Investigating the genetic architecture and adaptive relevance of complex traits in Cape Verde *Arabidopsis*

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät der

Universität zu Köln

vorgelegt von

Célia Carolina Martins Neto

aus Porto, Portugal

Köln, 2021

Max Planck Institute for
Plant Breeding Research

MAX-PLANCK-GESELLSCHAFT

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln in der Arbeitsgruppe von Dr. Angela M. Hancock angefertigt.

This dissertation has been accepted by the Faculty of Mathematics and Natural Sciences of the University of Cologne.

Erster Gutachterin und Prüferin	Dr. Angela M. Hancock
Zweite Gutachterin und Prüferin	Prof. Dr. Ute Höcker
Vorsitzende der Prüfungskommission	Prof. Dr. Joachim Krug
Beisitzer	Dr. Emmanuel Tergemina

Tag der mündlichen Prüfung: 21.01.2022

Table of Contents

Thesis outline

This thesis is composed by two main chapters and a general introduction and discussion. The introduction explains key concepts necessary for the unroll of the chapters and sets the main goals for the thesis, describing the major gaps in the research field. The discussion brings together the two chapters and allows for an overview of the main findings in a broader context.

The first chapter 'Parallel reduction in flowering time from new mutations enabled evolutionary rescue in colonizing Arabidopsis lineages' presents the system used in this project and introduces important elements needed for the deeper examination done in Chapter II. This chapter is a submitted and under review paper, in which I am a first co-author. My contributions to this paper include conceptualization, resources and data collection, investigation, validation and data curation, data analysis and writing. Specifically, I was responsible for collection genetic material and climatic information in the field, reviewing literature in Cvi-0 and its QTLs, designing all experiments and collecting phenotypic information, data analysis, quantitative genetic analysis, and trait mapping.

The second chapter 'Population history impacts the genetic architecture of flowering time after colonization of a novel environment' takes results from Chapter I and examines them deeper with a more quantitative approach.

Summary

Understanding how organisms adapt to new environments is a key goal of evolutionary biology. Populations subject to abrupt environmental change must adapt quickly to avoid extinction. Small populations are especially vulnerable to habitat changes, confronting high extinction risk due to limited genetic variation and low efficiency of selection. Theory predicts that the age of a population and its long-term effective size should influence adaptation and trait architecture.

Here, we investigate the mechanisms of adaptation after a sudden shift to a more arid climate using natural populations of *Arabidopsis thaliana* in Cape Verde (CVI). CVI *Arabidopsis* is found on two islands (Santo Antão and Fogo) and represents diverged, monophyletic lineages based on the near absence of shared polymorphisms with each other or the continent.

Time to flowering was reduced in parallel on the islands, causing a consequent increase in fitness, and allowing adaptation to the arid CVI. This change was mediated by convergent *de novo* loss of function of two core flowering time genes: *FRI* in Santo Antão and *FLC* in Fogo. Our results reveal a case where expansion of the new populations coincided with the emergence and proliferation of these novel variants, consistent with models of rapid adaptation and evolutionary rescue.

We further contrast the genetic architecture of flowering time in the recently formed small N_e *Arabidopsis* lineages from Cape Verde with their much older, larger N_e progenitor – the Moroccan population. We find that polygenicity is severely reduced in the colonizing populations and effect sizes of candidate loci are exponentially distributed, consistent with fitness measures showing evidence for directional selection in the islands. In addition to the major effect variants *FRI* K232X and *FLC* R3X, we identify candidate variants from core flowering time pathways as well as those that indirectly affect flowering time, including nutrient processing and light sensing. Surprisingly we find no effect of the well-known Cvi-0-EDI (*CRY2* V367M) variant in the natural population. Our results provide a particularly clear empirical example of the effect of demographic history has on trait architecture.

Zusammenfassung

Zu verstehen, wie sich Organismen an neue Umgebungen anpassen, ist ein wichtiges Ziel der Evolutionsbiologie. Populationen, die abrupten Umweltveränderungen unterliegen, müssen sich schnell anpassen, um das Aussterben zu vermeiden. Kleine Populationen sind besonders anfällig für Habitat-Veränderungen und sind aufgrund begrenzter genetischer Variation und geringer Selektionseffizienz einem hohen Aussterberisiko ausgesetzt. In der Theorie sollten das Alter einer Population und ihre langfristig effektive Größe die Anpassung und die Merkmalsarchitektur beeinflussen.

Hier untersuchen wir die Mechanismen der Anpassung nach einem plötzlichen Wechsel zu einem trockeneren Klima mit natürlichen Populationen von *Arabidopsis thaliana* in Kap Verde (CVI). CVI *Arabidopsis* kommt auf zwei Inseln (Santo Antão und Fogo) vor und repräsentiert divergierende, monophyletische Abstammungslinien, die auf einer nahezu vollständigen Abwesenheitgemeinsamer Polymorphismen untereinander oder auf dem Kontinent beruhen.

Parallel dazu wurde die Blütezeit auf den Inseln verkürzt, was eine konsequente Steigerung der Fitness bewirkte und eine Anpassung an die trockene CVI ermöglichte. Diese Veränderung wurde durch einen konvergenten *de novo* Funktionsverlust von zwei Kernblütezeitgenen herbeigeführt: *FRI* in Santo Antão und *FLC* in Fogo. Unsere Ergebnisse zeigen einen Fall, in dem die Expansion der neuen Populationen mit dem Aufkommen und der Verbreitung dieser neuen Varianten zusammenfiel, was mit Modellen der schnellen Anpassung und evolutionären Rettung übereinstimmt.

Wir kontrastieren die genetische Architektur der Blütezeit in den kürzlich gebildeten kleinen N_e *Arabidopsis*-Linien von den Kapverden mit ihrem viel älteren, größeren Ne-Vorläufer - der marokkanischen Population. Wir stellen fest, dass die Polygenität in den kolonisierenden Populationen stark reduziert ist und die Effektstärken der Kandidaten-Loci exponentiell verteilt sind, was mit Fitnessmessungen übereinstimmt, die Hinweise auf eine Richtungsselektion auf den Inseln zeigen. Zusätzlich zu den Haupteffektvarianten *FRI* K232X und *FLC* R3X identifizieren wir Kandidatenvarianten aus den Hauptblütezeitwegen sowie solche, die die Blütezeit indirekt beeinflussen, einschließlich Nährstoffverarbeitung und Licht Sensorik. Überraschenderweise finden wir keine Wirkung der bekannten Cvi-0-EDI (*CRY2* V367M) Variante in der natürlichen Population. Unsere Ergebnisse liefern ein besonders anschauliches empirisches Beispiel für den Einfluss der demografischen Geschichte auf die Merkmalsarchitektur.

INTRODUCTION

The genetic architecture of adaptation

A key goal of evolutionary biology is to understand how organisms adapt to new environments. To do so, one needs to examine the genetic architecture of adaptive traits, which can be described as monogenic, oligogenic, or polygenic, if, respectively, one, few, or many genetic variants are contributing to the phenotypic variance. Genetic architecture inference comprehends mapping the genetic basis of a trait, estimating the number of alleles affecting it, and their respective effect sizes and allele frequencies. However, and despite all the efforts, how adaptive traits evolve in response to environmental changes remains a poorly understood process [1–3].

Over the years, innumerous contributions have helped shedding some light into the genetic architecture of adaptation, by identifying candidate loci, genes and, in few cases, even the functional variant underlying the phenotypic divergence. Some of the most well-known examples include the extensively studied case of adaptation involving color patterning in the peppered moth *Biston betularia* [4–6] and in mice, such as *Chaetodipus intermedius* and *Peromyscus polionotus* [7,8]; in body shape in the threespine stickleback fish *Gasterosteus aculeatus* [9–12]; in flowering time in *Zea mays* [13]; or in development in *Drosophila* [14,15].

The "industrial melanism" in *Biston betularia* is a textbook example of evolutionary change in response to shifts in the environment. With the industrial revolution, the frequency of darker pigmented individuals increased in highly industrial areas as their camouflage matched the polluted background better, which made darker morphs fitter than the rest of the population. For over 50 years, the genetic basis of this adaptive trait was unknown, beyond the fact that it was caused by a single dominant allele. Now, we know that the single large effect locus underlies a transposon causing the difference in color. However, the exact mechanism behind it is still to be discovered [4–6].

A similar example is color patterning in mice, which is strongly correlated with the color of the substrate on which the different populations live. Across species and populations, living on pale coastal sand dunes, dark lava or mainland environments, two genes have recurrently shown association with dorsal pelage – *Mc1r* and *Agouti* –, which together explain less than 40% of variation in pigmentation traits. Despite functional variants have been identified in these two genes, the information gathered around this system is based on crosses between two populations or candidate gene approaches. These methods have the drawback of reducing the amount of variation examined and decreasing the pool of possible genetic contributors [7,8].

Another well-studied system is body shape variation in threespine stickleback fish in response to colonization of new environments: marine sticklebacks who colonized and adapted to freshwater habitats have developed repeated changes in body shape. Similar to the example above, based on crosses between two populations, two large effect loci and a few more small effect loci have been identified as underlying the divergence between marine and freshwater stickleback [9–12].

However, and despite Herculean efforts over the years on these examples and many others [16–24], the genetic basis of adaptation is known in only a handful of cases because difficulties have arisen from (i) the limited genetic information available for some species; (ii) the problematic identification of ecologically important traits; (iii) their effect on fitness; and (iv) the complex genetic architecture of fitness-related traits.

1. The limited genetic information available on some species

Even though many studies have tried to fill the gap in the understanding of adaptation, the lack of genetic information available for some species prevents a deeper knowledge of the genetic basis of adaptation, the causal relationship between phenotype and genotype, and the identification of functionally relevant genetic variation contributing to adaptation.

A model species is therefore a good place to start. *Arabidopsis thaliana*, long the workhorse of plant molecular geneticists, is an annual selfing plant with a wide geographical distribution, which exposes the species to a variety of ecological and climatic conditions. These likely exert very different selective pressures on the wild populations, making the diverse naturally inbred lines – accessions – a good model for studying local adaptation, as demonstrated previously [25–35].

A. thaliana makes an exceptional model organism to study natural variation due to the large collection of more than 1300 wild accessions collected worldwide and that are publicly available in seed stock centers. Natural variation panels have been developed over the years, comprising accessions from very different climatic and geographical locations that were then fully-sequenced and explored, covering a wide geographical and genetical distribution. These panels have allowed for a systematic characterization of genome-wide polymorphisms, large-scale surveys of genotypic and phenotypic diversity, investigation of demographic histories and translation of genetic variation into phenotypic variation [33,36–41].

Adding to the rich germplasm collection and the vast genomic information, this species still provides other mapping resources, essential for the discovery of genes involved in the traits of interest.

Examples include mutagenesis-derived populations – which allow for the phenotypic analysis of mutants and provide a direct measure of a gene's contribution to the different traits; and recombinant populations derived from crosses between two or more accessions (e.g., recombinant (RIL) and near-isogenic inbred lines (NIL), AMPRIL, MAGIC) [42–51].

Since it is a selfer, *A. thaliana* is also experimentally tractable and genetically identical individuals can be compared across growth conditions and field sites. Decades of work have provided a deep knowledge on physiological and developmental processes, that when coupled with the extensive genetic tools and the genomic resources developed facilitate functional and mechanistic characterization of adaptive alleles [42,52].

2. The problematic identification of ecologically important traits and their effects on fitness

How the different traits affect fitness and contribute to adaptation is another important question when understanding the genetic basis of adaptation. Nonetheless, for all living organisms, the timing of the initiation of reproductive development is a crucial component of the life cycle, and in nature it can be a primary determinant of an individual's lifetime fitness [13,16,26,31,53]. In plants, if flowering occurs too early, floral tissues may be damaged by late frosts, the environmental conditions may not yet be appropriate, or other flowering individuals may not yet be abundant enough to ensure fertilization. If flowering occurs too late, a plant may encounter conditions unfavorable for seed maturation or dispersal, fail to set seed before dying in season-ending frosts or droughts, or leave offspring in poor growth environments [16,53–56]. Therefore, flowering time is a highly tractable model trait for examining adaptation processes – examples including *Lythrum salicaria* [25], *Mimulus guttatus* [16], *Brassica rapa* [53] and *Arabidopsis thaliana* [26,27,30,55,57].

As a complex trait, flowering time results from interactions between the complex network of genes that controls development with specific environmental cues, such as day length, light level, temperature, time of snow-melt, nutrient level, and water availability [16,42,55,58,59]. In *A. thaliana*, genetic clines in flowering time have been studied considering elevation, latitude, or other climate parameters. The results hint towards adaptation to the different environments across its geographical distribution, due to the responsiveness of flowering to vernalization, photoperiod, and ambient temperature [57,60–63] . These clues can be further inspected thanks to the extensive genomic resources in this species.

3. The complex genetic architecture of quantitative fitness-related traits and the factors that shape it

When quantitative fitness-related traits are identified, their genetic architecture is typically thought to be polygenic and therefore discussed in a quantitative genetic context based on Fisher's infinitesimal model [64]. This model of adaptation proposes that effectively infinite loci with small effects contribute to the phenotypic variation and phenotypic evolution occurs continuously and gradually towards a fitness optimum [2,64,65].

However, adaptive traits could follow a different pattern. Orr has extended Fisher's geometrical model bringing together different concepts and ideas developed over the years [66–68]. Orr's view states that, apart from small effect mutations, adaptation relies on mutations of large effect when a population faces a new fitness optimum. The population will 'walk' – a concept introduced years before by Maynard Smith [69] – towards the new optimum, in a diminishing returns fashion, by accumulating large effect mutations at the beginning of the walk – a theory proposed by Gillespie [70,71]– and small effect mutations later [66,67]. In Orr's rendition of Fisher's model, the effect sizes of mutations in bouts of adaptation follow a nearly exponential distribution, with few alleles showing large effect while many more small effects [66–68].

Empirical data on adaptation of different species fall along the spectrum: some traits follow a polygenic architecture, e.g., cold hardiness in conifers [17], flowering time in maize [13], others are defined by only a few loci of large effect – an oligogenic architecture, e.g., pelvic girdle [9] and armor plating [11] in sticklebacks, flower architecture in *Mimulus* spp. [72,73], coloration in beach mice [7,74] and in peppered moth [4,5], and response to salinity in *A. thaliana* [75]; and others even by loci with a range of effect sizes – e.g., body shape in sticklebacks [76]. Mapping studies over the years have suggested that there is considerable heterogeneity in genetic architecture among species and among traits.

Cape Verde *Arabidopsis* as a model system to study adaptation

Studies of adaptation and the alleles that underlie it have long lacked reliable model systems that combine genetic tractability with ecological relevance. Here, we use natural populations of *Arabidopsis thaliana* in Cape Verde as a clear and simple system to study adaptation to a new environment with strong selective pressures after long-range colonization.

1. Challenges to mapping in global diversity panels of *Arabidopsis thaliana*

As seen above, *A. thaliana* represents a good model to study adaptation because it is an annual plant with a wide geographical distribution, growing in a variety of ecological conditions, with vast genomic resources that produced deep knowledge on physiological processes and massive amounts of information on fitness-related traits [31,32,37,42,52,54].

Despite all the resources available, some problems have been identified when mapping the genetic basis of fitness-related traits was attempted. Traditional linkage mapping can be an effective tool for identifying genes underlying natural variation [26,49,77,78]. However, the genes identified by this method are restricted to the ones segregating in the cross under consideration. Genome-wide association studies (GWAS) overcome this limitation by taking all segregating variation in a population. Yet, although mapping panels that make use of the broad geographic and genetic diversity are usually advantageous because they capture in theory all the species' variation, they also expose analyses to a variety of geographically-driven confounding forces [79–82].

Drawbacks from global diversity panels may include population structure, genetic heterogeneity, allelic heterogeneity, and epistasis. Population structure can generate false positives when the phenotypic trait of interest is also correlated with the underlying population structure. Including relatedness matrices in the model can control for these spurious associations but will reduce the power to detect causal variants whose geographic distributions overlap with patterns of population structure. Genetic heterogeneity is another problem in large diverse populations because the genetic basis of a trait may vary across regions and therefore alleles of interest will be restricted to specific regions. This is even a bigger problem for rare alleles, which will be 'lost' in the analysis. Another issue of very diverse populations is allelic heterogeneity, i.e., when a locus presents more than two alleles in the population, since GWAS typically uses biallelic markers only because more alleles will decrease statistical power. Finally, another difficulty of GWAS is dealing with epistasis. Different genotypic combinations will occur in different geographic regions and different alleles will interact in different ways. GWAS is less powerful when a causal variant's effect is weakened in some genetic backgrounds due to epistasis, since standard GWAS models are formulated to detect average additive effects across genetic backgrounds [79–82]. Therefore, in some cases, mapping in local panels may be more advantageous, provided that adequate phenotypic and genetic variation is present.

2. The Cape Verde representative: Cvi-0

Across the natural variation panels developed for *A. thaliana*, one accession has consistently stood out: Cvi-0. This accession was collected almost 40 years ago from Cape Verde (CVI) [83] and has since been part of all major panels due to its phenotypic and genotypic divergence [43–45,48,49]. It has also been repeatedly used as a parental line in several recombinant populations mapping a variety of traits, including insect and herbivory resistance [84], freezing and cold tolerance [85,86], circadian rhythm [87,88], flowering time [28,44,57,77,78,89], seed traits [90–96], and water use efficiency [97,98]. In some instances, the functional variant underlying Cvi-0's phenotypic divergence has been identified and functionally validated, as it is the case of, for example, *FRI* K232X [99,100], *CRY2* V367M [101,102], *CBF2* promotor deletion [85,86] or *ZTL* P35T [88].

However, broader conclusions on ecological impact and evolution, and follow-up work from these studies have been limited by the lack of population samples and information about the native environment.

3. The Cape Verde harsh environment as a strong selective pressure

Here, we will use natural populations of *Arabidopsis thaliana* in Cape Verde. CVI is a volcanic archipelago composed by 10 islands and located 570 km off the Western coast of Africa, between 14.80 and 17.20 degrees North of the Equator.

The environment is arid tropical because of their proximity to the Sahara, but on islands with high mountains and farther away from the coast, the humidity is much higher. Due to the relief of some islands, orographically induced precipitation allows for rich and luxuriant vegetation to grow where the humid air condenses. Most rainfall precipitation is due to condensation of the ocean mist, and therefore northeastern slopes often receive more rain than southwest ones. It rains irregularly between August and October, with frequent brief heavy downpours, creating a short and variable rainy season and a very long dry season.

4. Islands facilitate the interpretation of complex evolutionary processes

On top of the vast genomic resources that the study of *A. thaliana* offers and the rich body of information stemming from decades of work on Cvi-0, the use of islands eases the study of adaptation. When compared to mainland populations, islands are easier case studies, almost depleted of the complexity that comes from gene flow, population contractions and expansions, and secondary contacts.

Since Darwin and Wallace, islands have been regarded as laboratories of evolution and have proven to be very powerful systems to investigate the mechanisms of evolutionary change [103].

Due to their discrete nature, islands can be seen as independent replicates and act as Nature's test tubes – Nature places species in different replicates and lets them evolve in parallel. Besides, they are usually small in size, with simplified biotas, relatively young and geographically isolated, which make it easier for evolutionists to observe and interpret patterns of evolution [103]. When colonized from a distant population, island populations can allow the investigation of adaption to a novel environment, far from the previous optimum, after a sudden shift. The resulting isolated populations provide an opportunity to examine evolutionary processes in the absence of admixture and secondary contact, which is common in most continental populations.

Following a major environmental change, populations must adapt recurring to mutations, which, to contribute to adaptation, must enter the population at an appreciable rate, be beneficial and escape stochastic loss. Once a mutation is present in the population, its allele frequency trajectory will be determined by the interplay between genetic drift, demographic processes, and the natural selection acting on it [2,3,104,105]. Mutations can arise *de novo* or can be already present in the population and become beneficial after the environmental change. The former case – the standing genetic variation model – is expected to be more common in large populations, since these populations are more diverse, with more variation available [65,105,106]. Neutral or slightly deleterious alleles are maintained in the population until the environment changes, at which point they become beneficial. Newly adaptive alleles are then already present in the population at higher initial frequencies, making them less likely to be lost by genetic drift compared to new mutations.

One possible consequence of an environmental change – being by climate alteration or colonization of new habitats – is the reduction in size of a population [107]. The population bottleneck will also cause a reduction in genetic diversity and allele frequencies at some loci to differ from those of the parent population. The resulting smaller population will be subject to strong genetic drift, which will erode variation by randomly driving alleles to fixation or extinction. Hence, this new population facing a new environment will present lower levels of diversity caused by the bottleneck but also because new mutations are less frequent in small populations and more likely to be lost by genetic drift. Therefore, adaptation in small sized populations requires new mutations with effect sizes large enough to escape drift. Rare beneficial mutations will escape loss when selection is strong enough to overcome drift, and to ensure the elimination of deleterious mutations [66,67,107–109].

The richness of information for Cvi-0 together with the isolation of CVI provides an exceptional case to connect the genetic basis of adaptation with ecological drivers and evolutionary processes.

References

1. Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. Nat. Rev. Genet. 1–13 (2020) doi:10.1038/s41576-020-0250-z.

2. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. Nat. Rev. Genet. 3, 11–21 (2002).

3. Sella, G. & Barton, N. H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. Annu. Rev. Genomics Hum. Genet. 20, 461–493 (2019).

4. Buckler, E. S. et al. The Genetic Architecture of Maize Flowering Time. Science 325, 714–718 (2009).

5. Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. & Carroll, S. B. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433, 481–487 (2005).

6. Rebeiz, M., Pool, J. E., Kassner, V. A., Aquadro, C. F. & Carroll, S. B. Stepwise Modification of a Modular Enhancer Underlies Adaptation in a Drosophila Population. Science 326, 1663–1667 (2009).

7. Saccheri, I. J., Rousset, F., Watts, P. C., Brakefield, P. M. & Cook, L. M. Selection and gene flow on a diminishing cline of melanic peppered moths. Proc. Natl. Acad. Sci. 105, 16212–16217 (2008).

8. Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial Melanism in British Peppered Moths Has a Singular and Recent Mutational Origin. Science 332, 958–960 (2011).

9. Hof, A. E. van't et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534, 102–105 (2016).

10. Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. Science 313, 101–104 (2006).

11. Nachman, M. W., Hoekstra, H. E. & D'Agostino, S. L. The genetic basis of adaptive melanism in pocket mice. Proc. Natl. Acad. Sci. 100, 5268–5273 (2003).

12. Shapiro, M. D. et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428, 717–723 (2004).

13. Colosimo, P. F. et al. Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. Science 307, 1928–1933 (2005).

14. Chan, Y. F. et al. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. Science 327, 302–305 (2010).

15. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55–61 (2012).

16. Hall, M. C. & Willis, J. H. Divergent Selection on Flowering Time Contributes to Local Adaptation in Mimulus Guttatus Populations. Evolution 60, 2466–2477 (2006).

17. Howe, G. T. et al. From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. Can. J. Bot. (2011) doi:10.1139/b03-141.

18. Christians, J. K. & Senger, L. K. Fine mapping dissects pleiotropic growth quantitative trait locus into linked loci. Mamm. Genome 18, 240–245 (2007).

19. Gray, M. M. et al. Genetics of Rapid and Extreme Size Evolution in Island Mice. Genetics 201, 213–228 (2015).

20. Prasad, K. V. S. K. et al. A Gain-of-Function Polymorphism Controlling Complex Traits and Fitness in Nature. Science 337, 1081–1084 (2012).

21. Dong, L. et al. Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. Nat. Commun. 12, 5445 (2021).

22. Comadran, J. et al. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. Nat. Genet. 44, 1388–1392 (2012).

23. Zhang, Z. et al. Natural variation in CTB4a enhances rice adaptation to cold habitats. Nat. Commun. 8, 14788 (2017).

24. Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. Science 354, 54–59 (2016).

25. Olsson, K. & Ågren, J. Latitudinal population differentiation in phenology, life history and flower morphology in the perennial herb Lythrum salicaria. J. Evol. Biol. 15, 983–996 (2002).

26. Ågren, J., Oakley, C. G., Lundemo, S. & Schemske, D. W. Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping. Evolution 71, 550–564 (2017).

27. Brachi, B. et al. Investigation of the geographical scale of adaptive phenological variation and its underlying genetics in Arabidopsis thaliana. Mol. Ecol. 22, 4222–4240 (2013).

28. Le Corre, V., Roux, F. & Reboud, X. DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: extensive nonsynonymous variation is consistent with local selection for flowering time. Mol Biol Evol 19, 1261–1271 (2002).

29. 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 65, 6457–6469 (2014).

30. Fournier-Level, A. et al. Paths to selection on life history loci in different natural environments across the native range of Arabidopsis thaliana. Mol Ecol 22, 3552–3566 (2013).

31. Fournier-Level, A. et al. A map of local adaptation in Arabidopsis thaliana. Science 334, 86–89 (2011).

32. Hancock, A. M. et al. Adaptation to climate across the Arabidopsis thaliana genome. Science 334, 83–86 (2011).

33. 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).

34. Long, Q. et al. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat. Genet. 45, 884–890 (2013).

35. Exposito-Alonso, M. et al. Natural selection on the Arabidopsis thaliana genome in present and future climates. Nature 573, 126–129 (2019).

36. Alonso-Blanco, C. et al. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 166, 481–491 (2016).

37. Nordborg, M. et al. The pattern of polymorphism in Arabidopsis thaliana. PLOS Biol. 3, e196 (2005).

38. Durvasula, A. et al. African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. Proc Natl Acad Sci USA 114, 5213 (2017).

39. Bomblies, K. et al. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet 6, e1000890 (2010).

40. Zou, Y.-P. et al. Adaptation of Arabidopsis thaliana to the Yangtze River basin. Genome Biol. 18, 239 (2017).

41. Fulgione, A., Koornneef, M., Roux, F., Hermisson, J. & Hancock, A. M. Madeiran Arabidopsis thaliana Reveals Ancient Long-Range Colonization and Clarifies Demography in Eurasia. Mol. Biol. Evol. 35, 564–574 (2018).

42. Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. Naturally occurring genetic variation in Arabidopsis thaliana. Annu. Rev. Plant Biol. 55, 141–172 (2004).

43. Keurentjes, J. J. B. et al. Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. Genetics 175, 891–905 (2007).

44. Alonso-Blanco, C. et al. Development of an AFLP based linkage map of Ler, Col and Cvi Arabidopsis thaliana ecotypes and construction of a Ler/Cvi recombinant inbred line population: AFLP based linkage map of Arabidopsis. Plant J. 14, 259–271 (1998).

45. Simon, M. et al. Quantitative Trait Loci Mapping in Five New Large Recombinant Inbred Line Populations of Arabidopsis thaliana Genotyped With Consensus Single-Nucleotide Polymorphism Markers. Genetics 178, 2253–2264 (2008).

46. Kover, P. X. et al. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. PLOS Genet. 5, e1000551 (2009).

47. Huang, X. et al. Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. Proc. Natl. Acad. Sci. 108, 4488–4493 (2011).

48. Brock, M. T., Rubin, M. J., DellaPenna, D. & Weinig, C. A Nested Association Mapping Panel in Arabidopsis thaliana for Mapping and Characterizing Genetic Architecture. G3 Genes Genomes Genet. 10, 3701–3708 (2020).

49. Balasubramanian, S. et al. QTL Mapping in New Arabidopsis thaliana Advanced Intercross-Recombinant Inbred Lines. PLOS ONE 4, e4318 (2009).

50. Krysan, P. J. et al. Characterization of T-DNA insertion sites in Arabidopsis thaliana and the implications for saturation mutagenesis. Omics J. Integr. Biol. 6, 163–174 (2002).

51. Rosso, M. G. et al. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant Mol. Biol. 53, 247–259 (2003).

52. Gaut, B. Arabidopsis thaliana as a model for the genetics of local adaptation. Nat. Genet. 44, 115–116 (2012).

53. Franks, S. J., Sim, S. & Weis, A. E. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. Proc. Natl. Acad. Sci. 104, 1278–1282 (2007).

54. Postma, F. M. & Ågren, J. Early life stages contribute strongly to local adaptation in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. 113, 7590–7595 (2016).

55. Bloomer, R. H. & Dean, C. Fine-tuning timing: natural variation informs the mechanistic basis of the switch to flowering in Arabidopsis thaliana. J. Exp. Bot. 68, 5439–5452 (2017).

56. Elzinga, J. A. et al. Time after time: flowering phenology and biotic interactions. Trends Ecol. Evol. 22, 432–439 (2007).

57. Stinchcombe, J. R. et al. A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci USA 101, 4712–4717 (2004).

58. Andrés, F. & Coupland, G. The genetic basis of flowering responses to seasonal cues. Nat. Rev. Genet. 13, 627–639 (2012).

59. Mouradov, A., Cremer, F. & Coupland, G. Control of flowering time: interacting pathways as a basis for diversity. Plant Cell 14, S111–S130 (2002).

60. Hopkins, R., Schmitt, J. & Stinchcombe, J. R. A latitudinal cline and response to vernalization in leaf angle and morphology in Arabidopsis thaliana (Brassicaceae). New Phytol. 179, 155–164 (2008).

61. Lewandowska-Sabat, A. M., Fjellheim, S., Olsen, J. E. & Rognli, O. A. Local Populations of Arabidopsis thaliana Show Clear Relationship between Photoperiodic Sensitivity of Flowering Time and Altitude. Front. Plant Sci. 8, 1046 (2017).

62. Singh, A. & Roy, S. High altitude population of Arabidopsis thaliana is more plastic and adaptive under common garden than controlled condition. BMC Ecol. 17, 39 (2017).

63. Lutz, U. et al. Modulation of Ambient Temperature-Dependent Flowering in Arabidopsis thaliana by Natural Variation of FLOWERING LOCUS M. PLOS Genet. 11, e1005588 (2015).

64. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. Earth Environ. Sci. Trans. R. Soc. Edinb. 52, 399–433 (1919).

65. Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. Nat. Rev. Genet. 1–13 (2020) doi:10.1038/s41576-020-0250-z.

66. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52, 935–949 (1998).

67. Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. Evolution 56, 1317–1330 (2002).

68. Orr, H. A. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6, 119–127 (2005).

69. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. Nature 225, 563–564 (1970).

70. Gillespie, J. H. Some properties of finite populations experiencing strong selection and weak mutation. Am. Nat. 121, 691–708 (1983).

71. Gillespie, J. H. Molecular evolution over the mutational landscape. Evolution 38, 1116–1129 (1984).

72. Bradshaw, H. D., Wilbert, S. M., Otto, K. G. & Schemske, D. W. Genetic mapping of floral traits associated with reproductive isolation in monkeyflowers (Mimulus). Nature 376, 762–765 (1995).

73. Bradshaw, H. D., Jr., Otto, K. G., Frewen, B. E., McKay, J. K. & Schemske, D. W. Quantitative Trait Loci Affecting Differences in Floral Morphology Between Two Species of Monkeyflower (Mimulus). Genetics 149, 367–382 (1998).

74. Booker, T. R. & Keightley, P. D. Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome. Mol. Biol. Evol. 35, 2971–2988 (2018).

75. Baxter, I. et al. A Coastal Cline in Sodium Accumulation in Arabidopsis thaliana Is Driven by Natural Variation of the Sodium Transporter AtHKT1;1. PLOS Genet. 6, e1001193 (2010).

76. Albert, A. Y. K. et al. The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. Evol. Int. J. Org. Evol. 62, 76–85 (2008).

77. Dittmar, E. L., Oakley, C. G., Ågren, J. & Schemske, D. W. Flowering time QTL in natural populations of Arabidopsis thaliana and implications for their adaptive value. Mol. Ecol. 23, 4291–4303 (2014).

78. Brachi, B. et al. Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature. PLOS Genet. 6, e1000940 (2010).

79. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9, 29 (2013).

80. Borevitz, J. O. & Nordborg, M. The Impact of Genomics on the Study of Natural Variation in Arabidopsis. Plant Physiol. 132, 718–725 (2003).

81. Weigel, D. Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics1[W][OA]. Plant Physiol. 158, 2–22 (2012).

82. Zhao, K. et al. An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet. 3, e4 (2007).

83. W. Lobin. The occurrance of Arabidopsis thaliana in the Cape Verde Islands. in vol. 20 119–123 (1983).

84. Kliebenstein, D., Pedersen, D., Barker, B. & Mitchell-Olds, T. Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in Arabidopsis thaliana. Genetics 161, 325–332 (2002).

85. Alonso-Blanco, C. et al. Genetic and molecular analyses of natural variation indicate CBF2 as a candidate gene for underlying a freezing tolerance quantitative trait locus in Arabidopsis. Plant Physiol 139, 1304–1312 (2005).

86. McKhann, H. I. et al. Natural variation in CBF gene sequence, gene expression and freezing tolerance in the Versailles core collection of Arabidopsis thaliana. BMC Plant Biol. 8, 105 (2008).

87. Edwards, K. D., Lynn, J. R., Gyula, P., Nagy, F. & Millar, A. J. Natural allelic variation in the temperature-compensation mechanisms of the Arabidopsis thaliana circadian clock. Genetics 170, 387–400 (2005).

88. Kim, T.-S., Wang, L., Kim, Y. J. & Somers, D. E. Compensatory mutations in GI and ZTL may modulate temperature compensation in the circadian clock. Plant Physiol. 182, 1130–1141 (2020).

89. Bandaranayake, C. K., Koumproglou, R., Wang, X. Y., Wilkes, T. & Kearsey, M. J. QTL analysis of morphological and developmental traits in the Ler × Cvi population of Arabidopsis thaliana. Euphytica 137, 361–371 (2004).

90. Alonso-Blanco, C., Bentsink, L., Hanhart, C. J., Vries, H. B. & Koornneef, M. Analysis of natural allelic variation at seed dormancy loci of Arabidopsis thaliana. Genetics 164, 711–729 (2003).

91. Alonso-Blanco, C., Vries, H. B., Hanhart, C. J. & Koornneef, M. Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana. PNAS 96, 4710–4717 (1999).

92. Bentsink, L. et al. Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of Arabidopsis. Plant Physiol. 124, 1595–1604 (2000).

93. Bentsink, L. et al. Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways. PNAS 107, 4264–4269 (2010).

94. Bentsink, L., Yuan, K., Koornneef, M. & Vreugdenhil, D. The genetics of phytate and phosphate accumulation in seeds and leaves of Arabidopsis thaliana, using natural variation. Theor Appl Genet 106, 1234–1243 (2003).

95. Gilliland, L. U. et al. Genetic basis for natural variation in seed vitamin E levels in Arabidopsis thaliana. PNAS 103, 18834–18841 (2006).

96. Laserna, M. P., Sánchez, R. A. & Botto, J. F. Light-related loci controlling seed germination in Ler x Cvi and Bay-0 x Sha recombinant inbred-line populations of Arabidopsis thaliana. Ann Bot 102, 631–642 (2008).

97. Marais, D. L. D. et al. Variation in MPK12 affects water use efficiency in Arabidopsis and reveals a pleiotropic link between guard cell size and ABA response. PNAS 111, 2836–2841 (2014).

98. McKay, J. K., Richards, J. H. & Mitchell-Olds, T. Genetics of drought adaptation in Arabidopsis thaliana: Pleiotropy contributes to genetic correlations among ecological traits. Mol Ecol 12, 1137–1151 (2003).

99. Gazzani, S., Gendall, A. R., Lister, C. & Dean, C. Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. Plant Physiol 132, 1107–1114 (2003).

100. Shindo, C. et al. Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. Plant Physiol 138, 1163 (2005).

101. El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J. M., Raz, V. & Koornneef, M. A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. Nat. Genet. 29, 435–440 (2001).

102. El-Din El-Assal, S. et al. The Role of Cryptochrome 2 in Flowering in Arabidopsis. Plant Physiol. 133, 1504–1516 (2003).

103. Losos, J. B. & Ricklefs, R. E. Adaptation and diversification on islands. Nature 457, 830–836 (2009).

104. Vahdati, A. R. & Wagner, A. Population Size Affects Adaptation in Complex Ways: Simulations on Empirical Adaptive Landscapes. Evol. Biol. 45, 156–169 (2018).

105. Dittmar, E. L., Oakley, C. G., Conner, J. K., Gould, B. A. & Schemske, D. W. Factors influencing the effect size distribution of adaptive substitutions. Proc. R. Soc. B Biol. Sci. 283, 20153065 (2016).

106. Höllinger, I., Pennings, P. S. & Hermisson, J. Polygenic adaptation: From sweeps to subtle frequency shifts. PLOS Genet. 15, e1008035 (2019).

107. Willi, Y., Buskirk, J. & Hoffmann, A. A. Limits to the Adaptive Potential of Small Populations. Annu. Rev. Ecol. Evol. Syst. 37, 433–458 (2006).

108. Bell, G. & Gonzalez, A. Evolutionary rescue can prevent extinction following environmental change. Ecol. Lett. 12, 942–948 (2009).

109. Whitlock, M. C. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. Evolution 54, 1855–1861 (2000).

CHAPTER I

Parallel reduction in flowering time from new mutations enabled evolutionary rescue in colonizing Arabidopsis lineages

Andrea Fulgione[1,2,3†], Célia Neto[1†], Ahmed F. Elfarargi[1], Emmanuel Tergemina[1], Shifa Ansari[1], Mehmet Göktay[1], Herculano Dinis[4,5], Nina Döring[1], Pádraic J. Flood[1‡], Sofia Rodriguez-Pacheco[1], Nora Walden[6§], Marcus A. Koch[6], Fabrice Roux[7], Joachim Hermisson[2], Angela M. Hancock[1,2*]

1 Max Planck Institute for Plant Breeding Research, Cologne, Germany.

2 Mathematics and Bioscience, Department of Mathematics and Max F. Perutz Labs, University of Vienna, Vienna, Austria.

3 Vienna Graduate School for Population Genetics, Vienna, Austria.

4 Parque Natural do Fogo, Direção Nacional do Ambiente, Praia, Santiago, Cabo Verde.

5 Associação Projecto Vitó, São Filipe, Fogo, Cabo Verde.

6 Centre for Organismal Studies (COS) Heidelberg, Biodiversity and Plant Systematics, Heidelberg University, Heidelberg, Germany.

7 LIPME, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France.

‡ Current address: Plant Breeding, Wageningen University, Wageningen, The Netherlands.

§ Current address: Biosystematics, Wageningen University, Wageningen, The Netherlands.

† Equal contribution

* Correspondence to: hancock@mpipz.mpg.de

Abstract

Populations subject to abrupt environmental change must adapt quickly to avoid extinction. We sequenced genomes of 335 newly collected wild *Arabidopsis* samples from the Cape Verde Islands to investigate the mechanisms of adaptation after a sudden shift to a more arid climate. We found that the island populations represent diverged, monophyletic lineages based on the near absence of shared polymorphisms with each other or the continent. Patterns of genetic variation within and between islands revealed a polygenic response to multivariate selection. Time to flowering was reduced in parallel across islands, substantially increasing fitness, and this change was mediated by convergent *de novo* loss of function of two core flowering time genes: *FRI* on one island and *FLC* on the other. Evolutionary reconstructions reveal a case where expansion of the new populations coincided with the emergence and proliferation of these novel variants, consistent with models of rapid adaptation and evolutionary rescue.

Introduction

One in eight of the world's existing plant and animal species are at risk of extinction due to human-mediated environmental change [1]. To forecast and mitigate risk, it is necessary that we understand the mechanisms of adaptation to novel environmental challenges. On the one extreme, adaptation can be highly polygenic, with contributions from many small effect variants [2–5]. Conversely, when selection pressures are very strong and existing genetic variation is low, large effect variants are expected to provide a crucial contribution to adaptation [6–8]. Theoretical models show the importance of genetic diversity and the strength of selection for shaping the architecture of adaptive response [6,7,9–14].

In practice, reconstructing detailed adaptive histories in natural populations is challenging. However, long-range colonization events can represent powerful natural experiments where populations are deposited in replicate in a new environment [9,15–19]. The resulting isolated populations provide an opportunity to examine evolutionary processes in the absence of confounding from admixture and secondary contact. We collected new wild *Arabidopsis* populations living in a climatic extreme of the species range – the Cape Verde Islands (CVI) – to investigate the molecular and ecological basis of adaptation to sudden and extreme environmental change.

The CVI archipelago consists of ten islands located between 14.80 and 17.20 degrees north of the equator and 570 km off the coast of Senegal. The flora in CVI is a mix of native species that reached the islands via long range dispersal from mainland Africa and Macaronesia and species introduced since 1456, when humans first settled in the islands [20,21]. Precipitation in CVI is limited and unpredictable – so that plants must grow quickly and reproduce in the short time when water is available [20]. A single *Arabidopsis* line from Cape Verde (Cvi-0) was collected 37 years ago [22] and has since been studied extensively both at the phenotypic and genetic levels. The wealth of information for Cvi-0 together with the isolation of *Arabidopsis* in CVI provides a powerful case where it is possible to connect the genetic basis of adaptive change with ecological drivers and fitness differentials. Further, this case can inform us about more general mechanisms of adaptation after rapid environmental change.

Results

1. Reconstructing demographic history of CVI Arabidopsis from genome-wide patterns of variation

We collected *Arabidopsis* across its distribution in CVI (Fig. 1a, Supplementary Fig. 1, Supplementary Table 1), where it is limited to the islands Santo Antão and Fogo, and sequenced complete genomes of 335 lines. Compared to Eurasian and Moroccan collection locations, the *Arabidopsis* habitat in Cape Verde is more arid (median aridity index in CVI: 0.21, Morocco: 0.25, Eurasia: 0.78; Mann-Whitney-Wilcoxon (MWW) for CVI-Eurasia: $p=3.41x10^{-35}$ and CVI-Morocco: $p=5.97x10^{-4}$) with higher precipitation seasonality (median in CVI: 144.24, Morocco: 54.00, Eurasia: 25.94; MWW CVI-Eurasia: $p=2.01x10^{-36}$ and CVI-Morocco: $p=3.8x10^{-11}$), and a shorter growing season (median in CVI: 3.5 months, Morocco: 8 months, Eurasia: 8 months; MWW CVI-Eurasia: $p=2.72x10^{-35}$ and CVI-Morocco: $p=4.13x10^{-12}$) (Supplementary Fig. 2, Supplementary Table 2). The strong climatic divergence of CVI suggests nascent CVI populations may have been subject to strong selection.

We reconstructed the colonization history of CVI *Arabidopsis* by analysing CVI genomes together with published data [23]. Genome-wide, the two Cape Verde islands cluster tightly together and are nested within the Moroccan clade (Fig. 1b). Diversity within islands is 73.3- and 62.3-fold reduced compared to the continent (θ_W (Santo Antão) = $7.59x10^{-5}$, θ_W (Fogo) = $8.93x10^{-5}$, θ_W (Morocco) = $5.56x10^{-3}$; Supplementary Table 3) and there is almost no shared variation between the islands and Morocco or between the two Cape Verde Islands (Fig. 2a-b). Genome-wide, 99.9% of variants in CVI are absent in Morocco and 99.4% of variants segregating in Cape Verde are private to a single island. Similarly, at 4-fold degenerate sites, 99.9% are private to Cape Verde and 98.2% are private to only one island (Fig. 2b). Linkage disequilibrium decays rather rapidly in each island population (Supplementary Fig. 3), consistent with the near-complete loss of segregating variation with colonization (i.e., lack of deep population structure) and subsequent population expansion [24,25].

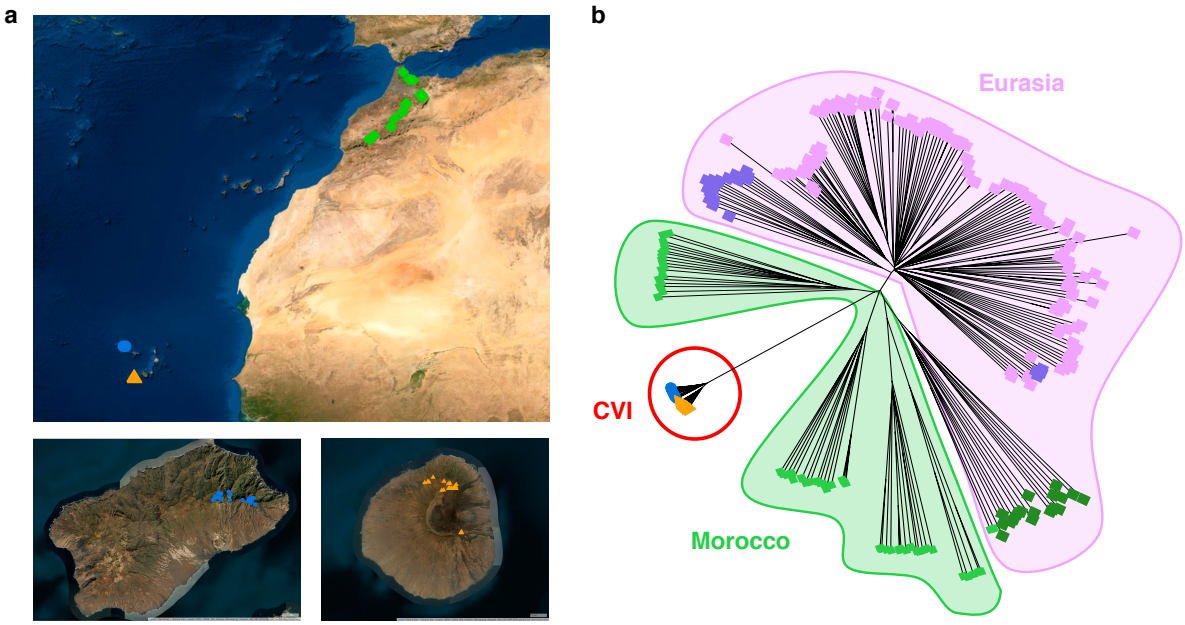


Figure 1. Population structure of Cape Verde Arabidopsis. a, Sample locations of sequenced lines from Morocco (green, n=64) and the Cape Verde Islands, Santo Antão (blue, n=189) and Fogo (orange, n=146). b, Neighbour-joining tree showing relationship of CVI (n=336, in blue and orange) to worldwide samples. Morocco: n=64, in green; Eurasia: n=180, including 20 samples, each from 9 representative clusters [26]. Divergent Iberian lines, or relicts, are shown in dark green, other Iberian lines (non-relicts) are shown in purple, and all other Eurasian lines in magenta.

These levels of differentiation between CVI and the Moroccan mainland as well as between CVI islands are striking. This divergence is higher than that observed between species pairs in the *Arabidopsis* genus, which ranges from 72.6% to 96.9% private 4-fold degenerate segregating variants [27]. As a result, each Cape Verde Island population forms a diverged, monophyletic group and is thus phylogenetically distinct, and will be treated as such here for the purposes of genetic analysis. Further, the patterns we observe for these lineages are analogous to those inferred for most named endemic species in Cape Verde, which have clear ecogeographic separation [20,21,28] and often retain inter-compatibility [20], so that the CVI *Arabidopsis* lineages could serve as a useful model for island endemic species more generally.

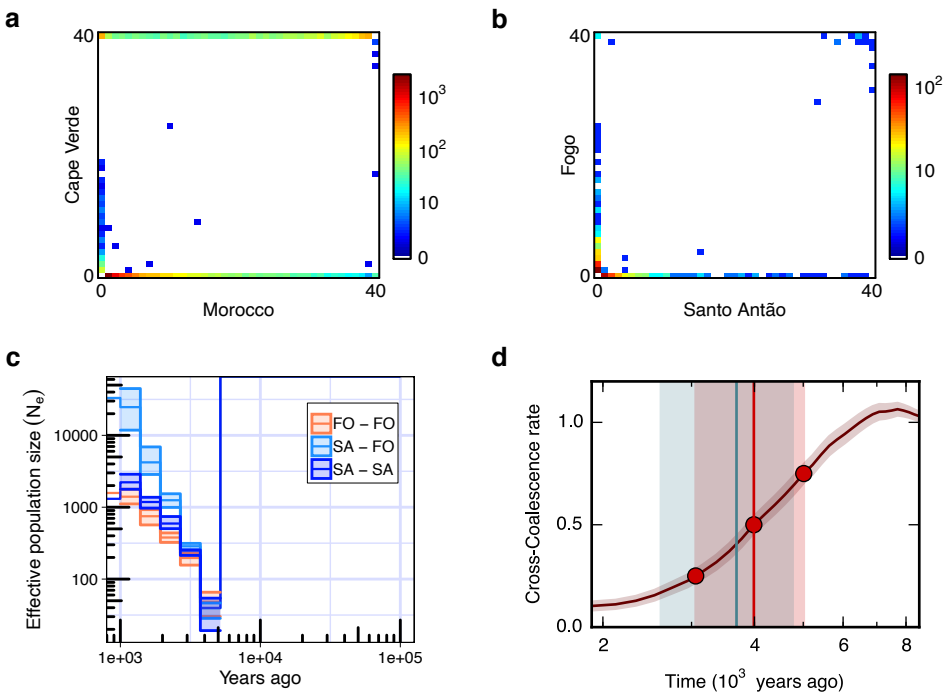


Figure 2. Demographic history of Cape Verde *Arabidopsis*. a, Joint site frequency spectrum between Cape Verde and Morocco and, b, between the two islands, Santo Antão and Fogo. The colour scale represents the number of variants in each frequency class. c, Historical population size trajectories within (SA-SA, FO-FO) and between islands (SA-FO) inferred from RELATE. Shaded areas represent the 95% CI calculated from the genome-wide distribution of coalescence events. d, Estimated split times between islands from MSMC-CCR (point estimate at the vertical red line (4.0 kya), 0.25 – 0.75 cross-coalescence rate quantiles shown as shaded red area (3.1 – 5.0 kya)), and dadi (point estimate at the vertical blue line (3.7 kya), 95% CI as shaded blue area (2.6 - 4.8 kya)).

Although the Moroccan High Atlas population is genetically most similar to CVI across the genome (61%), there are prominent examples where it is not – including the chloroplast and the S-locus (Supplementary Fig. 4-6, Supplementary Results) – suggesting that an unsampled 'ghost' population best represents the outgroup. To obtain an upper (i.e., more ancient) bound on colonization time, we modelled the split between CVI and this 'ghost' population. We used multiple complementary approaches, including inference based on the joint site frequency spectrum, reconstruction of coalescence events across the genome, and comparisons to forward simulations [29–32]. These analyses revealed an initial separation between the Moroccan population and the CVI progenitor 'ghost' population at 40-60 kya, followed by colonization of CVI from the 'ghost' population as early as 7-10 kya (Supplementary Fig. 7, Supplementary Results).

To obtain a lower (i.e., more recent) bound on colonization time, we next examined coalescence time within CVI. Historical reconstruction [30,33] indicated that both islands were colonized through strong bottlenecks, which eliminated nearly all pre-existing variation (Fig. 2a-b). Using haplotype coalescence events we estimated the number of colonizers [32] and confidence intervals around these [34]. The estimated number of founders was 40 individuals (95% CI: 19-54) in Santo Antão and 48 individuals (95% CI: 30-66) in Fogo [32] (Fig. 2c). After the initial colonization, random effects of allele sampling (i.e., genetic drift) would have resulted in further reduction in diversity and sharing with ancestral populations. To quantify this effect, we ran simulations based on the inferred effective population sizes over time starting with 40 founders (Supplementary Note). These revealed that in the present-day population only 1.7 (95% CI: 0.6-3) variants in 10,000 are expected to have come from the original founding population. This implies that nearly all variation segregating in CVI results from new mutations that occurred after colonization.

Between the two islands, patterns of variation differ, with Santo Antão displaying a higher proportion of private variation at segregating sites and Fogo displaying a higher proportion of private fixed variants (Fig. 2b). Consistent with this, we found evidence for deep population structure and restricted gene flow in Santo Antão, based on haplotype divergence among sub-populations. The overall pattern suggests early population subdivision followed by later population expansion across the island, with N_e increasing sharply in the past 3 ky (Fig. 2c). In Fogo, the more arid island, there is no evidence of early separation into sub-populations. Rather, we find a clear signal that after an initial moderate expansion (from approx. 48 individuals to 400 individuals) the population remained panmictic and restricted in size for approx. 830-940 years after colonization (Fig. 2c, Supplementary Fig. 8). Overall, our inference supports a model in which Santo Antão was colonized first (approximately 5-7 kya), and Fogo was colonized from Santo Antão approximately 3-5 kya [29,30,32] (Fig. 2c-d, Supplementary Fig. 8, Supplementary Results). Our inferences clearly place the initial colonization of CVI well before colonization by humans, which only occurred approx. 560 years ago, implying that colonization occurred by natural (non-human) dispersal, e.g., by wind-mediated transport. Figure 3 provides a schematic of the history that combines results from the different population genetic analyses.

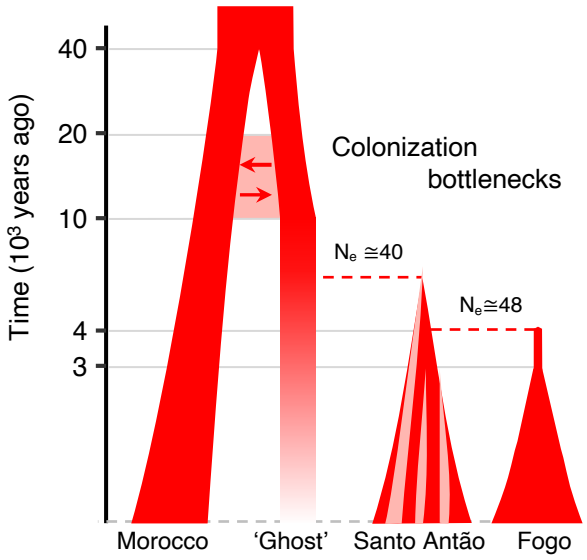


Figure 3. Schematic of the inferred history of CVI *Arabidopsis*. N_e: effective population size; arrows denote migration. The y-axis is log_{10} transformed.

2. Moroccan climatic niche and suitability of CVI landscape

To infer the suitability of the CVI climate to the colonizers when they initially arrived, we modelled the climatic niche of Moroccan *A. thaliana* and predicted suitability in CVI based on this model. We used Maxent [35] to model the factors that limit the distribution of *Arabidopsis* in Morocco based on georeferenced collection locations (Fig. 4a) and the set of bioclimatic variables listed in Supplementary Table 2. The main contributors to the model were the length of the growing season (38.7%), isothermality (20.2%), minimum temperature in the coldest month (18.4%) and maximum temperature in the warmest month (14.5%); (model AUC: 0.938 (std dev = 0.088); Fig. 4b; Supplementary Table 4-5). We predicted suitability of the CVI environment by projecting this model onto the CVI landscape. This analysis identified no suitable regions for Moroccan *Arabidopsis* in CVI (Fig. 4c). This may be expected given that distributions of climate variables taken from CVI collection locations are often outside of the range of those at Moroccan collection locations (Supplementary Figure 2, Supplementary Table 2). Therefore, we also used an approach to examine the multivariate environmental similarity surface. The regions with highest climatic similarity from this analysis (Fig. 4d) are those where *Arabidopsis* can be found in Santo Antão and Fogo (Fig. 1a, Supplementary Fig. 1a-b). Although there is the possibility that at the time of colonization the climates were somewhat more similar or that the Moroccan population extended into more extreme climatic zones, based on our results using present-day data, there are large differences in

many aspects of climate in CVI relative to Morocco. The overall low suitability and similarity of the CVI environment compared to that of the Moroccan population is thus consistent with the idea that the initial colonizers would have been challenged by multiple aspects of the novel CVI environment.

3. Evidence for adaptation based on functional genetic divergence and differential fitness

Both drift and positive selection can contribute to genetic divergence. We used two approaches to investigate the role of adaptive evolution in CVI. The first is based on patterns of polymorphism and divergence within and between lineages and the second on an experimental test of relative reproductive success under CVI versus Moroccan conditions.

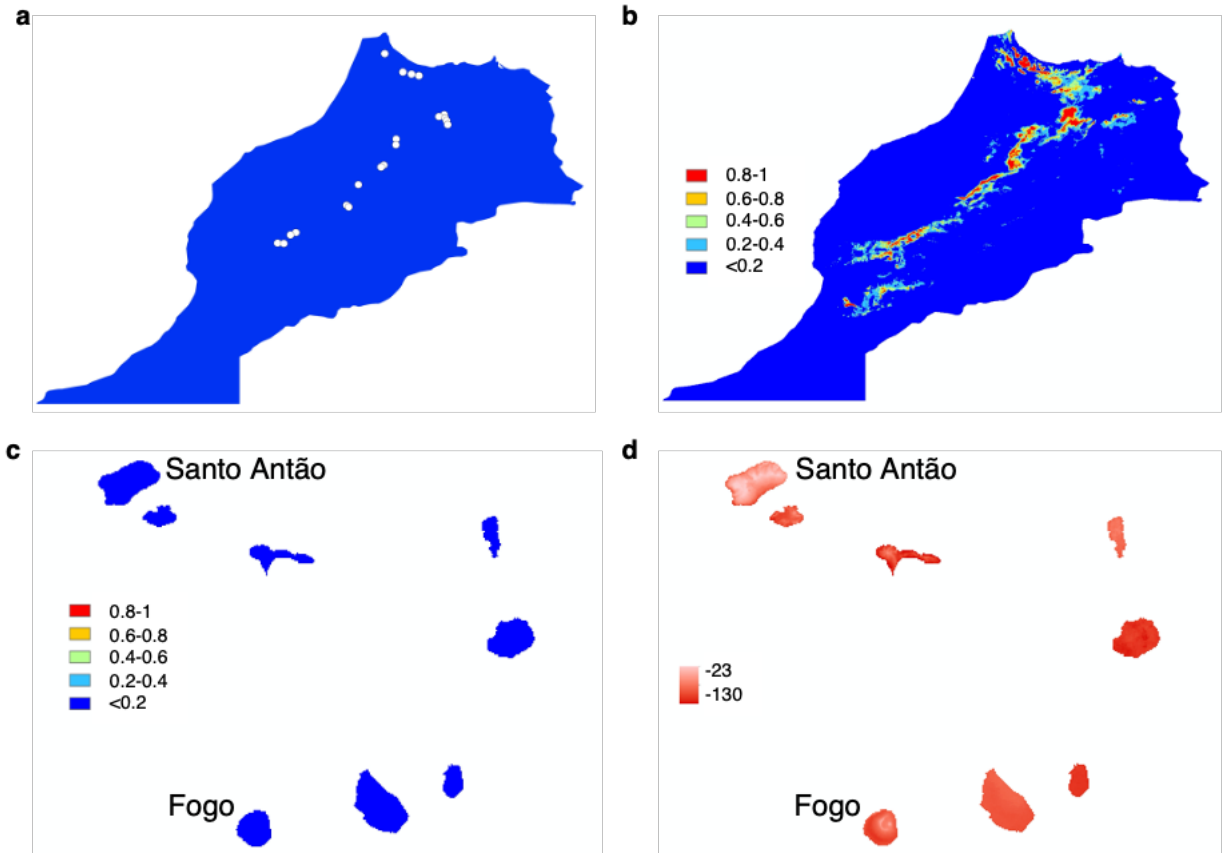


Figure 4. Moroccan and CVI predicted distributions. a, Moroccan *A. thaliana* occurrence locations and b, predicted climate envelope within Morocco. In (b-c) colours represent the predicted probability of occurrence and habitat suitability, with blue indicating low probability and red high. c,

Predicted climatic suitability for Moroccan *Arabidopsis* in CVI and d, predicted similarity of CVI climate to the Moroccan *A. thaliana* climate envelope expressed as a percentage of how dissimilar each point is in relation to the range of values used in the model. More negative (red) values indicate higher dissimilarity relative to Morocco.

We further inferred the distribution of fitness effects (DFE) [36] based on segregating variation, or more specifically, the discretised distribution of scaled selection coefficients ($S=4N_es$, where N_e is the effective population size and *s* the selection coefficient). The DFE contained large peaks corresponding to nearly neutral effects ($-1<S<0$) and smaller peaks corresponding to strongly positive ($1<S<10$) and negative effects ($S<-10$) (Fig. 5c). In Fogo, fixed non-synonymous mutations were prominent in the DFE, representing a classic signature of positive selection at the clade level, while in Santo Antão, non-synonymous mutations at intermediate to high frequency were more prominent, consistent with population stratification and/or local adaptation [37]. Consistent with strong positive selection acting in CVI, the estimated proportion of fixed adaptive substitutions, alpha, was extremely high in CVI (70% in Santo Antão, 62% in Fogo) relative to the Moroccan population, where the estimated proportion is effectively zero. It should be noted that population history can impact estimates of alpha so that these may be somewhat inflated due to possible fixation of deleterious variants under rapid population expansion [38,39]. Conversely, in Morocco, alpha may be underestimated due to recent population bottlenecks. While the limited numbers of fixed and segregating sites in the relatively young CVI lineages necessarily leads to large confidence intervals on our estimates (Fig. 5b-c), the results are consistent with strong positive selection after a shift to a new adaptive optimum in the nascent CVI lineages.

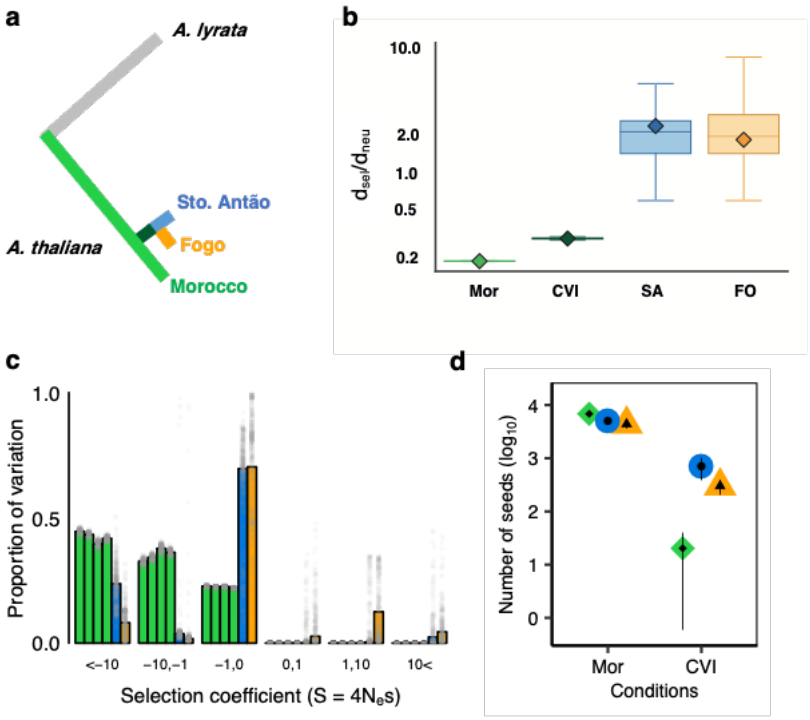


Figure 5. Population genetic signatures of adaptive evolution in CVI. a, Schematic of phylogeny separating branches examined in d_{sel}/d_{neu} analysis. b, Evolutionary rate ratios d_{sel}/d_{neu} across populations (observed data shown as diamonds) with 500 bootstrap resampling replicates showing median (line) and 95% confidence interval (whiskers). Mor: Morocco; CVI: branch between Morocco and CVI; SA: branch from the island split to Santo Antão; FO: branch from the island split to Fogo. c, Distribution of fitness effects for Morocco (green), Santo Antão (blue), and Fogo (orange). Grey dots represent estimates from 500 bootstrap analyses. d, Fitness scaled to seed number in CVI and Moroccan lines under CVI and Moroccan conditions. Medians per population are shown by the dots, and 95% CI by the whiskers. Y-axis values are $\log_{10}$ transformed.

Although population genetic approaches can provide evidence for positive selection, they make several assumptions. Therefore, we also tested for evidence of local adaptation in CVI and Moroccan clades based on evidence for higher relative fitness in local versus foreign environments. We propagated CVI and Moroccan lines in growth chambers set to match CVI and Moroccan environments (Supplementary Fig. 9a-b) and scored fitness (number of seeds produced). These experiments aimed to examine the fitness effects of climatic factors that differentiate CVI and Morocco and would not capture biotic or edaphic factors important for fitness. We tested for population, environment and population by environment effects using negative binomial GLM to correct for overdispersion. In the CVI environment, we found CVI lines performed significantly better than Moroccan lines ($b_{population}$=2.90, P=3.58x10^{-4}). In the Moroccan environment, all lines performed better compared to the CVI environment, ($b_{pop\text{-}CVI}$=2.63,

P=0.0151; $b_{pop\text{-}Mor}$=5.86, P<2x10^{-16}). There was no significant difference in fitness for the Moroccan and CVI lines in the Moroccan simulated environment (b=0.337, P=0.679). (Fig. 5d, Supplementary Table 6). Taken together these results highlight the challenging climatic conditions plants would have faced upon colonization of CVI, consistent with the results from the climate niche analysis (Fig. 4).

4. Evidence for ongoing multi-variate adaptation in Santo Antão

Next, we examined the nature of adaptation in Cape Verde by capitalizing on over twenty years of studies on Cvi-0. We identified QTL, candidate genes and specific functional variants from a meta-analysis of 129 QTL mapping studies and associated fine-mapping studies conducted in a recombinant population produced from a cross between Cvi-0 and L*er*-0^{40} (Fig. 6a, Supplementary Table 7). This data set allowed us to ask whether genetic polymorphisms that underlie the observed trait divergence between Cvi-0 and other worldwide lines (with L*er*-0 as the European representative) were present in the colonizing population or whether they represent variation that arose from *de novo* mutations after colonization. Based on the deep divergence between the RIL parents (Cvi-0 and L*er*-0), we expected that most or all of the variants would be found on the long divergence branch that separates the two Cape Verde islands from continental populations. This expectation can be quantified based on the background level of variation: genome-wide, 99.23% of the variants that segregate between Cvi-0 and L*er*-0 are fixed in CVI and therefore may have been present in the colonizing population. The remaining 0.77% are private to Santo Antão (the island of origin of Cvi-0; Supplementary Fig. 1) and absent in Fogo, and therefore can be inferred to have originated in CVI as new mutations (Fig. 6b). The null expectation was that only a small proportion of functional variation (roughly equal to the genome-wide level) would be private to Santo Antão.

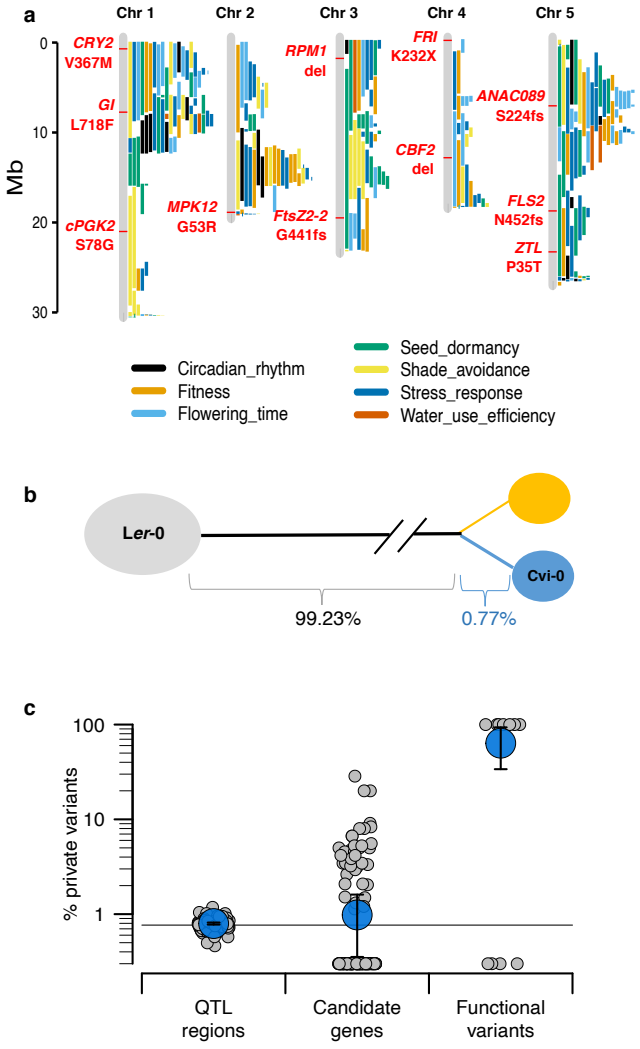


Figure 6. Evidence of multi-variate selection from 129 QTL mapping analyses using Cvi-0. a, Representative subset of previously identified QTL in Cvi-0 x *Ler*-0 RILs, with validated functional variants in red. Each segment along the chromosomes represents one QTL region with colours representing phenotypic classes. b, Schematic of the relationship between the RILs parents, with branch lengths proportional to the percentage of genome-wide variation fixed in CVI and segregating in Santo Antão. c, Percentage of variation private to Santo Antão in QTL regions, candidate genes and functional variants. Horizontal line shows the genome-wide average. Each grey dot represents a single QTL, candidate gene or functional variant, blue dots represent the average, and the whiskers the 95% CI.

At QTL mapping intervals, which cover most of the genome, we found very slight and non-significant enrichment of private variation relative to the genome-wide proportion (1.02-fold enrichment, Poisson test P=0.2723; Fig. 6c). This increased at candidate genes (1.30-fold enrichment, Poisson test

P=0.078) and became strongly significant at validated functional variants (87-fold enrichment, Poisson test P=1.417x10^{-10}). Functional variants private to Santo Antão affect core genes involved in flowering and light signalling (*CRY2* V367M [41], *FRI* K232X [42], *GI* L718F [43,44]), immunity against bacterial pathogens (*FLS2* N452fs [45]), stomatal aperture and water use efficiency (*MPK12* G53R [46]), chloroplast size (*FtsZ2-2* G441fs [47]), and fructose sensitivity similar to ABA- and ethylene-signalling mutant phenotypes (*ANAC089* S224fs [48]). These variants all segregate within Santo Antão at intermediate to high frequencies (between 0.43 and 0.89) and most are involved in functions that could underlie adaptation to the more drought-prone environment plants colonizing CVI would face. This suggests that adaptation on these variants is ongoing in Santo Antão. The strong enrichment of functional variation private to and segregating within Santo Antão implies that CVI *Arabidopsis* is adapting using novel variation that arose after colonization rather than variation inherited from North African ancestors. Further, the absence of these variants in *Arabidopsis* populations in Fogo implies that different genetic variants are involved in adaptation there.

To assess the effects of these seven private functional variants on fitness, we conducted a linear regression with these as predictors of fitness. All together they explain 22.58% of the within-island variation in fitness, which was significantly more than expected based on randomly sampled sets of seven variants across an LD-pruned genome (empirical P = 4.99x10^{-4}). Then, we used stepwise regression to identify the variants with the strongest effects on fitness. The best model based on the RMSE over 1000 bootstrap replicates explained 22.04% of the within-island variation and included two variants in flowering time pathway genes with significant effects, *FRI* K232X and *GI* L718F (Supplementary Table 8). Cvi-0 is known for its fast flowering time relative to many other populations [40,49]. Based on this, we focused specifically on the flowering time trait.

5. Mapping and historical reconstruction reveal convergent genetic adaptation to reduce flowering time

We scored flowering time as days to bolting in plants grown in simulated CVI conditions. We found that plants from both islands flowered significantly earlier than Moroccans (MWW test, W = 1620, P < 2.2x10^{-16}; Fig. 7a) and the majority of Moroccan lines never bolted in CVI conditions, resulting in a strong negative association between flowering time and fitness (Spearman's rho = -0.85, P < 2.2x10^{-16}; Fig. 7b). This is consistent with previous suggestions that reducing flowering time may allow escape from drought and provide an important fitness advantage [50]. To ask whether early flowering in the two islands results from the same or different variants, we examined segregation in three inter-island F2 populations (Fig. 7c). In each of these, flowering time was transgressive with some individuals flowering as early or earlier

than the parents and some flowering much later (two-tailed Dunnett's tests with Fisher's method, S=67.187, P = 1.54×10^{-12}). Taken together, these results imply that flowering time was reduced in CVI by convergent evolution involving mutations at different loci in the two islands.

To identify the loci responsible for reduced flowering time, we performed GWAS using a linear mixed model (LMM) to account for population structure [51]. In the Santo Antão population, we identified a single peak containing a nonsense variant, K232X, in *FRIGIDA* (*FRI*, AT4G00650), which results in faster flowering through loss of the vernalization (cold) requirement [42] (Fig. 7d). This variant explained 46.4% of the genetic variance in flowering time and 11.4% of the heritable variance in fitness. In the natural population, *FRI* 232X was associated with a 34-day decrease in flowering time (MWW test, W = 7, P < 2.2×10^{-16}), and a 140-fold increase in seed number (+387 seeds; MWW test, W = 4541, P = 7.18×10^{-14}; Fig. 7e). To further test whether loss of *FRI* was likely responsible for this effect, we compared a Col-0 transgenic line with a functional *FRI* allele to that with a non-functional *FRI* allele in the same environment and measured flowering time. We found that the effect is similar to that of the Santo Antão *FRI* 232X variant, (flowering time: -27 days, fitness: +669 seeds; MWW test W = 0, P = 3.85×10^{-3}; W = 37.5, P = 8.86×10^{-3}, respectively; Fig. 7e), further supporting the role of *FRI* 232X in flowering time reduction. *FRI* 232X is present at high frequency across all populations in Santo Antão except the early-diverging Cova de Paúl population, where it is completely absent (Supplementary Fig. 10). Coalescent reconstruction [32] of the history of *FRI* 232X indicated that the allele arose between 2.14 kya (95% CI: 1.62-2.72 kya) and 2.9 kya (95% CI: 2.14-3.74 kya) and rapidly spread across the island, with fixation likely restricted by barriers to gene flow (Supplementary Fig. 10). Based on the inferred frequency trajectory, we estimated that selection was maximized at 2-4 kya with a selection coefficient of s = 4.56% (Supplementary Table 9, Supplementary Results). The timing of the spread of *FRI* 232X is roughly coincident with the inferred expansion of *Arabidopsis* into the drier Espongeiro region of the island [32,52] (Fig. 7f, Supplementary Fig. 8e).

In Fogo, the more arid island, all individuals flowered early with low variance (mean time to flowering=29.05 days, SD=5.33 days). This suggested that at least one genetic variant underlying reduced flowering time was fixed in Fogo. Trait segregation in an inter-island F2 population (where *FRI* 232X was absent) exhibited a bimodal distribution with a 1:3 ratio (Fig. 7c top) and not major peaks in GWAS (Supplementary Fig. 11), indicating the presence of a single large effect early flowering allele. Sequencing the bulk of early flowering F2 individuals revealed a single region where the frequency of Fogo alleles reached 100%, corresponding to *FLOWERING LOCUS C* (*FLC*, AT5G10140; Fig. 7g). *FLC* is a central floral repressor that regulates genes responsible for the transition from the vegetative to the reproductive state

and is regulated by *FRI*[53]. We identified a premature truncation mutation in *FLC* (R3X), which is fixed in Fogo and absent from Santo Antão, and confirmed by qRT-PCR and genetic complementation that this mutation causes loss of function (Supplementary Fig. 12, Supplementary Results). This variant decreased flowering time by 27 days (based on the difference in modes in the F2 population, MWW test, W = 0, P < $2.2x10^{-16}$), comparable to Col-0 $FRI^{+}FLC^{-}$ (-31 days; MWW test, W = 25, P = 0.0107; Fig. 7h). Similarly, loss of function in the Col-0 background (Col-0 $FRI^{+}FLC^{-}$) resulted in higher seed production relative to Col-0 $FRI^{+}FLC^{+}$ in simulated CVI conditions (+1498 seeds; MWW test, W = 0, P = $7.5x10^{-3}$). Coalescent reconstructions and inferred frequency trajectories of *FLC* 3X indicated that it arose soon after colonization (between 3.31 kya (95% CI: 2.82-3.96 kya) and 4.72 kya (95% CI: 3.56-6.66 kya)) and was associated with strong positive selection [32,52] (*s* = 9.27%; Fig. 7i, Supplementary Fig. 13, Supplementary Table 10, Supplementary Results).

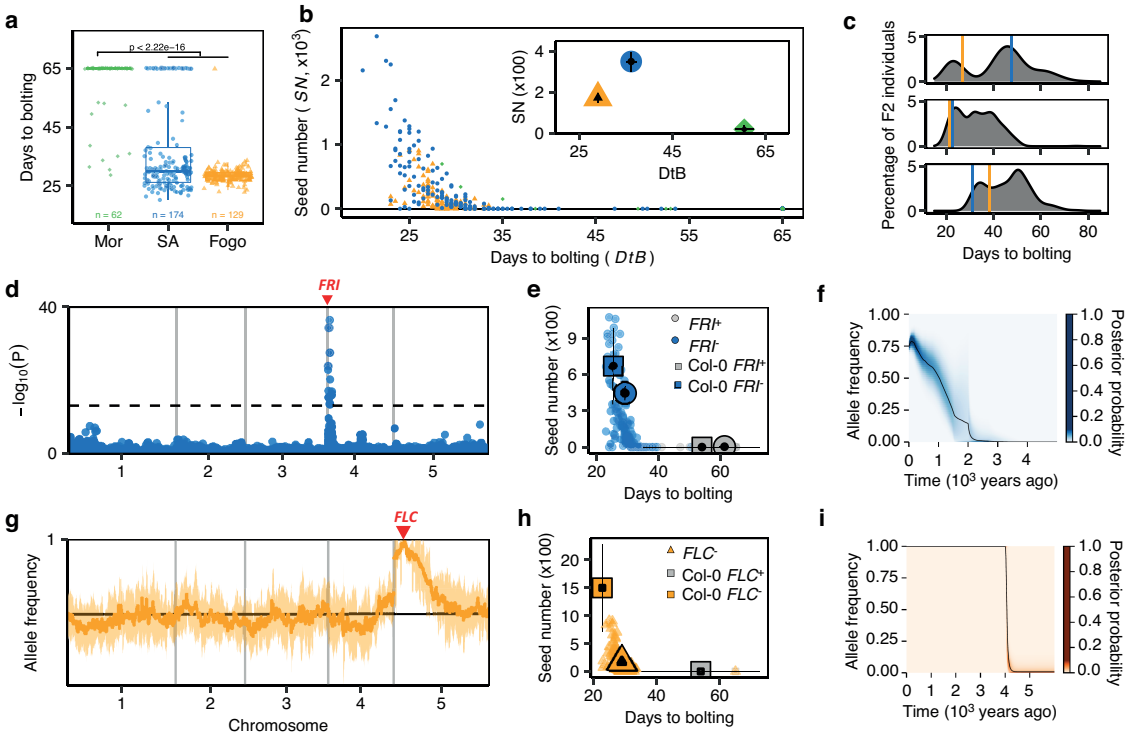


Figure 7. Adaptation through parallel reduction in flowering time in CVI. a, Bolting time in CVI relative to Morocco with MWW test to compare distributions. Boxplots show median, 1st and 3rd quartiles. b, Days to bolting versus fitness (seed number); inset: means and 95% CIs. c, Bolting time is transgressive across islands in three inter-island F2 populations. Vertical lines represent medians of days to bolting across replicates of the Santo Antão (blue) and Fogo (orange) parents. d, Bolting time GWAS in Santo Antão. Dashed line represents Bonferroni-corrected genome-wide significance. e,

Effects of *FRI* alleles on bolting time and seed number under simulated CVI conditions. Small symbols represent individual lines, large symbols population means with 95% CI (whiskers). f, Inferred allele frequency trajectory of the derived *FRI* 232X in Santo Antão. The black curve represents the posterior mean of the allele frequency and the coloured area the posterior distribution. g, Frequency distribution of the Fogo allele in an early flowering bulk from an inter-island F2 population reveals a peak at *FLC*. Median frequency per window (line) and one standard deviation (shading) are shown. h, Effects of *FLC* alleles on bolting time and seed number under simulated CVI conditions as in (e). i, Inferred allele frequency trajectory of the derived *FLC* 3X allele in Fogo, as in f.

In summary, loss of function mutations that greatly reduced flowering time appeared independently in Santo Antão (*FRI* 232X) and Fogo (*FLC* 3X) and their origins are temporally associated with initial increases in effective population size on the two islands (Fig. 2c). Because we take the inferred change in population size into account in our estimates of selection coefficients, these would be underestimated in the case that the variants themselves allow establishment and spread of populations across CVI. This may explain why the selection differentials estimated in simulated CVI environments for *FRI* and *FLC* loss of function variants are larger than the selection coefficients inferred from population genetic data. In Santo Antão, *FRI* 232X appears to have provided a strong selective advantage (Fig. 7e-f), likely enabling population expansion into drier regions of the island. In the more arid Fogo environment, the initial population appears to have been highly constrained in both size and breadth (Supplementary Results) and there is a remarkable overlap in the estimate of the time when *FLC* 3X arose and fixed in Fogo and the initial increase in population size there (Supplementary Fig. 14-15). The early appearance of these *de novo* variants is consistent with a role in evolutionary rescue of the nascent populations through reduced time to flowering.

6. Extinction risk and adaptation via new large effect mutations

Colonization of a new environment brings with it multiple challenges. Colonization events are often associated with strong bottlenecks, reducing standing genetic variation available for adaptation. When combined with a sudden and severe change in the selection regime, as may often accompany long-range colonization, extinction risk is high [7,54,55]. This is because the expected waiting time for a new beneficial mutation is likely to be greater than the expected time to extinction in a small maladapted colony [55]. Escape from extinction under this scenario is possible but relies on chance mutational events.

Theory predicts that when selection is strong and mutational input is low (i.e., a strong selection weak mutation (SSWM) regime), the first steps of adaptation are likely to occur through large effect

mutations [8,56–61]. Conversely, when mutational input is high and selection is weak (i.e., a weak selection strong mutation (WSSM) regime), adaptation is likely to occur through more, smaller effect variants. Specifically, the SSWM model is expected to hold when (i) the total number of new mutations that enter a population each generation is limited ($U_b \ll 1/4N_e$, where N_e is the effective population size and U_b is the genome-wide per-individual beneficial mutation rate for the focal trait) and (ii) selection is strong relative to drift ($s >> 1/4N_e$).

We asked where the CVI case fits in relation to the SSWM and WSSM models. First, we approximated the genome-wide mutation rate for the adaptive phenotype: very early flowering through loss of vernalization. Then, we applied our inferences about historical population size and selection coefficients to examine the fit of adaptation in CVI to these models (details in Supplementary Methods). We collated molecular information about the focal trait to produce a rough approximation of U_b for coding and regulatory changes (Supplementary Note), resulting in an estimated $U_b = 1.54 \times 10^{-6}$ mutations per site per generation. Estimates of s from reconstructed frequency trajectories were well above $1/4N_e$, and estimates of U_b were well below $1/4N_e$ in both Fogo ($s = 0.093$ and $1/4N_e = 5.21 \times 10^{-3}$) and Santo Antão ($s = 0.046$, $1/4N_e$ ranging from 2.5×10^{-4} to 5×10^{-4}; Supplementary Note), implying a SSWM regime. We also conducted forward simulations modelled after the Fogo population that incorporated the stochastic effects of drift across a range of plausible selfing rates (90%-99%; Supplementary Fig. 14, Supplementary Table 11). Taken together, our results imply that the scenarios in CVI are predictable and consistent with the SSWM regime, where mutation is limited and adaptation and establishment after initial colonization relies on sweeps of large effect alleles [5,8,56,62].

Discussion

We found several lines of evidence that adaptation was crucial for establishment of *A. thaliana* in CVI. First, early colonists from Morocco faced severe climatic challenge (Fig. 4). Second, population genetic data revealed an increased rate of nonsynonymous substitution on the branches leading to the current island populations (Fig. 5b) as well as an excess of validated functional variants within Santo Antão (Fig. 6c). Finally, we found evidence for higher relative fitness of Cape Verdean accessions compared to Moroccans in simulated conditions (Fig. 5d). The time to flowering was strongly associated with this fitness differential (Fig. 7b). Mapping (Fig. 7d,g) and evolutionary reconstructions (Fig. 7f,i) revealed that in each island, a variant that drastically reduced flowering time through loss of the vernalization (cold) requirement (*FRI* 232X, *FLC* 3X) was driven to high frequency by strong positive selection. Overall, the dynamics for both *FRI* and *FLC* mutations are consistent with a *strong selection, weak mutation* regime [56–58], where adaptation occurred by convergent loss of the vernalization requirement (Supplementary Note).

In Santo Antão, strong selection favoured early flowering (Fig. 7a,f) and was likely linked to establishment across the drier regions of the island. In more arid Fogo, population size increased in the same time frame when *FLC* 3X arose and fixed (Supplementary Fig. 15). Given the clear fitness advantage of reduced flowering time in CVI (Fig. 5d), this concordance strongly suggests that *FLC* 3X enabled escape from extinction in Fogo (Supplementary Figure 14-15).

Functional variation in *FRI* and *FLC* is widespread in natural populations of *A. thaliana* [42,63–68] and in homologues across species [69–75]. Adaptive mechanisms have been suggested to explain the prevalence of non-synonymous variation in *FRI* [76] and clinal patterns in flowering time in European *A. thaliana* populations [68,77,78]. Here, at the southern extreme of the *Arabidopsis* species distribution, the natural experiment in the isolated Cape Verde Islands allowed us to definitively connect mutations that occurred in parallel at *FRI* and *FLC* with adaptive divergence. Evolutionary convergence in this case highlights the importance of these two genes in adaptation to growing season length and aridity.

Our population genetic analyses (Fig. 5a-b) and investigation of patterns at known functional loci (Fig. 6c) further suggest that adaptation in Cape Verde was multivariate and involved many loci and traits. Some of these would be reflected in fitness differentials in the simulated CVI and Moroccan environments. But others – such as differences in biotic and edaphic factors – would not be captured in our simulated conditions. Future work in these new *Arabidopsis* island lineages will be necessary to better characterize the multivariate history of adaptation here.

Detailing the mechanisms of adaptation after a sudden environmental shift provides useful information for forecasting and ameliorating risk for vulnerable populations and species. Small, isolated populations that confront abrupt environmental change face high extinction risk [7,12,54,55]. Adaptive escape from extinction in these cases is a race with the clock, in particular when standing variation is not available. Adaptation in CVI fits well with the theoretical concept of an adaptive walk [56–58,79–81], in which a small, mutation-limited population faced a new environment far from its previous adaptive optimum and, due to the lack of standing variation, initially relied on new beneficial mutations to adapt (Supplementary Note). This is in-line with models of rapid adaptation and evolutionary rescue from new large effect mutations [6,59,79,80]. Our findings are reminiscent of work in laboratory-based microbial experiments showing that independent bouts of evolution often use the same paths [60,82–88]. Further, they suggest that adaptation to increasing aridity and shorter growing seasons – which are expected to be common under global climate change – is predictable. Therefore, our findings could also be relevant in efforts to tailor crops to drought-prone environments.

Methods

Methods described here represent a summary. Additional details can be found in the Supplementary Note.

1. Plant material

We collected plants over a series of field expeditions between 2012 and 2019 on Santo Antão and Fogo, the two islands where *A. thaliana* had been documented in herbarium records. In total, we present data for 335 lines from CVI (Supplementary Table 1, Fig. 1a), including 189 lines from 26 stands across four regions in Santo Antão (Cova de Paúl, Lombo de Figueira, Pico da Cruz and Espongeiro), and 146 lines from 18 stands across three regions in Fogo (Lava, Monte Velha and Inferno). The 62 Moroccan lines used in the study were first presented in [89] and were sequenced in [23].

2. Climate data

Climate data used in our analyses were retrieved from the Worldclim Project [90] and CGIAR Consortium (CGIAR-CSI) [91].

3. Sequencing

We sequenced the 335 new Cape Verde Islands lines and Cvi-0 using Illumina Hi-Seq and HiSeq3000 machines. Genomic DNA was extracted using the DNeasy Plant Mini kits (Qiagen), fragmented using sonication (Covaris S2), and libraries were prepared with Illumina TruSeq DNA sample prep kits (Illumina), NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs) and NEBNext Ultra II DNA Library Prep Kit (New England Biolabs). Libraries were immobilized and processed onto a flow cell with cBot (Illumina) and subsequently sequenced with 2x 100-150 bp paired end reads. We assessed DNA quality and quantity via capillary electrophoresis (TapeStation, Agilent Technologies) and fluorometry (Qubit and Nanodrop, Thermo Fisher Scientific). Due to changes in product availability over time, there were some slight differences among sequencing runs.

4. SNP identification and genotyping

We aligned the raw Illumina sequence data to the *Arabidopsis* TAIR10 reference genome and we called variants with three different pipelines. Two pipelines were previously used and described in [8] to call genotypes in the 1135 Eurasian and 64 Moroccan sequences. Here, we used the same parameters, settings, and software versions in order to analyse in a common framework the Cape Verdean and worldwide sequences (https://github.com/HancockLab). Further, for trait mapping, we used a pipeline

based on GATK [9] for the additional analyses of short indels using a modified version of the best practices workflows for germline short variant discovery (https://github.com/HancockLab/SNP_and_Indel_calling_Arabidopsis_GATK4). Average coverage across samples was 19.4x (range from 9.3x to 51.7x) after alignment to the TAIR10 reference genome.

5. Plant growth and phenotyping

For all experiments, seeds were stratified in the dark in Petri dishes on water-soaked filter paper for one week at 4°C prior to sowing. After stratification, seeds were sown in 7x7cm pots containing a standard potting compost mix. Four seeds were sown per pot and plants were thinned to one plant per pot, after germination.

5.1. CVI and Moroccan simulated conditions

We simulated the CVI growing season in a custom Bronson growth chamber based on hourly environmental data at a collection site (Supplementary Figure 9), where we measured air and soil temperature, air humidity and precipitation using data loggers. The experiment began with September 1st 2016 conditions, when we observed plants germinating at the field site. Photoperiod was set to track daylength (number of sunlight hours) in CVI. We simulated dawn and dusk by increasing light intensity by 50 µM every 15 minutes until 200µM (full light) and decreasing it by 50µM every 15 minutes until dark, respectively. At the same time points, far-red light decreased from 50 to 0 µM at dawn and increased from 0 to 50 µM at dusk. Based on precipitation data from the field, we withheld water starting 26 days after sowing. To mimic the gradual decrease in soil moisture levels we observed in the field, we used capillary mats to buffer the drought. Moroccan conditions were simulated based on matching to temperature and photoperiod in relevant locations within the Moroccan Atlas mountains [89] (https://www.worldweatheronline.com/morocco-weather.aspx). For this condition, photoperiod was set to 12 hours and plants were submitted to an eight-week cold period (4ºC) starting two weeks after sowing, to match winter temperatures.

5.2. Trait measurement

In CVI simulated conditions, we propagated 174 Santo Antão and 129 Fogo lines in four replicates each, and 64 Moroccan lines in two replicates each. Based on results from a preliminary pilot experiment, two mutants were included: Col-0 with a functional *FRI* introgressed from the Sf-2 line (Col-0 *FRI*-Sf2, shown as Col-0 $FRI^{+}FLC^{+}$) [53], and Col-0 *FRI*-Sf2 with a non-functional *FLC* allele (Col-0 *FRI*-Sf2 *flc*-3, shown as Col-0 $FRI^{+}FLC^{-}$) [53] as well as Col-0 as a control. The plants were organized in a randomized block design and Aracon tubes were added when the plants flowered to allow for the total set of seeds to be collected.

We scored flowering time, bolting time, time to anthesis, number of days until the stem reached 3 cm, and the number of rosette leaves at bolting, as in [92] as well as fitness. For downstream analyses, bolting time was used as a proxy for flowering time. The experiment was terminated ten weeks after sowing, when plants no longer produced new flowers or seeds. Plants that had not bolted at the end of the experiment were conservatively scored as bolting at 65 days (following [89]). Total number of seeds per individual was scored as a measure of fitness. Seeds were counted using the Germinator plugin [93] implemented in ImageJ v.1.40 [94]. In Moroccan-simulated conditions, we propagated the 64 Moroccan lines in four replicates together with a set of eight representative Cape Verdean lines (four from Santo Antão and four from Fogo) in eight replicates each. To assess fitness differences between populations under CVI and Moroccan-simulated conditions, we collected the complete sets of seeds produced per individual. In the CVI simulated conditions, where total seed numbers were limited, we counted the seeds, and from the Moroccan conditions we weighed seeds and estimated the counts based on the weight of 100 seeds.

6. Population structure, diversity and demographic reconstruction

6.1. Clustering

We evenly subsampled the 13 genetic clusters identified previously on the continents (nine in Eurasia [10], four in Africa [8]) and the two Cape Verdean Islands populations to 20 samples per cluster to avoid biases due to differences in sample size across populations. The only exceptions were the Moroccan Rif, North Middle Atlas and High Atlas populations where fewer samples are available (respectively, 8, 13 and 16). We pruned the data set for short-range linkage disequilibrium <--indep-pairwise 50 10 0.1>, and for missing data <--geno 0> using PLINK v.1.90 and removed multi-allelic variants. We produced neighbour-joining trees using the R package *ape* v.3.5 [96] (https://github.com/HancockLab/CVI).

6.2. Measures of diversity and inference of the joint site frequency spectra (JSFS)

We used custom scripts to estimate nucleotide diversity (θ) in CVI, Morocco and Eurasia by computing Tajima's (θ_π) and Watterson's estimators (θ_w), as well as for deriving the site frequency spectra (SFS) (https://github.com/HancockLab/CVI). The joint site frequency spectrum (JSFS) between islands was computed on a subsampled set of 40 individuals per island. We excluded sites with more than 5% missing data, CpG sites, due to their hypermutable nature, pericentromeric regions, which are rich in satellite repeats, and other repeat regions identified with Heng Li's SNPable approach (http://bit.ly/snpable). The JSFS between CVI versus Morocco was computed using both CVI islands together and was polarized to the outgroup species *Arabidopsis lyrata*. We aligned short-read data for 27

A. lyrata genomes to the *A. thaliana* reference genome (TAIR10) and retained for analyses only SNPs that were not polymorphic in *A. lyrata* and for which there were no missing data. To polarize the JSFS between islands, we reconstructed the most likely ancestral state at every SNP based on variation in Morocco, the best modern representative of the original colonizing lineage. At sites that were fixed in Cape Verde, a state was assigned as ancestral if it was found anywhere in Morocco; otherwise, it was assigned as derived. We used the same approach for sites that were polymorphic in Cape Verde. In cases where both alleles were found in Morocco, a missing value was assigned for the ancestral state.

6.3. Identifying tracts of shared ancestry

We inferred haplotypes across the genome, separated by historical recombination events, and screened a set of potential donor populations for the closest relative at each haplotype using Chromopainter v.0.0.4 [98]. We used a representative subset of 148 CVI genomes from the two islands. As donors, we used the 13 mainland clusters previously identified (nine in Eurasia [26], four in North Africa [23]). Each donor population was randomly subsampled to 20 samples 100 times, and for each subsampling we ran Chromopainter ten times for a total of 1000 replicated analyses of each Cape Verdean genome (https://github.com/HancockLab/CVI).

6.4. Demographic reconstruction

We narrowed down colonization time by obtaining an upper bound based on the minimum coalescence time between CVI and Morocco, and a lower bound based on the maximum coalescence time within the CVI clade.

6.4.1. Inference of split times and colonization times

We inferred split times between the two Cape Verde Islands, among subpopulations within islands and between CVI and Morocco using the cross-coalescence rate (CCR) statistic in the MSMC2 framework [17,18] as well as with dadi v.2.1.0 [30], which derives estimates for parameters based on fitting the JSFS. For both methods, we assumed a generation time of one year and a mutation rate of 7.1×10^{-9} [(99)]. MSMC2-CCR consists of comparing the rate of inferred coalescences between groups to the average rate within groups across time. CCR decays from one towards zero as populations split from each other. For analyses with MSMC2-CCR, we combined the effectively haploid genomes to produce artificial diploids. Diploids were created by combining lines from the same stand to avoid biases due to structure. We used the eight-haplotype implementation of MSMC2, which has the best resolution for recent events (up to approx. 1 kya in our system). For the inference of split parameters in dadi v.2.1.0 [30], we used intergenic JSFS, which are less likely to evolve under strong selection. We estimated parameters between the two

Cape Verde islands and between CVI and Morocco using four demographic models. For each model and population pair, we conducted the analysis 1000 times with up to 50 iterations to infer confidence intervals.

6.4.2. Demography within CVI

We used three complementary approaches to model the demographic history within the archipelago including the timing of colonization and severity of the associated bottlenecks. First, we ran RELATE [32] and COLATE [34] under a haploid model using the module 'EstimatePopulationSize' to reconstruct N_e over time based on inferred coalescence events within each island population. In addition, we fit a model to the data using forward-in-time, individual-based simulations from Slim3 [21]. We also conducted inference based on phylogenetic analysis of the non-recombining chloroplast locus to check for agreement at this locus.

7. Testing for evidence of adaptive evolution

7.1. d_{sel}/d_{neu} and the distributions of fitness effects

We used custom scripts (https://github.com/HancockLab/CVI) to compute the d_{sel}/d_{neu} ratio, defined as the rate ratio of 0-fold non-synonymous to 4-fold synonymous substitutions, scaled by the number of sites at risk for each category. We used the spectra at zero- and four-fold degenerate sites to infer the distribution of fitness effects (DFE) and the proportion of adaptive substitutions (alpha) with polyDfe v.2.0 [36] using default parameters <-m C -o bfgs>. We ran the analysis independently for the two CVI islands (11 samples in Fogo and 13 in Santo Antão), and Morocco. For both analyses, confidence intervals were estimated based on resampling.

8. Identifying QTLs, candidate genes and functional variants

We conducted a literature review of studies that used the Cvi-0 x L*er*-0 RILs and, based on these studies together with fine mapping and downstream functional analyses, we compiled lists of candidate genes and validated functional variants.

9. Trait mapping

9.1. GWAS

We conducted genome-wide association analysis (GWAS) using a univariate linear mixed model while accounting for population structure with a mean-centred kinship matrix <-gk 1> using the flag <-lmm 4> in GEMMA [98]. Input files for this analysis were generated on GATK genotypes, which included indel calls, using VCFtools [99] and PLINK [11]. Mapping was conducted based on the median phenotype across

replicates per genotype (https://github.com/HancockLab), since no block effect was detected across the chamber.

9.2. Bulked segregant analysis

We propagated an inter-island F2 population (S5-10 x F13-8, n=488), in which the ancestral allele *FRI* K232 was fixed), under simulated CVI conditions. Because early flowering segregated at an approximately 1:3 ratio (indicating a single recessive locus), we sampled leaf tissue from the 25% early tail of the F2 (n=108). We extracted DNA using a DNeasy Plant Mini kit (Qiagen), assessed DNA quality and quantity with Qubit and Nanodrop (Thermo Fisher Scientific), prepared a single library using NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs) and sequenced it to 50x coverage using the Illumina HiSeq3000 platform. We called variants against the TAIR10 reference assembly using a GATK pipeline [100] (https://github.com/HancockLab/CVI), retaining only biallelic variants. We identified window(s) where the median allele frequency was greater than 95% and annotated variants within candidate region(s) using SnpEff v.3.0 [101]. These are listed in Supplementary Table 12.

9.3. Functional validation

9.3.1. *FLC* RNA quantification

We measured *FLC* expression in a representative set of eight Cape Verdean and six Moroccan lines as well as in the Col-0 reference line, a modified Col-0 with a functional *FRI* introgressed (Col-0 *FRI*-Sf2, shown as Col-0 $FRI^{+}FLC^{+}$), since *FRI* affects *FLC* mRNA levels [64,65], and Col-0 *FRI*-Sf2 with an *FLC* knock-out (Col-0 *FRI*-Sf2 *flc-3*, shown as Col-0 $FRI^{+}FLC^{-}$) [53]. We grew three replicates of each genotype under CVI simulated conditions (12h light, 20ºC at day, 14ºC at night) and assessed mRNA levels by qRT-PCR on a LightCycler 480 instrument (Roche) using the $2^{-\Delta\Delta Ct}$ method (Applied Biosystems) and PP2A (AT1G13320) as a reference gene. Primers used in this experiment are listed in Supplementary Table 13.

9.3.2. *FLC* complementation test

We performed genetic complementation tests for *FLC* by crossing four individuals from Fogo (each with the *FLC* 3X allele) to Col-0 *FRI*-Sf2 plants with and without a functional *FLC* allele (Col-0 *FRI*-Sf2, referred to as Col-0 $FRI^{+}FLC^{+}$, and Col-0 *FRI*-Sf2 *flc-3* [53], referred to as Col-0 $FRI^{+}FLC^{-}$, respectively). We also crossed the mutants (Col-0 background) to obtain a heterozygous F1 at *FLC*. We grew four replicates of each parent and F1 per cross and scored bolting and flowering time in 12h standard greenhouse conditions.

10. Historical reconstruction of evolution of *FRI* and *FLC* loci

10.1. Reconstructing marginal genealogical trees and inferring selection coefficients

We used RELATE v1.1.4 [32] to infer the genealogical trees for the derived alleles *FRI* 232X (Chr4:269719) and *FLC* 3X (Chr5:3179333) and we used CLUES [52] to infer the frequency trajectory and selection coefficient for the derived *FRI* 232X and *FLC* 3X alleles. Selection coefficients were inferred relative to the reconstructed demographic history for each island (Supplementary Tables 14-15).

10.2. Fit to SSWM and WSSM models of evolution

We calculated the fit to strong selection weak mutation (SSWM) and weak selection strong mutation (WSSM) models of evolution [56–58] using an estimate of the genome-wide mutational target size based on molecular studies [64,77,102–104] and inferences from our population genetic analyses. The logic and details can be found in the Supplementary Note.

10.3. Simulations of adaptation

We conducted forward simulations in SLiM [33] under a Wright-Fisher model based on parameter estimates from the Fogo population to examine the probabilities of fixation of an adaptive variant (i.e., one that abolishes the vernalization requirement for flowering) taking into account the stochastic effects of drift. The selection coefficient (*s*) was set to 0.09273. Each simulation was run for a maximum of 6000 generations but was terminated earlier if a beneficial mutation arose and fixed. Mutation rate was set to $7x10^{-9}$ and the probability of a beneficial mutation was set to match our estimate of $U_b=1.54x10^{-6}$ (Supplementary Note). We used three different plausible estimates for the degree of selfing (90%, 95% and 99%) based on estimates from *Arabidopsis* populations [105] and conducted 200 simulations for each case. From these, we calculated the proportion of runs where populations adapted, the proportions of potentially adaptive variants that are lost or fixed in all runs, and the times to fixation or loss.

11. Statistical analyses

11.1. Comparison of climate variable distributions in Morocco and CVI

Differences in the climate distributions were evaluated using the two-tail Wilcoxon rank sum test/Mann Whitney U test (hereafter MWW test) with the *wilcox.test()* function in R (https://github.com/HancockLab/CVI).

11.2. Testing for evidence of adaptive evolution

11.2.1. d_{sel}/d_{neu} and distribution of fitness effects

We computed the d_{sel}/d_{neu} ratio and the distribution of fitness effects (DFE) with polyDfe v.2.0 [109] for the two CVI island populations and Morocco. To estimate uncertainty around these parameters, we bootstrapped frequency spectra 500 times with polyDfe and calculated an empirical p-value for the d_{sel}/d_{neu} ratio and the discretized DFE categories based on the bootstrapped data. The large variance in the bootstrapped data stems from the low number of variants segregating in CVI.

11.2.2. Testing for population and condition-dependent fitness effects

We tested deme, habitat and deme x habitat interaction effects of Moroccan and CVI lines in the CVI and Moroccan simulated environments. To correct for over-dispersion, we employed a negative binomial transformation using the *glm.nb()* function from the package MASS v.7.3-51.4 in R (https://github.com/HancockLab/CVI).

11.3. Testing for multi-variate adaptation in Santo Antão

11.3.1. Testing the percentage of private variation across functional categories

To compute the proportion of private variants we counted the mutations that distinguish Cvi-0 from L*er*-0 and calculated the proportion which are private to Santo Antão and segregating there. This calculation was repeated for the whole genome, QTL and candidate genes. Because functional variants represent single mutations, in this case each variant was either fixed in CVI and denoted with 0% private, or segregating in Santo Antão and denoted with 100% private. For every functional category, we compared the rate of private variation to the genome wide expectation (419466 variants differentiating Cvi-0 from *Ler*-0, of which 3214 private ones), using a two-tailed Poisson test implemented in R (*poisson.test()*).

11.3.2. Modelling effects of validated functional variants on fitness

To assess the effects of the seven functional variants segregating in Santo Antão on fitness, we used forward-backward stepwise regression (i.e., sequential replacement) approach in a linear model framework using the R package caret v.6.0-86 [106]. Significance of models was assessed based on the root mean squared error (RMSE) by 1000 bootstrap samples. To test whether the explanatory power of the seven functional variants was higher than randomly selected genomic variants, we resampled 2000 sets of seven randomly chosen variants from an LD-pruned genome (PLINK [95] command: <--indep-pairwise 50 10 0.1>) and conducted stepwise regression on each of these sets, exactly as we had done on the seven

functional variants. We obtained an empirical p-value by comparing the observed R^2 to the resampled null distribution (https://github.com/HancockLab).

11.3.3. Testing for differences in flowering time between populations

We tested for differences in the distributions of bolting time between CVI and Moroccan populations using two-tail MWW tests on the medians per genotype with the *wilcox.test()* function in R (https://github.com/HancockLab/CVI). 95% confidence intervals were calculated using function *ci()* implemented in the R package *gmodels* v.2.18.1 [107].

11.4. Testing for evidence of transgressive segregation in F2s

To determine whether there was transgressive segregation in inter-island crosses, we tested each F2 population against their corresponding parental lines. Each parental line was grown in 12 replicates, except for Cvi-0 and F9-2 (4 replicates per lines), and the F2s had 488, 598 and 636, respectively for the crosses S5-10 x F13-8, Cvi-0 x F9-2 and S15-3 x F3-2. We used Dunnett's tests on each individual cross, using the *DunnettTest* function implemented in the R package *DescTools*[110] (https://github.com/HancockLab), and a Fisher's combined p-value test on the set of crosses, using the function *fisher.method* implemented in the R package *metaseqR*[111] (https://github.com/HancockLab).

11.5. GWAS

We conducted likelihood ratio tests in GEMMA [112] to test associations between markers and the median bolting time per natural line. Manhattan plots show p-values $-\log_{10}$ transformed on the y axis.

11.6. *FLC* RNA experiment

We tested the difference in *FLC* expression and bolting time between genotypes with the Kruskal-Wallis method implemented in the R package *agricolae* (https://github.com/HancockLab). We applied the $2^{-\Delta\Delta Ct}$ (Applied Biosystems) on the median across three technical replicates per genotype.

11.7. *FLC* complementation test

We tested phenotypic complementation of F1 hybrids by comparing their phenotypic distributions to parental lines using the *wilcox.test()* function implemented in R (https://github.com/HancockLab), on four replicates of each of the parental lines and eight replicates of each F1 line. We tested for phenotypic complementation of Col-0 background F1 hybrids by comparing their phenotypic distribution to Col-0 *FRI*-Sf2 *flc-3* (*FRI*$^+$*FLC*$^-$) and Col-0 *FRI*-Sf2 (*FRI*$^+$*FLC*$^+$) using the *wilcox.test()* function implemented in R (https://github.com/HancockLab/CVI).

Data and code availability

All data generated in this study are included in this article and its Supplementary Materials files. Raw sequencing reads are available in the European Nucleotide Archive (ENA) under the project accession number PRJEB39079 (ERP122550), including whole genome sequences of 335 *A. thaliana* lines. The genomic variant calls are available in the European Variation Archive (EVA), with project accession number PRJEB44201 (ERZ1886920). All code used in analyses and data visualization is available at https://github.com/HancockLab/CVI.

Acknowledgements

We thank Martin Koornneef, Nick Barton, Christian Brochmann, George Coupland and four anonymous reviewers for valuable discussions and comments, and we thank Wolfram Lobin for sharing herbarium records. Logistical support in the field, field assistance and advice were provided by Natural Parks in Santo Antão and Fogo, Â. Moreno and S. Gomes at the Instituto Nacional de Investigação e Desenvolvimento Agrário (INIDA), Cape Verde, and Arlindo Martins. The project was supported by the Marie Curie CIG 304301, Vienna International Postdoctoral Program for Molecular Life Sciences (VIPS), NSF IRFP (1064766), Max Planck Society Funding, and ERC CVI_ADAPT 638810 to A.M.H., FWF DK W1225-B20 (A.F.), Laboratoire d'Excellence (LABEX) entitled TULIP (ANR-10-LABX-41) to F.R., DFG FOR 1078 to J.H. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. All sample collection was made with appropriate field permits (PERMIT NUMBERS No.12/2012, 01/2015, 112/2018).

Author contributions

Conceptualization: J.H., A.M.H.; Methodology: A.F., C.N., A.M.H; Software: A.F.; Investigation, validation and data curation: C.N., A.F., A.F.E., E.T., M.G., N.W., N.D., A.M.H.; Formal analysis: S.R., A.F., N.W., J.H., S.A., A.F.E., E.T., A.M.H.; Resources: H.D., C.N., E.T., P.J.F., A.F.E., A.F., F.R., A.M.H.; Writing-first draft: C.N., A.F., A.M.H.; Writing-reviewing and editing: all authors; Project administration: A.M.H.; Supervision and funding acquisition: M.K., F.R., J.H. and A.M.H.

Supplementary Methods

1. Sample collection

We collected plants over a series of field expeditions between 2012 and 2019 on Santo Antão and Fogo, the two islands where *Arabidopsis thaliana* had been recorded in herbarium records (personal communication, Wolfram Lobin). In addition, we explored possible locations in the two other islands with the most similar landscape (Santiago and São Nicolau) but found no evidence of *A. thaliana* there, consistent with a lack of herbarium records. In total, we present data for 335 accessions from the Cape Verde Islands (Supplementary Table 1), including 189 accessions from 26 stands across four regions in Santo Antão (Cova de Paúl, Lombo de Figueira, Pico da Cruz and Espongeiro), and 146 accessions from 18 stands across three regions in Fogo (Lava, Monte Velha and Inferno).

2. Climate data

We downloaded gridded data for climatic and bioclimatic variables at 30 second resolution (~ 1 km^2) in GeoTiff file format for the temporal range of 1970-2000 from WorldClim version 2.1 [90], including monthly climate data for average temperature and precipitation (12 data layers, i.e. 1 layer for each month) and 19 bioclimatic variables (1 data layer each) which are temperature and rainfall derived datasets. In addition, we downloaded a raster file for Aridity Index (1 data layer) from CGIAR Consortium for Spatial Information (CGIAR-CSI) [91], which is the ratio of precipitation to potential evapotranspiration, where higher values correspond to more humid conditions and lower to more arid conditions. In addition, we estimated growing season length using the monthly average temperature and precipitation data obtained from WorldClim. Months for which mean temperature ≥ 4°C and mean precipitation ≥ 2 * mean temperature were summed to produce an estimate of the growing season length [108] using 'Raster Calculator' of ArcGIS. We extracted values for sites where CVI, Moroccan and Eurasian samples had been collected in ArcGIS and compared distributions of climate variables across regions using Mann Whitney Wilcoxon (MWW) tests.

3. Sequencing

We sequenced 335 newly collected Cape Verde Islands accessions and Cvi-0 using Illumina Hi-Seq and HiSeq3000 machines. We extracted genomic DNA using DNeasy Plant Mini kits (Qiagen), fragmented using sonication (Covaris S2), and prepared libraries with Illumina TruSeq DNA sample prep kits (Illumina), NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs) and NEBNext Ultra II DNA Library Prep Kit (New England Biolabs). Libraries were immobilized and processed onto a flow cell with cBot (Illumina) and

subsequently sequenced with 2x 100-150 bp paired end reads. We assessed DNA quality and quantity via capillary electrophoresis (TapeStation, Agilent Technologies) and fluorometry (Qubit and Nanodrop, Thermo Fisher Scientific). Due to changes in product availability over time, sample preparation differed slightly between subsets of the sequenced accessions. Sample IDs 12766 to 35519 were prepared with Illumina TruSeq DNA sample prep kits (Illumina, San Diego, CA), samples in projects 4073 and 3968 were prepared with NEBNext Ultra II FS DNA Library Prep Kits (Illumina, New England Biolabs) including four cycles of PCR amplification, and samples from projects 3619, 3541, 3536, and 2876 were prepared with NEBNext Ultra II DNA Library Prep Kit (Illumina, New England Biolabs) with five cycles of PCR amplification.

4. SNP identification and genotyping

We aligned raw Illumina sequence data to the *A. thaliana* TAIR10 reference genome and called variants with three different pipelines. Two pipelines were previously used to call genotypes in the 1135 Eurasian and 64 Moroccan sequences [23]. Here, we used the same parameters, settings, and software versions in order to analyse the Cape Verdean and worldwide sequences in a common framework (https://github.com/HancockLab/CVI) . We used the SHORE pipeline [109] for all analyses except MSMC where we used the more conservative pipeline described in [23]. To call short indels, we used a modified version of the GATK4 [100] best practices workflow for germline short variant discovery and genotyping. We included biallelic variants only and converted heterozygous sites to missing data to mask possible false positives. Average coverage across samples was 19.4x (range from 9.3x to 51.7x; Supplementary Table 1) after alignment to the reference.

For all downstream analyses, we retained variants with coverage greater than 3 and base quality greater than 25 in the SHORE calls, and 2 and 30 in the GATK calls (https://github.com/HancockLab/CVI). To call S-locus haplogroups we followed the procedure used in [23]. We added to the reference genome (TAIR10), which represents haplogroup A, the sequences of S-locus haplogroups B and C (from Cvi-0 [110] and Lz-0 [111], respectively). We called variants against this modified reference and assigned to each CVI sample the S-locus haplogroup that had the highest proportion of sites with non-zero coverage, after quality filtering.

5. Plant growth and phenotyping

5.1. Plant growth and phenotyping: flowering time measurement

In the flowering time experiment, we scored flowering time, bolting time, time to anthesis, number of days until the stem reached 3 cm, and the number of rosette leaves at bolting, as in [92]. We measured correlation between the four flowering traits scored in the simulated CVI conditions experiment

using the function *cor()* implemented in R (https://github.com/HancockLab/CVI). These phenotypes were strongly correlated, with Spearman's rho of at least 0.96 in Santo Antão and Morocco between flowering time and bolting time. In Fogo, where flowering time is more difficult to score due to differences in petal morphology, the correlation was somewhat lower (rho=0.83), likely due to increased error for the flowering time trait here. Therefore, bolting time results were used as a proxy for flowering time in downstream analyses.

6. Linkage disequilibrium

Linkage disequilibrium (LD) was assessed by computing the correlation (r^2) in frequency across pairs of SNPs. The r^2 values between pairs of SNPs were calculated using the command < --ld-window 999 --ld-window-kb 10 --ld-window-r2 0 --r2 --snps-only > in PLINK [95]. Calculations were done between pairs of SNPs to a distance of 10 kb. LD decay analysis were conducted by division of marker pairs within the 10-kb region into bins of 1 kb and r^2 values within each bin were averaged. To visualize the result, the r^2 values were sorted and plotted against the physical distance, using loess smoothing.

7. Demographic reconstruction

7.1. Split times

Using MSMC-CCR [29], we inferred split times by computing the mean across combinations of sets of samples and the confidence interval of the mean (±1.96*standard error of the mean). For the split between Santo Antão and Fogo, we used a total of 63 combinations of eight accessions from each island. For the split to Morocco, we used 357 combinations (separately for the High Atlas, South and North Middle Atlas Moroccan populations). For splits within Santo Antão, we used 12 combinations to examine pairwise splits between the Figueira, Cova, Espongeiro, and Pico da Cruz populations. As suggested in [29], we inferred split times when CCR reached 0.5, with an uncertainty interval between $0.25 \leq CCR \leq 0.75$.

For the inference of split parameters in dadi v.2.1.0 [30], we used joint site frequency spectra (JSFS) based on intergenic SNPs, which are less likely to evolve under strong selection than coding regions. We estimated parameters between the two Cape Verde islands and between CVI and Morocco using four demographic models: 1) a simple two-population split model with no migration and constant population size (N_e); 2) the same model with a bottleneck at the split; 3) a split with exponential changes in N_e after the split and no migration; and 4) an isolation with migration model (IM): a split with exponential changes in N_e and asymmetric migration.

For each demographic model and population pair, we replicated the analysis 1000 times with a maximum of 50 iterations. In each replicate run we used starting values for all parameters drawn

randomly from predefined ranges. The parameter boundaries were (10^{-3}*N_{ref}; 20*N_{ref}) for effective population sizes (N_e), (0; 20/N_{ref}) for migration rate, and (0; 10*N_{ref}) for the split time, where N_{ref} is the size of the ancestral population. Among the 1000 runs per model, we selected the parameter combination that resulted in the highest likelihood. We identified the model with the best support using the Akaike information criterion (AIC), and for each resulting best model, we calculated confidence intervals for parameters using 100 000 bootstrapped data sets and the Godambe Information Matrix implemented in dadi. For the best-supported models, the 5% runs with highest likelihood all converged to the same parameter set.

7.2. Colonization time

We narrowed down colonization time by obtaining an upper bound based on the minimum coalescence time between CVI and Morocco, and a lower bound based on the maximum coalescence time within the CVI clade. First, we ran coalescent simulations of the CVI-Moroccan split in msprime v.0.4.0 [31] with split times drawn from a uniform distribution of times between 5-50 kya. To account for the confounding effect of purifying selection, which reduces the rate at which new mutations are introduced in the genome, we scaled mutation rate across simulated genomic windows as $\mu_{scaled}=\theta_{local}/4*N_e$, where θ_{local} was estimated as θ_π in the Moroccan population within each window and N_e was fixed to the genome-wide average ($N_e=\theta_{genome}/4*\mu$) so that $\mu_{scaled}=\mu*(\theta_{local}/\theta_{genome})$. Then, we inferred coalescence times between simulated and observed CVI and Moroccan genomes across genomic windows based on the density of mutations. We obtained 95% confidence intervals based on the standard error (SE) estimated by non-parametric bootstrap resampling of observed and simulated data. By fitting the simulated cumulative proportion of genomic windows with different inferred ages to observed data, we obtained a conservative estimate of the upper bound of colonization time. We inferred coalescence times within Cape Verde, across genomic windows (0.1 Mbp, non-overlapping), based on the density of mutations and used the 95th percentile as a lower bound for colonization time.

7.3. Chloroplast phylogenetic analysis

We constructed a time-calibrated chloroplast phylogeny to examine divergence in the chloroplast genome and to compare these to patterns at nuclear loci. First, we aligned the chloroplast sequences with outgroups from other *Arabidopsis* species, *Capsella grandiflora*, *Capsella bursa-pastoris* and *Camelina sativa* [112], excluding the Inverted Repeat region. All indels were removed. Identical sequences were excluded from the alignment and a maximum likelihood (ML) phylogenetic tree was reconstructed with RAxML v.8.1.16 [113] using the GTR+Γ+I model of rate heterogeneity and setting the clade of *Capsella* and

Camelina as outgroup. Rapid bootstrapping followed by a thorough ML search was applied with 1000 bootstrap replicates.

Divergence time was estimated using BEAST v.1.8.3 [114]. Three secondary calibration points were included from the literature [115]: the root height (split between genus *Arabidopsis* and the *Capsella*/*Camelina* clade) was set to 8.1627 million years (my), the split between *Capsella* and *Camelina* was set to 7.3572 my, and the crown age of genus *Arabidopsis* was set to 5.9685 my; a standard deviation of 1.0 was used for all three calibration points. The GTR+Γ+I model of rate heterogeneity with 4 Gamma categories was used as substitution model with an uncorrelated relaxed lognormal clock [116] and tree prior Speciation: Birth-Death Process [117]. Two independent MCMC runs with chain length $1x10^9$ were combined in LogCombiner v.1.8.3 [114], discarding the first 10% of each run as burn-in, and the median heights from the remaining 18002 trees were annotated onto the maximum clade credibility tree in TreeAnnotator v.1.8.3 [114].

7.4. Additional inference of demography within CVI

We used forward, individual-based simulations in SLiM v.3.3.2 [33] to model the demographic history within the archipelago, including the colonization events and consequent bottlenecks. Under this model, the initial propagule founded a population on one island in the archipelago. The population grew following an exponential function that varies across simulations (growth rate varies with final size between 100 and 2500 individuals) until the time of the split between islands (4.0 kya), when the second island was colonized from the first. Both populations grew exponentially until they reached final N_e (10K, inferred from θ_π). On the second island, as inferred with dadi[30], we simulated a 1000 year-long bottleneck with 400 individuals.

In order to determine which island was colonized first, we fit simulations to the observed data using the difference in the proportion of variants that are fixed in one island and segregating in the other (proportion of variants segregating in Santo Antão and fixed in Fogo minus the proportion of variants segregating in Fogo and fixed in Santo Antão) as a summary statistic. The value of this statistic is positive if Santo Antão was colonized first and negative if Fogo was colonized first. In addition, we used three-way (Morocco, Santo Antão, Fogo) JSFS modelling in dadi to compare the relative fit of Santo Antão-first and Fogo-first models and estimated the length of the Fogo bottleneck period [30]. For each demographic model we replicated the analysis 100 times with a maximum of 50 iterations. In each replicated run, we used starting values for all parameters drawn randomly from predefined ranges. The parameter boundaries were (10^{-3}*N_{ref}; 20*N_{ref}) for effective population sizes (N_e), and (0; 10*N_{ref}) for split times, where N_{ref} is

the size of the ancestral population. We identified the model with the best support using the Akaike information criterion (AIC).

8. Niche modeling

We used niche modeling in MaxEnt [35] to predict the suitability across the Cape Verde archipelago for colonization by *A. thaliana* from the Moroccan range, and to identify the regions within Cape Verde that are most similar to the Moroccan habitat. In the first step, we produced a predictive model using collection locations in Morocco and the bioclimatic variables described above and listed in Supplementary Table 2. We considered supplementing the collection locations with information from herbarium collection records from GBIF (https://www.gbif.org), but only 'fuzzy matches' existed in the data base and GPS coordinates were thus unreliable. We conducted climatic niche modelling in Morocco using occurrence data based on [89]. Data for the Moroccan region was extracted using the 'extract by mask' function in ArcToolbox. To avoid overfitting, climatic variables were pruned so that no two variables were correlated with Pearson correlation coefficient > 0.75. The model we present uses a set of variables chosen based on ecological and biological relevance, but the predicted suitability of CVI habitat for Moroccan accessions did not change across these different variable selection regimes. We ran a Maxent under the standard default parameters with jackknife resampling to estimate the importance of each variable on the model. Model fit was inferred based on the area under the curve (AUC) for the model output and a cross-validation approach in which the data were split into equal sized subsets. Then, we projected this model onto the CVI landscape to predict the suitable range of Moroccan samples in CVI. We further identified the regions within CVI that were most similar to the Moroccan *A. thaliana* habitat. Since the approach we used for pruning variables is somewhat subjective and the CVI suitability result was extreme, we also tried other approaches for pruning correlated variables but found no change in suitability in CVI across variable selection regimes including random selection of variables with Pearson's $r < 0.75$ and a variance inflation factor approach. This robustness is likely due to the fact that nearly all climate variable values for CVI lie outside those found in the Moroccan presence data.

9. Testing for evidence of adaptive evolution

9.1. d_{sel}/d_{neu}, inference of DFE and the proportion of adaptive variation

We used custom scripts (https://github.com/HancockLab/CVI) to compute the d_{sel}/d_{neu} ratio, defined as the rate ratio of 0-fold non-synonymous to 4-fold synonymous substitutions, scaled by the number of sites at risk for each category:

$$\frac{dsel}{dneu} = \frac{\frac{number\ of\ 0-fold\ substitutions}{number\ of\ sites\ at\ risk\ for\ 0-fold\ substitutions}}{\frac{number\ of\ 4-fold\ substitutions}{number\ of\ sites\ at\ risk\ for\ 4-fold\ substitutions}}$$

For this, we first constructed an artificial variant call format (VCF) file with all possible variants at all sites in the genome and annotated them with SnpEff v.3.0.7 [101]. Then we used a custom script for the calculation of JSFS (https://github.com/HancockLab/CVI) to compute the number of zero- and four-fold degenerate substitutions, as proxies for selected and neutral sites, respectively. Note that this is a simplified approach to estimating dN/dS, which excludes 2- and 3-fold degenerate sites. We used this approach because estimating the expected changes at these classes of sites is problematic due to asymmetries in substitution rates. We scaled these to the number of sites in the genome at risk for each substitution type and deducted the positions with more than 5% missing data. For continental clades, the spectra were polarized to *A. lyrata* samples. Due to the long divergence time and genomic rearrangements between species, the alignment of *A. lyrata* to *A. thaliana* reduced the number of bases for the analyses to 70676280. For the CVI populations, we defined substitutions as variants derived in comparison to Morocco, fixed in one island and absent from the other. To estimate uncertainty, we bootstrapped frequency spectra 500 times in polyDfe v.2.0 [36] and calculated empirical p-values based on the bootstrapped data. The large variance in the bootstrapped data stems from the low number of total variants fixed in the two island populations. If the number of four-fold substitutions was zero (in real or bootstrapped data), we conservatively added one to avoid dividing by zero. For the Moroccan clade, we used *Arabidopsis lyrata* samples as an outgroup. We used the spectra at zero- and four-fold degenerate sites to infer the distribution of fitness effects (DFE) and the proportion of adaptive substitutions (alpha) with polyDfe v.2.0 [36] using default parameters <-m C -o bfgs>. We ran the analysis independently for the two CVI islands (11 samples in Fogo and 13 in Santo Antão), and the four Moroccan clusters.

10. Evidence for ongoing multi-variate adaptation in Santo Antão

10.1. Identifying QTL, candidate genes and functional variants

Since its collection 37 years ago [22], a single plant from CVI (Cvi-0) was studied extensively. Many mapping studies have used recombinant inbred line (RIL) populations (Cvi-0 x L*er*-0 and Cvi-0 x Col-0) [40,118], and near inbred introgression lines (NILs) of Cvi-0 into the L*er*-0 genome [119]. The island of origin of Cvi-0 was previously unknown, but we found it clusters tightly with the Espongeiro population in Santo Antão, indicating that it was collected in this region (Supplementary Fig. 1). We conducted a literature review of studies that used the Cvi-0 x L*er*-0 RILs. We identified 47 QTL-mapping studies (Fig. 6a) that mapped 129 traits that we grouped into 23 major trait-categories. These studies localized 717 QTL intervals. Based on

these studies and follow-up fine mapping, we compiled a set of 135 candidate genes [40–48,92,118–146,146–180]. In eleven cases, the actual mutation responsible for an effect on phenotype was found and validated (by complementation tests, transgenics, sequence analyses). These variants include two large deletions, three small indels (frameshifts) and six SNPs (non-synonymous amino acid changes and truncating variants). To genotype large deletions in the natural population, we computed average coverage across 100 bp windows overlapping the deletions and flanking regions. The phenotypes affected by the functional variants range from flowering time and light signalling (*FRI* K232X [42], *CRY2* V367M [41], *GI* L718F [43,44]), circadian clock regulation (*ZTL* P35T [44]), stomatal aperture (*MPK12* G53R [46]), freezing tolerance (*CBF2* promoter deletion [181]), pathogen resistance (cPGK2 S78G [182] and RPM1 whole gene deletion [162]), chloroplast morphology (FtsZ2-2 G441fs [47]), fructose signalling (*ANAC089* S224fs [48]), and innate immunity (*FLS2* N452fs [45]). In one case, a functional variant responsible for copper detoxification was identified (*HMA5* N923T [141]), but it likely arose in Cvi-0 in the laboratory, since it is completely absent from the sampled natural population.

10.2. Modelling effects of validated functional variants on fitness

To assess the effects of the seven functional variants segregating in Santo Antão on fitness, we used forward-backward stepwise regression (i.e., sequential replacement) approach in a linear model framework using the R package caret v.6.0-86 [106]. For the forward case, we started with a model with no predictors (only an intercept), iteratively added functional variant predictors, and stopped when the improvement was no longer statistically significant based on the change in root mean squared error (RMSE). For the backward case, modelling started with the full model (intercept plus all functional variants), iteratively removed the predictors that contributed the least, and stopped when all predictors were statistically significant. Significance of models was assessed based on the root mean squared error (RMSE), by bootstrap resampling (1000 times).

To test whether the explanatory power of the seven functional variants was higher than randomly selected genomic variants, we resampled 2000 sets of seven randomly chosen variants from an LD-pruned genome (PLINK [95] command: <--indep-pairwise 50 10 0.1>) and conducted stepwise regression on each of these sets, exactly as we had done on the seven functional variants. We calculated the model R^2 to produce a null distribution and obtained an empirical p-value by comparing the observed R^2 value to this using the formula: $(1 + \text{sum}(s \geq s_0)) / (N + 1)$, where s is the R^2 value per draw, s_0 the observed R^2 value, and N the number of draws (https://github.com/HancockLab).

11. Trait mapping

11.1. Flowering time segregation of inter-island F2 populations

We generated three inter-island F2 populations (S5-10 x F13-8 (n=488), S15-3 x F3-2 (n=636), and Cvi-0 x F9-2 (n=598)). These were grown in Bronson climatic growth chambers, with settings to match CVI conditions: 20ºC during the day and 14ºC at night, with a 12h photoperiod and 70% humidity. We scored bolting and flowering time in all F2 individuals and 12 replicates per parental line, except for Cvi-0 and F9-2, for which only four replicates were grown.

To determine whether there was transgressive segregation in each of these populations against their corresponding parental lines, we used the *DunnettTest()* function implemented in the R package *DescTools* v.0.99.37 [183]. We used Fisher's method [184] to calculate a combined p-value across the set of crosses, using the function *fisher.method()* implemented in the R package *metaseqR* v.1.26.0 [185] (https://github.com/HancockLab/CVI).

11.2. Bulked segregant analysis

We propagated an inter-island F2 population (S5-10 x F13-8, n=488), in which the ancestral allele *FRI* K232 was fixed), under simulated CVI conditions. Because early flowering segregated at approximately a 1:3 ratio (indicating a single recessive locus), we sampled leaf tissue from the 25% early tail of the F2 (n=108). We extracted DNA using a DNeasy Plant Mini kit (Qiagen), assessed DNA quality and quantity with Qubit and Nanodrop (Thermo Fisher Scientific), prepared a single library using NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs) and sequenced it to 50x coverage using the Illumina HiSeq3000 platform. We called variants against the TAIR10 reference assembly using a GATK pipeline[100] (https://github.com/HancockLab/CVI), retaining only biallelic variants. We identified window(s) where the median allele frequency was greater than 95% and annotated variants within candidate region(s) using SnpEff v.3.0 [101].

12. FLC RNA quantification

We measured *FLC* expression in a representative set of eight Cape Verdean, and six Moroccan accessions. We also measured *FLC* expression in the Col-0 reference strain, as well as a modified Col-0 with a functional *FRI* introgressed (Col-0 *FRI*-Sf2, shown as Col-0 $FRI^{+}FLC^{+}$), since *FRI* affects *FLC* mRNA levels [64,65], and Col-0 *FRI*-Sf2 with an *FLC* knock-out (Col-0 *FRI*-Sf2 *flc-3*, shown as Col-0 $FRI^{+}FLC^{-}$) [53]. We grew three replicates of each genotype under CVI simulated conditions (12h light, 20ºC day, 14ºC night). We collected and immediately froze 2-3 rosette leaves each from 2-week-old plants and ground them with the TissueLyser II (Qiagen). We extracted RNA with TRIzol (Invitrogen) and treated 2µg with the DNA-

free DNA Removal Kit (Invitrogen) for 1h at 37ºC. We generated cDNA using the Superscript II reverse transcriptase (Invitrogen) together with oligo(dT) primer for 2h at 42ºC. We assessed mRNA levels by qRT-PCR on a LightCycler 480 instrument (Roche) with the EvaGreen dye (Biotium) using the $2^{-\Delta\Delta Ct}$ method (Applied Biosystems) and PP2A (AT1G13320) as a reference gene. Primers used in this experiment are listed in Supplementary Table 13. Differences in *FLC* expression between genotypes were tested with the Kruskal-Wallis method implemented in the R package *agricolae* v.1.3-2 [186] (https://github.com/HancockLab/CVI).

13. FLC complementation test

We performed genetic complementation tests for *FLC* by crossing four individuals from Fogo (each with the *FLC* 3X allele) to Col-0 *FRI*-Sf2 plants with and without a functional *FLC* allele (Col-0 *FRI*-Sf2, referred to as Col-0 $FRI^{+}FLC^{+}$, and Col-0 *FRI*-Sf2 *flc-3* [53], referred to as Col-0 $FRI^{+}FLC^{-}$, respectively). We also crossed the mutants (Col-0 background) to obtain a heterozygous F1 at *FLC*. We grew four replicates of each parent and F1 per cross and scored bolting and flowering time in 12h standard greenhouse conditions.

14. Historical reconstruction of evolution of *FRI* and *FLC* loci and fit to models of adaptation

14.1. Producing marginal genealogical trees with RELATE

We used RELATE v1.1.4 [32] to infer the genealogical trees for the derived alleles *FRI* 232X (Chr4:269719) and *FLC* 3X (Chr5:3179333). We used bcftools v1.9 [187] to filter the VCF file for quality, removed non-biallelic SNPs, retained segregating sites, and filtered out missing data with the following execution <bcftools view -m2 -M2 -v snps –min-ac=1 -i 'MIN(FMT/DP)>3 & MIN(FMT/GQ)>25 & F_MISSING=0'>. For *FLC* 3X, because the derived allele is fixed in Fogo, we included S1-1 from Lombo de Figueira, Santo Antão, as the outgroup. Within RELATE, we used the command RelateFileFormats (using --mode ConvertFromVcf) to convert the VCF file into haplotype and sample files. To infer the genome-wide genealogies, we first ran the command Relate (using –mode All) per chromosome and defined parameters of the recombination map (--map), mutation rate (--m), and the coalescence file (--coal) for N_e over time. The estimated mutation rate ($7x10^{-9}$) for *A. thaliana* [99] was corrected for the percentage of missing data for each region and -m set to $2.512x10^{-9}$ for *FRI* 232X and $2.1x10^{-9}$ for *FLC* 3X. We used a published recombination map [188] corrected for the estimated outcrossing rate of 5% estimated in natural populations [105] by dividing the genetic distances by 20. We then used the output to estimate coalescence rates using the script EstimatePopulationSize.sh across all 5 chromosomes with generation time set to

one year, running the algorithm for 10 iterations. To obtain 95% confidence intervals for RELATE-inferred coalescence rates, we used genome-wide genealogies and coalescence rates as inputs into the COLATE package [34] (https://github.com/leospeidel/Colate), which uses a block bootstrap (100x) over genomic regions.

To infer the local genealogies for *FRI* 232X and *FLC* 3X, we ran RELATE and used the genome-wide coalescence rates (--coal) inferred previously. To produce genealogical trees for *FRI* 232X and *FLC* 3X variants with confidence intervals for the estimated ages based on 200 samples from the MCMC (derived using SampleBranchLengths.sh --format a, and using default settings), we used the script TreeViewSample.sh, with 10*N steps (N is the number of haplotypes) and 1000 burn-in iterations.

14.2. Inference of allele frequency trajectories and selection with CLUES

We used CLUES [52] to infer the frequency trajectory and selection coefficient for the derived *FRI* 232X (Chr4:269719) and *FLC* 3X (Chr5: 3179333) alleles. CLUES uses importance sampling over trees generated in RELATE to produce a posterior distribution of trees from which a frequency trajectory can be inferred. From the output from RELATE, we used the command <./SampleBranchLengths.sh --format b> to obtain 200 samples from the MCMC. For *FRI* 232X, we integrated a pseudo-ancestor individual (from our inferred CVI ancestral states) as an outgroup for the Santo Antão population. Then, we inferred genome-wide and local genealogies and conducted importance sampling in RELATE after adjusting mutation rate (-m) to 3.24×10^{-10} and 4.193×10^{-10} based on the proportion of missing data removed for the *FRI* and *FLC* regions, respectively. We obtained estimates of the posterior distributions of allele frequencies over time using a recessive model (--dom 0): <inference.py --popFreq 0.7513 --tCutoff 5000 --coal relate.popsize.coal --sMax 1 --df 100 --dom 0 > for *FRI* 232X, and <inference.py --popFreq 0.99 --tCutoff 7000 --coal relate.popsize.coal --sMax 1 --df 100 --dom 0 > for *FLC* 3X (https://github.com/HancockLab/CVI). As in other analyses, we assumed one generation per year. We inferred selection coefficients jointly across two-time bins (epochs) for *FRI* 232X (0-2 and 2-4 kya) and three-time bins for *FLC* 3X (0-2, 2-4 and 4-6 kya) between the present day and the time in the past when the variants arose.

14.3. Fit to SSWM and WSSM models of evolution

We calculated the fit to strong selection weak mutation (SSWM) and weak selection strong mutation (WSSM) models of evolution [56–58] using an estimate of the genome-wide mutational target size based on molecular studies [64,77,102–104] and inferences from our population genetic analyses. The logic and details can be found in the Supplementary Note.

14.4. Simulations of adaptation

We conducted forward simulations in SLiM [33] under a Wright-Fisher model based on parameter estimates from the Fogo population to examine the probabilities of fixation of an adaptive variant (i.e., one that abolishes the vernalization requirement for flowering) taking into account the stochastic effects of drift. The (constant) population size was set to N=48 based on the estimate from RELATE and the selection coefficient was set to s = 0.09273 based on the estimate from CLUES under a model where the reconstructed N_e was used in RELATE/CLUES. The final number of generations depended on when the variant fixed with a maximum of 6000 generations. We simulated two genomic elements: one of size 1.5 Mbp where neutral mutations could arise and another of 1 bp where the selected variant could arise. We used three different plausible estimates for the degree of selfing (90%, 95% and 99%) based on estimates from natural population [105] and conducted 200 simulations for each case. From these, we calculated the proportion of runs where populations adapted, the proportions of potentially adaptive variants that are lost or fixed in all runs, and the times to fixation or loss.

Supplementary Results

1. Population history reconstruction

1.1. Identifying the outgroup to the CVI populations

We used Chromopainter [98] to identify the closest ancestor to CVI across the genome by matching haplotypes to sequenced African and European individuals [23,26]. We found that the Moroccan High Atlas population was the closest relative for the majority of the genome (approx. 61%), followed by the North Middle Atlas population (approx. 7%) and then other Moroccan and European populations (Supplementary Fig. 4). Some of the variance visible in Supplementary Fig. 4 matching across populations may be due to the lack of a strong (close) match, so that multiple nearly equivalent distant matches can be often found. We next examined two specific large-scale loci (the chloroplast and the S-locus), where interpretation of ancestral sharing may be simpler due to the very low probability of recombination at these loci. At the chloroplast, we found that Cvi-0 clusters most closely with individuals from the South Middle Atlas population (Supplementary Fig. 4-5).

The S-locus is a well-characterized region responsible for self-incompatibility in *A. thaliana*. In the species, three deeply diverged haplogroups segregate (A, B, C) as well as an ancient recombinant (A/C) haplotype [189–191]. Due to the deep divergence between the major haplogroups, recombination between these is suppressed and thus exceedingly rare in this region [192] so that matching to major haplogroups is clear. We classified S-locus haplogroups in all CVI samples following [23] and found that they all carried haplogroup B (Supplementary Fig. 4). In the continental sample, this haplogroup is present only in three samples and only in the northernmost Moroccan population in the Rif Mountains [23,193].

Taken together, these patterns show that while Morocco is the continental population genetically closest to CVI, there is no single Moroccan sample or population that is consistently closest to CVI across the genome. Instead, our findings suggest that CVI was colonized by a 'ghost' population that is not represented well by any modern-day sampled population.

Previously, based on the timing of coalescence events, we inferred that Moroccan populations expanded and contracted over time [23]. This could have led to loss and/or re-sorting of lineages among populations and could help to explain why we were unable to identify a single best representative of the colonizing population [23].

1.2. Estimating CVI colonization time

We narrowed down the CVI colonization time by obtaining an upper bound based on the minimum coalescence time between CVI and Morocco and a lower bound based on the maximum coalescence time within the CVI clade.

1.2.1. Estimating the upper bound of colonization time using haplotype coalescences between Morocco and CVI

To estimate the split time between Moroccan and CVI populations, we calculated the relative ratio of between-group coalescence events to within-group coalescences (i.e., the cross-coalescence rate (CCR) statistic) implemented in MSMC [29,194]. Given that we are unable to identify the closest Moroccan population, we expected this analysis to overestimate the divergence time from the actual continental ancestor. The CVI population exhibited initial divergence from all present-day Moroccan populations at approximately 40-60 kya (Supplementary Fig. 7), which we interpret to represent the split time between the present-day Moroccan population and the 'ghost' ancestor of the CVI populations.

CCR between CVI and High Atlas shows a somewhat different pattern compared to other Moroccan populations. The trajectory of the statistic does not monotonically decay as would be expected under a simple split model but rather inflects and reaches a local maximum between 10 and 20 kya (Supplementary Fig. 7). As a result, the 0.25 - 0.75 CCR quantiles for the CVI-High Atlas split are consistent with a wide range of split times from 60 kya until as recently as 10 kya. This inflection could potentially be explained by a complex relationship between the ancestors of the Moroccan and CVI lineages, i.e., secondary contact between the ancestors of the 'ghost' and the High Atlas populations. This resulted in the presence of some haplotypes across the High Atlas genomes that represent the population that originally colonized CVI.

1.2.2. Estimating the lower bound of colonization time based on haplotype coalescences within CVI

Although we were unable to identify a close outgroup population to CVI, we can use information about the coalescences within CVI to obtain a lower bound on the colonization time. We examined historical coalescence events in Santo Antão, Fogo and between the two islands. Coalescence rates spike around 10 kya (Supplementary Fig. 7) and decline sharply in Santo Antão starting at approximately 7 kya, when we infer population structure begins to develop within this island (Supplementary Fig. 8e). In agreement with this time estimate, and based on the density of mutations, 95% of genomic windows between islands (0.1 Mb, non-overlapping) coalesced by 7.1 kya (Supplementary Fig. 8). Then, at

approximately 4-5 kya, there is a strong signature of reduced coalescence between islands, consistent with a split at this time (Supplementary Fig. 8e).

The long gap between the coalescence of lineages within CVI (5-7 kya) and the coalescence times between present-day Moroccan and CVI populations (40-60 kya) is consistent with a model in which a now extinct or unsampled 'ghost' population was the actual founding population of CVI. Based on our analyses, we hypothesize that this population split from the present-day Moroccan population approximately 35-50 kya (Supplementary Fig. 7) and that island colonization likely occurred approximately 7 kya, when population structure becomes apparent in CVI (Supplementary Fig. 7).

1.2.3. Using the distribution of haplotype ages to test for secondary contact between the ancestors of the Moroccan and 'ghost' populations

To further explore the colonization dynamics and to assess evidence for colonization by a 'ghost' population in different time frames, we compared the distribution of ages of genomic windows (haplotypes) to those from simulations of a 'ghost'-CVI split at different time points (5, 10, 20, 30, 40 and 50 kya). These simulations were conducted under the assumption that variable mutation rate and purifying selection reduce diversity by the same extent as divergence (the same rationale as in the HKA test). In that way, we were able to capture the genomic variance in the combined $N_e\mu$ parameter by scaling the simulations to diversity in Morocco across genomic regions. This approach was based on the logic that the inflection in the MSMC-CCR plot (Supplementary Fig. 7) could be due to secondary contact between the 'ghost' population and the Moroccan population (i.e., gene flow back into the Moroccan population from the 'ghost'). In this case, we could use the distribution of window ages (inferred based on the density of SNP variants) to estimate the timing of the secondary contact event. This timing inference would better represent the split between CVI and the 'ghost', although it is still likely to be an overestimate. We found that the cumulative tail of recent coalescence times in observed data fits best with a 10 ky old split and upper bound of colonization time (Supplementary Fig. 7).

1.2.4. Estimating the CVI colonization time by fitting demographic models to the joint site frequency spectrum (JSFS)

The site frequency spectrum provides complementary (largely independent) information from signatures of haplotype coalescences. We ran dadi to infer the split time between Morocco and the ancestor of the CVI population. We used five demographic models including 1) a simple split model, 2) a model that included an exponential population size change in CVI after the split, 3) an isolation-with-migration model, 4) a model that included a bottleneck in CVI after the split, followed by an instantaneous

size change, and 5) a model that included bottlenecks in both the CVI and Moroccan populations after the split and a subsequent size change in CVI. Details of the model parameters are shown in Supplementary Fig. 7. The best performing model (based on AIC) is the two-bottleneck model. This model includes a bottleneck in CVI after the split, followed by an instantaneous population increase (Supplementary Fig. 7). Similar to the haplotype coalescence analysis, the parameter estimates under this JSFS-based model capture the signal of an early split between the ancestor of the CVI colonist ('ghost' population) and the ancestor of the current Moroccan population, placing this split time at 49.7 kya, with a 43.5 kya bottleneck. This scenario fits well with the inferences from haplotype coalescence times (Supplementary Fig. 7) where the long-term effective population size of the 'ghost' population was small in the interval between the split of the parental populations and expansion within CVI. In the JSFS-based model, the time when the CVI population is inferred to increase in size (6 kya) is consistent with an expansion beginning sometime after the colonization of the islands (Supplementary Fig. 7d-e). Neither the haplotype coalescence approach nor the JSFS approach between Morocco and CVI can reveal specific information about the propagule size to CVI due to confounding with the 'ghost' population. However, the very low number of shared variants between present-day Moroccan populations and CVI (0.1%) suggests that the colonizing population was depleted of variation and that current trait variation in the islands occurred via new variants.

1.3. Colonization of Fogo from Santo Antão

1.3.1. Order of island colonization

Isolation by distance leads to a reduction in variation with distance from the starting population and to a pattern in which a proportion of derived variants that segregate in one (parent) population will be fixed in the child population [195,196]. Therefore, we focused on the subset of variants that are segregating in one island and fixed derived in the other to infer colonization order. First, we examined our power for this approach using simulations. Here, we found that as long as the time between the colonization of the first and second island, or the size of the island colonized first is large enough (larger than approximately 500 individuals at the split), the island colonized first will have a lower proportion of fixed mutations that segregate in the other island, independent of the number of colonizers. In the observed data, the Fogo population has a higher proportion of fixed mutations that segregate in Santo Antão compared to the converse (a positive statistic in Supplementary Fig. 8d), supporting initial colonization of Santo Antão, followed by Fogo (from Santo Antão). When we fit the statistic estimated from data to forward simulations, we inferred that population size in Santo Antão grew quite slowly from colonization until the split from Fogo, with a final size of only about 1000 plants when Fogo was colonized.

Further, we used dadi [30] to fit three-population models to the observed join site frequency spectra. The models were simple 3-populations splits with constant population sizes, but they differed in which island was colonized first. In the first model, Santo Antão was colonized from Morocco, and later Fogo was colonized from Santo Antão. In the second model, Fogo was colonized from Morocco, and later Santo Antão was colonized from Fogo. The model that best fit observed data (lowest AIC, 612 vs 628; highest likelihood, -301.3 vs -309.1) was the one in which Santo Antão was colonized first from Morocco, and Fogo was colonized later from Santo Antão.

1.3.2. Dynamics of island colonization (split time and initial population size in Fogo)

Next we inferred the split time between islands and other aspects of historical population dynamics using the JSFS (dadi [30]). We estimated parameters under a range of models (simple split, exponential, isolation with migration, bottleneck) and compared model AICs to identify the best fitting model of the historical dynamics of the Fogo population. We found that the best model was one in which the island colonization event was accompanied by a bottleneck lasting 930 years. The population size (N_e) during the bottleneck period was approximately 400 individuals (Supplementary Fig. 8a-b). We did not attempt to estimate the number of founders separately from the bottleneck size because diffusion models generally do not perform well, such as dadi, generally does not perform well to infer the N_e before a bottleneck [30].

The estimated split time is in rough agreement with a simpler estimate based on the distribution of pairwise differences across windows within and between islands (Supplementary Fig. 8c). We calculated the mean pairwise differences among CVI individuals across 100 kb windows of the genome and found that coalescence time between islands was centred around 4.5 kya (mean: 4.6 kya median: 4.5 kya) and that 95% of the windows coalesced by 7.1 kya. The complete distribution of pairwise differences is shown in Supplementary Fig. 8c. Variation within Santo Antão was centred around 3.0 kya (mean: 3.0 kya; median: 3.0 kya) with 95% of windows coalescing by 4.8 kya and in Fogo around 2.4 kya (mean: 2.5; median: 2.4) with 95% of windows coalescing by 3.9 kya.

The bottleneck duration (T_s) agrees with a simple calculation based on the proportion of fixed variants in Fogo. The Fogo population carries 135 fixed derived mutations out of a total of 23 Mbp of sequence. Using T_s = number of fixed mutations/(μ*L), we can estimate about 840 generations, in which Fogo remained a single connected population (about 840 years) after colonization before structure built up preventing mutations from fixing in the island. The same calculation based on intergenic sites only,

results in a very similar bottleneck duration (36 fixed derived intergenic mutations out of 4.8 Mbp: about 1070 years).

The complete lack of population structure for approximately 870-1070 years for the Fogo population is striking. Seed-dispersed plant populations tend to be highly structured [197] including *A. thaliana* populations [198–201]. However, we find no evidence of structure accumulating in the long post-colonization bottleneck period in Fogo. This implies that the nascent population was poorly adapted to its new environment and therefore severely limited in size during this waiting period until one or more necessary mutations arose that increased fitness and allowed the population to expand.

1.4. The probability of secondary migration events after initial colonization is low

Allowing for migration (gene flow) between the Moroccan and 'ghost' populations after the split led to a poor model fit (based on AIC; Supplementary Fig. 7), implying a lack of migration after the split. This is not surprising given that multiple independent migration events into CVI would likely lead to much higher genetic variation than we observe in the archipelago. The average pairwise differences between two Moroccan individuals is 82.4-fold higher relative to the average pairwise differences within CVI (θ_π (Morocco) = 5.38×10^{-3}; θ_π (CVI) = 6.53×10^{-5}). Even within a single Moroccan region, the average pairwise differences is 54.2-fold higher than in CVI (on average θ_π (Moroccan regions) = 3.54×10^{-3}; Supplementary Table 3). Further, we observed extremely low sharing of variation between CVI and Morocco (0.1%). Based on this, it is difficult to imagine a scenario in which multiple independent dispersal events could have contributed to the present CVI populations.

Similarly, given the almost complete lack of shared variants genome-wide between Santo Antão and Fogo (0.6%), it is highly unlikely there was any subsequent migration after the initial colonization of Fogo. In further support of this assertion, a single chloroplast haplotype is fixed in CVI, with only 18 variants segregating there. Similar to clustering from genomic variation, a chloroplast network clearly separates the two islands (Supplementary Fig. 6). Consistent with the lack of secondary migration between islands, demographic inference with dadi finds the best support for a model with no migration (Supplementary Fig. 8d-e). Further, given the inferred small initial population size in Fogo, a secondary migration event from Santo Antão would likely result in much higher shared variation between islands than observed.

2. Population history within islands

Trajectories of coalescence rates within Santo Antão and Fogo (MSMC) as well as matching to patterns of polymorphism in simulations (Supplementary Fig. 8d) imply that Santo Antão was colonized

first and Fogo second from Santo Antão. Trajectories of coalescence rates over time for individual Santo Antão populations and inferred split times with MSMC-CCR show that the Cova de Paúl population best represents the early colonists (Supplementary Fig. 8e). Based on CCR analysis, we estimated that Cova de Paúl split from Lombo de Figueira at approximately 7 kya. At this time, the coalescence (MSMC) trajectory for Santo Antão enters a period of intense reduction (Supplementary Fig. 7b), which likely corresponds to the formation of population structure as *Arabidopsis* expanded its range. The most recent population splits are between Espongeiro and Pico da Cruz. In Fogo, the more arid island, we found evidence for a bottleneck that lasted approximately 930 years after colonization (Supplementary Fig. 8a-b). Once the population did begin to expand in Fogo, it dispersed to three regions (Monte Velha, Inferno and Lava) and structure developed among these.

3. Niche modelling to predict suitability of Cape Verde habitat based on Moroccan distribution

The variables in the final Maxent model were the length of the growing season, isothermality, maximum temperature of the warmest month, minimum temperature of the coldest month, temperature annual range, mean temperature of the wettest quarter and precipitation seasonality. When we projected the Moroccan niche model onto the Cape Verde archipelago, we found that there was no habitat predicted to be suitable for colonization from Morocco. As is generally the case in niche modelling, correlations between environmental variables across the range result in alternative possible models. The model we present uses a set of variables chosen based on ecological and biological relevance. However, predicted suitability of CVI habitat for Moroccan accessions did not change across different variable selection regimes: none predicted suitable habitat for Moroccan *A. thaliana* establishment in CVI. This is likely due to the fact that the climate variable values for CVI lie outside or nearly outside the distribution of most Moroccan climatic variables. As a result, the predicted (dis-)similarity may be more informative, which shows that much of Santo Antão and a band around the highest altitude Bordeira region of Fogo has the highest predicted similarity. These most similar regions encompass locations where we found populations of *Arabidopsis* in CVI.

4. Evidence of positive selection in CVI

4.1. d_{sel}/d_{neu}, DFE and proportion of adaptive variation (alpha)

We computed the relative rate of fixation of 0-fold non-synonymous to 4-fold synonymous variants (d_{sel}/d_{neu}) for the Moroccan and CVI lineages. In the absence of selection $d_{sel}/d_{neu}=1$, as non-synonymous and synonymous mutation have the same probability to arise and fix. Purifying selection

reduces the probability of fixation of non-synonymous mutations resulting in $0 \leq d_{sel}/d_{neu} < 1$. A relaxation of purifying selection will move the d_{sel}/d_{neu} ratio from $0 \leq d_{sel}/d_{neu} < 1$ towards neutrality, $d_{sel}/d_{neu}=1$. Positive selection is in principle the only force that can result in higher substitution rates of functional compared to neutral mutations, resulting in $d_{sel}/d_{neu} > 1$. However, multiple forces are likely to be acting at any one time across loci in the genome, so that the observed d_{sel}/d_{neu} is expected to represent a composite of the neutral evolution, purifying selection and adaptive evolution. Therefore $d_{sel}/d_{neu} > 1$ implies that there are many advantageous mutations fixed on the branch leading to the clade.

Consistent with the large-scale effect of purifying selection in continental *A. thaliana*, the d_{sel}/d_{neu} ratio in Morocco, polarized to *A. lyrata*, is 0.18 (Fig. 5a). For Santo Antão and Fogo, we analysed the long branch of divergence from the continents, on which mutations are found fixed derived in CVI as a whole, as well as the two short branches separating the two islands, where mutations are fixed derived in one or the other island. The long branch of divergence mainly represents the continental history in the 'ghost' population but it is also confounded with the early history in CVI after colonization. In this case, we obtained a d_{sel}/d_{neu} ratio of 0.276, slightly higher than the Moroccan population. Then we examined the short branches separating the two islands, which correspond to the past 4-10 ky of evolution in isolation within Santo Antão and Fogo. In this case the genome-wide d_{sel}/d_{neu} ratios (Santo Antão: $d_{sel}/d_{neu}=2.2$, Fogo: $d_{sel}/d_{neu}=1.7$) are much higher than for the Moroccan population and also higher than unity, consistent with strong positive selection acting across the genomes of the island populations. Due to the very shallow history within each island, few mutations had the time to arise and fix in each functional category; as a consequence, confidence intervals around these estimates are necessarily high, but nonetheless consistent with exceptionally high d_{sel}/d_{neu} values in CVI. To further investigate these patterns, we used the software polyDfe[36] to estimate the statistic alpha, the proportion of non-synonymous substitutions driven by positive selection or linked to positively selected alleles. Consistent with the high d_{sel}/d_{neu} ratio, we estimate that in Santo Antão 70.7% and in Fogo 62.8% of non-synonymous substitutions were driven to fixation by positive selection or by linkage to positively selected alleles.

Additionally, we estimated the distribution of fitness effects (DFE) of variants segregating in the CVI and Moroccan populations with polyDfe [36]. The DFE in the two Cape Verde islands are enriched both in neutral and slightly deleterious variants (-1, 0 category), as well as in variants of large positive effects on fitness (all positive categories) compared to Moroccan populations (Fig. 5b). This is consistent with a reduction in efficiency of purifying selection relative to the continent combined with strong positive selection in response to the novel CVI environment. The two islands differ somewhat in the inferred parameters of positive selection. In Santo Antão, the percentage of variants estimated to have a positive

effect on fitness (p_b) was 2.9%, with an average effect of S_b = 54.7. In Fogo, the percentage of beneficial variants was greater, p_b = 19.7%, but with a smaller average inferred effect on fitness, S_b = 6.8.

5. Parallel adaptation by reduced time to flowering

5.1. GWAS in Santo Antão reveals large effect of *FRI* K232X on flowering time

We mapped flowering time in Santo Antão using a linear mixed model (LMM) that controls for population stratification [51]. We used the median per genotype across replicates as phenotype, since no block effect was detected in the simulated CVI conditions experiment (Supplementary Fig. 9). All typed variants in the genome explained 99.997% of the observed phenotypic variance (PVE; also known as 'chip heritability' or 'SNP heritability') and we identified one clear genome-wide significant peak on top of chromosome 4. This peak included *FRIGIDA* (*FRI*, AT4G00650; likelihood ratio test, P = 5.468×10^{-35}), a major flowering time determining gene. The nonsense mutation *FRI* K232X in Cvi-0 truncates the protein and was previously shown to strongly reduce flowering time [42]. Adding *FRI* K232X to our model as a covariate allowed us to quantify the association with the phenotype. We found that when *FRI* K232X was added to the model, the percentage of phenotypic variance explained by all remaining markers decreased to 53.59%, with an estimated effect size for the covariate of -35.27 ± 1.50 days. This means that this single SNP is able to reduce flowering time in this population in about 35 days and explain 46.41% of the phenotypic variance. Concordant with this estimate, in the natural population, *FRI* 232X is associated with a decrease in flowering time of 34 days (MWW test, W = 7, P < 2.2×10^{-16}), and a 140-fold increase in seed number (+387 seeds; MWW test, W = 4541, P = 7.179×10^{-14}; Fig. 6e). In the Col-0 background, under CVI conditions, a non-functional *FRI* allele was responsible for a decrease in flowering time of 27 days (MWW test, W = 0, P = 0.00384) and an increase in fitness of 669 seeds (MWW test, W = 37.5, P = 0.008856; Fig. 6e).

5.2. Identification and functional analysis of *FLC* R3X

5.2.1. Bulked segregant analysis

We did not find any statistically significant association in GWAS for flowering time in the Fogo population (Supplementary Fig. 11), suggesting that the genetic variant(s) underlying the uniformly reduced flowering time were fixed or at high frequency (mean=28.72 days, SD=3.76). Taking this into consideration, we scored flowering time in an inter-island F2 population in which the ancestral allele *FRI* K232 was fixed. We observed early flowering individuals segregating at approximately 1:3 ratio, indicating that a single recessive locus is causing the early flowering phenotype. After bulking and sequencing the early tail of this distribution (n=108), we identified a single region on chromosome 5 between 2 Mbp and

3.3 Mbp where the frequency of the Fogo alleles was greater than 95% in the sequenced pool. Among the 33 variants in this region (Supplementary Table 12), 14 are at a frequency greater than 90% in the natural population and therefore they are stronger candidates to explain the uniform early flowering observed in Fogo. Of these 14, two are predicted by SnpEff [101] to have moderate impact, and one to have high impact.

The two moderate impact variants affect AT5G09930, an ABC transporter protein, causing a missense mutation (V193F), and AT5G07520, a glycine-rich protein expressed only in flowers during a specific developmental stage (flower stage 12), with an 11 amino acid deletion (A202_A213del). The high impact variant is predicted as causing a premature stop codon in AT5G10140 (R3X). AT5G10140 is *FLOWERING LOCUS C* (*FLC*), a MADS-box protein central to the flowering time pathway. *FLC* is regulated by *FRI* and vernalization, contributes to temperature compensation of the circadian clock, and acts as a repressor of floral transition [202,203]. Due to its central function in flowering time, *FLC* is the best candidate for the early flowering observed in Fogo.

5.2.2. Functional characterization of FLC nonsense mutation

Although the flowering time gene *FLC* contains a premature truncation variant at the third amino acid fixed in Fogo, the gene could be functional, e.g., due to an alternative start codon or transcriptional read-through. To determine whether *FLC* is functional in Fogo, we quantified *FLC* mRNA levels in the natural population, and compared them to *FLC* mRNA levels in Santo Antão, Morocco, the Col-0 reference strain, and Col-0 *FRI*-Sf2 (functional *FRI* in Col-0 background) with and without a functional *FLC* [53] (noted as $FRI^{+}FLC^{+}$ and $FRI^{+}FLC^{-}$, respectively; Supplementary Fig. 12a).

Compared to $FRI^{+}FLC^{+}$, transcript levels of *FLC* were reduced in Fogo individuals (Kruskal-Wallis, $P < 1x10^{-4}$), as expected if *FLC* R3X results in non-functional *FLC*. Similarly, *FLC* transcription was reduced in $FRI^{+}FLC^{-}$ and Col-0 wild-type (non-functional *FRI* and functional *FLC*) compared to $FRI^{+}FLC^{+}$ (Kruskal-Wallis, $P < 1x10^{-4}$), which was consistent with previous findings [64]. In contrast, all accessions from Santo Antão showed high levels of *FLC* expression (comparable to $FRI^{+}FLC^{+}$; Kruskal-Wallis, $P > 0.6783$). This result is in agreement with previous work showing a higher baseline expression of *FLC* in Cvi-0 [42,64]. Within Santo Antão, the single sample tested with a functional *FRI* (S5-10) had higher *FLC* expression compared to the other samples from Santo Antão (Kruskal-Wallis, p-values: Cvi-0=0.0006; S1-1=0.0039; S15-3=0.0008), consistent with an effect of *FRI* on *FLC* expression even in the CVI genetic background[42,64]. In Moroccan accessions, we observed high levels of *FLC* expression across populations, with much larger variation.

The two loci, *FRI* and *FLC*, explain the differences in bolting time in the tested accessions (Supplementary Figure 12b). In the Col-0 background, when both loci have functional alleles ($FRI^{+}FLC^{+}$), the plants bolted later than when either of the loci had one non-functional allele (non-functional *FRI* in

Col-0 and non-functional *FLC* in $FRI^{+}FLC^{-}$; Kruskal-Wallis, P = 0.0047 and P = 0.0002, respectively). In Santo Antão, all individuals bolted early, comparable to Col-0 (Kruskal-Wallis, for S1-1, S15-3, Cvi-0: P = 1) and $FRI^{+}FLC^{-}$ (Kruskal-Wallis, for S15-3, Cvi-0: P = 1, for S1-1: P = 0.2376), as all accessions carry a non-functional *FRI*. The exception was S5-10, which bolted later (Kruskal-Wallis, p-values: Cvi-0=0.0005; S1-1=0.5811; S15-3=0.0026) due to a functional *FRI* allele. This accession also showed higher levels of *FLC* expression, consistent with a role of regulation of *FLC* mRNA levels by *FRI* [42,64]. In Fogo, the non-functional *FLC* present in all accessions is consistent with the early bolting and the low levels of *FLC* mRNA, both comparable to the non-functional *FLC* mutant in the Col-0 background (Kruskal-Wallis, for F10-1-3, F13-8, F9-2: P = 1, for F3-2: P = 0.2014). In Morocco, the levels of *FLC* mRNA did not completely explain the bolting time, suggesting that other loci may influence this trait in this population.

5.2.3. FLC genetic complementation test

To further test whether *FLC* is responsible for the early flowering phenotype in Fogo, we crossed 4 individuals from Fogo (all with a potential non-functional *FLC* allele) to Col-0 FRI-Sf2 with and without a functional *FLC* (noted as $FRI^{+}FLC^{+}$ and $FRI^{+}FLC^{-}$ [53], respectively; Supplementary Fig. 12).

If the allele in Fogo was non-functional, when crossed to $FRI^{+}FLC^{-}$, all F1 individuals would flower as early as $FRI^{+}FLC^{-}$, as they would carry non-functional *FLC* alleles from both parents. In this case, F1 individuals homozygous for the null-allele at *FLC* would also flower earlier than heterozygous individuals (F1 from the cross between $FRI^{+}FLC^{+}$ and $FRI^{+}FLC^{-}$), given that the functional *FLC* allele should be dominant [53,64]. On the other hand, if the *FLC* allele in Fogo was functional, the F1 individuals from this cross would have one functional allele and flower later than $FRI^{+}FLC^{-}$, and similar to the heterozygous individuals (F1 from the cross between $FRI^{+}FLC^{+}$ and $FRI^{+}FLC^{-}$). In the inverse cross – between Fogo and $FRI^{+}FLC^{+}$ – if the Fogo allele is non-functional, we expect the F1 to be heterozygous and flower as late as $FRI^{+}FLC^{+}$ and similarly to the F1 resultant from the cross between $FRI^{+}FLC^{+}$ and $FRI^{+}FLC^{-}$.

For the crosses between Fogo individuals and $FRI^{+}FLC^{-}$, we found that all F1 individuals flowered as early as the parental line $FRI^{+}FLC^{-}$ (MWW test, W = 42, P = 0.3611), much earlier than the $FRI^{+}FLC^{+}$ (MWW test, W = 0, P = 0.00248) and the F1 heterozygous between $FRI^{+}FLC^{+}$ and $FRI^{+}FLC^{-}$ (MWW test, W = 4, P = 0.0002342). On the other hand, all the F1 individuals from the cross between Fogo and $FRI^{+}FLC^{+}$ flowered as late as the *FLC* functional parental line $FRI^{+}FLC^{+}$ (MWW test, W = 34, P = 0.8814), much later than the Fogo parents (MWW test, W = 256, P = 1.301×10^{-6}) and also later than the F1 between Fogo and $FRI^{+}FLC^{-}$ (MWW test, W = 116.5, P= 0.001227). These results together suggest that the *FLC* 3X allele found in the natural population in Fogo is indeed non-functional.

5.3. Reconstructed histories of *FRI* 232X and *FLC* 3X

We inferred the coalescent genealogies of *FRI* 232X and *FLC* 3X using RELATE [32]. The coalescent tree for *FRI* 232X is bifurcated rather than hierarchical at the time when the variant is estimated to have arisen, indicating that some branches were likely lost (extinct) or unsampled in the modern populations. The time of the bifurcation is estimated at approximately 2.9 kya, with 95% CI estimated from 200 samples from the MCMC approximately 2.14 kya – 3.74 kya. The tMRCA, which represents the lower bound on the age, is 2.14 kya (95% CI: 1.62-2.72 kya). Based on the coalescent reconstruction for *FLC* 3X and 200 sampled trees, the allele arose between the tMRCA at 3.3 kya (95% CI: 2.82-3.96 kya) and the split from the outgroup at 4.72 kya (95% CI: 3.56-6.66 kya).

We inferred the frequency trajectories for *FRI* 232X and *FLC* 3X using CLUES [52], which uses importance sampling over trees generated in RELATE to infer the frequency trajectories and selection coefficients for the functional variants. For each variant, we defined time bins (epochs) between the present and the emergence of the allele (based on the inferred age of the allele in RELATE when approximately 97.5% of the coalescent trees for the region support the existence of the allele). For *FRI* 232X, these epochs were 0-2 kya and 2-4 kya. For this variant, we found that *s* was maximized in the epoch 2-4 kya years ago, with a selection coefficient of 4.56% (Supplementary Table 9). For *FLC* 3X, these epochs were 0-2 kya, 2-4 kya, and 4-6 kya, with the inferred selection coefficient maximized 4-6 kya with $s = 9.27\%$ (Supplementary Table 10).

Supplementary Note

1. Assessing the fit of adaptation in CVI to models of selection

Theory predicts that when mutational input is low and selection is strong (**SSWM** regime), the first steps of adaptation are likely to occur through large effect mutations, whereas when mutational input is high and selection is weak (**WSSM** regime), adaptation is likely to occur through more, smaller effect variants [8,56–61]. We examine the CVI case in the context of SSWM versus WSSM regimes. First, we approximate the genome-wide mutation rate for the adaptive phenotype (very early flowering through loss of vernalization) and then we apply the inferences we made about population history and selection coefficients to examine the fit of adaptation in each of the two Cape Verde islands to these two models.

The SSWM model is expected to hold when the total number of new mutations that enter a diploid population each generation is small such that $4NU_b \ll 1$ (where N is population size and U_b is the genome-wide beneficial mutation rate for the focal trait) and when selection is strong $4Ns >> 1$. In this scenario, single new beneficial mutations arise and overcome genetic drift ($s >> 1/4N$) and subsequently rise in frequency without interference (by linkage or epistasis) from beneficial alleles at other loci [56–58]. Note that small population size (as in a founder population) plays a double role. On one hand, it enables SSWM type adaptation by restricting the mutation input; on the other hand, it can inhibit adaptation altogether unless selection is strong. The combination of small population size and strong selection is expected to result in a "selective-sweep type" architecture of adaptation [5] where one or few variants with large effects underlie the adaptive phenotype. Alternatively, adaptation in the WSSM regime occurs through small frequency shifts at a large number of alleles with small individual effect. Theory shows that $4NU_b = 1$ is indeed the threshold that separates the sweep-like from highly polygenic architectures [5,62].

Time to flowering has been studied extensively in *A. thaliana* and much is known about its molecular basis. We used available molecular and functional analyses of major flowering time loci to produce rough approximations of U_b. First, we reviewed the literature to identify the loci and mutational effects that lead to effects on flowering time. Many genes can contribute to variation in flowering time (at least 174 genes are thought to be involved in the flowering time pathway (https://www.mpipz.mpg.de/14637/Arabidopsis_flowering_genes [204–207]). However, very few of these cause large or even moderate changes in flowering time. A more relevant (specific) phenotype in this case is the loss of the vernalization requirement (i.e., cold period needed to induce flowering), which results in a large reduction in flowering time. The loci most often implicated in this trait are *FRI* and *FLC*, both of which are essential for vernalization response. Loss of function of either of these genes results in loss of

the vernalization requirement for flowering. When diverse rapid flowering accessions were examined, 85% turned out to have clear evidence of functional mutations in *FRI*, *FLC* or both [64,205]. In the absence of a vernalization treatment, complete loss of function of these genes results in a reduction of flowering time of approximately 35 days relative to wild type (Fig. 7e). In nature, variation in *FRI* occurs primarily by loss of function mutations in coding regions [64,102], while most putative functional variation in *FLC* is found in the first intron [77,103], which contains a well-characterized regulatory element [104].

Based on this information about the genetic basis of large effect changes in flowering time (vernalization), we can roughly estimate U_b, the genome-wide per individual mutation rate of beneficial mutations that act through the focal phenotype (the number of nucleotides whose changes would cause a large shift in phenotype).

First, we focus on loss of function mutations due to SNPs in coding regions of the genes involved in complete loss of vernalization (*FRI* and *FLC*). For these, the total length of the two genes is 805 amino acids. We estimate that on average three out of 64 mutations could lead to a premature stop codon. This results in a mutational target size for loss of function through SNPs in coding regions of 805*3/64. We assume a mutation rate by SNP changes of 7.1×10^{-9} [(99)]. The rate of introduction of loss of function mutations by premature stop codons per individual and per generation is therefore:

$U_{b\,(coding\_SNPs)} = 805 \times 3/64 \times 7.1 \times 10^{-9} = 2.68 \times 10^{-7}$

We can additionally account for mutation by indels, using the mutation rate estimated from mutation accumulation lines. We used the per base mutation rate estimated for 1-3 bp long deletions, 4.0×10^{-10} [(99)]. The mutational target size is then 805 amino acids * 3 nucleotides/amino acid. The rate of introduction of indel mutations in the coding region is therefore:

$U_{b(coding\_indels)} = 805 \times 3 \times 4 \times 10^{-10} = 9.66 \times 10^{-7}$

Combining the probability of mutation by SNPs or indels in coding regions gives:

$U_{b(coding)} = 2.68 \times 10^{-7} + 9.66 \times 10^{-7} = 1.23 \times 10^{-6}$

Regulatory mutations in *FLC* and *FRI* also have the potential to impact flowering time. To include the possibility of regulatory mutations in our estimate of U_b, we estimated the probability that a mutation has a major regulatory effect based on the literature. Regulatory elements can include promoter regions as well as conserved non-coding elements upstream, downstream or within introns. For *FRI* and *FLC* these are well-studied. Core promoters are on average approximately 75 bp in *A. thaliana* and are tightly packed with regulatory elements [103]. Given a moderate level of nucleotide-level functional redundancy, we assumed that one in 10 possible changes in the core promoter could have major functional effects. In

addition, we assumed that larger 5' regions with interspersed transcription factor binding sites would add 300 bp to the core [208]. In these regions regulatory motifs are generally more dispersed and redundancy is likely to be rather high at the nucleotide level. We assumed that on average the probability of a mutation resulting in a strong functional effect in these regions would be one in 50. In *FLC* there is evidence that motifs exist within a 'vernalization response element' in the first intron (289 bp) that are crucial to function [104]. Here, we assumed that one in 20 mutations may result in large changes in flowering time.

While we attempted to come to a meaningful approximation of the number of variants (mutational target size) that might have *major* effects on regulatory function of *FLC* or *FRI*, assumptions about this mutational target size are necessarily less certain. However, the final estimate would not change much if other biologically informed estimates were used.

$U_{b(\text{regulatory FRI/FLC\_SNPs})} = (150 \times 0.1 + 600 \times 0.02 + 289 \times 0.05) \times 7.1\text{x}10^{-9} = 2.94\text{x}10^{-7}$

And including indels in regulatory regions gives us:

$U_{b(\text{regulatoryFRI/FLC\_indels})} = (41.45) \times 4*10^{-10} = 1.66\text{x}10^{-8}$

So that the combined probability of a major effect mutation in regulatory regions by SNPs or indels is:

$U_{b(\text{regulatory})} = 2.94\text{x}10^{-7} + 1.66\text{x}10^{-8} = 3.11\text{x}10^{-7}$

And the combined probability of any mutation with a major effect on the vernalization requirement is:

$U_b = 1.23\text{x}10^{-6} + 3.11\text{x}10^{-7} = 1.54\text{x}10^{-6}$

Using the estimate for U_b and an assumption about past population size, we can infer the waiting time for a mutation in the natural population. Then, using our inference about selection coefficients from the allele frequency trajectory in each island, we can assess how each estimate fits with SSWM and WSSM models. We initially focus on Fogo because the population history there is simpler and resolved with higher certainty. Then we turn to Santo Antão, and make some approximations and an estimate for that population.

For N, we used the estimate of N_e from RELATE/COLATE in the Fogo population (N_e=48 individuals at colonization). With that, the per generation population mutation rate for a strong effect functional variant in *FRI* or *FLC* would be

$\Theta_{(\text{FRI/FLC})} = 4N_eU_b = 4 \times 48 \times 1.54\text{x}10^{-6} = 2.95\text{x}10^{-4} << 1.0$

There is of course uncertainty in this result from our rough estimate of both U_b and N_e. However, as our derivation leaves considerable room for error of almost four orders of magnitude, the conclusion that adaptation of our focal trait (loss of vernalization requirement) is mutation limited ($\Theta_{(FRI/FLC)} \ll 1$) appears to be highly plausible. Our estimate corresponds to an expected waiting time for the occurrence of any mutation that abolishes function of the genes of $1/\Theta = 3775$ generations. If we break this down into coding and non-coding we find that the estimated waiting time for a coding change specifically is $(1/4N_eU_{b(coding\_FLC/FRI)}) = 4727$ generations and the waiting time for loss or major reduction of function through a regulatory change specifically is estimated at $(1/4N_eU_{b(regulatory\_FLC/FRI)}) = 18{,}694$ generations. Based on this, if such a variant was required to escape eventual extinction, extinction risk would be very high. Further, given the much lower non-coding adaptive mutation rate, an adaptive mutation from variation in the coding region would be much more likely in this case.

Under the SSWM model, strong selection is required in order to escape drift. Here, we apply a selection coefficient based on the reconstructed frequency trajectory of *FLC* 3X in the time just after it is estimated to have arisen ($s = 0.0927$). If *FLC* 3X resulted in a population size increase, this estimate would be highly conservative because the population size change is used as the null model here. Recall that SSWM is expected to hold when $4NU_b \ll 1$ and $s >> 1/4N$. Even with the conservative estimate of s, we find that this is well above $1/4N_e$: $s = 0.0927$ and $1/4N_e = 5.21 \times 10^{-3}$. This implies that a new mutation that obliterates the vernalization requirement in a Fogo-like environment would tend to escape drift. We conclude that the scenario is consistent with the SSWM regime, where adaptation relies on sweeps of large-effect alleles [5,8,56,62].

Of course, variance on the estimates of the time for a variant to arise and fix in the population would be high due to the stochastic nature of mutation and the uncertainty of establishment. We conducted simulations under a model designed to fit the Fogo population to quantify the probability of fixation under a constant-size Wright-Fisher model with three plausible estimates of the selfing coefficient (90%, 95% and 99%). The model ignores the possibility of extinction, a risk that may be high under individual and/or temporal (e.g., due to climate) variance in reproductive success. After 6000 generations, 13.5 - 24% % runs resulted in fixation of the adaptive variant (a variant that eradicates the vernalization requirement) (Supplementary Table 11). The time of the simulated trajectories was inverted to a backwards in time model to follow the same structure as the inferred trajectory from CLUES. Supplementary Figure 14 shows the trajectories of functional variants that arise under the three different selfing coefficients along with the inferred trajectory.

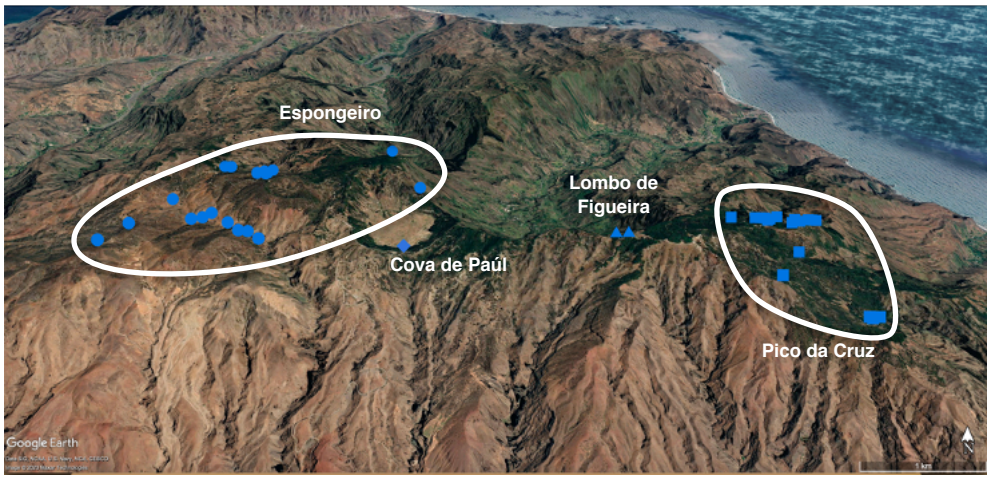
Santo Antão fits somewhat less well with the idealized model due to its more complex history. The colonization event in Santo Antão is confounded with the much earlier split from the 'ghost' population. Further, population structure appears to have developed in Santo Antão well before *FRI* 232X appeared, consistent with a more permissive landscape where late-flowering plants could survive to reproduce potentially with moderate to high success in some years. We consider a range of possible N_e from 500 to 1000 based on the estimated N_e in Santo Antão at the time when *FRI* 232X arose. We examine the scenario with *s* set to 0.046 based on the selection coefficient inferred from the frequency trajectory of *FRI* 232X in Santo Antão during the peak of selection. U_b is the same as in the Fogo case.

With these assumptions $4N_eU_b$ would range from $3.08x10^{-3}$ to $6.16x10^{-3}$, which is again much less than 1.0, indicating limited mutational input for adaptation. And *s* of 0.046 is much greater than $1/4N_e$, which ranges from $5x10^{-4}$ to $2.5x10^{-4}$, consistent with expectations for a SSWM model, albeit less extreme than the Fogo case.

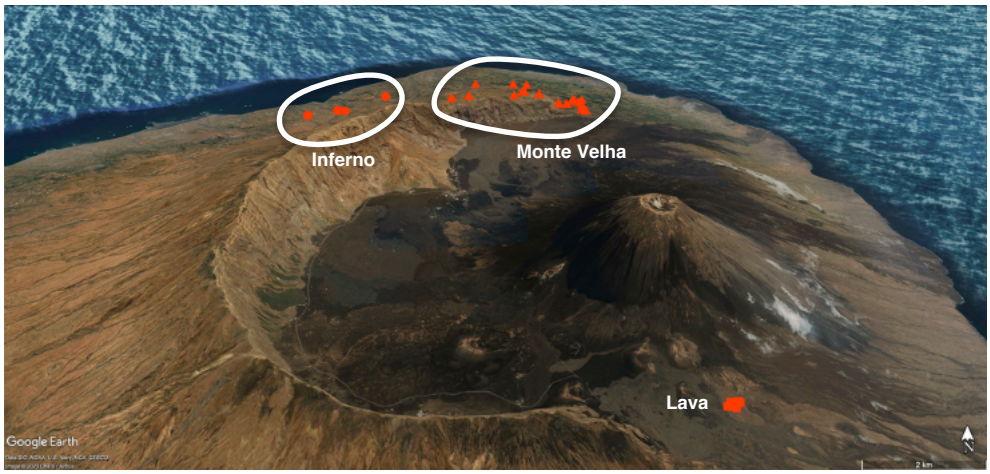
Given that most observed large changes in flowering time in nature can be attributed to changes in *FRI* and *FLC* (85%) [63,64,66,76,205,209], it seems reasonable to focus on these genes in our analysis of mutational target size. But there are many genes across the genome that act in the flowering time pathway and through which variation could affect flowering time. If we broadened the phenotypic definition to include variants with less extreme effects on flowering time, the architecture of the trait would be more polygenic and U_b would be larger. A few other genes are known to have fairly large effects on flowering time (ranging from 6 to 10 days) and act through vernalization (e.g., *FLM*, genes in the *MAF2-5* gene cluster, and *SVP*). We could attempt to calculate expected waiting times for any change in flowering time under progressively more polygenic architectures. While we have no estimate of *s* to which we could compare in these cases, we can assume that it would be smaller, and that drift would then play a much more important role in the probability that mutations in these genes would be established in the population. In a larger population, where new variants were less susceptible to loss by drift, the first steps of adaptation may be more likely to include variants in genes with weaker effects.

Supplementary Figures

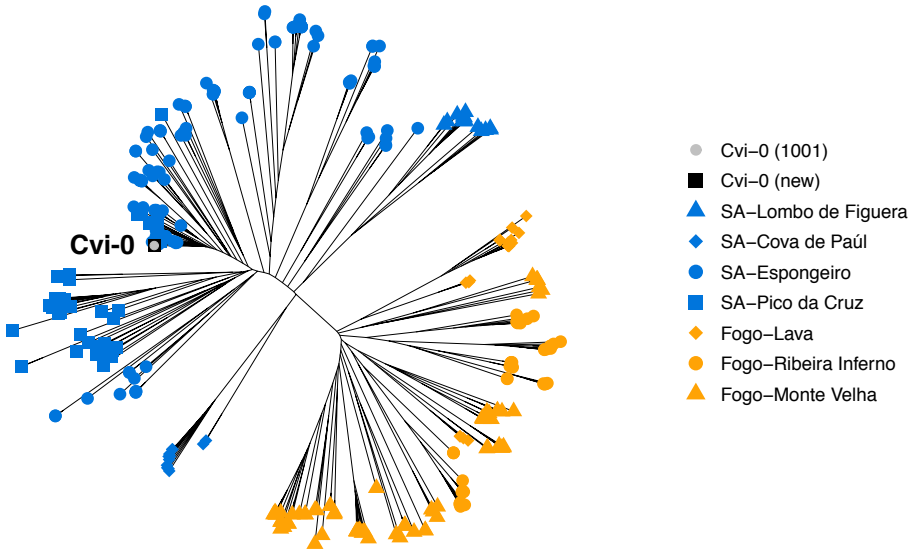
a



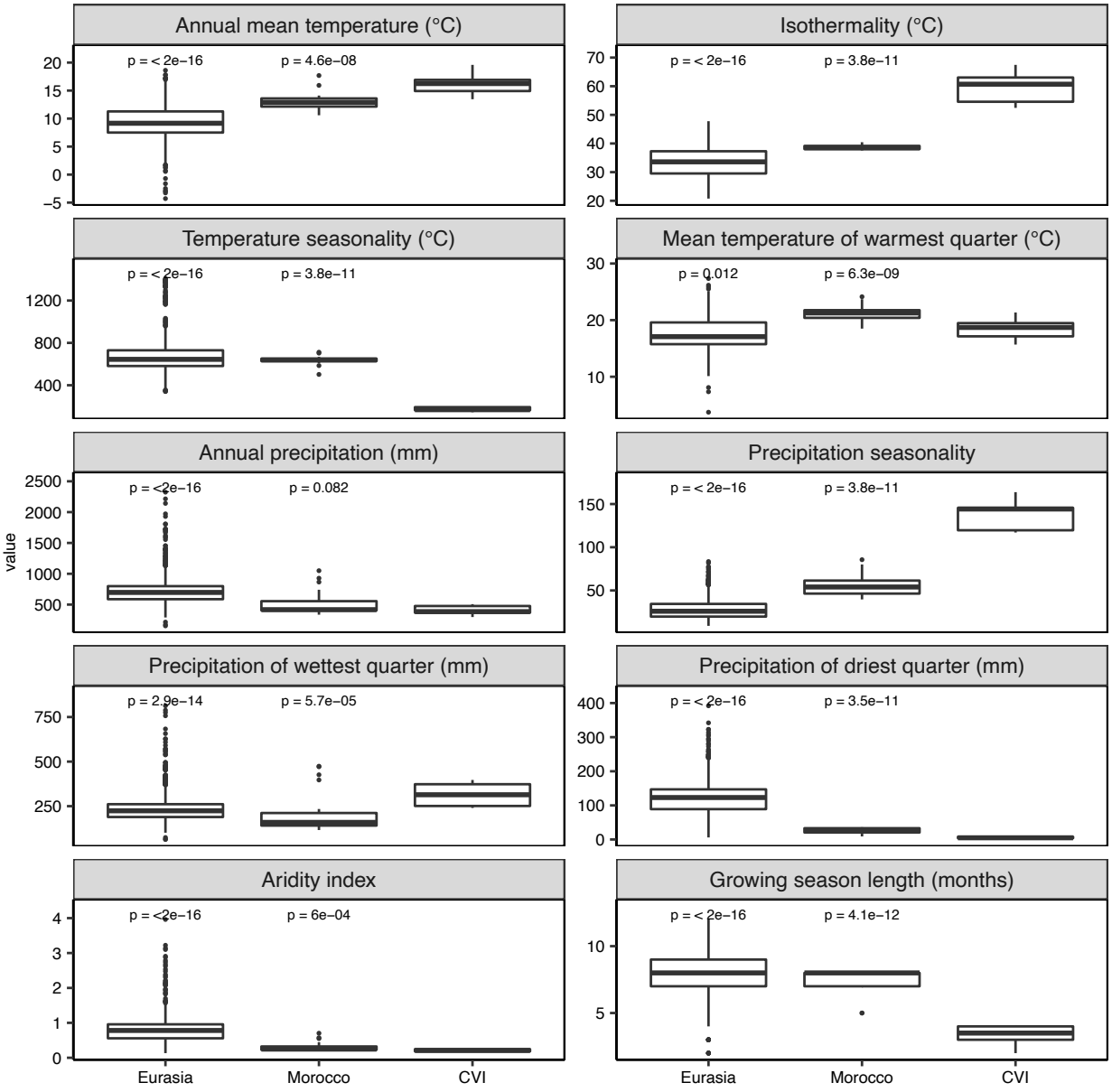
b



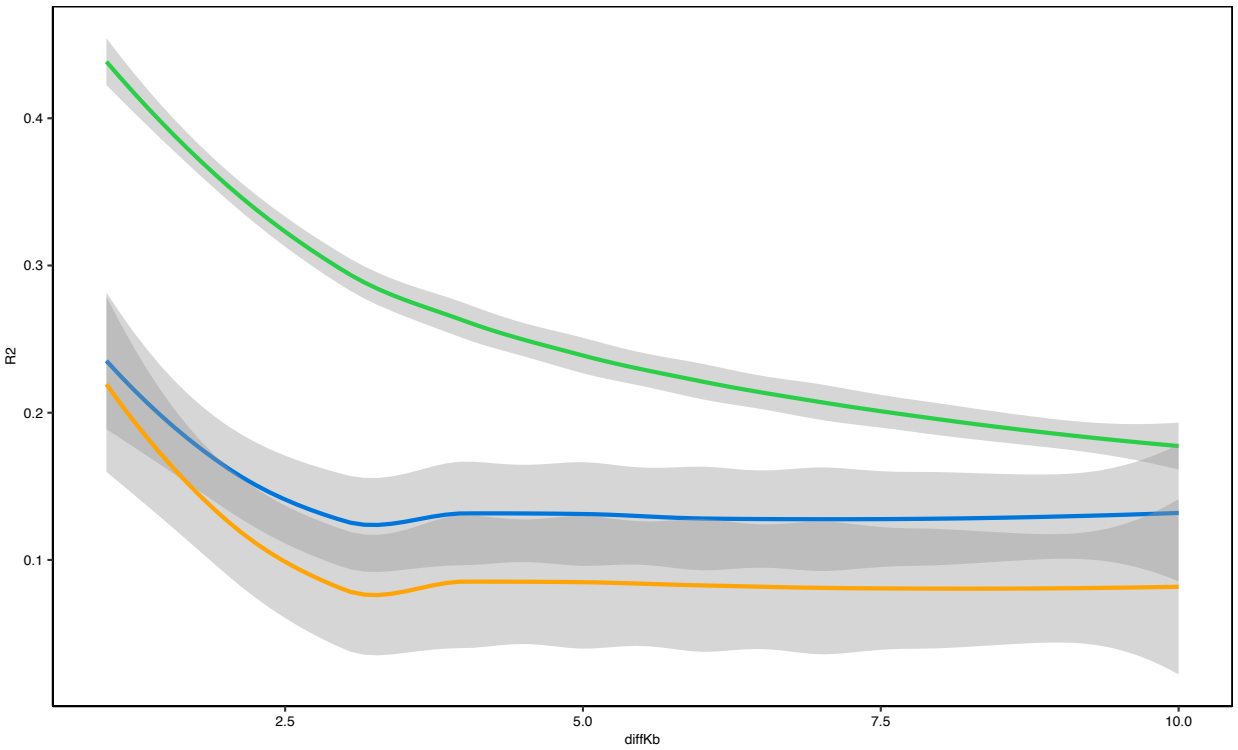
c



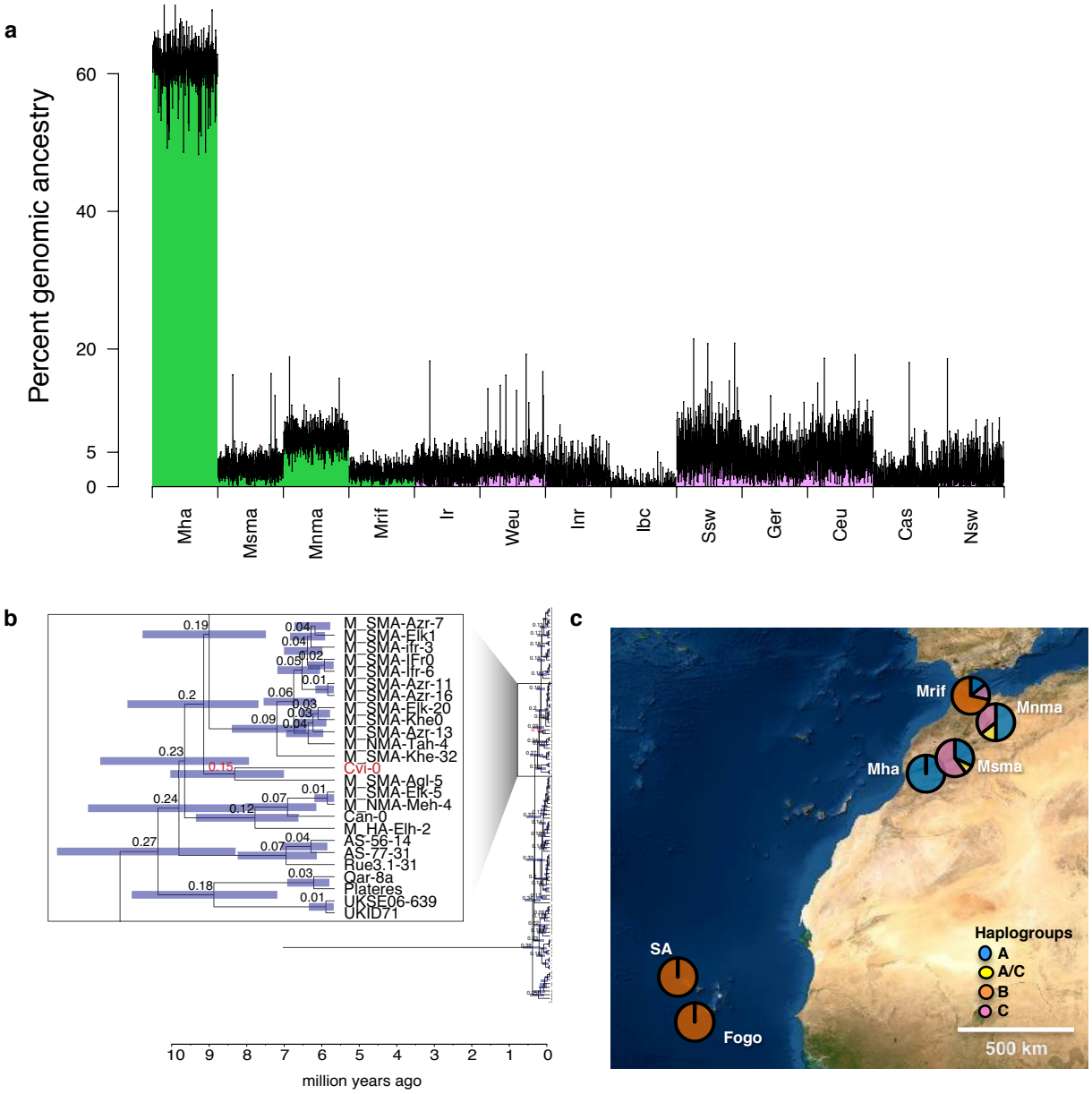
Supplementary Figure 1. Geographic locations and genetic clustering of CVI samples. Maps of sub-populations in a, Santo Antão and b, Fogo. c, Neighbour-joining tree showing deep separation between islands and clustering of island sub-populations. Cvi-0 clusters with the Santo Antão population. Samples from Santo Antão are shown in blue, with sub-populations denoted as follows: Lombo de Figueira (triangles), Cova de Paúl (diamonds), Espongeiro (circles), and Pico da Cruz (squares). Samples from Fogo are shown in orange, with sub-populations denoted as Lava (diamonds), Ribeira Inferno (circles), and Monte Velha (triangles). Cvi-0 (1001) represents the Cvi-0 sequenced in the 1001 Genomes Project for *Arabidopsis thaliana*, while Cvi-0 (new) represents the Cvi-0 re-sequenced in this study.



Supplementary Figure 2. Climate at collection sites in CVI relative to Moroccan and Eurasian sites. The three populations are shown on the x-axis, while the values for each climatic variable are shown on the y-axis. Each variable is presented with the respective unit. P-values for Wilcoxon tests are shown for Morocco and Eurasia relative to CVI.

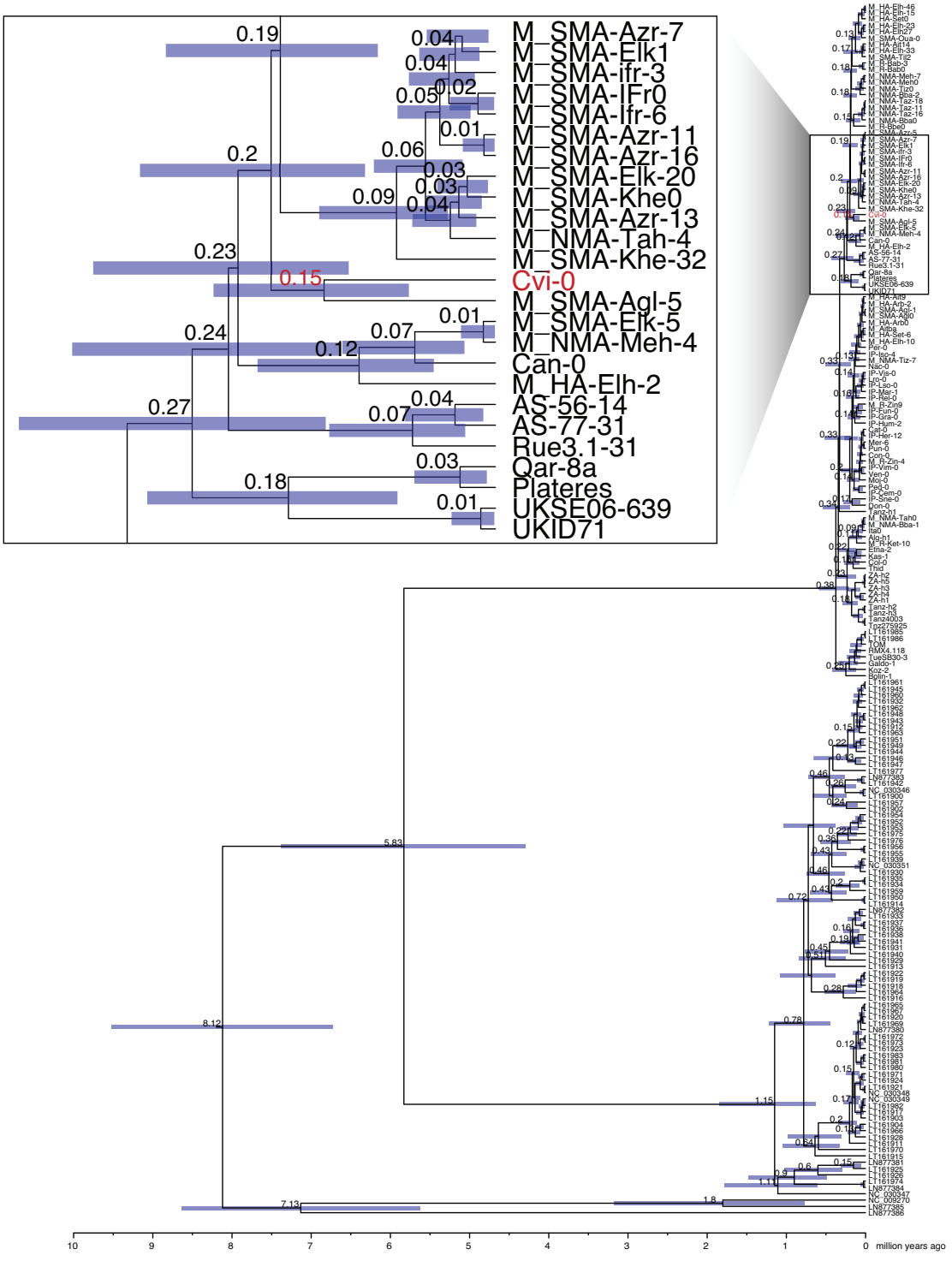


Supplementary Figure 3. Linkage disequilibrium (LD) decay in all three populations. X-axis shows distance between SNPs in kbp, and the y-axis the corresponding r^2 value. Decay of LD is shown for Santo Antão (blue), Fogo (orange) and Morocco (green). Lines were smoothed with locally weighted scatterplot smoothing (LOESS) and the shaded grey area represents the 95% confidence interval.

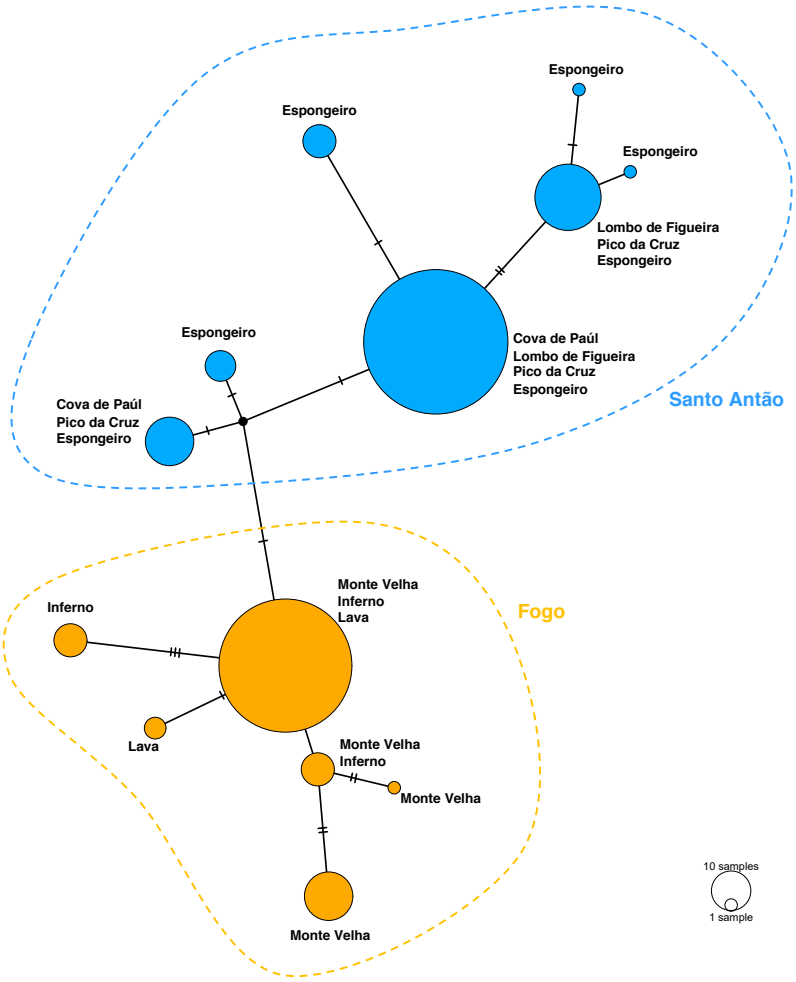


Supplementary Figure 4. Local ancestry of CVI genomes. a, The percentage of CVI genomes for which the closest relative is found in each continental population, inferred with Chromopainter. Averages across runs are shown as coloured bars, 95% confidence interval (CI) as vertical black lines, one for each of 148 CVI genomes. Genome-wide, the Moroccan population in the High Atlas Mountain is the closest relative for 61% of CVI genomes. b, Region of the chloroplast phylogeny with Cvi-0 and worldwide accessions (Full phylogeny in Supplementary Fig. 5). Cvi-0 clusters with the Moroccan population from South Middle Atlas with an estimate of 150 ky divergence time from the closest Moroccan individual. Numbers at the nodes are divergence time in million years. Purple shades represent the 95% CI. c, Geographic distribution of S-locus haplogroups in Morocco and CVI. All individuals from the islands carry haplogroup B, which is only found in the Rif population in Morocco. Different colours in the pie charts represent different haplogroups: A in blue, A/C in yellow, B in orange,

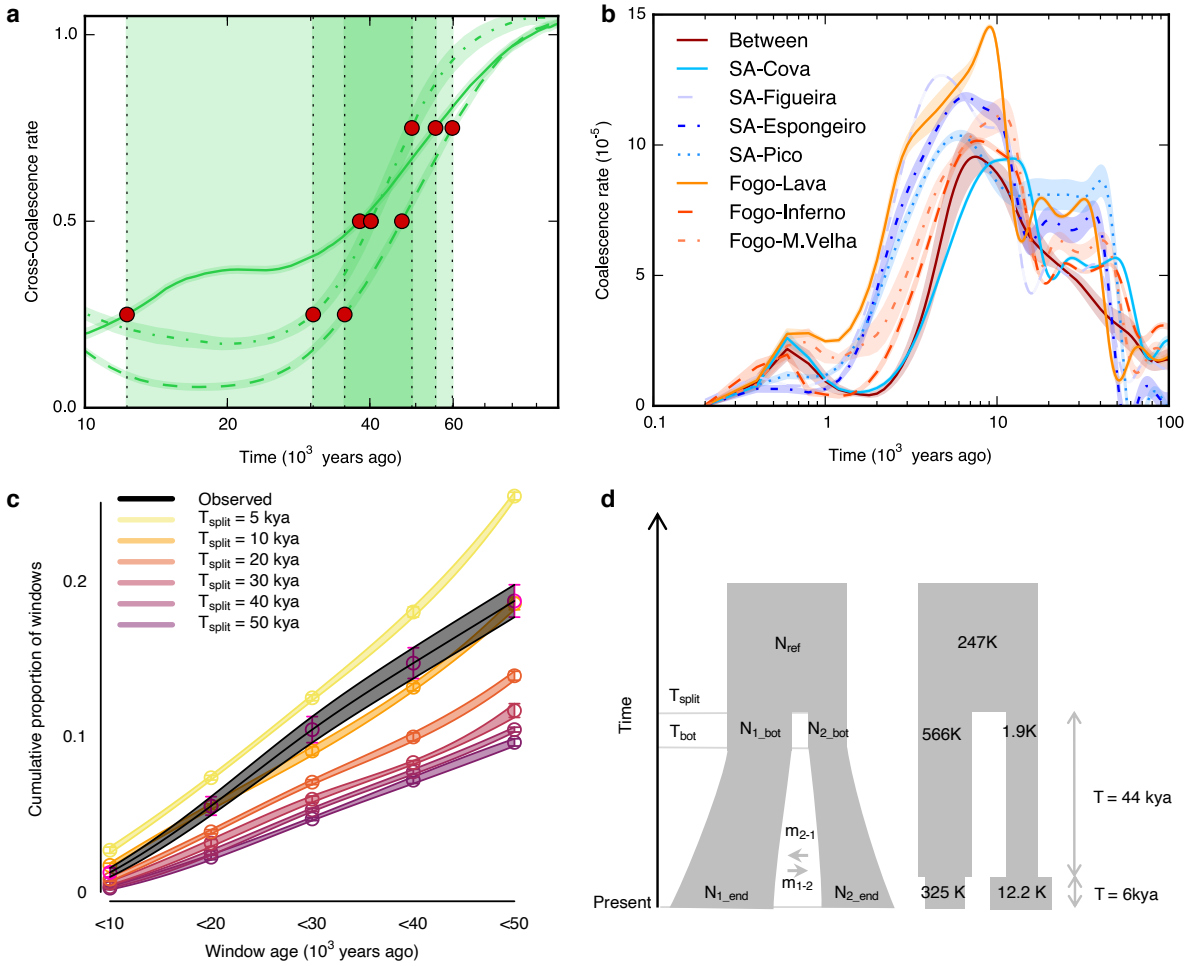
and C in pink. Abbreviations: Mha, Moroccan High Atlas; Msma, Moroccan South Middle Atlas; Mnma, Moroccan North Middle Atlas; Mrif, Moroccan Rif; Ir, Iberian relicts; Weu, Western Europe; Inr, Iberian non-relicts; Ibc, Italy, Balkans and Caucasus; Ssw, South Sweden; Ger, Germany; Ceu, Central Europe; Cas, Central Asia; Nsw, North Sweden; SA, Santo Antão.



Supplementary Figure 5. Time-calibrated chloroplast phylogeny showing the location of Cvi-0 relative to representatives of *A. thaliana*, and other *Arabidopsis* species, *Capsella grandiflora*, *Capsella bursa-pastoris* as well as *Camelina sativa* for calibration. Inset shows the location of Cvi-0.



Supplementary Figure 6. Chloroplast network for all CVI accessions. The size of the nodes is proportional to the number of samples, with the corresponding cluster name. Blue represents haplotypes in Santo Antão, orange in Fogo.



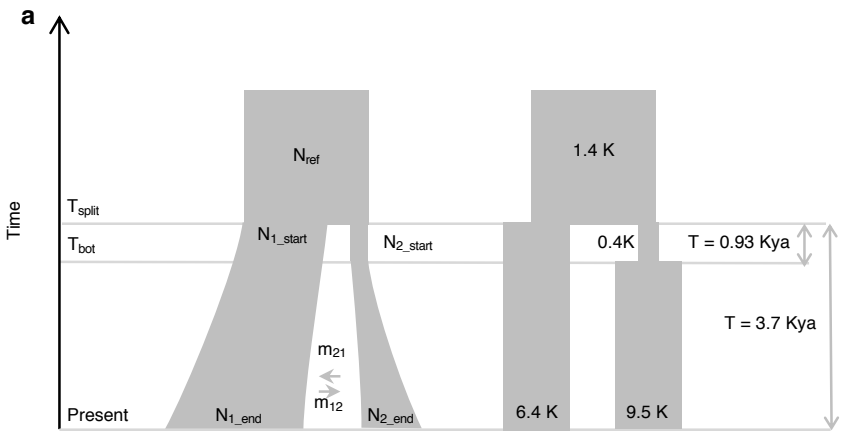
e

Model	MaxLik	AIC	Theta	N_{ref}	N_{1_bot}	N_{2_bot}	N_{1_end}	N_{2_end}	T_{split}	m_{12}	m_{21}	T_{bot}
Simple split	-461	928	1.9K	239K	447K	2.9K	x	x	70K	x	x	x
Exponential	-376	758	2.0K	260K	259K	0.8K	556K	12.0K	56K	x	x	x
Isolation-with-migration	-1051	2112	1.7K	220K	47K	215K	448K	1.5K	1397K	8E-07	6E-06	x
Bottleneck	-396	800	1.9K	236K	526K	0.3K	526K	6.7K	51K	x	x	17K
2-sided bottleneck	-348	708	1.9K	247K	566K	1.9K	325K	12.2K	50K	x	x	44K

Supplementary Figure 7. Modelling of Morocco-CVI split. a, Split time estimated with MSMC-CCR for the Moroccan populations and CVI. Solid line represents the High Atlas population, dashed line the North Middle Atlas and dash-dotted line the South Middle Atlas populations. 0.25 – 0.75 cross coalescence rate quantiles are shown as shaded areas, with point estimates at 0.25, 0.5 and 0.75 highlighted by red dots. b, Coalescence rate as a function of time within Santo Antão and Fogo clusters, and between islands ('Between'), estimated with MSMC2 in 8-haplotypes mode. Lines represent means across octets, shaded areas the 95% CI. c, Cumulative proportion of genomic windows with different inferred ages based on the density of mutations. Simulations were run with a split at different times (5-50 kya) between a Morocco-like and CVI-like population. Points represent observed data and averages across simulations. Whiskers represent 95% CI based on the SE estimated with ordinary non-parametric bootstrap for observed data and on the SE across simulations. Points were interpolated with cubic splines. Simulations with a split at 10 kya most closely match the observed data, supporting a

colonization of CVI from a 'ghost' population more recent than 10 kya. d, Overview of models run in dadi (left) and best model (right) and e, model statistics. Each row corresponds to a different model: Simple split, a population split with constant population sizes; Exponential, a population split followed by exponential changes in population sizes; Isolation with migration, a population split followed by exponential changes in population sizes and asymmetric migration; Bottleneck, a population split with a colonization bottleneck of varying severity (duration and size), followed by constant population sizes; 2-sided bottleneck, a population split with a bottleneck in both populations following the split. MaxLik, maximum log-likelihood obtained across dadi runs; AIC, Akaike Information Criterion; Theta, population mutation parameter inferred in dadi; N_{ref}, size of the ancestral population; N_{1_bot}, N_{2_bot}, size of population one and two at the split, respectively; only in the bottleneck models, N_{1_bot} and N_{2_bot} remain constant for the duration of the bottleneck, T_{bot}; N_{1_end}, N_{2_end}, population sizes at present; T_{split}, split time; m_{12}, m_{21}, migration rates. Estimates of the split time with MSMC-CCR reveal a complex scenario, likely due to the absence of a direct outgroup to CVI in our sample. In these analyses, there is evidence of a split at 40-50 kya, followed by secondary contact or low-level migration between 10 and 20 kya. This initial split and secondary contact likely happened on the continents between sampled Moroccan populations and an unsampled 'ghost' population (the true closest continental outgroup to CVI).

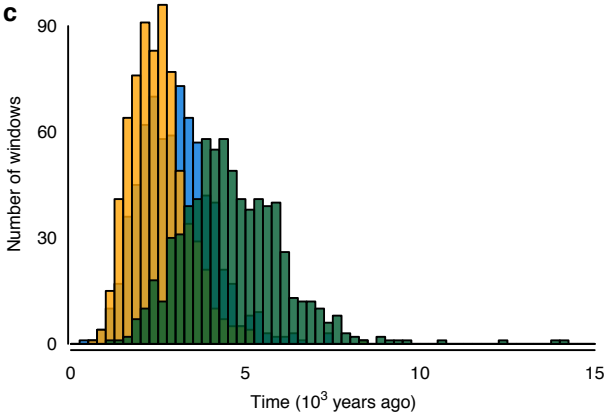
a



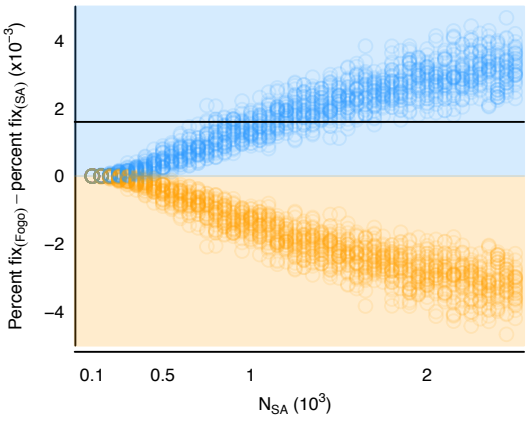
b

Model	MaxLik	AIC	Theta	N_{ref}	N_{1_start}	N_{2_start}	N_{1_end}	N_{2_end}	T_{split}	m_{12}	m_{21}	T_{bot}
Simple split	-389.1	784.3	53.4	776	5487	7750	x	x	4754	x	x	x
Exponential	-313.5	635.0	142.1	2067	1579	488	9268	17167	3715	x	x	x
Isolation with migration	-304.0	620.0	126.6	1841	1427	414	9186	17069	3996	$4.3e^{-6}$	$6.0e^{-7}$	x
Bottleneck	-281.9	573.8	98.6	1435	6434	397	6434	9511	3685	x	x	931

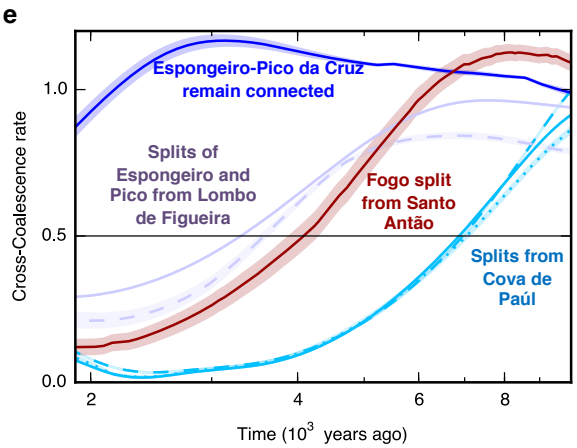
c



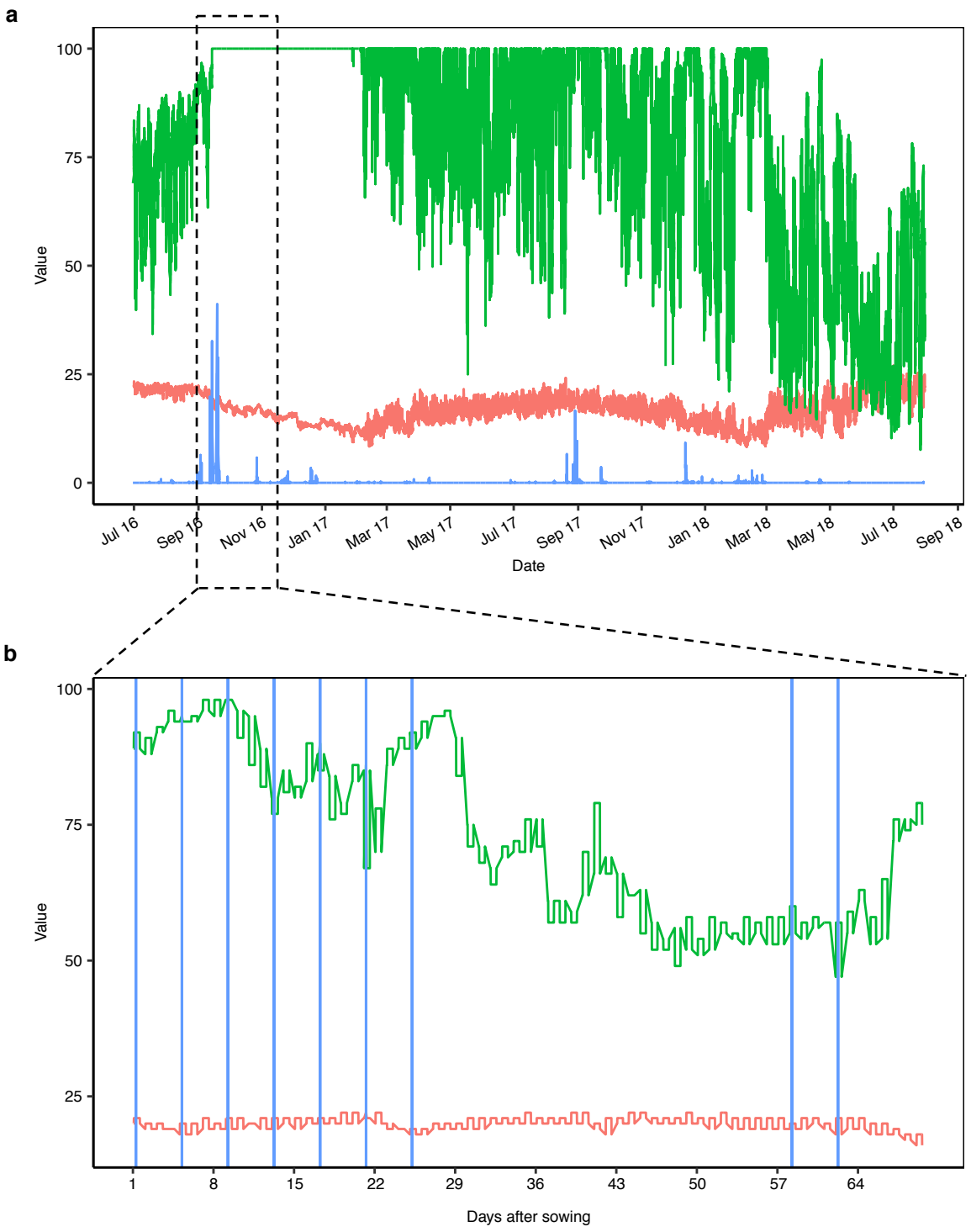
d



e

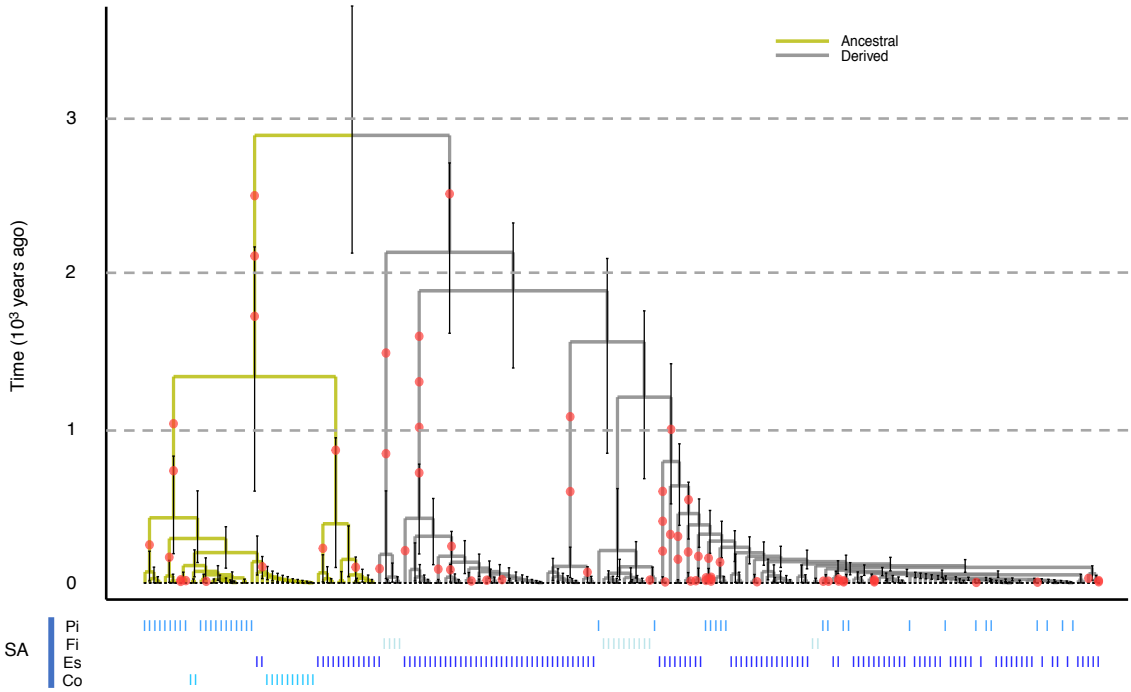


Supplementary Figure 8. Modelling the dynamics within CVI. a, Overview of models run in dadi (left) and best model (right), and b, model statistics. Each row corresponds to a different model: Simple split, a population split with constant population sizes; Exponential, a population split followed by exponential changes in population sizes; Isolation with migration, a population split followed by exponential changes in population sizes and asymmetric migration; Bottleneck, a population split with a colonization bottleneck of varying severity (duration and size), followed by constant population sizes. MaxLik, maximum log-likelihood obtained across dadi runs; AIC, Akaike Information Criterion; Theta, population mutation parameter inferred in dadi; N_{ref}, size of the ancestral population; N_{1_start}, N_{2_start} , size of population one and two at the split, respectively; only in the bottleneck model, N_{2_start} remains constant for the duration of the bottleneck, T_{bot}; N_{1_end}, N_{2_end}, population sizes at present; T_{split}, split time; m_{12}, m_{21}, migration rates. c, Coalescence time within CVI across genomic windows (size=0.1Mbp) provides a lower bound for colonization timing at approximately 7 kya (95^{th} percentile of coalescence times between islands). Comparisons within island are in blue (Santo Antão) and orange (Fogo), and comparisons between islands in green. d, Forward simulations fitted to observed data support a scenario in which Santo Antão was colonized prior to Fogo. The percentage of fixed differences in Fogo compared to Santo Antão (SA) minus the percentage of fixed differences in Santo Antão compared to Fogo (y-axis) varies as a function of which island was colonized first (positive if Santo Antão, blue shade; negative if Fogo, orange shade) and of the population size at the split (x-axis). Circles represent the results from simulations (n=1516): blue if Santo Antão was colonized first, orange if Fogo was colonized first. Simulations in which Santo Antão was colonized first with a population size of approximately 1K at the time of the split, and in which Fogo was colonized later from Santo Antão, fit best the observed data (horizontal black line). e, Cross-coalescence rates (CCR) among populations in Santo Antão as inferred by MSMC-CCR. The estimated between-population split time corresponds to CCR=0.5 (horizontal line). Dark red shows CCR between Santo Antão and Fogo for reference. Lines represent means across octets, shaded areas the 95% CI.

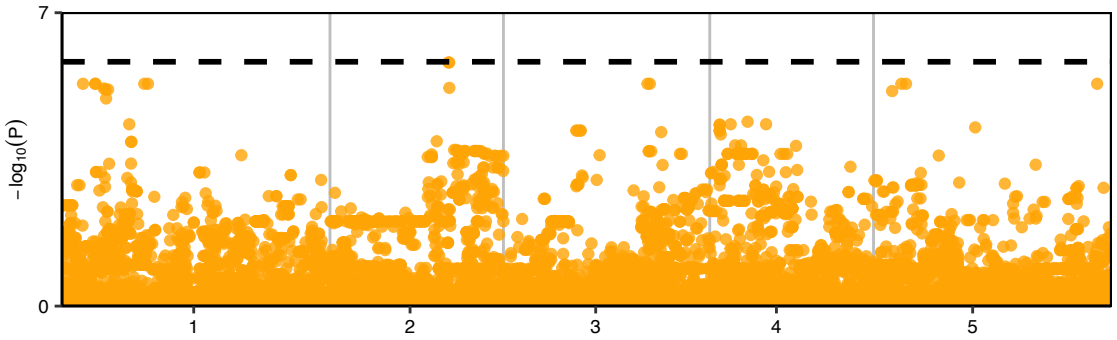


Supplementary Figure 9. Field climate data and simulated conditions. a, Field site measurements for precipitation (blue), humidity (green) and temperature (red) using loggers in the Espongeiro field site over two years (July 2016 to July 2018). Y-axis shows values in mm for precipitation, in percentage for humidity and in degrees Celsius for temperature. The dashed line highlights the period of time simulated in panel b. b, Chamber conditions during the experiment for humidity (green) and temperature (red). Watering is shown as blue vertical lines. Y-axis shows values in percentage for

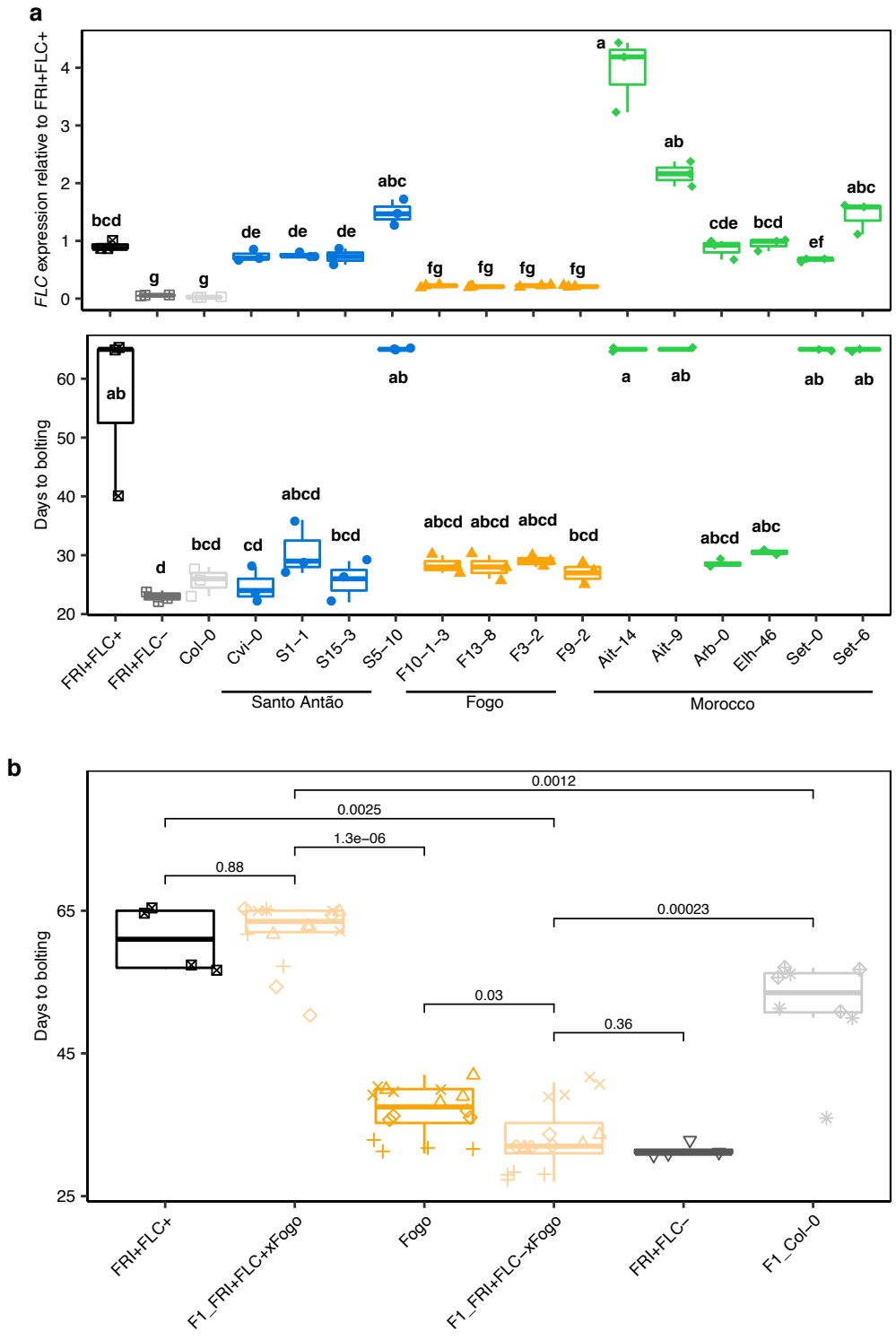
humidity and in degrees Celsius for temperature, and x-axis shows days after sowing (day 1: 1st Sept 2016).



Supplementary Figure 10. Evolutionary history of *FRI* K232X. Marginal genealogical tree estimated in RELATE for *FRI* (Chr4:269719). Individuals are shown across the x-axis with their ancestral and derived carriers coloured in mustard and grey, respectively. The estimated time to coalescence is shown on the y-axis. Confidence intervals indicate the 0.025 and 0.975 quantiles of the posterior density of coalescence times and are shown by the vertical black lines on the tree. Red dots represent mapped SNPs on their corresponding branches. Abbreviations: SA, Santo Antão; Pi, Pico da Cruz; Fi, Lombo de Figueira; Es, Espongeiro; and Co, Cova de Paúl.

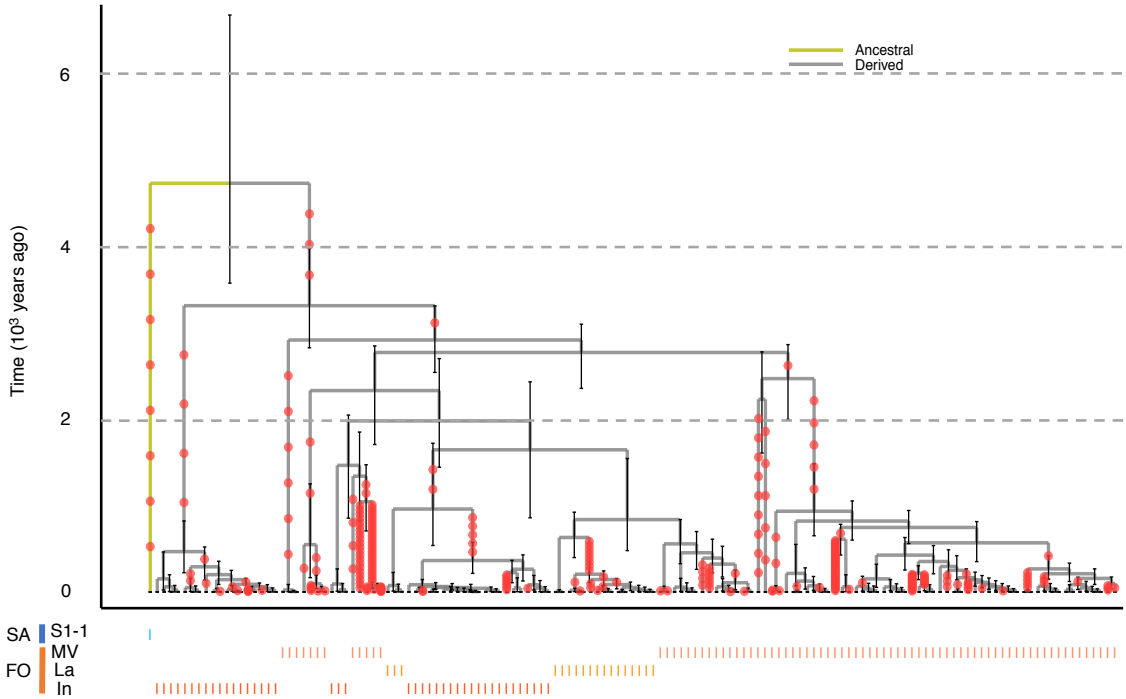


Supplementary Figure 11. GWAS in Fogo. Manhattan plot showing results of GWAS for bolting time in the Fogo population (n=129). Dashed line represents the Bonferroni significance threshold.

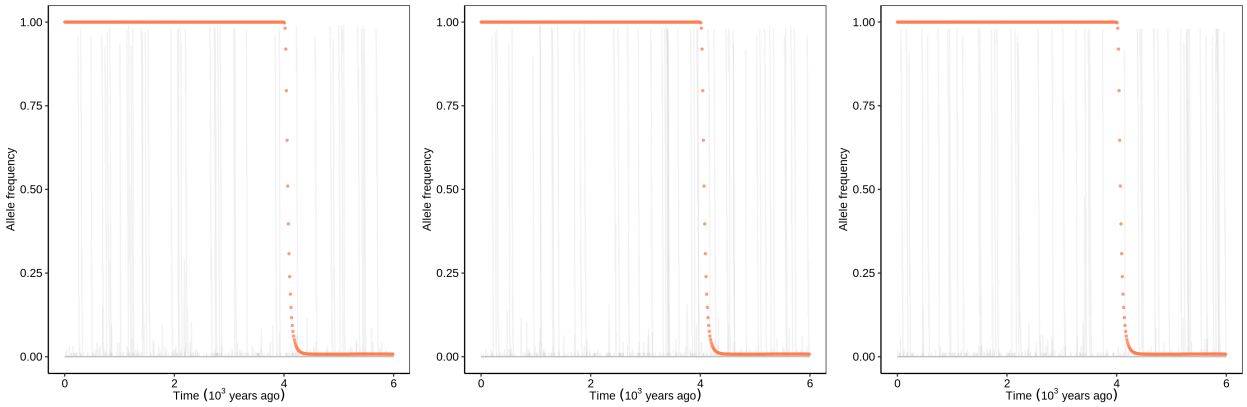


Supplementary Figure 12. Functional characterization of *FLC* nonsense mutation. a, *FLC* effect on natural accessions from CVI and Morocco, and in mutant lines in the Col-0 background. *FRI* K232X segregates among the Santo Antão individuals shown. Cvi-0, S1-1 and S15-3 carry the *FRI* 232X allele, while S5-10 carries the ancestral *FRI* K232 allele. All Fogo individuals carry the derived *FLC* 3X allele. Top: *FLC* mRNA expression levels, with y-axis showing the expression levels relative to functional *FRI*

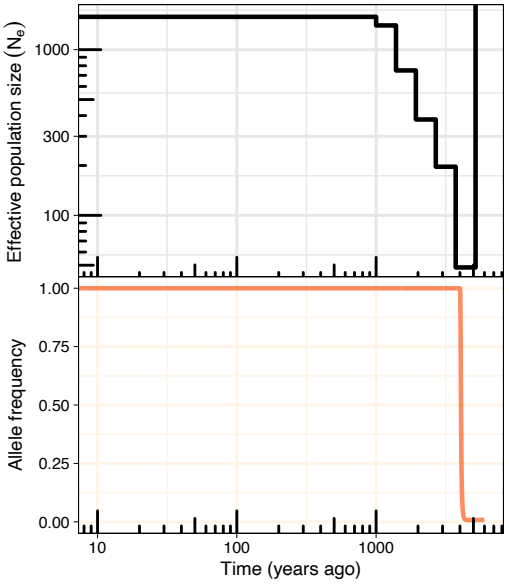
***FLC (FRI^{+}FLC^{+})*. Bottom: Days to bolting (shown on the y-axis). In both panels, dots represent replicates per genotype (n=3), and letters above boxplot represent statistical groups from Kruskal Wallis tests. The central line in the boxplots represents the median; box limits are first and third quartiles. b, Bolting time differences in the complementation test. Each dot represents one replicate per genotype (n=4), and different symbols represent different accessions, with symbols matching between Fogo and F1 for the parental lines. The central line in the boxplots represents the median; box limits are first and third quartiles. P-values are shown for Wilcoxon test. Throughout the figure, *FRI^{+}FLC^{+}* represents the accession Col-0 *FRI*-Sf2, with both functional *FRI* and *FLC; FRI^{+}FLC^{-}* represents the accession Col-0 *FRI*-Sf2 *flc-3*, carrying a functional *FRI* and a non-functional *FLC*; F1_*FRI^{+}FLC^{+}*xFogo represents F1 individuals from crosses between *FRI^{+}FLC^{+}* and Fogo accessions; F1_*FRI^{+}FLC^{-}*xFogo represents F1 individuals from crosses between *FRI^{+}FLC^{-}* and Fogo accessions; F1_Col-0 represents F1 individuals from crosses between *FRI^{+}FLC^{+}* and *FRI^{+}FLC^{-}*.**



Supplementary Figure 13. Marginal genealogical tree estimated in RELATE for FLC 3X (Chr5: 3179333). Individuals are shown across the x-axis with the ancestral and derived carriers coloured in mustard and grey, respectively. The estimated coalescence times are shown on the y-axis. Confidence intervals indicate the 0.025 and 0.975 quantiles of the posterior density of coalescence times and are shown by the vertical black lines on the tree. Red dots represent mapped SNPs on their corresponding branches. Abbreviations: FO, Fogo; MV, Monte Velha; La, Lava; In, Inferno; SA, Santo Antão; and S1-1 accession from Lombo de Figueira subpopulation in Santo Antão (used as the outgroup).



Supplementary Figure 14. Allele frequency trajectories of variants arising in a simulated Fogo colonizing population. Simulated allele frequency trajectories of variants arising in Fogo (starting population size=48) with selection coefficients of 9.2% (the estimate from CLUES with variable N_e). Each grey line represents a simulated trajectory, and the orange line represents the trajectory inferred from CLUES for *FLC* R3X, across selfing rates (left to right: 90%, 95%, 99%).



Supplementary Figure 15. *FLC* 3X appeared in Fogo soon after colonization, and rose in frequency rapidly, consistent with evolutionary rescue. Top: reconstructed change in effective population size over time in Fogo. Bottom: inferred allele frequency trajectory of *FLC* 3X.

References

1. Díaz, S. *et al.* Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *IPBES* (2019).

2. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1918).

3. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nat. Rev. Genet.* **3**, 11–21 (2002).

4. Hancock, A. M., Alkorta-Aranburu, G., Witonsky, D. B. & Di Rienzo, A. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2459–2468 (2010).

5. Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.* 1–13 (2020) doi:10.1038/s41576-020-0250-z.

6. Orr, H. A. & Unckless, R. L. The population genetics of evolutionary rescue. *PLOS Genet.* **10**, e1004551 (2014).

7. Orr, H. A. & Unckless, R. L. Population extinction and the genetics of adaptation. *Am. Nat.* **172**, 160–169 (2008).

8. Orr, H. A. Theories of adaptation: what they do and don't say. *Genetica* **123**, 3–13 (2005).

9. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).

10. Wright, S. Physiological genetics, ecology of populations, and natural selection. *Perspect. Biol. Med.* **3**, 107–151 (1959).

11. Kimura, M. & Ohta, T. On the rate of molecular evolution. *J. Mol. Evol.* **1**, 1–17 (1971).

12. Whitlock, M. C. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* **54**, 1855–1861 (2000).

13. Uecker, H., Otto, S. P., Hermisson, J., Rice, A. E. S. H. & Day, E. T. Evolutionary rescue in structured populations. *Am. Nat.* **183**, E17–E35 (2014).

14. Bell, G. & Gonzalez, A. Evolutionary rescue can prevent extinction following environmental change. *Ecol. Lett.* **12**, 942–948 (2009).

15. Wallace, A. R. On the law which has regulated the introduction of species. **16**, 184–196 (1855).

16. Darwin, C. The Origin of Species by means of natural selection. (J. Murray, 1859).

17. Losos, J. B., Warheitt, K. I. & Schoener, T. W. Adaptive differentiation following experimental island colonization in Anolis lizards. *Nature* **387**, 70–73 (1997).

18. Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).

19. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).

20. Brochmann, C., Rustan, Ø. H., Lobin, W. & Kilian, N. *The endemic vascular plants of the Cape Verde Islands, W Africa*. (Botanical Garden and Museum, Univ. of Oslo, 1997).

21. Romeiras, M. M., Monteiro, F., Duarte, M. C., Schaefer, H. & Carine, M. Patterns of genetic diversity in three plant lineages endemic to the Cape Verde Islands. *AoB PLANTS* **7**, (2015).

22. W. Lobin. The occurrance of *Arabidopsis thaliana* in the Cape Verde Islands. in vol. 20 119–123 (1983).

23. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **114**, 5213 (2017).

24. Pritchard, J. K. & Przeworski, M. Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).

25. Rogers, A. R. How Population Growth Affects Linkage Disequilibrium. *Genetics* **197**, 1329–1341 (2014).

26. Alonso-Blanco, C. *et al.* 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).

27. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

28. Franzke, A., Sharif Samani, B.-R., Neuffer, B., Mummenhoff, K. & Hurka, H. Molecular evidence in Diplotaxis (Brassicaceae) suggests a Quaternary origin of the Cape Verdean flora. *Plant Syst. Evol.* **303**, 467–479 (2017).

29. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919–925 (2014).

30. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genet.* **5**, e1000695 (2009).

31. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016).

32. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

33. Haller, B. C. & Messer, P. W. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol* **36**, 632–637 (2019).

34. Speidel, L. *et al.* Inferring population histories for ancient genomes using genome-wide genealogies. *Mol. Biol. Evol.* (2021) doi:10.1093/molbev/msab174.

35. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231–259 (2006).

36. Tataru, P. & Bataillon, T. polyDFE: inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data. *Methods Mol Biol* **2090**, 125–146 (2020).

37. Wright, S. I. & Andolfatto, P. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu. Rev. Ecol. Evol. Syst.* **39**, 193–213 (2008).

38. Rousselle, M., Mollion, M., Nabholz, B., Bataillon, T. & Galtier, N. Overestimation of the adaptive substitution rate in fluctuating populations. *Biol. Lett.* **14**, 20180055 (2018).

39. Eyre-Walker, A. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**, 2017–2024 (2002).

40. Alonso-Blanco, C. *et al.* Development of an AFLP based linkage map of Ler, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line population: AFLP based linkage map of *Arabidopsis*. *Plant J.* **14**, 259–271 (1998).

41. El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J. M., Raz, V. & Koornneef, M. A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. *Nat. Genet.* **29**, 435–440 (2001).

42. Gazzani, S., Gendall, A. R., Lister, C. & Dean, C. Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. *Plant Physiol* **132**, 1107–1114 (2003).

43. Edwards, K. D., Lynn, J. R., Gyula, P., Nagy, F. & Millar, A. J. Natural allelic variation in the temperature-compensation mechanisms of the *Arabidopsis thaliana* circadian clock. *Genetics* **170**, 387–400 (2005).

44. Kim, T.-S., Wang, L., Kim, Y. J. & Somers, D. E. Compensatory mutations in GI and ZTL may modulate temperature compensation in the circadian clock. *Plant Physiol.* **182**, 1130–1141 (2020).

45. Dunning, F. M., Sun, W., Jansen, K. L., Helft, L. & Bent, A. F. Identification and mutational analysis of *Arabidopsis* FLS2 leucine-rich repeat domain residues that contribute to flagellin perception. *Plant Cell* **19**, 3297–3313 (2007).

46. Marais, D. L. D. *et al.* Variation in MPK12 affects water use efficiency in *Arabidopsis* and reveals a pleiotropic link between guard cell size and ABA response. *PNAS* **111**, 2836–2841 (2014).

47. Kadirjan-Kalbach, D. K. *et al.* Allelic variation in the chloroplast division gene FtsZ2-2 leads to natural variation in chloroplast size. *Plant Physiol.* **181**, 1059–1074 (2019).

48. Li, P. *et al.* Fructose sensitivity is suppressed in *Arabidopsis* by the transcription factor ANAC089 lacking the membrane-bound domain. *PNAS* **108**, 3436–3441 (2011).

49. Alonso-Blanco, C., El-Assal, S. E.-D., Coupland, G. & Koornneef, M. Analysis of natural allelic variation at flowering time loci in the Landsberg ererecta and Cape Verde Islands ecotypes of *Arabidopsis thaliana*. *Genetics* **149**, 749 (1998).

50. McKay, J. K., Richards, J. H. & Mitchell-Olds, T. Genetics of drought adaptation in *Arabidopsis thaliana*: Pleiotropy contributes to genetic correlations among ecological traits. *Mol Ecol* **12**, 1137–1151 (2003).

51. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821–824 (2012).

52. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genet.* **15**, e1008384 (2019).

53. Michaels, S. D. & Amasino, R. M. FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* **11**, 949 (1999).

54. Gomulkiewicz, R. & Holt, R. D. When does evolution by natural selection prevent extinction? *Evolution* **49**, 201–207 (1995).

55. Holt, R. D. & Gomulkiewicz, R. How Does Immigration Influence Local Adaptation? A Reexamination of a Familiar Paradigm. *Am. Nat.* **149**, 563–572 (1997).

56. Gillespie, J. H. Some properties of finite populations experiencing strong selection and weak mutation. *Am. Nat.* **121**, 691–708 (1983).

57. Gillespie, J. H. Molecular evolution over the mutational landscape. *Evolution* **38**, 1116–1129 (1984).

58. Gillespie, J. H. *The causes of molecular evolution*. (Oxford University Press, 1991).

59. Osmond, M. M., Otto, S. P. & Martin, G. Genetic paths to evolutionary rescue and the distribution of fitness effects along them. *Genetics* **214**, 493–510 (2020).

60. Szendro, I. G., Franke, J., de Visser, J. A. G. M. & Krug, J. Predictability of evolution depends nonmonotonically on population size. *Proc. Natl. Acad. Sci.* **110**, 571–576 (2013).

61. Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56**, 1317–1330 (2002).

62. Höllinger, I., Pennings, P. S. & Hermisson, J. Polygenic adaptation: From sweeps to subtle frequency shifts. *PLOS Genet.* **15**, e1008035 (2019).

63. Johanson, U. *et al.* Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290**, 344–347 (2000).

64. Shindo, C. *et al.* Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*. *Plant Physiol* **138**, 1163 (2005).

65. Werner, J. D. *et al.* FRIGIDA-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. *Genetics* **170**, 1197–1207 (2005).

66. Michaels, S. D., He, Y., Scortecci, K. C. & Amasino, R. M. Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. *Proc. Natl. Acad. Sci.* **100**, 10102–10107 (2003).

67. Lempe, J. *et al.* Diversity of flowering responses in wild *Arabidopsis thaliana* strains. *PLoS Genet.* **1**, 109–118 (2005).

68. Méndez-Vigo, B., Picó, F. X., Ramiro, M., Martínez-Zapater, J. M. & Alonso-Blanco, C. Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in *Arabidopsis*. *Plant Physiol.* **157**, 1942–1955 (2011).

69. Schranz, M. E. *et al.* Characterization and effects of the replicated flowering time gene FLC in *Brassica rapa*. *Genetics* **162**, 1457–1468 (2002).

70. Tadege, M. *et al.* Control of flowering time by FLC orthologues in *Brassica napus*. *Plant J. Cell Mol. Biol.* **28**, 545–553 (2001).

71. Guo, Y.-L., Todesco, M., Hagmann, J., Das, S. & Weigel, D. Independent FLC mutations as causes of flowering-time variation in *Arabidopsis thaliana and Capsella rubella*. *Genetics* **192**, 729–739 (2012).

72. Okazaki, K. *et al.* Mapping and characterization of FLC homologs and QTL analysis of flowering time in *Brassica oleracea*. *TAG Theor. Appl. Genet. Theor. Angew. Genet.* **114**, 595–608 (2007).

73. Albani, M. C. *et al.* PEP1 of *Arabis alpina* is encoded by two overlapping genes that contribute to natural genetic variation in perennial flowering. *PLoS Genet.* **8**, e1003130 (2012).

74. Kemi, U. *et al.* Role of vernalization and of duplicated FLOWERING LOCUS C in the perennial *Arabidopsis lyrata*. *New Phytol.* **197**, 323–335 (2013).

75. Lee, C.-R., Hsieh, J.-W., Schranz, M. E. & Mitchell-Olds, T. The functional change and deletion of FLC homologs contribute to the evolution of rapid flowering in *Boechera stricta*. *Front. Plant Sci.* **9**, 1078 (2018).

76. Le Corre, V., Roux, F. & Reboud, X. DNA polymorphism at the FRIGIDA gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. *Mol Biol Evol* **19**, 1261–1271 (2002).

77. Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. & Purugganan, M. D. Epistatic interaction between *Arabidopsis* FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proc Natl Acad Sci USA* **101**, 15670–15675 (2004).

78. Stinchcombe, J. R. *et al.* A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. *Proc Natl Acad Sci USA* **101**, 4712–4717 (2004).

79. Orr, H. A. & Coyne, J. A. The genetics of adaptation: a reassessment. *Am. Nat.* **140**, 725–742 (1992).

80. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6**, 119–127 (2005).

81. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).

82. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).

83. Silander, O. K., Tenaillon, O. & Chao, L. Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS Biol.* **5**, e94 (2007).

84. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).

85. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**, 571–574 (2013).

86. Woods, R. J. *et al.* Second-order selection for evolvability in a large *Escherichia coli* population. *Science* **331**, 1433–1436 (2011).

87. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).

88. Bataillon, T., Zhang, T. & Kassen, R. Cost of adaptation and fitness effects of beneficial mutations in *Pseudomonas fluorescens*. *Genetics* **189**, 939–949 (2011).

89. Brennan, A. C. *et al.* The genetic structure of *Arabidopsis thaliana* in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. *BMC Plant Biol.* **14**, 17 (2014).

90. Fick, S. E. & Hijmans, R. J. WorldClim2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).

91. Antonio Trabucco & Robert J. Zomer. Global aridity index and potential evapo-transpiration (ET0) climate database v2. (2019).

92. Salomé, P. A. *et al.* Genetic architecture of flowering-time variation in *Arabidopsis thaliana*. *Genetics* **188**, 421–433 (2011).

93. Joosen, R. V. L. *et al.* germinator: a software package for high-throughput scoring and curve fitting of *Arabidopsis* seed germination. *Plant J.* **62**, 148–159 (2010).

94. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

95. 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).

96. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

97. Hämälä, T. & Savolainen, O. Genomic patterns of local adaptation under gene flow in *Arabidopsis lyrata*. *Mol Biol Evol* **36**, 2557–2571 (2019).

98. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).

99. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).

100. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

101. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).

102. Zhang, L. & Jiménez-Gómez, J. M. Functional analysis of FRIGIDA using naturally occurring variation in *Arabidopsis thaliana*. *Plant J.* **103**, 154–165 (2020).

103. Sheldon, C. C., Conn, A. B., Dennis, E. S. & Peacock, W. J. Different regulatory regions are required for the vernalization-induced repression of FLOWERING LOCUS C and for the epigenetic maintenance of repression. *Plant Cell* **14**, 2527–2537 (2002).

104. Sung, S. *et al.* Epigenetic maintenance of the vernalized state in *Arabidopsis thaliana* requires LIKE HETEROCHROMATIN PROTEIN 1. *Nat. Genet.* **38**, 706–710 (2006).

105. Bomblies, K. *et al.* Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet* **6**, e1000890 (2010).

106. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, (2008).

107. Warnes, G. gmodels: various R programming tools for model fitting. (2007).

108. Lasky, J. R. *et al.* Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Mol. Ecol.* **21**, 5512–5529 (2012).

109. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).

110. Tang, C. *et al.* The evolution of selfing in *Arabidopsis thaliana*. *Science* **317**, 1070–1072 (2007).

111. Dwyer, K. G. *et al.* Molecular characterization and evolution of self-incompatibility genes in *Arabidopsis thaliana*: the case of the Sc haplotype. *Genetics* **193**, 985–994 (2013).

112. Novikova, P. Y. *et al.* Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

113. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma. Oxf. Engl.* **30**, 1312–1313 (2014).

114. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

115. Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**, 2770–2784 (2015).

116. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).

117. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).

118. Simon, M. *et al.* Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. *Genetics* **178**, 2253–2264 (2008).

119. Keurentjes, J. J. B. *et al.* Development of a near-isogenic line population of *Arabidopsis thaliana* and comparison of mapping power with a recombinant inbred line population. *Genetics* **175**, 891–905 (2007).

120. Alonso-Blanco, C., El-Assal, S. E., Coupland, G. & Koornneef, M. Analysis of natural allelic variation at flowering time loci in the Landsberg erecta and Cape Verde Islands ecotypes of *Arabidopsis thaliana*. *Genetics* **149**, 749–764 (1998).

121. Alonso-Blanco, C. *et al.* Genetic and molecular analyses of natural variation indicate CBF2 as a candidate gene for underlying a freezing tolerance quantitative trait locus in *Arabidopsis*. *Plant Physiol* **139**, 1304–1312 (2005).

122. Alonso-Blanco, C., Vries, H. B., Hanhart, C. J. & Koornneef, M. Natural allelic variation at seed size loci in relation to other life history traits of *Arabidopsis thaliana*. *PNAS* **96**, 4710–4717 (1999).

123. Bandaranayake, C. K., Koumproglou, R., Wang, X. Y., Wilkes, T. & Kearsey, M. J. QTL analysis of morphological and developmental traits in the Ler × Cvi population of *Arabidopsis thaliana*. *Euphytica* **137**, 361–371 (2004).

124. Bentsink, L. *et al.* Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of *Arabidopsis*. *Plant Physiol.* **124**, 1595–1604 (2000).

125. Bentsink, L. *et al.* Natural variation for seed dormancy in *Arabidopsis* is regulated by additive genetic and molecular pathways. *PNAS* **107**, 4264–4269 (2010).

126. Bentsink, L., Yuan, K., Koornneef, M. & Vreugdenhil, D. The genetics of phytate and phosphate accumulation in seeds and leaves of *Arabidopsis thaliana*, using natural variation. *Theor Appl Genet* **106**, 1234–1243 (2003).

127. Borevitz, J. O. *et al.* Quantitative trait loci controlling light and hormone response in two accessions of *Arabidopsis thaliana*. *Genetics* **160**, 683–696 (2002).

128. Botto, J. F., Alonso-Blanco, C., Garzarón, I., Sánchez, R. A. & Casal, J. J. The Cape Verde Islands allele of Cryptochrome 2 enhances cotyledon unfolding in the absence of blue light in *Arabidopsis*. *Plant Physiol.* **133**, 1547–1556 (2003).

129. Buescher, E. *et al.* Natural genetic variation in selected populations of *Arabidopsis thaliana* is associated with ionomic differences. *PLOS ONE* **5**, e11081 (2010).

130. Coneva, V. & Chitwood, D. H. Genetic and developmental basis for increased leaf thickness in the *Arabidopsis* Cvi ecotype. *Front Plant Sci* **9**, (2018).

131. Conte, M., de Simone, S., Simmons, S. J., Ballaré, C. L. & Stapleton, A. E. Chromosomal loci important for cotyledon opening under UV-B in *Arabidopsis thaliana*. *BMC Plant Biol.* **10**, 112 (2010).

132. Darrah, C. *et al.* Analysis of phase of LUCIFERASE expression reveals novel circadian quantitative trait loci in *Arabidopsis*. *Plant Physiol.* **140**, 1464–1474 (2006).

133. DeRose-Wilson, L. & Gaut, B. S. Mapping salinity tolerance during *Arabidopsis thaliana* germination and seedling growth. *PLOS ONE* **6**, e22832 (2011).

134. Fournier-Level, A. *et al.* Paths to selection on life history loci in different natural environments across the native range of *Arabidopsis thaliana*. *Mol Ecol* **22**, 3552–3566 (2013).

135. Gilliland, L. U. *et al.* Genetic basis for natural variation in seed vitamin E levels in *Arabidopsis thaliana*. *PNAS* **103**, 18834–18841 (2006).

136. Hobbs, D. H., Flintham, J. E. & Hills, M. J. Genetic control of storage oil synthesis in seeds of *Arabidopsis*. *Plant Physiol.* **136**, 3341–3349 (2004).

137. Kasulin, L., Agrofoglio, Y. & Botto, J. F. The receptor-like kinase ERECTA contributes to the shade-avoidance syndrome in a background-dependent manner. *Ann Bot* **111**, 811–819 (2013).

138. Keurentjes, J. J. *et al.* Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in *Arabidopsis thaliana*. *Genome Biol.* **9**, R129 (2008).

139. Kliebenstein, D. J., Gershenzon, J. & Mitchell-Olds, T. Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics* **159**, 359–370 (2001).

140. Kliebenstein, D., Pedersen, D., Barker, B. & Mitchell-Olds, T. Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in *Arabidopsis thaliana*. *Genetics* **161**, 325–332 (2002).

141. Kobayashi, Y. *et al.* Amino acid polymorphisms in strictly conserved domains of a P-type ATPase HMA5 are involved in the mechanism of copper tolerance variation in *Arabidopsis*. *Plant Physiol* **148**, 969–980 (2008).

142. Laserna, M. P., Sánchez, R. A. & Botto, J. F. Light-related loci controlling seed germination in Ler x Cvi and Bay-0 x Sha recombinant inbred-line populations of *Arabidopsis thaliana*. *Ann Bot* **102**, 631–642 (2008).

143. Lee, S., Sergeeva, L. I. & Vreugdenhil, D. Quantitative trait loci analysis of hormone levels in *Arabidopsis* roots. *PLOS ONE* **14**, e0219008 (2019).

144. Luquez, V. M. C. *et al.* Quantitative trait loci analysis of leaf and plant longevity in *Arabidopsis thaliana*. *J Exp Bot* **57**, 1363–1372 (2006).

145. Moore, C. R., Gronwall, D. S., Miller, N. D. & Spalding, E. P. Mapping quantitative trait loci affecting *Arabidopsis thaliana* seed morphology features extracted computationally from images. *G3 Bethesda* **3**, 109–118 (2013).

146. Nguyen, T.-P., Keizer, P., Eeuwijk, F. van, Smeekens, S. & Bentsink, L. Natural variation for seed longevity and seed dormancy are negatively correlated in *Arabidopsis*. *Plant Physiol.* **160**, 2083–2092 (2012).

147. Sergeeva, L. I. *et al.* Histochemical analysis reveals organ-specific quantitative trait loci for enzyme activities in *Arabidopsis*. *Plant Physiol* **134**, 237–245 (2004).

148. Sergeeva, L. I. *et al.* Vacuolar invertase regulates elongation of *Arabidopsis thaliana* roots as revealed by QTL and mutant analysis. *PNAS* **103**, 2994–2999 (2006).

149. Snoek, B. L. *et al.* Genetic dissection of morphometric traits reveals that Phytochrome B affects nucleus size and heterochromatin organization in *Arabidopsis thaliana*. *G3 Bethesda* **7**, 2519–2531 (2017).

150. Swarup, K. *et al.* Natural allelic variation identifies new genes in the *Arabidopsis* circadian system. *Plant J.* **20**, 67–77 (1999).

151. Symonds, V. V. *et al.* Mapping quantitative trait loci in multiple populations of *Arabidopsis thaliana* identifies natural allelic Variation for trichome density. *Genetics* **169**, 1649–1658 (2005).

152. Teng, S., Keurentjes, J., Bentsink, L., Koornneef, M. & Smeekens, S. Sucrose-specific induction of anthocyanin biosynthesis in *Arabidopsis* requires the MYB75/PAP1 gene. *Plant Physiol.* **139**, 1840–1852 (2005).

153. Tessadori, F. *et al.* PHYTOCHROME B and HISTONE DEACETYLASE 6 control light-induced chromatin compaction in *Arabidopsis thaliana*. *PLOS Genet.* **5**, e1000638 (2009).

154. Ungerer, M. C., Halldorsdottir, S. S., Modliszewski, J. L., Mackay, T. F. C. & Purugganan, M. D. Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. *Genetics* **160**, 1133–1151 (2002).

155. 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* **65**, 6457–6469 (2014).

156. Vaughn, L. M. & Masson, P. H. A QTL study for regions contributing to *Arabidopsis thaliana* root skewing on tilted surfaces. *G3 Genes Genomes Genet.* **1**, 105–115 (2011).

157. Velázquez, I., Valencia, S., López-Lera, A., de la Peña, A. & Candela, M. Analysis of natural allelic variation in in vitro organogenesis of *Arabidopsis thaliana*. *Euphytica* **137**, 73–79 (2004).

158. Ward, J. K. *et al.* Identification of a major QTL that alters flowering time at elevated [CO2] in *Arabidopsis thaliana*. *PLOS ONE* **7**, e49028 (2012).

159. Waters, B. M. & Grusak, M. A. Quantitative trait locus mapping for seed mineral concentrations in two *Arabidopsis thaliana* recombinant inbred populations. *New Phytol.* **179**, 1033–1047 (2008).

160. Juenger, T. E. *et al.* Identification and characterization of QTL underlying whole-plant physiology in *Arabidopsis thaliana*: δ13C, stomatal conductance and transpiration efficiency. *Plant Cell Environ.* **28**, 697–708 (2005).

161. Juenger, T., Pérez-Pérez, J. M., Bernal, S. & Micol, J. L. Quantitative trait loci mapping of floral and leaf morphology traits in *Arabidopsis thaliana*: evidence for modular genetic architecture. *Evol Dev* **7**, 259–271 (2005).

162. Aranzana, M. J. *et al.* Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* **1**, (2005).

163. Bikard, D. *et al.* Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**, 623–626 (2009).

164. Brachi, B. *et al.* Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLOS Genet.* **6**, e1000940 (2010).

165. Egli, B., Kölling, K., Köhler, C., Zeeman, S. C. & Streb, S. Loss of cytosolic phosphoglucomutase compromises gametophyte development in *Arabidopsis*. *Plant Physiol.* **154**, 1659–1671 (2010).

166. Ehrenreich, I. M., Stafford, P. A. & Purugganan, M. D. The genetic architecture of shoot branching in *Arabidopsis thaliana*: A comparative assessment of candidate gene associations vs. quantitative trait locus mapping. *Genetics* **176**, 1223–1236 (2007).

167. Fulcher, N. *et al.* Genetic architecture of natural variation of telomere length in *Arabidopsis thaliana*. *Genetics* **199**, 625–635 (2015).

168. Kooke, R. *et al.* Genome-wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in *Arabidopsis*. *Plant Physiol.* **170**, 2187–2203 (2016).

169. Leskow, C. C. *et al.* Allelic differences in a vacuolar invertase affect *Arabidopsis* growth at early plant development. *J Exp Bot* **67**, 4091–4103 (2016).

170. Li, W. *et al.* Identification of quantitative trait loci controlling high calcium response in *Arabidopsis thaliana*. *PLOS ONE* **9**, e112511 (2014).

171. Ouibrahim, L. *et al.* Cloning of the *Arabidopsis* rwm1 gene for resistance to Watermelon mosaic virus points to a new function for natural virus resistance genes. *Plant J* **79**, 705–716 (2014).

172. Routaboul, J.-M. *et al.* Metabolite profiling and quantitative genetics of natural variation for flavonoids in *Arabidopsis*. *J Exp Bot* **63**, 3749–3764 (2012).

173. Seedat, N., Dinsdale, A., Ong, E. K. & Gendall, A. R. Acceleration of flowering in *Arabidopsis thaliana* by Cape Verde Islands alleles of FLOWERING H is dependent on the floral promoter FD. *J Exp Bot* **64**, 2767–2778 (2013).

174. Yano, R., Takebayashi, Y., Nambara, E., Kamiya, Y. & Seo, M. Combining association mapping and transcriptomics identify HD2B histone deacetylase as a genetic factor associated with seed dormancy in *Arabidopsis thaliana*. *Plant J* **74**, 815–828 (2013).

175. Yuan, W., Flowers, J. M., Sahraie, D. J. & Purugganan, M. D. Cryptic genetic variation for *Arabidopsis thaliana* seed germination speed in a novel salt stress environment. *G3 Genes Genomes Genet.* **6**, 3129–3138 (2016).

176. Alonso-Blanco, C., Bentsink, L., Hanhart, C. J., Vries, H. B. & Koornneef, M. Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics* **164**, 711–729 (2003).

177. Moore, C. R. *et al.* High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response. *Genetics* **195**, 1077–1086 (2013).

178. O'Neill, C. M. *et al.* Towards the genetic architecture of seed lipid biosynthesis and accumulation in *Arabidopsis thaliana*. *Heredity* **108**, 115–123 (2012).

179. Nguyen, T.-P., Cueff, G., Hegedus, D. D., Rajjou, L. & Bentsink, L. A role for seed storage proteins in *Arabidopsis* seed longevity. *J Exp Bot* **66**, 6399–6413 (2015).

180. Keurentjes, J. J. B. *et al.* Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *PNAS* **104**, 1708–1713 (2007).

181. McKhann, H. I. *et al.* Natural variation in CBF gene sequence, gene expression and freezing tolerance in the Versailles core collection of *Arabidopsis thaliana*. *BMC Plant Biol.* **8**, 105 (2008).

182. Poque, S. *et al.* Allelic variation at the rpv1 locus controls partial resistance to *Plum pox* virus infection in *Arabidopsis thaliana*. *BMC Plant Biol.* **15**, 159 (2015).

183. Signorell, A. DescTools: tools for descriptive statistics. (2020).

184. Fisher, R. A. Statistical methods for research workers. in *Breakthroughs in Statistics: Methodology and Distribution* (eds. Kotz, S. & Johnson, N. L.) 66–70 (Springer, 1992). doi:10.1007/978-1-4612-4380-9_6.

185. Moulos, P. & Hatzis, P. Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Res* **43**, e25–e25 (2015).

186. Mendiburu, F. Agricolae: statistical procedures for agricultural research. *R Package Version* **1**, 1–8 (2010).

187. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

188. Salomé, P. A. *et al.* The recombination landscape in *Arabidopsis thaliana* F2 populations. *Hered. Edinb* **108**, 447–455 (2012).

189. Boggs, N. A., Nasrallah, J. B. & Nasrallah, M. E. Independent S-locus mutations caused self-fertility in *Arabidopsis thaliana*. *PLOS Genet.* **5**, e1000426 (2009).

190. Sherman-Broyles, S. *et al.* S locus genes and the evolution of self-fertility in *Arabidopsis thaliana*. *Plant Cell* **19**, 94–106 (2007).

191. Shimizu, K. K., Shimizu-Inatsugi, R., Tsuchimatsu, T. & Purugganan, M. D. Independent origins of self-compatibility in *Arabidopsis thaliana*. *Mol. Ecol.* **17**, 704–714 (2008).

192. Shimizu, K. K. & Tsuchimatsu, T. Evolution of selfing: recurrent patterns in molecular adaptation. *Annu. Rev. Ecol. Evol. Syst.* **46**, 593–622 (2015).

193. Tsuchimatsu, T. *et al.* Patterns of polymorphism at the self-incompatibility locus in 1,083 *Arabidopsis thaliana* genomes. *Mol. Biol. Evol.* **34**, 1878–1889 (2017).

194. Schiffels, S. & Wang, K. MSMC and MSMC2: the multiple sequentially Markovian coalescent. *Methods Mol Biol* **2090**, 147–166 (2020).

195. Wright, S. Isolation by Distance. *Genetics* **28**, 114–138 (1943).

196. Malécot, M. Les mathématiques de l'hérédité. *Bull. Mens. Société Linn. Lyon* 203 (1948).

197. Loveless, M. D. & Hamrick, J. L. Ecological determinants of genetic structure in plant populations. *Annu. Rev. Ecol. Syst.* **15**, 65–95 (1984).

198. Schmid, K. J. *et al.* Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor. Appl. Genet.* **112**, 1104–1114 (2006).

199. Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLOS Genet.* **6**, e1000843 (2010).

200. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLOS Biol.* **3**, e196 (2005).

201. 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).

202. Coupland, G. FLOWERING LOCUS C isolation and characterization: two articles that opened many doors. *Plant Cell* **31**, 1190 (2019).

203. Salathia, N. *et al.* FLOWERING LOCUS C -dependent and -independent regulation of the circadian clock by the autonomous and vernalization pathways. *BMC Plant Biol* **6**, 10 (2006).

204. Cho, L.-H., Yoon, J. & An, G. The control of flowering time by environmental factors. *Plant J.* **90**, 708–719 (2017).

205. Bloomer, R. H. & Dean, C. Fine-tuning timing: natural variation informs the mechanistic basis of the switch to flowering in *Arabidopsis thaliana*. *J. Exp. Bot.* **68**, 5439–5452 (2017).

206. Mouradov, A., Cremer, F. & Coupland, G. Control of flowering time: interacting pathways as a basis for diversity. *Plant Cell* **14**, S111–S130 (2002).

207. He, Y., Chen, T. & Zeng, X. Genetic and epigenetic understanding of the seasonal timing of flowering. *Plant Commun.* **1**, (2020).

208. Korkuć, P., Schippers, J. H. M. & Walther, D. Characterization and identification of cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism information. *Plant Physiol.* **164**, 181–200 (2014).

209. Schläppi, M. RNA levels and activity of FLOWERING LOCUS C are modified in mixed genetic backgrounds of *Arabidopsis thaliana*. *Int. J. Plant Sci.* **162**, 527–537 (2001).

CHAPTER II

Population history impacts the genetic architecture of flowering time after colonization of a novel environment

Célia Neto[1], Angela M. Hancock[1*]

[1] Max Planck Institute for Plant Breeding Research, Cologne, Germany.

Abstract

Theory predicts that the age of a population and its long-term effective size (N_e) should influence trait architecture. We contrast the genetic architecture of flowering time in young, small N_e *Arabidopsis* lineages with their much older, larger N_e progenitor population. We use a simple yet powerful system in which island populations were formed recently by long-range colonization of the Cape Verde Islands (CVI, approx. 5 kya) from a North African progenitor population represented by the modern Moroccan population (coalescence at approx. 1 mya). We find that polygenicity is severely reduced in the colonizing populations relative to the Moroccan population. Further, effect sizes of trait-associated loci are exponentially distributed in the island populations and uniformly distributed in the Moroccan population, consistent with fitness measures showing evidence for directional selection in the islands. In addition to the major effect variants *FRI* K232X and *FLC* R3X, we identify candidate loci and variants from core flowering time pathways as well as those that indirectly affect flowering time, including nutrient processing and light sensing. At *ATX2* we find a nonsynonymous variant associated with reduced *FLC* expression and a 6-day reduction in flowering time. Surprisingly we find no effect of the well-known Cvi-0-EDI (*CRY2* V367M) variant in the natural population. Instead, we find evidence that the effect of this variant on flowering time is dependent on genetic background, indicating an epistatic effect. Our results provide a particularly clear empirical example of the effect of demographic history has on trait architecture.

Introduction

Understanding how organisms adapt to new environments is a key goal of evolutionary biology. Since Darwin and Wallace – and the origin of evolutionary theory – there has been a debate over the relative importance of large vs small effects in evolution. This began with references to the geological concepts of Gradualism and Saltation. With the integration of genetics into evolutionary theory, this morphed into a caustic debate between 'Mendelians', who saw a role for 'sports' or large effect loci in trait evolution, and the 'Biometricians', who were convinced that Mendel's discoveries were irrelevant for complex traits [1]. This debate resolved with the introduction of Fisher's infinitesimal theory, which states that an effectively infinite number of loci, each with a small effect, could combine to produce the quantitative trait variation observed in population data [2]. However, this 'infinitesimal model' still did not explain many of the patterns observed in nature, where large, Mendelian effects do seem to play an important role in adaptation [3–6].

Building on Fisher's model and subsequent theoretical advancements, Orr formulated a new model of the genetics of adaptation that approximates theoretical and empirical data [7]. Orr's model considers that, after a sudden change in the environment (e.g., following a major environmental perturbation or after colonization of a new habitat), a small, mutation-limited population will move towards the new fitness optimum through an 'adaptive walk'. The walk consists of a series of fixation events of beneficial mutations with an exponential distribution of effect sizes. At the beginning of the walk, few mutations, each of relatively large effect, will move the population towards the new optimum, followed by many of relatively small effect later [8].

To contribute to adaptation, mutations must enter the population, be beneficial, and escape stochastic loss by genetic drift. In large populations, multiple beneficial mutations may appear simultaneously in different lineages due to the high mutational input, or, since many variants are available at any given time, mutations can become beneficial in the new environment after a change in selection pressure. In small populations, however, genetic drift is stronger and mutational input is reduced. As a result, either no beneficial mutation may arise, or sub-optimal beneficial mutations may be used in adaptation [9–15]. Although, theory predicts that in populations which follow a strong selection – weak mutation regime of adaptation, selection will be strong enough to prevent the stochastic loss of mutations. Under this model, the first steps of adaptation are likely to occur through few large effect mutations that overcome genetic drift and rise in frequency without interference [7,8,11,16–19].

Flowering time is an excellent model trait for connecting ecological and evolutionary factors with their quantitative genetic basis. The timing of reproduction is a crucial component of any organism's life cycle, and in nature it can be a primary determinant of an individual's lifetime fitness [20,21]. In plants, flowering time has been shown to be an important adaptive trait – examples include *Lythrum salicaria* [22], *Mimulus guttatus* [20], *Brassica rapa* [23] and *Arabidopsis thaliana* [24–26] – and it strongly depends on the environment. Flowering too early, before an individual has had time to accumulate sufficient resources, could result in limited capacity for seed production, while flowering too late might cause the plant to reach the end of the growing season without any descendants [27].

Extensive knowledge of the molecular pathways that underlie flowering time determination provides a valuable resource for studies of the genetic factors that contribute to flowering time variation in natural populations. External stimuli are integrated with endogenous signals through a complex network of several genetically defined pathways to directly and indirectly impact flowering time [28,29]. In *A. thaliana*, dozens of genes distributed across several pathways (including vernalization (i.e., cold requirement to flower), circadian clock/photoperiod, autonomous, light-sensing, hormones biosynthesis and signaling) are known to be involved in the control of flowering time, revealing the polygenic basis of this trait [28–32]. However, almost three-quarters of flowering time variation among Eurasian accessions is dependent on two genes alone, *FRIGIDA* (*FRI*) and *Flowering Locus C* (*FLC*), both central genes in the vernalization pathway [33–38]. Therefore, and although there are many genetic factors that contribute to flowering time determination, few major effect variants can play a large role in trait variation, accompanied by many others with smaller effects.

Here we make use of a well-defined island colonization event – *Arabidopsis* in the Cape Verde (CVI) archipelago – to examine the effects of colonization on genetic architecture of an ecologically and evolutionarily important trait in plants: flowering time. The archipelago, which lies at the geographic and climatic edge of the species distribution, has been represented for the past 40 years by a single accession – Cvi-0. Due to its distinct origin, far outside the core range of *A. thaliana*, Cvi-0 has been included in many phenotypic studies across panels of natural accessions and as a parental line in recombinant inbred (RIL) and a near-isogenic (NIL) population over the years. As a result, extensive information is available for the Cvi-0 accession, including large-scale QTL and in some cases, specific candidate genes or even functional variants based on follow-up fine-mapping and functional validation. This makes it an excellent model for studying the genetic architecture of quantitative traits.

For the flowering time trait, there have been several interesting findings in mapping populations that use Cvi-0. Two of the largest effect QTL identified – EDI and FLF [39]– have been carefully and extensively examined. While the former has been fine-mapped to a Cvi-0-specific dominant gain of function mutation at *Cryptochrome 2* (*CRY2* V367M) that accelerates flowering [40], the latter remains a mystery. A possible candidate gene at this locus is *FLC*, for which the Cvi-0 allele would delay flowering [39]. In fact, FLC mRNA levels in Cvi-0 are surprisingly high for such an early flowering accession [34,41], but no candidate variant for this intriguing behavior has been identified. By contrast, and using natural variation panels, a major effect loss of function mutation in *FRI* (K232X), private to CVI and decreasing flowering time, has been functionally validated and shown to affect fitness [35,41,42].

However, deeper examination of Cvi-0 complex traits and their genetic architecture in the ecological context of the natural population has not been possible due to the lack of a population samples and information about their native environment. Here, we make use of a large panel of natural accessions [42] that we sampled across the native geographical distribution in Cape Verde, to bridge this gap.

Arabidopsis thaliana is found in two Cape Verde islands – Santo Antão and Fogo, which were colonized in the last 5-7 ky by long-range dispersal from a North African progenitor (Fig. 1, Supplementary Fig. 1), to which the modern Moroccan population is genetically most similar [42]. The Moroccan population contains high levels of genetic diversity and a large long-term effective population size (N_e) compared to CVI [42,43]. θ_W estimates in the Moroccan population are 73.3 and 62.3 times higher than in the Santo Antão and Fogo, respectively ($\theta_{W\ Santo\ Antão} = 7.59x10^{-5}$, $\theta_{W\ Fogo} = 8.93x10^{-5}$, $\theta_{W\ Morocco} = 5.56x10^{-3}$), and N_e based on coalescence reconstruction is also 26.6 times greater in the mainland population ($N_{e\ Morocco}$ = 325K, $N_{e\ CVI}$ = 12.2K) [42,43]. The near-complete bottlenecks during colonization eliminated nearly all (99.9%) of the shared variation with the continent, producing phylogenetically distinct populations (incipient species) in each of the two islands [42]. Upon colonization, CVI *Arabidopsis* experienced a sudden shift in several environmental factors, including increased precipitation seasonality, increased aridity, reduced growing season length, reduced photoperiod and changes in edaphic factors [42]. We previously showed that under CVI simulated conditions flowering time was strongly reduced in CVI populations compared to Morocco (Supplementary Fig. 1) and that this reduction increased fitness [42]. This phenotypic shift was largely mediated by loss of function of two major genes in the vernalization pathway: *FLC* 3X, which is fixed in Fogo, and *FRI* 232X, which segregates in Santo Antão. However, as a complex trait, the basis of

flowering time is expected to be polygenic, with contributions from many loci genome-wide. Here, we use variation in the natural populations to examine the polygenic architecture of flowering time in the young CVI populations and to compare it to trait architecture in the much older Moroccan outgroup population.

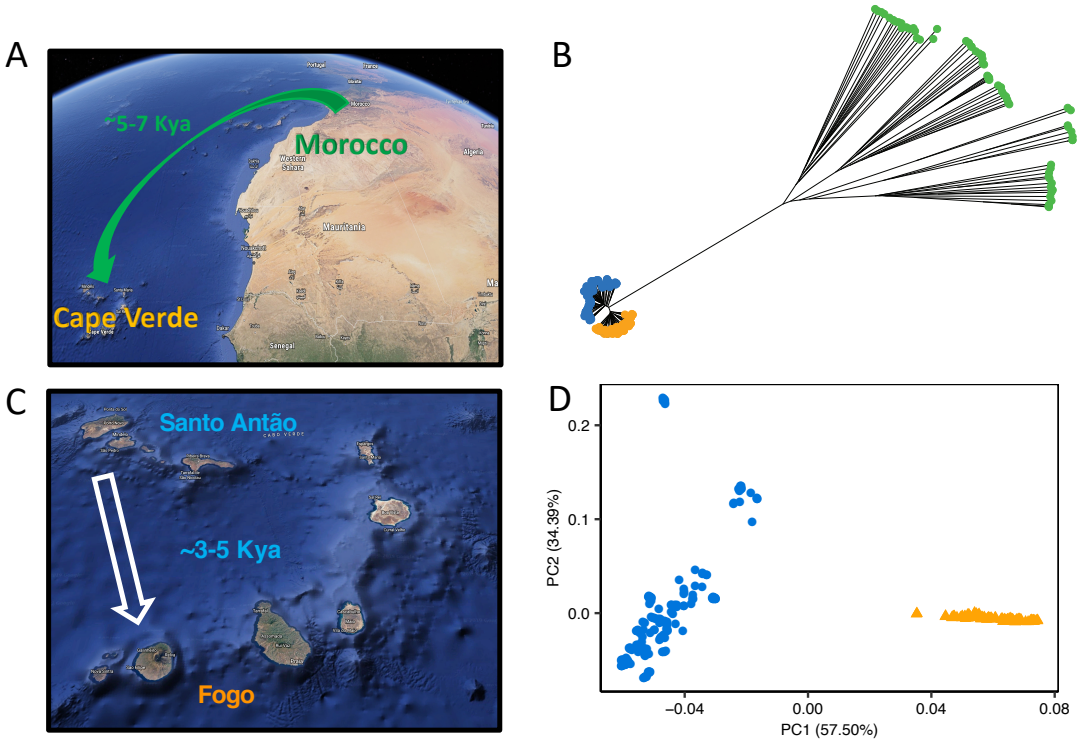


Figure 1. Geographical location and population history of CVI *Arabidopsis*. A. Schematic of the colonization of the Cape Verde archipelago 5-7 kya from a North African population, to which the modern Moroccan population is genetically the closest. B. Neighbour-joining tree showing the relationship between Moroccan (green dots) and Cape Verdean individuals (Santo Antão in blue, and Fogo in orange). C. Schematic of the colonization of Fogo from Santo Antão 3-5 kya. D. Principal component analysis (PCA) showing the differentiation between Santo Antão (blue dots) and Fogo individuals (orange triangles).

Results

1. Time to flowering is reduced in CVI relative to Morocco

We expected directional selection may have acted on flowering time in CVI due to the reduction in growing season length relative to Morocco. To test this, we grew population samples from the archipelago (Santo Antão and Fogo) together with those from Morocco under CVI and Moroccan simulated conditions, which tracked daily daylength and precipitation, as well as hourly temperature and humidity records (Supplementary Fig. 2). We scored bolting time as the number of days from germination to differentiation of the inflorescence and used it as the measure for flowering time. Both broad-sense heritability (based on repeatability across replicates) and narrow-sense heritability (based on the proportion of variance in phenotypes explained by relatedness, PVE; [44]) were high in all populations in both conditions (

Table 1), implying that a large proportion of the variation in flowering time is genetic.

Under both CVI and Moroccan conditions, CVI accessions flowered earlier than Moroccans on average, with the Fogo population being the fastest (Fig. 2, Table 1, Mann-Whitney test (MW), CVI conditions: Mor-CVI: W = 131959, p-value < $2.2x10^{-16}$; Moroccan conditions: Mor-CVI: W = 105190, p-value < $2.2x10^{-16}$). Most Moroccan individuals never bolted or flowered under CVI simulated conditions. One potential reason for this could be that these accessions require a period of cold (vernalization) to induce flowering, consistent with previous observations [45]. CVI individuals flowered earlier under CVI simulated conditions than under simulated Moroccan conditions (MW, mean under CVI conditions = 33.26 days, mean under Moroccan conditions = 42.33 days, W = 8923.5, p-value < $2.2x10^{-16}$), since the conditions were more permissive and the strong selective pressures from Cape Verde were absent. These results suggest that directional selection acted on *Arabidopsis* in Cape Verde to reduce flowering time in the shorter and warmer growing season.

Table 1. Summary statistics for flowering time under CVI and Moroccan simulated conditions.

Condition	Population	n[1]	mean[2]	sd[3]	Repeatability[4]	PVE[5] (95% CI) [6]
CVI conditions	Morocco	62	60.37	11.07	88.67	99.96 (99.8-99.99)
	SA	174	36.17	15.07	95.65	98.3 (97.6 – 98.8)
	Fogo	129	29.05	5.33	80.49	95.2 (92.5 – 97.6)
Moroccan conditions	Morocco	62	62.05	13.20	93.53	93.8 (78.1-99.8)
	SA	174	42.21	6.70	88.44	87.90 (79.1-94.4)
	Fogo	129	43.65	5.08	55.45	45.12 (52.4-87.3)

[1] number of genotypes assessed per condition and population

[2] average bolting time

[3] standard deviation

[4] broad-sense heritability calculated across replicates

[5] percentage of phenotypic variance explained by relatedness (narrow-sense heritability)

[6] 95% confidence interval across 10 runs in BSLMM

We next assessed whether the phenotypic shift observed was caused by genetic variation already present in the colonizing population, and therefore fixed on the archipelago, or by *de novo* island-specific variation. Under the assumption that most segregating variants would have recessive effects, these would be complemented in crosses between islands, and inter-island F1s would display later flowering times than the island populations. To test this, we produced intra- and inter-islands crosses and assessed whether there were significant differences between the resulting F1 individuals and the parental lines. We observed that inter-island F1 hybrids flowered later than the parental lines

(Fig. 2), implying that different loci underlie flowering time reduction on the two islands. The inter-island F1s also recapitulated the ancestral late flowering phenotype, suggesting that recessive variants causing early flowering arose *de novo* on the islands. The Fogo intra-island F1s showed uniformly early flowering with low phenotypic variance (mean=29.5, sd=1.8 days). This is consistent with results in the natural population [42] and indicates that a large effect variant or variants might be fixed or at high frequencies on the population. In contrast, the Santo Antão intra-island F1s presented two modes on the phenotypic distribution, demonstrating that more than one segregating variant may affect flowering time on this population.

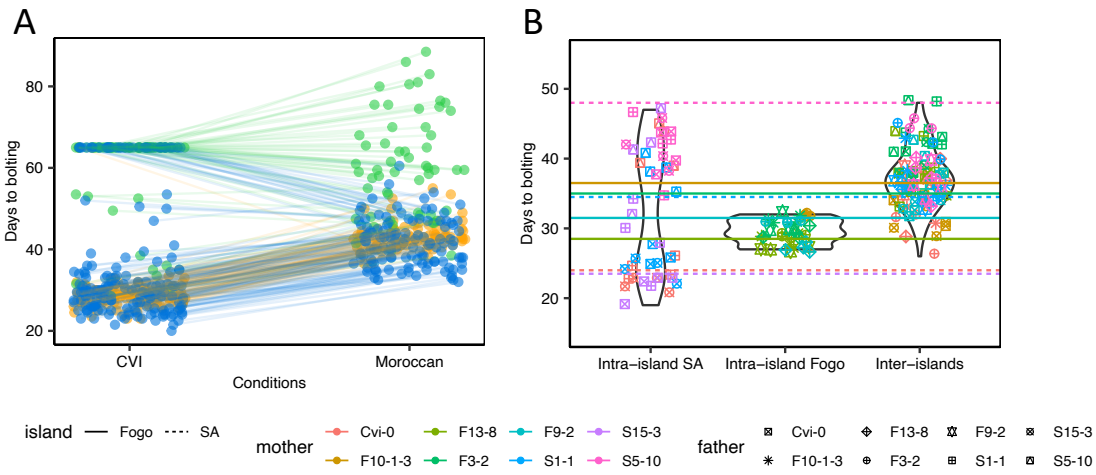


Figure 2. Flowering time in Cape Verde. A. Bolting time in days (y-axis) under both simulated conditions (x-axis). Each dot represents the median across replicates per genotype, with Moroccan individuals in green, Santo Antão in blue and Fogo in orange. Under CVI simulated conditions, individuals that did not bolt by the end of the experiment were scored with 65 days. B. Bolting time in days (y-axis) per cross (x-axis). Each dot represents one F1 hybrid colored by the parental line used as the mother in the cross and shaped by the parental line used as the father. The horizontal lines show average bolting time across replicates of each parental line colored by genotype and with line type matching the island of origin.

Our results are in agreement with previous findings that showed that two *de novo* independent large effect mutations are responsible for a large proportion of flowering time reduction in the two CV islands: *FRI* K232X, which is segregating in Santo Antão at 77% frequency and *FLC* R3X, which is fixed in Fogo. These are concordant with the inferred near-complete colonization bottleneck, which erased (almost) all pre-existing variation and the lack of shared variation with the mainland [42].

2. Simpler genetic architecture of flowering time in CVI compared to Morocco

The CVI system provides the rare possibility to directly compare genetic architecture between young, recently formed populations with low long-term N_e and their outgroup population, where long-term N_e is much higher. Here, we use this system to investigate how the architecture of flowering time is affected by population history.

We used two complementary approaches to infer the genetic architecture of flowering time, mainly focusing on variation with moderate effects. In CVI, this variation would correspond to the variance not explained by the two large effect mutations previously identified (*FRI* K232X in Santo Antão and *FLC* R3X in Fogo [42]). Since *FRI* K232X still segregates in Santo Antão, this variant was included in the model as a covariate to account for its effect and increase power to identify additional potentially causative loci whenever possible.

The first method we used, BSLMM, creates a Bayesian sparse linear mixed model and estimates trait architecture assuming a mixture of large effects and small to infinitesimal effects [44]. Since this method does not allow for covariates, the inferences in Santo Antão will include the large effect *FRI* K232X.

The second method we used – LMM, a standard linear mixed model – estimates the genetic architecture of a trait, assuming that every genotyped variant affects the trait, with effect sizes normally distributed [46]. We used two approaches to estimate the number of independent loci underlying the trait. Both approaches use linkage disequilibrium (LD) between pairs of markers to remove redundancy (Table 2): clumping and the local score [47]. The local score method focus on small effect loci, which are detected by aggregating association signals based on LD between markers within a genomic region. P-values are then turned into scores, defined as the maximum of a Lindley process over the sequence of scores [48]. The clumping strategy forms clumps of markers around central 'index variants' based on the p-value of association, their physical distance and statistical correlation.

Both statistical approaches inferred that many more loci were involved in flowering time in the Moroccan population than on the islands, and this was true under both simulated conditions (CVI and Moroccan; Fig. 3, Table 2, Supplementary Fig. 3). Even though we estimated different numbers of loci across methods, a general trend was observed. Our results suggest a more polygenic architecture for flowering time in the continent and a more oligogenic architecture on the archipelago, implying that

the genetic architecture of this trait was simpler in the younger populations with smaller long-term N_e (Santo Antão and Fogo) compared to the older Moroccan population with much larger long-term N_e.

Table 2. Number of variants affecting flowering time per population under the two simulated conditions per method used.

	CVI conditions			Moroccan conditions		
	BSLMM[1]	Clump	LocalScore	BSLMM[1]	Clump	LocalScore
Morocco	39	6666	353	58	2775	49
Santo Antão	10	104	6	10	57	3
Fogo	3	159	26	15	101	5

[1] Bayesian sparse linear mixed model

3. Genetic bases of flowering time in the natural populations

We next investigated specific candidate loci based on genome-wide association mapping (GWAS). We conducted GWAS using a linear mixed model approach that controls for population history through the inclusion of a relatedness matrix [46] to identify peaks of association. We further narrowed candidate regions by correcting for LD by employing the local score approach [47] (Fig. 3, Supplementary Fig. 3).

3.1. Santo Antão

In Santo Antão, we included the previously identified *FRI* K232X SNP as a covariate in the model to improve our power to identify more moderate effect loci. Using the local score approach, we identified five significantly associated regions (Fig. 3) containing seven genes (Supplementary Table 1). Among these, *NRT1*, *AT59*, *HKT1*, and two transposable elements (AT1G45035 and AT2G12770) were the strongest candidates to influence flowering time. *NRT1* (AT1G12110) is a particularly good candidate, since it encodes a nitrate transporter that contributes to nitrate, cadmium, amino acid, and

auxin uptake, and affects stomatal opening, germination, and flowering time via interaction with *FLC* in a nitrate-independent manner [49–53].

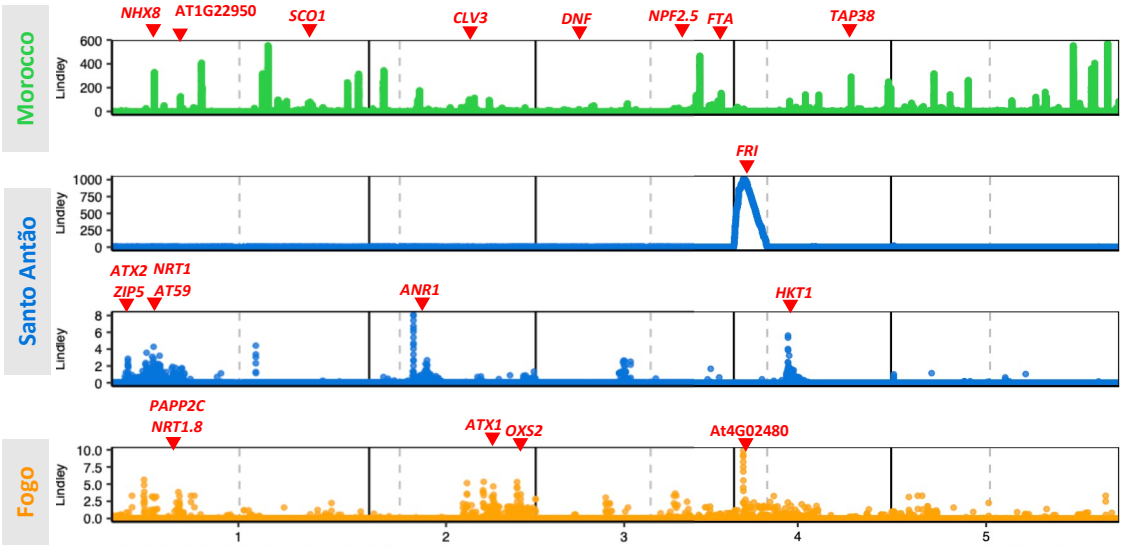


Figure 3. Genetic bases of flowering time. Manhattan plots for the three populations under CVI simulated conditions, with chromosome position on the x-axis and Lindley score from the local score approach on the y-axis. Candidate genes in association are shown by the red arrows.

As a complementary approach, we iteratively added top markers as covariates, in addition to *FRI* K232X. Using this method, we again identified *AT59* and *NRT1* (as the local score method), and also found signals of association for four additional genes: *ANR1*, *ATX2, ZIP5* and *ANAC036* (Supplementary Table 1). Among these, *ANR1* (AT2G14210) and *ATX2* (AT1G05830) were the strongest candidates because their functions are directly linked to the flowering time pathway. *ANR1* is a MADS-box transcription factor that interacts with *SOC1* (a core flowering time gene) and positively regulates lateral root growth in response to nitrate deprivation in an *NRT1*-dependent manner [54,55].

ATX2 is an H3K4-specific methyltransferase that affects flowering time in a *FRI*-dependent way by controlling levels of FLC mRNA. *atx2* mutants show decreased levels of FLC mRNA and reduced flowering time, with a stronger effect on a *FRI* functional background [56–58]. In the natural population, we identified a missense variant in this gene, L125F, that segregates at 66.67% frequency and that, under simulated CVI conditions and after correcting for *FRI* K232X, decreases flowering time by 4 days (LMM in GWAS, effect size = -3.84 days, p-value = 8.88×10^{-3}). Because the effect of *ATX2* is dependent

on *FRI* [56], we looked at all possible allelic combinations between these two loci and how they affect flowering time and FLC mRNA levels in a CVI context. We found no flowering time effect of *ATX2* L125F in the *FRI* functional background but a strong effect on the derived background, reducing flowering time by about 6 days (linear model (LM) and Tukey test, *FRI-ATX2*, Anc-Der/Anc-Anc: diff= -1.89, p-value=0.937, Der-Der/Der-Anc: diff= -5.43, p-value=4.82×10^{-10}). *ATX2* L125F was also responsible for a 9.7% decrease in FLC mRNA levels in a non-functional *FRI* background but no effect was detected in the functional background (LM and Tukey test, *FRI-ATX2*, Anc-Der/Anc-Anc: diff=-0.14, p-value = 0.994, Der-Der/Der-Anc: diff= -0.56, p-value = 0.0009; Supplementary Fig. 4). Our results are in agreement with previous work showing a strong effect of *ATX2* mutants on the number of leaves at bolting (as a proxy for flowering time) but contradict expectations related to the *FRI* background effect [56]. In Santo Antão, no effect of *ATX2* was observed in the functional *FRI* background which could be explained by epistatic effects in the different genetic backgrounds (CVI vs Col-0), or by an underpowered statistical analysis, since only one individual in the natural population presents both ancestral alleles.

Previous work based on RILs and NILs between Cvi-0 and Ler-0 have identified a large effect QTL (EDI, *early day-length insensitive*) on top of chromosome 1 controlling flowering time [39]. Further work has linked this QTL to a Cvi-0-specific functional variant at Cry*ptochome 2* (*CRY2* V367M, [40]), a blue-light photoreceptor whose enhanced activity could function as an activator of downstream floral pathway integrators, therefore bypassing the repression conferred by the high FLC mRNA levels observed in Cvi-0. The missense gain of function variant decreases flowering time (scored as a reduction in total leaf number) with a more pronounced response under short-days [39,40].

Since Cvi-0 originated in Santo Antão [42], we considered *CRY2* V367M in the context of the natural population. Using GWAS, no effect was detected (LMM in GWAS, p-value = 0.945; Fig. 3, Supplementary Fig. 5) under simulated CVI conditions, even after correction for the large effect *FRI* 232X. This may be due to *CRY2* V367M high frequency (90%) and strong association with population structure (Supplementary Fig. 6), being fixed in almost all populations, including the most ancestral Cova de Paúl.

Next, we examined the segregation of *FRI* K232X and *CRY2* V367M to examine potential interaction effects in the CVI population. We found no plants that carried both the ancestral *FRI* and *CRY2* variants and therefore we could not assess the effect of *CRY2* alone in the ancestral *FRI* background. In the derived non-functional *FRI* background, the *CRY2* derived allele showed once again

no effect (LM and Tukey, *FRI-CRY2*, Der-Der/Der-Anc: diff= -1.87, p-value = 0.31; Supplementary Fig. 5).

Since previously published work on the Cvi-0 *CRY2* allele was done in European genetic backgrounds [40,59–65] (Ler-0 and Col-0), the discrepancies observed with our results could potentially be explained by differences in the genetic background. To distinguish between a lack of signal due to genetic background effects (GxG) versus environmental conditions (GxE), we examined two NILs that carried the CVI-EDI locus in a Ler-0 background (LCN1-2.5, LCN1-2.8; [59]) grown under simulated CVI conditions. We observed a reduction in flowering time compared to Ler-0, implying that the Cvi-0 *CRY2* region alone in the Ler-0 genetic background was able to reduce flowering time also under CVI simulated conditions (MW, LCN1-2.5: W = 0, p-value = 0.01554, LCN1-2.8: W = 0, p-value = 0.01554; Fig. 4, Supplementary Fig. 5). These results suggest that the effect of *CRY2* V367M may be due to interactions with other variants that differentiate the natural CVI populations relative to the Ler-0 genetic background. Overall, our results demonstrate that the *CRY2* V367M effect is background-dependent, a result that has previously been observed [64]. Even though we were able to recapitulate the previously shown effect of *CRY2* in the Ler-0 background, even under CVI simulated conditions, we were not able to detect its effect in flowering time in the CVI genetic background.

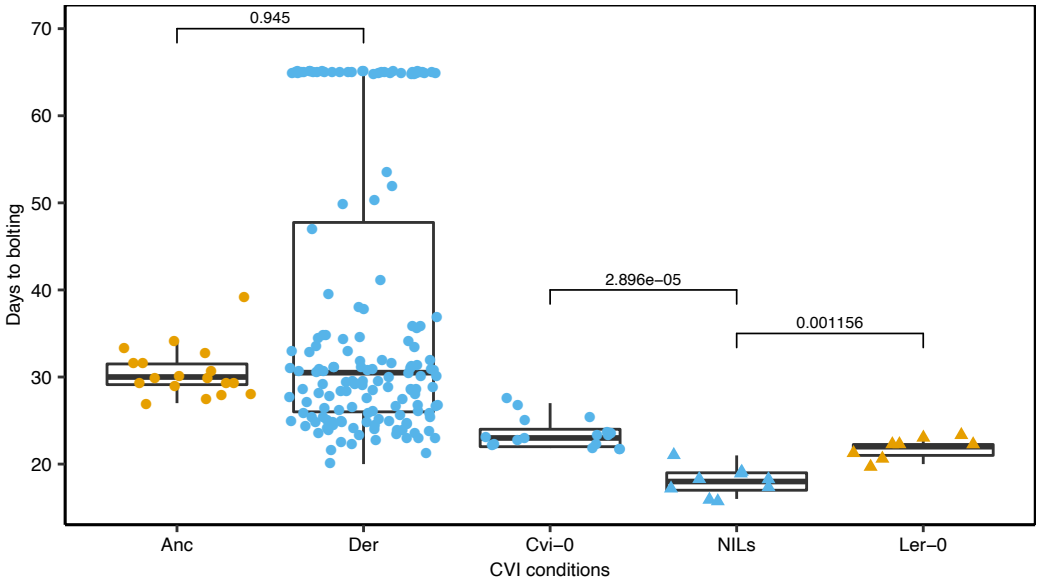


Figure 4. *CRY2* V367M effect under CVI simulated conditions. Boxplot of bolting time in days (y-axis) per genotypes (x-axis). *Anc* and *Der* refer respectively to individuals carrying the ancestral

and the derived *CRY2* V367M allele in the natural population of Santo Antão (each dot represents the median across replicates per genotype). *Cvi-0* and *Ler-0* refer to these genotypes, and *NILs* to two Cvi-0 x Ler-0 NILs with the EDI locus (each dot represents one replicate). Orange and blue represent carriers of ancestral and derived *CRY2* alleles, respectively. Circles represent the CVI genetic background and triangles the Ler-0 genetic background. P-values shown from the linear models.

3.2. Fogo

We employed the same strategy as above to find loci that contribute to flowering time in Fogo on addition to *FLC* R3X. We found signals of association upstream of genes in the flowering time pathway, such as *PAPP2C* (AT1G22280) – a phosphatase that interacts with phytochromes A and B [66] – and *OXS2* (AT2G41900) – a stress-induced transcription factor which interacts with *SOC1*, *FT* and *FD* (three proteins known to induce floral transition) [67]. The strongest association signal however underlaid AT4G02480, an AAA-type ATPase protein. Although its function is unknown, AT4G02480 has been shown to interact with *FKF* and *GI*, which subsequently regulate CO protein stability for photoperiodic control of flowering [68].

Interestingly, we also identified associations in two other candidate genes that fall into gene families that contained members that were associated with flowering time in Santo Antão: *ATX1* (AT2G31650) and *NPF5.15* (AT1G22570), which contains a with premature stop codon predicted to truncate the protein. The former is a histone-lysine N-methyltransferase involved in the formation, placement and identity of floral organs, and the epigenetic control of *FLC*. It also affects seed germination, stomatal aperture, water loss and sensitivity to dehydration stress [56,69]. *NPF5.15* belongs to the NRT1/PTR protein family which is composed by nitrate, di/tri-peptide and hormone transporters [70].

4. Flowering time on the archipelago follows Orr's model of adaptation

Theory predicts that mutation-limited populations facing an abrupt environmental change that leads to displacement from its optimum – as in the colonization of the Cape Verde archipelago from North Africa – will adapt through an 'adaptive walk' towards the new optimum [7,8]. In such cases, the 'adaptive walk' is expected to consist in larger effect mutations in the beginning of the walk and smaller ones later [7,8,71]. We can use the information gained about the genetic associations with flowering time

in the natural populations of CVI to further examine the genetic architecture of this trait across populations and to make inferences about its evolutionary history.

4.1. Effect size distributions follow an exponential distribution on the archipelago

We investigated the distribution of effect sizes in each population (Fig. 5) under CVI simulated conditions, using one representative SNP per candidate locus, identified using the local score approach. Then, we calculated the absolute effect sizes of each of these representative SNPs. We manually added information about *FLC* R3X, which is fixed in Fogo. In the Moroccan population, the 297 candidate loci identified had a median effect size of 14.09 days (range: 6.15 – 25.90), while in Santo Antão the median effect size of the 5 candidate loci was 5.90 days (range: 3.83 – 35.27) and in Fogo the 26 candidate loci had a median of 5.94 days (range: 0.35 – 27).

To investigate whether the effect sizes fit a uniform distribution, as predicted under Fisher's infinitesimal model [2], or an exponential distribution, as predicted by Orr's model of adaptation [7], we compared the fit of these two classes of distributions to the effect size distribution per population using a maximum likelihood method. The distribution of the effect size in Morocco best fit a uniform distribution (AIC: exponential: 21.87.02, uniform: 1775.99), while in Santo Antão and Fogo, an exponential distribution had a better fit (AIC $_{\text{Santo Antão}}$: exponential: 36.00, uniform: 38.48; AIC $_{\text{Fogo}}$: exponential: 141.97, uniform: 168.15). While the Moroccan population fit Fisher's infinitesimal model, the pattern for both islands is more consistent with Orr's model, in which few mutations of relatively large effect together with many of relatively small effect underlie adaptation to a new distant optimum. Similarly, the distribution of effect sizes estimated based on QTL mapping of Cvi-0 x Ler-0 RILs (across studies, including very different growing conditions and traits as proxies for flowering time; Fig. 5) follows an exponential distribution (AIC: uniform: 119.76, exponential: 100.60).

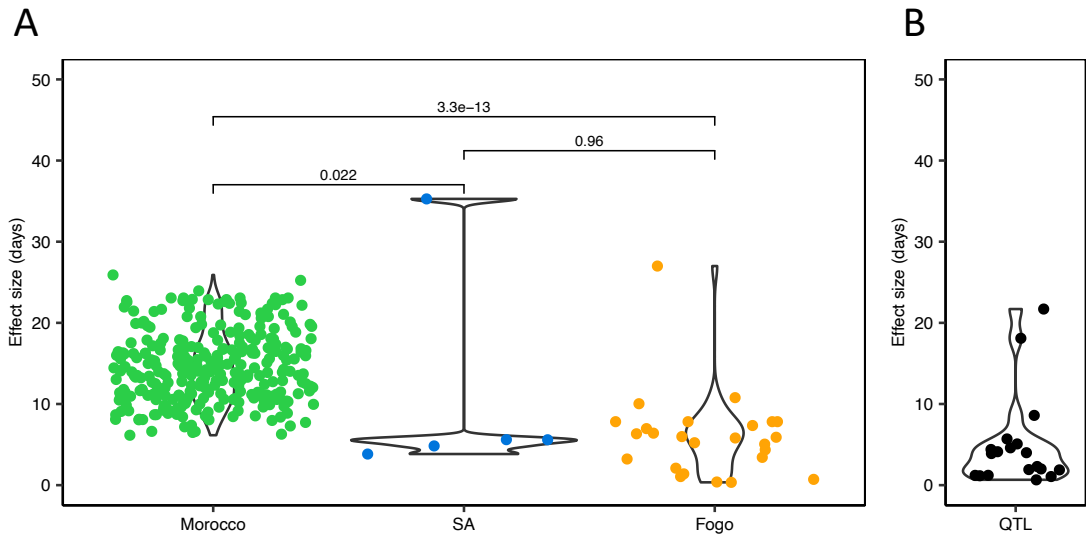


Figure 5. Effect size distributions of flowering time-associated variants. Absolute effect size in days (y-axis) per population (x-axis). A. Each dot represents one SNP per candidate loci identified with the local score approach in the natural populations grown under CVI simulated conditions. *FLC* R3X was added to the plot for completeness B. Each dot represents one QTL previously identified in Cvi-0 x Ler-0 RILs. P-values shown from MU test.

Our results on the natural populations fit theoretical expectations of effect size distributions in populations with different long-term N_e [14,72]: large, diverse populations, such as the Moroccan, follow a uniform distribution of effect sizes, while small, mutation-limited populations, such as the ones in the archipelago, follow an exponential distribution. Although we did not phenotype Eurasian populations, we used available data from the 1001 Genomes Project [73,74] to conduct the same analysis in that population (Supplementary Fig. 7). Similar to the Moroccan population, we found that variants associated with flowering time on this large, diverse population ($\theta_{W\ Eurasia} = 6.42x10^{-3}$) also followed a uniform distribution (AIC: uniform: 117.77, exponential: 166.27).

4.2. Directional selection acting on the archipelago to reduce flowering time

We next explored the relationship between effect size and allele frequency, using the same representative sets of SNPs. Moroccan alleles were polarized against the *A. lyrata* outgroup to identify which allele is derived and therefore assess the direction of the effect size (i.e., negative effect size/early flowering or positive effect size/late flowering) and the respective allele frequency. CVI alleles were polarized against Col-0 since we assume all variation in Cape Verde arose *de novo*, based

on the lack of shared variation with the mainland previously observed [42]. In Morocco, we found that effect size was positively correlated with allele frequency for loci increasing flowering time (t-test, t = 8.08, df = 195, p-value = 6.54×10^{-14}, Pearson's R = 0.50; Fig. 6). For loci decreasing flowering time, no significant correlation was observed (t-test, t = 0.50915, df = 98, p-value = 0.6118, Pearson's R = 0.05) since large effect loci are at very high and very low frequencies in the population. Under Moroccan simulated conditions, allele frequencies were positively correlated with effect sizes for loci that increase flowering time (t-test, t = 2.5479, df = 11, p-value = 0.02709, Pearson's R = 0.61), and negatively correlated for loci decreasing flowering time (t-test, t = -4.3586, df = 23, p-value = 0.0002303, Pearson's R = -0.67). On the islands, under CVI simulated conditions, loci accelerating flowering were mainly at high frequencies while loci delaying flowering were at low frequencies, suggesting a history of directional selection. Similar results were observed when CVI alleles were polarized based on allelic presence in Morocco (data not shown).

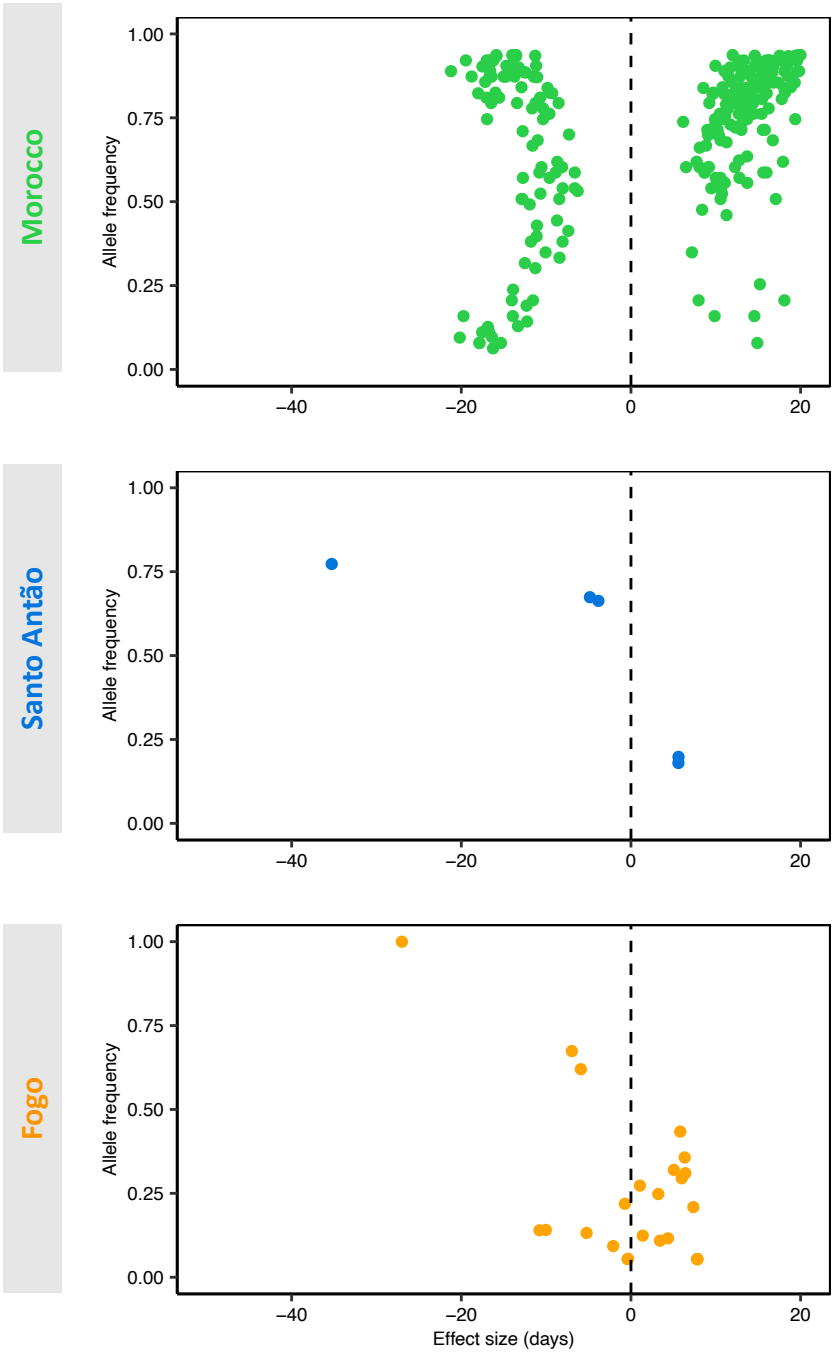


Figure 6. Relationship between allele frequency and effect size for flowering time-associated variants on the three populations. Effect size in days (x-axis) for each variant tagging a candidate loci (each dot) and its respective allele frequency on the population (y-axis).

4.3. The more polygenic architecture in the mainland is predominately associated with non-coding variants

We were also interested in asking whether variants associated with flowering time were more likely to result in coding or non-coding changes. In order to investigate how these variants might influence the flowering time trait in the different populations, we used predictions of effects based on the sequence change using SnpEff [75]. SnpEff classifies variants into four main groups: 'high impact', 'moderate impact, 'low impact' or 'modifier'. In the following analysis, the severity of substitution refers to the predicted severity of the changed such that high impact > moderate > modifier > low (from most severe to least).

In all cases, we found variants with higher frequencies and effects sizes were associated with stronger predicted effects. In Morocco, 70% of the variants tagging candidate loci were predicted to be modifiers, 12.1% were predicted to have moderate impacts and 17.8% were predicted to have low impacts (Fig. 7). No predicted high impact variants were found among the set of Moroccan variants. Both allele frequency and absolute effect size were positively correlated with the severity of the mutation in Morocco (LM, allele frequency: adjusted $R^2 = 0.08$, p-value = $0.1.667x10^{-6}$; absolute effect size: adjusted $R^2 = 0.069$, p-value = $1.07x10^{-5}$). In Santo Antão, 20% of the variants tagging candidate loci were predicted to have high impacts (*FRI* K232X), 20% were predicted to have moderate impacts and 60% were predicted to be modifiers. Absolute effect sizes were positively correlated with the severity of mutation, but we found no correlation between severity and allele frequencies (LM, absolute effect size: adjusted $R^2 = 0.999$, p-value = 0.00053; allele frequency: adjusted $R^2 = 0.034$, p-value = 0.483). In Fogo, 4% of the candidate loci were predicted to have high impacts (*FLC* R3X), 12% were predicted to have moderate impacts, 76% were predicted to have low impacts and 8% were predicted to be modifiers. Allele frequency and absolute effect size were again positively correlated with the severity of the mutation (LM, allele frequency: adjusted $R^2 = 0.426$, p-value = 0.0020; absolute effect size: adjusted $R^2 = 0.647$, p-value = $1.39x10^{-5}$). These results should be interpreted carefully due to the low number of variants in some severity categories.

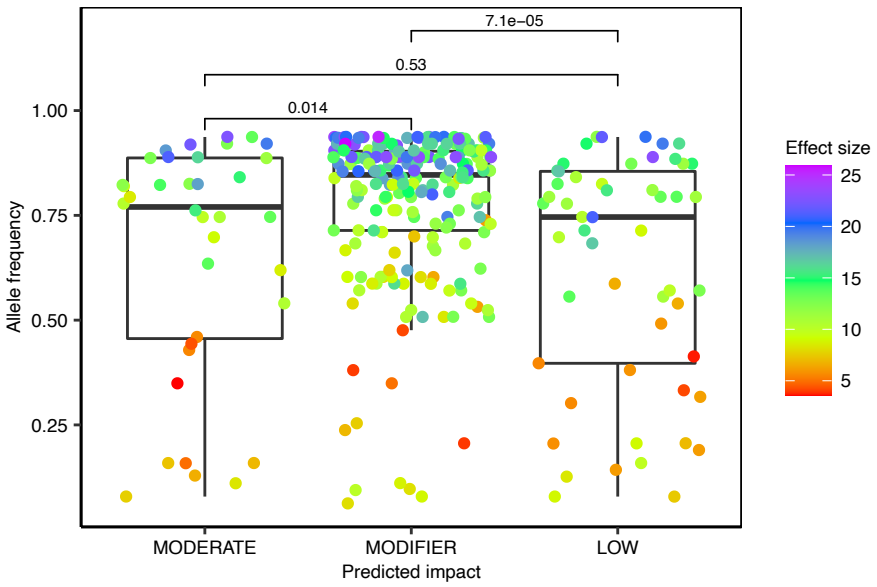


Figure 7. Predicted impact of flowering time-associated variants in Morocco. Boxplot of allele frequency (y-axis) per category of predicted impact (x-axis) for the variants tagging candidate loci (each dot). Each SNP is colored by absolute effect size in days estimated under CVI simulated conditions, following the legend. P-values shown from MU test.

4.4. Decrease of flowering time on the archipelago through an adaptive walk

The near-complete loss of variation that occurred with the colonization of the two Cape Verde islands provides a rare opportunity to examine how variation in a quantitative trait builds up over time. To investigate this, we estimated the ages of the variants tagging candidate loci identified in GWAS using RELATE [76] and compared them to the genomic background (Fig. 8). An important note here is that RELATE will eliminate any site where there is missing data even for a single individual, so that some positions in the genome including in the trait-associated loci will be lost.

In both islands, the distributions of ages of candidate loci differed from the genomic background distributions, with more older variants than expected among the trait-associated loci (Kolmogorov-Smirnov test, Santo Antão: mean background age = 205.91 years, mean candidate loci age = 1778.08 years, D = 0.81199, p-value = 0.0001987; Fogo: mean background age = 133.68 years, mean candidate loci age = 370.54 years, D = 0.38706, p-value = 0.006843). In Santo Antão, the oldest candidate locus, a small effect (effect size = -4 days) non-coding variant in a transposable element-rich region, is inferred to have arisen approx. 4000 years ago, and this was soon followed by the appearance of *FRI* K232X, a knock-out mutation of large effect [42]. *FRI* 232X may have enabled population expansion

on the island by allowing plants to colonize drier areas [42]. Smaller effect loci, mainly representing modifier mutations, appeared more recently, consistent with an Orr 'adaptive walk' model, in which large effect mutations tend to appear early in the walk towards a new fitness optimum. In Fogo, this pattern is even more striking with a positive correlation between absolute effect size and the age of the flowering time-associated variants (Pearson's R = 0.864, p-value = 1.76×10^{-6}).

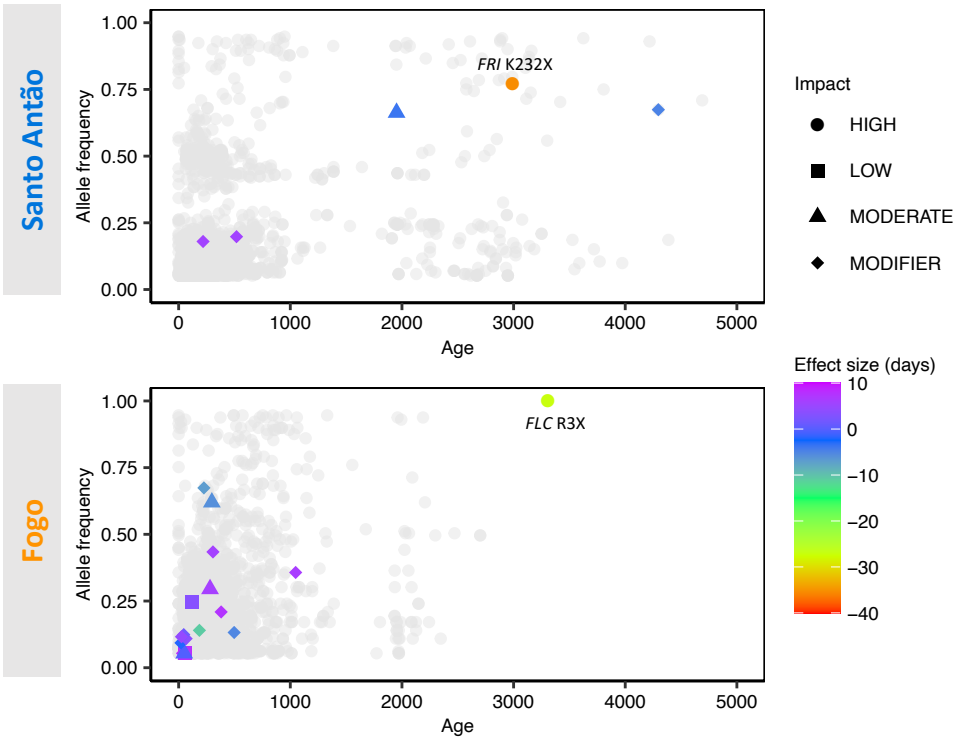


Figure 8. Relationship between age and allele frequency of flowering time-associated variants in CVI. Age estimates (x-axis) per variant in CVI and its allele frequency (y-axis) are shown. Each grey dot represents one SNP in the genome. Colored SNPs represent associated variants, with colors matching their estimated effect size in days and shape their predicted impact, following the legend.

Discussion

The long-range colonization of the CVI archipelago from North of Africa represented a sudden change in the environment. In the islands, precipitation seasonality and aridity increased, growing season length was reduced, photoperiod variation was lost and edaphic factory likely changed substantially [42]. We previously showed that a phenotypic shift towards early flowering on the islands allowed adaptation through two independent *de novo* loss of function large effect mutations at core genes in the vernalization pathway: *FRI* K232X and *FLC* R3X [42]. Here, we further showed that the genetic architecture of flowering time also differs between the island and mainland populations, with a more polygenic architecture on the mainland to an oligogenic architecture on the archipelago. The large effect mutations in CVI appear to represent early steps in the adaptive process, that were then followed by many small effect variants, for which associated candidates affect more peripheral pathways connected to flowering time.

The differences between mainland and island populations go farther. CVI was colonized about 5 kya through a complete bottleneck, which created a small population with low genetic diversity ($N_{e\ CVI}$ = 12.2K, $\theta_{W\ CVI}$ 1.48x10^{-4}). The Moroccan population, on the other hand, coalesces around 1 mya and presents high levels of genetic diversity and high N_e ($\theta_{W\ Morocco}$ = 5.56x10^{-3}, $N_{e\ Morocco}$ = 325K) [42,43]. Therefore, the striking difference in genetic architecture observed between the large continental and small island populations was expected [14,72]. Large populations, such as the Moroccan and the Eurasian populations, with large N_e and high mutational input, contain large amounts of standing genetic variation, which may include beneficial mutations at their disposal. After a change in the environment, adaptation in these populations is likely to occur through standing genetic variation [13,14,17]. However, in the case of young, small populations facing very strong selective pressures, such as the colonizing population of Cape Verde, adaptation tends to be rapid and to occur through few large effect loci, prevailing a more oligogenic architecture – the basis of Orr's model of adaptation [7,8,12,72].

Orr's 'adaptive walk' model, an extension of Fisher's geometric model, states that adaptation mostly relies on mutations of large fitness effects, with a pattern of diminishing returns. Under this model, a population placed far from its local optimum will accumulate large effect mutations at the beginning of the walk and move closer to the new optimum. Then, the next mutation will, on average, have a smaller effect, and so on, so that the effect size distribution at the end follows an exponential distribution [7,8]. In our system, both island populations – Santo Antão with *FRI* K232X at ~4000 years

ago and Fogo with *FLC* R3X at ~3000 years ago, followed by several smaller effect variants provide an example of an 'adaptive walk' in parallel in a natural case, with an exponential distribution of effects. In Morocco, on the other hand, the more polygenic architecture follows Fisher's infinitesimal model, with many small to moderate effects fitting a uniform distribution of effect sizes.

Furthermore, our results are also in agreement with the expectation that large effect alleles caused by loss of function of key genes are more likely to contribute after a sudden change to a distant optimum [11,18,19,72]. While in the continent we found mainly non-coding, possibly regulatory, mutations, on the archipelago we found two large effect independent loss of function mutations that arose and swept to high frequency. Nevertheless, the walk towards the new optimum is expected to comprise several steps, especially for a complex trait such as flowering time. GWAS mapping in natural variation panels have revealed association with large numbers of genes, implying polygenicity in Eurasian populations (from ~30 loci, explaining ~50% of the phenotypic variation [31,32] to >100 loci [77,78]).

On the archipelago, in addition to the two major effect variants, many other variants with smaller effect were also associated with flowering time. Among these, we identified candidate loci involved in flowering time and belonging to directly connected pathways (e.g., *ATX2*), but also in more peripheral and indirect pathways, such as nutrients/metal uptake and regulation (e.g. *NRT1*, *NRT1.8*, *ANR1*, *ZIP5*), and light sensing (e.g. *PAPP2C*, AT4G02480). Even though genes involved in these pathways were expected to affect flowering time indirectly [28,29], their major role may be related to adaptation to other selective pressures such as the volcanic soil, constant photoperiod and the higher light intensity. Under an omnigenic model, association signals from other pathways other than the flowering time itself are expected based on the complex network of molecular interactions with components of the flowering time pathway [79]. This omnigenic model proposes that essentially any gene in any peripheral pathway can contribute to a complex trait by some indirect effect. By the small world property of networks, most expressed genes are only a few steps from the nearest core gene and thus may have non-zero effects on the trait. Since core genes are hugely outnumbered by peripheral genes, a large fraction of the total genetic contribution to complex traits is expected to result from variants in peripheral genes that do not play direct roles on the trait [79]. It is also interesting to note a parallel between islands, where genes involved in drought response, epigenetic control and nutrients/minerals uptake showed association with flowering time.

Lastly, it is surprising that we find no significant effect of the well-known Cvi-0-EDI locus on flowering [39,40,80]. This large effect locus (*Early Day-length Insensitive*) was initially detected in QTL

mapping using RILs derived from a cross between Cvi-0 and Ler-0 [39] and later linked to a Cvi-0-specific allele at *CRY2* [40]. A dominant gain of function mutation in this gene (V367M) was identified as the responsible variant underlying the EDI QTL [40] and has been the focus of extensive work over the years [39,40,61,80]. This Cvi-0-specific allele has been shown to be responsible for a large reduction in the days to flower in the Ler-0 genetic background. Measured as total leaf number, Cvi-0 *CRY2* reduces flowering time in about 3 to 5 leaves under long-day photoperiods and in about 18 to 22 leaves under short-day photoperiods [39,40,80]. Although we capitulated this finding in the Cvi-0 x Ler-0 NILs under CVI simulated conditions, we detected no significant effect of *CRY2* V367M in the CVI genetic background in the natural populations, under CVI simulated conditions. Our results suggest one or more variants that are differentiated between the CVI and the Ler-0 genetic backgrounds result in an epistatic effect on the manifestation of a *CRY2* effect on flowering time. This finding also reflects the importance of being cautious when extrapolating effects, functions and interactions observed in laboratory and in specific lines, outside the natural population and ecological contexts. Regardless, the *CRY2* V367M allele is at very high frequency in the natural population of Santo Antão, suggesting that it may have effects in other traits. In fact, variation in *CRY2* has been shown to have pleiotropic effects on several fitness-related traits, such as fruit length, ovule number per fruit, and percentage of unfertilized ovules [61,62], but also in temperature and light responsiveness [64,65] and chromatin condensation [63]. Further work on the effects of *CRY2* V367M on other traits in relevant CVI genetic backgrounds is needed to fully understand its function and relevance under natural conditions.

With the CVI *Arabidopsis*-Morocco contrast, our findings exemplify the effects of population history on the genetic architecture of a quantitative trait after colonization of a new environment. In the contrast between Cape Verde and Morocco, they further provide an empirical example of Orr's theoretical expectations. Adaptation to a new far fitness optimum by a small, mutation-limited population was accomplished by few large effect mutations at the beginning of the 'adaptive walk', followed by more smaller effect variants in several intrinsically connected pathways.

Material and methods

1. Genomic data and population structure analyses

All genomic data on the CVI natural populations and the Moroccan population have been previously published in [42,43], respectively. Neighbor-joining tree was calculated using the R package *ape* and the principal component analysis using the flag <--pca> in PLINK v.1.90 [81] (scripts available at https://github.com/hancocklab).

2. Experimental conditions and accessions

2.1. CVI simulated conditions experiment

We used phenotypic data published in [42]. In short, four replicates of 365 accessions (174 from Santo Antão, 129 from Fogo, and 62 from Morocco) were grown in a custom Bronson growth chamber set to track hourly environmental data (temperature, humidity, photoperiod, and precipitation) from Cape Verde, simulating the growing season. Following the precipitation loggers in the field, water was withheld 26 days after sowing.

Prior to sowing, seeds were stratified in the dark in Petri dishes on water-soaked filter paper for one week at 4°C, and then sowed in 7x7cm pots containing a standard potting compost mix, supplemented with iron. Four seeds were sown per pot and plants were thinned to one plant per pot, after germination. The plants were organized in a randomized block design.

Bolting time, i.e., the number of days from sowing to the appearance of the differentiated floral bud, was scored per individual and the median across replicates was taken as the phenotype per genotype. The term flowering time is used throughout the paper to represent the scored bolting time, as the latter is a proxy for the former.

We also grew two Cvi-0 x Ler-0 NILs – LCN1-2.5 and LCN1-2.8 – and the two parental lines Cvi-0 and Ler-0 to assess the effect of *CRY2* – an *a priori* candidate –, under CVI simulated conditions. Each line was grown in four replicates. Phenotyping was conducted as indicated above.

2.2. Moroccan simulated conditions experiment

To simulate the growing season in Morocco, we collected climate data from weather stations in Setti Fatma (from https://www.worldweatheronline.com) as the representative Moroccan conditions. This site was selected because it is located in the High Atlas region, from where the closest

genome-wide Moroccan accessions to CVI had been identified [42]. We used these data to set up a custom Bronson growth chamber and simulated temperature, humidity, and precipitation. We started the experiment with data from January 2016, when we estimated would be the beginning of the growing season in the field. Photoperiod was set to 11h and dawn and dusk were simulated by increasing light intensity by 50 µM every 15 minutes until 200µM (full light) and decreasing it by 50µM every 15 minutes until dark, respectively. At the same time points, far-red light decreased from 50 to 0 µM at dawn and increased from 0 to 50 µM at dusk.

Seed stratification and scoring of bolting time were done in the same manner as in the CVI simulated conditions experiment. The experimental design and the accessions from the natural populations included were also as in the previous experiment.

2.3. Production of inter- and intra-islands F1 hybrids

To produce inter- and intra-islands F1 hybrids, we selected four geographical and genetical representative individuals from each island (Cvi-0, S1-1, S5-10 and S15-3 from Santo Antão, and F3-2, F9-2, F13-8 and F10-1-3 from Fogo) and cross them in all possible combinations. Resulting seeds were stratified, sowed, thinned and scored for bolting time as described above. F1 plants and parental lines were grown in four replicates, under standard greenhouse conditions with 12h photoperiod.

Replicates that did not pass Tukey fence method for outlier detection were discarded from the analysis (scripts available at https://github.com/hancocklab).

3. Mapping the genetic bases of flowering time

To map flowering time, we conducted genome-wide association studies (GWAS) using GEMMA v.0.94 [46], with some *a posteriori* modifications (see below). We used SNP and InDel variants called with the GATK pipeline and published in [42]. All input files were generated using VCFtools v.0.1.14 [82] and PLINK v.1.90 [81], and the median bolting time across replicates per genotype was used as the phenotype (scripts available at https://github.com/HancockLab). For mapping, in Santo Antão, we used 172 individuals, in Fogo 129, and in Morocco 62.

Broad-sense heritability (H^2) was calculated in R using the function *repeatability()* from package *heritability* [83], and narrow-sense heritability in GEMMA [46] with BSLMM [44] (see below). Candidate genes were identified based on allele frequency in the natural population, estimated effect size (beta), and gene function. Regions in association and candidate genes were annotated and examined in more detail with IGV [84], ThaleMine [85] and SnpEff [75]. For the analysis of *ATX2* influence on

FLC mRNA levels, we used RNA-seq data from unpublished work in the lab (scripts available at https://github.com/HancockLab).

3.1. GEMMA

First, we employed BSLMM [44], a Bayesian sparse linear mixed model, implemented in GEMMA [44], by using <-bslmm 1>. The model estimates the number of loci with major effects involved in the phenotype using a Markov chain Monte Carlo (MCMC) with <-s 10000000> sampling steps and <-w 2500000> burn-in iterations. Median and 95% confidence interval (CI) PVE (proportion of the variance explained) and gamma (number of variants with sparse effects) were calculated across ten runs (scripts available at https://github.com/HancockLab).

We also used a univariate linear mixed model (LMM) that accounts for population structure with a centered kinship matrix <-gk 1> in GEMMA v.0.94. We used the flag <-lmm 4> which estimates beta per marker as the effect size of an allele and used the likelihood ratio test to assess association. Any time a covariate was added to the model, it was done so with the command <-c >.

3.2. Local score

To focus on small to moderate effect QTLs, we used the local score approach, which accounts for linkage disequilibrium (LD) when estimating association [47]. The significant p-values are assessed together through LD between markers in a short genomic region containing a causal variant and are then turned into scores, defined as the maximum of a Lindley process over the sequence of scores [48]. We used available scripts to compute the Lindley scores from the LMM output (from https://forge-dga.jouy.inra.fr/projects/local-score/documents).

3.3. Iteratively adding covariates

In a similar way and to account for linkage between markers and association to the phenotype, we clumped the results from the association test with the LMM using PLINK [81] and the flags <--clump>, <--clump-p1 0.01>, <--clump-kb 1000>, and <--clump-r2 0.8>. Clumps were formed around central 'index variants' which are significant at α=0.01, and by variants that are within 1 Mbp distance and with $r^2 > 0.8$ with the index marker. We then added these index markers to the model iteratively, using the flag <-c> in GEMMA and a custom script (scripts available at https://github.com/HancockLab). We took the top 10% of index SNPs and added them iteratively as covariates to identify the markers that decreased the PVE the most.

4. Genetic architecture inference

For all analyses in this section, only SNP markers were used and filtered from the initial VCF file using the command < --remove-indels> in vcftools [82]. Any time we needed *FLC* R3X in the analysis, it was added manually since the SNP is fixed on the population. Values for age and effect size were collected from [42].

4.1. Allele frequency and effect size estimation and impact prediction of candidate loci

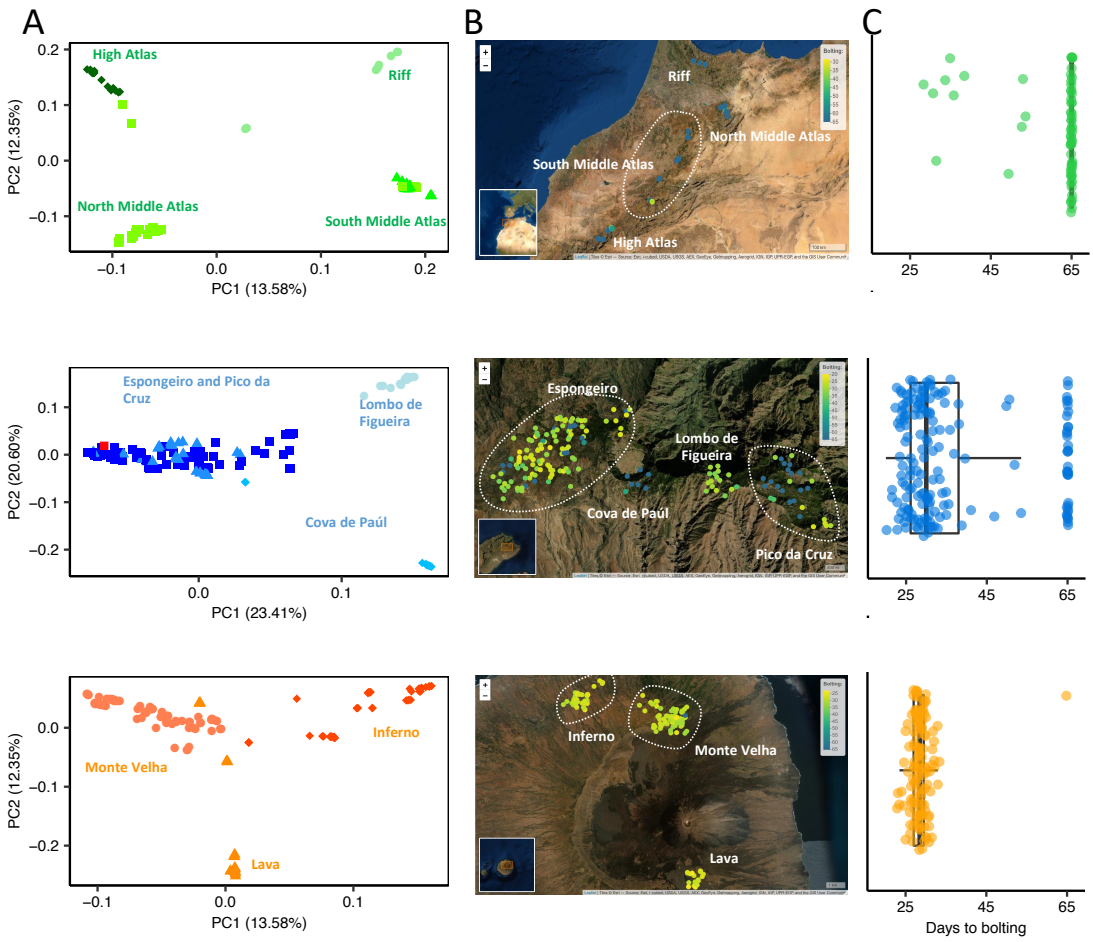
For each candidate locus identified with the local score approach, we took one representative SNP and re-calculated allele frequency and effect size, based on the estimates outputted by GEMMA (with flag <-lmm 4>). GEMMA takes the minor allele as the effect allele. However, in some cases, the allele of interest, usually the derived allele, may be at higher frequencies. In order to assess which allele is the ancestral/derived and therefore infer the direction of the effect size and the respective allele frequency, we polarized Moroccan alleles to the *A. lyrata* outgroup and CVI alleles to Col-0. Estimates were re-calculated whenever necessary. Candidate loci, represented by a single SNP, were then annotated using SnpEff [75].

Effect sizes for previously published QTL were retrieved from the original study (data in https://github.com/HancockLab) and for the Eurasian population in the 1001 Genome Project were calculated following the same method as for natural populations in our study using available phenotypic data [73,74].

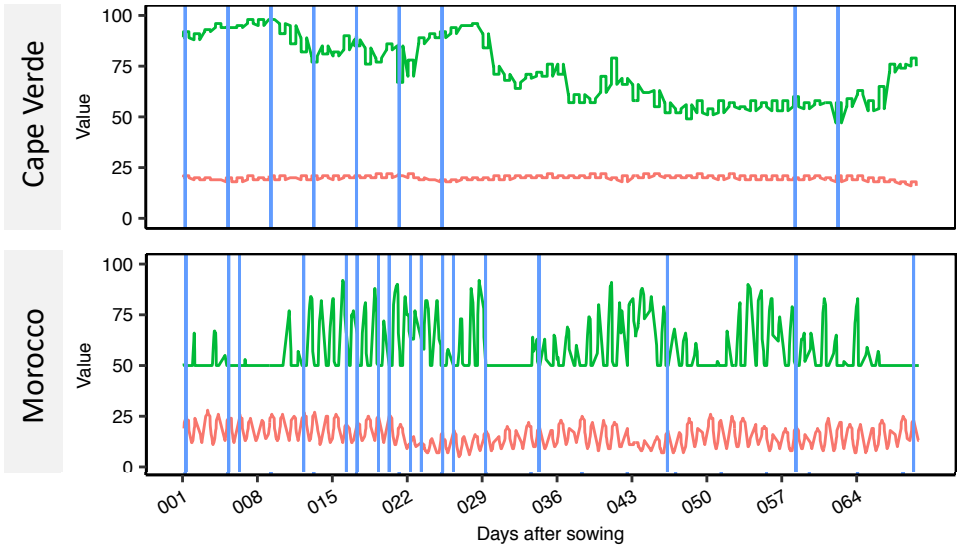
4.2. Age estimation of candidate loci

First, we pruned the genomes in Santo Antão and Fogo using PLINK and the flag < --indep-pairwise 50 10 0.3>. Then, we estimated age for all SNPs in the pruned genome in the CVI populations using RELATE [76] and compared background ages to candidate loci's. To do so, we followed [42] (scripts are available at https://github.com/HancockLab). Mutation rate was corrected for missing data across the entire genome and the recombination map was taken from [30] with a correction for the outcrossing rate of 5% in the natural populations [86].

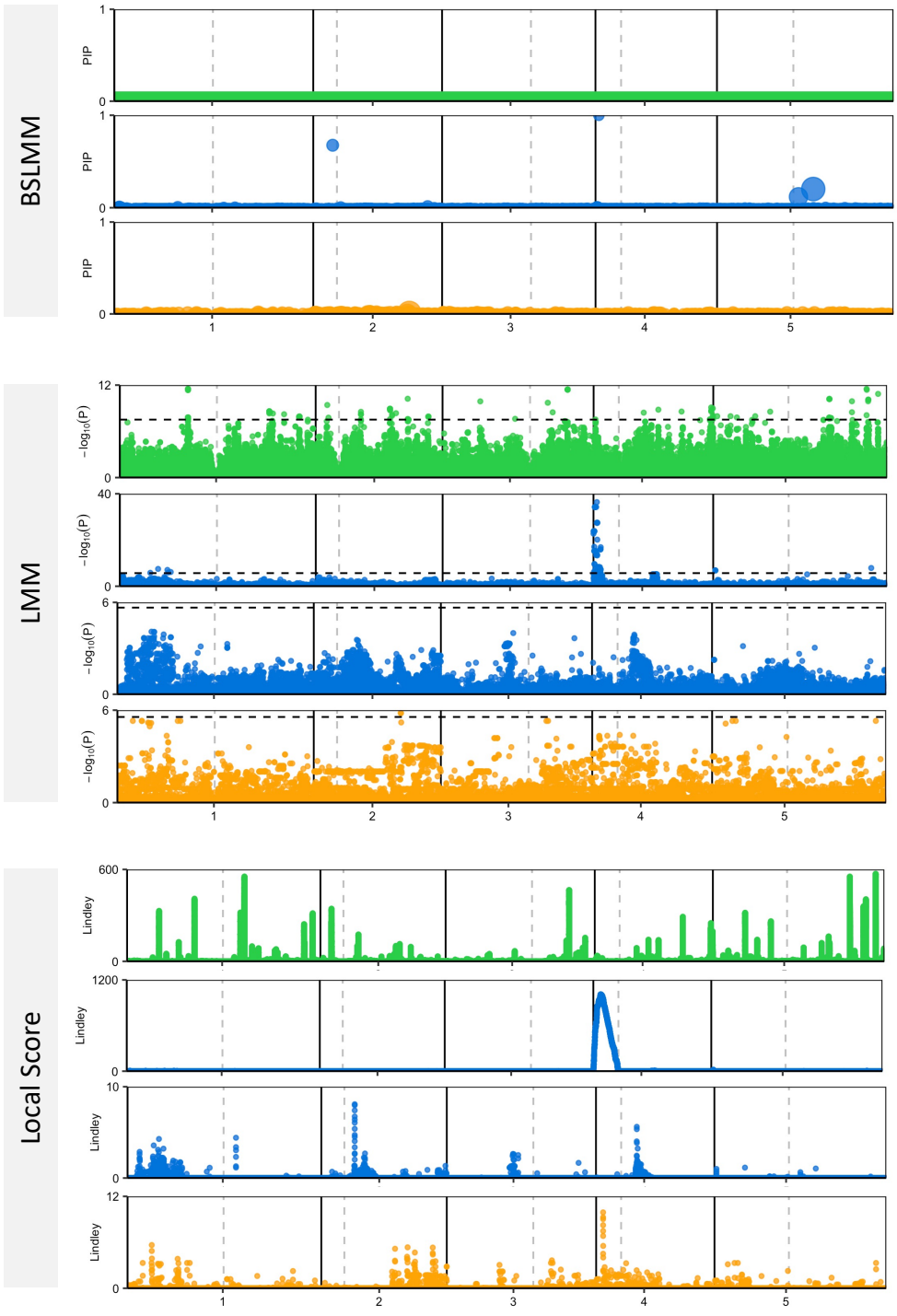
Supplementary Figures



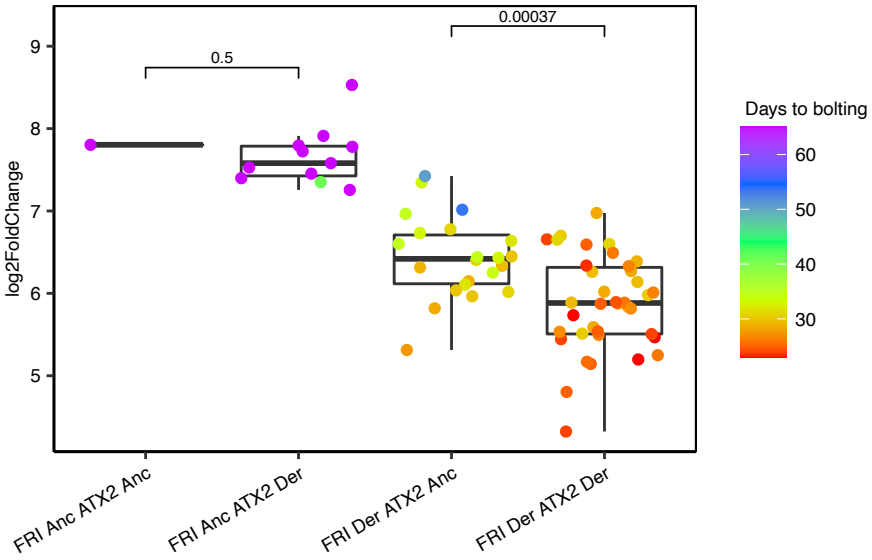
Supplementary Figure 1. Population cluster differentiation and flowering time differences. A. PCA showing cluster differentiation in Morocco (top, in green), Santo Antão (middle, in blue) and Fogo (bottom, in orange). Each dot represents one individual colored and shaped according to the geographical cluster it belongs. The red dot in the Santo Antão PCA shows the Cvi-0 position relative to the natural population. B. Maps with the geographical distribution of the accessions used annotated by cluster. Each dot represents one genotype colored by the flowering time scored under CVI simulated conditions. C. Bolting time in days (x-axis) per population (Morocco in green, Santo Antão in blue, and Fogo in orange) under CVI simulated conditions. Each dot represents the median across replicates.



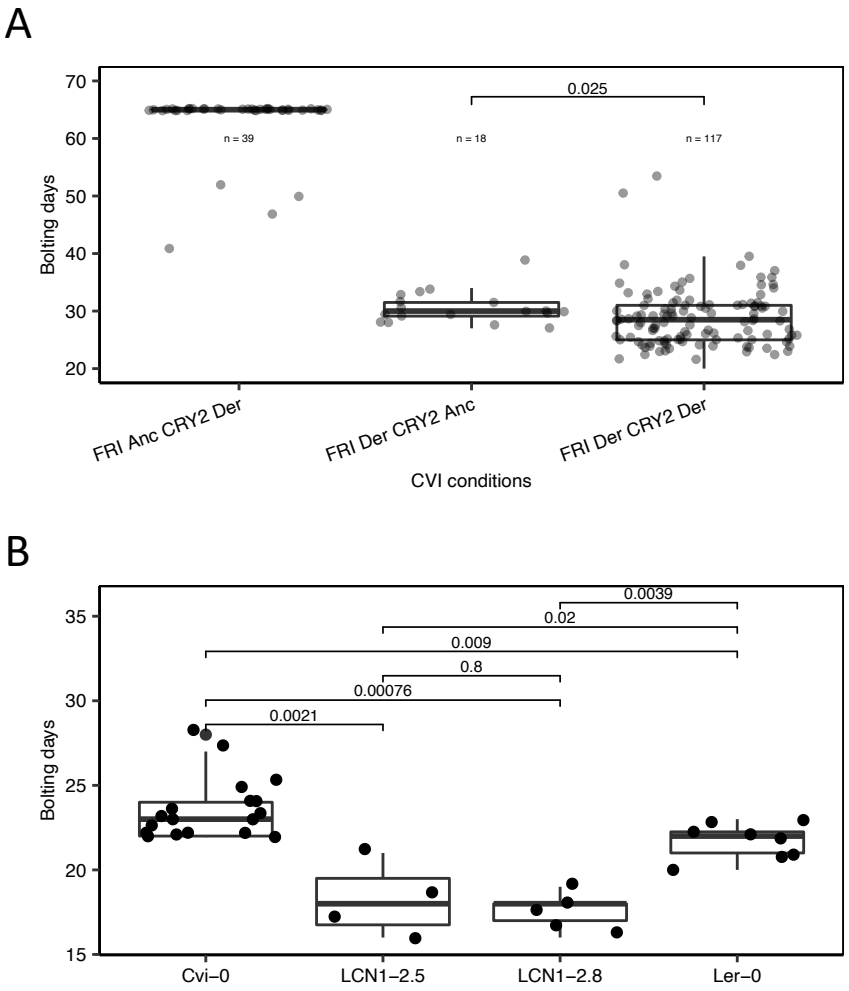
Supplementary Figure 2. Cape Verde (top) and Morocco (bottom) simulated conditions. Days after sowing are shown on the x-axis and temperature (in red, in °C) and humidity values (in green, in %) on the y-axis. Vertical blue lines represent watering events.



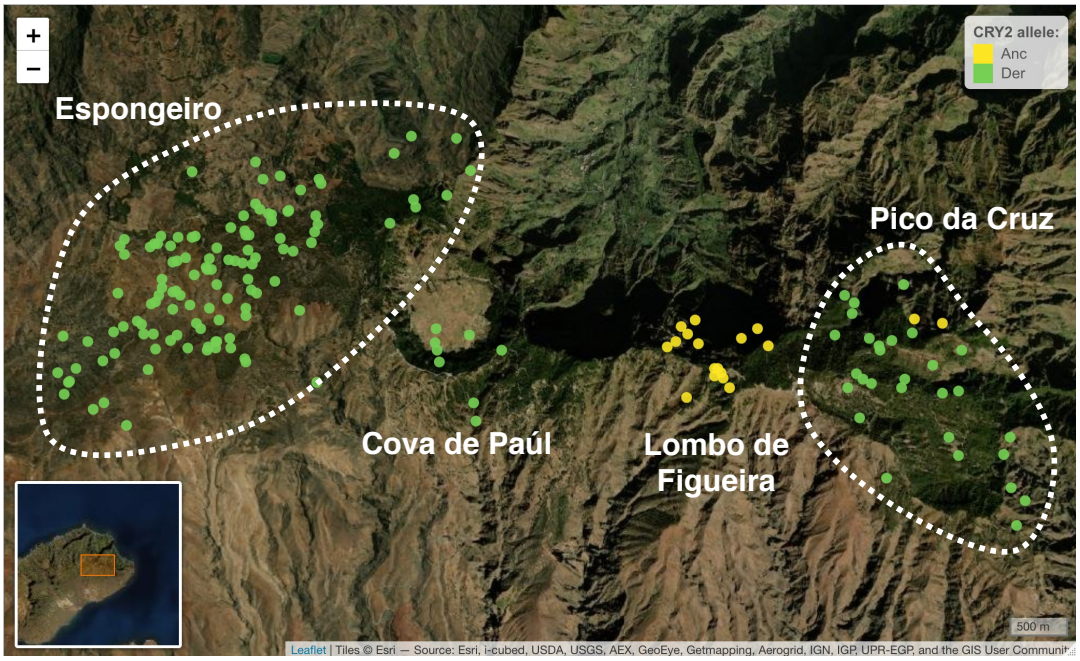
Supplementary Figure 3. Genetic bases of flowering time. Manhattan plots for flowering time on the three populations (Morocco in green, Santo Antão in blue, and Fogo in orange) across methods (BSLMM, LMM and local score) with genomic position on the x-axis and significance statistic on the y-axis. For Santo Antão, the two Manhattan plots show mapping with all genotyped markers on top and with *FRI* K232X as a covariate on the bottom.



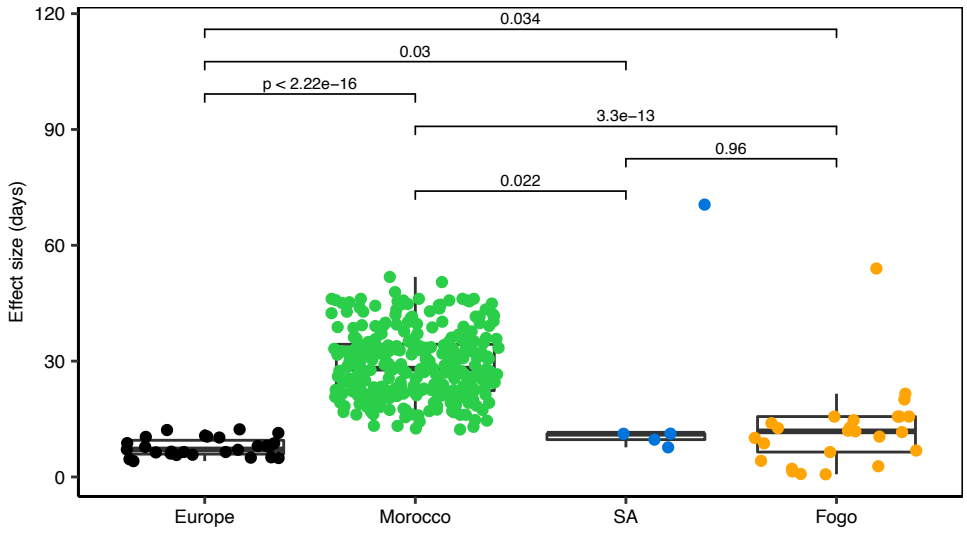
Supplementary Figure 4. FLC mRNA levels in Santo Antão. X-axis shows all possible allelic combinations of *FRI* K232X and *ATX2* L125F and y-axis the log2-fold change on FLC mRNA levels, compared to Col-0. Each dot represents one accession (3 replicates) colored by days to bolting under CVI simulated conditions following the legend. P-values shown from MU test.



Supplementary Figure 5. *CRY2* V367M effect in the Santo Antão natural population and in Cvi-0 x Ler-0 NILs. A. Days to bolting (y-axis) per *FRI* K232X and *CRY2* V367M allelic combination (x-axis) scored under CVI simulated conditions. Each dot represents the median across replicates per genotype. B. Days to bolting (y-axis) under CVI simulated conditions on two Cvi-0 x Ler-0 NILs containing the EDI locus and the parental lines (x-axis). Each dot represents one replicate. P-values shown from MU test.



Supplementary Figure 6. *CRY2* V367M geographical distribution in Santo Antão. Each dot represents one accession in the natural population with clusters highlighted. Yellow dots show individuals carrying the ancestral *CRY2* V367 allele and green dots the derived *CRY2* 367M allele.



Supplementary Figure 7. Effect size distribution of candidate loci affecting flowering time. Absolute effect size in days (y-axis) per population (x-axis) is shown for candidate loci which are represented by one SNP (each dot). Morocco, Santo Antão and Fogo effect sizes were assessed under CVI simulated conditions. Europe comprehends the Eurasian individuals from the 1001 Genome Project previously phenotyped in [73,74]. P-values shown from MU test.

Supplementary Tables

	Chromosome	Position	RefAllele	AltAllele	Variant	Impact	Gene	GeneID
	1	4105098	C	G	upstream_gene_variant	MODIFIER	NPF6.3	AT1G12110
	1	4110031	TC	T	downstream_gene_variant	MODIFIER	NPF6.3	AT1G12110
	1	4113448	G	A	downstream_gene_variant	MODIFIER	NPF6.3	AT1G12110
	1	4937852	CA	C	upstream_gene_variant	MODIFIER	AT59	AT1G14420
	1	4937980	T	C	upstream_gene_variant	MODIFIER	AT59	AT1G14420
	1	4938379	G	A	upstream_gene_variant	MODIFIER	AT59	AT1G14420
	1	17025853	G	GGA	upstream_gene_variant	MODIFIER	AT1G45035	AT1G45035
	1	17025859	TAA	T	upstream_gene_variant	MODIFIER	AT1G45035	AT1G45035
	1	17025880	TCA	T	upstream_gene_variant	MODIFIER	AT1G45035	AT1G45035
	1	17025905	C	CAACA	upstream_gene_variant	MODIFIER	AT1G45035	AT1G45035
	2	5239388	A	AACATGT	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
	2	5239389	T	A	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
	2	5239392	T	TC	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
	2	5239396	A	G	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
	2	5239397	A	ATCC	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
	2	5239398	ACG	A	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
	2	5239404	TAG	T	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
Santo Antão	2	5239408	AC	A	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760

2	5239411	CTT	C	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
2	5239416	GATAAT	G	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
2	5239422	A	G	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
2	5239424	T	TCC	upstream_gene_variant	MODIFIER	AT2G12760	AT2G12760
4	6385229	C	CTTG	upstream_gene_variant	MODIFIER	AT4G10290	AT4G10290
4	6394594	GT	G	splice_region_variant&intron_variant	LOW	HKT1	AT4G10310
4	6396971	C	CCATGAATCCAAATAATACATGATAT	downstream_gene_variant	MODIFIER	HKT1	AT4G10310
4	6405104	G	C	upstream_gene_variant	MODIFIER	AT4G10320	AT4G10320
1	2390463	T	C	upstream_gene_variant	MODIFIER	AT1G07710	AT1G07710
1	3771082	A	C	upstream_gene_variant	MODIFIER	AT1G11240	AT1G11240
1	3775490	T	A	upstream_gene_variant	MODIFIER	STP1	AT1G11260
1	3808292	T	C	missense_variant	MODERATE	AT1G11320	AT1G11320
1	4755623	T	C	upstream_gene_variant	MODIFIER	PAP2	AT1G13900
1	4962070	GC	G	3_prime_UTR_variant	MODIFIER	AL7	AT1G14510
1	5238976	C	T	upstream_gene_variant	MODIFIER	ABCG35	AT1G15210
1	7872125	G	C	upstream_gene_variant	MODIFIER	PAPP2C	AT1G22280
1	7872126	T	G	upstream_gene_variant	MODIFIER	PAPP2C	AT1G22280
1	7951954	CA	C	downstream_gene_variant	MODIFIER	ATL15	AT1G22500
1	7977152	T	A	missense_variant	MODERATE	NPF5.15	AT1G22570
1	7977153	C	A	stop_gained	HIGH	NPF5.15	AT1G22570
1	7982918	T	C	upstream_gene_variant	MODIFIER	NPF5.15	AT1G22570

Fogo

1	7984084	G	A	3_prime_UTR_variant	MODIFIER	AGL87	AT1G22590
1	7993635	A	C	upstream_gene_variant	MODIFIER	AT1G22600	AT1G22600
1	7996626	C	G	missense_variant	MODERATE	AT1G22610	AT1G22610
1	9354932	T	A	upstream_gene_variant	MODIFIER	AT1G26950	AT1G26950
1	9766388	G	A	splice_region_variant&intron_variant	LOW	ABCB14	AT1G28010
2	11544690	C	T	upstream_gene_variant	MODIFIER	AGO4	AT2G27040
2	11545505	A	G	upstream_gene_variant	MODIFIER	AGO4	AT2G27040
2	11559156	C	T	missense_variant	MODERATE	ARR13	AT2G27070
2	11573890	C	T	upstream_gene_variant	MODIFIER	AT2G27090	AT2G27090
2	13461335	A	C	upstream_gene_variant	MODIFIER	SAD2	AT2G31660
2	13463702	T	C	upstream_gene_variant	MODIFIER	ATX1	AT2G31650
2	13528769	C	A	upstream_gene_variant	MODIFIER	AT2G31820	AT2G31820
2	13546302	C	A	3_prime_UTR_variant	MODIFIER	AT2G31862	AT2G31862
2	13548572	T	G	upstream_gene_variant	MODIFIER	AT2G31860	AT2G31860
2	14495701	T	G	3_prime_UTR_variant	MODIFIER	AT2G34350	AT2G34350
2	14497242	C	T	missense_variant	MODERATE	AT2G34355	AT2G34355
2	14676400	C	T	downstream_gene_variant	MODIFIER	MEE22	AT2G34780
2	14679208	A	C	upstream_gene_variant	MODIFIER	MEE23	AT2G34790
2	14680964	G	A	upstream_gene_variant	MODIFIER	MEE23	AT2G34790
2	14682581	G	GAA	upstream_gene_variant	MODIFIER	AT2G34800	AT2G34800
2	14682587	T	A	upstream_gene_variant	MODIFIER	AT2G34800	AT2G34800
2	14682591	TTC	T	upstream_gene_variant	MODIFIER	AT2G34800	AT2G34800

2	14683061	T	A	upstream_gene_variant	MODIFIER	AT2G34800	AT2G34800
2	17489244	A	C	upstream_gene_variant	MODIFIER	AT2G41900	AT2G41900
2	17489246	T	A	upstream_gene_variant	MODIFIER	AT2G41900	AT2G41900
2	17490597	T	C	5_prime_UTR_variant	MODIFIER	AT2G41900	AT2G41900
2	17496401	C	T	upstream_gene_variant	MODIFIER	AT2G41910	AT2G41910
2	17926691	T	TC	upstream_gene_variant	MODIFIER	AT2G43120	AT2G43120
2	17926692	T	A	upstream_gene_variant	MODIFIER	AT2G43120	AT2G43120
2	17926694	AAAAG	A	upstream_gene_variant	MODIFIER	AT2G43120	AT2G43120
2	17926700	TA	T	upstream_gene_variant	MODIFIER	AT2G43120	AT2G43120
2	17926702	CG	C	upstream_gene_variant	MODIFIER	AT2G43120	AT2G43120
3	8265110	AT	A	upstream_gene_variant	MODIFIER	AT3G23160	AT3G23160
3	8276089	A	T	upstream_gene_variant	MODIFIER	AT3G23172	AT3G23172
3	16348761	A	G	upstream_gene_variant	MODIFIER	AT3G44805	AT3G44805
3	16495461	G	A	upstream_gene_variant	MODIFIER	AT3G45090	AT3G45090
3	16497595	TA	T	upstream_gene_variant	MODIFIER	AT3G45090	AT3G45090
3	16497599	A	T	upstream_gene_variant	MODIFIER	AT3G45090	AT3G45090
3	16616372	C	G	upstream_gene_variant	MODIFIER	SYP72	AT3G45280
4	1091742	GAA	G	upstream_gene_variant	MODIFIER	AT4G02480	AT4G02480
4	1091745	A	G	upstream_gene_variant	MODIFIER	AT4G02480	AT4G02480
4	1091747	GAGAGAGAGG TGA	G	upstream_gene_variant	MODIFIER	AT4G02480	AT4G02480
4	4702473	T	C	upstream_gene_variant	MODIFIER	AT4G07874	AT4G07874
4	4702475	G	GC	upstream_gene_variant	MODIFIER	AT4G07874	AT4G07874

5	2093955	T	A	3_prime_UTR_variant	MODIFIER	EML2	AT5G06780
5	3146556	G	A	missense_variant	MODERATE	AT5G10060	AT5G10060
5	3661841	A	G	upstream_gene_variant	MODIFIER	AT5G11460	AT5G11460
5	25358142	A	G	upstream_gene_variant	MODIFIER	AT5G63220	AT5G63220

References

1. Provine, W. B. *The Origins of Theoretical Population Genetics*. (University of Chicago Press, 2001).

2. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1919).

3. Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. *Science* **313**, 101–104 (2006).

4. Buckler, E. S. *et al.* The Genetic Architecture of Maize Flowering Time. *Science* **325**, 714–718 (2009).

5. van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial Melanism in British Peppered Moths Has a Singular and Recent Mutational Origin. *Science* **332**, 958–960 (2011).

6. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).

7. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).

8. Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56**, 1317–1330 (2002).

9. Desai, M. M. & Fisher, D. S. Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics* **176**, 1759–1798 (2007).

10. Frenkel, E. M., Good, B. H. & Desai, M. M. The Fates of Mutant Lineages and the Distribution of Fitness Effects of Beneficial Mutations in Laboratory Budding Yeast Populations. *Genetics* **196**, 1217 (2014).

11. Arjan G., J. *et al.* Diminishing Returns from Mutation Supply Rate in Asexual Populations. *Science* **283**, 404–406 (1999).

12. Willi, Y., van Buskirk, J. & Hoffmann, A. A. Limits to the Adaptive Potential of Small Populations. *Annu. Rev. Ecol. Evol. Syst.* **37**, 433–458 (2006).

13. Sella, G. & Barton, N. H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).

14. Vahdati, A. R. & Wagner, A. Population Size Affects Adaptation in Complex Ways: Simulations on Empirical Adaptive Landscapes. *Evol. Biol.* **45**, 156–169 (2018).

15. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nat. Rev. Genet.* **3**, 11–21 (2002).

16. Szendro, I. G., Franke, J., de Visser, J. A. G. M. & Krug, J. Predictability of evolution depends nonmonotonically on population size. *Proc. Natl. Acad. Sci.* **110**, 571–576 (2013).

17. Dittmar, E. L., Oakley, C. G., Conner, J. K., Gould, B. A. & Schemske, D. W. Factors influencing the effect size distribution of adaptive substitutions. *Proc. R. Soc. B Biol. Sci.* **283**, 20153065 (2016).

18. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243–1247 (2009).

19. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).

20. Hall, M. C. & Willis, J. H. Divergent Selection on Flowering Time Contributes to Local Adaptation in Mimulus Guttatus Populations. *Evolution* **60**, 2466–2477 (2006).

21. Ågren, J., Oakley, C. G., Lundemo, S. & Schemske, D. W. Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping. *Evolution* **71**, 550–564 (2017).

22. Olsson, K. & Ågren, J. Latitudinal population differentiation in phenology, life history and flower morphology in the perennial herb Lythrum salicaria. *J. Evol. Biol.* **15**, 983–996 (2002).

23. Franks, S. J., Sim, S. & Weis, A. E. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proc. Natl. Acad. Sci.* **104**, 1278–1282 (2007).

24. Stinchcombe, J. R. *et al.* A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4712–4717 (2004).

25. Fournier-Level, A. *et al.* Paths to selection on life history loci in different natural environments across the native range of Arabidopsis thaliana. *Mol. Ecol.* **22**, 3552–3566 (2013).

26. Brachi, B. *et al.* Investigation of the geographical scale of adaptive phenological variation and its underlying genetics in Arabidopsis thaliana. *Mol. Ecol.* **22**, 4222–4240 (2013).

27. Elzinga, J. A. *et al.* Time after time: flowering phenology and biotic interactions. *Trends Ecol. Evol.* **22**, 432–439 (2007).

28. Mouradov, A., Cremer, F. & Coupland, G. Control of flowering time: interacting pathways as a basis for diversity. *Plant Cell* **14**, S111–S130 (2002).

29. Andrés, F. & Coupland, G. The genetic basis of flowering responses to seasonal cues. *Nat. Rev. Genet.* **13**, 627–639 (2012).

30. Salomé, P. A. *et al.* Genetic Architecture of Flowering-Time Variation in Arabidopsis thaliana. *Genetics* **188**, 421–433 (2011).

31. Zan, Y. & Carlborg, Ö. A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected Arabidopsis thaliana. *Mol. Ecol. Resour.* **18**, 798–808 (2018).

32. Zan, Y. & Carlborg, Ö. A Polygenic Genetic Architecture of Flowering Time in the Worldwide Arabidopsis thaliana Population. *Mol. Biol. Evol.* **36**, 141–154 (2019).

33. Lempe, J. *et al.* Diversity of flowering responses in wild *Arabidopsis thaliana* strains. *PLoS Genet.* **1**, 109–118 (2005).

34. Shindo, C. *et al.* Role of FRIGIDA and FLOWERING LOCUS C in Determining Variation in Flowering Time of Arabidopsis. *Plant Physiol.* **138**, 1163 (2005).

35. Johanson, U. *et al.* Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science* **290**, 344–347 (2000).

36. Werner, J. D. *et al.* FRIGIDA-Independent Variation in Flowering Time of Natural Arabidopsis thalianaAccessions. *Genetics* **170**, 1197–1207 (2005).

37. Zhang, L. & Jiménez-Gómez, J. M. Functional analysis of FRIGIDA using naturally occurring variation in *Arabidopsis thaliana*. *Plant J.* **103**, 154–165 (2020).

38. Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. & Purugganan, M. D. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15670–15675 (2004).

39. Alonso-Blanco, C., El-Assal, S. E.-D., Coupland, G. & Koornneef, M. Analysis of Natural Allelic Variation at Flowering Time Loci in the Landsberg erecta and Cape Verde Islands Ecotypes of Arabidopsis thaliana. *Genetics* **149**, 749 (1998).

40. El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J. M., Raz, V. & Koornneef, M. A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. *Nat. Genet.* **29**, 435–440 (2001).

41. Gazzani, S., Gendall, A. R., Lister, C. & Dean, C. Analysis of the Molecular Basis of Flowering Time Variation in Arabidopsis Accessions. *Plant Physiol.* **132**, 1107–1114 (2003).

42. Fulgione, A. *et al.* Parallel reduction in flowering time from new mutations enabled evolutionary rescue in colonizing Arabidopsis lineages. *Rev.* (2022).

43. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci.* **114**, 5213 (2017).

44. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genet.* **9**, e1003264 (2013).

45. Brennan, A. C. *et al.* The genetic structure of Arabidopsis thaliana in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. *BMC Plant Biol.* **14**, 17 (2014).

46. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

47. Bonhomme, M. *et al.* A local score approach improves GWAS resolution and detects minor QTL: application to Medicago truncatula quantitative disease resistance to multiple Aphanomyces euteiches isolates. *Heredity* **123**, 517–531 (2019).

48. Mercier, S. & Daudin, J. J. Exact distribution for the local score of one i.i.d. random sequence. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **8**, 373–380 (2001).

49. Guo, F. Q., Wang, R., Chen, M. & Crawford, N. M. The Arabidopsis dual-affinity nitrate transporter gene AtNRT1.1 (CHL1) is activated and functions in nascent organ development during vegetative and reproductive growth. *Plant Cell* **13**, 1761–1777 (2001).

50. Guo, F.-Q., Young, J. & Crawford, N. M. The nitrate transporter AtNRT1.1 (CHL1) functions in stomatal opening and contributes to drought susceptibility in Arabidopsis. *Plant Cell* **15**, 107–117 (2003).

51. Glass, A. D. M. & Kotur, Z. A reevaluation of the role of Arabidopsis NRT1.1 in high-affinity nitrate transport. *Plant Physiol.* **163**, 1103–1106 (2013).

52. Jian, S. *et al.* NRT1.1 Regulates Nitrate Allocation and Cadmium Tolerance in Arabidopsis. *Front. Plant Sci.* **10**, 384 (2019).

53. Teng, Y., Liang, Y., Wang, M., Mai, H. & Ke, L. Nitrate Transporter 1.1 is involved in regulating flowering time via transcriptional regulation of FLOWERING LOCUS C in Arabidopsis thaliana. *Plant Sci.* **284**, 30–36 (2019).

54. Gan, Y., Filleur, S., Rahman, A., Gotensparre, S. & Forde, B. G. Nutritional regulation of ANR1 and other root-expressed MADS-box genes in Arabidopsis thaliana. *Planta* **222**, 730 (2005).

55. Walch-Liu, P. *et al.* Nitrogen Regulation of Root Branching. *Ann. Bot.* **97**, 875–881 (2006).

56. Pien, S. *et al.* ARABIDOPSIS TRITHORAX1 dynamically regulates FLOWERING LOCUS C activation via histone 3 lysine 4 trimethylation. *Plant Cell* **20**, 580–588 (2008).

57. Saleh, A. *et al.* The Highly Similar Arabidopsis Homologs of Trithorax ATX1 and ATX2 Encode Proteins with Divergent Biochemical Functions. *Plant Cell* **20**, 568–579 (2008).

58. Shafiq, S., Berr, A. & Shen, W.-H. Combinatorial functions of diverse histone methylations in Arabidopsis thaliana flowering time regulation. *New Phytol.* **201**, 312–322 (2014).

59. Keurentjes, J. J. B. *et al.* Development of a Near-Isogenic Line Population of Arabidopsis thaliana and Comparison of Mapping Power With a Recombinant Inbred Line Population. *Genetics* **175**, 891–905 (2007).

60. Simon, M. *et al.* Quantitative Trait Loci Mapping in Five New Large Recombinant Inbred Line Populations of Arabidopsis thaliana Genotyped With Consensus Single-Nucleotide Polymorphism Markers. *Genetics* **178**, 2253–2264 (2008).

61. Botto, J. F., Alonso-Blanco, C., Garzarón, I., Sánchez, R. A. & Casal, J. J. The Cape Verde Islands Allele of Cryptochrome 2 Enhances Cotyledon Unfolding in the Absence of Blue Light in Arabidopsis. *Plant Physiol.* **133**, 1547–1556 (2003).

62. el-Assal, S. E. D., Alonso-Blanco, C., Hanhart, C. J. & Koornneef, M. Pleiotropic effects of the Arabidopsis cryptochrome 2 allelic variation underlie fruit trait-related QTL. *Plant Biol. Stuttg. Ger.* **6**, 370–374 (2004).

63. Tessadori, F., Schulkes, R. K., Driel, R. van & Fransz, P. Light-regulated large-scale reorganization of chromatin during the floral transition in Arabidopsis. *Plant J.* **50**, 848–857 (2007).

64. Sanchez-Bermejo, E. *et al.* Genetic Architecture of Natural Variation in Thermal Responses of Arabidopsis1[OPEN]. *Plant Physiol.* **169**, 647–659 (2015).

65. Balasubramanian, S., Sureshkumar, S., Lempe, J. & Weigel, D. Potent Induction of Arabidopsis thaliana Flowering by Elevated Growth Temperature. *PLOS Genet.* **2**, e106 (2006).

66. Phee, B.-K. *et al.* A novel protein phosphatase indirectly regulates phytochrome-interacting factor 3 via phytochrome. *Biochem. J.* **415**, 247–255 (2008).

67. Blanvillain, R., Wei, S., Wei, P., Kim, J. H. & Ow, D. W. Stress tolerance to stress escape in plants: role of the OXS2 zinc-finger transcription factor family. *EMBO J.* **30**, 3812–3822 (2011).

68. Song, Y. H. *et al.* Distinct roles of FKF1, Gigantea, and Zeitlupe proteins in the regulation of Constans stability in Arabidopsis photoperiodic flowering. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17672–17677 (2014).

69. Ding, Y., Avramova, Z. & Fromm, M. The Arabidopsis trithorax-like factor ATX1 functions in dehydration stress responses via ABA-dependent and ABA-independent pathways. *Plant J. Cell Mol. Biol.* **66**, 735–744 (2011).

70. Léran, S. *et al.* A unified nomenclature of NITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER family members in plants. *Trends Plant Sci.* **19**, 5–9 (2014).

71. Connallon, T. & Hodgins, K. A. Allen Orr and the genetics of adaptation. *Evolution* **n/a**, (2021).

72. Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.* 1–13 (2020) doi:10.1038/s41576-020-0250-z.

73. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–631 (2010).

74. Alonso-Blanco, C. *et al.* 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).

75. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).

76. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

77. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Res.* **44**, D1167–D1171 (2016).

78. Lopez-Arboleda, W. A., Reinert, S., Nordborg, M. & Korte, A. Global Genetic Heterogeneity in Adaptive Traits. *Mol. Biol. Evol.* **38**, 4822–4831 (2021).

79. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

80. El-Din El-Assal, S. *et al.* The Role of Cryptochrome 2 in Flowering in Arabidopsis. *Plant Physiol.* **133**, 1504–1516 (2003).

81. 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).

82. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

83. Kruijer, W. *Heritability*. (2019).

84. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

85. Krishnakumar, V. *et al.* ThaleMine: A Warehouse for Arabidopsis Data Integration and Discovery. *Plant Cell Physiol.* **58**, e4 (2017).

86. Bomblies, K. *et al.* Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. *PLoS Genet.* **6**, e1000890 (2010).

DISCUSSION

We presented here a model system to study adaptation and map the genetic basis of a fitness-related trait – flowering time – making use of *Arabidopsis thaliana* natural populations in the Cape Verde archipelago.

First, we reconstructed the demographic history of CVI based on genome-wide variation of 335 newly collected accessions of A. thaliana, from two Cape Verdean islands: Santo Antão and Fogo. Genetic clustering analyses suggested that a North African population, genetically similar to the modern Moroccan population, might have been the colonizing population of CVI. The initial colonization took place 5-7 kya, possibly aided by the more favorable conditions during the last pluvial period in Africa, from North Africa to Santo Antão. Soon after that, Fogo was colonized from Santo Antão. The initial colonization bottleneck was so strong that erased >99% of all pre-existing variation, suggesting that all variation found today in CVI resulted from new mutations. The differentiation between the mainland and island populations is similar to that observed between species pairs in the Arabidopsis genus 1. Since the split, the two island populations have evolved in isolation and therefore share less than 1% of segregating variants. The severe colonization bottlenecks allied to the small number of colonizers (<50 individuals in both islands) may help to explain the low genetic diversity observed in the current populations. Therefore, the resulting small mutation-limited populations would need to rely on de novo variation to adapt to their new fitness optimum after colonization 2,3.

Long-range colonization events represent a sudden change in the environment for the colonizing population and bring with them multiple challenges. In our system, climate variables such as precipitation, aridity, growing season or temperature showed significant differences between the Moroccan habitat and CVI, suggesting that the initial colonizers would have been challenged by multiple aspects of the novel CVI environment. This is consistent with results from species distribution modeling which predicted no suitable regions for Moroccan *Arabidopsis* in Cape Verde and significant dissimilarities between the two habitats. Adaptation to these strong selective pressures left genomic signals. We found signals of positive selective from population genetic analyses and examination of known functional variants in Cvi-0 (compared to Ler-0), suggesting that CVI *Arabidopsis* is adapting to the harsh Cape Verde environment. Adding on these results, we also found higher fitness on the island populations compared to the outgroup, under CVI simulated conditions, implying that indeed the island populations are adapting to the novel environment. Adaptation to the much shorter growing season in Cape Verde was regulated by a reduction in flowering time, which happened in parallel on both islands and is negatively correlated with seed production. Plants that flowered early enough were

able to produce descendants by avoiding the long dry season in CVI and consequently producing more seeds.

The adaptive reduction in flowering time was achieved through convergent evolution using novel variation that arose after colonization. Vernalization requirement (the need of a cold period to induce flowering) was lost by knocking out two core genes in this pathway: *FRI* K232X in Santo Antão and *FLC* R3X in Fogo. Functional variation in *FRI* and *FLC* is widespread in natural populations of Eurasian *A. thaliana* [4–10]. Variation in *FRI* is primarily reigned by loss of function mutations in coding regions [6,11], while most putative functional variation in *FLC* is found in the first intron, which contains a well-characterized regulatory element [9,12]. Loss of function of either of these genes results in loss of the vernalization requirement and is responsible for 85% of flowering time variation in Eurasian accessions [7,13]. Despite functional variants at these core genes being well characterized and associated with clinal patterns of phenotypic and climatic variation, suggesting an adaptive effect, specific alleles have not yet been confidently associated with fitness differentials due to many confounding factors in Eurasian populations, such as the complex demography of most populations, population structure, or heterogenous and spatially differentiated ecological drivers. Our unique natural experiment in the isolated Cape Verde islands allowed us to definitively connect mutations that occurred in parallel at *FRI* and *FLC* with adaptive divergence.

In parallel, both CVI-specific mutations were driven to high frequency by strong selection. In Santo Antão, strong selection favored early flowering through *FRI* K232X, which then permitted the establishment of new populations across the drier regions of the island. In the more arid Fogo, population size increased in the same time frame when *FLC* 3X arose and fixed, suggesting that this mutation enabled evolutionary rescue in Fogo. Our results show that adaptation in CVI fits the theoretical concept of an adaptive walk proposed by Orr [2,3]. Under this model, a small, mutation-limited population facing a new environment far from its previous adaptive optimum, initially relies on *de novo* large effect mutations to adapt and escape extinction. When selection is strong enough to overcome drift, so that the beneficial mutations can escape stochastic loss, these will rise in frequency until fixation. This is particularly important in small populations where genetic drift is stronger [2,3,14,15].

Although these large effect mutations likely represented the first steps in the adaptive process, others with smaller effects were also expected based on the polygenicity of flowering time [16–20]. On the archipelago, we found an oligogenic architecture for flowering time, with few large effect mutations and many small effect candidate loci affecting the trait, following an exponential

distribution of effect sizes. In the outgroup Morocco, on the other hand, we found a very polygenic architecture, with many loci affecting the trait and following a uniform distribution of effects.

The observed differences in genetic architecture between the continental and island populations were expected based on long-term N_e and genetic diversity differences between the two [21,22]. Large populations, such as the outgroup Moroccan, have more beneficial mutations at their disposal due to the large N_e, high mutational input and high diversity. When the environment changes – being by colonization of new habitats or climatic changes –, these populations can effectively adapt relying on standing genetic variation [23–25]. However, these events are often associated with strong bottlenecks that reduce standing genetic variation available for adaptation [26,27]. The resulting small populations, such as the CVI populations after colonization, tend to adapt rapidly through few *de novo* large effect mutations, because the low N_e, low mutational input and low genetic diversity reduce the number of beneficial mutations available

Orr's model also predicts that, once the initial large effect mutations reach fixation, other with smaller effects will arise and rise in frequency in a diminishing returns fashion, generating an exponential distribution of effect sizes [2,3]. These expectations were met in CVI but not in Morocco. Consistent with Orr's take on Fisher's geometrical model, the candidate loci identified in the small CVI populations followed an exponential distribution of effect sizes, while the continental Moroccan population followed a uniform distribution, in line with Fisher's infinitesimal model. Moreover, the two populations also presented differences in terms of impact of variants and frequency: on the archipelago, two independent large effect loss of function mutations rose to high frequencies and were accompanied by a small number of moderate impact and modifier mutations. Contrastingly, flowering time variation in the mainland population is regulated by several non-coding, possibly regulatory, mutations at intermediate frequencies. These results fit with expectations that large effects alleles caused by loss of function of key genes are more likely to contribute after a sudden change to a distant optimum and to rise to fixation [23,28–30].

Our CVI *Arabidopsis* system provides useful information for forecasting risk for vulnerable populations and species. This is particularly important for small, isolated populations that face higher extinction risk and need to escape extinction in a race with the clock [26,31]. Adaptation in CVI fits well with models of rapid adaptation and evolutionary rescue, in-line with Orr's theoretical concept of an adaptive walk [2,3,31]. Under this model, a small, mutation-limited population facing a new environment

far from its previous adaptive optimum will rely on *de novo* large effects mutations to adapt. When selective pressures are strong, these mutations will be able to escape the strong effect of genetic drift and rise in frequency [14,26]. The large effect mutations that arise at the beginning of the adaptive process will then be followed by smaller effect variants, in which is described by Orr as an 'adaptive walk' [2,3,32], following an exponential distribution of effects. Moreover, with the CVI *Arabidopsis*-Morocco contrast, our findings exemplify the effects of population history on the genetic architecture of a quantitative trait. Further, this system provides an illustration of adaptation after a sudden environmental shift in a natural set-up and an example of genetic architecture characterization of a fitness-related trait, linking specific functional variants to fitness differentials.

References

1. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

2. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).

3. Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56**, 1317–1330 (2002).

4. Johanson, U. *et al.* Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290**, 344–347 (2000).

5. Korves, T. M. *et al.* Fitness effects associated with the major flowering time gene FRIGIDA in *Arabidopsis thaliana* in the field. *Am Nat* **169**, E141-157 (2007).

6. Le Corre, V., Roux, F. & Reboud, X. DNA polymorphism at the FRIGIDA gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. *Mol Biol Evol* **19**, 1261–1271 (2002).

7. Shindo, C. *et al.* Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*. *Plant Physiol* **138**, 1163 (2005).

8. Stinchcombe, J. R. *et al.* A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. *Proc Natl Acad Sci USA* **101**, 4712–4717 (2004).

9. Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. & Purugganan, M. D. Epistatic interaction between *Arabidopsis* FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proc Natl Acad Sci USA* **101**, 15670–15675 (2004).

10. Méndez-Vigo, B., Picó, F. X., Ramiro, M., Martínez-Zapater, J. M. & Alonso-Blanco, C. Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in *Arabidopsis*. *Plant Physiol.* **157**, 1942–1955 (2011).

11. Zhang, L. & Jiménez-Gómez, J. M. Functional analysis of FRIGIDA using naturally occurring variation in *Arabidopsis thaliana*. *Plant J.* **103**, 154–165 (2020).

12. Sheldon, C. C., Conn, A. B., Dennis, E. S. & Peacock, W. J. Different regulatory regions are required for the vernalization-induced repression of FLOWERING LOCUS C and for the epigenetic maintenance of repression. *Plant Cell* **14**, 2527–2537 (2002).

13. Bloomer, R. H. & Dean, C. Fine-tuning timing: natural variation informs the mechanistic basis of the switch to flowering in *Arabidopsis thaliana*. *J. Exp. Bot.* **68**, 5439–5452 (2017).

14. Gillespie, J. H. Some properties of finite populations experiencing strong selection and weak mutation. *Am. Nat.* **121**, 691–708 (1983).

15. Gillespie, J. H. *The causes of molecular evolution*. (Oxford University Press, 1991).

16. Zan, Y. & Carlborg, Ö. A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected Arabidopsis thaliana. *Mol. Ecol. Resour.* **18**, 798–808 (2018).

17. Zan, Y. & Carlborg, Ö. A Polygenic Genetic Architecture of Flowering Time in the Worldwide Arabidopsis thaliana Population. *Mol. Biol. Evol.* **36**, 141–154 (2019).

18. Andrés, F. & Coupland, G. The genetic basis of flowering responses to seasonal cues. *Nat. Rev. Genet.* **13**, 627–639 (2012).

19. Mouradov, A., Cremer, F. & Coupland, G. Control of flowering time: interacting pathways as a basis for diversity. *Plant Cell* **14**, S111–S130 (2002).

20. Salomé, P. A. *et al.* Genetic architecture of flowering-time variation in *Arabidopsis thaliana*. *Genetics* **188**, 421–433 (2011).

21. Dittmar, E. L., Oakley, C. G., Conner, J. K., Gould, B. A. & Schemske, D. W. Factors influencing the effect size distribution of adaptive substitutions. *Proc. R. Soc. B Biol. Sci.* **283**, 20153065 (2016).

22. Vahdati, A. R. & Wagner, A. Population Size Affects Adaptation in Complex Ways: Simulations on Empirical Adaptive Landscapes. *Evol. Biol.* **45**, 156–169 (2018).

23. Barghi, N., Hermisson, J. & Schlötterer, C. Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.* 1–13 (2020) doi:10.1038/s41576-020-0250-z.

24. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nat. Rev. Genet.* **3**, 11–21 (2002).

25. Höllinger, I., Pennings, P. S. & Hermisson, J. Polygenic adaptation: From sweeps to subtle frequency shifts. *PLOS Genet.* **15**, e1008035 (2019).

26. Whitlock, M. C. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* **54**, 1855–1861 (2000).

27. Willi, Y., Buskirk, J. & Hoffmann, A. A. Limits to the Adaptive Potential of Small Populations. *Annu. Rev. Ecol. Evol. Syst.* **37**, 433–458 (2006).

28. Arjan G., J. *et al.* Diminishing Returns from Mutation Supply Rate in Asexual Populations. *Science* **283**, 404–406 (1999).

29. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243–1247 (2009).

30. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).

31. Orr, H. A. & Unckless, R. L. Population extinction and the genetics of adaptation. *Am. Nat.* **172**, 160–169 (2008).

32. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 563–564 (1970).

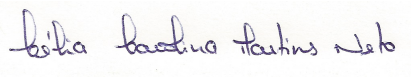
Erklarung

Erklärung zur Dissertation

gemäß der Promotionsordnung vom 12. März 2020

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht."

November 2021



Célia Carolina Martins Neto

Acknowledgments

This PhD and this entire journey are a huge part of my life now. Not only for everything I learned scientifically (and looking back now it was massive!) but also because it exposed me to new realities, contexts, and people, pushing me outside my comfort zone. However long (and sometimes tortuous), this path was worth it. So, to everyone (even slightly) linked to the past 5 years of my life, thank you for making me grow!

First of all, to Angela. It is not an exaggeration to say that without you this wouldn't be possible: not only you gave me the chance to be part of this exciting and challenging project, but also because you were always there, helping, supporting, guiding me and trying to make me a better scientist. There was a lot of ups and downs, trials and errors, but at the end I think we managed to do good science together. Thank you.

To my defense committee, Prof. Dr. Ute Höcker and Prof. Dr. Joachim Krug, for accepting to be a part of this process. I hope this goes well…

To my TAC committee, Prof. Dr. George Coupland and Prof. Dr. Benjamin Stich, for their support and guidance. George specially has played a great part in this project with his always-insightful comments and feedback. Thank you for always being available for me.

To the MPIPZ community for the friendly and good scientific environment but also to the institute for all the facilities. Specially, to the Coupland department for the scientific discussions; to the PhD Coordinators, Johanna and Stephan, for the care, the support and some sense of stability and organization; and to the greenhouse staff, without whom this project would have taken several years more. Thank you for caring about the plants – physiologically but also worrying every time we wanted to kill them… -- and for being always available to try new things, even though many of these come from really crazy ideas.

To all the Hancocks. Thank you for the scientific (and not so much) discussions, for always being available for my annoying questions and for the coffee/tea/cake breaks. To Pádraic, who adopted me and took me into the plant world. We fought but I had fun! I need someone to defy me. To Andrea, who also adopted me but pushed me to the PopGen world. Hope you have enjoyed as much as I did the long hours in the chamber or over an endless paper (which is not done yet, BTW), filled with overwhelming scientific questions. To Nina, who kept us on our toes, organized and prepared. To Mehmet, who is not only an amazing colleague but also a good friend, always ready for some uncalled fun fact about tea. And to my fellow Cape Verde adventurers, Manu and Ahmed. You guys were a solid

scientific help in Cologne, but also great fieldtrip companions. Despite everything – from sand storms, lost luggage, malaria to F3 or F4 --, those were fun times that made me stronger.

Talking about Cape Verde, an unmeasurable thanks to all the friends I made there. To Herculano and Lindim, for the more serious part of the fieldtrip – which was never serious at all. Your knowledge about Cape Verde and its fauna and flora was valuable and very needed. To Zenita and David, and Alain and Luci, who always treated us with more hospitality than simple guests and made us feel a little bit at home so far from home. Obrigada pela morabeza!

To the Storchenweg community. To every one of you (and you guys are many!!), thank you. You gave me a home and a family, that – as pretty much all families – went through a lot and survived. You enriched my world by bringing in with you a little bit of your own world and circumstances, and made me grow as a social being (I'm still working on this…). Even though every single one of you played a special role in the last 4 years, some have gained a particular place. Theresa, Xuan, Marco, Jack, Mehmet, Chloé, Rigel and Ale, thank you for everything. You have accepted me despite all my flaws and taught me how to be a better friend. From pushing the boundaries of Science to simply folding Aracons, from ferocious games of Exploding Kittens to just listening to my complaints and rantings over a glass of wine at the end of an endless day, you are a part of me. As someone wise once said, to Marco!

Aos meus amigos de outras andanças, que começaram por ser científicas, mas acabaram em muito mais. Embora o contacto não seja diário, entre Christmas markets e Rebeldes em fofocas pela noite fora, ajudaram-me a manter-me sã e a continuar.

Ao Renato. Não há muito para te dizer que já não te tenha dito. Obrigada por estares sempre aqui – não tanto como gostaria fisicamente -- , mas estás sempre por perto. Obrigada pelo apoio e pela força, pelo fortalecer da minha autoestima mas também por me mandares a um sítio quando preciso. Obrigada por me aturares há tanto tempo – passados já tantos graus académicos – e por teres resistido mais um bocadinho e sobrevivido ao meu doutoramento também. Espero ter-te comigo na próxima graduação....

Por fim, mas de todo menos importante, ao meus pais e à Lena. Mãe, pai, obrigada pelo apoio incondicional e pela segurança que me dão e que me permite ser livre e ir atrás do que eu quero. Lena, obrigada por me mostrares que há mais coisas além da Ciência – a maioria é parvoíce mas existem.

Obrigada!