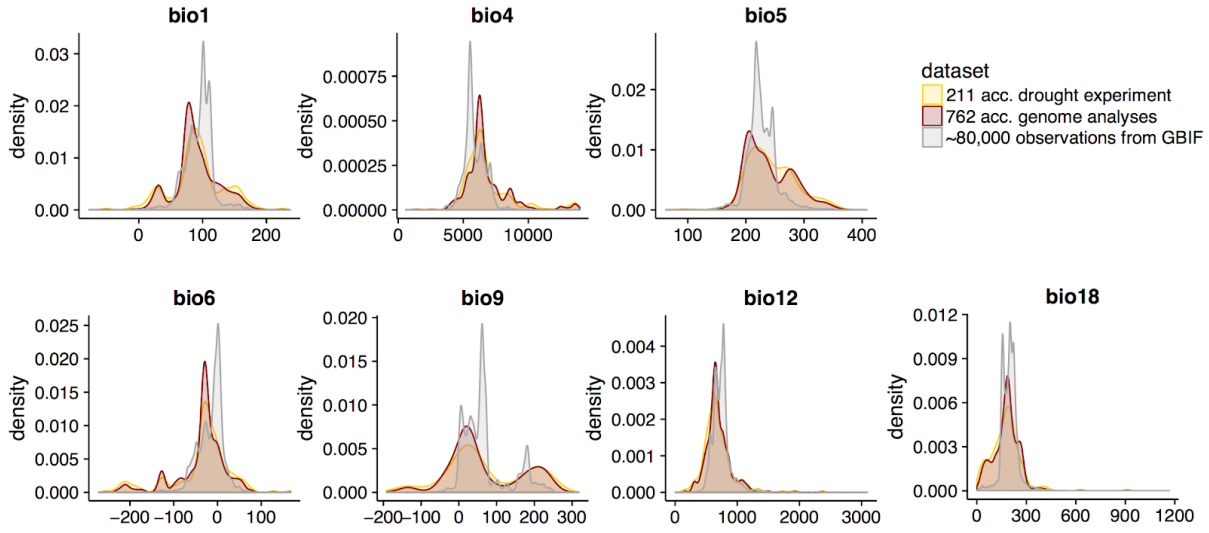
Mean Q_{st} (95% Highest Posterior Density) and median F_{st} (95% percentile) for all pairs of genetic groups.

SUPPLEMENTARY FIGURES



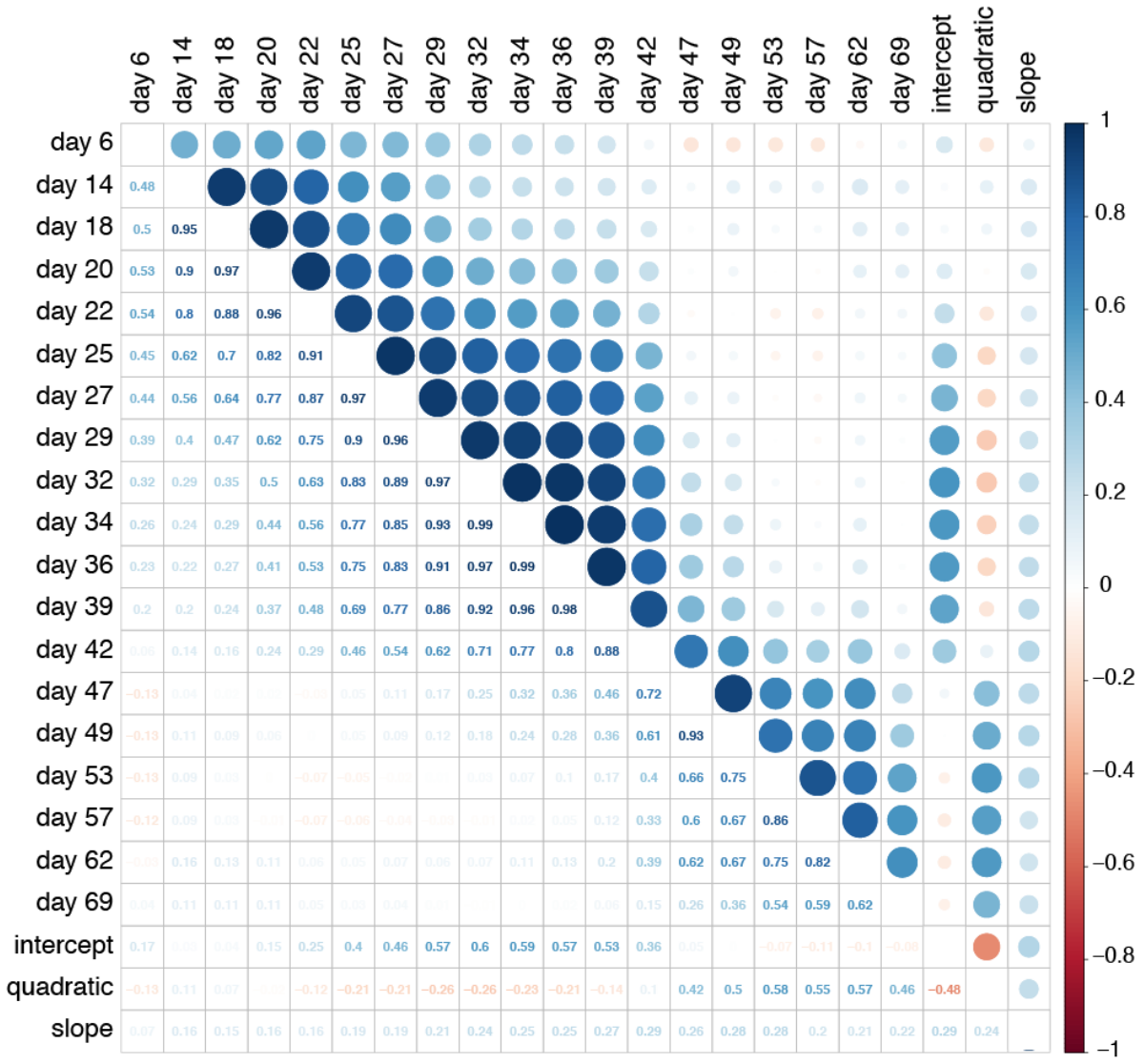
Supplementary Figure 1. Global *Arabidopsis thaliana* distribution.

The world map shows ca. 80,000 records from the Global Biodiversity Information Facility (GBIF, www.gbif.org) (grey), the 762 *A. thaliana* accessions used for genetic analyses (red), and the 211 accessions used for phenotyping experiments (yellow).



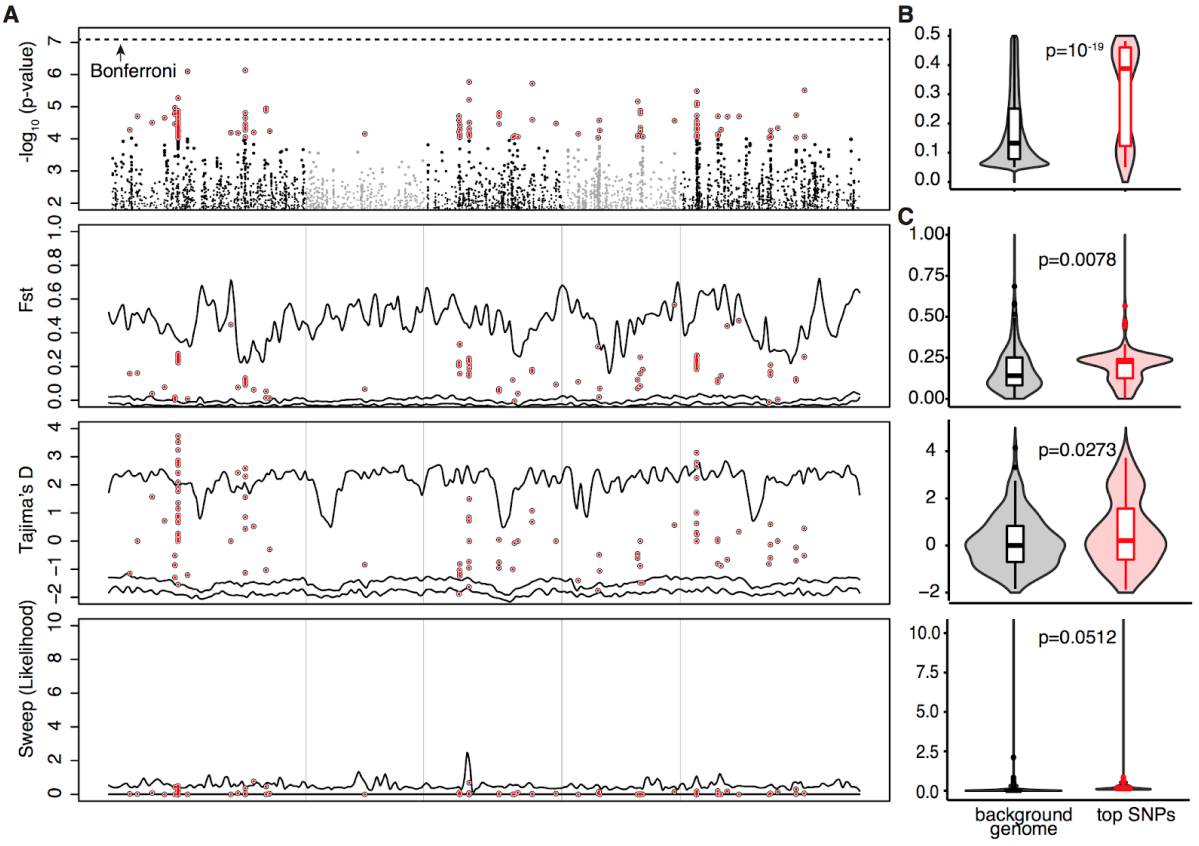
Supplementary Figure 2. Environmental ranges of *Arabidopsis thaliana*.

The range in key environmental variables for the three datasets in Supplementary Fig. 1 is shown. The set of accessions used in our analyses not only covered the range of the species as estimated from GBIF data, but also revealed that these accessions have a more even distribution throughout the environmental ranges. The bioclimatic variables are: annual precipitation (bio12), precipitation of the warmest quarter (bio18), annual mean temperature (bio1), temperature seasonality (bio4), maximum temperature of the warmest month (bio5), minimum temperature of the coldest month (bio6), and mean temperature of the driest quarter (bio9). See Supplementary Table 5 for more information.



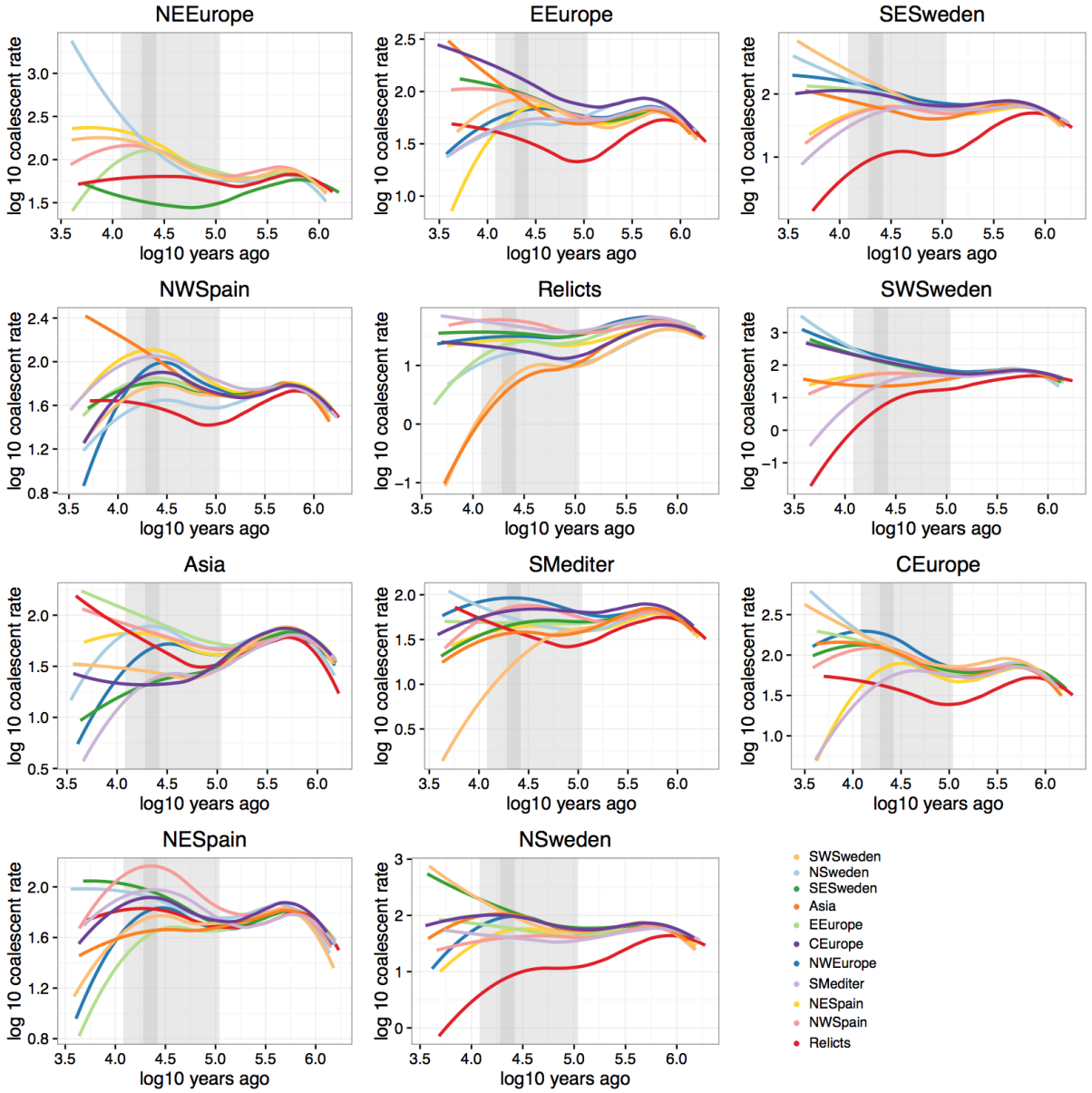
Supplementary Figure 3. Correlation between rosette areas and model parameters.

Pearson product-moment correlation coefficients between the three drought-index parameters and the 'raw' number of green pixels per pot and per day photographed. Sizes of circles indicate strength and colors arithmetic signs of association, shown as numbers in the lower triangle.

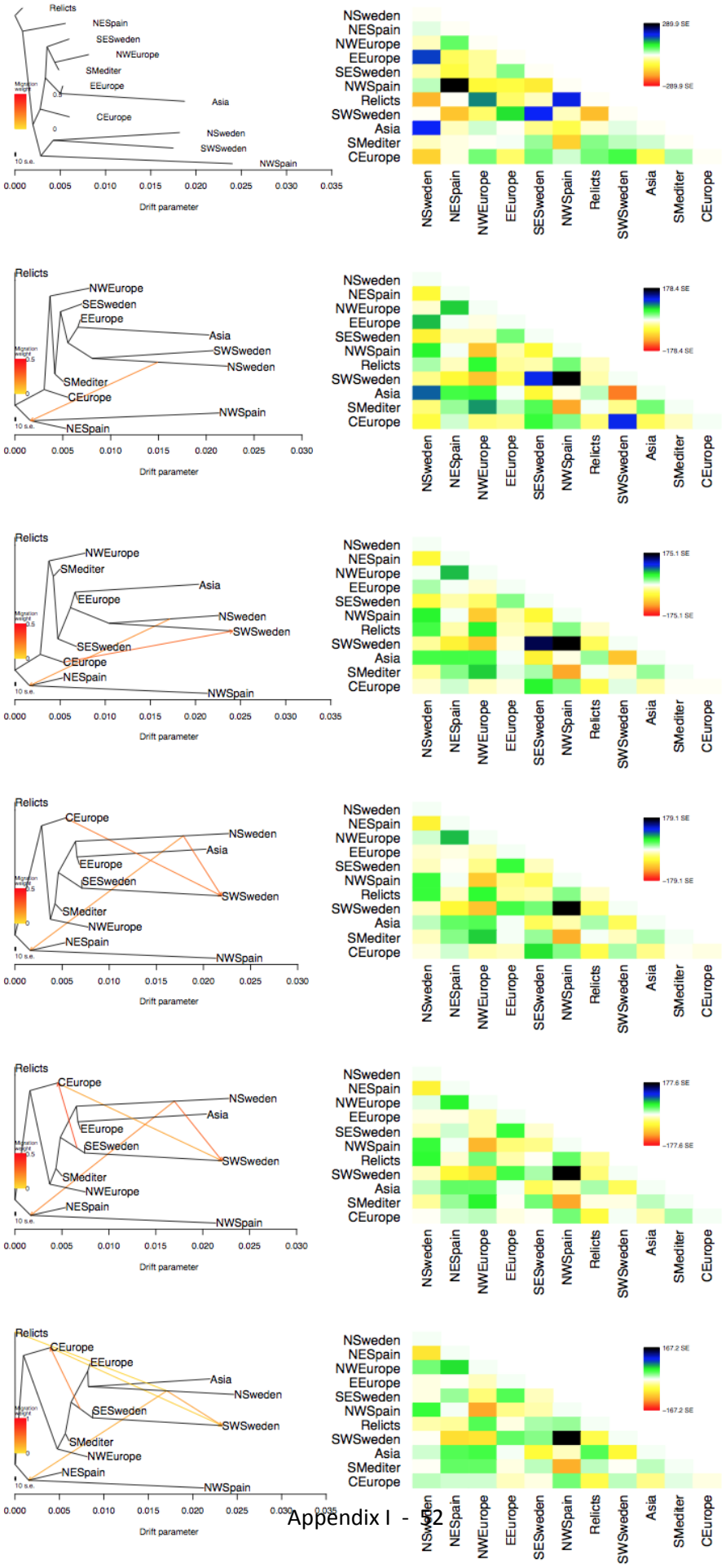


Supplementary Figure 4. GWA with drought survival and population genetic statistics.

(**A**) Manhattan plot of drought survival index GWA, F_{st}, Tajima's D, and selective sweeps. (**B**) Violin and box plots of allele frequency, and (**C**) F_{st}, Tajima's D, and selective sweeps of the top 150 SNPs (red) vs frequency-matched 150 SNPs from a random genome background (grey). GWA was calculated using EMMAX. F_{st} across populations (see Fig. 1) and Tajima's D were calculated using PLINK. Sweep likelihood was calculated using SWEED software. Median p-values from Wilcoxon tests with 100 bootstrap replicates are indicated.

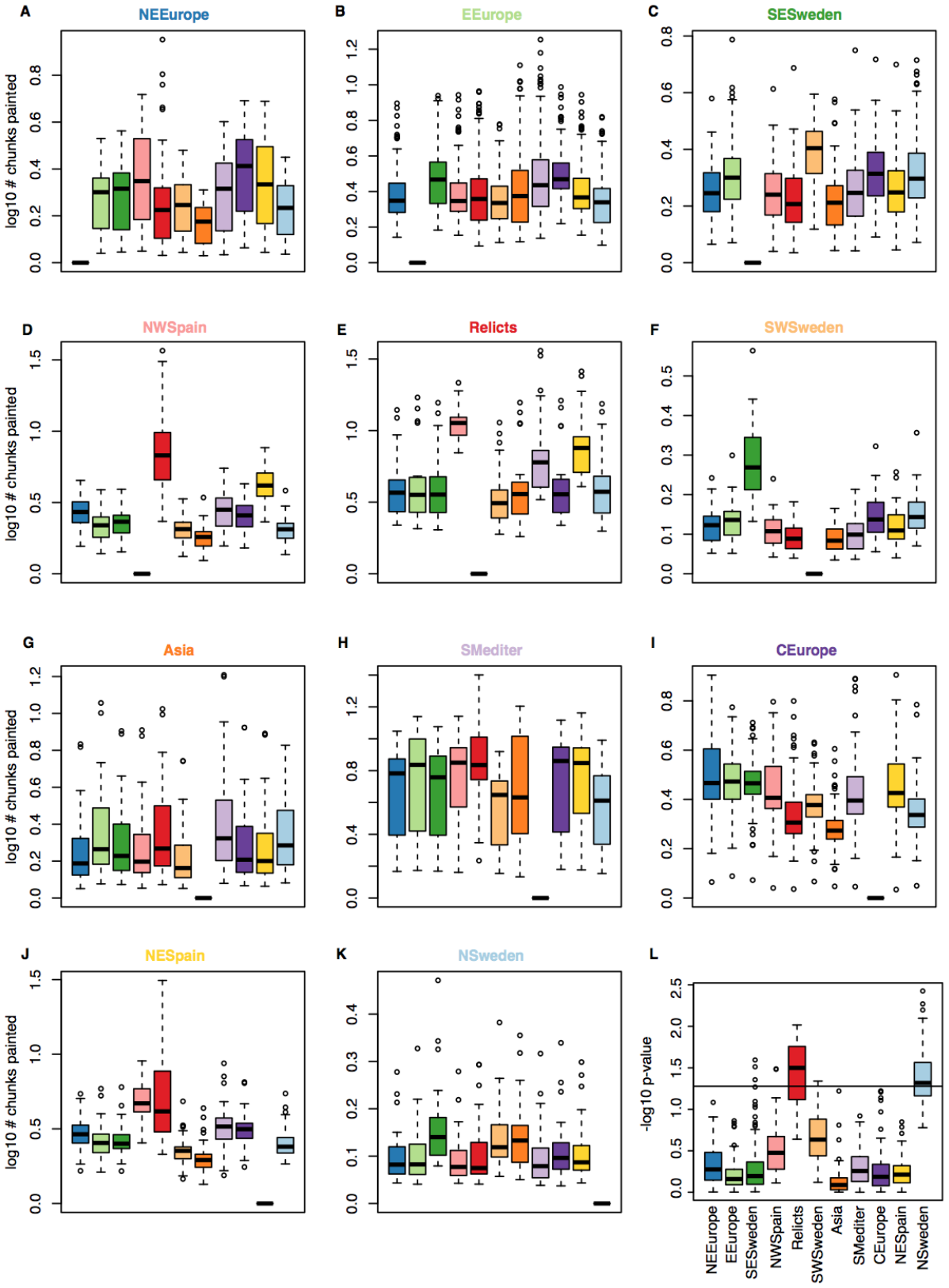


Supplementary Figure 5. Cross-coalescent rates between populations inferred by MSMC. Joint coalescent rates of each of the 11 ADMIXTURE genetic groups are (see Fig. 1 and Supplementary Fig. 12) compared to the other groups. Each line is a smoothed loess of 6 replicated runs. Light grey indicates the extent of the last glacial maxima (100-10 kya) and dark grey the peak of the last glaciation (22 kya). Note that the N. Swedish group is the first to separate from the relicts.


Relicts
NESpain
SESweden
NWEurope
SMediter
EEurope
Asia
CEurope
NSweden
SWSweden
NWSpain
Migration weight
10 s.e.
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035
Drift parameter
NSweden NESpain NWEurope EEurope SESweden NWSpain Relicts SWSweden Asia SMediter CEurope
289.9 SE
−289.9 SE
178.4 SE
−178.4 SE
175.1 SE
−175.1 SE
179.1 SE
−179.1 SE
177.6 SE
−177.6 SE
167.2 SE
−167.2 SE

Supplementary Figure 6. Treemix with different migration rates.

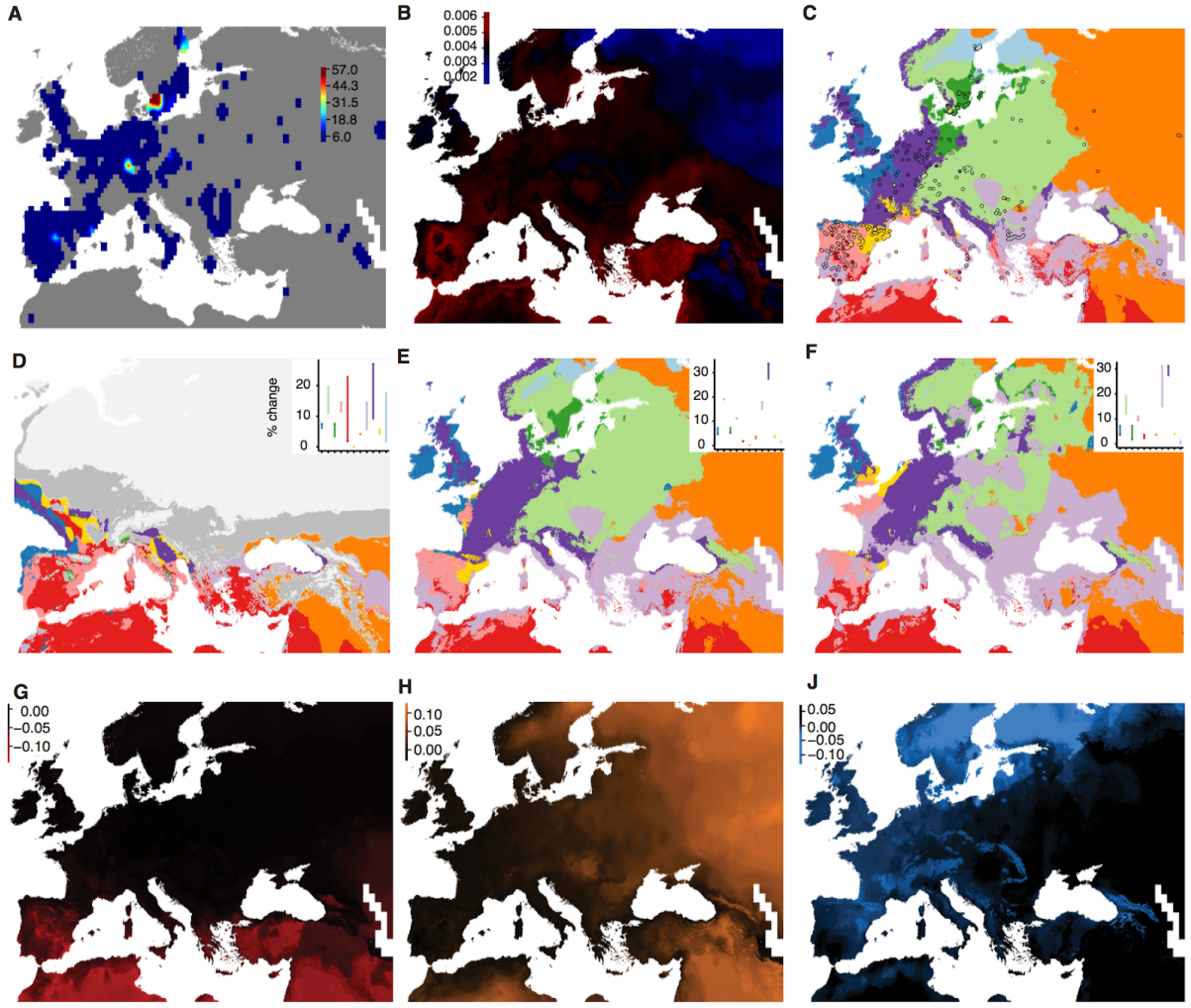
Maximum likelihood (ML) population trees from Treemix (left). Analyses with zero to five migration edges are presented. Heatmaps with the residual fit of the ML trees are shown on the right. Note that the unexpected closeness of NW. Spain and Sweden without migration is resolved with one migration edge. With this, a more parsimonious tree that adheres to geographic locations is uncovered.



Supplementary Figure 7. Genomic ChromoPainter chunks per population.

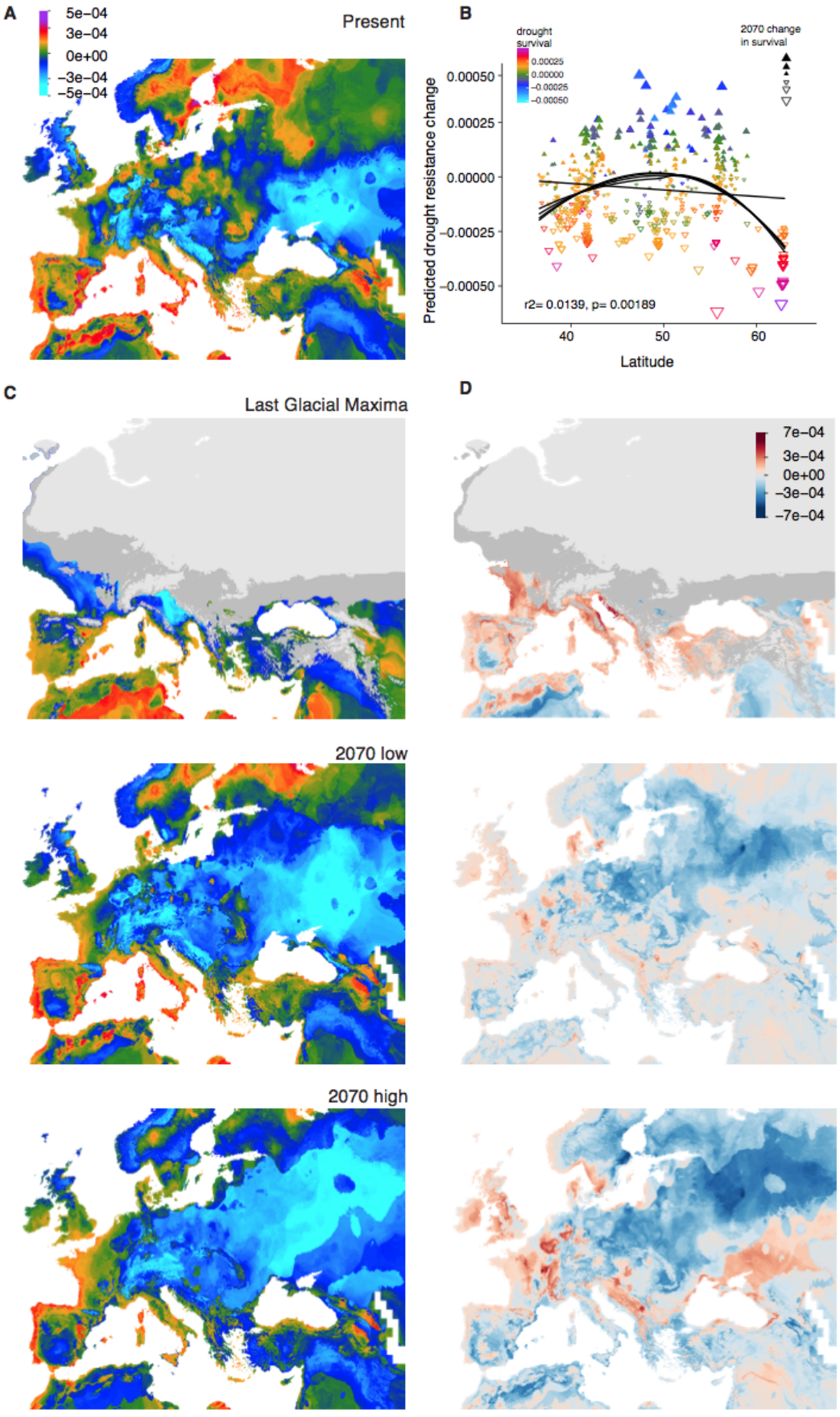
(**A-K**) Summary of the number of ChromoPainter chunks inherited from other genomes that had

been assigned to ADMIXTURE groups. Each graph summarizes the information of all the genomes from an admixture group. (**L**) The p-value of the Pearson correlation test between an accession's drought survival index and the number of chunks received from another genome. The p-value distributions of genomes from the same ADMIXTURE group are grouped in a box plot. Intuitively this can be interpreted as how well the number of chunks inherited from a specific donor predicts the drought survival of the receiver. The black line indicates the 5% significance threshold, which is passed by most relict and N. Swedish groups. Therefore, chunks that have N. Swedish and/or relict ancestry explain the drought survival of other individuals well.

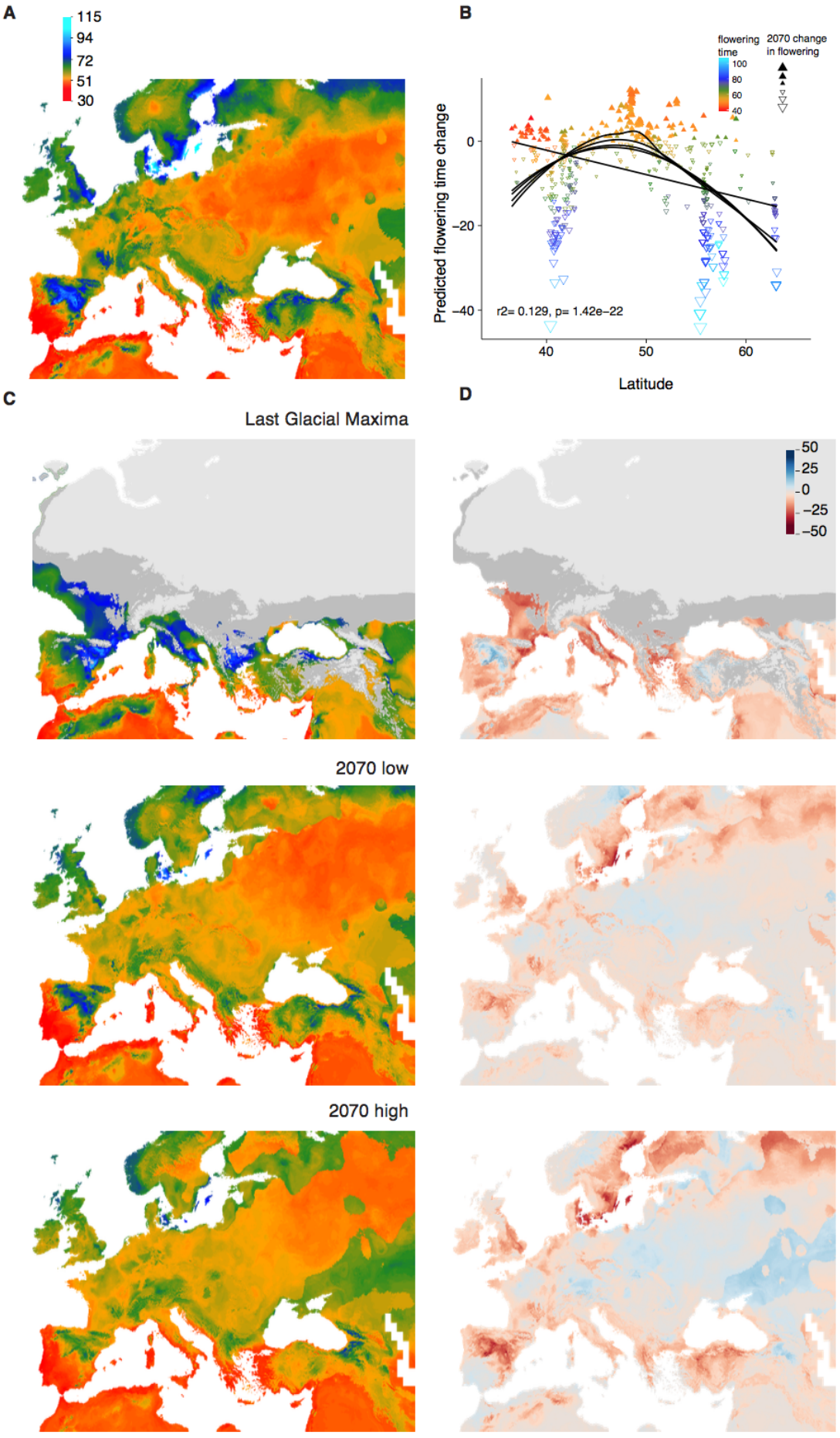


Supplementary Figure 8. Environmental niche model (ENM) of population structure. (**A**) Distribution of 762 accessions from the 1001 Genomes project used for environmental niche modeling of genetic diversity and analysis of population structure. Colors indicate number of accessions within a 1°x1° latitude x longitude grid. (**B**) Random forest environment niche models using estimates of pairwise nucleotide diversity (π) of each accession with its closest 10 geographic neighbours. The trained model was used to predict diversity based on environmental data. (**C**) Random forest environment model of the 11 genetic groups (see Fig. 1). Locations with accessions are shown as points filled with the actual genetic group assigned, and are used for model training as in (B). The trained model was used to predict a raster of environmental variables and is shown in the background. When the circle is filled with the same color as the background, the model succeeds in the prediction. The trained models were also used to predict the change in overall area covered by each genetic group from present to the last glacial maxima (**D**) and for 2070 under low (**E**) and high (**F**) CO_2 concentration scenarios (panels in the upper-right corner) (see Fig. 1 for color keys). (**G-J**) The

first three genome-wide principal components from Fig. 2 were modeled based on environmental variables. Later, these were used as covariates of GEMs (Supplementary Fig. 13-14).

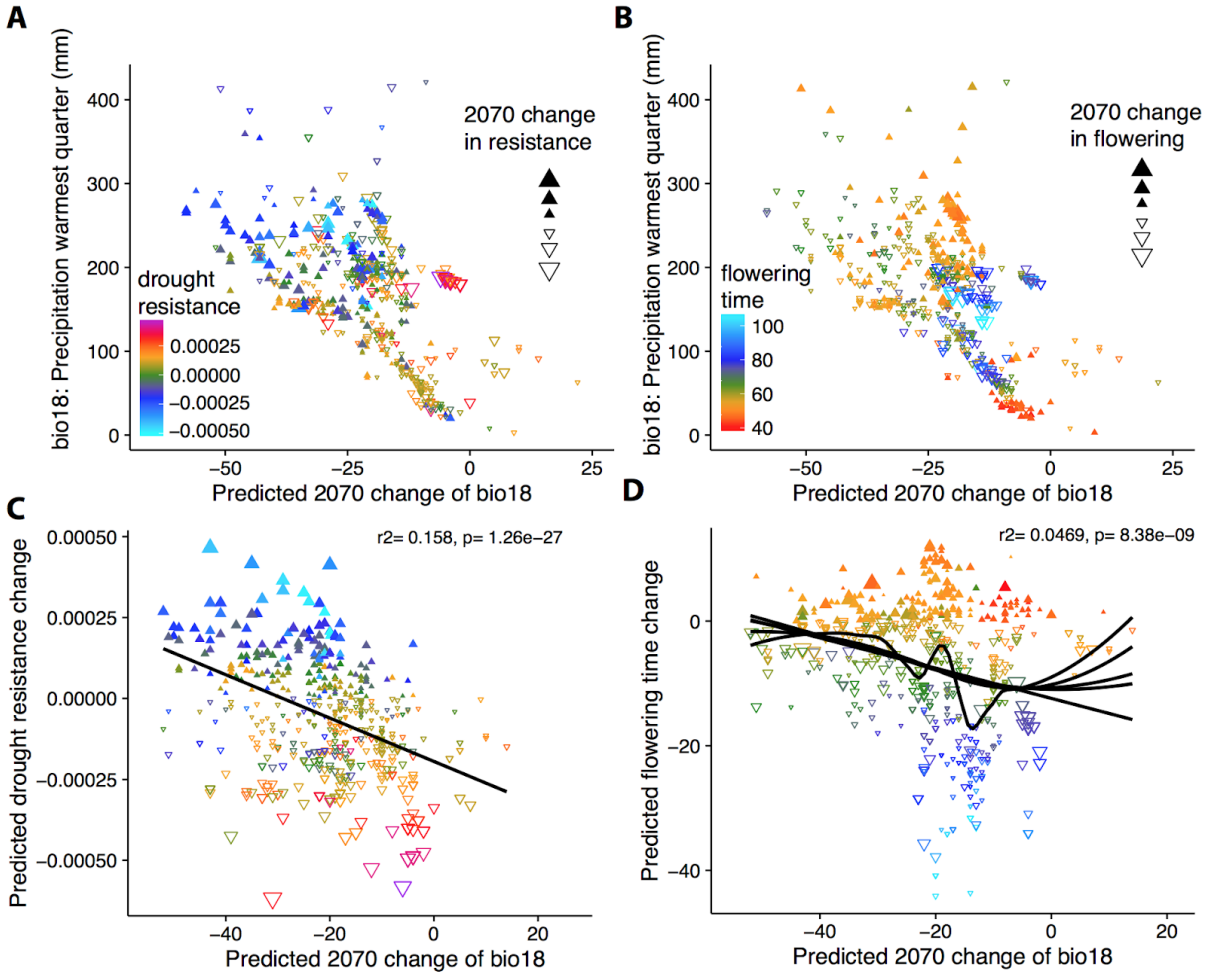

A
5e−04
3e−04
0e+00
−3e−04
−5e−04
Present
B
drought survival
0.00025
0.00000
−0.00025
−0.00050
2070 change in survival
Predicted drought resistance change
0.00050
0.00025
0.00000
−0.00025
−0.00050
r2= 0.0139, p= 0.00189
40
50
60
Latitude
C
Last Glacial Maxima
2070 low
2070 high
D
7e−04
3e−04
0e+00
−3e−04
−7e−04

Supplementary Figure 9. Environmental niche model (ENM) of drought survival index.
(**A**) Present geographic prediction of drought survival index from a random forest ENM trained on experimentally determined phenotypes for 211 accessions. Note that the highest drought survival indices are inferred for the Mediterranean as well as N. Sweden. (**B**) Correlation of phenotypic change in 2070 under a high CO_2 scenario with latitude; colors indicate present drought survival values, lines indicate linear (r^2 and p value in figure) or loess models. (**C**) The trained model is also used to predict drought survival index under the last glacial maximum, and for two 2070 scenarios of low and high CO_2 concentrations. (**D**) For the three scenarios, the change is shown relative to the current date prediction for easier comparison.


A
115
94
72
51
30
B
flowering time
2070 change in flowering
100
80
60
40
Predicted flowering time change
0
−20
−40
r2= 0.129, p= 1.42e−22
40
50
60
Latitude
C
Last Glacial Maxima
2070 low
2070 high
D
50
25
0
−25
−50

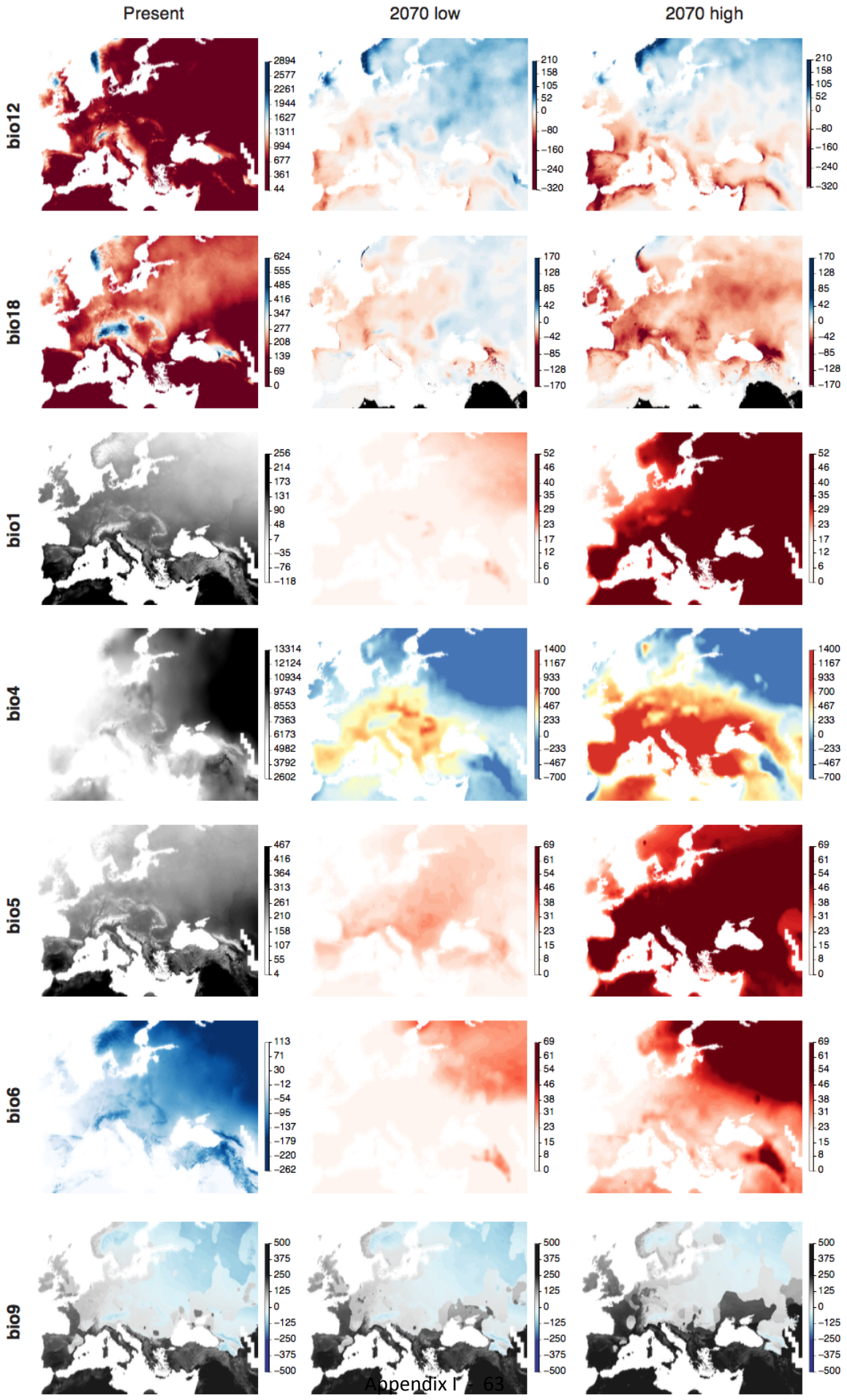
Supplementary Figure 10. Environmental niche model (ENM) of flowering time.

Same models as in Supplementary Fig. 9, but for flowering time.



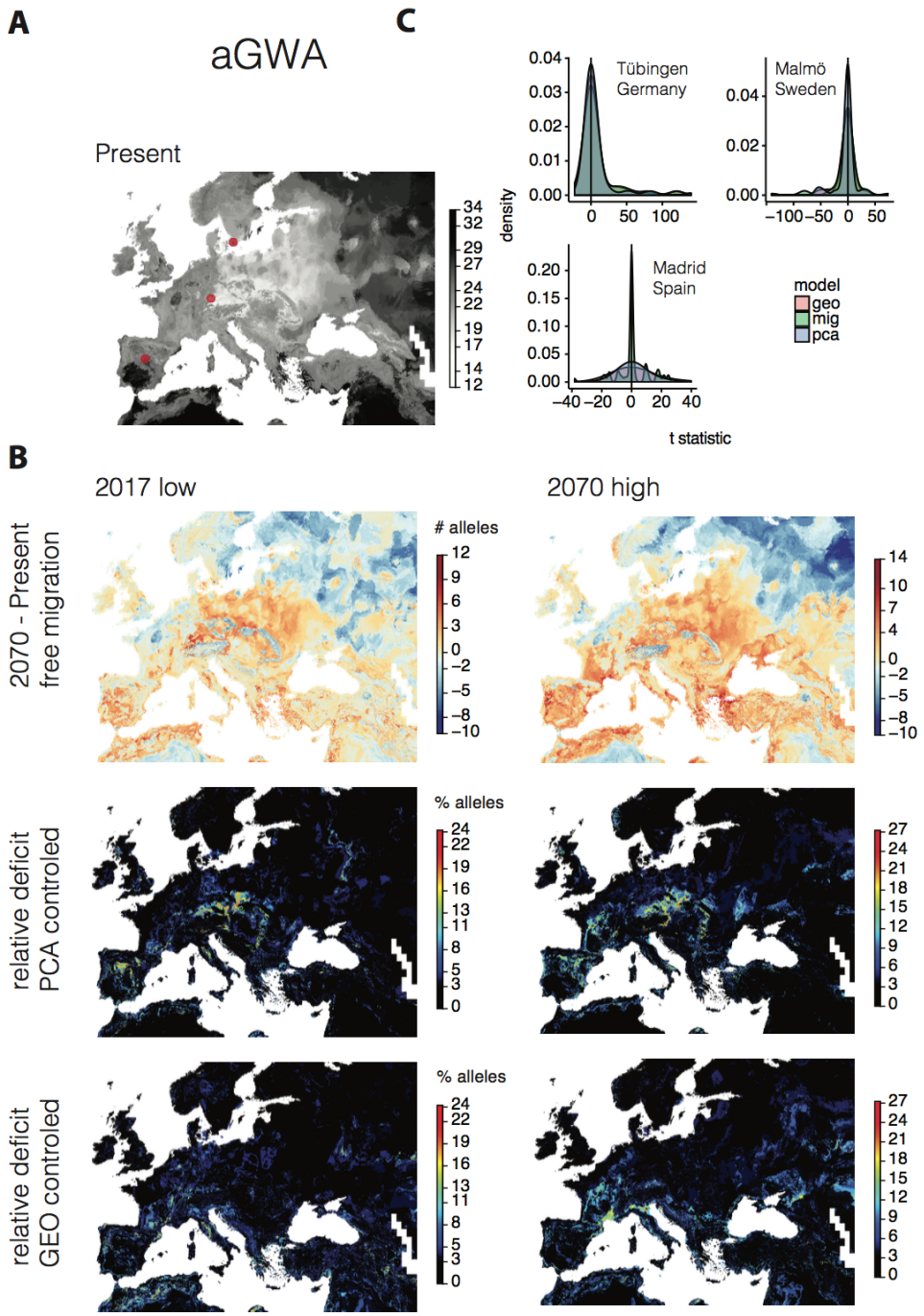
Supplementary Figure 11. Profile of phenotypic change under climate change.

(**A, B**) Correlation of precipitation during the warmest quarter today and in 2070 under a high CO_2 scenario. Colors indicate current drought survival (**A**) or flowering time (**B**), and symbol shapes indicate increase or decrease in trait values for 2070. (**C, D**) Regression of the predicted change in drought survival index (**C**) and flowering time (**D**) on the predicted change in precipitation in 2070. Note that areas with already low precipitation will not have large decreases in precipitation in 2070 (A-B). Note also the linear relationship between decreased precipitation in 2070 and predicted increase of drought survival in (C). Flowering will be on average faster in 2070 (D), but the relationship between precipitation reduction and flowering time change is not linear, which suggests that areas with a moderate reduction in precipitation will have accelerated flowering (rather than increased drought survival).


Present
2070 low
2070 high
bio12
2894 2577 2261 1944 1627 1311 994 677 361 44
210 158 105 52 0 −80 −160 −240 −320
bio18
624 555 485 416 347 277 208 139 69 0
170 128 85 42 0 −42 −85 −128 −170
bio1
256 214 173 131 90 48 7 −35 −76 −118
52 46 40 35 29 23 17 12 6 0
bio4
13314 12124 10934 9743 8553 7363 6173 4982 3792 2602
1400 1167 933 700 467 233 0 −233 −467 −700
bio5
467 416 364 313 261 210 158 107 55 4
69 61 54 46 38 31 23 15 8 0
bio6
113 71 30 −12 −54 −95 −137 −179 −220 −262
69 61 54 46 38 31 23 15 8 0
bio9
500 375 250 125 0 −125 −250 −375 −500

Supplementary Figure 12. Maps of the most important climatic variables.

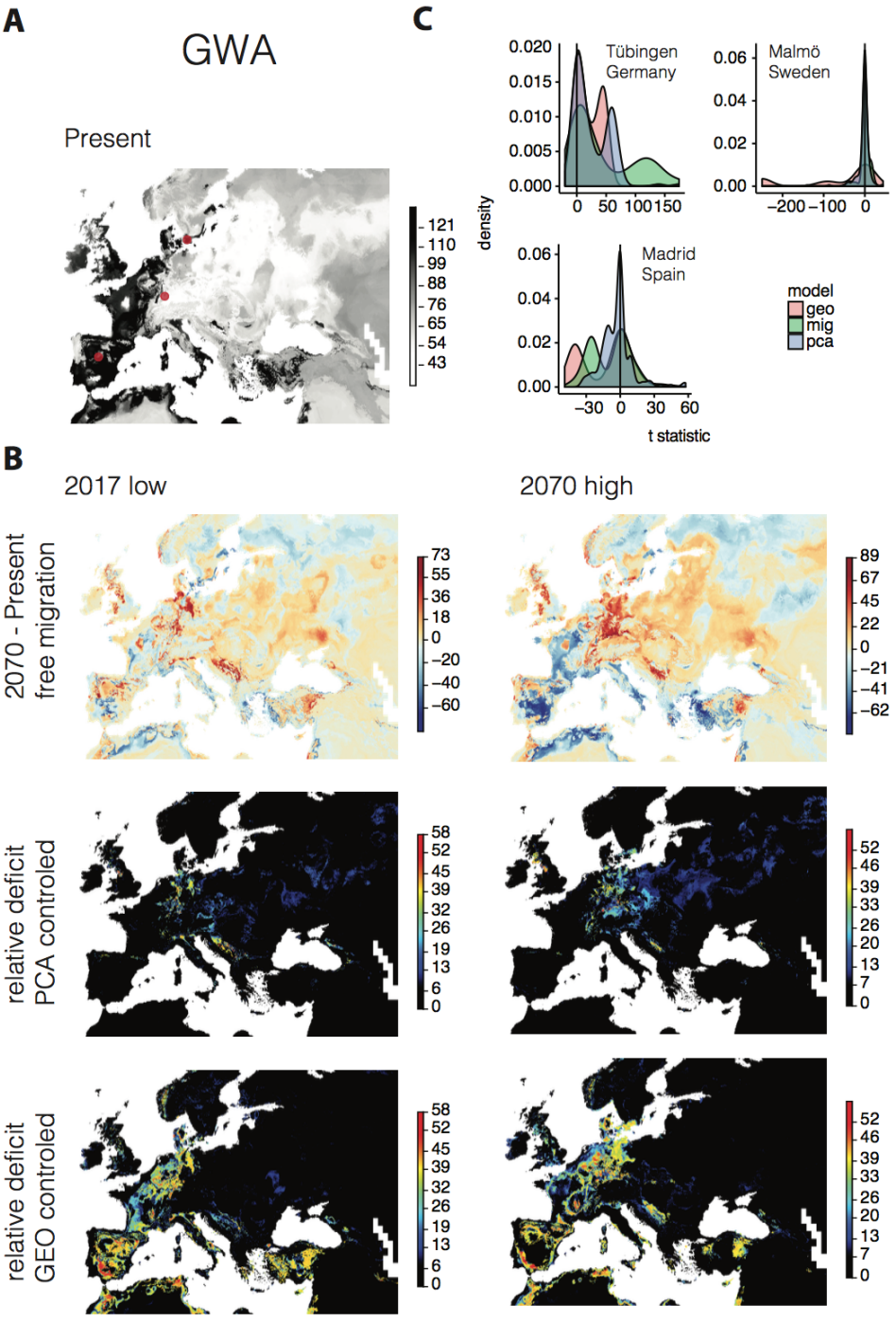
The bioclimatic variables (www.worldclim.org) that typically had more importance in phenotypic and genome environmental models are shown as an aid for interpretation of the results from our study. bioclim variables shown are annual precipitation (bio12), precipitation of the warmest quarter (bio18), annual mean temperature (bio1), temperature seasonality (bio4), maximum temperature of the warmest month (bio5), minimum temperature of the coldest month (bio6), and mean temperature of the driest quarter (bio9). The columns show distributions at present, in 2070 under a scenario of low CO_2 concentration, and in 2070 under a scenario of high CO_2 scenario. Except for bio9, the values for future scenarios were expressed as future-present difference to highlight geographic areas that will change the most. Note the bimodality of bio9: areas in black are summer drought (Mediterranean climate) areas, whereas blue areas indicate winter drought. Also note that bio18 is predicted to change mostly along the transition from Mediterranean to non-Mediterranean climate. For bio18, areas that will have lower precipitation than any current location of *A. thaliana* are shown in black, to highlight that most areas will remain within the range of current precipitation across the species range.



Supplementary Figure 13. aGWA Genome Environment Models (GEMs).

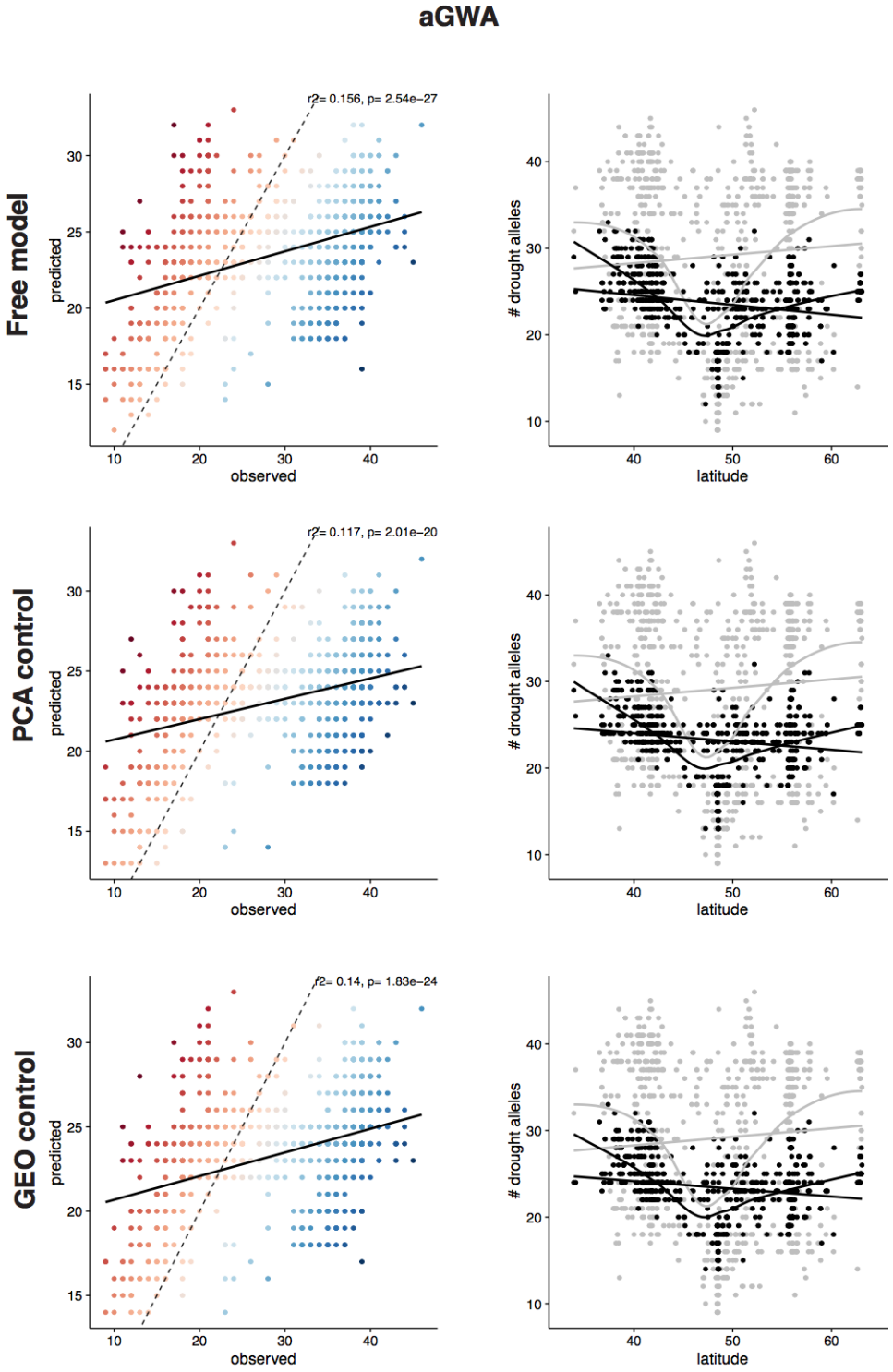
(**A**) We ran GEMs to describe the geographic distribution of alleles at the 70 aGWA top loci. Concatenating all maps, we produced a map of the count of all drought-survival alleles that a genotype is expected to have in a given location today. (**B**) The trained model from (A) was used to predict distribution of drought survival alleles in the future. The difference to numbers inferred for today (A) corresponds to the alleles that will have been gained or lost in 2070 in a given location. Two additional models were trained which included a genome background (PCA) correction and latitudinal and longitudinal (GEO) correction of the allele distributions. The percentage of gained

alleles from the “free” model that were not present in the corrected models is shown as a deficit in percentage. (**C**) For three highly sampled locations, Madrid (Spain), Tübingen (Germany) and Malmö (Sweden), we calculated allele frequency differences between today and 2070 (under high CO_2) and calculated a t-statistic to describe the effect size of the change. A skew towards the right (increase) is observed for Tübingen only.



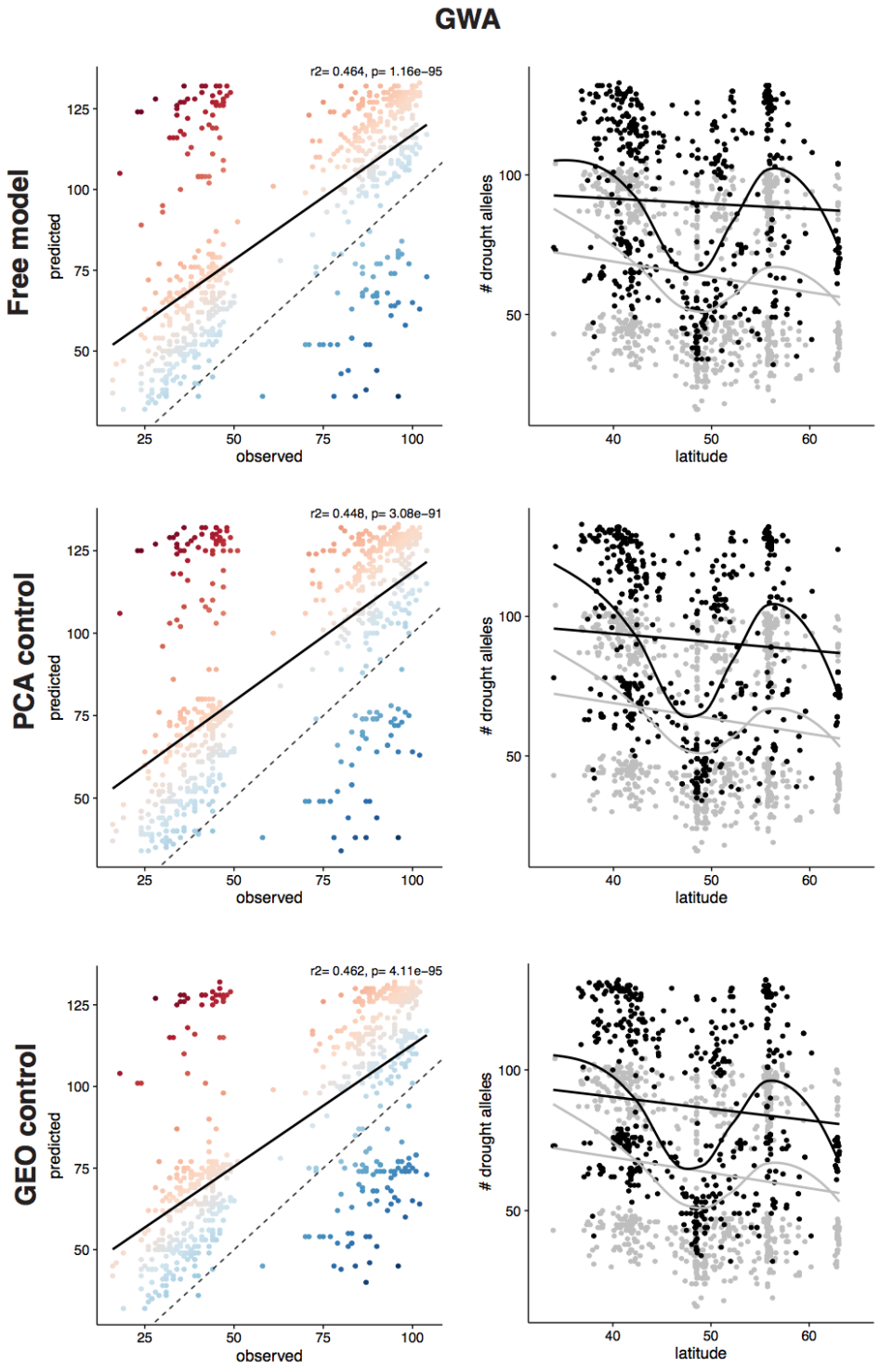
Supplementary Figure 14. GWA Genome Environment Models (GEMs).

See Supplementary Fig. 13 for legend.



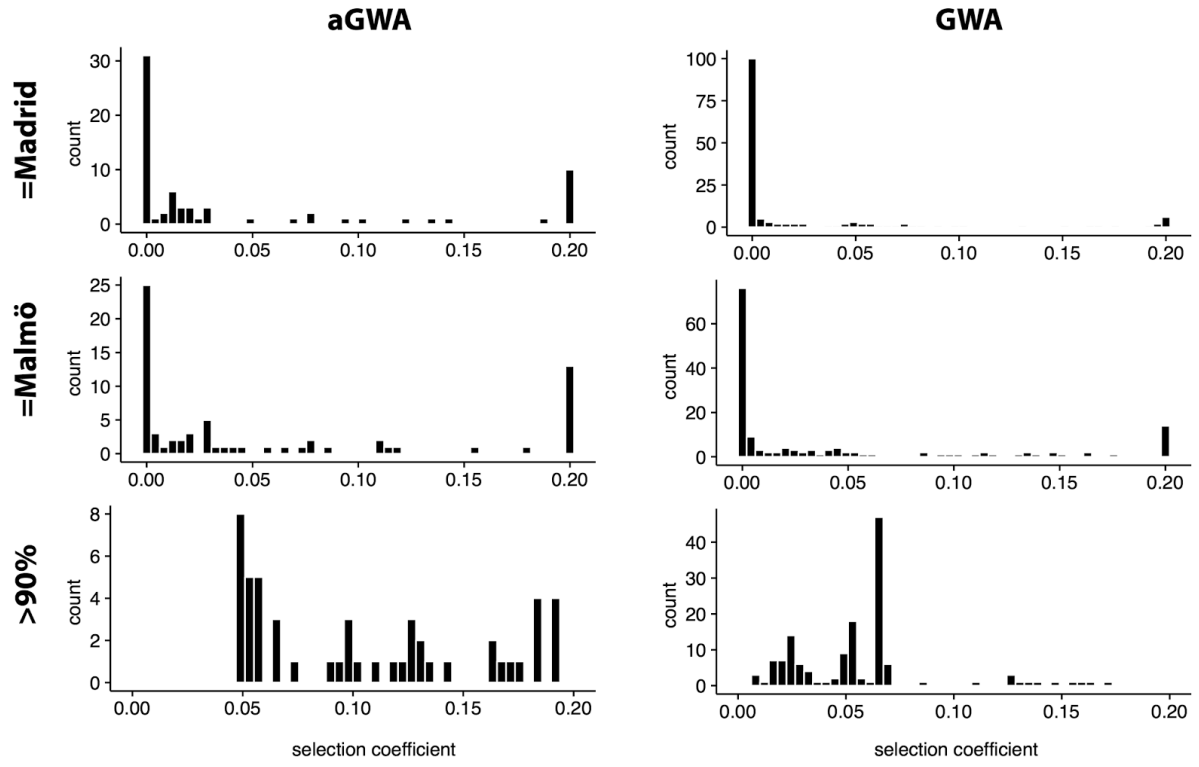
Supplementary Figure 15. aGWA GEM residuals.

For each GEM, we plotted predicted against observed (empirical) number of drought-associated alleles at each sampled location. Red color indicates overestimation and blue underestimation (dashed line is the one-to-one relationship; solid line is the true regression). Latitudinal trends of predicted (grey) and observed (black) are shown (right). Note that variance of predictions is larger than the empirical observations, probably due to the discrete nature of random forests.



Supplementary Figure 16. GWA GEM residuals

See Supplementary Fig. 15 for legend.



Supplementary Figure 17. Population genetics simulations.

We ran Wright-Fisher population simulations of 70 (aGWA) or 151 (GWA) independent loci for 50 generations of evolution under mutation-selection balance, starting with the current allele frequencies in the Tübingen population, and repeating each simulation with an array of selection coefficients from 0.0001 to 0.2 (relative fitness advantage) for each locus. The distributions shown correspond to the positive selection coefficients that are required for the drought survival alleles to rise to the frequency at which they are currently found in Malmö (top) or Madrid (center), or to at least 90% (bottom), which is close to fixation.

SUPPLEMENTARY VIDEO in Supplementary_Video.gif

19-frames time series of green-segmented images for one exemplary tray is available at:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41559-017-0423-0/MediaObjects/41559_2017_423_MOESM4_ESM.gif

Thesis Appendix II

“A map of climate change-driven natural selection in *Arabidopsis thaliana*”

Exposito-Alonso, M., 500 Genomes Field Experiment Team, Burbano, H. A., Bossdorf, O., Nielsen, R., Weigel, D. (2018) *bioRxiv*, https://doi.org/10.1101/321133.

Exposito-Alonso, M., Rodríguez, R.G., Barragán, C., Capovilla, G., Chae, E., Devos, J., Dogan, E.S., Friedemann, C., Gross, C., Lang, P., Lundberg, D., Middendorf, V., Kageyama, J., Karasov, T., Kersten, S., Petersen, S., Rabbani, L., Regalado, J., Reinelt, L., Rowan, B., Seymour, D.K., Symeonidi, E., Schwab, R., Tran, D.T.N., Venkataramani, K., Van de Weyer, A.-L., Vasseur, F., Wang, G., Wedegärtner, R., Weiss, F., Wu, R., Xi, W., Zaidem, M., Zhu, W., García-Arenal, F., Burbano, H.A., Bossdorf, O., Weigel, D., (2017). *bioRxiv*, https://doi.org/10.1101/186767.

A map of climate change-driven natural selection in *Arabidopsis thaliana*

Moises Exposito-Alonso[1], 500 Genomes Field Experiment Team[2], Hernán A. Burbano[3], Oliver Bossdorf[4], Rasmus Nielsen[5], Detlef Weigel[1*]

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.

[2] See author contributions section

[3]Research Group of Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.

[4]Institute of Evolution and Ecology, University of Tübingen, 72076 Tübingen, Germany.

[5]Departments of Integrative Biology and Statistics, University of California Berkeley, Berkeley, CA 94720, USA. Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 København K, Denmark

*correspondence to: weigel@weigelworld.org

Keywords: *Arabidopsis thaliana*, climate change, environmental niche models, field experiments, genetic natural selection, selection scan.

Running title: A map of climate change-driven natural selection

Through the lens of evolution, climate change is an agent of natural selection that forces populations to change and adapt, or face extinction. Current assessments of the biodiversity risks associated with climate change[1,2], however, do not typically take into account that natural selection can dramatically impact the genetic makeup of populations[3]. We made use of extensive genome information in *Arabidopsis thaliana* and measured how rainfall-manipulation affected the fitness of 517 natural lines grown in Spain and Germany. This allowed us to directly infer selection at the genetic level[4]. Natural selection was particularly strong in the hot-dry Spanish location, killing 63% of lines and significantly changing the frequency of ~5% of all genome-wide variants. A significant proportion of this selection over variants could be predicted from the climate (mis)match between experimental sites and the geographic areas where variants are found (R^2=29-52%). Field-validated predictions across the species range indicated that Mediterranean and Western Siberia populations — at the edges of the species' environmental limits — currently experience the strongest climate-driven selection, and Central Europeans the weakest. With more frequent droughts and rising temperatures in Europe[5], we forecast an increase in directional selection moving Northwards from the South range, putting many native *A. thaliana* populations at evolutionary risk.

To predict the future impact of climate change on biodiversity, the typical starting point has been climatic tolerances inferred from the current species distributions. These tolerances are usually treated as static, and risks are assessed based on whether species' environmental niches will shrink[1,2] or shift faster than the species can migrate[1,6]. However, these approaches do not account for within-species genetic variation, and for natural selection causing species to genetically change and adapt over time[3,7]. To predict the "evolutionary impact" of climate change on a species, i.e. how much genetic change is required for adaptation to climate change, we thus need to quantify and model environment-driven natural selection at the genetic level. Thanks to species-wide genome scans[8–10], as well as genome associations with climate of origin[11–16], we increasingly understand the genomic basis of past selection and climate adaptation, which has been used to estimate future "genomic vulnerability" of populations[11,12].

Natural selection, however, is only indirectly inferred in the types of analyses discussed above. The best way to directly quantify selection in a specific environment is provided by field experiments in which multiple genotypes of a species are grown together in a common environment[17–19]. With such experiments, relative fitness can be directly associated with genetic variation across populations[4,20–22]. Ideally, one would carry out such field experiments at many

different sites throughout the species range, but this is rarely practical. Nevertheless, an emergent finding is that individuals are normally locally adapted and that local genotypes are often positively selected over foreigners in their "home" environment, while negatively selected in their "away" environments[23,24]. From this knowledge, it should be possible to derive a metric of how selection would change in a future home environment that is altered by climate change, and in turn, a metic of future adaptation deficit of local populations. Here we combine high-throughput associations of genome and current climate variation with experimentally quantified *in situ* natural selection in the plant *Arabidopsis thaliana*. We exploit these associations to forecast natural selection driven by future climate change, and how it impacts the genomic variation of a species across its geographic range — what we interpret as a new metric of evolutionary risk of populations.

To study climate change-driven natural selection in the annual plant *A. thaliana*, we performed two common garden experiments for one generation in two climatically distinct field stations, at the warm edge of the species distribution in Madrid (Spain, 40.40805ºN -3.83535ºE), and at the distribution center in Tübingen (Germany, 48.545809ºN 9.042449ºE) (for details see Supplemental Appendix II). At each site, we simulated high precipitation typical of a wet year in Germany, and low precipitation typical of a dry year in Spain (we used four flooding tables with a split replicated design of two wet and two dry treatments in each site, see Fig. SII.2, Table SII.1). In fall of 2015 we sowed over 300,000 seeds of 517 natural lines capturing species-wide genomic diversity[25] and randomized within treatment (Dataset 1-2). For each line, we prepared seven pots in which only a single plant was retained after germination *in situ*, and five pots with exactly 30 seeds that were allowed to germinate and grow without intervention throughout the experiment. At the end of the experiment in June 2016, we had collected data from 23,154 pots, consisting of survival to the reproductive stage, the number of seeds per surviving plant (fecundity), and lifetime fitness (the product of survival and individual fecundity) (Dataset 3-4). Heritability of fitness varied across environments and between survival and fecundity. It was generally highest in the most stressful environment ($H^2_{survival}$=0.551; Table SI.3), as defined by reduced survival, i.e., in Spain under low precipitation and at high plant density. In this environment, only 193 of the 517 accessions survived, whereas in Germany at least a few plants of each accession reproduced (Table SI.1).

In each experimental environment, we quantified genome-wide selection at the genetic level based on the difference in relative fitness of lines with the minor and the major allele at each genomic position (1,353,386 biallelic SNPs across 515 lines with high-quality genome information, see Supplemental Appendix I section IV). Because *A. thaliana* is a selfer species with extensive population structure, our approach quantifies selection both in causal variants, as well as many more

variants that are in significant linkage disequilibrium (LD) with causal variants[26,27] — a phenomenon behind the concepts of background selection or genetic hitchhiking. We use the term *total selection coefficient* (s, following the interpretation and methods in Gompert et al. [27]), to denote the realized selection affecting each SNP resulting from the combination of selection acting directly on the focal variant, and the indirect effects due to selection on causal SNPs that are in LD with the focal variant. This total selection coefficient best reflects the increase or decrease of frequency of a variant after one generation of selection (see simulations Fig. SI.15). Using a Genome-Wide Association (GWA) approach with Linear Models (LM-GEMMA, ref. [28]) to calculate total selection coefficients, we found a total of 421,962 SNPs below a 0.05 significance threshold (Benjamini & Hochberg FDR correction) in at least one of the eight environments (Fig. 1) (Table SI.3). Using the more stringent Bonferroni correction ($<7x10^{-7}$), we still detected 6,538 SNPs distributed throughout the genome, suggesting that the polygenic model of natural selection[29] prevails in this climate-manipulation experiment (Fig. 1). These high numbers are not surprising, given that we expect to capture many SNPs that are only indirectly selected. Thinking about our experiment as studying a population of plants with multiple genotypes, the change of allele frequencies in response to one generation of selection would be up to 10% in Spain and low precipitation, while it would not exceed 2% in the benign high-precipitation environment in Germany (Table SI.4, Fig. SI.9, Supplemental Appendix I section IV). While variants inferred to be under positive or negative selection after Bonferroni-correction were overall more likely to be located in intergenic regions than in genes (Fisher's Exact test Odds ratio [Odds]=1.11, $P=7x10^{-30}$), such variants were enriched for nonsynonymous mutations (Odds=1.05, $P=2x10^{-4}$). The large number of variants affected by selection implies a strong turnover of variation across the entire genome as a response to the environment[30,31], and a potentially significant demographic decimation — what Haldane called "the demographic cost of natural selection"[32].

Changes in allele frequency are not only determined by the adaptive value of a variant but also the alleles it is linked to. We therefore improved the detection of direct targets of selection by correcting for LD-driven effects[27,33] using Bayesian Sparse Linear Mixed Model associations with relative fitness (BSLMM-GEMMA, ref. [33]), see Supplemental Appendix I section IV). This analysis estimated that the most likely number of causal loci ($$n\gamma$$) was in the range of 7 to 89, depending on the experiment (although this hyperparameter tends to be underestimated, see Gompert et al. [27]). Because this number was much smaller than the total number of variants that experience total selection (tens to thousands, Table SI.3), our results indicate that selection must be mostly indirect, i.e. via linkage disequilibrium[4,26,34].

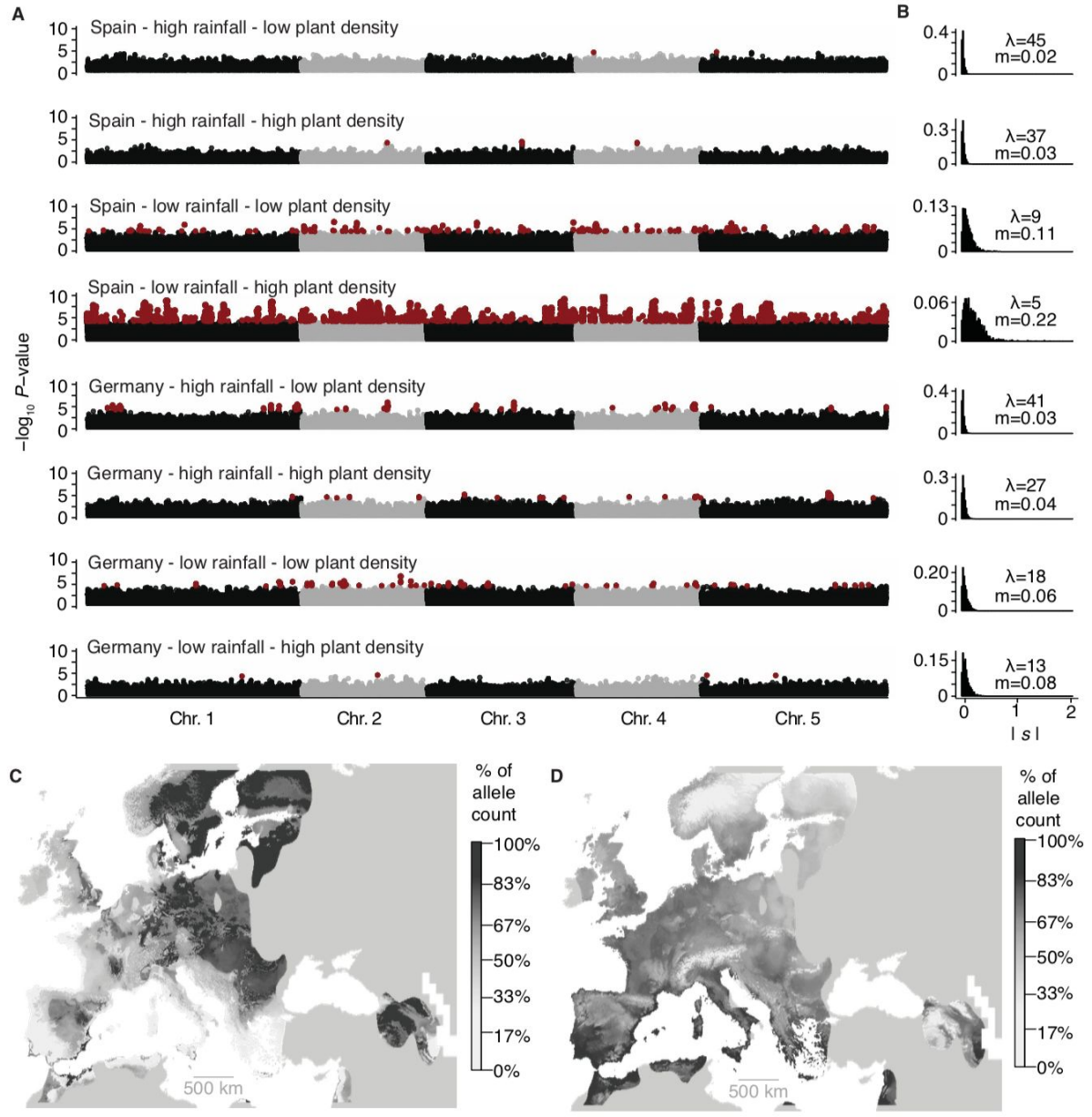


Fig. 1 A genome map of total selection coefficients. (A) Manhattan plots of SNPs significantly associated with relative lifetime fitness in eight different environments. SNPs significant after FDR (black and grey) or Bonferroni correction (red) are shown. For genome-wide scans of survival and fecundity fitness see Fig. SI.4 and SI.5. (B) Distribution of absolute total selection coefficients $|s|$ per experiment. λ denotes maximum likelihood-inferred parameter of an exponential distribution, and *m* denotes the mean total selection coefficient. (C, D) Genome-wide Environmental Niche Models for the most significant SNPs in each 0.5 Mb window of the genome. Color scale indicates the % of the total number of positive alleles locally present. (C) 424 windows had significant SNPs in high-precipitation experiments. (D) 279 windows had significant SNPs in low-precipitation experiments.

We studied whether alleles selected in one environment were typically selected in other environments it it was rather different genetic variants that were selected in each environment.

Alleles that were positively selected under low precipitation tended to be negatively selected under high precipitation, and vice versa, so-called antagonistic pleiotropy[21] (Fig. 2, Fisher's exact test Odds Ratios >1.31, $P<4x10^{-24}$) — an observation that is particularly clear when comparing the two most "natural" conditions, low precipitation in Spain and high precipitation in Germany (Odds Ratio=6.72). In contrast, when we compared the same precipitation condition between the two locations, selection was either in the same direction (0.23<Pearson's r<0.51), or there was selection in one environment and neutrality in the other, displaying conditional neutrality (All Odds ratio<1, $P<10^{-16}$). Together, this indicates opposite selection across precipitation but not temperature gradients. This is an important observation, because it tells us that nature might not be able to select for generalist genotypes that are successful in a wide range of precipitation environments.

To study whether short-term selection in our experiments aligns with genomic footprints of past selection (see Supplemental Appendix I section II, V and VI, we searched for selective sweeps[35], for outlier allele frequency differentiation (F_{ST}) between eleven previously defined *A. thaliana* genetic groups[11,25], and for climate-genome associations[13,14] (GWA with 1960-1990 climate averages, worldclim.org, ref. [36]). Comparing frequency-matched background SNPs with Bonferroni-corrected significant SNPs for total selection coefficients, we found that the latter had higher average F_{ST} values (0.39 compared to 0.14, Wilcoxon test, $P<10^{-16}$), but were not any more likely to have experienced a selective sweep (P=0.2) (Fig. 2, Fig. SI.7-8). Absolute values of total selection coefficients were significantly higher for strongly climate-correlated SNPs (e.g. annual precipitation [bio1] and temperature [bio12]: Spearman's rho=0.12, $p<10^{-16}$). The 1% top hits for climate associations also had higher F_{ST} values than frequency-matched background SNPs (e.g. bio1 and bio12: $P<10^{-5}$), but no differences in sweep likelihood (P=0.9). Implementing genome-wide environmental niche models[11] (see Supplemental Appendix I section VI), we found that alleles selected in Germany and high precipitation were more likely to come from higher latitudes (Fig. 1C), while the opposite was true for alleles selected in Spain and low precipitation (Fig. 1D). In agreement, alleles coming from regions with precipitation regimes to the experimental site, tended to be positively selected (Fig. 2D). All in all, the fact that there is a genome-wide signal of correlations between total selection coefficients with allele frequency shifts across population lineages and climate regions is most easily reconciled with a polygenic model of natural selection rather than with a selective sweep model[30].

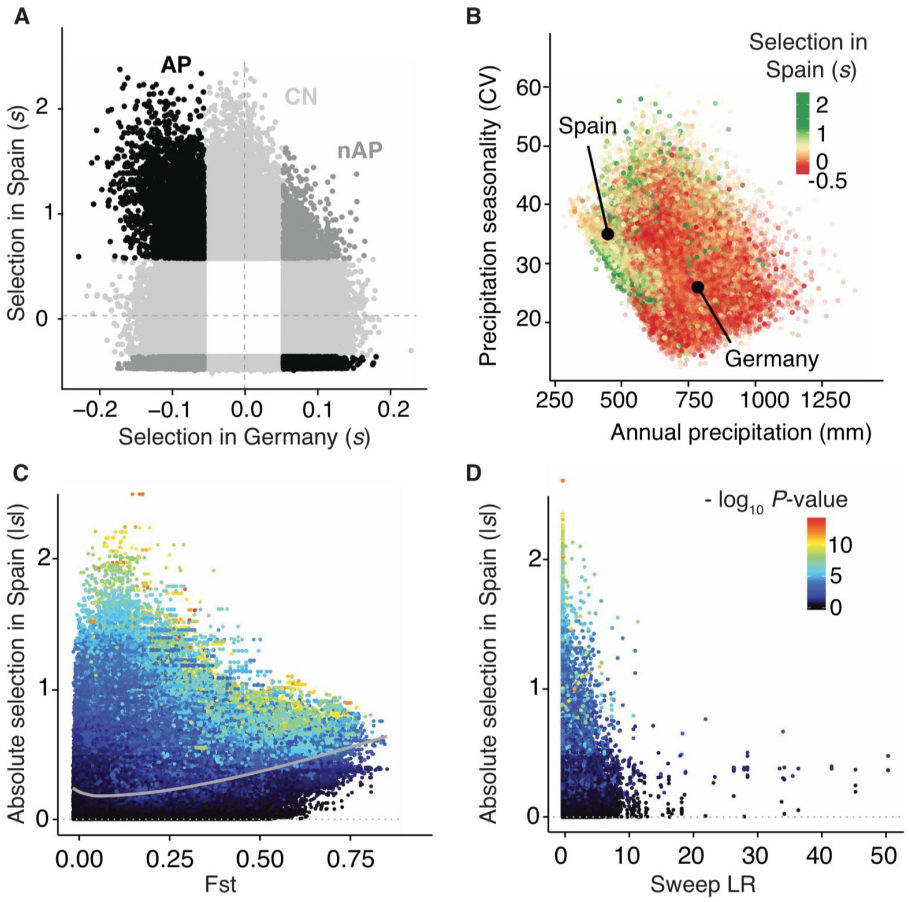


Fig. 2 Selection trade-offs and the signal of environmental local adaptation. (A) 5% extreme tails of total selection coefficients across two contrasting environments; Spain with low precipitation and high population density, and Germany with high precipitation and low population density. In light grey are conditionally neutral alleles for either environment (CN, n=265436), in black are alleles behaving as antagonistic pleiotropic (AP, n=20503), and in dark grey are alleles non-antagonistic pleiotropic (nAP, n=9681). (B) Mean annual precipitation and precipitation seasonality in the geographic areas of origin of SNPs (n=1,353,386 SNPs). Black circles indicate the average climate values of Spain (left) and Germany (right). (C) Relationship between field absolute total selection coefficients and F_{ST} values across 11 lineages, and (D) the likelihood ratio of selective sweeps (n=1,353,386 SNPs).

We finally aimed to build an environmental model that can predict total selection coefficients based on the climate and diversity patterns. We used a regression with decision trees using Random Forests to build what we call Genome-wide Environment Selection (GWES) models. The response variable was total selection coefficients, and we used as predictors the per-allele associations with climate of origin from multiple climatic GWAs (β_{clim}), the local climate at the experimental facility (E_{clim}), the signatures of past selection at each SNP (F_{ST}, π, and sweep likelihood ratio LR_S), and their genome annotation (An). Our model thus learned the function:

$s = f(\beta_{clim}, E_{clim}, F_{ST}, \pi, LR_S, An)$. All predictors were derived from public databases (worldclim.org, 1001genomes.org, arabidopsis.org) (see Supplemental Appendix I section VII). Conceptually, GWES models are similar to Environmental Niche Models (ENMs), but instead of training them with presence/absence data of a genetic variant[11,12], we trained them with our measured total selection coefficients. This provided a means to predict whether alleles should increase/decrease in frequency in a certain climate, instead of merely an indication of whether alleles are likely to be present, which is the indirect ENMs' version [11]. By training models on total selection coefficients in Spain and Germany (10,000 SNPs), testing the accuracy of models using cross-validation (i.e. 10,000 other SNPs) and the confidence intervals with bootstrapping (100 samples of 100 SNPs), we confirmed that total selection coefficients were correctly predicted, with a high correlation accuracy (0.56 < Pearson's r_{cv} < 0.7) and explaining a large proportion of variance (R^2_{cv}= 29—52%) (Fig. 3A, for variable importance see Table SI.7) (further details in Supplemental Appendix I section VII). To further cross-validate the predictive accuracy of our models in other unknown environments, we made use of published fitness data for partially-overlapping sets of natural lines that had been grown at different locations in Spain, Germany and England[37,38]. Using these data and GWES predictions based on the climate at those locations, we confirmed moderate predictability (7%<R^2_{cv}<36%) (Fig. 3A, Table SI.8) (for further discussion on null expectations and cases of apparently low predictability, see Fig. SI.11, SI.12, and Supplemental Appendix I section VIII).

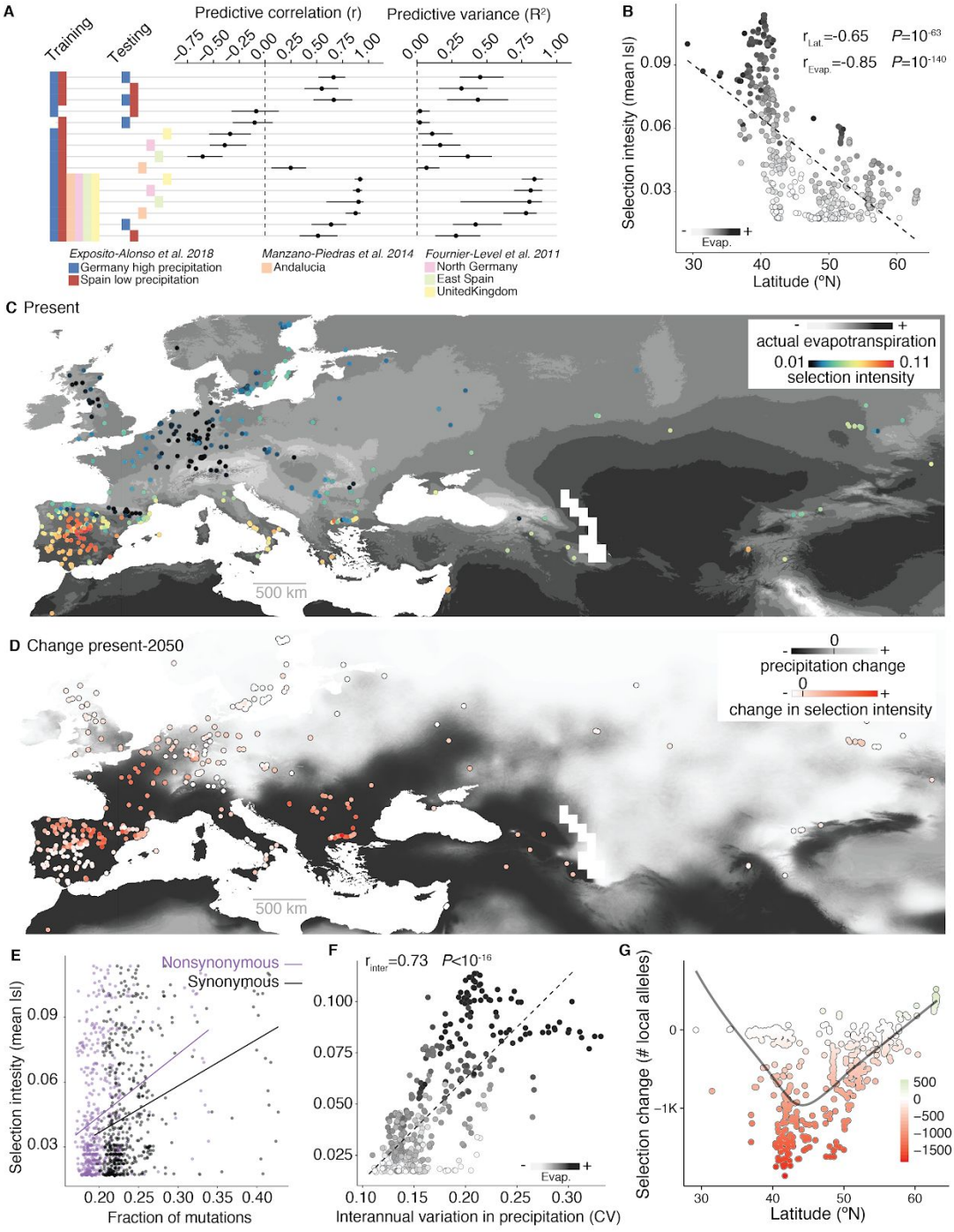


Fig. 3 A geographic map of climate-driven selection and its predictability. (A) Genome-wide Environment Selection (GWES) models trained and tested with different combinations of our data from Germany and Spain, or previously published field experiments (accuracy was estimated using cross-validation and the 95% confidence intervals using bootstrap, see Supplemental Appendix I section VII). (B-C) Mean GWES-predicted total selection coefficients ("selection intensity"; n= 10,752 SNPs, one random SNP per 10 kb windows) in known locations of *A. thaliana* populations in relationship to latitude and evapotranspiration in summer (ref. [36]). (D) Predicted changes in selection intensity using climate projections to 2050 as a proxy of a sudden climate change (2050 MP rcp 8.5, ref. [36]). (E) Relationship between selection intensity and synonymous and nonsynonymous polymorphisms present at each location. (F) Relationship between selection intensity and

interannual variation in precipitation from 1958-2017 (ref. [39]). (G) Number of local alleles (of the total 10,752 SNPs) whose selection is predicted to positively or negative change >5% in relative fitness in 2050 across the latitudinal range.

Using the trained GWES models, we then predicted genetic natural selection at hundreds of locations, simulating field experiments in which the same set of diverse natural lines is challenged by different local climates (Fig. 3). The intensity of selection, i.e. genome-wide average total selection coefficients, was strongest towards the environmental limits of the species, i.e. in hot (annual temperature, Spearman's rank correlation rho=0.62, $P<10^{-16}$), dry (annual precipitation, rho=-0.457, $P=10^{-27}$), and high evapotranspiration locations (actual evapotranspiration in August, rho=0.86, $P<10^{-16}$) (Fig. 3B-C, Table SI.11). High selection intensity coincided with locations where natural lines have a lower-than-average ratio of nonsynonymous to synonymous polymorphisms (Fig. 3E, r=-0.276, $3x10^{-10}$, Fig. SI.14), high local genetic diversity π (rho=0.187, $P=2.63x10^{-5}$) and elevated Tajima's D (rho=0.161, $P=3x10^{-4}$). Various demographic scenarios could partially explain some of these patterns in isolation, i.e. bottlenecks can reduce the nonsynonymous polymorphisms due to they are typically at low frequency or high diversity might be found in old, large populations. These patterns are, however, congruent with stronger selection having acted more efficiently over nonsynonymous mutations. In addition, high diversity could also be driven by strong natural selection fluctuating over time, with alternative polymorphisms having been selected in each period[40]. To test that we inspected precipitation data[39] from 1958 to 2017 revealed that locations where we had inferred strong selection, low nonsynonymous substitutions and high diversity also suffered from highly variable climate (rho=0.22, $P<8x10^{-7}$; Fig. 3F). The overlap of temporal climate stochasticity and climate extremes has a number of evolutionary consequences, namely the interruption of adaptive walks towards theoretical optimums[41], the maintenance of multiple genotypes per population[40], and the evolution of bet-hedging strategies[42]. Therefore these findings also highlight the importance of temporal resolution in climate databases for ecological predictive models. All in all, we did not find evidences that the warm edge of the geographic distribution of *A. thaliana* is limited by an increase in drift that causes lowly diverse small populations to accumulate nonsynonymous deleterious mutations, as some theories propose[43]. Rather, our observations and predictions (Fig. 3 C, E and F) indicate that a the species' warm geographic limit is primarily defined by the environmental tolerance limits, where climate-driven natural selection is the limiting factor for the survival of individuals and populations outside their range edges[44].

A sudden change in climate and increased climate variability[45,46] will obviously increase the magnitude of natural selection. Using climate projections of 2050 as a proxy for potentially abrupt changes in local climate (Intergovernmental Panel on Climate Change, www.ipcc.ch, ref. [5,36]), we

predict that selection intensity will likely increase in much of Southern-Central Europe, with an expected decrease in annual precipitation and increase in annual temperatures (Fig. 3D, Fig. SI.3, SI.10). To enable comparability across locations, our metric of selection intensity is by design standardized based on the same reference set of 515 accessions representing the whole species diversity. Therefore, it can be interpreted as the fraction of the species diversity suitable for survival and reproduction in a given environment, or as the magnitude of allele frequency changes and allele fixations in response to a single generation of selection (Fig. 3C-D, for a discussion on pitfalls and interpretations see Supplemental Appendix I section VII.3). Local populations, however, typically consist of more closely related lines that harbor only a subset of genetic variants, which may put these populations either in a better or worse position to respond to future climate than our global set of more diverse lines. We therefore looked for SNPs predicted to change most strongly in selection under the 2050's projected climate (fitness advantage or disadvantage changed over 5%), and evaluated whether the allele positively changing in selection is the one locally present or rather the opposite. We found that most local alleles will become more negatively selected if climate would suddenly change (Fig. 3G, Fig. SI.13). We therefore predict that many native populations — specifically those in transition zone from the Mediterranean to Temperate regions[47] — could suffer a negative demographic impact due to a diminished degree of local adaptation and an increased intensity of natural selection. As Southern Mediterranean populations are already locally adapted to low precipitation regimes, gene flow from those could catalyze evolutionary rescue of more vulnerable, central populations[48].

Conclusion

The expected changes in climate during the 21st century will threaten the survival of many species. Because the distribution of genetic diversity is so well characterized in *A. thaliana*, we have used it to address the challenge of predicting the effects of climate-driven natural selection over genomic variation across a species' range. Integration of genome-climate associations with direct fitness observations allowed us to build models that predict selection at the genetic level rather than mere probability of presence/absence of variants. This information enabled us to infer range-wide evolutionary risk in the face of rapid climate change. The first two steps in our project, assembling a worldwide collection and genome sequencing of diverse lines, are in reach for many species of plants. A greater challenge is generating fitness data, but this can be partially solved by identifying particularly informative field sites — as we have done in our study – and by exploiting the immense technological progress in grassland, forest, or farm monitoring at different scales[49,50]. Combining such observations with our new genome-wide environment modeling approach will help us to fully incorporate evolution into predicting the impacts of climate change on biodiversity.

ADDITIONAL INFORMATION

Data availability Phenotypic datasets are available as supplemental material with doi:. Genomes are available at http://1001genomes.org/data/GMI-MPI/releases/v3.1/. The seed collection can be obtained from the Arabidopsis Biological Resource Center (ABRC) under accession CS78942. The GWA scans for fitness and climate variables will be deposited at aragwas.1001genomes.org.

Author contribution MEA, HAB and DW conceived the project outline. MEA designed, implemented and coordinated the project. MEA carried out statistical analyses with advice from RN. HAB, OB, RN, and DW supervised the project and discussed analyses interpretation. MEA prepared the first draft and the final manuscript was written by MEA, HAB, OB, RN, and DW. MEA carried out the experiment in Tübingen and in Madrid with technical support of the 500 Genomes Field Experiment Team: Moises Exposito-Alonso[1], Rocío Gómez Rodríguez[2], Cristina Barragán[1], Giovanna Capovilla[1], Eunyoung Chae[1], Jane Devos[1], Ezgi S. Dogan[1], Claudia Friedemann[1], Caspar Gross[1], Patricia Lang[1], Derek Lundberg[1], Vera Middendorf[1], Jorge Kageyama[1], Talia Karasov[1], Sonja Kersten[1], Sebastian Petersen[1], Leily Rabbani[1], Julian Regalado[1], Lukas Reinelt[1], Beth Rowan[1], Danelle K. Seymour[1], Efthymia Symeonidi[1], Rebecca Schwab[1], Diep Thi Ngoc Tran[1], Kavita Venkataramani[1], Anna-Lena Van de Weyer[1], François Vasseur[1], George Wang[1], Ronja Wedegärtner[1], Frank Weiss[1], Rui Wu[1], Wanyan Xi[1], Maricris Zaidem[1], Wangsheng Zhu[1], Fernando García-Arenal[2], Hernán A. Burbano[1], Oliver Bossdorf[3], and Detlef Weigel[1] ([1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany. [2]Center for Plant Biotechnology and Genomics, Technical University of Madrid, Pozuelo de Alarcón, Spain. [3]Institute of Ecology and Evolution, University of Tübingen, Tübingen, Germany). See author contributions to the field experiments in Supplemental Appendix II.

Acknowledgements We gratefully thank Patricia Lang, Angela Hancock, and Talia Karasov for comments on the manuscript, and the Weigel and Burbano labs for discussions. We also thank Xavi Picó for advice on experimental design, Ilja Bezrukov for advice on image processing replicability, and Belen Mendez-Vigo, Carlos Alonso-Blanco, Antolín López Quirós, Marisa López Herránz and Miguel Ángel Mora Plaza for assistance during sowing in Madrid.

Funding statement This work was funded by an EMBO ST fellowship (MEA), ERC Advanced Grant IMMUNEMESIS and the Max Planck Society (DW).

Disclosure statement The authors declare no competing financial interests. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Urban, M. C. Accelerating extinction risk from climate change. *Science* **348,** 571–573 (2015).
2. Warren, R., Price, J., Graham, E., Forstenhaeusler, N. & VanDerWal, J. The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C. *Science* **360,** 791–795 (2018).
3. Hoffmann, A. A. & Sgrò, C. M. Climate change and evolutionary adaptation. *Nature* **470,** 479–485 (2011).
4. Thurman, T. J. & Barrett, R. D. H. The genetic consequences of selection in natural populations. *Mol. Ecol.* **25,** 1429–1448 (2016).
5. Intergovernmental Panel on Climate Change. *Climate Change 2013 - The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge University Press, 2014). doi:10.1017/CBO9781107415324
6. Jezkova, T. & Wiens, J. J. Rates of change in climatic niches in plant and animal populations are much slower than projected climate change. *Proc. R. Soc. B* **283,** 20162104 (2016).
7. Anderson, J. T., Inouye, D. W., McKinney, a. M., Colautti, R. I. & Mitchell-Olds, T. Phenotypic plasticity and adaptive evolution contribute to advancing flowering phenology in response to climate change. *Proceedings of the Royal Society B: Biological Sciences* **279,** 3843–3852 (2012).
8. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15,** 1566–1575 (2005).
9. Horton, M. W. *et al.* Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.* **44,** 212–216 (2012).
10. Bonhomme, M. *et al.* Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics* (2010). doi:10.1534/genetics.110.117275
11. Exposito-Alonso, M. *et al.* Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. *Nat Ecol Evol* **2,** 352–358 (2018).
12. Bay, R. A. *et al.* Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science* **359,** 83–86 (2018).
13. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185,** 1411–1423 (2010).
14. Hancock, A. M. *et al.* Adaptation to climate across the Arabidopsis thaliana genome. *Science* **334,** 83–86 (2011).
15. Lasky, J. R. *et al.* Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate. *Mol. Ecol.* **21,** 5512–5529 (2012).
16. Fitzpatrick, M. C. & Keller, S. R. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecol. Lett.* **18,** 1–16 (2015).
17. Kingsolver, J. G. *et al.* The strength of phenotypic selection in natural populations. *Am. Nat.* **157,** 245–261 (2001).
18. Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nat. Rev. Genet.* **14,** 807–820 (2013).
19. Wang, T., O'Neill, G. A. & Aitken, S. N. Integrating environmental and genetic effects to predict responses of tree populations to climate. *Ecol. Appl.* **20,** 153–163 (2010).
20. Gompert, Z. *et al.* Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* **17,** 369–379 (2014).
21. Anderson, J. T., Lee, C.-R. & Mitchell-Olds, T. Strong selection genome-wide enhances fitness

trade-offs across environments and episodes of selection. *Evolution* **68,** 16–31 (2014).
22. Price, N. *et al.* Combining population genomics and fitness QTLs to identify the genetics of local adaptation in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 201719998 (2018). doi:10.1073/pnas.1719998115
23. Hereford, J. A quantitative survey of local adaptation and fitness trade-offs. *Am. Nat.* **173,** 579–588 (2009).
24. Leimu, R. & Fischer, M. A meta-analysis of local adaptation in plants. *PLoS One* **3,** e4010 (2008).
25. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166,** 481–491 (2016).
26. Kojima, K. & Lewontin, R. C. Evolutionary Significance of Linkage and Epistasis. in *Mathematical Topics in Population Genetics* 367–388 (Springer, Berlin, Heidelberg, 1970). doi:10.1007/978-3-642-46244-3_12
27. Gompert, Z., Egan, S. P., Barrett, R. D. H., Feder, J. L. & Nosil, P. Multilocus approaches for the measurement of selection on correlated genetic loci. *Mol. Ecol.* **26,** 365–382 (2017).
28. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11,** 407–409 (2014).
29. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet.* **10,** e1004412–e1004412 (2014).
30. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20,** R208–15 (2010).
31. Wittmann, M. J., Bergland, A. O., Feldman, M. W., Schmidt, P. S. & Petrov, D. A. Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E9932–E9941 (2017).
32. Haldane, J. B. S. The Cost of Natural Selection. *Genetics* **55,** 511–524 (1957).
33. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9,** e1003264 (2013).
34. Charlesworth, B. The effects of deleterious mutations on evolution at linked sites. *Genetics* **190,** 5–22 (2012).
35. DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* **32,** 1895–1897 (2016).
36. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).
37. Fournier-Level, A. *et al.* A map of local adaptation in Arabidopsis thaliana. *Science* **334,** 86–89 (2011).
38. Manzano-Piedras, E., Marcer, A., Alonso-Blanco, C. & Picó, F. X. Deciphering the adjustment between environment and life history in annuals: lessons from a geographically-explicit approach in Arabidopsis thaliana. *PLoS One* **9,** e87836–e87836 (2014).
39. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Sci Data* **5,** 170191 (2018).
40. Levene, H. Genetic Equilibrium When More Than One Ecological Niche is Available. *Am. Nat.* **87,** 331–333 (1953).
41. Bell, G. Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365,** 87–97 (2010).
42. Venable, D. L. Bet hedging in a guild of desert annuals. *Ecology* **88,** 1086–1090 (2007).
43. Henry, R. C., Bartoń, K. A. & Travis, J. M. J. Mutation accumulation and the formation of range limits. *Biol. Lett.* **11,** 20140871 (2015).
44. Lee-Yaw, J. A. *et al.* A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. *Ecol. Lett.* **19,** 710–722 (2016).
45. Giorgi, F., Bi, X. & Pal, J. Mean, interannual variability and trends in a regional climate change experiment over Europe. II: climate change scenarios (2071–2100). *Clim. Dyn.* **23,** 839–858

(2004).
46. Samaniego, L. *et al.* Anthropogenic warming exacerbates European soil moisture droughts. *Nat. Clim. Chang.* **8,** 421–426 (2018).
47. Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T. & Prentice, I. C. Climate change threats to plant diversity in Europe. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 8245–8250 (2005).
48. Aitken, S. N. & Bemmels, J. B. Time to get moving: assisted gene flow of forest trees. *Evol. Appl.* **9,** 271–290 (2016).
49. Shakoor, N., Lee, S. & Mockler, T. C. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr. Opin. Plant Biol.* **38,** 184–192 (2017).
50. Asner, G. P., Nepstad, D., Cardinot, G. & Ray, D. Drought stress and carbon uptake in an Amazon forest measured with spaceborne imaging spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6039–6044 (2004).

Supplemental Information Guide for

Exposito-Alonso et al.: **A map of climate change-driven natural selection in *Arabidopsis thaliana***

Table of Content

SUPPLEMENTAL APPENDIX I: Extended statistical methods for "A map of climate-change driven natural selection in *Arabidopsis thaliana*"

Moises Exposito-Alonso[1], 500 Genomes Field Experiment Team[2], Hernán A. Burbano[3], Oliver Bossdorf[4], Rasmus Nielsen[5], Detlef Weigel[1*]

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [2] See author contributions section. [3]Research Group of Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [4]Institute of Evolution and Ecology, University of Tübingen, 72076 Tübingen, Germany. [5]Departments of Integrative Biology and Statistics, University of California Berkeley, Berkeley, CA 94720, USA. Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 København K, Denmark.

I. 1001 Genomes Project data

We used VCFtools v.0.1.12b (ref. [51]) to subset and filter the 1001 Genomes VCFv4.1 (available at: http://1001genomes.org/data/GMI-MPI/releases/v3.1/). We used vcftools with the flags: --maf 0.01 --max-alleles 2 --min-alleles 2 --max-missing 0.95. The resulting high-quality dataset was a genome matrix of 515 individuals by 1,353,386 variants for which we did not impute the small number of missing data points.

We annotated the 1001 Genomes VCF using the package SnpEff 4.3p (ref. [52]). We then manually curated a set of eight categories of variants: intergenic, intron, UTR3, UTR5, exon, synonymous, nonsynonymous, exon noncoding.

II. F_{st} and selective sweep signatures from polymorphism data

We used the genetic groups previously defined for the same accessions[11] and computed F_{ST} using PLINK version 1.9 (ref. [53]). We also used PLINK to calculate π and Tajima's D using PLINK in windows of 100 SNPs across the genome.

We used SweepFinder2 (ref. [35]) to scan the genome for deviations of the Site Frequency Spectrum (SFS) that might be caused by selective sweeps. We used all 11,769,920 biallelic SNPs from the 1001 Genomes Project (without the filters of 1% MAF and maximum missing data of 5%, which were applied to generate the variants used in the GWA [see section V]).

II.1 Geographic proxies of diversity metrics

In order to estimate π and a proxy of Tajima's D at a regional scale, we used the 4 closest neighbouring accessions in our set (same patterns were observed with different sets of neighbours within a geographic area of 5° latitude-longitude radius), and computed the total number of polymorphisms P in the subset and the sum of all pairwise Hamming differences, H. Then we calculated θ, π and the proxy of D as:

$$\pi = \frac{H}{6 \times G} \frac{N_{full}}{N_{all}}$$

$$\theta = \frac{P}{1.8666 \times G} \frac{N_{full}}{N_{all}}$$

$$\hat{D} = \pi - \theta$$

Where G is the genome size, N_{full} are all SNPs with full information that were used to count polymorphisms and distances, and N_{all} are all SNPs of the genome matrix. In the denominators, 6 is the number of pairwise comparisons of four genomes, and 1.8666 is the harmonic number of 4. Although D is normally divided by the standard error, we only wanted to rank our natural lines so we used the difference between π and θ as a proxy of D.

III. Heritability of fitness

To estimate how much variance in fitness is related to the genotypes of the lines, we used generalized linear mixed models using the R package MCMCglmm (ref. [54]). We used fitness estimates per replicate and, appart from including the natural line ID, we controlled for block (growing tray) and position within the block (longitudinal, latitudinal, and the interaction). As this is a Bayesian approach, we used flat priors, we used 10,000 MCMC steps, a burnin of 10%, and confirmed that this was sufficient for convergence of the chain. For survival proportion we used a Binomial link, for number of seeds we used a Poisson link, and for the combined lifetime relative fitness we used a Gaussian link. The mode and 95% Highest Posterior Density of the posterior distribution of each random effect were extracted (Table SI.3).

IV. The special case of fitness GWA and the consequences of natural selection in allele frequencies

IV.1 Natural selection on correlated genotypes

Genome-Wide Association (GWA) approaches were first used in the quantitative genetics field to study common human diseases, with the main aim to determine a limited number of most important loci that ultimately would have clinical or other utility — strongly favouring true positives and neglecting false negatives[55].

Genome-Wide Association (GWA) approaches, which were first developed by quantitative geneticists to identify loci responsible for human diseases[55], have been applied in recent years in other disciplines such as functional ecological genomics. This, for example, helped identifying important loci controlling ecologically relevant traits such as animal coat color or flowering time in plants[56,57]. For a number of organisms, it is also possible to directly measure lifetime fitness of an individual. This opens an opportunity of linking quantitative genetics, where the focus is the identification of the most important genetic players of a phenotype, with population genetics, where the focus is understanding how populations genetically change over time. The link here is because variation in fitness cannot be treated as a nuisance of developmental noise, but, if heritable, it all has consequences in changing the allele frequencies of the population. Recently, in a seminal paper, Gompert and colleagues [27] discussed how one can borrow methodological advances in linear model and statistical software used in GWA to carry out genome-wide scans of selection when individual fitness is available, and discuss thoroughly the subtle differences in interpretation of GWA estimates in contrast to phenotypic GWA effects.

In their paper, Gompert et al. begin by comparing fitness—SNPs with fitness—phenotype associations, as a large body of theory already exists to understand the different aspects of phenotypic natural selection. The most used approach to quantify natural selection on multiple phenotypes comes from Arnold & Lande's classic Evolution paper "The Measurement of Selection on Correlated Characters"[58]. From their manuscript, the formulation of total selection over a trait zi is represented by: $s = Cov[w, zi]$; where w is the relative fitness (absolute lifetime fitness divided by the mean fitness of the population). Because other phenotypic traits can covary with zi, one cannot be sure from this approach that all the selection experienced by zi is from direct effects, but a sum of indirects effects from n other traits: $s = \sum_{j=1}^{n} Cov[zi, zj]\beta_j$; where β represents the direct selection and can be calculated as $\beta = P^1 s$, where P is the n-dimensional variance-covariance

matrix. Gompert et al discuss that the same approach can be applied to genetic loci instead of phenotypes. This framework is interesting, as it is often the case that alleles of multiple SNPs in the genome are also correlated, that is, they are in linkage disequilibrium.

Gompert et al. discuss that different GWA models might also allow to calculate total and direct effects of selection over SNPs (although this requires solving some additional problems as the large amount of SNPs to compute associations with, which can be solved with advanced GWA approaches). We can consider the most simple case, where there are two SNPs, x_1 and x_2. For mathematical convenience we assume that the response variable fitness, y, as well as the predictors, are mean centered and variance scaled. From the univariate approach, where the effect of a SNP x_1 is estimated marginally or independently from x_2, the total effect in selection would be:

$$\beta_{x_1} = \frac{cov(x_1, y)}{var(x_1)}$$

The same calculation would be repeated for SNP x_2. In a multivariate regression framework, the regression coefficient, called conditional or partial coefficient, β^*, is corrected by the correlative indirect effect of the other predictor, $r_{x_1 x_2}$. In this way, effects driven by linkage disequilibrium to another loci are removed from β^*, so that only direct effects are measured. The formula would be as:

$$\beta^*_{x_1} = \frac{\beta_{x_1} - r_{x_1 x_2} \times \beta_{x_2}}{\sqrt{(1 - \beta^2_{x_1})(1 - \beta^2_{x_2})}}$$

Sensu Gompert et al., β would capture the “total selection” and thus can be called a “total selection coefficient”, as it is in essence $$\beta ~ s = w_11 - w_00$$, where $$w_11$$ and $$w_00$$ represent the true (noise-free) fitness of a plant carrying an alternative allele at the giving SNP. On the other hand, β^* corrects out indirect (or linked) effects, thus can be called “direct selection coefficient”.

Gompert et al. use the statistical GWA package GEMMA (ref. [33]), to being able to compute the above coefficients efficiently for thousands to millions of SNPs.

GEMMA implements a single-marker marginal linear models GWA (LM) of the form: $y = \mu + \beta_i x_i + \epsilon$; This provided us with allele effects, β, on relative fitness per SNP. We also run in GEMMA a Bayesian Sparse Linear Mixed model GWA (BSLMM), to calculate direct accurately pinpoint casual positions. This model accommodates both poly- and oligogenic architectures and by jointly fitting all SNPs (n=1,353,386) it statistically corrects for LD arising from population structure

and/or low recombination. It models two effect hyperparameters, a basal effect, α, that captures the fact that many SNPs contribute to the phenotype, and an extra effect, β, that captures the stronger effect of only a subset of SNPs. An internal parameter measuring the probability of having another extra effect, γ, can be used to prioritize SNPs. In BSLMM the overall effect of an allele is $= \alpha + \gamma\beta$, which in the simplistic example above corresponds to the β^*. The full model specification is:

$$y = 1_n\mu + X\beta + X\alpha + \epsilon;$$
$$\beta_i \sim \pi N(0, \sigma_a^2\tau^{-1}) + (1-\pi)\delta_0 ;$$
$$\alpha_i \sim N(0, \sigma_b^2/(p\tau));$$
$$\epsilon \sim MVN_n(0, \tau^{-1}I_n) .$$

The BSLMM model is also used to calculate the proportion of variance explained (PVE or 'chip heritability'). To do this, we used the last 1,000 samples of the MCMC chain and calculated the 95% Highest Posterior Density Interval (95% HPD), for which we report the median and the 2.5% and 97.5% percentiles. The BSLMM model is an improvement over the classic GBLUP or kinship-based (population structure correction) GWA, a form of linear mixed model where one corrects-out population structure or general relatedness between individuals by having the u random effect term with a given covariance structure that is the kinship matrix K, therefore the calculated β is conditioned on genomic background effects:

$$y = 1_n\mu + X\beta + Zu + \epsilon;$$
$$u \sim MVN(0, \sigma_a^2K)$$

The two estimates above, the direct (or conditional) effect β^*, and the total effect β (or s), provide thus differently useful insights on the nature of selection. As already argued in Gompert et al. and othres, it is the total selection coefficient, s, that best predicts the change in population allele frequency in one generation as a response to selection. We show this with simulations in the next section.

IV.2 Proof-of-concept simulations on the importance of total selection coefficients

To illustrate the differences of GBLUP (population structure corrected) GWA or marginal GWA when using fitness, we carried out simulations (Fig. SI.15) (for step-by-step code and intermediate plots see https://github.com/MoisesExpositoAlonso/selectioncorrelatedgenotypes).

We began by subsetting our dataset of 515 genomes of *A. thaliana* to 1000 SNPs from chromosome 1 (to keep intact the linkage structure. Note: results also hold simulating a genome matrix with random linkage). We then simulated 1000 selection coefficients following a Normal distribution with mean zero and standard deviation 0.1. To get the fitness of a plant of genotype j, we sum selection coefficients along the genome as: $\sum_{i=1}^{1000} s_i x_{ij}$; where x_{ij} would indicate whether the haplotype has the reference (0) or alternative allele (1) in the given i SNP (we generated some artificial noise so heritability would be 0.9; conclusions hold the same with intermediate heritability). We then inferred total selection coefficients using marginal GWA and direct selection coefficients using GBLUP GWAs. Our results comparing true and estimated effects show how even when we attempt to estimate true (direct) selection coefficients by the GBLUP method that tries to correct by background effects, we largely fail (Fig. SI.15A). It is also important to notice that the estimates from GBLUP GWA are one order of magnitude smaller than the true values. Because the architecture of fitness here is rather polygenic and SNPs are in linkage, most of the fitness variation is assigned to the kinship term in the GBLUP GWA rather than to specific SNPs. In agreement with this the kinship-based random effect accumulates 99% of the true heritability ($Vg/Vg + Ve$ = 0.89). In our field experiment, we also saw high values of kinship-based heritability (our median h^2 was 0.7; Table SI.2).

As discussed earlier, if interest is in the consequences of natural selection in allele frequencies rather than true (direct) selection coefficients, total selection coefficient are the adequate approach. To show this, we run a individual-based simulation, drawing genotypes proportionally to their relative fitness to generate the population of offspring one generation after selection (with constant population size). We then compared the change in frequency in the simulated population with the marginal GWA and GBLUP GWA estimates. Because allele frequency changes are driven by both direct and indirect selection pressures, the marginal GWA estimates correlate best with the changes of frequency in one generation (Fig. SI.15B). In fact, we can try to predict directly the change in frequency in one generation (Δq) if we not only use inferrred selection coefficients but also the original starting frequency, as the change is proportional also on how frequent is an allele originally: $\Delta q = s(1-p)p$. Plugging in the marginal GWA and GBLUP GWA estimates into the equation we show that marginal GWA estimates allow prediction of allele frequency change with accuracy of R^2=0.97, while GBLUP GWA estimates perform poorly, R^2=0.08 (Fig. SI.15C).

A theoretical concern of studying total selection coefficients rather than direct selection coefficients is that the first are thought to be contingent on the specific allele frequency and linkage

structure of the population of analyzed; something that is allegedly ameliorated if one corrects by population structure with GBLUP GWA. Therefore, it is common to think that GBLUP GWA estimates are more informative if one aims to extrapolate findings across populations. Purposedly, our experimental population of 515 accessions (subset of the 1001 Genomes Project) aimed to maximize geographic as well as genetic coverage of the species so interpolations to smaller, less diverse subpopulations, would be possible.

We then studied to what degree measurements of natural selection hold across populations, we selected 50 Spanish genotypes out of the 515. Spanish are known to belong to very distinct lineages. We then run the one-generation individual-based simulation to compute allele frequency changes in the 50 Spanish accessions. We then plugged again the marginal and GBLUP GWA estimates calculated in the 515 genomes into the equation $\Delta q = s(1-p)p$ and correlated them with the simulations (Fig. SI.15C). We showed that marginal GWA effects calculated in the 515 genomes panel also predict well the frequency changes after selection in the 50 genomes subset (R^2=0.85; for GBLUP GWA R^2=0.04) (We also confirmed our conclusion with other subset populations, such as 10 Spanish accessions or 10 low diverse USA accessions).

IV.3 Further notes on the interpretation of population structure in wild species

Much of the discussion in the previous section is about the special interpretation of fitness marginal GWA estimates as total selection coefficients, which have the property of predicting allele frequency changes. Oppositely, there is not much insight gained from marginal GWA estimates for phenotypes, where we advocate some type of population structure or linkage disequilibrium aware GWA approach (e.g. refs. [11,59]). Nevertheless, it is worth pointing out that there are two other reasons in favour of applying kinship-correction in GWA studies in humans and breeding with no clear analogy or interest for wild species:

1) Human genome-wide association are based on data collected typically in different countries, where individuals were born and rise in different health systems, cultures, and or other environmental inputs. Because those inputs are correlated with spatial location and genetic ancestry of human populations, the population structure correction also corrects for this structured environmental confounder. In experimentally-tractable wild species, such confounders are avoided by controlled experiments and replication; otherwise, if data comes directly from field observations as in humans, population structure correction is paramount.

2) In crop and animal breeding, correcting for family or population structure in associations is also highly interesting to avoid pleiotropic effects of markers during marker-assisted breeding. As in clinical cases, the aim for GWA here is to prioritize a small subset of true positive, unlinked markers. Because different breeds or varieties might differ in a variety of traits, genetic variants associated to breed or variety background are not reliable for breeding. In wild species, this case might apply if one wants to follow up a GWA hit for molecular characterization and genetic engineering with no side-effect phenotypes.

In wild species, there might be cases when population history and differentiation coincides with historical events of adaptation. This can challenge the use of population structure correction to correctly identify valuable SNPs, as much of the true positive SNPs will be positively correlated with historic population lineages and thus a lot (if not most) of the important variation will be assigned to the kinship factor. In those cases, it is up to the researcher to decide what is the best method in a case-specific manner. Some approaches such as the BSLMM approach tries to solve background effect confounders more elegantly than the kinship approach. Another potentially useful approach when a trait is expected to be directly associated with a population ancestry is admixture mapping[11,60].

IV.4 Trade-offs of selection

IV.1.1 Across field experiments

In order to test the two most prevalent hypothesis of local adaptation driven by selection trade off, conditional neutrality vs antagonistic pleiotropy[21,37,61], we do pairwise comparisons of total selection coefficients in two environments. We devised two tests: The first test discriminates between pleiotropy (selected either in the same direction or in different directions) and conditional neutrality (only selected in one environment). We use the extreme 5% selection coefficients at each tail, similar following Anderson et al. [61] to generate the contingency table:

Not selected	Selected in Spain (5% left and 5% right tails)
Selected in Germany (5% left and 5% right tails)	Selected in both environments (both 5% left and 5% right tails)

Because this test does not distinguish between those pleiotropic variants that are selected in opposite directions (antagonistic pleiotropy) or in the same direction (non-antagonistic or synergistic

pleiotropy), we do another test only for the direction of those variants selected in both environments:

Negative in Spain Negative in Germany	Positive in Spain Negative in Germany
Negative in Spain Positive in Germany	Positive in Spain Positive in Germany

We report the Odds Ratio for both tests (Table SI.5) as well as the Spearman's rho correlation between each pair of environments (Table SI.6).

IV.1.1 Across life history stages

Calculating total selection coefficients for survival and fecundity separately, we found no correlation between survival-only and fecundity-only estimates (r<0.07, Fig. SI.4-5), consistent with different stages of a plant being differentially affected by environmentally imposed selection[6263].

IV.5 Intensity of selection

The distribution of absolute total selection coefficients, $|s|$, has a shape resembling that of an exponential function. We calculated the expected rate using Maximum Likelihood optimization in R, which can also can be approximated as the inverse of the mean:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

For this, we use $\hat{\lambda}$ or the mean of $|s|$ as a metric of the overall intensity of selection (Fig. 1B, Fig. 3D).

V. Climate Genome-Wide Association

Similarly to our GWA with relative fitness, we run a GWA with each climate variable m (see Section VII.1) as response variable y_m in a LM model using GEMMA (ref. [33], see Section V):

$$y_m = \mu + \beta_i x_i + \epsilon;$$

This β coefficient for SNP i, which reflects the correlation of the alternative allele's presence and a climate variable, was used later in our predictive models (Section VIII). As this is a raw correlation between allele presence and climate variables, it will capture both past signatures of climate adaptation and historic population migration and differentiation, and is only used to capture how environmentally separated are typically found the two alleles of a SNP.

VI. Climate and modeling

VI.1. Climate layers

We used the classic bioclim variables (n=19), plus monthly data of minimum and maximum temperature, and precipitation (n=12 x 3) (worldclim.org). From these we estimated monthly evapotranspiration rates using the R package EcoHydRology v. 0.4.12 (ref. [64]) and actual monthly evapotranspiration using a bucket model [65] (n=12 x 2). Based on ref. [15] we calculated whether *A. thaliana* can grow in a given month based on temperature and precipitation (n=12), and derived from this the length of the potential growing season (n=1). Over the potential growing season, we calculated minimum and maximum temperature, and total precipitation (n=3). Finally, using the mean and variance flowering time (=lifespan) across all our field experiments per accession, and based on their climate of origin using the above variables, we used an environmental niche model to generate a map surface of the most likely plant lifespan (n=2). This provides an estimate of the actual growing season, which we subtracted from the potential growing season to generate one more composite variable (n=1). Each variable is further described in Table SI.9. A total of 98 raster layers are available as .gri/.grd files (native R format) from: github.com/MoisesExpositoAlonso/araenv, with doi:.

VI.2. Environmental Niche Models

Genome-wide Environmental Niche Models (GEMs) were fit using decision trees with presence/absence of SNPs as response variable and the climate variables described in the previous section and latitude and longitude as predictors as described previously[11]. To fit the models we used an Stochastic Gradient Boosting approach with the R package caret (ref. [66]). The parameters used to fit the model were: 50 decision trees, an interaction depth of 2, a shrinkage of 0.1, and a minimum of observations at end nodes of 10. This set of parameters was determined after running our GEMs for some exemplary SNPs and confirming that this set was typically optimal for reducing residual-mean squared error in a Repeated Cross-Validation approach.

We used these models to predict from raster maps of the climate layers a probability between 0 and 1 that the alternative allele was in a map cell. We judge this as a more appropriate output than a discrete 0/1 outcomes, as sometimes alleles were widespread or at intermediate frequencies in many regions and thus their environment niche was not strictly defined.

Areas outside the high-density areas of *A. thaliana* (Fig S.I.1) were excluded from the GEM training and projections, as our information of populations, for instance, from Siberia is limited.

Nevertheless, the few samples there had a relatively high fitness in Spain and low precipitation (Dataset 3).

VI.3. Climate variability

To study spatial climate variability, for each *A. thaliana* natural line, we extracted the 19 bioclim variables (Table SI.9) in a 50 Km buffer where they were originally collected from and calculated the coefficient of variation (CV) across grid cells.

To study temporal variability, we used climate data[39] from 1958-2017 to calculate annual precipitation values for each population, from which we in turn derived the inter-annual CV..

VII. Predictions of total selection coefficients from sequence features and environment of origin

VII.1 The model

We used a decision tree approach with Random Forest using the R package randomForest (ref. [67,68]) to predict the vector (n=1,353,386) of GWA results with relative fitness in one environment, which we call total selection coefficients s, from a 1,353,386 x 98 matrix of GWA associations with climate variables, β_{clim} (Table SI.9, section V). We also included as predictors a 1,353,386 x 5 matrix, μ, of genetic diversity and frequency metrics: minimum allele frequency, π diversity, Tajima's D, selective sweep likelihood ratio, and selective sweep alpha value (section II). In addition, we included as predictors a 1,353,386 x 8 matrix θ of non mutually exclusive variables taking values of 0 or 1 indicating genomic annotations: intergenic, intron, UTR3, UTR5, exon, synonymous, nonsynonymous, exon noncoding (section I). A total of 112 variables were thus used as predictors: $s = f(\beta_{clim}, \mu, \theta)$. In the cases where we trained models with two environments, we also included the 2 x 98 x_{clim} climate variables at our field stations: $s = f(x_{clim}, \beta_{clim}, \mu, \theta)$.

VII.2 Genome-wide cross-validation

Because training a Random Forest with the full dataset would be computationally expensive, we only trained with 10,000 observations (with smaller and larger SNP sets, we had determined that training with more than 10,000 observations did not improve predictions). To test accuracy and bias we used a different set of 10,000 SNPs, divided into 100 bootstrap samples, and we report the intervals of the 95% bootstrap distribution. The results presented in Fig. 3 were produced with 10,000 randomly drawn SNPs across the genome. To confirm that there was no confounding from non-independent samples in the training and testing SNPs, we repeated all analyses, training with 10,000 random SNPs

from chromosome 1 and testing with 10,000 random SNPs from the four other chromosomes. There were no substantial changes in predictability.

Several combinations of training and testing were performed to validate the predictions of "unobserved" environments (Table SI.8).

VII.3 Note on interpretations and limitations of GWES

As in any predictive exercise, our geographic projections of intensity of selection have limitations (discussed below). We nevertheless firmly believe that they are indispensable to move forward in the field of forecasting climate impacts. Models such as ours are tremendously useful for subsequent experimental validation (as we are currently doing through an experimental evolution network: GrENE-net.org) or with *in situ* observations collected as we move into the future (e.g. iNaturalist.org, iSpot.org). This iterative prediction ⇿ validation process will be key to advancing the complex field of predicting the effects of climate change on biodiversity.

Below we discuss a list of points describing potential pitfalls of the GWES, and what is their interpretation.

A. Selection is a "relative force". The selection of an allele depends on the other alternative allele, and at what frequency both are found. Thus, the exact value of total selection coefficients might vary depending on the GWA panel. A *reductio ad absurdum* case would be that of a GWA panel where many specific positions in the genome allegedly under selection in other population, are invariant. Therefore, in such a case one could not calculate total selection coefficients for that invariant site, although that does not mean the population is not under natural selection, which could lead to extinction if only disadvantageous alleles present. As we discuss in Section IV.2 and show with simulations in Fig. SI.15D, by using a diverse reference GWA panel to calculate total selection coefficients, we can interpolate to subset populations. Therefore the GWES projections are useful for relative trends of selection in the species across its geographic range.
B. Our GWES projections are not long-term population projections.
C. Short-term total selection coefficients (over one generation, ecological times) do not necessarily reflect long-term selection coefficients (i.e. over evolutionary times), which are an integration of selection events over time.

D. Over longer timescales, immigration of genotypes, admixture, and recombination, can alter the efficiency of selection.
E. Demographic dynamics are ultimately determined both by natural selection and stochastic demographic forces (drift). Therefore, the knowledge of total selection coefficients in a generation is necessary but not sufficient to determine the fate of a population over multiple generations. To do so, explicit demographic models are needed which also take into account nuances such as bet-hedging strategies like seedbanks, and overlapping generations.
F. We used climate projections of 2050 (of different CO_2 scenarios and years) to feed into the GWES models only as proxies of plausible magnitudes of climate change. Demographic processes year to year will interact with the local gradual or stochastic changes in climate and ultimately determine the extinction or persistence of populations. A useful way to think of our climate projections from models trained in the warm edge (Spain) and the distribution center (Germany), is to think how climate change might put German populations under similar selection pressures to Spain.

VIII. Re-analysis of published data from common garden experiments

VIII.1 Environment cross-validation

In order to cross-validate our model on independent environments, we re-analyzed published data. This approach is an environmental cross-validation on top of cross-validation of SNPs. That is, we train in a subset of 10,000 SNPs in Spain and Germany, and test our model in another subset of 10,000 SNPs using the previously-published experiments of Spain, Germany and England[37,38]. For a conceptual diagram of predictability (and extrapolability) validation with external common garden experimental datasets, see Fig. SI.11. Note that a partial overlap of natural lines and genomic data is required for the following re-analysis (predictions on common gardens with recombinant inbred lines or non-overlapping natural lines would require further adjustments in our approach).

VIII.2 Manzano-Piedras et al. 2014

Manzano-Piedras and colleagues[38] planted exactly 60 seeds per line in pots. They monitored how many plants established at the rosette stage and later on became reproductive adults (survival proportion). From these, they counted the number of fruits per pot and divided them by the number of reproductive adults (reproduction, seed set). We computed lifetime fitness as the product of survival and reproduction.

VIII.3 Fournier-Level et al. 2011

Fournier-Level and colleagues[37] germinated seeds in greenhouses, and two weeks after germination (established seedling stage), they transplanted seedlings to outdoor field stations where one plant was transplanted in one pot. They counted how many transplanted seedlings survived to reproduction (partial survival proportion), and the number of fruits per plant (reproduction, seed set). We again computed lifetime fitness as the product of partial survival and reproduction.

We excluded the experiment in Finland in downstream analyses because only 58 natural lines were planted there in the original publication[37] and because later we verified the imputation accuracy was very low (Pearson's r<0.008).

VIII.4 1001 Genomes x RegMap panel phenotype imputation

The 1001 Genomes panel (http://1001genomes.org/, ref. [25]) includes 1,135 natural lines with 11,769,222 biallelic SNPs from Illumina sequencing. The RegMap panel (http://arabidopsis.gmi.oeaw.ac.at:5000/DisplayResults, ref. [9]) included 1,307 natural lines with 214,051 biallelic SNPs from array hybridization. The two populations shared 413 lines. Of these, 185 were shared with the 515 lines used in the field experiments.

Of the 157 accessions of Fournier-Level *et al.*, all were part of the RegMap panel, 89 were part of the 1001 Genomes, and 50 overlapped with our lines. Of the 279 accessions of Manzano-Piedras *et al.*, 150 were part of the 1001 Genomes, and 131 overlapped with our field lines.

Because fitness is heritable, we tried to impute missing data based on the overall genomic relationships among all of the 2,029 natural lines belonging to 1001 Genomes and RegMap panels. After downloading and transforming the RegMap dataset to PLINK format, we overlapped genome-wide SNPs and filtered them for a genotyping rate of 95%, which yielded 154,090 biallelic SNPs. Given the linkage disequilibrium and genome size of *A. thaliana*, this easily suffices for generating a relationship matrix A (related to a kinship matrix), which we computed using the R package rrBLUP (ref. [69]). The data of survival, reproduction, and lifetime fitness was an average per genotype, so we fit a classic GBLUP: $y = Zg + \epsilon$; where y is the fitness trait of interest, Z is a design matrix of genotypes and g is a random effect factor with covariance matrix equal to the relationship matrix $g \sim MVN(0, A\sigma_g^2)$. Heritability of traits and imputation accuracy from the Manzano *et al.* and Fournier-Level *et al.* experiments is given in Table SI.10.

VIII.5 Sanity checks for imputation and geographic predictions

We carried out sanity checks to ensure that the imputed fitness from other experiments was not just an artifactual phenotype with the same structure as the relationship matrix. This would mislead us to think there is predictability, as we would expect that total selection coefficient calculated in such artifactual phenotype would depend on population structure and thus would likely be predictable from climate structure alone.

We shuffled the genotype identities from Fournier-Level *et al.* and Manzano-Piedras *et al.* with their fitness values. Then we repeated the GBLUP analysis with 50 rounds of shuffling and computed heritabilities and prediction accuracies. We confirmed that heritability with shuffled data was negligible ($1x10^{-9}<h^2<1.6^{-3}$) and so was the accuracy of imputation ($-0.047< r <0.070$). This indicated that in the absence of true heritable variation, imputation of fitness would be random and not an artifact of the relationship matrix.

We also were concerned that geographic predictions could be driven by some underlying bias in our analyses, i.e. bias inherent to geographic sampling, population history of genotypes chosen, etc. In other words, we were concerned that the null expectation of predictability would be non-zero. As before, we randomized fitness values with genotypes for all six environments (Fournier-Level *et al.*, Manzano-Piedras *et al.*, and ours). Then, we repeated the GWA to estimate total selection coefficients (as Fig. 1), and trained different combinations of GWES models to re-predict total selection coefficients at each location based on climate (as Fig. 3). We confirmed that, differently from the analyses of real data presented in Fig. 3, there was no significant predictability (Fig. SI.12).

VIII.6 An explanation for "inverse predictability"

We noticed that using only our two experiments for model training, there was "inverse predictability" for the three experiments from ref. [37]. While the sign of inferred total selection coefficients was the opposite of the observed values ($-0.33<r_{cv}<-0.51$, $P<0.001$), the magnitude of selection was correctly inferred ($15\%<R^2_{cv}<25\%$, Fig. 3A). Such a phenomenon could arise for several reasons and has already been observed in studies with evolving *Drosophila* populations where seasonal environments vary from year to year[70], as well as in timema insects[71]. In our case, the worldclim.org climate averages (1960-1990) at 2.5 arc-minutes resolution might strongly deviate from the truly experienced environmental conditions in the years the experiments were conducted.

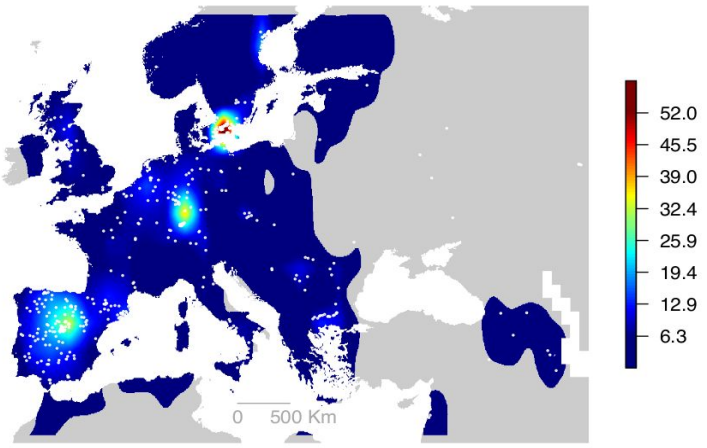
Such climate variability can exert opposite selection in different years[72]. Second, differences in experimental design could lead to different lifetime fitness estimates. In the Fournier-level *et al.* experiments, early survival of seedlings was not measured at all, as only seedlings that had survived for two weeks in the greenhouse were transplanted into the field. In the Southern Spain experiment from Manzano-Piedras *et al.* [38], seeds were sown directly in the field, as in our own experiments, and accordingly, we had "positive predictability" (r=0.24, Bootstrap CI=0.09—0.41). In further support of this experimental design confounder, when we trained GWES models with only reproduction-based total selection coefficients in our experiment of high precipitation in Southern Germany, i.e., excluding early survival from lifetime fitness, we correctly predicted the sign of total selection coefficients in Fournier's Northern Germany experiment (r=0.392, Bootstrap CI= 0.20—0.57) (for null expectations see Supplemental Methods IX.4).

The differences in predictions between two- and six-environment-trained models did not yield differences in downstream conclusions from Fig. 3 (correlation between predictions, r=0.56, $P<10^{-16}$), but predictability increased with the number of experiments included in the training set (6 environments, r= 0.746 [Bootstrap CI= 0.667— 0.800], R^2= 0.517 [0.445—0.640]).

We preferred to show geographic predictions (Fig. 3) with GWES trained with our two environments so we only rely on highly-replicated fitness estimates from over 500 accessions that were grown in carefully controlled precipitation and temperature environments.

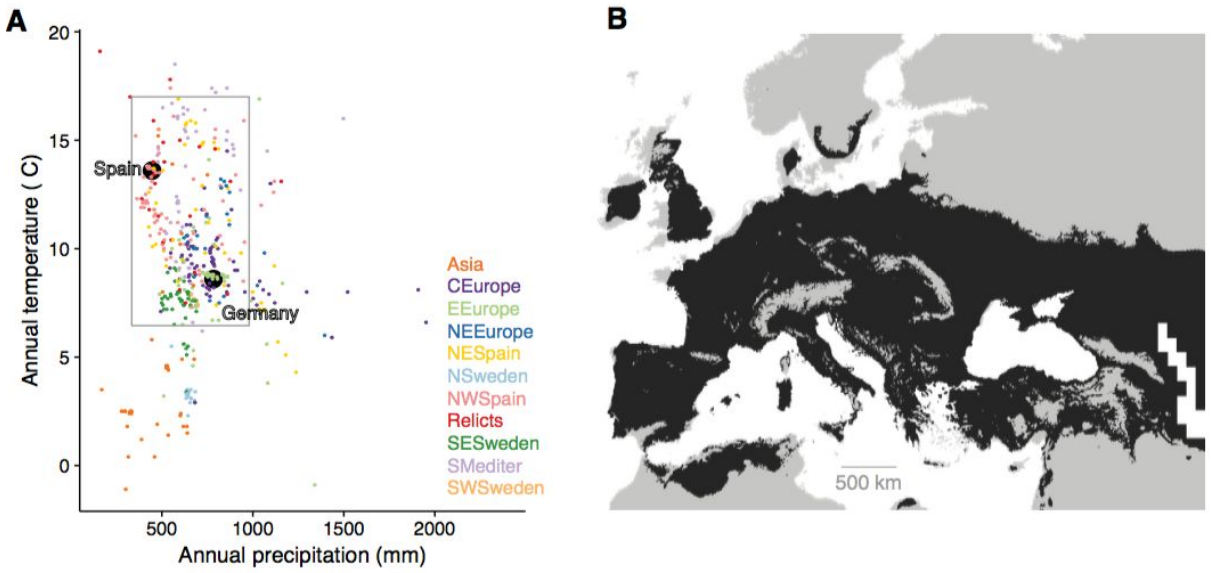
SUPPLEMENTAL FIGURES I

Figure SI.1. Map of abundance of *Arabidopsis* samples



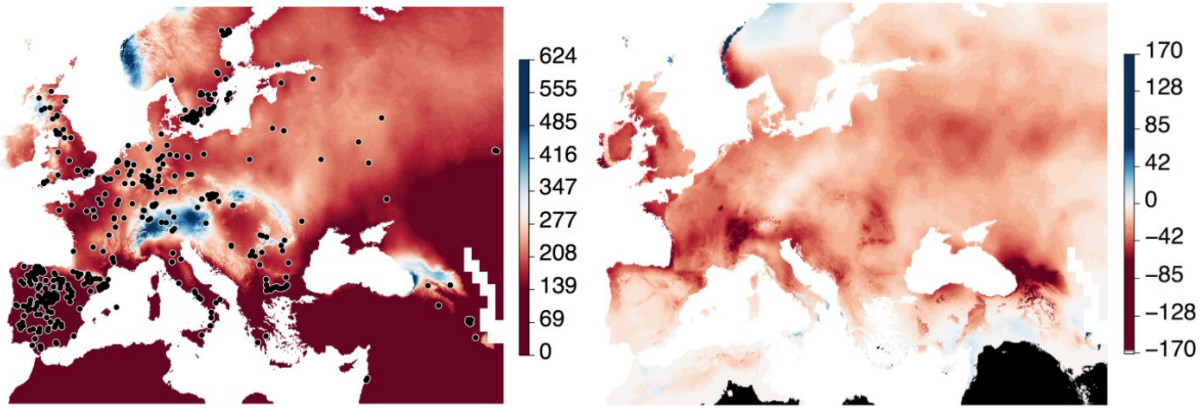
Points indicate the locations where the 517 *A. thaliana* accessions were collected. The color gradient is the density of samples from our study in squares of approximately 200 km x 200 km. The limits of the colored area were determined using a combined density grid from gbif.org and 1001 Genomes records. The density was generated in a grid of 125 min resolution and by applying a bilinear and then Gaussian smoothing. The threshold was chosen to be the 50% of the upper distribution, which roughly corresponds to 10 records per 200 km x 200 km square. Regions outside the colored were excluded from future climate change predictions, as we prefer to make predictions only in regions where the presence of *A. thaliana* is rather likely and continuous (Fig. 1).

Figure SI.2. Environment ranges



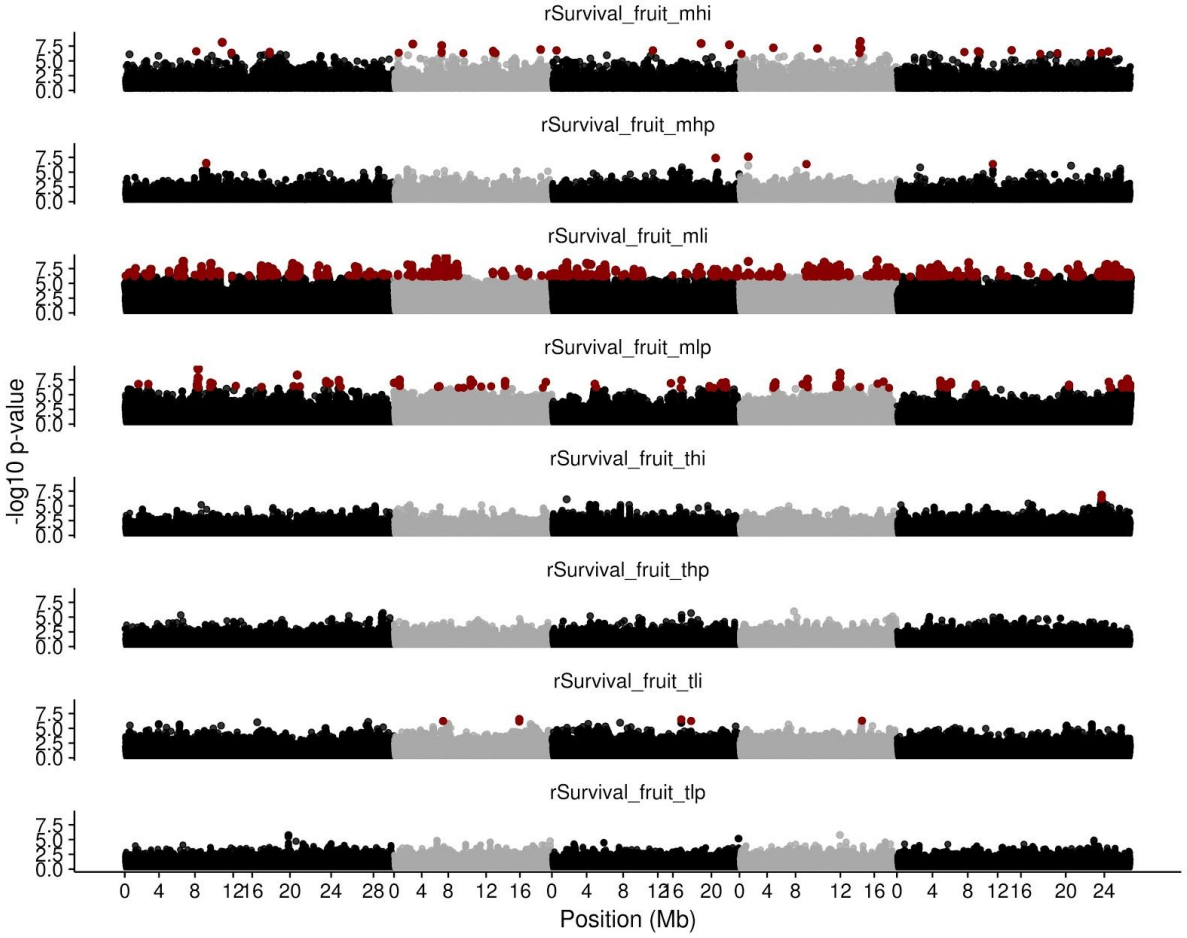
(**A**) Classic biplot of precipitation *vs.* temperature of origin of accessions (black dots) and field experiment of Spain (sepia) and Germany (green). Grey box indicates locations where precipitation was at least 70% of Spain and no more than 130% of Germany, and where temperature was no less than 70% of Germany and no more than 130% of Spain. (**B**) Areas that would be within the grey box in (A). Colored population groups based on previously calculated genetic clusters[11].

Figure SI.3. Map of predicted precipitation change



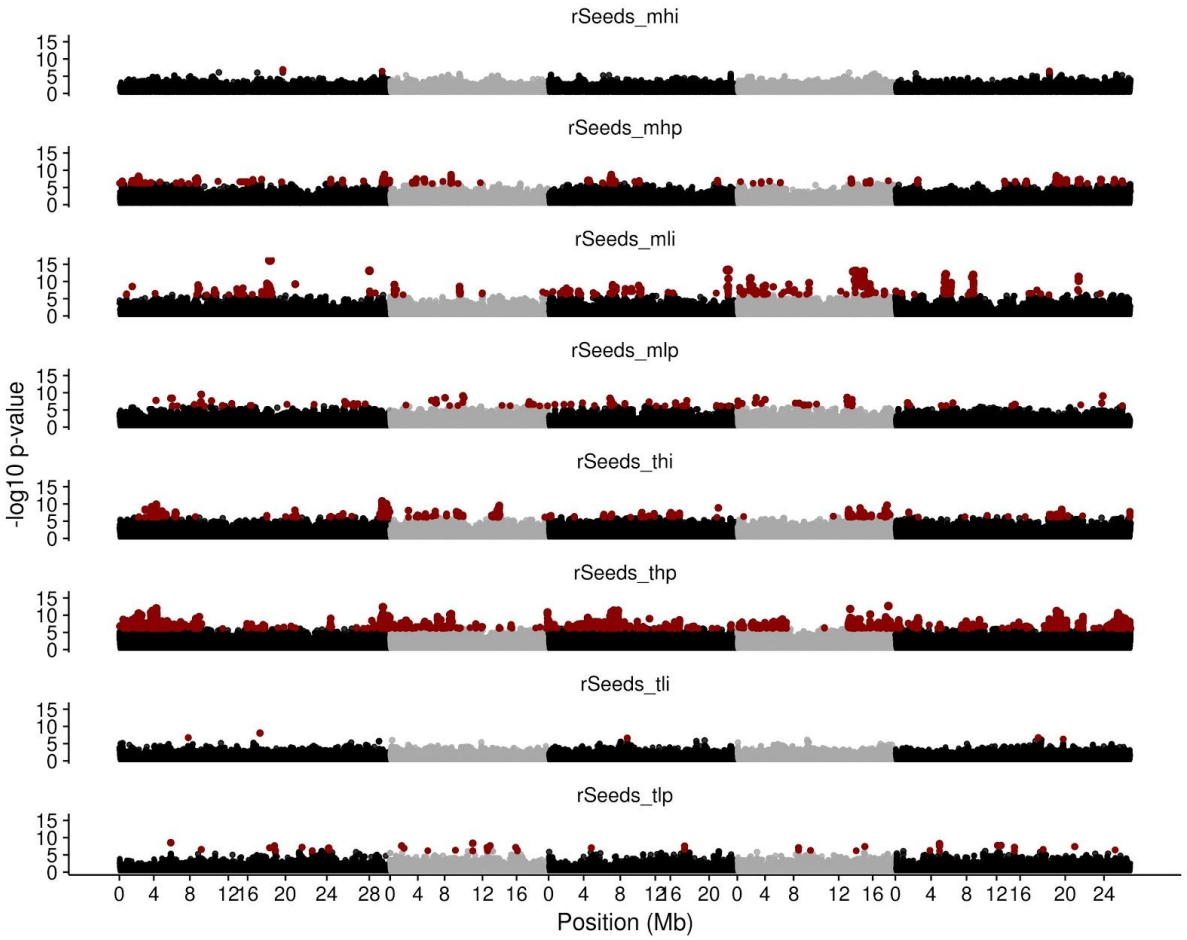
Precipitation during the warmest quarter (bio18, left), and its change predicted for 2070 (rcp 8.5) (right) (worldclim.org). Black areas indicate regions where precipitation will be lower than any area where *A. thaliana* has been currently sampled (black dots, left).

Figure SI.4. Genome maps of survival



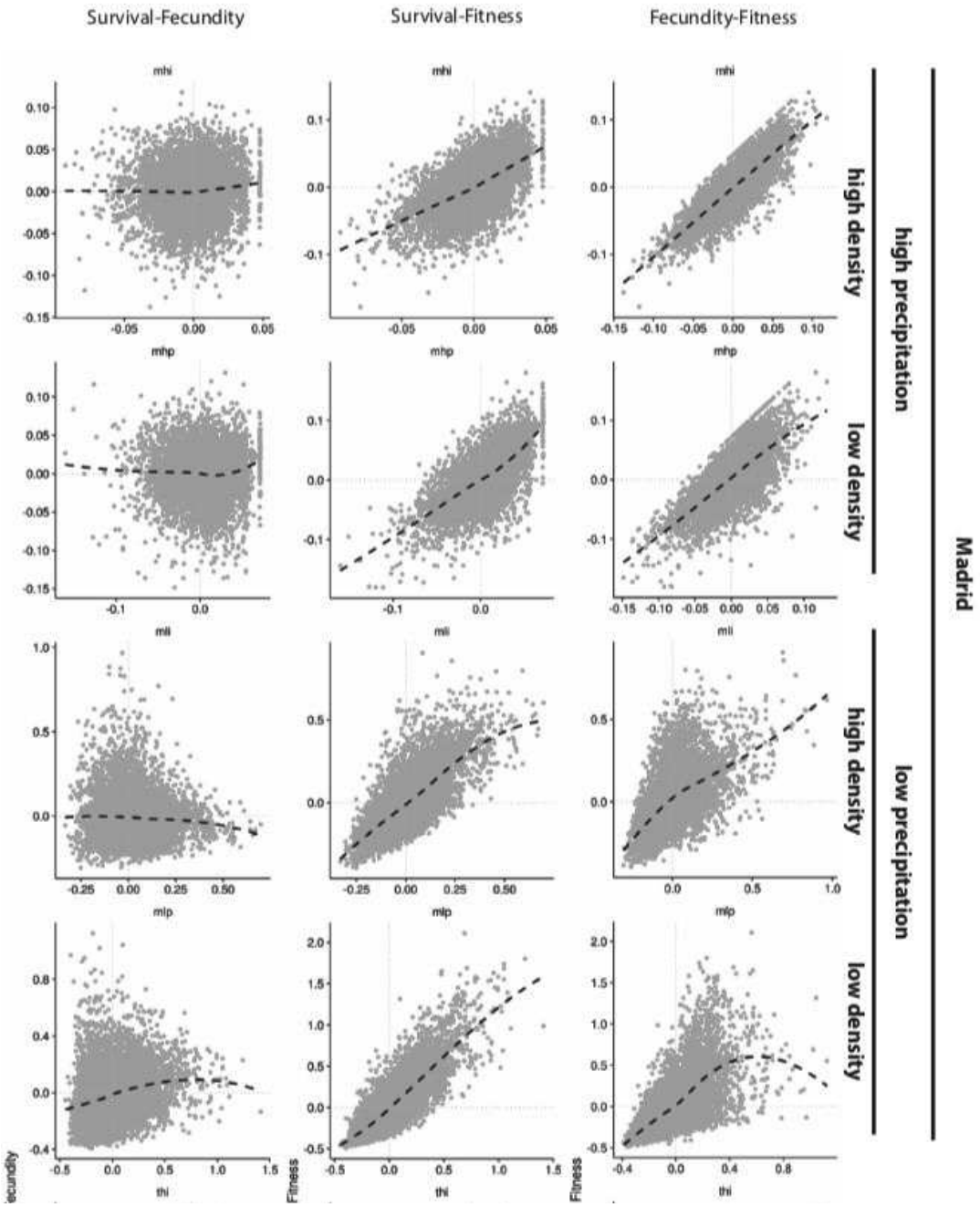
Same as Fig. 1, but only using the survival component of fitness. [Abbreviations: The three characters of the codes: MLI, MLP, MHI, MHP, TLI, TLP, THI, TLP; indicate M=Madrid (Spain), T=Tübingen (Germany), L=Low precipitation, H=High precipitation, I=Individual replicates (one plant per pot), P=Population replicates (up to 30 plants per pot)].

Figure SI.5. Genome maps of fecundity

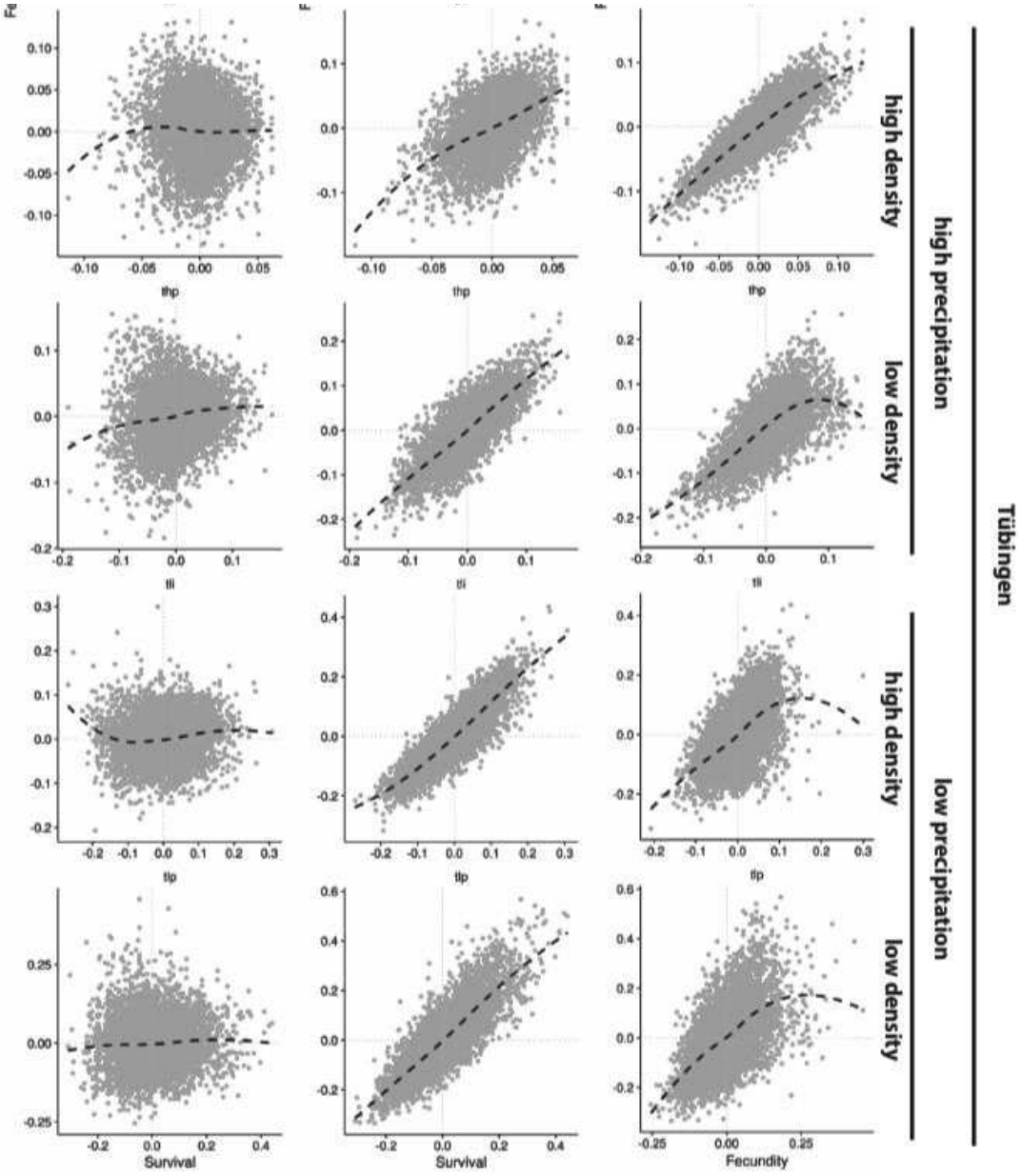


Same as Fig. 1, but only using the fecundity component of fitness. [Abbreviations: The three characters of the codes: MLI, MLP, MHI, MHP, TLI, TLP, THI, TLP; indicate M=Madrid (Spain), T=Tübingen (Germany), L=Low precipitation, H=High precipitation, I=Individual replicates (one plant per pot), P=Population replicates (up to 30 plants per pot)].

Figure SI.6. Trade-offs in survival and fecundity

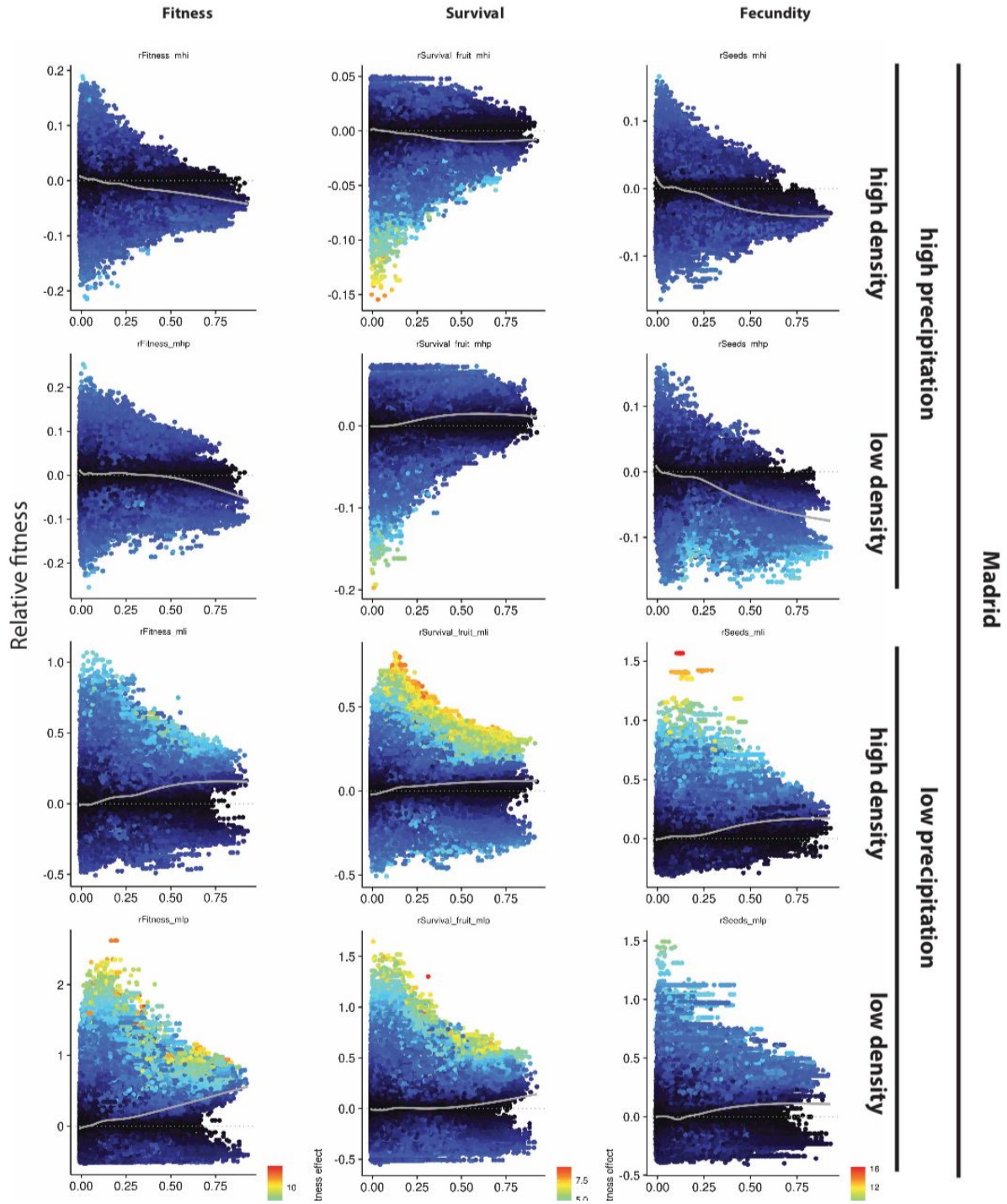


(Fig. S6 continued)

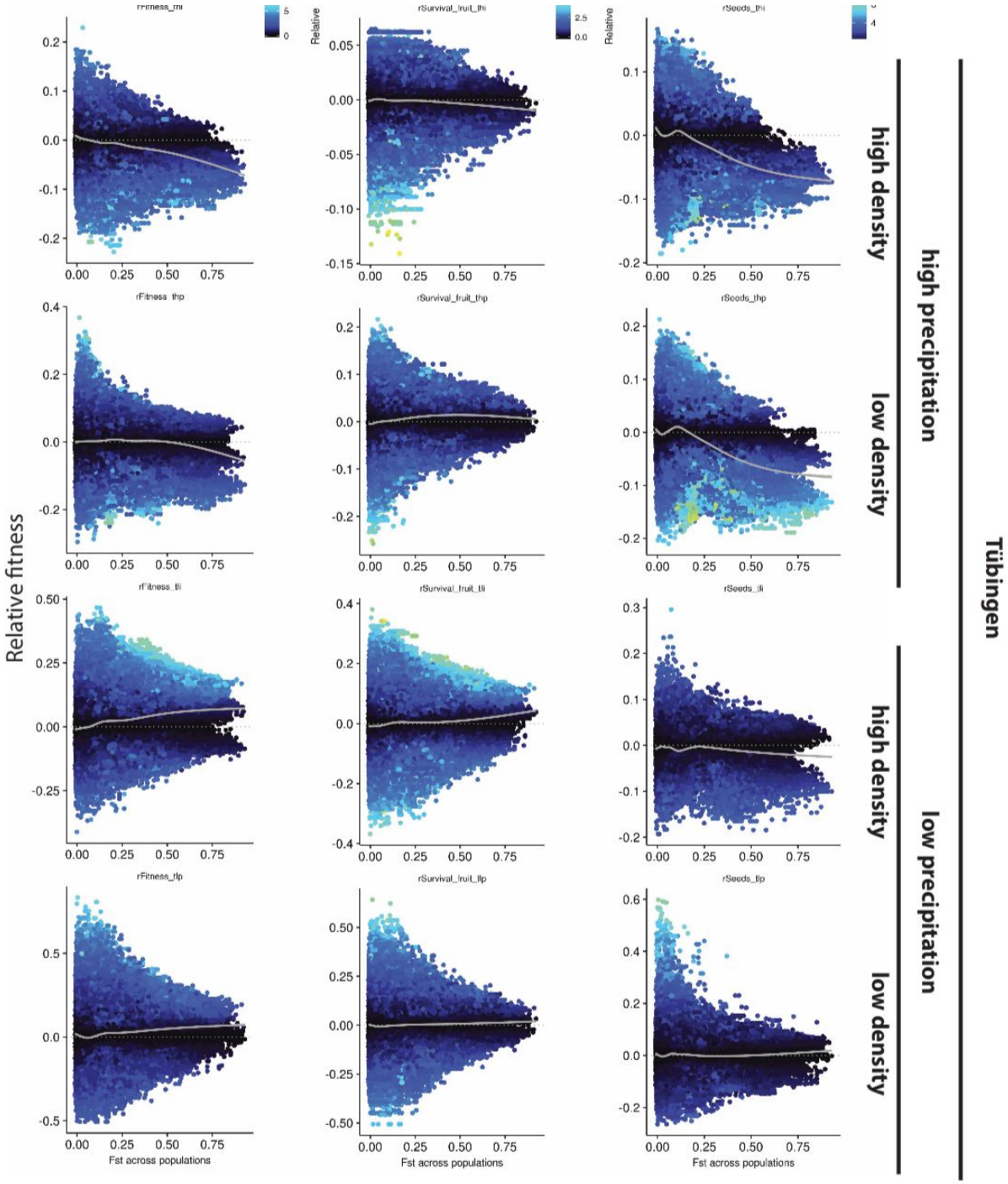


Comparisons of total selection coefficients computed only with the survival component, only with the fecundity component, and with lifetime fitness. All environment combinations are plotted: Madrid (Spain) and Tübingen (Germany), high and low precipitation treatments, and high and low plant density treatments.

Figure SI.7. F_{st} and empirical selection

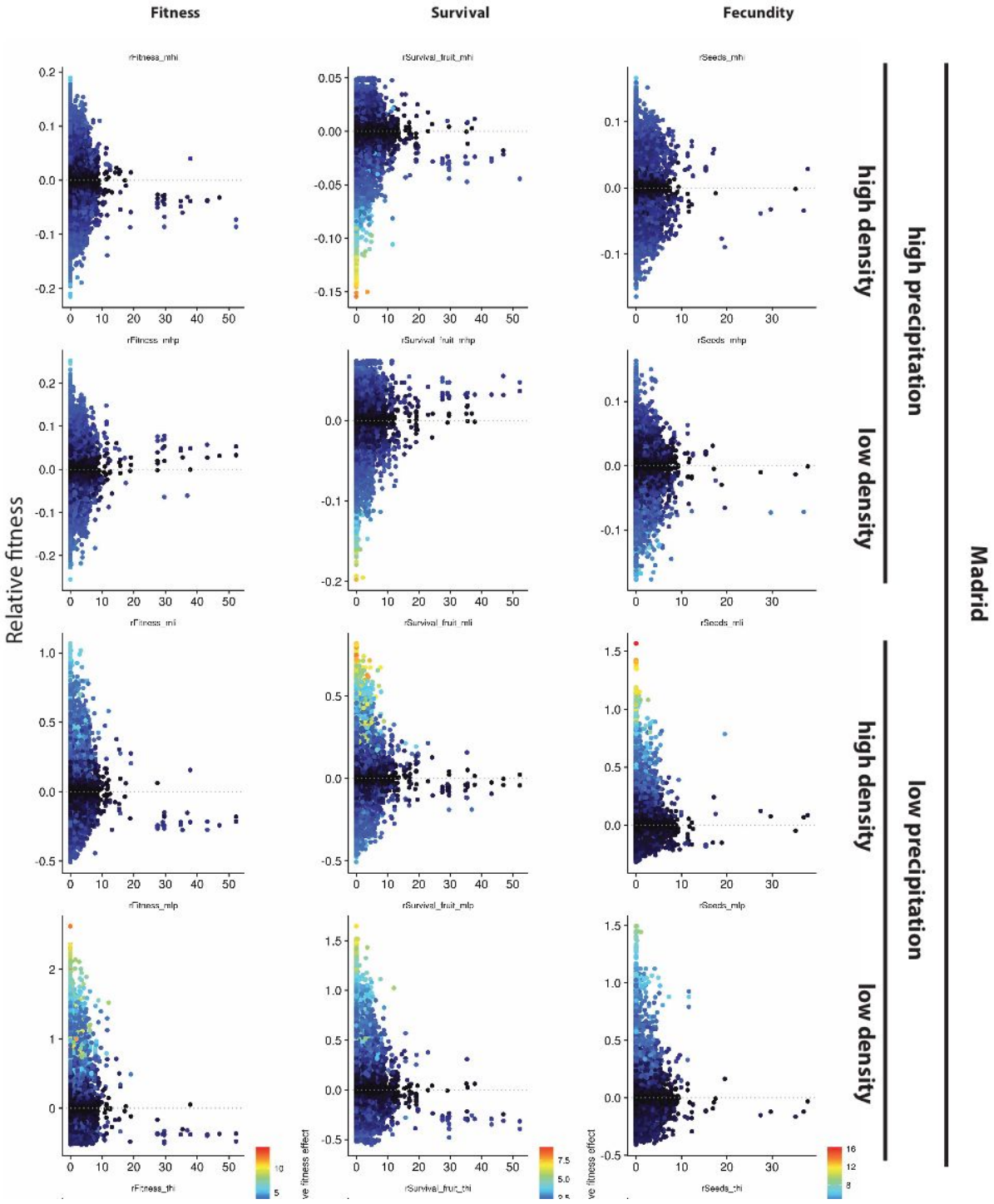


(Fig. SI.7 continued)

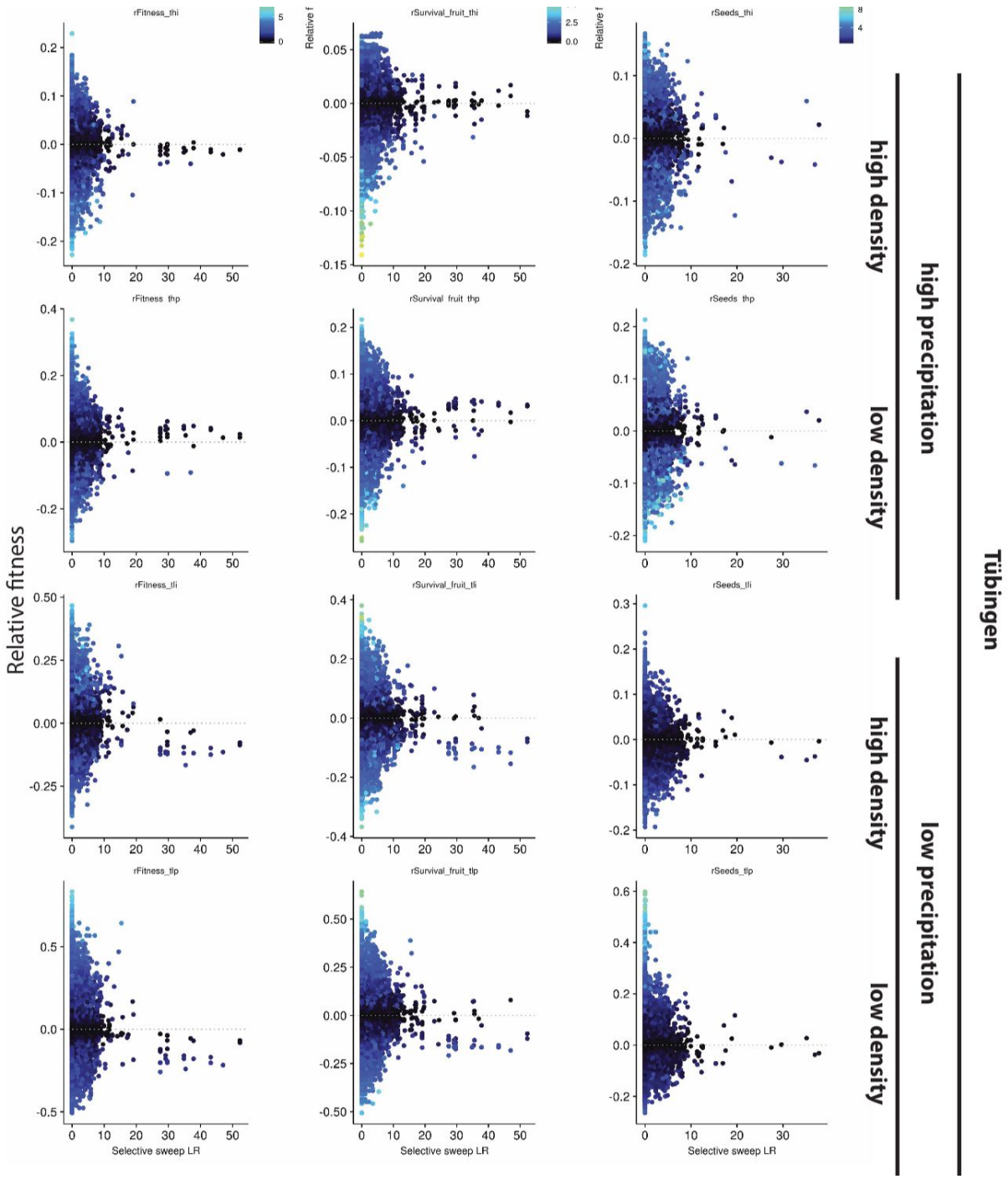


As Fig. 2C, for all environments: Madrid (Spain) and Tübingen (Germany), high and low precipitation treatments, and high and low plant density treatments.

Figure SI.8. Sweeps and empirical selection

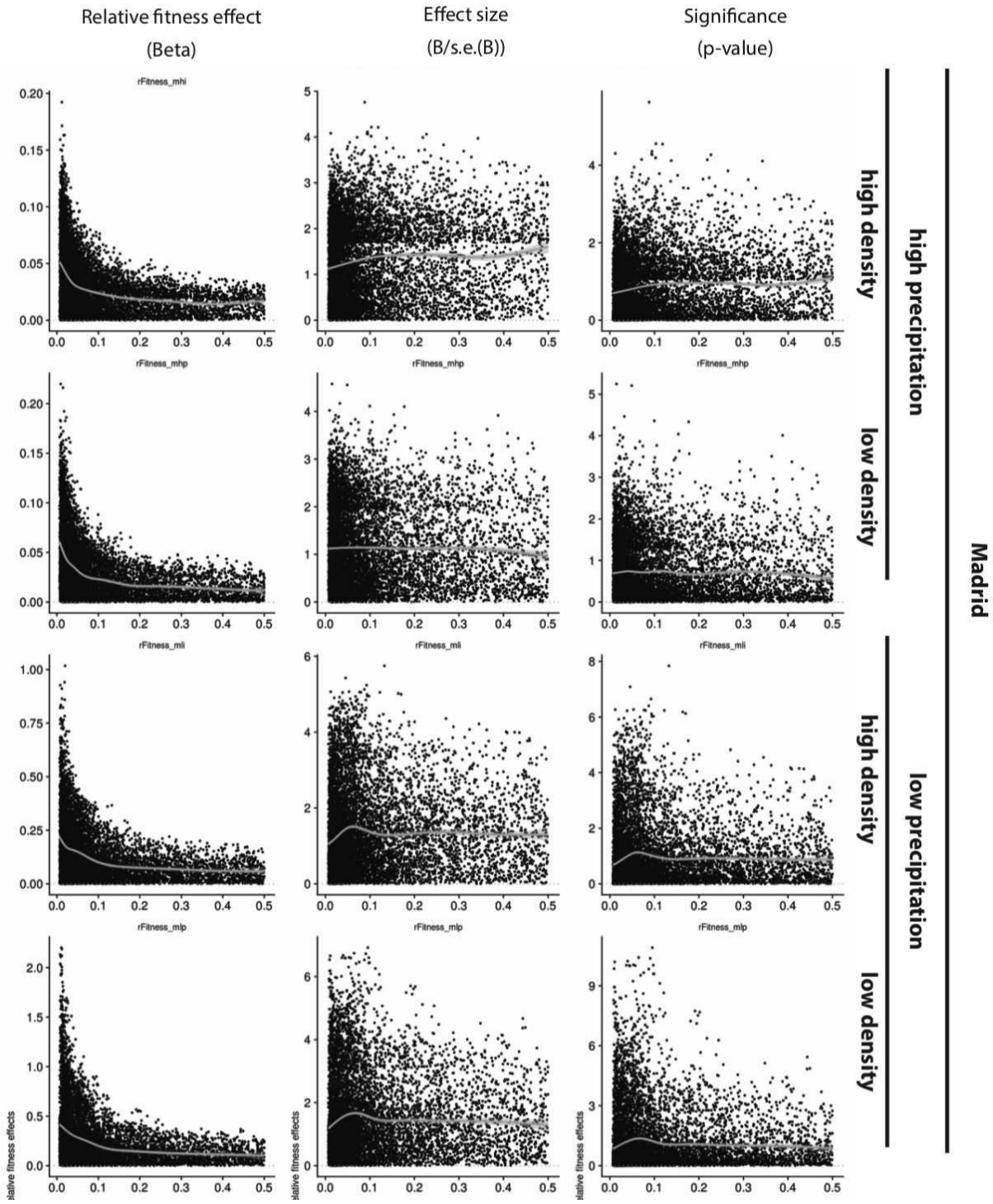


(Fig. S8 continued)

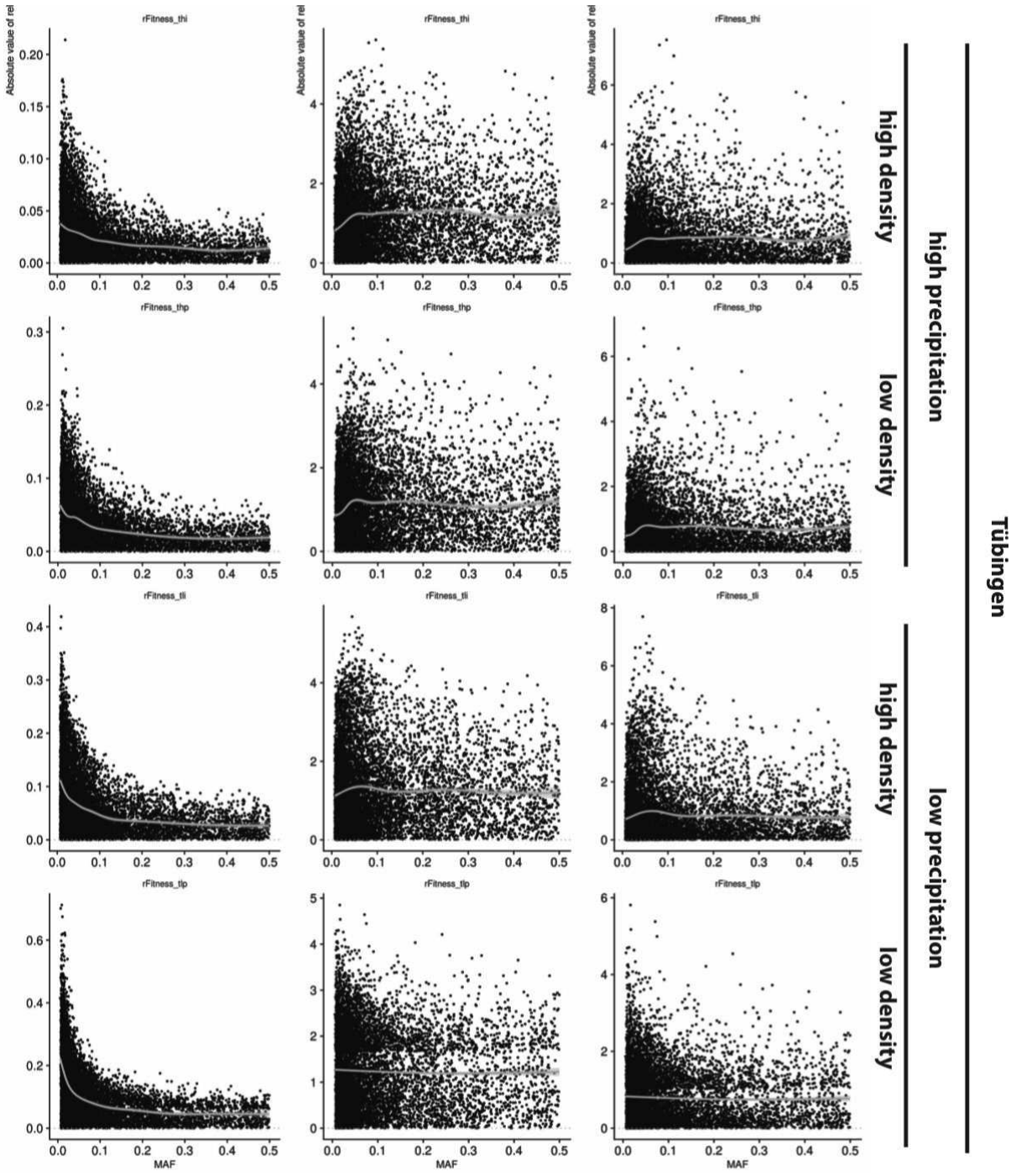


As Fig. 2D, for all environments: Madrid (Spain) and Tübingen (Germany), high and low precipitation treatments, and high and low plant density treatments.

Figure SI.9. Allele frequency and empirical selection



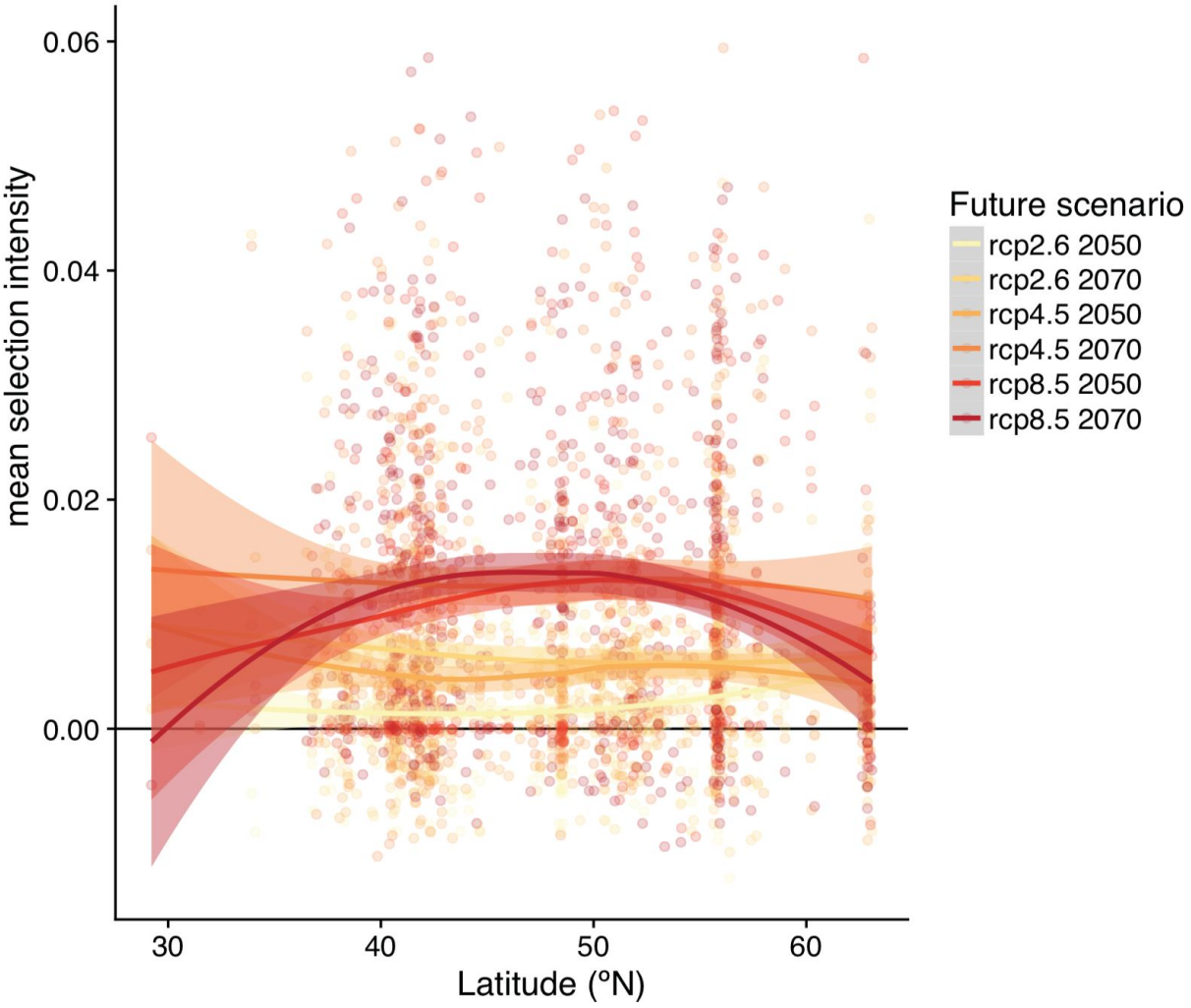
(Fig. S9 continued)



Relationships between relative fitness effect, relative fitness effect size, and *P*-values (calculated from GWA with relative fitness) and minor allele frequency of alleles for all environments: Madrid (Spain) and Tübingen (Germany), high and low precipitation treatments, and high and low plant density treatments.

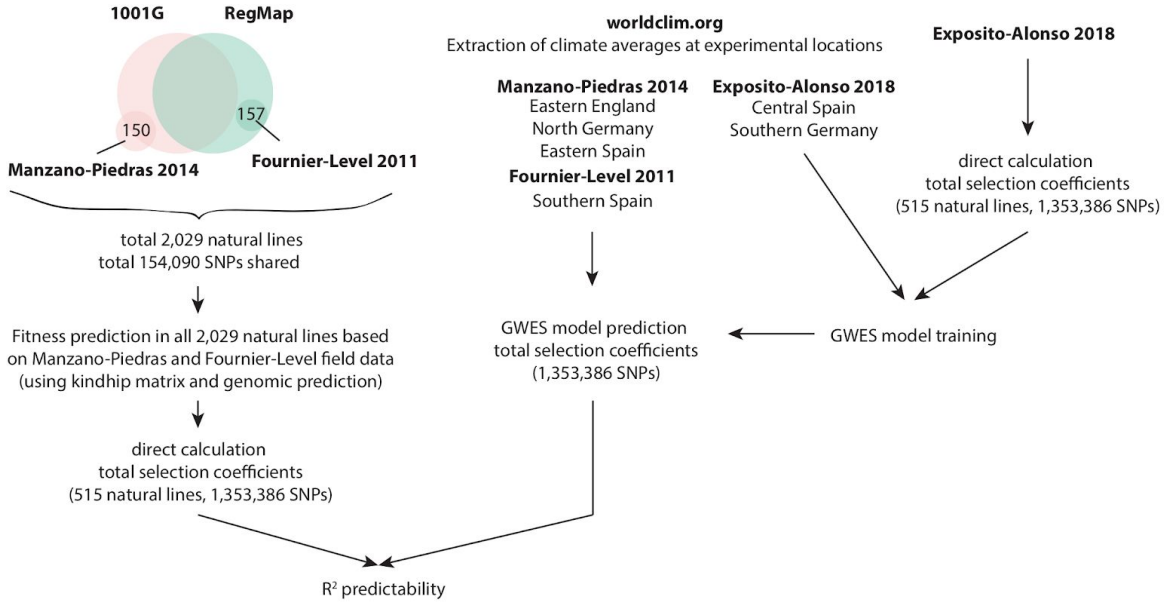
.

Figure SI.10. Future change in selection for different climate change scenarios



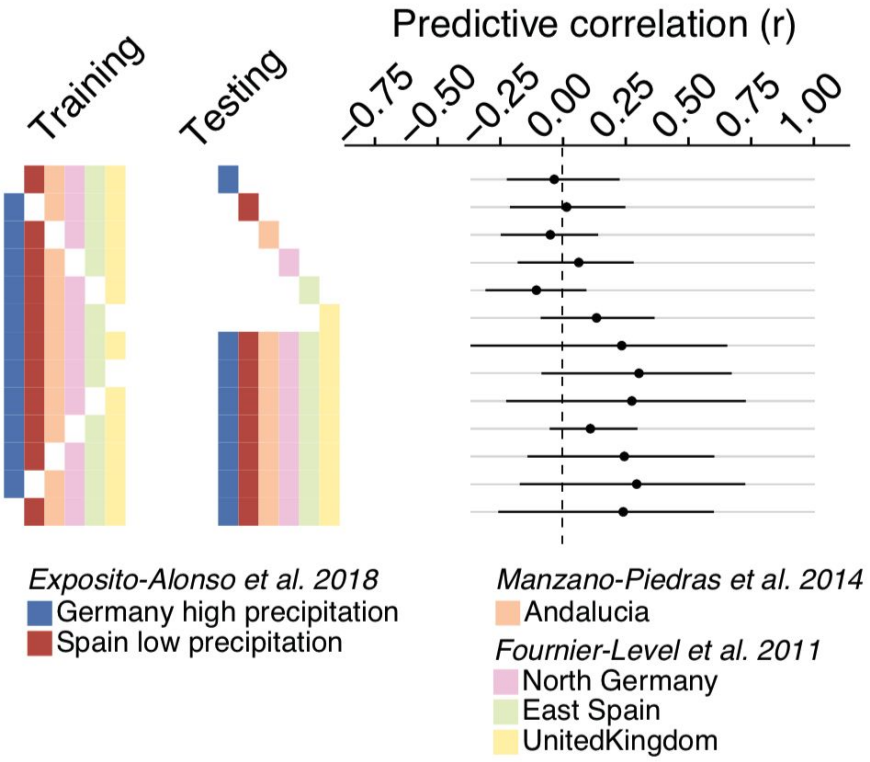
Same as Fig. 3G, but for different climate change scenarios. The higher the predicted CO_2 emissions (rcp, representative concentration pathway), the stronger the predicted increase in selection intensity.

Figure SI.11. Field validation conceptual chart



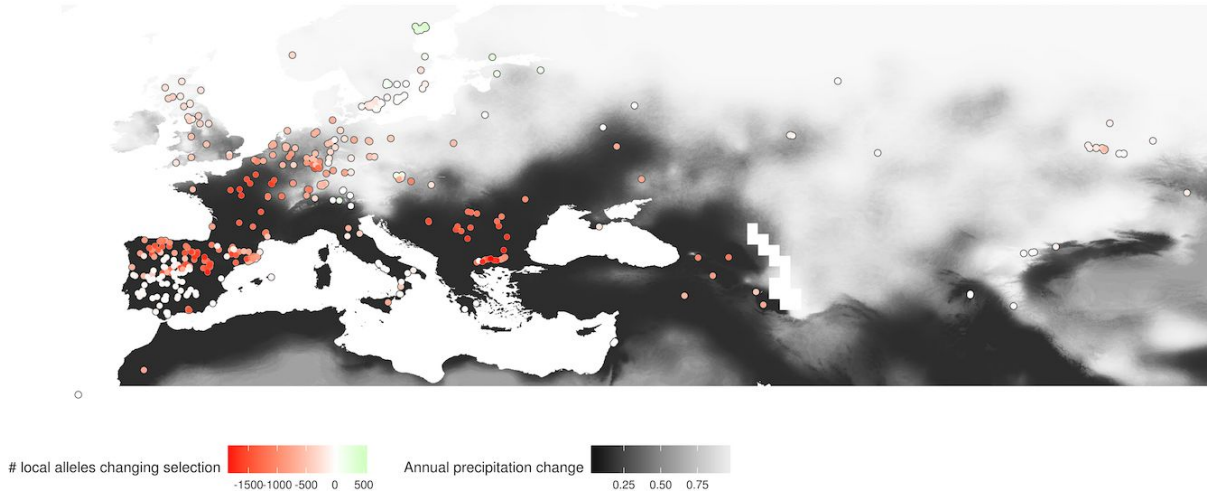
Conceptual workflow on field validation procedure with data from published experiments (section VIII).

Figure SI.12. Null expectation of predictability



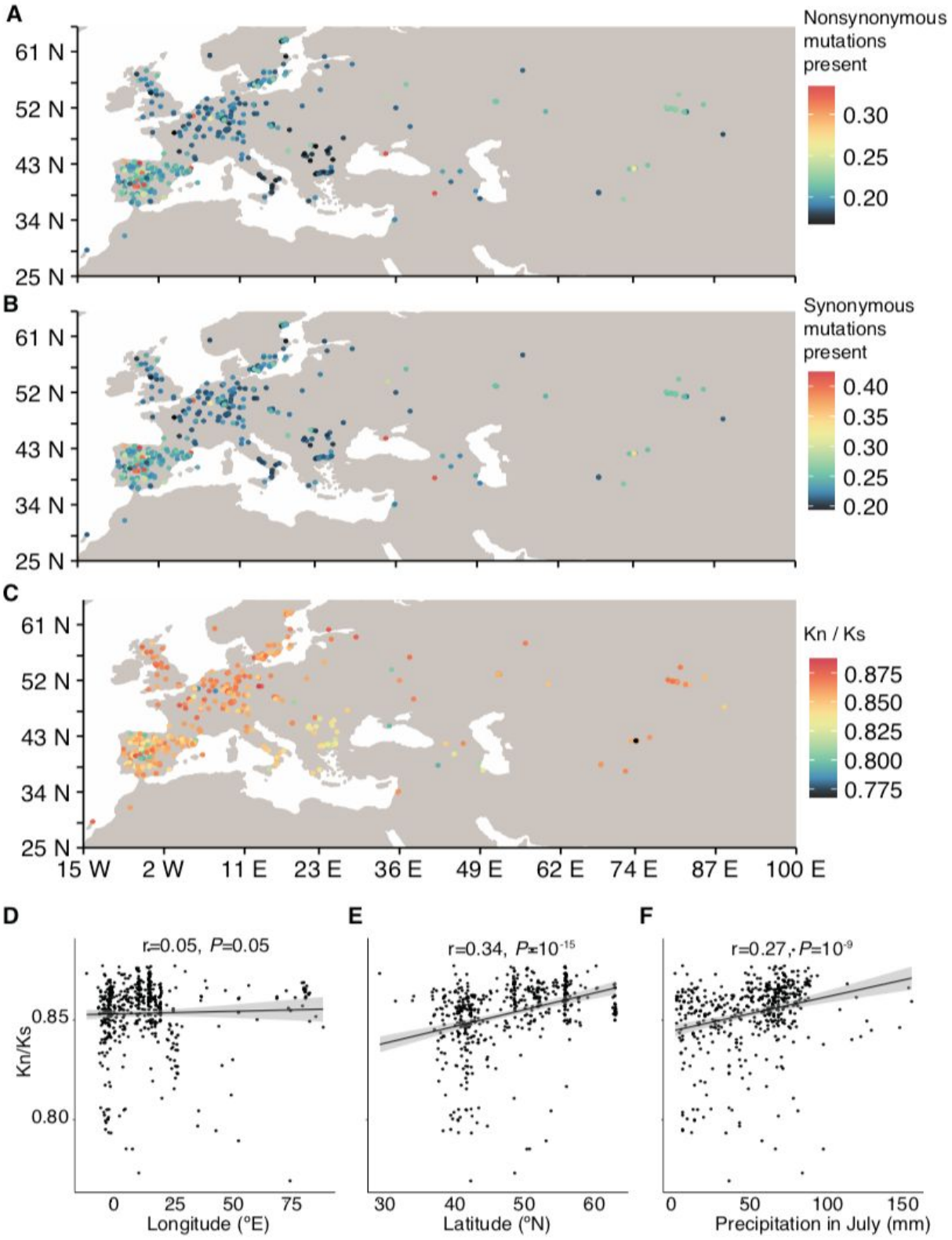
Same as Fig. 3, but with randomized fitness values associated to genotypes (section VIII). We could not find any model combination that had non-zero predictability (95% bootstrap confidence overlaps with zero). This proof of concept indicates that the predictability we find must have a biological basis, in which the combination of climate of origin for a genetic variant and the local climate allows to infer selection over such a variant.

Figure SI.13. Change in selection relative to local diversity



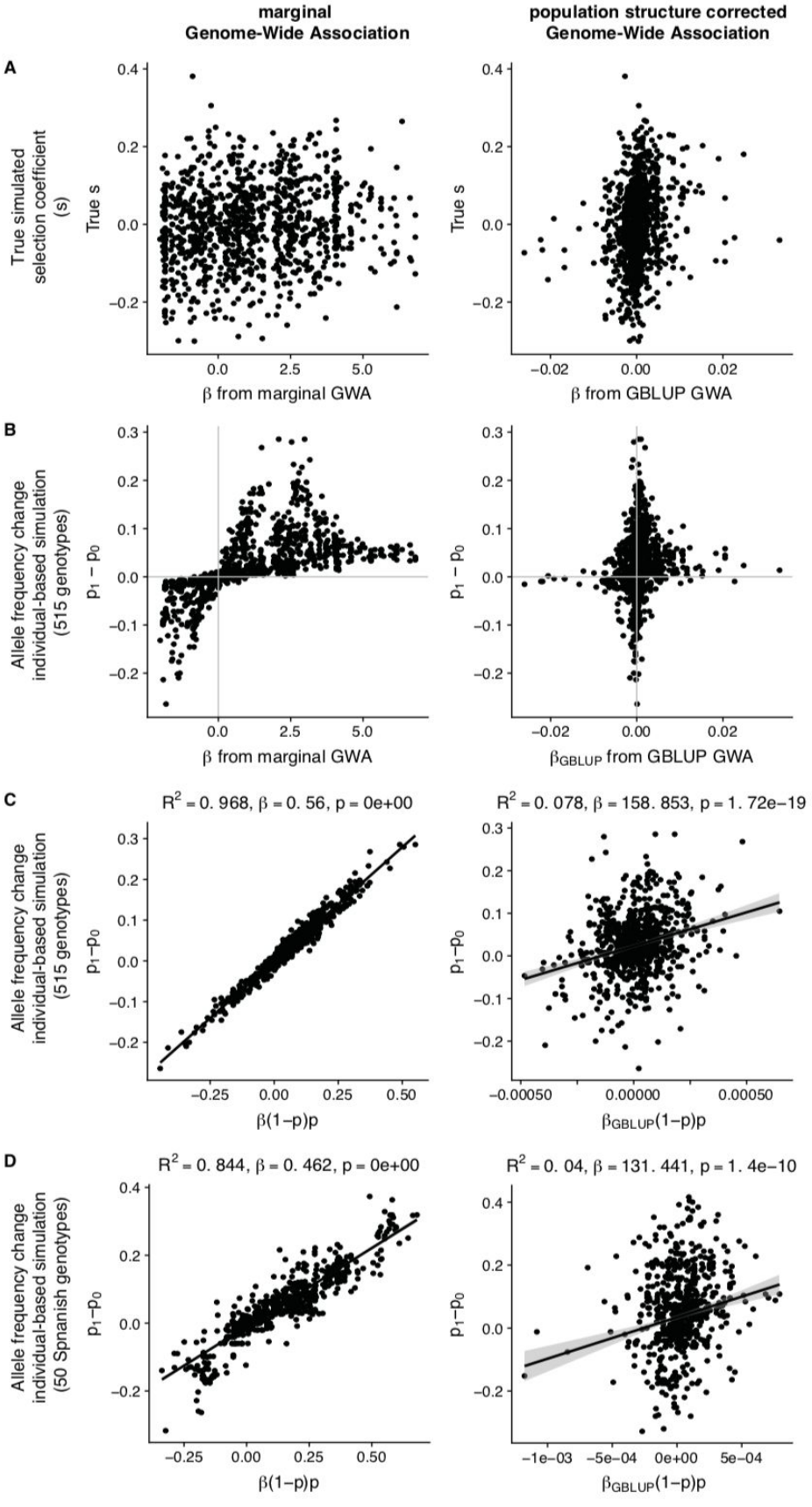
Same as Fig. 3D, but counting the number of local alleles increasing or decreasing in selection (total n=10,752 SNPs). Only changes with more than 5% advantage/disadvantage were considered (defined *a posteriori* from Bonferroni-significant alleles, which generated at least 5% effect in fitness).

Figure SI.14. Deleterious and neutral mutations across space



Fraction of all genome-wide nonsynonymous (A) and synonymous (B) mutations present in the local genotype. (C) Ratio of nonsynonymous and synonymous fraction, i.e. KnKs. Correlation of KnKs with degrees longitude, latitude, and precipitation in July (associated to selection intensity in Fig. 3)

Figure SI.15. GWA model comparison in a simulation study of selection



We simulated fitness of 515 plants that differ in 1000 SNPs (subset of the original 1,353,386 genome matrix used throughout the manuscript) with selection coefficients drawn from a normal distribution around zero (more details and code are available at https://github.com/MoisesExpositoAlonso/selectioncorrelatedgenotypes/, with DOI: https://doi.org/10.5281/zenodo.1408095). (A) Comparison of simulated (true) values of selection coefficients and estimates from marginal GWA and GBLUP-based GWA. In the main text, these are called "total selection coefficients" and "direct selection coefficients", respectively. (B) Genotypes were sampled based on their relative fitness values to produce a population one generation after selection. Genome-wide allele frequency changes from generation zero (p_0) to one generation after selection (p_1) are compared to marginal and GBLUP GWA estimates. (C) We plug in GWA estimates into the theoretical equation of allele frequency change based on selection coefficients, $\Delta p = p(1-p)s$, and compare the theoretical and the simulated allele frequency changes in one generation. (D) In order to demonstrate extrapolability, we repeated (C) but instead of running the one-generation simulation of allele frequency with the 515 genotypes, we do so with only 50 Spanish genotypes. We repeat again the comparison of theoretical frequency changes based on GWA estimates with 515 genotypes, with the simulated allele frequency changes with 50 genotypes. All in all, the comparisons above indicate that marginal GWA estimates are appropriate to understand the consequences of selection in changing allele frequencies, even when extrapolating to other populations with slightly different allele frequencies and linkage.

SUPPLEMENTAL TABLES I

Supplemental tables are available in the online version of the paper x. And are also deposited at Figshare wit doi: https://doi.org/10.6084/m9.figshare.6756836.

Table SI.1. Summary of fitness data

Average survival, fecundity, and lifetime fitness. Total number of genotypes with at least one surviving replicate per experiment.

[Abbreviations: The three characters of the codes: MLI, MLP, MHI, MHP, TLI, TLP, THI, TLP; indicate M=Madrid (Spain), T=Tübingen (Germany), L=Low precipitation, H=High precipitation, I=Individual replicates (one plant per pot), P=Population replicates (up to 30 plants per pot)].

Table SI.2. Heritability of fitness

Broad sense heritability and the 95% Highest Posterior Density Interval per trait (variance explained by line genotype), as calculated from a generalized linear mixed model using MCMCglmm, is reported as: σ_g/σ_{Total}. The proportion of variance explained by nuisance factors such as block (tray), position of the tray within a treatment block, and position of plant within a tray are reported in the same way. Proportion of Variance Explained (chip-heritability) and the 95% Highest Posterior Density Interval per trait, as calculated from a Bayesian Sparse Linear Mixed Model (BSLMM-GEMMA) using genotype means per trait and using a kinship/relationship matrix.

[Abbreviations: The three characters of the codes: MLI, MLP, MHI, MHP, TLI, TLP, THI, TLP; indicate M=Madrid (Spain), T=Tübingen (Germany), L=Low precipitation, H=High precipitation, I=Individual replicates (one plant per pot), P=Population replicates (up to 30 plants per pot)].

Table SI.3. Number of SNPs with significant total selection coefficients

All significant variants from marginal GWA after FDR and Bonferroni correction and all variants with non-zero probability of inclusion from conditional GWA, and sharing of significant variants across experiments.

Table SI.4. Expected allele frequency changes in response to selection

Summaries of allele frequency changes per experiment.

Table SI.5. Odds ratio of pleiotropic selection and conditional neutrality

Table SI.6. Correlation of total selection coefficients across environments

Table SI.7. Variable importance of predictive models

Sharing of significant variants across experiments.

Table SI.8. Predictability of environmental models

After training GWES models with a set of experiments, we inferred total selection coefficients on another set of experiments and compared those with the real total selection coefficients. We calculated Pearson's product-moment correlation r_{cv} and percentage of variance explained R^2_{cv} using a regression. 95% confidence intervals were calculated with 100 bootstrap replicates.

[Abbreviations: ml= Central Spain and low precipitation (both high and low plant density treatments combined); th= South Germany and high precipitation (both high and low plant density treatments combined), Andalucia= South Spain from Manzano-Piedras *et al.* (2014), Germany= North Germany from Fournier-Level *et al.* (2011), Spain= South East Spain from from Fournier-Level *et al.* (2011), UnitedKingdom= East England from Fournier-Level *et al.* (2011)].

Table SI.9. Description of climate variables

Climate variables used for environmental models are described and their sources reported.

Table SI.10. GBLUP heritability and imputation accuracy of published field data

We used GBLUP to impute fitness from Fournier-Level *et al.* (2011) and Manzano-Piedras *et al.* (2014) into our 517 global accessions. We report heritability, Pearson's r between GBLUP predicted fitness and real fitness, and the significance of the correlation test.

Table SI.11. Correlation between inferred natural selection intensity and other variables

Spearman's rho between selection intensity and diversity metrics or climate metrics is given.

SUPPLEMENTAL REFERENCES I

51. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
52. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6,** 80–92 (2012).
53. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
54. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33,** 1–22 (2010).
55. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
56. Weigel, D. & Nordborg, M. Population Genomics for Understanding Adaptation in Wild Plant Species. *Annu. Rev. Genet.* **49,** 315–338 (2015).
57. Weigel, D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.* **158,** 2–22 (2012).
58. Lande, R. & Arnold, S. J. THE MEASUREMENT OF SELECTION ON CORRELATED CHARACTERS. *Evolution* **37,** 1210–1226 (1983).
59. Vasseur, F. *et al.* Adaptive diversification of growth allometry in the plant *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 201709141 (2018). doi:10.1073/pnas.1709141115
60. Shriner, D. Overview of admixture mapping. *Curr. Protoc. Hum. Genet.* **1,** 1.23.1–1.23.6 (2013).
61. Anderson, J. T., Lee, C.-R., Rushworth, C. A., Colautti, R. I. & Mitchell-Olds, T. Genetic trade-offs and conditional neutrality contribute to local adaptation. *Mol. Ecol.* **22,** 699–708 (2013).
62. Mitchell-Olds, T. Pleiotropy Causes Long-Term Genetic Constraints on Life-History Evolution in Brassica rapa. *Evolution* **50,** 1849–1858 (1996).
63. Wadgymar, S. M. *et al.* Identifying targets and agents of selection: innovative methods to evaluate the processes that contribute to local adaptation. *Methods Ecol. Evol.* **8,** 738–749 (2017).
64. Fuka, D.R., Walter, M.T., Archibald J.A., Steenhuis T.S., and Easton Z.M. EcoHydRology: A community modeling foundation for Eco-Hydrology. (2014).
65. Golicher, D. Implementing a bucket model using WorldClim layers. (2012). Available at: https://rpubs.com/dgolicher/2964.
66. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles* **28,** 1–26 (2008).
67. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
68. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).
69. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **4,** 250–255 (2011).
70. Machado, H. *et al.* Broad geographic sampling reveals predictable and pervasive seasonal adaptation in Drosophila. *bioRxiv* 337543 (2018). doi:10.1101/337543
71. Nosil, P. *et al.* Natural selection and the predictability of evolution in Timema stick insects. *Science* **359,** 765–770 (2018).
72. Exposito-Alonso, M., Brennan, A., Alonso-Blanco, C. & Picó, F. X. Spatio-temporal variation in fitness responses to contrasting environments in Arabidopsis thaliana. *Evolution (in press)* (2018).
73. Siepielski, A. M. *et al.* The spatial patterns of directional phenotypic selection. *Ecol. Lett.* **16,** 1382–1392 (2013).
74. Siepielski, A. M., DiBattista, J. D. & Carlson, S. M. It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecol. Lett.* **12,** 1261–1276 (2009).

SUPPLEMENTAL APPENDIX II: A rainfall-manipulation experiment with 517 *Arabidopsis thaliana* accessions

Moises Exposito-Alonso[1], Rocío Gómez Rodríguez[2], Cristina Barragán[1], Giovanna Capovilla[1], Eunyoung Chae[1], Jane Devos[1], Ezgi S. Dogan[1], Claudia Friedemann[1], Caspar Gross[1], Patricia Lang[1], Derek Lundberg[1], Vera Middendorf[1], Jorge Kageyama[1], Talia Karasov[1], Sonja Kersten[1], Sebastian Petersen[1], Leily Rabbani[1], Julian Regalado[1], Lukas Reinelt[1], Beth Rowan[1], Danelle K. Seymour[1], Efthymia Symeonidi[1], Rebecca Schwab[1], Diep Thi Ngoc Tran[1], Kavita Venkataramani[1], Anna-Lena Van de Weyer[1], François Vasseur[1], George Wang[1], Ronja Wedegärtner[1], Frank Weiss[1], Rui Wu[1], Wanyan Xi[1], Maricris Zaidem[1], Wangsheng Zhu[1], Fernando García-Arenal[2], Hernán A. Burbano[1], Oliver Bossdorf[3], Detlef Weigel[1].

[1] Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany. [2] Center for Plant Biotechnology and Genomics, Technical University of Madrid, Pozuelo de Alarcón, Spain. [3] Institute of Ecology and Evolution, University of Tübingen, Tübingen, Germany.

I. Background & Summary

The gold standard for studying natural selection and adaptation in the wild is to quantify lifetime fitness of individuals from natural populations that have been grown together in a common garden, or that have been reciprocally transplanted. Natural selection over morphological, physiological or other traits has been studied in a wide range of organisms[17,23,73,74] using observational and experimental fitness measurements of multiple individuals in field conditions. However, studies that combine such measurements with knowledge on genome-wide variation are, in comparison, very rare[4,20,21]. This is surprising, given that they would enable the translation of selection to the genetic level and thus ultimately help us to understand whether traits will evolve over generations.

With climate change, the study of adaptation to the environment has acquired new importance. Predictions of climate change indicate not only that temperature will rise, but that also precipitation regimes will be altered, leading to more frequent and extreme droughts[75] and seriously threaten the persistence of plant communities[3,76]. Field experiments where climate variables such as rainfall are manipulated can be used to address this question[77].

Here we present a high-throughput field experiment with 517 whole-genome sequenced natural lines of *Arabidopsis thaliana*[25]. This experiment was designed to be of a sufficiently large scale to enable powerful genome-wide association analyses[78] and to maximize the replicability of species-wide patterns, which has been shown to increase with the diversity of genotypes included in an experiment[79]. The experiments were conducted in two field stations with contrasting climate, in the Mediterranean (Spain) and in Central Europe (Germany), where we built rainout shelters and simulated high and low rainfall. Using custom image analysis we quantified fitness- and phenology-related traits for 23,154 pots, which contained about 14,500 plants growing independently, and over 310,000 plants growing in small populations (max. 30 plants per pot). Three measurements of fitness were produced: survival from seed to reproductive adult (proportion 0—1) and the average fecundity per reproductive adult (inflorescence skeleton lengths ranged from 18,400 to 1,622,000 pixels, which approximately corresponds to 1 to 6,127 seeds per plant). Fecundity was only measured for plants with at least one fruit. We finally calculated an integrated lifetime fitness value by multiplying the survival proportion to adulthood with the total offspring produced. This dataset will be invaluable for the study of natural selection and adaptation in the context of global climate change at the genetic level, building on the genetic catalog of the 1001 Genomes Project[25] and complementing the already published extensive set of traits measured in controlled growth chamber or greenhouse conditions[80,81].

II. Selection of accessions from the 1001 Genomes Project

The 1001 Genomes (1001G) Project[25] has provided information on 1,135 natural lines or accessions and 11,769,920 SNPs and small indels called after re-sequencing. To select the most genetically and geographically informative 1001G lines, we applied several filters: (1) First we removed the accessions with the lowest genome quality. We discarded those with < 10X genome coverage of Illumina sequencing reads and < 90% congruence of SNPs called from MPI and GMI pipelines[25]. (2) We removed near-identical individuals. Using Plink software[53] we computed identity by state across the 1,135 accessions. For pairs of accessions with < 0.01 differences per SNP (<100,000 variants approx.), we randomly selected one accession to include in our study. (3) Finally, we reduced geographic sampling ascertainment bias, as the sampling for 1001G was performed in neither a random nor a regularly structured scheme. Some laboratories provided several lines per location whereas others provided lines that were collected at least several hundred kilometres apart. Using each accession's collection location, we computed Euclidean distances across the 1,135 accessions and identified all pairs that were apart less than 0.0001 Euclidean distance in degrees latitude and longitude (<< 100 meters). From such pairs, we randomly selected one accession to remain. After

applying criteria (1), (2), and (3), we obtained a final set of 523 accessions (Datasets 1 and 2). To bulk seeds for our rainfall-manipulation experiment and control for maternal effects, we first propagated accessions in controlled conditions. We stratified the seeds one week at 4°C, we sowed them in trays with industrial soil (CL-P, Einheitserde Werkverband e. V., Sinntal-Altengronau Germany) and placed them in a growth room with 16 h light and 23°C for one week. Trays were vernalized for 60 days at 4°C and 8 h daylength. After vernalization, trays were moved back to 16 h light and 23°C for final growth and reproduction. This generated sufficient seeds for 517 accessions, which were later grown in the field in two locations (Fig. SII.1). Seeds originating from the same parents can be ordered from the 1001G seed stock at the Arabidopsis Biological Resource Center (CS78942).

III. Field experiment design

III.1 Rainout shelter, watering, and block design

We built two 30 m x 6 m tunnels of PVC plastic foil to fully exclude rainfall in Madrid (Spain, 40.40805ºN -3.83535ºE) and in Tübingen (Germany, 48.545809ºN 9.042449ºE) (Fig. SII.2A-B). The foil tunnels are different from a regular greenhouse in that they are completely open on two sides. Thus, ambient temperatures vary virtually as much as outside the foil shelter (see Environmental sensors section). In each location, we supplied artificial watering in two contrasting regimes: abundant watering and reduced watering. Inside each tunnel, we created a 4% slope, and four flooding tables (two for high and two for low precipitation) (1 m x 25 m, Hellmuth Bahrs GmbH & Co KG, Brüggen, Germany) covered with soaking mats (4 l/m^2, Gärtnereinkauf Münchingen GmbH, Münchingen, Germany). The flooding tables were placed on the ground in parallel to the slope. Water was able to drain at the lower end of the flooding table (Fig. SII.2A-B). A watering gun was used to manually simulate rainfall from the top.

Our experimental design is a split-plot design (Fig. SII.2C), with precipitation treatments replicated twice in each location and the genotypes randomized within precipitation treatment in a total of 8 spatial blocks. This ensured that all genotypes would be equally evenly distributed within the foil tunnel, and that we could robustly measure consistent fitness responses to water deprivation across precipitation replicates.

On top of the flooding tables, we used potting trays with 8x5 cells (5.5 cm x 5.5 cm x 10 cm size) and industrial soil (CL-P, Einheitserde Werkverband e.V., Sinntal-Altengronau Germany). Each cell would correspond to a genotype, excluding corner cells, to avoid extreme edge effects. We grew a total of 12 replicates per genotype per treatment: Five replicates were grown at high density, with

30 seeds per cell and without further intervention ("population replicate"). The remaining seven replicates were at low density (ca. 10 seeds) and one seedling was selected at random after germination ("individual replicate"). Excess individuals were culled. While the population replicates should more faithfully reflect survival from seed to reproduction, the individual replicates were useful to more accurately monitor flowering time and seed set.

III.2 Environmental sensors

Environmental variables — air temperature, photosynthetically active radiation (PAR) and soil water content — were monitored every 15 minutes for the entire duration of the experiment using multi-purpose sensors (Flower Power, Parrot SA, Paris, France). This enabled us to adjust watering depending on the degree of local evapotranspiration during the course the experiment. The sensors outside of the tunnel in Madrid (i.e. only natural rainfall) showed an interquartile range between 1% and 17% soil water content. This overlapped with the range of 10 to 22% water content of the drought treatment that we artificially imposed inside the tunnels in Madrid and Tübingen. The lower range of measurements in Madrid (outside sensor) is due to a lack of natural rainfall during the first two months of the experiment (Fig. SII.2E, Table SII.1). In contrast, the sensor outside the tunnel in Tübingen recorded an interquartile range of soil water content percentage of 22 to 27%, which was comparable to the high watering treatments in Tübingen and Madrid (from 20 to 33%) (Fig. SII.2E, Table SII.1). These values confirmed that our low and high watering treatment were not only different, but also that they mimicked natural soil water content at the two contrasting locations. Mean daily air temperatures (measured by the sensors at 5-10 cm above the soil surface every 15 minutes) were overall higher in Madrid (8-10°C) than in Tübingen (5-6°C), and the difference in temperature between the sensors inside and outside the tunnels was in both locations on average only 1°C (Fig. SII.2F, Table SII.1). The photosynthetically active radiation (PAR, wavelengths from 400 to 700 nm) had a median of 0.1 mol m^{-2} day^{-1} at night for all experiments. At mid-day (11:00-13.00 hrs), the median PAR in Madrid was 57.8 mol m^{-2} day^{-1} outside, and 45.7 mol m^{-2} day^{-1} inside the tunnel. In Tübingen, the median values were 29.0 outside, and 30.9 mol m^{-2} day^{-1} inside the tunnel.

III.3 Sowing and quality control

During sowing, contamination of neighboring pots with adjacent genotypes can occur for multiple reasons. In order to avoid such contamination, we chose a day with no wind and sowed seeds at 1-2 cm height from the soil. Additionally, we took care during the first days to be particularly gentle when using the watering gun to avoid seed-carryover (bottom watering by flooding was done regularly). We also tried to remove human error during sowing by preparing and randomizing 2 ml

plastic tubes containing the seeds to be sown in the same layouts (5x8) as the destination trays. During sowing, each experimenter took a box at random and went to the corresponding labeled and arranged tray in the field (Fig. SII.2). This reduced the possibility of sowing errors. Sowing occurred on November 16 2015 in Madrid and on October 22 2015 in Tübingen. During vegetative growth, we could identify seedlings that resembled their neighbors or were located in the border between two pots and removed such plants as potential contaminants. We also used the homogeneity of flowering within a pot in the population replicates as a further indicator for contamination (Fig. SII.3A). When a plant had a completely different flowering timing or vegetative phenotypes did not coincide with the majority of plants in the pot, this plant was removed. After sowing and quality control, the total number of pots was 24,747 instead of the original 24,816 pots (99.7%) (Dataset 3).

IV. Field monitoring

IV.1 Image analysis of vegetative rosettes

Top-view images were acquired every four to five days (median in both sites) with a Panasonic DMC-TZ61 digital camera and a customized closed dark box, the "Fotomatón" (Fig. SII.3A), at a distance of 40 cm from each tray. In total, we imaged each tray at 20 timepoints throughout vegetative growth. The implemented segmentation was the same as in Exposito-Alonso *et al.*[11], which relies on the Open CV Python library[82]. We began by transforming images from RGB to HSV channels. We applied a hard segmentation threshold of HSV values as (H=30-65, S=65-255, V=20-220). The threshold was defined after manually screening 10 different plants in order to capture the full spectrum of greens both of different accessions and of different developmental stages. This was followed by several iterations of morphology transformations based on erosion and dilation. For each of the resulting binary images we counted the number of green pixels.

During field monitoring, we noticed that some pots were empty because seeds had not germinated. In these cases, we left a red marker in the corresponding pots (Fig. SII.3A), which could be detected in a similar way as the presence of green pixels (with threshold H=150-179, S=100-255, V=100-255). These pots were excluded from survival analysis as they did not contain any plants (Fig. SII.3A). The resulting raw data consist of green and red pixel counts per pot (Fig. SII.3B). In order to detect the red markers automatically, we performed an analysis of variance between pots above and below a threshold of red pixels and finding the threshold that maximized this separation (Fig. SII.3C). This provided us with the threshold of red pixels above which a pot had a red marker (indicating an empty pot). As expected, the distribution of pixels was bimodal, making this identification straightforward.

We estimated germination timing by analysing trajectories (Fig. SII.3B) of green pixels per pot, and identifying the first day that over 1,000 green pixels were observed in a pot (corresponding to a plant size of ~ 10 mm^2, Fig. SII.3) (Datasets 3). The final dataset contained data for 22,779 pots — after the removal of pots with red labels — with a time series of green pixel counts.

IV.2 Manual recording of flowering time

We visited the experimental sites every 1-2 days and manually recorded the pots with flowering plants. Flowering time was measured as the day when the first white petals could be observed with the unaided eye. This criterion was chosen as sufficiently objective to reduce experimenter error. To keep track of previous visits and avoid errors, we labeled the pots where flowering had already been recorded with blue pins. To calculate flowering time, we counted the number of days from the date of sowing to the recorded flowering date (we did not use the inferred day of germination to avoid introducing modeling errors in the flowering time metric). Fig. SII.4A shows the raw flowering time data per pot in the original spatial distribution and the distribution of flowering time per treatment combination. Note that grey boxes are pots with plants that did not survive until flowering. In total, we gathered data for 16,858 pots with flowering plants (Datasets 3).

IV.3 Image analysis of reproductive plants

Once the first dry fruits were observed, we harvested them and took a final 'studio photograph' of the rosette and the inflorescence (Fig. SII.5A). In total, we took 13,849 photographs. The camera settings were the same as for the vegetative monitoring, but here we included an 18% grey card approximately in the same location for each picture in case *a posteriori* white balance adjustments would be needed. We first used a cycle of morphological transformations of erode-and-dilate to produce the segmented image (Fig. SII.5C). This generated a segmented white/black image without white noise. Then, we used the thin (erode cycles) algorithm from the Mahotas Python library[83] to generate a binary picture reduced to single-pixel paths — a process called skeletonisation (Fig. SII.5C). Finally, to detect the branching points in the skeletonised image we used a hit-or-miss algorithm. We used customized structural elements to maximize the branch and end point detection (Fig. SII.5C). This resulted in four variables per image: total segmented inflorescence area, total length of the skeleton path, number of branching points, and number of end points (Fig. SII.5C) (Datasets 3-4).

IV.4 Estimation of fruit and seed number

Although the study of natural selection is based on studying relative fitness, and total reproductive area might provide a good relative estimate, sometimes it is useful to have a proxy of the absolute fitness. In order to provide an approximate number of how many seeds each plant had produced, we generated two allometric relationships by visual counting of fruits per plant and seeds per fruit. In order to be sure that the counts corresponded to single plants, we counted fruits and seeds of only individual replicates of accessions, not the population replicates (see Field experiment design section). Because a strong relationship had already been validated between inflorescence size and the number of fruits in a number of studies with *A. thaliana*[84–86], we decided that counting a few inflorescences of three sizes, reflecting the broad size spectrum, would be sufficient to establish a first allometric relationship with the four image-acquired variables (n=11 inflorescences, R^2=0.97, P=4×10^{-4}, Fig. SII.5B). To express fecundity as the number of seeds, we counted all seeds inside one fruit for each of the inflorescences used for the first allometric relationship (n=11 fruits), aiming for a wide range of fruit sizes. The mean was 28.3 seeds per fruit and the standard deviation was 11.2 seeds. The two aforementioned allometric relationships were used to predict, first, the number of fruits per inflorescence using the four image analysis variables, and second, the number of seeds corresponding to the number of fruits per inflorescence (Datasets 3-4).

V. Technical validations

Data processing

All images, from where fruits and leaf area were estimated, are backed up and stored at the Max Planck Institute for Developmental Biology and available through ftp transfer (ca. 2Tb) upon request to weigel@weigelworld.org. The Max Planck Society requires storage of publication-relevant data for a minimum of 10 years. The Python modules to process images for green area segmentation and inflorescence analyses are available at http://github.com/MoisesExpositoAlonso/hippo and http://github.com/MoisesExpositoAlonso/hitfruit, along with example datasets.

To document our data curation we created the R package dryAR (http://github.com/MoisesExpositoAlonso/dryAR with doi:).

Replicability of image processing

After testing different camera parameters, we used an exposure of -2/3 and an ISO of 100. White balance was set for flashlight. We used a dark box with all sides closed, so the flashlight was the only source of illumination. This ensured that the white balance and illumination were virtually consistent

from picture to picture, as shown before[11]. Photos were saved both in .jpeg and .raw to allow for *a posteriori* adjustments if needed. Using a calibration board with 1.3 cm x 1.3 cm white and dark squares, we examined the error between the inferred area from image analysis and the real 1.3 cm-side squares across the tray. This provided us with a median resolution estimate of 101.5 pixels mm^{-2}. The deviations from the true area were minimal, with a median of 2.7% and values of 1.4% / 4.2% for the 1st and 3rd quartile. The maximum area deviations were of 8 to 9% in the extreme corners of the tray, where we did not sow any seeds. We are confident that such small variation in retrieved area is compensated by the randomized locations of genotypes within the trays.

To further verify that our camera settings and segmentation pipeline produced replicable extractions of plant green area, we used images of trays that were photographed twice on the same day by mistake. In total there were 1,508 such pots distributed across 11 timepoints and different trays. By comparing the area of the same pot of two different camera shots and segmentation analyses, we could verify that the Spearman's rho of rank correlation was very high (r=0.97, n=1508, $P<10^{-16}$), confirming high replicability.

Because we ran the same segmentation and skeletonization software on both rosette and inflorescence images, we could leverage the clearly different image patterns that rosettes and inflorescences have to identify labeling errors (i.e. mistakes in manually inputting sample information of the pictures). To do this, we first trained a random forest model to predict the manually labeled "rosette" or "inflorescence" by the four image variables (Fig. SII.5). By fitting a Random Forest with all images, we find that the leave-one-out accuracy was 92.1%, i.e. ca. 2,000 images were incorrectly labeled by the algorithm. We manually checked whether these were mislabeled or rather whether they "looked similar" in terms of area or landmark points in the photo, e.g. when both rosette or inflorescences were diminute. We found that only 2.5% were incorrectly mislabeled (and corrected them) and are thus confident that the labeling error must be below 2.5%.

Experimental validation

Although repeating experiments in climatically-similar locations would be impractical, we could verify that survival in Madrid and low precipitation correlated with a preliminary drought experiment in the greenhouse[11] (Spearman's rho=0.17, n=211, P=0.01). On the other hand, reproductive allocation measured under optimal conditions in the greenhouse correlated with total seed output in the most similar field experiments, Tübingen high precipitation (Spearman's rho=0.27, n=211, $P=5\times10^{-5}$)[84].

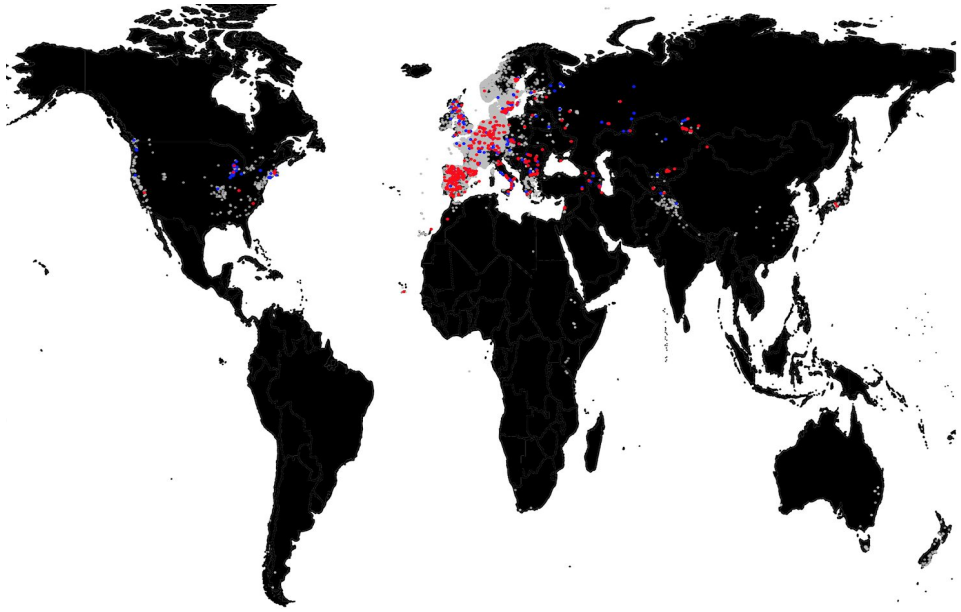
VI. Author contributions

MEA conceived and designed the project. MEA carried out the experiment in Tübingen. MEA and RGR carried out the experiment in Madrid. All authors contributed to specific tasks in the experiments (see detailed description below). OB provided the field site in Tübingen and FGA provided the site in Madrid. DW secured funding for the project. MEA carried out the analyses and wrote the first draft of the manuscript. All authors edited, commented and approved the manuscript.

AUTHOR	Conceived_idea	Funding	Advice	Coordination	Materials	Bulking_seeds	Seed_aliquoting	Field_setup	Pictures_plants	Sowing_Madrid	Sowing_Tuebingen	Thinning_seedlings	Field_care	Image_processing	Foil_tunnel_reparation	Fresh_harvesting_Madrid	Fresh_harvesting_Tuebingen	Dry_imaging_Madrid	Dry_imaging_Tuebingen	Flowering_monitoring	Image_processing	Data analysis/processing	Writing_first_draft
Moises Exposito-Alonso	x			x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x
Rocio Gomez Rodriguez							x	x	x	x			x			x		x		x			
Detlef Weigel		x	x		x												x						
Hernán A Burbano			x							x													
Oliver Bossdorf			x		x																		
Rebecca Schwab			x	x	x												x						
Fernando García Arenal			x		x																		
George Wang			x																				
François Vasseur			x								x					x							
Julian Regalado							x																
Derek Lundberg											x						x						
Ronja Wedegärtner							x	x	x		x		x				x		x	x			
Frank Weiss									x														
Danelle Seymour											x												
Beth Rowan											x				x		x						
Patricia Lang									x		x	x			x	x	x						
Jorge Kagayema											x												
Rui Wu											x				x		x						
Wanyan Xi											x												
Kavita Venkataramani											x				x	x	x						
Giovanna Capovilla												x			x		x						
Efthymia Symeonidi									x			x			x		x						
Vera Middendorf												x					x		x	x			
Anna-Lena Van de Weyer												x											
Jane Devos												x											
Diep Thi Ngoc Tran												x											
Sonja Kersten					x						x				x								
Wangsheng Zhu												x			x								
Maricris Zaidem															x								
Sebastian Petersen																	x						
Ezgi Dogan																	x						
Claudia Friedemann																x	x						
Talia Karasov																x							
Cristina Barragán																x							
Leily Rabbani																	x						
Caspar Gross																	x		x	x			
Lukas Reinelt													x						x	x			
Eunyoung Chae																	x						

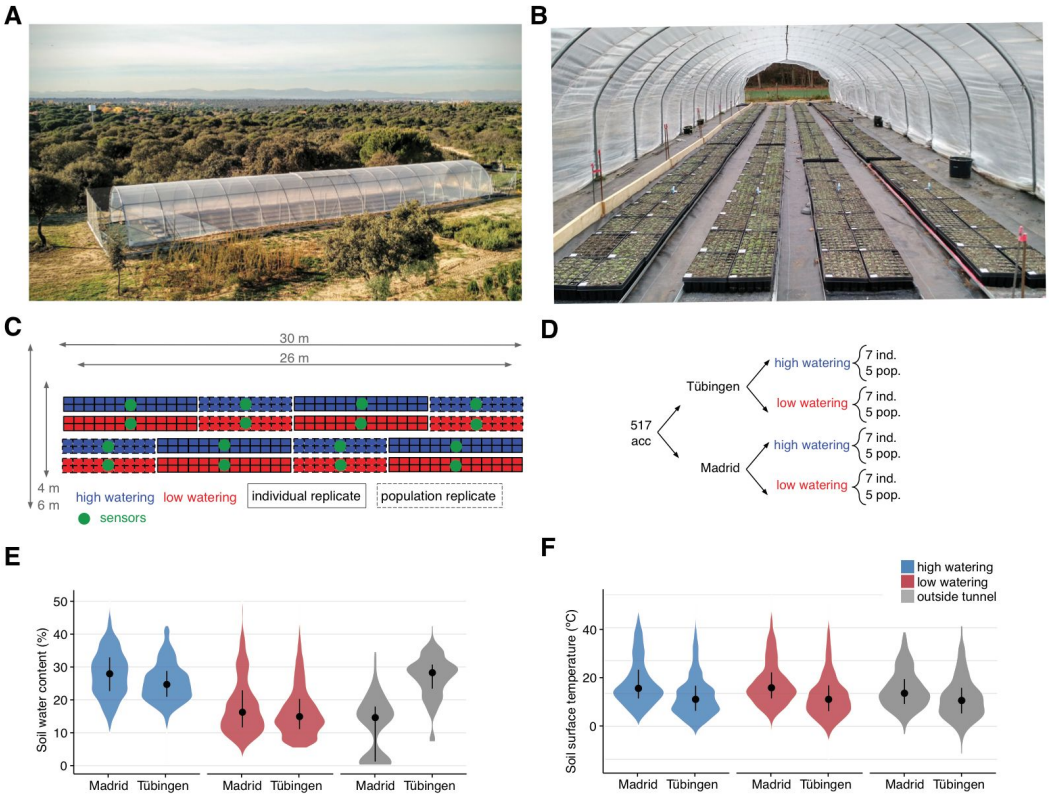
SUPPLEMENTAL FIGURES II

Figure SII.1. Geographic distribution of accessions



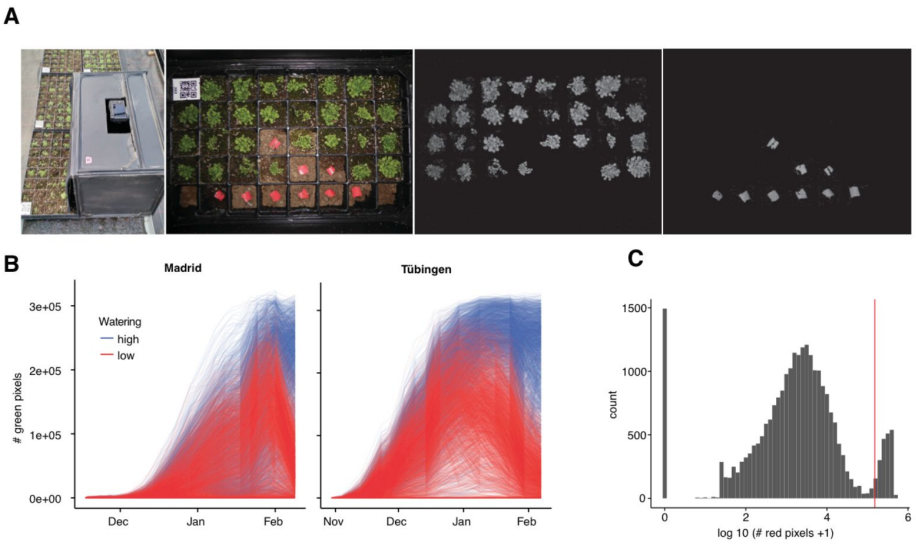
Locations of *Arabidopsis thaliana* accessions used in this experiment (red), 1001G accessions (blue), and all sightings of the species in gbif.org (grey).

Figure SII.2. Field experiment design



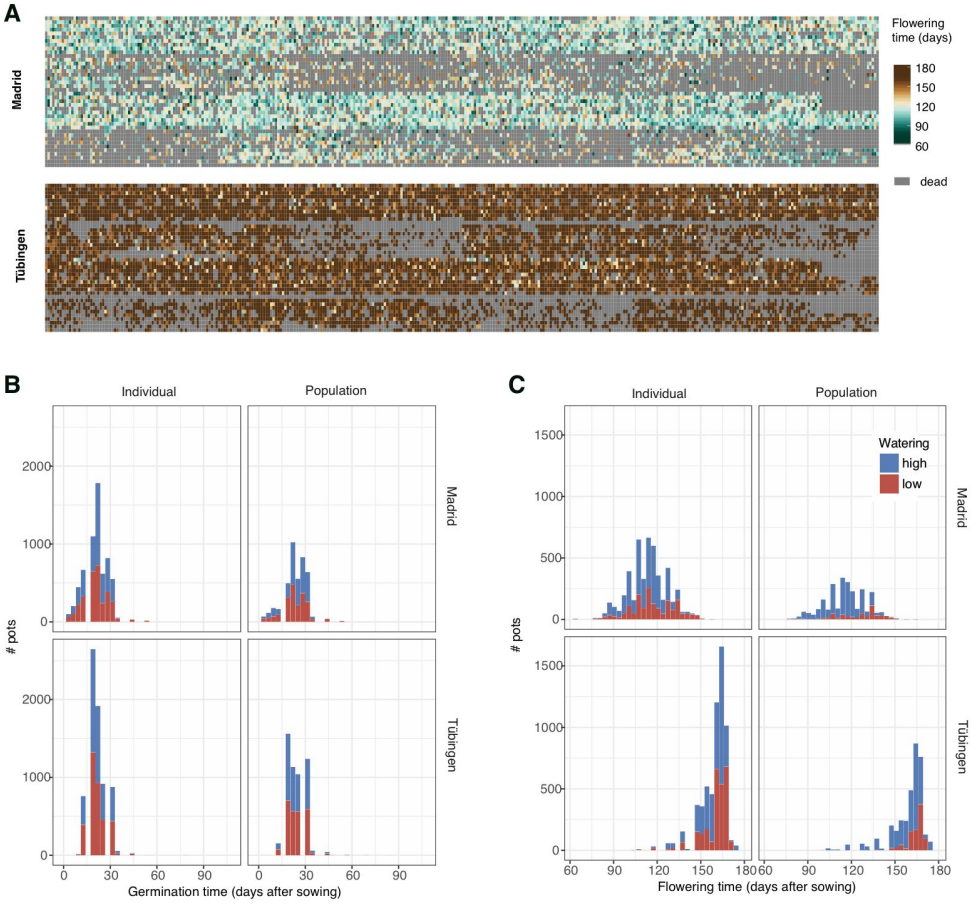
(A) Aerial view of foil tunnel settings in Madrid and (B) view inside the foil tunnel in Tübingen. (C) Spatial distribution of blocks and replicates and (D) experimental design. (E) Soil water content and (F) soil surface temperature from the 34 sensors monitoring each experimental block and conditions outside the tunnel.

Figure SII.3. Rosette monitoring



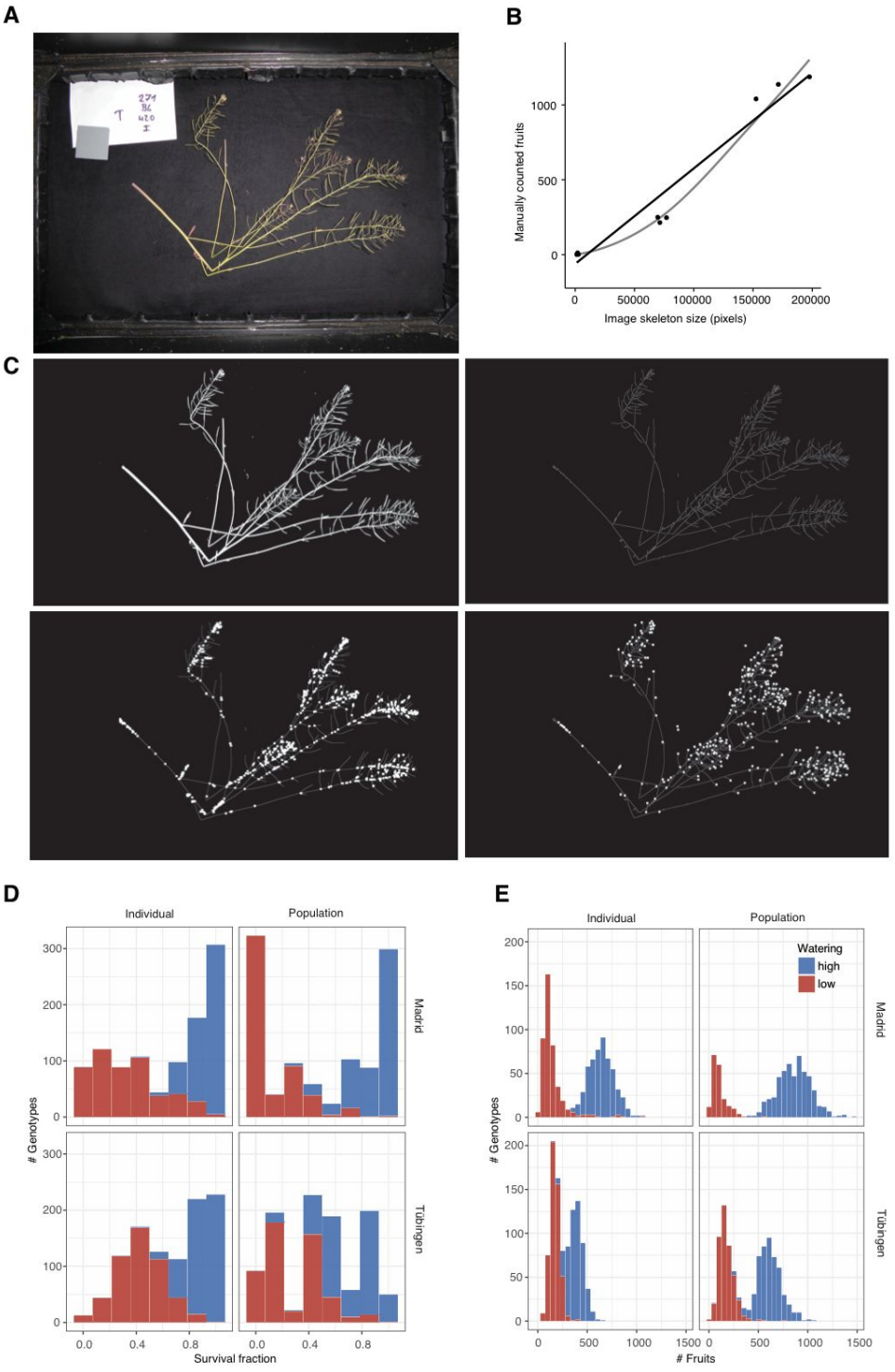
(A) Customized dark box ("Fotomatón") for image acquisition and example tray with the corresponding green and red segmentation. (B) Trajectories of number of green pixels per pot, indicating rosette area, for Madrid and Tübingen. (C) Distribution of the sum of red pixels per pot over all time frames. The red vertical line indicates the heuristically chosen threshold to define whether the pot actually had a red marker.

Figure SII.4. Flowering time distributions



(A) Flowering times per pot in the same spatial arrangement as in each tunnel (see Fig. SII.2). (B) Distribution of germination times. (C) Distribution of flowering times.

Figure SII.5. Inflorescence and seed set estimation



(A) Representative inflorescence picture. (B) Regression between the fruits of a few manually counted inflorescences and the inflorescence size calculated based on image processing. The four variables inferred in (C) accurately predicted the visually counted inflorescences as example (R^2=0.97, n=11, P=10^{-4}). (C) Resulting variables from image processing of (A): total segmented area (upper-left), skeletonized inflorescence (upper-right), branching points (lower-left), and endpoints (lower-right). Distribution of survival to reproduction (D) and fruits per plant (E) in the four environments.

SUPPLEMENTAL TABLES II

Table SII.1 Summaries of environmental sensor measurements

A total of 34 sensors were placed in the different treatment blocks (low/high) as well as outside (out) of the foil tunnels. The median (interquartile) values of all sensors per treatment and location are shown.

Site	Rainfall	Soil water content (%)	Air temperature (ºC)
Madrid	out	14.5 (1.09, 17.46)	8.5 (5.34, 12.39)
Madrid	low	16.1 (11.38, 22.51)	10.0 (6.95, 15.13)
Tuebingen	low	14.7 (10.76, 20.09)	6.6 (3.27, 10.78)
Tuebingen	out	27.7 (22.82, 30.50)	5.6 (2.44, 9.54)
Tuebingen	high	24.6 (20.73, 29.02)	6.6 (3.27, 10.78)
Madrid	high	27.8 (22.62, 33.00)	9.8 (6.82, 15.13)

Table SII.2 Variable descriptions

Variable names and their descriptions and units are reported (see Datasets). All datasets share a common accession identification number.

Dataset	Variable	Information
D1&D2& D3&D4	id	Unique numeric ID assigned to the accessions included in the 1001 Genomes Project
D1&D2	name	Classic accession name assigned by original collector
D1&D2	country	Country of collection
D1&D2	sitename	Toponym of the location of collection
D1&D2	latitude	Degrees North of the location of origin (°N)
D1&D2	longitude	Degrees East of the location of origin (°E)
D1&D2	collector	Original researcher that collected the accession
D1&D2	collectiondate	Calendar date of collection
D1&D2	CS_number	Stock number in the Arabidopsis Biological Resource Center (abrc.osu.edu)
D1&D2	Q_SNPcongruency	Pass/no pass of thresholds for genome quality and SNP calling congruency
D1&D2	Q_geneticsdist	Pass/no pass of the filter for almost identical accessions
D1&D2	Q_geodist	Pass/no pass of filter for geographically close accessions
D1&D2	is_relict	Belongs to the Mediterranean "relict" lineage
D1&D2	finalset	Included in the final 517 set for the field experiment
D3&D4	site	Field station site. m=Madrid(Spain), t=Tübingen(Germany)
D3&D4	water	Rainfall/watering treatment. h=high rainfall, l=low rainfall
D3&D4	indpop	Density of plants per pot. i=single plant selected after germination, p=population of 30 seeds growing undisturbed
D3&D4	qpblock	Identification number of quickpot (tray) within treatment block (rainfall row x replicate block)
D3&D4	qp	Identification number of quickpot tray in the whole experiment
D3&D4	qp_x	Pot position in x axis within the quickpot tray
D3&D4	qp_y	Pot position in y axis within the quickpot tray
D3&D4	pos	Pot x,y coordinate within the quickpot tray
D3&D4	rep	Replicate number
D3&D4	trayid	Identification of the tray combining block and treatments
D3&D4	potindex	Identification of pot combining site, tray, and position within the tray
D3&D4	Germination_time	Inference of germination time based on the day that rosette area was over 1,000 pixels size (days after sowing)
D3&D4	Green	Sum of all green areas per pot throughout the experiment (# pixels). This helps to identify successfully growing pots.
D3&D4	Red	Sum of all red areas per pot throughout the experiment (# pixels). This helps to identify red tags placed on pots that failed throughout the experiment
D3&D4	Survival_flowering	Survival until reproduction (i.e. production of flowers)
D3&D4	Flowering_date	Date that the first flowers had developed
D3&D4	Flowering_time	Time from sowing until the date of flowering (days)
D3&D4	Inflorescence_size	Area of inflorescence (#pixels)
D3&D4	Survival_num	Number of surviving plants until fruit set. Only applies to "population" pots.
D3&D4	Survival_fruit	Survival until fruit set (i.e. produced fruits)

D3&D4	Fruits	Number of fruits inferred from the function between visually counted fruits and inflorescence area, total path, branching points, and ending points.
D3&D4	Seeds	Number of seeds inferred from the average number of seeds per fruit and number of fruits.
D3&D4	Infloresncence_byind	Area of inflorescence (#pixels) divided by total number of plants per pot. Only applies to "population" pots.
D3&D4	Fruits_byind	Number of fruits divided by total number of plants per pot. Only applies to "population" pots.
D3&D4	Seeds_byind	Number of seeds divided by total number of plants per pot. Only applies to "population" pots.
D3&D4	Fitness	Lifetime fitness (number of seeds / seed planted). This metric integrates survivorship and reproduction.

DATASETS

Supplemental datasets are available in the online version of the paper x and are also deposited at Figshare with doi: https://doi.org/10.6084/m9.figshare.6480599. A detailed descriptions of each Dataset's columns can be found in Table SII.2.

Dataset 1 Quality-based selection of the original 1,135 accessions

We report the 1001 Genome identification numbers, the quality filters that each accession passed during the selection of the 517 set.

Dataset 2 Description of the 517 accessions

We report the final set of 517 accessions that were used in the field experiment.

Dataset 3 All traits measured per replicate

For each pot replicate, we report all raw data as well as composite variables.

Dataset 4 Curated means per accession

For each accession, we report averages of all data as well as composite variables.

SUPPLEMENTAL REFERENCES II

75. Dai, A. Increasing drought under global warming in observations and models. *Nat. Clim. Chang.* **3,** 52–58 (2012).
76. Schwalm, C. R. *et al.* Global patterns of drought recovery. *Nature* **548,** 202–205 (2017).
77. Tielbörger, K. *et al.* Middle-Eastern plant communities tolerate 9 years of drought in a multi-site climate manipulation experiment. *Nat. Commun.* **5,** 5102 (2014).
78. Burghardt, L. T., Young, N. D. & Tiffin, P. A Guide to Genome-Wide Association Mapping in Plants. *Current Protocols in Plant Biology* 22–38 (2017).
79. Milcu, A. *et al.* Genotypic variability enhances the reproducibility of an ecological study. *Nat Ecol Evol* **2,** 279–287 (2018).
80. Seren, Ü. *et al.* AraPheno: a public database for Arabidopsis thaliana phenotypes. *Nucleic Acids Res.* **45,** D1054–D1059 (2017).
81. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465,** 627–631 (2010).
82. Itseez. Open Source Computer Vision Library. (2015). Available at: https://github.com/itseez/opencv.
83. Coelho, L. P. Mahotas: Open source software for scriptable computer vision. *Journal of Open Research Software* **1,** e3 (2013).
84. Vasseur, F., Wang, G., Bresson, J., Schwab, R. & Weigel, D. Image-based methods for phenotyping growth dynamics and fitness in Arabidopsis thaliana. *bioRxiv* (2018).
85. Brachi, B. *et al.* Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **112,** 4032–4037 (2015).
86. Roux, F., Gasquez, J. & Reboud, X. The dominance of the herbicide resistance cost in several Arabidopsis thaliana mutant lines. *Genetics* **166,** 449–460 (2004).

Thesis Appendix III

“The rate and effect of *de novo* mutation in a colonizing lineage”

Exposito-Alonso, M., Becker, C., Schuenemann, V.J., Reitter, E., Setzer, C., Slovak, R., Brachi, B., Hagmann, J., Grimm, D.G., Jiahui, C., Busch, W., Bergelson, J., Ness, R.W., Krause, J., Burbano, H.A., Weigel, D., (2018). *PLOS Genetics*, https://doi.org/10.1371/journal.pgen.1007155.

The rate and potential relevance of new mutations in a colonizing plant lineage

Moises Exposito-Alonso[1,2†], Claude Becker[1†], Verena J. Schuenemann[3,4], Ella Reiter[3], Claudia Setzer[5], Radka Slovak[5], Benjamin Brachi[6§], Jörg Hagmann[1§], Dominik G. Grimm[1§], Jiahui Chen[6,7], Wolfgang Busch[5§], Joy Bergelson[6], Rob W. Ness[8], Johannes Krause[3,4,9], Hernán A. Burbano[2,*], Detlef Weigel[1,*]

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[2]Research Group for Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[3]Institute of Archaeological Sciences, University of Tübingen, 72070 Tübingen, Germany

[4]Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, 72070 Tübingen, Germany

[5]Gregor Mendel Institute, Austrian Academy of Sciences, 1030 Vienna, Austria

[6]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

[7]Institute of Tibet Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[8]Department of Biology, University of Toronto Mississauga, Mississauga, Ontario L5L IC6, Canada.

[9]Max Planck Institute for the Science of Human History, 07743 Jena, Germany

[†]Co-first authors

[§]Current addresses: INRA, UMR 1202 Biodiversité Gènes & Communautés, 33610 CESTAS, France (B.B.); Computomics, 72072 Tübingen, Germany (J.H.); Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland (D.G.G); Salk Institute for Biological Studies, La Jolla, CA 92037, USA (W.B.).

*Correspondence to: hernan.burbano@tuebingen.mpg.de, weigel@weigelworld.org

Running title: *de novo* mutation rate in *A. thaliana*

Keywords: colonization, mutation, selection, herbarium genomes, aDNA, phylogenomics, population genomics, association mapping, *Arabidopsis thaliana*

ABSTRACT

By following the evolution of populations that are initially genetically homogeneous, much can be learned about core biological principles. For example, it allows for detailed studies of the rate of emergence of *de novo* mutations and their change in frequency due to drift and selection. Unfortunately, in multicellular organisms with generation times of months or years, it is difficult to set up and carry out such experiments over many generations. An alternative is provided by "natural evolution experiments" that started from colonizations or invasions of new habitats by selfing lineages. With limited or missing gene flow from other lineages, new mutations and their effects can be easily detected. North America has been colonized in historic times by the plant *Arabidopsis thaliana*, and although multiple intercrossing lineages are found today, many of the individuals belong to a single lineage, HPG1. To determine in this lineage the rate of substitutions – the subset of mutations that survived natural selection and drift –, we have sequenced genomes from plants collected between 1863 and 2006. We identified 73 modern and 27 herbarium specimens that belonged to HPG1. Using the estimated substitution rate, we infer that the last common HPG1 ancestor lived in the early 17th century, when it was most likely introduced by chance from Europe. Mutations in coding regions are depleted in frequency compared to those in other portions of the genome, consistent with purifying selection. Nevertheless, a handful of mutations is found at high frequency in present-day populations. We link these to detectable phenotypic variance in traits of known ecological importance, life history and growth, which could reflect their adaptive value. Our work showcases how, by applying genomics methods to a combination of modern and historic samples from colonizing lineages, we can directly study new mutations and their potential evolutionary relevance.

SUMMARY

A consequence of an increasingly interconnected world is the spread of species outside their native range — a phenomenon with potentially dramatic impacts on ecosystem services. Using population genomics, we can robustly infer dynamics of colonization and successful population establishment. We have compared hundred genomes of a single *Arabidopsis thaliana* lineage in North America, including genomes of contemporary individuals as well as 19th century herbarium specimens. These differ by an average of about 200 mutations, and calculation of the nuclear evolutionary rate enabled the dating of the initial colonization event to about 400 years ago. We also found mutations associated with differences in traits among modern individuals, suggesting a role of new mutations in recent adaptive evolution.

INTRODUCTION

Colonizing or invasive populations sampled through time [1,2] constitute “natural experiments” where it is possible to study evolutionary processes in action [3]. Colonizations, which are dramatically increasing in number [4,5], sometimes are characterized by strong bottlenecks and genetic isolation [6,7], and thus greatly facilitate the observation of new mutations and potentially their effects under natural population dynamics and selection [8]. Colonizations thus offer a complementary approach to other studies of new mutations, which often minimize natural selection, for example in laboratory mutation accumulation experiments [9] and parent-offspring comparisons [10]. The study of colonizations is also complementary to the investigation of genetic divergence over long time scales, e.g., between distant species [11], where the results are largely independent of short-term demographic fluctuations. There is broad interest in understanding how genetic diversity is generated [12], and how new mutations can provide a path for rapid adaptive evolution [13–15]. Additionally, accurate evolutionary rates permit dating historic population splits, which is fundamental to the study of population history [16].

The analysis of colonizing populations can also contribute to resolving the “genetic paradox of invasion” [17]. This paradox comes from the observation that colonizing populations can be surprisingly successful and spread very widely even when strongly bottlenecked, suggesting some level of adaptation to new environments that goes beyond the exploitation of unoccupied ecological niches [17]. Much of the work in plant ecology and evolution has focused on evidence that populations can rapidly adapt from standing variation [18]. In invasive lineages, initial standing variation may originate from incomplete bottlenecks, multiple introductions, or admixture with local relatives [19]. Much less work has been done with respect to the role of *de novo* mutations as a solution to the genetic paradox of invasion, although this has been proposed as an alternative explanation for rapid adaptation by colonizing lineages [3,17,20].

The self-fertilizing plant *Arabidopsis thaliana* is native to Africa and Eurasia [21,22] but has recently colonized N. America, where it likely experienced a strong founder effect [23]. At nearly half of N. American sites sampled during the 1990s and early 2000s, more than 80% of plants belong to a single haplogroup, HPG1, as inferred from genotyping with 149 intermediate-frequency markers evenly spread throughout the genome [23]. The HPG1 lineage has been reported from many sites along the East Coast and in the Midwest as well as at a few sites in the West [23] (Figure 1, Table S1). The great ubiquity of HPG1 in comparison to any other haplogroup could be due to either some

adaptive advantage, or, more parsimoniously, be the result of HPG1 being derived from one of the first arrivals of *A. thaliana* in the continent.

Here, we focus on 100 HPG1 individuals that do not show any evidence of outcrossing with other lineages. We combine genomes from herbarium specimens and live individuals, collectively covering the time span from 1863 to 2006, to infer mutation rates, to date the birth of the HPG1 lineage, and to investigate the evolutionary forces that shape genetic diversity. Our analyses of this lineage serves as a model for future studies of similar colonizing or otherwise recently bottlenecked plant populations, in order to better understand how diversity is generated and to which extent it contributes to adaptation in nature.

RESULTS AND DISCUSSION

Historic and modern genomes

In a self-fertilizing species, a single individual can give rise to an entire lineage of millions of offspring, which then diversify through new mutations and eventually intra-lineage recombination. If self-fertilization is much more common than outcrossing, the founder is likely to have been homozygous throughout almost the entire genome. Because it is so wide spread, HPG1 presents an opportunity to sample many natural populations that have been potentially derived from a common, very recent ancestor with such characteristics. In the best possible case, this would allow for new mutations to be directly observed through time. To test these assumptions and to better understand the evolution of HPG1, we sequenced two different groups of plants. The first group were live descendants of 87 plants that had been collected between 1993 and 2006 (Fig. 1; Table S1), and which had been identified as likely members of the HPG1 lineage with 149 genome-wide markers spaced at roughly 1-Mb-intervals [23]. We aimed for broad geographic representation, with at least two accessions per collection site, where available. The second group comprised 36 herbarium specimens, collected between 1863 and 1993, for which we had no a priori information whether they may or may not belong to the HPG1 lineage, but which were selected from the herbarium records to cover the full historical geographic range and overlap with modern samples when possible (Fig. 1).

The DNA from the herbarium specimens showed biochemical features typical of ancient DNA (aDNA) from plants, which we have previously described in detail [24]. Such DNA damage included a median fragment length of 60 bp, an excess of C-to-T substitutions of about 2.5% at the first base of sequencing reads and a 1.5 to 1.8 fold enrichment of purines at DNA breakpoints (Fig. S1**,**

Supplementary Text 2**).** The reads of repaired libraries are available at https://www.ebi.ac.uk/ena/data/view/PRJEB24619. To remove aDNA associated damage and produce high-quality genomes, chemically-repaired libraries (see Methods) were later sequenced. These reads were mapped against an HPG1 pseudo-reference genome [25], focusing on single nucleotide polymorphisms (SNPs) because the short sequence reads of herbarium samples preclude accurate calling of structural variants. Genome sequences were of high quality, with herbarium samples covering 96.8–107.2 Mb of the 119 Mb reference, and modern samples covering 108.0–108.3 Mb (Table S1).

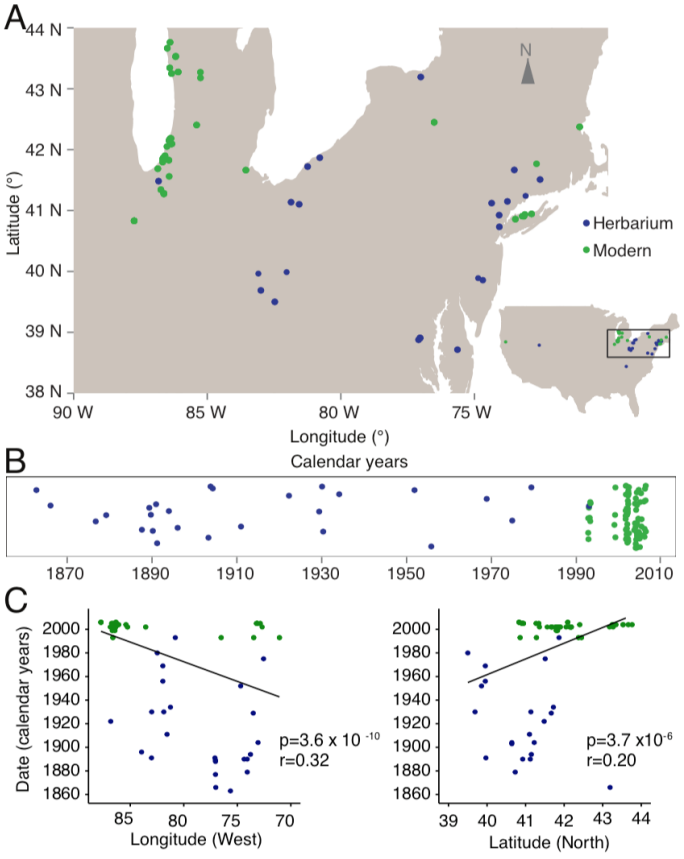


Figure 1. Geographic location and temporal distribution of HPG1 samples.

(A) Sampling locations of herbarium (blue) and modern individuals (green). **(B)** Temporal distribution of samples (random vertical jitter for visualization purposes). **(C)** Linear regression of latitude and longitude as a function of collection year (p-value of the slope and Pearson correlation coefficient are indicated)

Genetic diversity of HPG1 and delineation from other lineages

We visualized the relationships between the sequenced historic and modern plants building a neighbor joining tree of all 123 samples and confirmed that the majority fell within a almost-identical

clade, the HPG1 (Fig. 2A) [23]. Because any degree of introgression from other non-HPG1 lineages would confound the discovery of new mutations downstream, we removed all divergent samples and built a neighbour joining tree (n=103 samples), which revealed that the HPG1 samples were very similar to each other, with very little within-population structure (Fig. 2B). A parsimony network was used to detect recombinant genomes within this HPG1 clade (Fig. 2C), which led us to remove three potential intra-lineage recombinants. Repeating the parsimony network cleared all previously inferred reticulations due to recombinations (Fig. 2D). After such stringent filtering, we kept 27 of the 35 herbarium samples, and 73 of the 87 modern samples (Table S1). These constitute a set of non-admixed, non-recombined and quasi-identical HPG1 individuals.

Pairs of HPG1 herbarium genomes differed by 28-207 SNPs genome-wide, pairs of HPG1 modern genomes by 2-259 SNPs, and pairs of historic-modern HPG1 genomes by 56-244 SNPs. That is, whole-genome identity was at least 99.9997% in any pairwise comparison. Of the approximately five to six thousand segregating SNPs in the HPG1 population, the vast majority, about 95% (Supplementary Text 3), have not been reported outside of this lineage [21]. Importantly, the density of SNPs along the genome was low and evenly distributed (typically fewer than 20 SNPs / 100 kb) with no peaks of much higher frequency, which makes us confident that chunks of introgressions from other lineages do not exist in this putatively pure HPG1 set (Fig. 4). For comparison, random pairs of *A. thaliana* accessions from the native range or pairs of non-HPG1 typically differ by about 500 SNPs / 100 kb [21] (see scale in Fig. 2A).

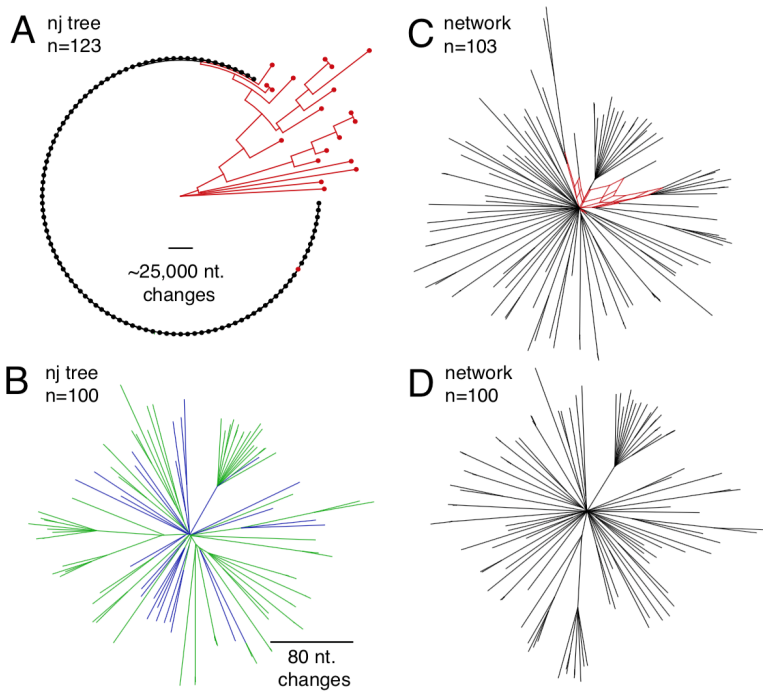


Figure 2. Relationship among herbarium and modern samples.

(A) Neighbor joining tree with all 123 samples (dots) and rooted with the most distant sample. The black clade of almost-identical samples is the HPG1 lineage. Scale line shows the equivalent branch length of over 25,000 nucleotide changes. **(B)** Neighbor joining tree only with the HPG1 black clade from (A). Colors represent herbarium (blue) and modern individuals (green). Scale line shows the equivalent branch length of 80 nucleotide changes. Note that no outgroup was included. **(C, D)** Network of samples using the parsimony splits algorithm, before **(C)** and after **(D)** removing three intra-HPG1 recombinants (in red). Note that the network algorithm returns in (D) a network devoid of any reticulation, which indicates absence of intra-haplogroup recombination.

There were no SNPs in mitochondrial nor chloroplast genomes, which already suggested a recent common origin, and genome-wide nuclear diversity (π = 0.000002, θ_W = 0.00001, with 5,013 full informative segregating sites) was two orders of magnitude lower than in the native range of the species (θ_W = 0.007) [21] (Table S1) (Supplementary Text 6). The population recombination parameter was also four orders of magnitude lower ($4N_e r = \rho = 3.0\text{x}10^{-6}$ cM bp^{-1}) than in the native range ($\rho = 7.5\text{x}10^{-2}$ cM bp^{-1}) [26] (Supplementary Text 6). While recombination occurs in every generation, regardless of self-fertilization or outcrossing, it is only observable after outcrossing between genetically non-identical individuals. We must stress that because *A. thaliana* can outcross at rates of several percent per generation [23,27], but because the HPG1 population is genetically so homogeneous, we are mostly "blind" to the consequences of outcrossing in this special case. The

lack of "observable recombination" in the genome is important, as it allows for the use of straightforward phylogenetic methods to calculate a mutation rate. The enrichment of low frequency variants in the site frequency spectrum (Tajima's *D* = -2.84; species mean = -2.04, [21]) and low levels of polymorphism are consistent with a recent bottleneck followed by population expansion (Fig. 3). The obvious explanation is that the strong bottleneck corresponds to a colonization founder event, likely by few closely related individuals or perhaps even a single plant.

Altogether these patterns indicate that the collection of HPG1 plants we investigated constitute a quasi-clonal and quasi-identical set of individual genomes, mostly devoid of observable recombination and population structure, and thus eminently suited for the study of naturally arising *de novo* mutations.

The genome-wide substitution rate

It is important to distinguish between the *mutation rate*, which is the rate at which genomes change due to DNA damage, faulty repair, gene conversion and replication errors, and *substitution rate*, which is the rate at which mutations survive and accumulate under the influence of demographic processes and natural selection [28,29]. Under neutral evolution, mutation and substitution rates should be equal [29]. The simple evolutionary history of the HPG1 population enables direct estimates of substitution rates, and the comparison of theses between different genome annotations, as well as with mutation rates from controlled conditions experiments, could reveal the role played by both demographic and selective forces.

To estimate the substitution rate in the HPG1 lineage, we used distance- and phylogeny-based methods that take advantage of the known collection dates (Supplementary Text 7). The distance method is independent of recombination and has been previously applied to viruses [30] and humans [31]. The substitution rate is calculated from correlation between differences in collection time in historic-modern sample pairs, and the number of nucleotide differences between those pairs relative to a reference (Fig. 3C), scaled to the size of the genome accessible to Illumina sequencing. This method resulted in an estimated rate of 2.11×10^{-9} substitutions site^{-1} year^{-1} (95% bootstrap Confidence Interval [CI]: $1.88–2.33 \times 10^{-9}$) using rigorous SNP calling quality thresholds. Relaxing the thresholds for base calling and minimum genotyped rate affects both the number of called SNPs and the length of the interrogated reference sequence [32]. These largely cancelled each other out, and the adjusted estimates were relatively stable, between $2.1–3.2 \times 10^{-9}$ substitutions site^{-1} year^{-1} (Table S3, Supplementary Text 3).

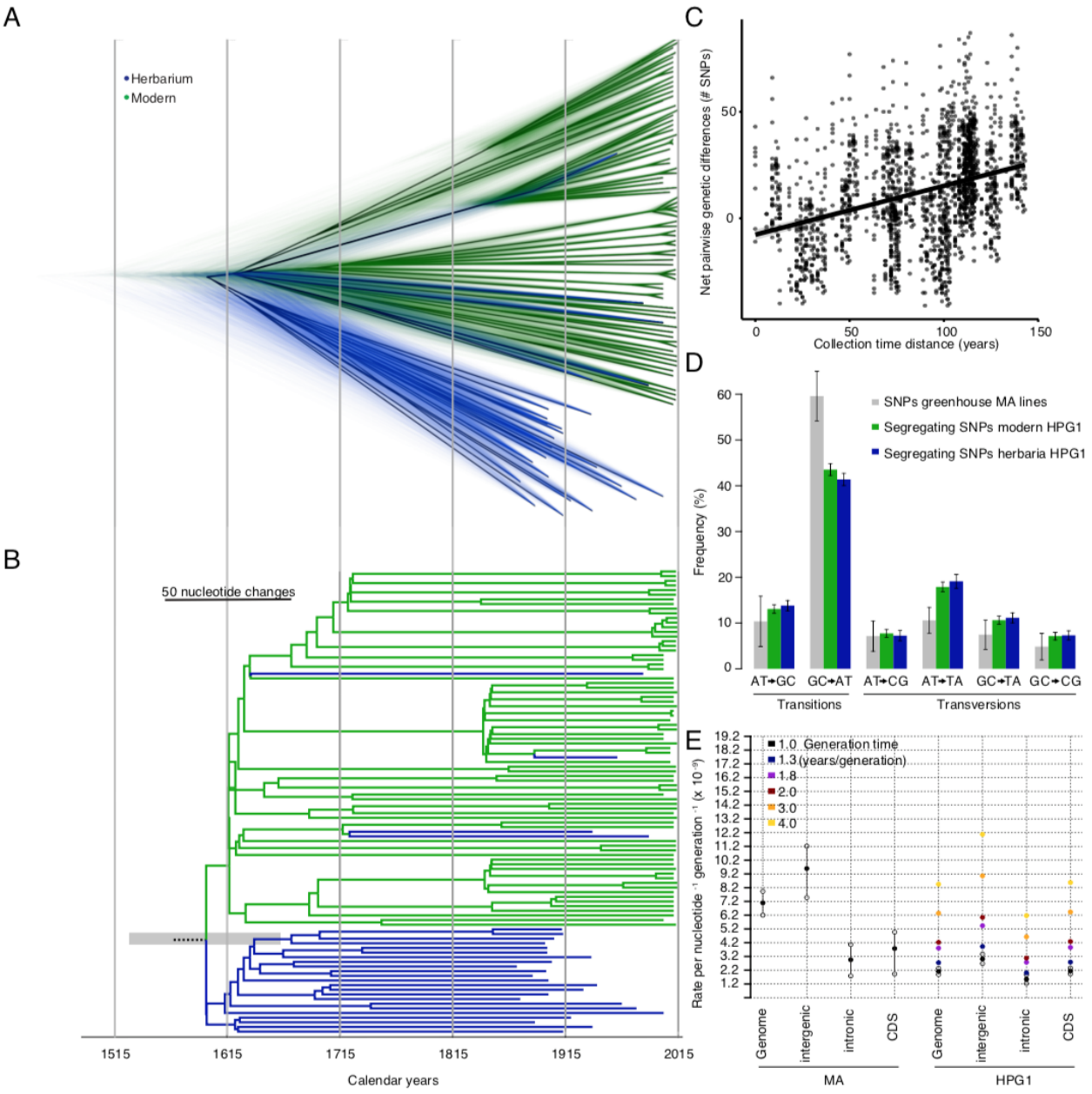


Figure 3. Substitution rates.

(A) Bayesian phylogenetic analyses employing tip-calibration. A total of 10,000 trees were superimposed as transparent lines, and the most common topology was plotted solidly. Tree branches were calibrated with their corresponding collection dates. **(B)** Maximum Clade Credibility (MCC) tree summarizing the trees in (A). Note the scale line shows the equivalent branch length of 50 nucleotide changes. The grey transparent bar indicates the 95% Highest Posterior Probability of the root date. **(C)** Regression between pairwise net genetic and time distances. The slope of the linear regression line corresponds to the genome substitution rate per year. **(D)** Substitution spectra in HPG1 samples, compared to greenhouse-grown mutation accumulation (MA) lines. **(E)** Comparison of genome-wide, intergenic, intronic, and genic substitution rates in HPG1 and mutation rates in greenhouse-grown MA lines. Substitution rates for HPG1 were re-scaled to a per generation basis assuming

different generation times. Confidence intervals in HPG1 substitution rates were obtained from 95% confidence intervals of the slope from 1,000 bootstraps (Table S4 for actual values).

The second method, a Bayesian phylogenetic approach, uses the collection years for tip-calibration and assumes a relaxed molecular clock. It summarizes thousands of plausible coalescent trees, and it has been extensively used to calculate evolutionary rates in various organisms [33–35]. This method yielded a substitution rate of 4.0×10^{-9}, with confidence ranges overlapping the above estimates (95% Highest Posterior Probability Density [HPPD]: $3.2–4.7 \times 10^{-9}$).

Based on the similar results obtained with two very different methods, we can confidently say that the substitution rate in the wild populations of HPG1 is between 2 and 5 $\times 10^{-9}$ site^{-1} year^{-1}.

To date the colonization of N. America by HPG1 *A. thaliana* and to improve the description of intra-HPG1 relationships compared to that from a NJ tree, we further used a Bayesian phylogeny. At first sight, the 73 modern samples appeared separated from the herbarium samples (Fig. 3B), but the superimposition of thousands of possible trees showed that the apparent separation of samples was less clear near the root (Fig. 3A). Long terminal branches reflected that the majority of the variants are singletons, typical of populations that expand after bottlenecks.

The mean estimate of the last common HPG1 ancestor, the average tree root, was the year 1597 (HPPD 95%: 1519–1660) (Fig. 3A, B), and an alternative non-phylogenetic method gave a similar estimate, 1625. Both estimates are older than a previously suggested date in the 19th century, using a laboratory mutation rate estimate and having no information from herbarium samples [25]. Because HPG1 appears to have been the most abundant lineage in N. America since the 1860s, we believe it could have been one of the first, if not the first colonizer that could establish itself in N. America. If that is true, the time of coalescence of the HPG1 diversity could be close to the time of HPG1 introduction to N. America. During the colonial period, many European immigrants settled on the East coast, consistent with N. American *A. thaliana* lineages being genetically closest to British and coastal West European populations [21]. Coincidently, the oldest herbarium samples (12 out of the 27) were HPG1 and came from the East Coast, and we found a significant correlation between collection date and both latitude and longitude (Fig. 1C). This could indicate that after the colonization they moved from the East Coast to the Midwest – the other main area of the distribution that experienced an agricultural expansion in the 19th century [36]. Still, these conclusions need to be treated with caution, since regardless of the robustness of the results and our attempts to sample evenly from available collections, there could be unknown biases in the 19th century herbaria.

Mutation spectra across genome annotations

Although for dating divergence events a substitution rate expressed in years is ideal, in order to compare substitution and mutation rates, both need to be expressed per generation. While *A. thaliana* is an annual plant, seed bank dynamics generate a delay of average generation time at the population scale. A comprehensive study of multiple *A. thaliana* populations in Scandinavia found that dormant seeds could wait for longer than a year in the seed bank, generating overlapping generations and an delayed average generation time of 1.3 years [37] with a notable variance across populations. Multiplication by the mean generation time led to an adjusted rate of $2.7x10^{-9}$ substitutions $site^{-1}$ $generation^{-1}$ (95% CI $2.4\text{-}3.0x10^{-9}$) (Fig. 3E). To be able to compare this rate with a reference, we also re-sequenced mutation accumulation (MA) lines in the Col-0 reference background grown under controlled conditions in the greenhouse that had been analyzed before with less advanced short read sequencing technology [38]. From the new re-sequencing data, we obtained an updated rate of $7.1x10^{-9}$ mutations $site^{-1}$ $generation^{-1}$ (95% CI $6.3–7.9x10^{-9}$) (Tables S2, S3, Supplementary Text 4 and 7). This mutation rate is two- to three-fold higher than the per-generation substitution rate estimate in the wild, but within the same order of magnitude. The same holds for rates in different genome annotations, i.e. genic, intronic and intergenic regions, but the confidence intervals overlapped in many cases (Table S3).

Differences in per-generation rates between laboratory and wild populations could stem from both methodological as well as biological causes. For instance, if the true average generation time was actually over 3 years / generation, the differences would cancel out (Fig. 3E). Limitations in mapping structural variation in non-reference samples could lower the substitution rate, which may explain why we calculated an atypically low substitution rate in regions with transposable elements (see Supplementary Text 7.2.1). Environmentally-driven effects that are not yet well understood, such as variable methylation status of cytosines, account for much of the variation in local substitution rates [39], and could increase or decrease the rate (see Supplementary Text 7.2.3, Fig. S4).

An alternative evolutionary explanation to the aforementioned laboratory and wild populations' rates differences is that purifying selection in the wild would slow down the accumulation of mutations by removing deleterious mutations (Fig. 3E). This has been observed before and is one of the accepted causes of the discrepancy between the so called long- and short-term substitution rates in a range of organisms [40].

In order to provide evidence for negative purifying selection acting in the wild, we performed three types of analyses involving comparisons across genomic annotations within the HPG1 dataset. Firstly, by calculating contingency tables and computing a Fisher's exact test, we compared the

deviation of expected and observed SNPs between coding regions (more likely under purifying selection), with intergenic regions, intronic regions, and all non-coding regions of genome. All three pairwise comparisons showed a depletion of coding SNPs and an enrichment of intergenic, intronic and non-coding SNPs (odds ratio>2, $p<10^{-16}$). An obvious explanation is that in genome annotations where a mutation is more likely to be deleterious, i.e. coding regions, the number of observed variants should be lower due to selection having removed them from the population before we could sequence them.

Secondly, we studied the Site Frequency Spectrum (SFS) of genetic variants. The rationale was that because purifying natural selection is more efficient at removing intermediate-frequency variants, variants that tend to be deleterious or slightly deleterious should be found at lower frequency than those that only suffer neutral drift [41]. We built contingency tables of coding, intergenic, intronic and non-coding variants segregating above and and below the conventional frequency cutoff of 5% to separate low- and intermediate-frequency variants [42]. We found that SNPs in coding regions were more likely to be at low frequency than those in intergenic (odds ratio=2.34, $p=3.09x10^{-11}$), intronic (odds ratio=1.48, p=0.02), and all non-coding regions (odds ratio=2.05, $p=1.29x10^{-8}$). We carried out the same analysis using nonsynonymous and synonymous SNPs, which are easily interpretable in terms of the selection regimes under which they evolve. We did not find an enrichment (p=0.67), perhaps due to an insufficient number of testable mutations (Table S3).

Thirdly, to verify that the full frequency spectrum of coding SNPs was shifted to lower frequencies (i.e. the results were not dependent on the arbitrary 5% frequency cutoff), we used the nonparametric Kolmogorov-Smirnov test for two samples. We found that the cumulative distribution of the site frequency spectrum (CD_{SFS}) of coding regions is above (i.e., the frequency distribution is overall skewed to lower values) both the intergenic CD_{SFS} ($p=3.25x10^{-6}$) and the non-coding regions CD_{SFS} (p=0.001), but not the intronic CD_{SFS} (p=0.60) (Fig. S5). As in our previous analysis, the comparison between the nonsynonymous and synonymous CD_{SFS} yielded, likely for similar reasons, no differences (p=0.53).

All in all, these results support that purifying selection is a force shaping to some degree the diversity across the HPG1 genome and might therefore as well contribute to the differences between HPG1 and MA rates.

Potentially advantageous *de novo* mutations

Finally, having discovered over 5,000 *de novo* mutations in the HPG1 lineage, we wondered whether there is any evidence for an adaptive role of these *de novo* mutations in the colonization of N.

America by HPG1. We noted that some new mutations had risen to intermediate or even high frequencies in the HPG1 samples. This might have been the consequence of drift from stochastic demographic processes, or it could have been caused by positive natural selection. To find direct evidence for the latter, we grew the modern accessions in a common garden and studied phenotypes of known importance in ecology of invasions [43], namely flowering time and root traits (see Supplementary Text 8). Using linear mixed models, we calculated the proportion of variance explained (also called narrow sense heritability, h^2) with a kinship matrix of all SNPs that had become common (>5%, n=391). We found significant heritable variation for multiple traits including the growth rate in length (h^2=0.64) and the average root gravitropic direction (h^2=0.54). As in our study mutations are the main source of genetic variants, these mutations — or mutations linked to them — should be responsible for significant quantitative variation in several traits (Table S4, Supplementary Text 10). The existence of mutation-driven phenotypic variation at least indicates that natural selection could have acted upon such phenotypic variation.

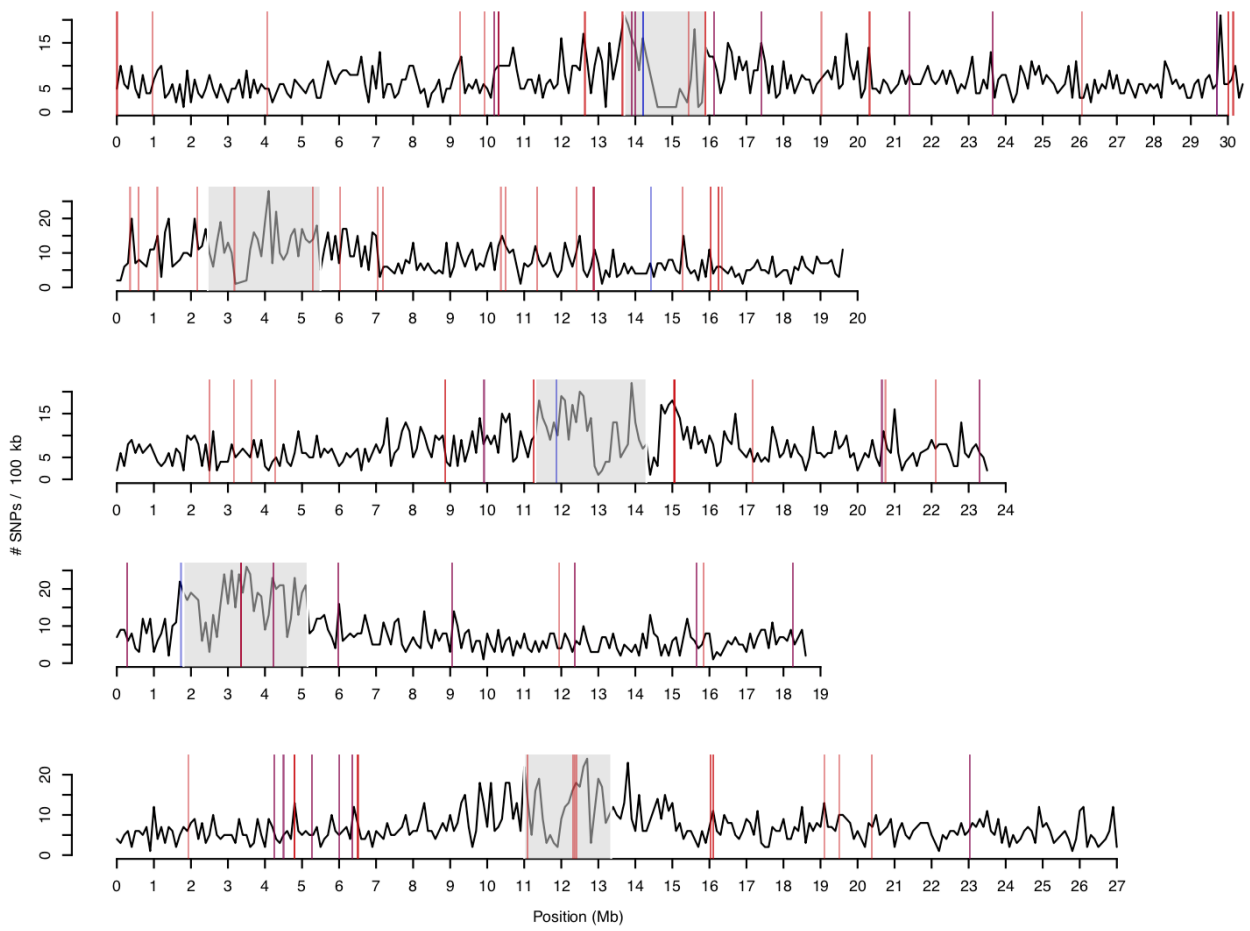


Figure 4. Density of SNPs along all chromosomes and location of GWAS hits

Black line shows number of SNPs per 100 kb window. Centromere locations are indicated by grey shading. Vertical lines indicate SNPs associated with root phenotypes (red) and climatic variables (blue) (Table 1 and Table S5).

Although linkage disequilibrium (LD) among SNPs is high, the fact that HPG1 genomes differ in very few SNPs greatly reduces the list of candidate loci that might generate the observed phenotypic variation (Fig. S7) [44]. With this reasoning in mind and understanding the limitations imposed by LD, we carried out a genome-wide association (GWA) analysis and found 79 SNPs associated with one or more root traits, mostly growth and directionality (Fig. 4). Twelve SNPs were in coding regions and seven resulted in nonsynonymous changes — some producing non-conservative amino-acid changes and thus likely to affect protein structure and/or function (Table 1, based on transition scores from [45]). Due to the aforementioned LD, in some cases the results of associations could not be confidently assigned to a specific SNP and thus we report the number of other associated mutations with $r^2 > 0.5$ (Table 1, Fig. S7). We note thatFor other cases,

we were able to pinpoint clear candidates that were not in LD with other SNPs and whose functional annotation had a strong connection to the phenotype (Table 1, Fig. S7). For example, one SNP associated with root gravitropism was not linked to any other SNP hit and it was found at 40% frequency (top 3% percentile). This SNP produces a cysteine to tryptophan change in AT5G19330, which is involved in abscisic acid response and confers salt tolerance when overexpressed [46]. Another nonsynonymous SNP associated with root growth is located in AT2G38910, which encodes a calcium-dependent kinase that is a factor regulating root hydraulic conductivity and phytohormone response *in vitro* [47,48].

Table 1. Genic SNPs associated with different traits.

For nonsynonymous SNPs, the amino acid change and the Grantham score (ranging from 0 to 215), which measures the physico-chemical properties of the amino acids, are reported. All SNPs in the table were significant ($p < 0.05$) after raw p-values were corrected by an empirical p-value distribution from a permutation procedure. * highlights those that also passed a double Bonferroni threshold, correcting by number of SNPs and number of phenotypes ($p < 0.0001$). LD corresponds to how many other SNP hits are in high linkage ($r^2>0.5$). Table S5 contains information on all significant SNPs and Table S4 for details on phenotypes and climatic variables.

Trait[T]	Location (chr-bp)	Gene	Anno–tation	Protein	aa change	LD	Bonf.
G	1-958,948	AT1G03810	nonsyn	Oligonucleotide binding	A>P, 27	53	
D	1-13,994,958	AT1G36933	transposon	Copia		49	
S	1-20,324,050	AT1G54440	intronic	RRP6-LIKE 1		11	*
D	1-23,648,407	AT1G63740	nonsyn	TIR-NLR family	Y>S, 144	46	
G	2-358,395	AT2G01820	syn	RLK family		43	*
G	2-585,918	AT2G02220	syn	PSKR1		42	*
G	2-6,034,545	AT2G14247	syn	Expressed protein		38	*
G	2-7,047,529	AT2G16270	nonsyn	Unknown protein	P>A, 27	37	*
G	2-7,186,220	AT2G16580	intronic	SAUR8		36	*
G	2-10,495,275	AT2G24680	intronic	B3 family		34	*
G	2-12,415,084	AT2G28900	intronic	OEP16		32	
S	2-16,039,488	AT2G38290	3' UTR	AMT2		8	*
S	2-16,247,290	AT2G38910	nonsyn	CPK20	A>G, 60	7	*
G	2-16,333,662	AT2G39160	nonsyn	Unknown protein	A>G, 60	29	
G	3-2,500,258	AT3G07830	syn	PGA3		28	*
G	3-3,629,794	AT3G11530	intronic	VPS55		26	*
G	3-4,269,626	AT3G13229	5' UTR	DUF868 domain		25	*
D	3-11,873,293	AT3G30219	transposon	Gypsy		0	
G & D	4-4,228,138	AT4G07440	transposon	Oligonucleotide binding		19	

G & D	4-9,046,942	AT4G15960	nonsyn	Alpha/beta-hydrolase	A>Q, 24	18	
G & D	4-15,646,341	AT4G32410	syn	ANY1		15	
G	4-15,845,001	AT4G32840	3' UTR	PFK6		14	
D	5-4,245,213	AT5G13260	syn	Unknown protein		12	
D	5-4,500,202	AT5G13950	nonsyn	Unknown protein	A>G, 60	11	
G	5-4,797,923	AT5G14830	transposon	Retrotransposon		10	
G	5-6,508,329	AT5G19330	nonsyn	ARIA	C>W, 215	0	
G	5-11,090,365	AT5G29037	transposon	Gypsy		4	
G	5-12,312,975	AT5G32630	pseudogene	–		3	
G	5-12,358,159	AT5G32825	transposon	CACTA		2	
S	5-16,024,197	AT5G40020	intronic	Thaumatin superfamily		2	*

†Traits with significant associations were root gravitropism (G), size (S), or low summer precipitation.

Nineteen other SNPs were associated with climate variables after correction for latitude and longitude (www.worldclim.org, Table S4), and generally tended to coincide with top root-associated SNPs (odds ratio = 3.9, Fisher's Exact test p = 0.002; Fig. 4, and Table S5). Specifically, this means that alleles increasing root length and gravitropic growth were present in areas with lower precipitation, and *vice versa* (Pearson's correlation r=0.85, p=0.003). This indicates that phenotypic variation generated by mutations coincides with environmental (and not geographic) gradients along the colonized areas. Compared to other mutations with matched allele frequencies, root-associated mutations are first found in older herbarium samples nearer to Lake Michigan (Fig. S6), the area in N. America that seems to be most densely populated by *A. thaliana* [21]. A moreThis could be explained by natural selection having maintained mutations with phenotypic effect for a longer time than neutral mutations or perhaps that these mutations were selected for in a new environment. All in all, our results are compatible with natural positive selection having already acted on root morphology variation that was generated by *de novo* mutations in this colonizing lineage. To confirm such hypotheses of local adaptation by *de novo* mutations, it will be necessary to grow collections of divergent HPG1 individuals in multiple contrasting locations over several years, and ideally revive historical specimens to compare performance [49].

Conclusions

In summary, we have exploited whole-genome information from historic and contemporary collections of a herbaceous plant to empirically characterize evolutionary forces during a recent colonization. With this natural time series experiment we could directly estimate the nuclear substitution rate in wild *A. thaliana* populations – a parameter difficult to characterize experimentally [9]. This allowed us to date the colonization time and spread of HPG1 in N. America. We provide

evidence that purifying selection has already changed the site frequency spectrum in the course of just a few centuries. Finally, we discovered that a small number of *de novo* mutations that rose to intermediate frequency can together explain quantitative variation in root traits across environments. This strengthens the hypothesis that some *de novo* variation could have had an adaptive value during the colonization and expansion process, a hypothesis that has been put forward as one of the possible solutions to the genetic paradox of invasion in plants [17]. This process might be more relevant in self-fertilizing plants, which typically have less diversity than outcrossing ones [50], but have higher growth rates [43] and account for the majority of successful plant colonizers [5]. While *A. thaliana* HPG1 is not an invasive, i.e. harmful, species, it can teach us about fundamental evolutionary processes behind successful colonizations and adaptation to new environments. Our work should encourage others to search for similar natural experiments and to unlock the potential of herbarium specimens to study "evolution in action".

METHODS

Sample collection and DNA sequencing

Modern *A. thaliana* accessions were from the collection described by Platt and colleagues [23], who identified HPG1 candidates based on 149 genome-wide SNPs (Table S1, Supplementary Text 1). Herbarium specimens were directly sampled by Max Planck colleagues Jane Devos and Gautam Shirsekar, or sent to us by collection curators from various herbaria (Table S1, Supplementary Text 1). Among the substantial number of specimens in the herbaria of the University of Connecticut, the Chicago Field Museum and the New York Botanical Garden, we selected herbarium specimens spaced in time so there was at least one sample per decade starting from the oldest record (1863). The differences in geographic biases of herbarium and modern collections are difficult to know [2], thus we did choose both historic and modern samples that were as regularly distributed in space as possible, and sample overlapping locations wherever possible. DNA from herbarium specimens was extracted as described [51] in a clean room facility at the University of Tübingen. Two sequencing libraries with sample-specific barcodes were prepared following established protocols, with and without repair of deaminated sites using uracil-DNA glycosylase and endonuclease VIII (refs. [52–54]) (Supplementary Text 2). We also investigated patterns of DNA fragmentation and damage typical of ancient DNA [24] (Supplementary Text 2). DNA from modern individuals was extracted from pools of eight siblings using the DNeasy plant mini kit (Qiagen, Hilgendorf, Germany). Genomic DNA libraries were prepared using the TruSeq DNA Sample or TruSeq Nano DNA sample prep kits (Illumina, San Diego, CA), and sequenced on Illumina HiSeq 2000, HiSeq 2500 or MiSeq instruments. Paired-end reads from modern samples were trimmed and quality filtered before mapping using the SHORE

pipeline v0.9.0 [25,55]. Because ancient DNA fragments are short (Fig. S1) we merged forward and reverse reads for herbarium samples after trimming, requiring a minimum of 11 bp overlap [51], and treated the resulting as single-end reads. Reads were mapped with GenomeMapper v0.4.5s [56] against an HPG1 pseudo-reference genome [25], and against the Col-0 reference genome, and SNPs were called with SHORE for the HPG1 pseudo-reference genome mappings [25,57] using different thresholds (Supplementary Text 3). Average coverage depth, number of covered genome positions, and number of SNPs identified per accession relative to HPG1 are reported in Table S1. We also re-sequenced the genomes of twelve Col-0 MA lines [57,58] (Table S2) (Supplementary text 4) to recalculate and update the laboratory mutation rate from Ossowski et al. [38] with the newer sequencing technologies.

Phylogenetic methods and genome-wide statistics

We used the Pegas, Ape and Adegenet packages in R [59–61] to manipulate and visualize the genetic distances of all samples as well as the HPG1 subset (Supplementary Text 7). We constructed parsimony networks using SplitsTree v.4.12.3 [62], with confidence values calculated with 1,000 bootstrap iterations. We built Maximum Clade Credibility Trees using the Bayesian phylogenetic tools implemented in BEAST v.1.8 [63] (see below).

Transforming the variant sites into a FASTA format, we estimated genetic diversity as Watterson´s θ [64] and nucleotide diversity π, and the difference between these two statistics as Tajimas's *D [65]* using DnaSP v5 [66]. Then we re-scaled the estimates using the sequencing-accessible genome sizes (Table S3). We estimated pairwise linkage disequilibrium (LD) between all possible combinations of informative sites, ignoring singletons, by computing r^2, *D* and *D*' statistics using DnaSP v5 [66]. For the modern individuals, we calculated the recombination parameter rho ($4N_e r$) also using DnaSP v5 [66].

Substitution and mutation rate analyses

Similarly as in Fu et al. [67], we used genome-wide nuclear SNPs to calculate pairwise "net" genetic distances using the equation $D'_{ij} = D_{ic}-D_{jc}$, where D'_{ij} is the net distance between a modern sample *i* and a herbarium sample *j*; D_{ic} the distance between the modern sample *i* and the reference genome *c*; and D_{jc} is the distance between a modern sample (j) and the reference genome (c). We calculated a pairwise time distance in years between the collection times, *T*'ij, and calculated the linear regression: $D' = a+bT'$. The slope coefficient *b* describes the number of substitution changes per year. We used either all SNPs or subsets of SNPs at different annotations (genic, intergenic etc.) appropriately scaled by accessible genome length. Because the points used to calculate the

regression are non-independent, a bootstrap has been recommended to overcome to a certain extent the anti-conservative confidence intervals [30] (Supplementary Text 7 and Fig. S3).

To fully account for the non-independence of points, we need to work with phylogenies. The Bayesian phylogenetics approach we used is implemented in BEAST v1.8 [63] and is called tip-calibration, and calculates a substitution rate along the phylogeny. Our analysis optimized simultaneously and in an iterative fashion using a Monte Carlo Markov Chain (MCMC) a tree topology, branch length, substitution rate, and a demographic Skygrid model (Supplementary Text 7). The demographic model is a Bayesian nonparametric one that is optimized for multiple loci and that allows for complex demographic trajectories by estimating population sizes in time bins across the tree based on the number of coalescent - branching - events per bin [68]. We also performed a second analysis run using a fixed prior for substitution rate of $3x10^{-9}$ substitutions $site^{-1}$ $year^{-1}$ based on our previous net distance estimate to confirm that the MCMC had the same parameter convergence, e.g. tree topology, as in the first "estimate-all-parameters" run.

Having a substitution rate per year we can estimate the time to the most common recent ancestor *L* solving $d = 2L \times \mu$ where *d* is the average pairwise genetic distance between our samples and μ is the calculated substitution rate from the distance method. This yielded 363 years, which subtracted to the average collection date of the samples, produced a point estimate of 1615. We compare this estimate with the inferred phylogeny root from the BEAST analysis.

Inference of genome-wide selection

We separately analyzed sequences at different annotations, since as they might be under different selection regimes (i.e. evolutionary constraints). We computed, using the HPG1 dataset, one-tailed Fisher's exact test using the base stats package in R [69] on contingency tables of the total number of base pairs against the number of SNPs, and those separated by positions being annotated as a coding against non-coding (intergenic, intronic, all other noncoding). The test returned whether coding regions have a lower number of SNPs than other reference annotation (intronic, interenic, all non-coding regions), as expected by the total number of positions in the genome annotated as such. We also constructed contingency tables to test whether SNPs annotated as coding compared to those annotated as non-coding were more likely to be found at low (<5%) or intermediate (5≥%) frequency.

Finally, we calculated the unfolded Site Frequency Spectrum (SFS) based on the order of appearance of genetic variants in the herbarium dataset. We then used the Kolmogorov–Smirnov two-samples test and 10,000 bootstrap resampling using the R package Matching v. 4.9-2 (ref. [70]) to calculate whether the frequency spectrum was lower for coding SNPs than for other SNPs.

Additionally, we also repeated these analyses comparing nonsynonymous and synonymous mutations instead of coding and non-coding regions.

Association analysis

We collected flowering, seed and root morphology phenotypes for 63 accessions (Supplementary Text 8). For associations with climate parameters, we followed a similar rationale as previously described [71]. We extracted information from the bioclim database (http://www.worldclim.org/bioclim) at a 2.5 degrees resolution raster and intersected it with geographic locations of HPG1 samples (n = 100). We performed association analyses under several models and *p*-value corrections using the R package GeneABEL [72] (Supplementary Text 8.2). To calculate the variance of the trait explained by all genetic variants, we used a linear mixed model: $y = Xb + Zu + \varepsilon$; where y is the phenotype or climate variable, X is the genotype states at a given SNP, b is the fixed phenotypic effect of such SNP, Z is the design matrix of genome identities, u is the random genome background effect informed by the kinship matrix and distributed as MVN (0, $\sigma_g A$), and ε is the random error term. The ratio of σ_g / σ_T is commonly called narrow sense heritability, "chip" heritability, or proportion of variance explained by genotype [73]. Only SNPs with MAF>5% (n=391) were used to build a kinship or relationship matrix A. Note that the differences between any two genotypes were of the order of one or few dozens of SNPs. While this approach is appropriate to calculate a chip heritability, it would not be very useful to detect significant SNP, as the random factor accumulates all the available variation (Table S4). We therefore run a regular GWA model without kinship matrix: $y = Xb + \varepsilon$; but generated a p-value empirical null distribution based on running such model over 1,000 permuted datasets, which lead to conservative significance calculation (Fig. S7, Data Appendix S1). The p-values from running the association in the real data that were below the 5% tail in the empirical distribution could be considered significant. However, we also established a conservative "double" Bonferroni correction, where the significant threshold was lowered to 0.01% (= 5% / [number of SNPs + number of phenotypes tested]). All significant SNPs are shown in Table S5, and a subset in Table 1. Although many phenotypic traits did not have significant SNPs, we show all the QQ plots in the Data Appendix S1 file.

Accession numbers. Short reads have been deposited in the European Nucleotide Archive under the accession number https://www.ebi.ac.uk/ena/data/view/PRJEB24619.

Online Content This article contains supplementary information including data sets, extended methods and supplementary figures at https://doi.org/10.1371/journal.pgen.1007155.

Acknowledgments For providing and retrieving herbarium specimens, we thank R. Capers, J. Devos, G. Shirsekar, M. S. Dossmann, J. Freudenstein, C. M. Herring, C. Niezgoda, C. A. McCormick, J. Peter and M. Thines. We thank X. Zhao and I. Henderson for recombination estimates, C. Lanz for sequencing support, C. Goeschl, B. Zierfuss and B. Wohlrab for help with root analyses, and P. Lang, D. Seymour, and D. Koenig for thorough proofreading and comments on the manuscript. We thank to Robert Colautti for useful comments on the theoretical framing of the manuscript, M. Nordborg for discussions and pointing us to the work of A.R. Templeton, K. Pruefer for input on data analysis, and the Weigel and Burbano labs for comments. Supported by the President's Fund of the Max Planck Society (project "Darwin"), ERC (AdG IMMUNEMESIS) and core funds of the Max Planck Society.

Author Contributions H.A.B. and D.W. conceived and supervised the project, and coordinated the collaborative effort. J.B. coordinated the collection of modern seed samples. C.J., B.B. and J.B. performed and analyzed flowering time and seed set greenhouse experiments. C.S. and R.S. performed and analyzed root assays and seed size measurements under the supervision of W.B.; C.B. and J.H. sequenced and curated modern samples, coordinated by D.W.; H.A.B. coordinated the collection and analysis of herbarium samples. J.K. coordinated the extraction of DNA and library preparation of herbarium samples. V.J.S. and E.R. prepared sequencing libraries from herbarium specimens. C.B. called variants in HPG1. J.H. called variants in mutation accumulation lines. M.E.A. performed the population and quantitative genomic analyses with supervision of R.N., C.B. and H.A.B. The first draft was written by M.E.A. and the final manuscript was written by M.E.A., C.B., H.A.B. and D.W. with comments from all coauthors.

Authors declare no conflict of interests.

REFERENCES

1. Green RE, Shapiro B. Human evolution: turning back the clock. Curr Biol. Elsevier; 2013;23: R286–8. doi:10.1016/j.cub.2013.02.050

2. Crawford PHC, Hoagland BW. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? J Biogeogr. 2009;36: 651–661. doi:10.1111/j.1365-2699.2008.02043.x

3. Colautti RI, Lau JA. Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. Mol Ecol. 2015;24: 1999–2017. doi:10.1111/mec.13162

4. van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, et al. Global exchange and accumulation of non-native plants. Nature. Nature Research; 2015;525: 100–103. doi:10.1038/nature14910

5. Razanajatovo M, Maurel N, Dawson W, Essl F, Kreft H, Pergl J, et al. Plants capable of selfing are more likely to become naturalized. Nat Commun. Nature Publishing Group; 2016;7: 13313. doi:10.1038/ncomms13313

6. Sax DF, Stachowicz JJ, Brown JH, Bruno JF, Dawson MN, Gaines SD, et al. Ecological and evolutionary

insights from species invasions. Trends Ecol Evol. 2007;22: 465–471. doi:10.1016/j.tree.2007.06.009

7. Gauze GF. The struggle for existence. Baltimore: The Williams & Wilkins company; 1934.

8. Hardouin EA, Tautz D. Increased mitochondrial mutation frequency after an island colonization: positive selection or accumulation of slightly deleterious mutations? Biol Lett. 2013;9: 20121123. doi:10.1098/rsbl.2012.1123

9. Halligan DL, Keightley PD. Spontaneous mutation accumulation studies in evolutionary genetics. Annu Rev Ecol Evol Syst. 2009;40: 151–172. doi:10.1146/annurev.ecolsys.39.110707.173437

10. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010;328: 636–639. doi:10.1126/science.1186802

11. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 1987;84: 9054–9058. Available: http://www.ncbi.nlm.nih.gov/pubmed/3480529

12. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 2012;10: e1001388. doi:10.1371/journal.pbio.1001388

13. Pennings PS, Hermisson J. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol Biol Evol. 2006;23: 1076–1084. doi:10.1093/molbev/msj117

14. Karasov T, Messer PW, Petrov DA. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genet. 2010;6: e1000924. doi:10.1371/journal.pgen.1000924

15. Rouco M, López-Rodas V, Flores-Moya A, Costas E. Evolutionary changes in growth rate and toxin production in the cyanobacterium Microcystis aeruginosa under a scenario of eutrophication and temperature increase. Microb Ecol. 2011;62: 265–273. doi:10.1007/s00248-011-9804-0

16. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. Nature Publishing Group; 2012;13: 745–753. doi:10.1038/nrg3295

17. Estoup A, Ravigné V, Hufbauer R, Vitalis R, Gautier M, Facon B. Is There a Genetic Paradox of Biological Invasion? Annu Rev Ecol Evol Syst. 2016;47: 51–72. doi:10.1146/annurev-ecolsys-121415-032116

18. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends Ecol Evol. 2008;23: 38–44. doi:10.1016/j.tree.2007.09.008

19. Dlugosch KM, Parker IM. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol. 2008;17: 431–449. doi:10.1111/j.1365-294X.2007.03538.x

20. Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. Mol Ecol. 2015;24: 2095–2111. doi:10.1111/mec.13183

21. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell. Elsevier; 2016;166: 481–491. doi:10.1016/j.cell.2016.05.063

22. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, et al. African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. Proceedings of the National Academy of Sciences. 2017; doi:10.1073/pnas.1616736114

23. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The scale of population structure in Arabidopsis thaliana. PLoS Genet. 2010;6: e1000843. doi:10.1371/journal.pgen.1000843

24. Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. Royal Society Open Science. The

Royal Society; 2016;3: 160239. doi:10.1098/rsos.160239

25. Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, et al. Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS Genet. 2015;11: e1004920–e1004920. doi:10.1371/journal.pgen.1004920

26. Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, et al. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat Genet. Nature Publishing Group; 2013;45: 1327–1336. doi:10.1038/ng.2766

27. Bomblies K, Yant L, Laitinen R a., Kim S-T, Hollister JD, Warthmann N, et al. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet. 2010;6: e1000890–e1000890. doi:10.1371/journal.pgen.1000890

28. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. Nature Publishing Group; 2013;14: 827–839. doi:10.1038/nrg3564

29. Kimura M. On the evolutionary adjustment of spontaneous mutation rates. Genet Res. Cambridge University Press; 1967;9: 23–23. doi:10.1017/S0016672300010284

30. Drummond A, Pybus OG, Rambaut A. Inference of viral evolutionary rates from molecular sequences. Adv Parasitol. 2003;54: 331–358. Available: http://www.ncbi.nlm.nih.gov/pubmed/14711090

31. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514: 445–449. doi:10.1038/nature13810

32. Ness RW, Morgan AD, Colegrave N, Keightley PD. Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. Genetics. 2012;192: 1447–1454. doi:10.1534/genetics.112.145078

33. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7: 214–214. doi:10.1186/1471-2148-7-214

34. Millar CD, Dodd A, Anderson J, Gibb GC, Ritchie PA, Baroni C, et al. Mutation and evolutionary rates in adélie penguins from the antarctic. PLoS Genet. journals.plos.org; 2008;4: e1000209. doi:10.1371/journal.pgen.1000209

35. Christin P-A, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. Molecular dating, evolutionary rates, and the age of the grasses. Syst Biol. academic.oup.com; 2014;63: 153–165. doi:10.1093/sysbio/syt072

36. Klein Goldewijk K, Ramankutty N. Land cover change over the last three centuries due to human activities: The availability of new global data sets. GeoJournal. 2004;61: 335–344. doi:10.1007/s10708-004-5050-z

37. Falahati-Anbaran M, Lundemo S, Stenøien HK. Seed dispersal in time can counteract the effect of gene flow between natural populations of Arabidopsis thaliana. New Phytol. 2014;202: 1043–1054. doi:10.1111/nph.12702

38. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010;327: 92–94. doi:10.1126/science.1180677

39. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43: 956–963. doi:10.1038/ng.911

40. Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, et al. Time-dependent rates of molecular evolution. Mol Ecol. 2011;20: 3087–3101. doi:10.1111/j.1365-294X.2011.05178.x

41. Charlesworth B, Charlesworth D. Elements of Evolutionary Genetics [Internet]. Roberts and Company Publishers; 2010. Available: https://books.google.de/books?id=dgNFAQAAIAAJ

42. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8: e1002822. doi:10.1371/journal.pcbi.1002822

43. van Kleunen M, Dawson W, Maurel N. Characteristics of successful alien plants. Mol Ecol. 2015;24: 1954–1968. doi:10.1111/mec.13013

44. Templeton AR, Sing CF, Kessling A, Humphries S. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics. 1988;120: 1145–1154. Available: http://www.ncbi.nlm.nih.gov/pubmed/3147219

45. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;185: 862–864. doi:10.1126/science.185.4154.862

46. Kim S, Choi H-I, Ryu H-J, Park JH, Kim MD, Kim SY. ARIA, an Arabidopsis arm repeat protein interacting with a transcriptional regulator of abscisic acid-responsive gene expression, is a novel abscisic acid signaling component. Plant Physiol. 2004;136: 3639–3648. doi:10.1104/pp.104.049189

47. Li G, Boudsocq M, Hem S, Vialaret J, Rossignol M, Maurel C, et al. The calcium-dependent protein kinase CPK7 acts on root hydraulic conductivity. Plant Cell Environ. 2015;38: 1312–1320. doi:10.1111/pce.12478

48. Choi H-I, Park H-J, Park JH, Kim S, Im M-Y, Seo H-H, et al. Arabidopsis calcium-dependent protein kinase AtCPK32 interacts with ABF4, a transcriptional regulator of abscisic acid-responsive gene expression, and modulates its activity. Plant Physiol. 2005;139: 1750–1761. doi:10.1104/pp.105.069757

49. Franks SJ, Weis AE. A change in climate causes rapid evolution of multiple life-history traits and their interactions in an annual plant. J Evol Biol. 2008;21: 1321–1334. doi:10.1111/j.1420-9101.2008.01566.x

50. Arunkumar R, Ness RW, Wright SI, Barrett SCH. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics. 2015;199: 817–829. doi:10.1534/genetics.114.172809

51. Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, et al. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. eLife. 2013;2: e00731. doi:10.7554/eLife.00731

52. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010: db.prot5448. doi:10.1101/pdb.prot5448

53. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010;38: e87. doi:10.1093/nar/gkp1163

54. Kircher M. Analysis of High-Throughput Ancient DNA Sequencing Data. In: Shapiro B, Hofreiter M, editors. Ancient DNA. Humana Press; 2011. pp. 197–228. doi:10.1007/978-1-61779-516-9_23

55. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res. 2008;18: 2024–2033. doi:10.1101/gr.080200.108

56. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, et al. Simultaneous alignment of short reads against multiple genomes. Genome Biol. 2009;10: R98. doi:10.1186/gb-2009-10-9-r98

57. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature. nature.com; 2011;480: 245–249. doi:10.1038/nature10555

58. Shaw RG, Byers DL, Darmo E. Spontaneous mutational effects on reproductive traits of arabidopsis thaliana. Genetics. 2000;155: 369–378. Available: http://www.ncbi.nlm.nih.gov/pubmed/10790410

59. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24:

1403–1405. doi:10.1093/bioinformatics/btn129

60. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20: 289–290. Available: https://www.ncbi.nlm.nih.gov/pubmed/14734327

61. Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. Oxford University Press; 2010;26: 419–420. doi:10.1093/bioinformatics/btp696

62. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23: 254–267. doi:10.1093/molbev/msj030

63. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29: 1969–1973. doi:10.1093/molbev/mss075

64. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7: 256–276. Available: http://www.ncbi.nlm.nih.gov/pubmed/1145509

65. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. Genetics; 1989;123: 585–595. Available: http://www.genetics.org/content/123/3/585

66. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25: 1451–1452. doi:10.1093/bioinformatics/btp187

67. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol. 2013;23: 553–559. doi:10.1016/j.cub.2013.02.044

68. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard M a. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2012;30: 713–724. doi:10.1093/molbev/mss265

69. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available: https://www.R-project.org/

70. Sekhon JS. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R [Internet]. Journal of Statistical Software. 2011. pp. 1–52. Available: http://www.jstatsoft.org/v42/i07/

71. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, et al. Adaptation to climate across the Arabidopsis thaliana genome. Science. 2011;334: 83–86. doi:10.1126/science.1209244

72. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007;23: 1294–1296. doi:10.1093/bioinformatics/btm108

73. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. Nature Research; 2010;42: 565–569. doi:10.1038/ng.608

Supplementary Information Guide for:

Exposito-Alonso, Becker et al.**: The rate and evolutionary relevance of new mutations in a colonizing plant lineage**

Text S1: Detailed methods and analyses

1. Sample collection and preparation

Seeds from modern accessions (Table S1) were bulked at the University of Chicago. Progeny for DNA extraction was grown at the Max Planck Institute for Developmental Biology. We used 2 to 8 mm^2 of dried tissue for destructive sampling from the herbarium specimens (Table S1).

2. Authenticity of aDNA

First, unrepaired sequencing herbarium libraries were screened for authenticity by sequencing at low coverage on Illumina HiSeq 2500 or MiSeq instruments. To verify the DNA retrieved from historical samples of *A. thaliana* was authentic, we checked the percentage of endogenous DNA of the sample (Fig. S1A) as well as typical postmortem DNA damages: high fragmentation of DNA (Fig. S1B), enrichment of substitution from C to T at the first base pair (Fig. S1C) as well as purine enrichment at breakpoints of DNA fragments (Fig. S1D) (for details see [1]). Sequencing to produce the final genomes (101 bp paired end) was carried out on an Illumina HiSeq 2000 instrument after DNA repair by uracil-DNA glycosylase [2–4]. For a detailed analysis of authenticity in a fraction of our samples, see Weiss et al. [1].

3. SNP calling thresholds

To assess the effect of SNP calling thresholds on the mutation rate, we employed three different SHORE v0.9.0 quality thresholds following previous work (see Table S4 from [5]): allowing at most one intermediate penalty in all strains (most stringent threshold; “32-32”); requesting that at least one strain had at most one intermediate penalty, while all others were allowed up to two high and one intermediate penalties (intermediate stringency, “32-15”); and finally allowing one high and one intermediate penalty for all strains (most lenient stringency, “24-24”). On top of that, we would either allow missing information per SNP in up to 50% of accessions, or request complete information (0% missing rate). Thus, the most rigorous case would be 32-32 quality and 0% missing rate, and the most relaxed 24-24 quality and 50% maximum missing rate. Substitution rate calculations (section 7.2) were done for datasets from all combinations of these quality parameters (Fig. S3), and we chose the regular 32_15 quality threshold and complete information for the final estimate (Fig 3 C, E).

4. Resequencing of Col-0 Mutation Accumulation lines

We also sequenced the genomes of twelve greenhouse-grown mutation accumulation (MA) lines, including ten that had been sequenced at lower coverage before [5,6] (Table S2). We called SNPs,

indels and structural variants (SVs), following the workflow and parameters described [7], but without iterations. This procedure resulted in 2,203 polymorphisms shared by all lines, indicating errors in the reference sequence (12% of variants replaced N's in the TAIR9 genome) or genetic differences in the founder plant of the MA population compared to the Col-0 reference genome. In addition, we identified 388 segregating variants across the twelve lines (Table S2), of which 350 were singletons. This analysis revealed on average 25.5 SNPs, 4.9 deletions and 3.2 insertions per MA line at the 31^{st} generation (Table S2), compared to 19.6 SNPs, 2.4 deletions and 1.0 insertions previously detected in the 30^{th} generation with shorter read length and lower read depth [8]. The genome length accessed in this sequencing effort, 115,954,227 bp, was used to scale the number of point mutations to a rate of 7.1×10^{-9} mutations $site^{-1}$ $generation^{-1}$ (Table S3, Fig. 3E).

5. Identification of *bona fide* HPG1 accessions and mutations

5.1 HPG1 and other haplogroups in North America

The modern samples had been originally selected based on previous genotyping efforts of about 2,000 N. American accessions with for 149 nuclear, intermediate-frequency SNPs. This work had pointed to there being a single haplogroup, HPG1, that was invariant at these 149 markers and that accounted for about half of N. American individuals genotyped [9]. We extracted from the 123 genomes we had completely sequenced the same 149 SNPs and built a neighbour joining tree (Fig. S1A). We also built the same tree with the whole-genome sequences (Fig. S1B), which was mostly in agreement with the 149 SNP tree.

The previous work had identified several other haplogroup in N. America [9]. Not surprisingly, HPG1 individuals outcross with other lineages, and this accounts for some of the individuals which we later removed, because they did not agree completely in all 149 markers with the HPG1 consensus.

5.2 North american private diversity

Having identified these *bona fide* HPG1 individuals, we wanted to confirm that the diversity has a legitimate origin from *de novo* mutations. For that we used the 1001 Genomes resource (www.1001genomes.org), which covers a sampling of populations from the native Eurasian and African range. Subsetting the genomes from this resource to only European accessions, and limiting the SNP set to those with ≥1% frequency of alternative alleles and a maximum of 50% missing data (the same quality rate as our HPG1 SNP call), there were 300 variants out of all 5,181 HPG1 variants that were also found in Europe or Asia (5.7%). Changing the maximum missing data to 10% we get a more conservative estimate of 1.8% overlap, while increasing the maximum missing data to 90%, we

get the anti-conservative estimate of 6.5% overlap. Only one of the reported SNPs associated with phenotypes (see Section 8) was among these shared variants.

There are several scenarios that can explain these shared SNPs. One is simply that there was not a single founding seed, but a few of closely related individuals coming from the native range. Other explanations are that parallel mutations occurred in North America and Eurasia, that HPG1 individuals were reintroduced to Europe, or that reversion-mutation occurred in some HPG1 individuals. The latter is not implausible given the large population size of the species and the fact that about 10% of all sites in the genome are SNPs in the 1001 Genomes collection. As explained in the main text, SNP sharing due to admixture with other lineages is extremely unlikely, as such cases should be evident as blocks of high SNP diversity along the genome (Fig. 4).

Finally, regarding chloroplast diversity, we did not find any SNP in the chloroplast of HPG1 individuals. This is probably because chloroplast mutation rates are much slower [10] and because the founder colonizers actually came from a small batch of seeds from an identical mother (chloroplast diversity in the native range is of 2,842 SNPs [11]).

6. Extent of linkage disequilibrium and recombination

We estimated pairwise linkage disequilibrium (LD) between all possible combinations of informative sites, ignoring singletons, by computing r^2, D and D' statistics. LD decay was estimated using a linear regression approach. Linkage disequilibrium parameter $|D'|$ did not decay with physical distance (intercept = 0.99, slope = 0.00) among all SNP pairs. Indeed 99.975% of pairwise SNP comparisons had $|D'|=1$ meaning that 99.975% of those comparisons only three out of the four possible gametes (ab, aB, Ab, AB) are found and thus mutation alone can explain their existence without the need of invoking recombination. In other words, such three gametes can be represented in a tree structure. LD and recombination related statistics were determined using DnaSP v5 [12].

7. Substitution and mutation rate analyses

7.1 Greenhouse grown MA lines

Mutation rates were estimated for each 31st generation greenhouse-grown MA line [5] as the number of mutations divided by the total bp length of the genome (or a given annotation) and by 31 generations (the two MA lines with only three generations were excluded from this analysis). Mean and confidence intervals across lines are reported (Table S3). The genome length was determined as all base pairs with coverage higher or equal to 3, and a SHORE mapping quality score of at least 32 in one sample (Table S2).

7.2 Natural populations of HPG1

7.2.1 Net distances

For the "net genetic distances" method, we computed confidence intervals of the *b* regression slope coefficient ($D' = a+bT'$) using a bootstrap with replacement of 1,000 samples to avoid over-confident confidence intervals due to lack of independence of points [13]. We used either all SNPs or SNPs at specific annotations to calculate different substitution rates and scaled the slope into a per-base rate using all positions (of the given annotation) that passed alternative or reference call quality thresholds rather than using a single value of genome length (Table S3). For all annotations we calculated substitution rates with three quality thresholds and either full information per SNP or allowing a maximum of 50% missing accessions per SNP (see Section 3 and Fig. S1C).

For some annotations substitution rates were not reliable. For instance, in 3' and 5' UTR regions, we did not have enough mutations (on average ~1 SNP difference between any pair), and thus do not report these regions' rates. We could also have less power to discover SNPs in annotations with extensive structural variation such as active transposable elements [14]. Transposons, which comprise ~8% of the genome and ~19% of all the SNPs in greenhouse MA lines, had fewer SNPs called than expected in HPG1. This would explain the atypically low transposon substitution rate (Table S3). Therefore, transposon substitution rates in HPG1 cannot be trusted.

7.2.2 Bayesian tip-calibration

For the second approach to estimate a substitution rate, the Bayesian phylogenetics tip-calibration approach, we performed systematic runs and chain convergence assessments of different demographic and molecular clock models. We found the Skygrid demographic model [15] and the lognormal relaxed molecular clock [16] the most appropriate models. Under a relaxed molecular clock, the substitution rate is allowed to vary across branches with a lognormal distribution. The prior used for molecular clock was a Continuous-Time Markov Chain (CTMC) [15,17]. The analysis was carried out remotely at CIPRES PORTAL (v3.1 www.phylo.org) using uninformative priors. The run took about 1,344 CPU hours and performed 1,000 million steps in a Monte Carlo Markov Chain (MCMC), sampling every 100,000 steps. Burn-in was adjusted to 10% of the steps. To visualize the tree output we produced a Maximum Clade Credibility (MCC) tree with a minimum posterior probability threshold of 0.8 and a 10% burn-in using TreeAnnotator (part of BEAST package), and visualized the MCC tree using FigTree (tree.bio.ed.ac.uk/software/figtree/) (Fig. 3B). Additionally, we used DensiTree [18] to simultaneously draw the 10,000 BEAST trees with the highest posterior probability (Fig. 3A). Since all trees were drawn transparently, agreements in both topology and branch lengths appear as densely colored regions, while areas with little agreement appear lighter.

7.2.3 Methylation status of mutated sites

As in many other species, the spectrum of *de novo* mutations in the greenhouse-grown *A. thaliana* MA lines is biased towards G:C→A:T transitions [8], leading to an inflated transition-to-transversion ratio (Ts/Tv). This bias is less pronounced in recent mutations in a Eurasian collection of natural accessions (Fig. 5A of [19] and in HPG1 accessions (Fig. 3D). A recent multigenerational salt stress experiment in the greenhouse also showed a more balanced Ts/Tv [20]. These findings indicate that less benign conditions might promote a lower Ts/Tv, and one possible cause are methylation patterns, known to change under different environments [21].

We interrogated the potential evolutionary role of cytosine methylation in the mutability of cytosine bases in the HPG1 accessions. For reference DNA methylation data, we used previously generated bisulfite-sequencing data of HPG1 strains [7] and of Col-0 MA lines [5], respectively. For both datasets, methylation status was calculated as the fraction of reads with methylated cytosines by the total number of reads at a certain cytosine position in the genome. Our rationale was that if methylation affected mutability, the degree of methylation at positions were we find a new mutation should be higher. To be sure that a given site in HPG1 was a new mutation, we only considered positions for which we could determine that state by alignment to the *A. lyrata* genome [22]. The "tested sites" were positions in HPG1 that had a mutation both from *A. lyrata* and *A. thaliana* Col-0. These positions can be of two kinds, "fixed" if all HPG1 individuals carry the alternative, or "segregating" if both reference and alternative alleles exist in HPG1. As control, "control set", we used cytosine positions that did not vary across HPG1, *A. lyrata* and *A. thaliana*. To produce the methylation distribution of the control set we randomly chose 1,000 invariant cytosine positions. For the test sets, we averaged the methylation degree and compared it with the control distribution.

Ancestral cytosines with higher methylation in both *A. thaliana* Col-0 reference and HPG1 pseudo-reference methylome datasets were more likely to mutate to thymines in HPG1 (Fig. S2 A-D). Additionally, the methylation degree at substitutions inside genes was higher in the HPG1 methylome (Fig. S2 B,D). While some C→T changes could be explained by higher spontaneous deaminations known to happen more often at methylated cytosines, also C→A/G substitutions were more likely to have been methylated. If this process is common enough, the Ts/Tv ratio should decrease. We are far from understanding differences in Ts/Tv in natural and controlled conditions, but definitely methylation status seems to have a strong statistical connection with mutability.

8. Phenotypic association analyses and dating of newly arisen mutations

8.1. Phenotyping

8.1.1 Root

Fifteen root phenotypes were scored for ≥ 10 replicates per genotype over a time-series experiment at the Gregor Mendel Institute in Vienna, using image analysis as described in detail elsewhere [23]. We used the means per genotypes and per time series for association analyses.

8.1.2 Seed size

We spread the seeds of given genotypes on separate plastic square 12 x 12 cm Petri dishes. For faster image acquisition we used a cluster of eight Epson V600 scanners. The scanner cluster was operated by the BRAT Multiscan image acquisition tool (www.gmi.oeaw.ac.at/research-groups/wolfgang-busch/resources/brat/). The resulting 1600 dpi images were analyzed in Fiji software. Scans were converted to 8-bit binary images, thresholded (parameters: setAutoThreshold("Default dark"); setThreshold(20, 255)) and particles analyzed (inclusion parameters: size=0.04-0.25 circularity=0.70-1.00). The 2D seed size was measured in square millimeters (parameters: distance=1600 known=25.4 pixel=1 unit=mm) for 2 plants per genotype, > 500 seeds per plant.

8.1.3 Flowering in the growth chamber

We estimated the flowering time in growth chambers under four vernalization treatments (0, 14, 28 and 63 days of vernalization). We grew 6 replicates per accession divided between two complete randomized blocks for each treatment. Seeds were sown on a 1:1 mixture of Premier Pro-Mix and MetroMix and cold stratified for 6 days (6°C, no light). We then let plants germinate and grow at 18°C, 14 hours of light, 65% humidity. After 3 weeks, we transferred the plants to vernalization conditions (6°C, 8 hours of light, 65% humidity). After vernalization, plants were transferred back to long day conditions. Trays were rotated around the growth chambers every other day throughout the experiment, under both vernalization and ambient conditions. Germination, bolting and flowering dates were recorded every other day until all plants had flowered. Days till flowering or bolting times were calculated from the germination date until the first flower opened and until the first flower bud was developed, respectively. The average flowering time and bolting time per genotype were used for association analyses.

8.1.4 Fecundity in the field

To investigate variation in fecundity in natural conditions, we grew three replicates of each accession in a field experiment following a completely randomized block design. Seeds were sown from 09/20/2012 to 09/22/2012 in 66-well trays (well diameter = 4 cm) on soil from the field site where plants were to be transplanted. The trays were cold stratified for seven days before being placed in a cold frame at the University of Chicago (outdoors, no additional light or heat, but watered as needed and protected from precipitation). Seedlings were transplanted directly into tilled ground at the Warren Wood field station (41.84° N., 86.63° W.), Michigan, USA on 10/13/2012 and 10/14/2012. Seedlings were watered-in and left to overwinter without further intervention. Upon maturation of all fruits, stems were harvested and stored between sheets of newsprint paper. To estimate the fecundity, stems were photographed on a black background and the size of each plant was estimated as the number of pixels occupied by the plant on the image. This measure correlates well with the total length of siliques produced, a classical estimator of fecundity in *A. thaliana* (Spearman's rho=0.84, *p*-value<0.001, data not shown).

8.2 Quantitative genetic analyses

For 63 modern accessions, we measured time to bolting and flowering, seeds per plant, seed size, and 15 root phenotypes in common chamber or common garden settings. For all 100 accessions, climatic information from the bioclim database (www.worldclim.org/bioclim) was extracted using their geographic coordinates. For historic samples, some locations were only known by county name. In this case we assigned the geographic coordinate location of the centroid of the county.

8.2.1 Heritability

We performed association analyses using the R package GenABEL [24], with measured phenotypes (p = 25) and climatic variables (c = 18) as response variables and SNPs as explanatory variables. A Minimum Allele Frequency (MAF) cutoff of 5% was used. The number of assessed SNPs was 391 in a dataset of only modern samples but with imputed genotypes for missing data using Beagle v4.0 [25], and 456 SNPs with a dataset of modern and historic samples, without imputation. For all associations, at least 63 individuals were genotyped for a specific SNP. We first investigated broad sense heritability (H^2) of each trait using ANOVA partition of variance between and within lines using replicates (Table S4). Significance was obtained by common F test in ANOVA. Secondly we used the *polygenic_hglm* function to fit a genome wide kinship matrix to calculate a narrow sense heritability estimate (h^2). This fits a model of the type $y = Zu + \varepsilon$ (see Main text Methods). Significance was calculated employing a likelihood ratio test comparing with a null model. In principle, h^2 is a component of H^2, then its values should theoretically be $h^2 < H^2$. That is not our case. Our result

cannot be interpreted in this framework, since the calculation of both was not done with the same samples: for the h^2 calculation we employed genotype means whereas for the H^2 we used multiple replicated measurements per genotype. The averaging of replicates per genotype in h^2 reduced environmental and developmental noise and thus we would expect $h^2>H^2$. We did this so the climatic estimates of h^2, for which we only have one value per genotype, would be comparable with the phenotypic h^2 ones (Table S4).

8.2.2 Linear Models

For association analyses we first employed a linear mixed model that fitted the kinship matrix using the *mmscore* function. This model is of the type: $y = Xb + Zu + \varepsilon$ (see Main text Methods) [26]. Only three significant SNP hits were discovered using a 5% significance threshold after False Discovery Rate correction (FDR). This was expected since we have few variants and these would have originated in an approximated phylogeny structure. We concluded that fitting the kinship matrix in our model was not appropriate since there would be no residual variation for association with specific SNPs. With this rationale we employed a fixed effects linear model using the *qtscore* function [27]. This model is of the type: $y = Xb + \varepsilon$; where no random effect of genome background is fit. To reduce the risk of having false-positives, we took a conservative permutation strategy by carrying out association with over 1,000 randomized datasets (permuting phenotypes across individuals) and used the resulting empirical p-value distribution to correct p-values estimated with the original dataset. SNPs with p-values below 5% in the empirical p-value distribution should be considered significant (but see next section). In climatic models, we included longitude and latitude as covariates to correct for any spurious association between SNPs and climate gradients created by the migratory pattern of isolation by distance.

8.2.3 Evaluation of significance

Significant SNPs were interspersed throughout the genome (Fig. 4) and their p-values and phenotypic effects did not correlate with the minimum age of the SNPs nor with their allele frequency, something that could have indicated that the significance was merely driven by the higher statistical power of intermediate frequency variants. Using QQ plots to assess inflation or deflation of p-values, we observed generally that permutation corrected p-values were deflated — another evidence of our conservative strategy. Straight horizontal series of points in QQ plots indicate that multiple SNPs have identical p-values, a pattern that we attributed to long range LD, i.e. lack of independence (see Text S2 for trait distributions and QQ plots from each association analysis).

To further ensure that we avoided false positive results, we also prioritized SNPs whose empirical p-value was not below 5% only but also below 5% / (number of SNPs + number of traits) = 0.01%. This "double" Bonferroni correction was very conservative (Table 1, Table S5).

8.2.4 Context of de novo mutations associated with phenotypes

For each SNP in our dataset, we determined the ancestral and derived states, by identifying which allele was found in the oldest herbarium samples. We compared the time of emergence and the centroid of geographic distribution of the alternative alleles of SNP hits to random draws of SNPs with the same MAF filtering (5%) (Fig. S1).

8.2.5 Functional information

On top of phenotypic and climatic associations of SNP hits, we also provide a likely functional effect employing a commonly used amino acid matrix of biochemical effects [28]. Functional information of gene name and ontology categorization of SNP hits was obtained from www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp and www.arabidopsis.org/tools/bulk/go/ (Table 1 and Table S5).

8.2.6 Proof of concept examples

We argue that the power of our association approach relies on the fact that HPG1 lines resemble Near Isogenic Lines (NILs) produced by experimental crosses [29] (Fig. S2A). Similar to genome-wide association studies (GWA), power depends on many factors, namely the noise of phenotype under study, architecture of phenotypic trait, quality of genotyping, population structure, sample diversity, sample size, allele frequency, and recombination. On one hand, association analyses in NILs suffer from large linkage blocks, but confident results can be achieved due to accurate measurement of phenotypes, limited genetic differences between any two lines, and high quality genotypes. In common GWA studies such as in humans, there are multiple confounding effects. Among the confounders are (1) that any two samples differ in hundreds of thousands of SNPs, and (2) that historical and geographic stratification produce non-random correlations among those SNP differences. This considerably complicates the identification of phenotypic effects at specific genes, and power relies greatly on large sample sizes to achieve the sufficient number of recombination between markers.

To provide support for the non-synonymous SNP on chromosome 5, at position 6,508,329 in AT5G19330, we looked for pairs of lines that carry the ancestral and the derived allele, but that differ in few (or no other) SNPs in the genome. When considering all genic substitutions with a minimum allele frequency of 5% (Fig. S2A), we identified 20 pairs of lines differing only in the AT5G19330 SNP

and another linked SNP (located on a different chromosome, association p-value > 0.4). The phenotypic differences in mean gravitropic score of these almost-identical pairs were significantly higher than phenotypic differences among all pairs of HPG1 lines, and genetically identical pairs attending to substitutions inside genes (Fig. S2A). Furthermore, this SNP was not in complete linkage with any other SNP hit (r^2 < 0.5) (Fig. S2D). The same approach was used to examine the SNPs in AT1G54440 (Fig. S2E) and AT2G16580 (Fig. S2F), which represent an intermediate and a high LD example.

References

1. Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. Royal Society Open Science. The Royal Society; 2016;3: 160239. doi:10.1098/rsos.160239

2. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010: db.prot5448. doi:10.1101/pdb.prot5448

3. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010;38: e87. doi:10.1093/nar/gkp1163

4. Kircher M. Analysis of High-Throughput Ancient DNA Sequencing Data. In: Shapiro B, Hofreiter M, editors. Ancient DNA. Humana Press; 2011. pp. 197–228. doi:10.1007/978-1-61779-516-9_23

5. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature. nature.com; 2011;480: 245–249. doi:10.1038/nature10555

6. Shaw RG, Byers DL, Darmo E. Spontaneous mutational effects on reproductive traits of arabidopsis thaliana. Genetics. 2000;155: 369–378. Available: http://www.ncbi.nlm.nih.gov/pubmed/10790410

7. Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, et al. Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS Genet. 2015;11: e1004920–e1004920. doi:10.1371/journal.pgen.1004920

8. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010;327: 92–94. doi:10.1126/science.1180677

9. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The scale of population structure in Arabidopsis thaliana. PLoS Genet. 2010;6: e1000843. doi:10.1371/journal.pgen.1000843

10. Wolfe KH, Sharp PM, Li W-H. Rates of synonymous substitution in plant nuclear genes. J Mol Evol. Springer-Verlag; 1989;29: 208–211. doi:10.1007/BF02100204

11. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. Elsevier; 2016;166: 481–491. doi:10.1016/j.cell.2016.05.063

12. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25: 1451–1452. doi:10.1093/bioinformatics/btp187

13. Drummond A, Pybus OG, Rambaut A. Inference of viral evolutionary rates from molecular sequences. Adv

Parasitol. 2003;54: 331–358. Available: http://www.ncbi.nlm.nih.gov/pubmed/14711090

14. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13: 36–46. doi:10.1038/nrg3117

15. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard M a. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2012;30: 713–724. doi:10.1093/molbev/mss265

16. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4: e88–e88. doi:10.1371/journal.pbio.0040088

17. Ferreira M a. R, Suchard M a. Bayesian analysis of elapsed times in continuous-time Markov chains. Can J Stat. 2008;36: 355–368. doi:10.1002/cjs.5550360302

18. Bouckaert RR. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics. 2010;26: 1372–1373. doi:10.1093/bioinformatics/btq110

19. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43: 956–963. doi:10.1038/ng.911

20. Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. Genome Res. 2014;24: 1821–1829. doi:10.1101/gr.177659.114

21. Wibowo A, Becker C, Marconi G, Durr J, Price J, Hagmann J, et al. Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. eLife. eLife Sciences Publications Limited; 2016;5: e13546. doi:10.7554/eLife.13546

22. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet. Nature Publishing Group; 2011;43: 476–481. doi:10.1038/ng.807

23. Slovak R, Göschl C, Su X, Shimotani K, Shiina T, Busch W. A scalable open-source pipeline for large-scale root phenotyping of Arabidopsis. Plant Cell. 2014;26: 2390–2403. doi:10.1105/tpc.114.124032

24. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007;23: 1294–1296. doi:10.1093/bioinformatics/btm108

25. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98: 116–126. doi:10.1016/j.ajhg.2015.11.020

26. Wilson AJ, Réale D, Clements MN, Morrissey MM, Postma E, Walling CA, et al. An ecologist's guide to the animal model. J Anim Ecol. 2010;79: 13–26. doi:10.1111/j.1365-2656.2009.01639.x

27. Aulchenko YS, de Koning D-J, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007;177: 577–585. doi:10.1534/genetics.107.075614

28. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;185: 862–864. doi:10.1126/science.185.4154.862

29. Weigel D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiol. 2012;158: 2–22. doi:10.1104/pp.111.189845

Text S2

For each trait employed in association analyses, we report the histogram
distribution and the QQ plot of p-values to ensure that no trait departs exaggeratedly from the normal distribution, and that no inflation of p-values is observed (when lambda ≤ 1, there is no inflation of false positives).

https://doi.org/10.1371/journal.pgen.1007155.s002

SUPPLEMENTARY TABLES

Table S1. HPG1 sample information.
https://doi.org/10.1371/journal.pgen.1007155.s003

Table S2. Sample information for Col-0 mutation accumulation lines.
https://doi.org/10.1371/journal.pgen.1007155.s004

Table S3. Mutation rate estimates for different annotations in HPG1 and mutation accumulation lines.
https://doi.org/10.1371/journal.pgen.1007155.s005

Table S4. Description of phenotypic and climatic variables for association mapping analyses.
https://doi.org/10.1371/journal.pgen.1007155.s006

Table S5. SNP hits from association analyses and several descriptors.
https://doi.org/10.1371/journal.pgen.1007155.s007

SUPPLEMENTARY FIGURES

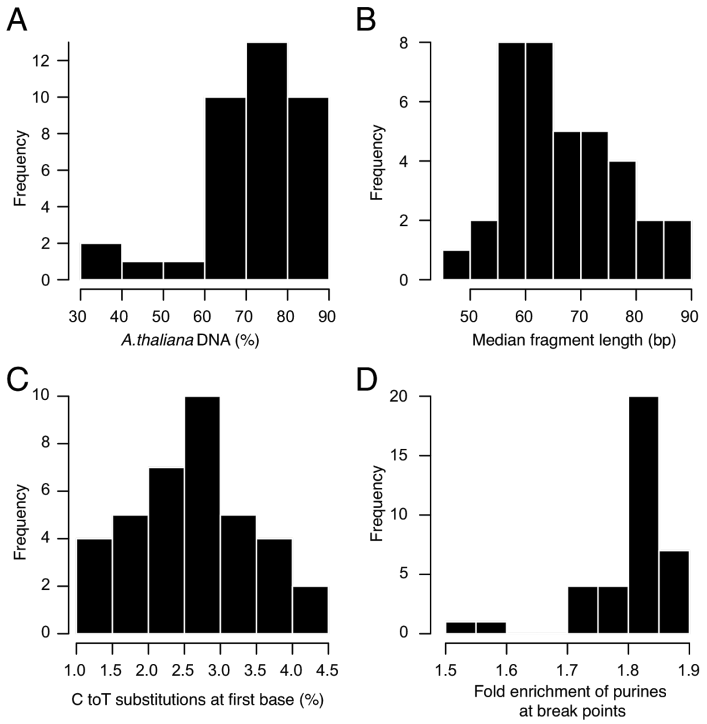


Fig S1. Ancient-DNA characteristics of unrepaired herbarium libraries.

(A) Fraction of *A. thaliana* DNA in sample. **(B)** Median length of merged reads. **(C)** Fraction of cytosine to thymine (C-to-T) substitutions at first base (5' end). **(D)** Relative enrichment of purines (adenine and guanine) at 5' end breaking points. Position -1 is compared with position -5 (negative numbers indicate genomic context before upstream reads' 5' end).

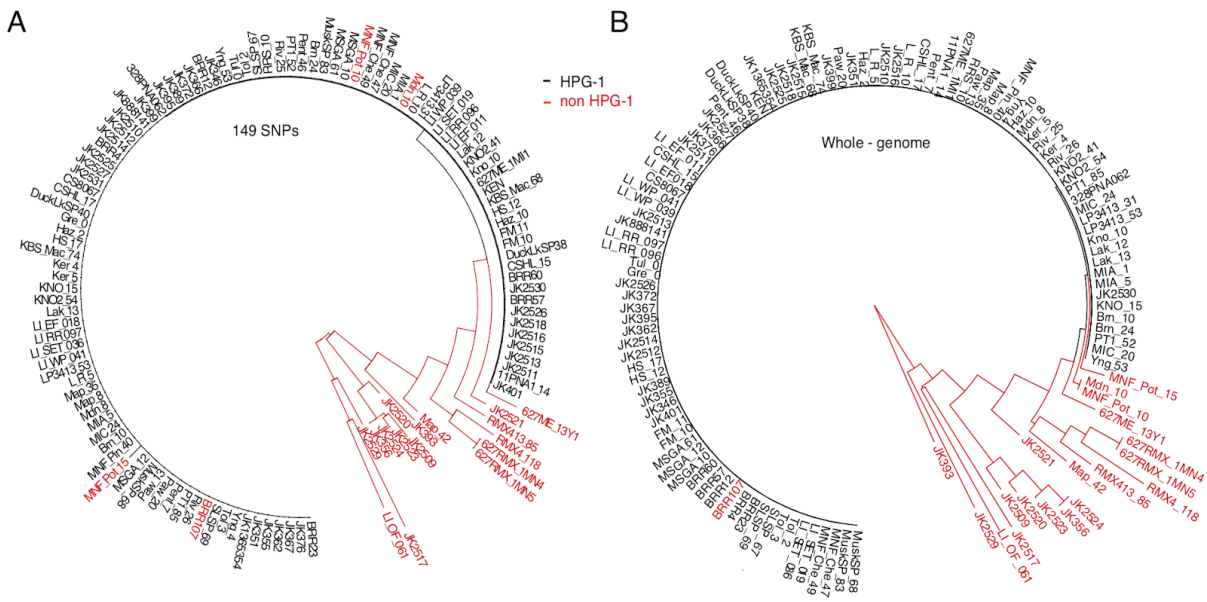


Fig S2. Separation between HPG1 and other North American lineages.

(A) Neighbor-joining tree built using Illumina-based SNP calls at the 149 genotyping markers originally used to identify HPG1 candidates. HPG1 accessions are shown in black, whereas other North American lineages are depicted in red (see explanation below for four HPG1-like accessions). **(B)** Neighbor-joining tree based on genome-wide SNPs. Accessions colored as in (A). Note that three accessions originally classified as HPG1 based on 149 SNPs (A) are placed outside this clade. A further accession (BRR7) within the HPG1 main branch was a recombinant removed from the analysis.

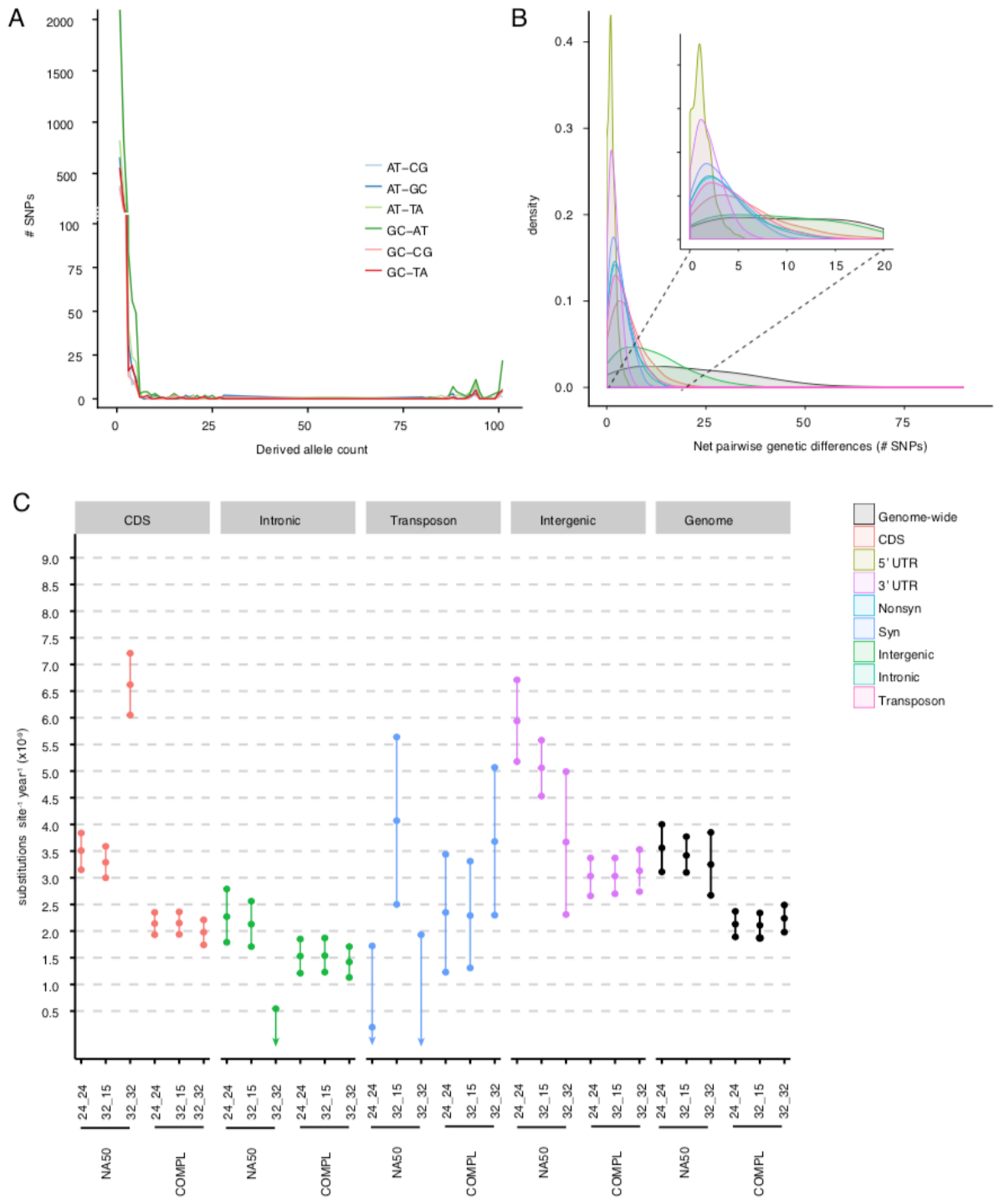


Fig S3. Substitution spectrum and rates.

(A) Site frequency spectrum for all transitions and transversions. **(B)** Distributions of "net" pairwise genetic distances between historic and modern samples used to calculate mutation rates per genomic annotation (from quality 32_15 and complete information per site). UTRs were excluded because of the small number of SNPs. **(C)** Mutation rates calculated for different genomic

annotations and quality thresholds (32_32, 32_15, 24_24) and missing values (NA50: maximum 50% missing data per SNP; COMPL: missing data 0%). Mean and 95% confidence intervals are shown.

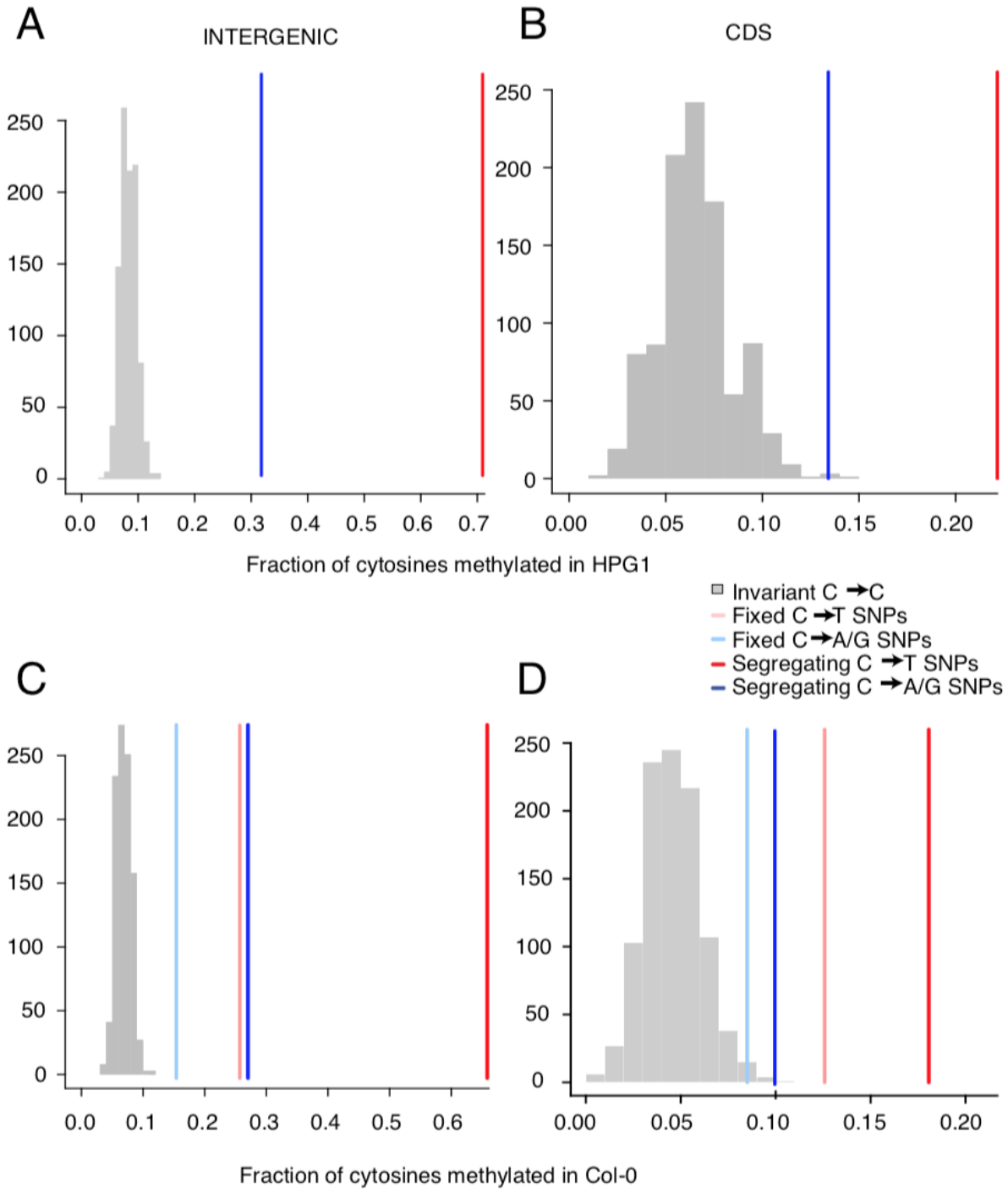


Fig S4. Relationship between methylation and substitutions.

(A, B) Fraction of methylation of cytosines in HPG1 pseudo-reference[7] at intergenic (A) or coding regions (B). **(C, D)** Fraction of methylation of cytosines in Col-0 reference genome[5] at intergenic (C) or coding regions (D). In each of the four comparisons, a grey histogram represents distribution of methylation of 1,000 random sets of invariant cytosines. Lines represent average methylation degree at those sites in HPG1 that changed from cytosine to thymine (red). We differentiate those

substitutions that are shared - fixed - across all individuals (light red) or whose allele are present at an intermediate - segregating - frequency (dark red). Likewise, average methylation is shown for sites that changed from cytosine to adenine (blue) that that are fixed (light blue) or segregating (dark blue). The fact that the average methylation is higher in new substitutions than in invariant positions supports a connection between methylation and mutability of sites.

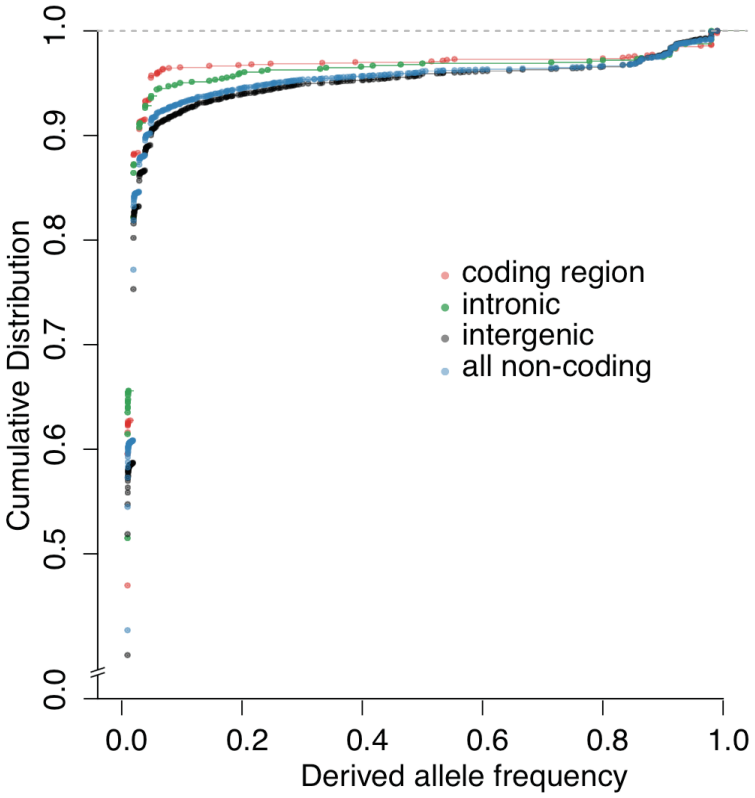


Fig S5. Comparison of Site Frequency Spectra across genomic annotations.

Cumulative empirical distribution, at different genomic annotations, of the unfolded Site Frequency Spectrum of SNPs oriented based on the order of appearance of alleles in the herbarium genomes. Note the steep slope at low frequency indicating large numbers of such variants.

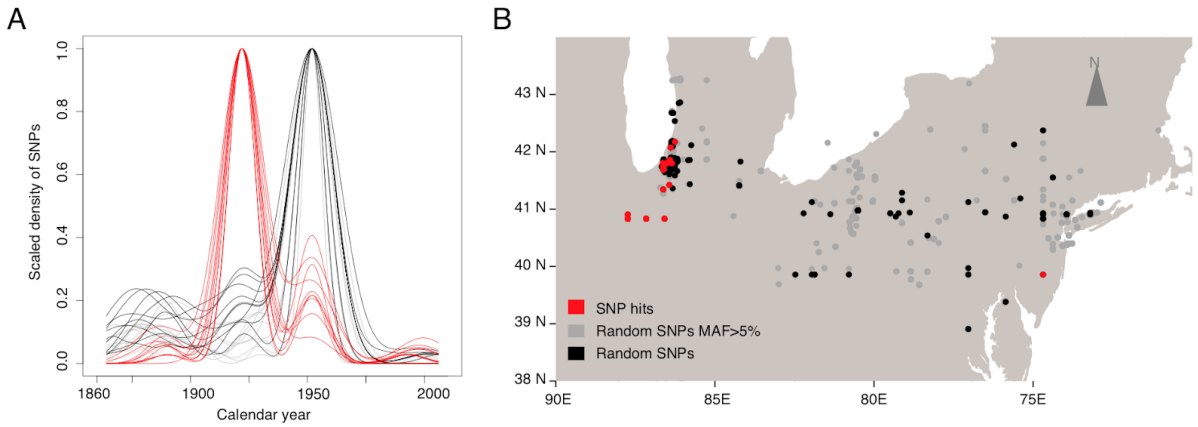


Fig S6. Spatial and temporal emergence of root-associated mutations.

(A) Age distribution of derived SNPs with a significant trait association (the herbarium sample in which they were first recorded) (red), compared with genome-wide SNPs with at least 5% minor allele frequency (grey), or without frequency cutoff (black). **(B)** Spatial centroid of all samples carrying a derived allele. Since it is an average location, centroids can be in a body of water. Ten random draws of 50 SNPs for each category were used to produce the density lines in (A) and points in (B).

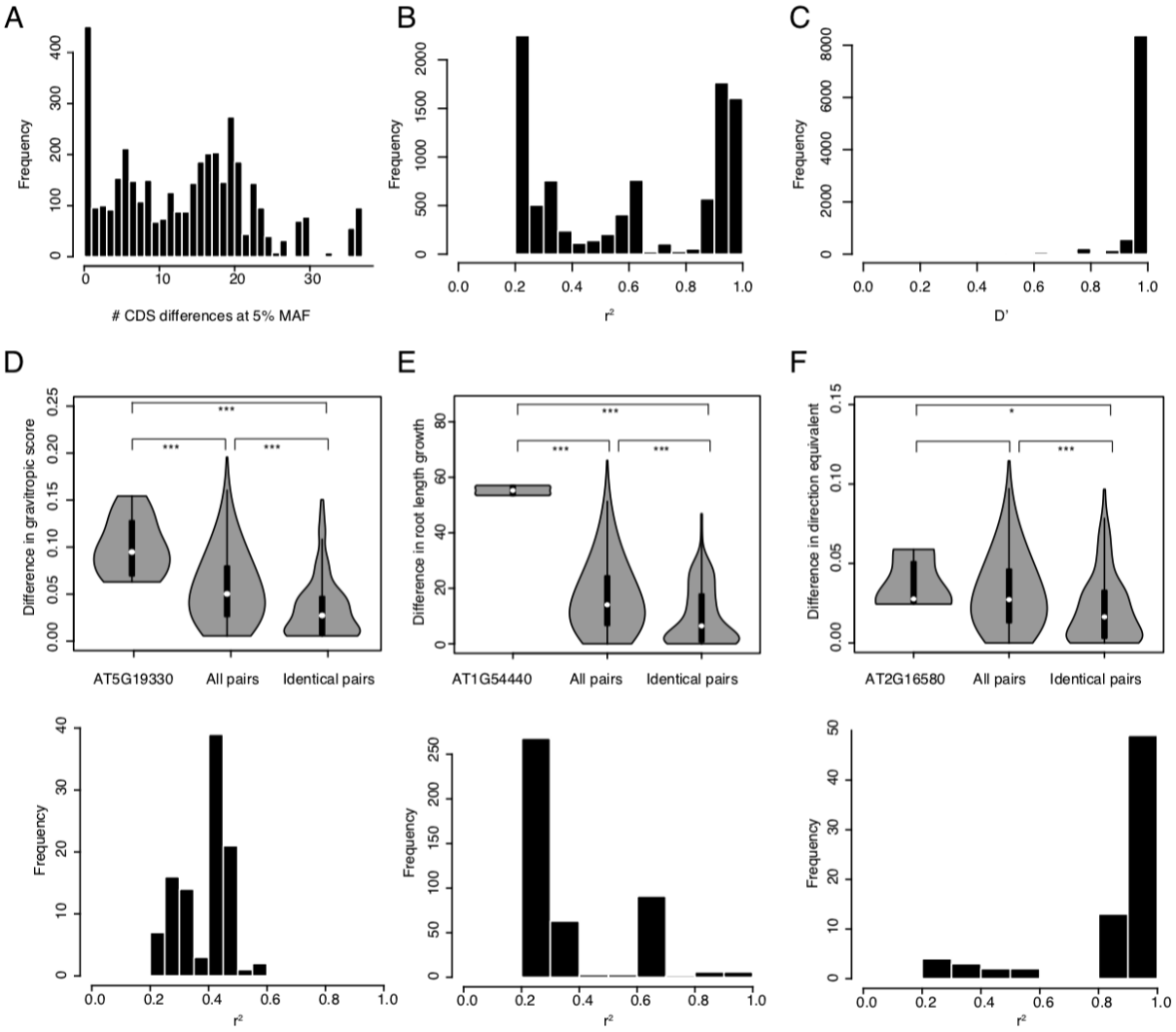


Fig S7. Linkage disequilibrium of significant SNPs.

(A-F) Linkage disequilibrium between SNPs with significant trait associations. Histogram of genetic distances **(A)** between samples when evaluating only coding regions at 5% minimum allele frequency. Linkage disequilibrium between SNP hits measured as r^2 **(B)** and *D*' **(C)**. Three significant SNPs were further studied to exemplify the power of association analyses with HPG1. For each, phenotypic differences between accessions that differ in the focal SNP and that are otherwise virtually genetically identical are compared both with all pairs of accessions and with pairs of accessions completely identical for coding regions. Below each violin plot is the histogram of linkage disequilibrium of the focal SNP with all other SNP hits. The three focal SNPs evaluated are located in AT5G19330 **(D)**, AT1G54440 **(E)** and AT2G16580 **(F)**.