The rate and potential relevance of new mutations in a colonizing plant lineage

Moises Exposito-Alonso, Claude Becker, Verena J. Schuenemann, Ella Reiter, Claudia Setzer, Radka Slovak, Benjamin Brachi, Jörg Hagmann, Dominik G. Grimm, Jiahui Chen, et al.

PLOS GENETICS

RESEARCH ARTICLE

The rate and potential relevance of new mutations in a colonizing plant lineage

Moises Exposito-Alonso[1,2‡], Claude Becker[1‡], Verena J. Schuenemann[3,4], Ella Reiter[3], Claudia Setzer[5], Radka Slovak[5], Benjamin Brachi[6¤a], Jörg Hagmann[1¤b], Dominik G. Grimm[1¤c], Jiahui Chen[6,7], Wolfgang Busch[5¤d], Joy Bergelson[6], Rob W. Ness[8], Johannes Krause[3,4,9], Hernán A. Burbano[2]*, Detlef Weigel[1]*

1 Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany, **2** Research Group for Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany, **3** Institute of Archaeological Sciences, University of Tübingen, Tübingen, Germany, **4** Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, Tübingen, Germany, **5** Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria, **6** Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America, **7** Institute of Tibet Plateau Research, Chinese Academy of Sciences, Beijing, China, **8** Department of Biology, University of Toronto Mississauga, Mississauga, Ontario, Canada, **9** Department of Archeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany

¤a Current address: INRA, UMR 1202 Biodiversité Gènes & Communautés, Cestas, Bordeaux, France
¤b Current address: Computomics, Tübingen, Germany
¤c Current address: Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
¤d Current address: Salk Institute for Biological Studies, La Jolla, California, United States of America
‡ MEA and CB share co-first authorship of this work.
* hernan.burbano@tuebingen.mpg.de (HAB); weigel@weigelworld.org (DW)

Citation: Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, et al. (2018) The rate and potential relevance of new mutations in a colonizing plant lineage. PLoS Genet 14(2): e1007155. https://doi.org/10.1371/journal.pgen.1007155

Editor: Jeffrey Ross-Ibarra, University of California Davis, UNITED STATES

Received: September 24, 2017

Accepted: December 13, 2017

Published: February 12, 2018

Data Availability Statement: Short reads have been deposited in the European Nucleotide Archive under the accession number PRJEB24619 and are available at https://www.ebi.ac.uk/ena/data/view/PRJEB24619.

Funding: This study was supported by the President's Fund of the Max Planck Society (project "Darwin") to HAB and by an ERC grant (AdG IMMUNEMESIS) and core funds of the Max Planck Society to DW. The funders had no role in

Abstract

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

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

to a combination of modern and historic samples from colonizing lineages, we can directly study new mutations and their potential evolutionary relevance.

Author summary

A consequence of an increasingly interconnected world is the spread of species outside their native range—a phenomenon with potentially dramatic impacts on ecosystem services. Using population genomics, we can robustly infer dynamics of colonization and successful population establishment. We have compared hundred genomes of a single *Arabidopsis thaliana* lineage in North America, including genomes of contemporary individuals as well as 19th century herbarium specimens. These differ by an average of about 200 mutations, and calculation of the nuclear evolutionary rate enabled the dating of the initial colonization event to about 400 years ago. We also found mutations associated with differences in traits among modern individuals, suggesting a role of new mutations in recent adaptive evolution.

Introduction

Colonizing or invasive populations sampled through time [1,2] constitute "natural experiments" where it is possible to study evolutionary processes in action [3]. Colonizations, which are dramatically increasing in number [4,5], sometimes are characterized by strong bottlenecks and genetic isolation [6,7], and thus greatly facilitate the observation of new mutations and potentially their effects under natural population dynamics and selection [8]. Colonizations thus offer a complementary approach to other studies of new mutations, which often minimize natural selection, for example in laboratory mutation accumulation experiments [9] and parent-offspring comparisons [10]. The study of colonizations is also complementary to the investigation of genetic divergence over long time scales, e.g., between distant species [11], where the results are largely independent of short-term demographic fluctuations. There is broad interest in understanding how genetic diversity is generated [12], and how new mutations can provide a path for rapid adaptive evolution [13–15]. Additionally, accurate evolutionary rates permit dating historic population splits, which is fundamental to the study of population history [16].

The analysis of colonizing populations can also contribute to resolving the "genetic paradox of invasion" [17]. This paradox comes from the observation that colonizing populations can be surprisingly successful and spread very widely and in multiple even when strongly bottlenecked, suggesting some level of adaptation to new environments that goes beyond the exploitation of unoccupied ecological niches [17]. Much of the work in plant ecology and evolution has focused on evidence that populations can rapidly adapt from standing variation [18]. In invasive lineages, initial standing variation may originate from incomplete bottlenecks, multiple introductions, or admixture with local relatives [19]. Much less work has been done with respect to the role of *de novo* mutations as a solution to the genetic paradox of invasion, although this has been proposed as an alternative explanation for rapid adaptation by colonizing lineages [3,17,20].

The self-fertilizing plant *Arabidopsis thaliana* is native to Africa and Eurasia [21,22] but has recently colonized N. America, where it likely experienced a strong founder effect [23]. At nearly half of N. American sites sampled during the 1990s and early 2000s, more than 80% of

plants belong to a single haplogroup, HPG1, as inferred from genotyping with 149 intermediate-frequency markers evenly spread throughout the genome [23]. The HPG1 lineage has been reported from many sites along the East Coast and in the Midwest as well as at a few sites in the West [23] (Fig 1, S1 Table). The great ubiquity of HPG1 in comparison to any other haplogroup could be due to either some adaptive advantage, or, more parsimoniously, be the result of HPG1 being derived from one of the first arrivals of *A. thaliana* in the continent.

Here, we focus on 100 HPG1 individuals that do not show any evidence of outcrossing with other lineages. We combine genomes from herbarium specimens and live individuals, collectively covering the time span from 1863 to 2006, to infer mutation rates, to date the birth of the HPG1 lineage, and to investigate the evolutionary forces that shape genetic diversity and potentially adaptive trait variation. Our analyses of this lineage serves as a model for future studies of similar colonizing or otherwise recently bottlenecked plant populations, in order to better understand how diversity is generated and to which extent it contributes to adaptation in nature.

Results and discussion

Historic and modern genomes

In a self-fertilizing species, a single individual can give rise to an entire lineage of millions of offspring, which then diversify through new mutations and eventually intra-lineage recombination. If self-fertilization is much more common than outcrossing, the founder is likely to have been homozygous throughout almost the entire genome. Because it is so wide spread, HPG1 presents an opportunity to sample many natural populations that have been potentially derived from a common, very recent ancestor with such characteristics. In the best possible case, this would allow for new mutations to be directly observed through time. To test these assumptions and to better understand the evolution of HPG1, we sequenced two different groups of plants. The first group were live descendants of 87 plants that had been collected between 1993 and 2006 (Fig 1; S1 Table), and which had been identified as likely members of the HPG1 lineage with 149 genome-wide markers spaced at roughly 1-Mb-intervals [23]. We aimed for broad geographic representation, with at least two accessions per collection site, where available. The second group comprised 36 herbarium specimens, collected between 1863 and 1993, for which we had no a priori information whether they may or may not belong to the HPG1 lineage, but which were selected from the herbarium records to cover the full historical geographic range and overlap with modern samples when possible (Fig 1).

The DNA from the herbarium specimens showed biochemical features typical of ancient DNA (aDNA) from plants, which we have previously described in detail [24]. Such DNA damage included a median fragment length of 60 bp, an excess of C-to-T substitutions of about 2.5% at the first base of sequencing reads and a 1.5 to 1.8 fold enrichment of purines at DNA breakpoints (S1 Fig, S2 Text). To remove aDNA associated damage and produce high-quality genomes, chemically-repaired libraries (see Methods) were later sequenced. These reads were mapped against an HPG1 pseudo-reference genome [25], focusing on single nucleotide polymorphisms (SNPs) because the short sequence reads of herbarium samples preclude accurate calling of structural variants. Genome sequences were of high quality, with herbarium samples covering 96.8–107.2 Mb of the 119 Mb reference, and modern samples covering 108.0–108.3 Mb (S1 Table).

Genetic diversity of HPG1 and delineation from other lineages

We visualized the relationships between the sequenced historic and modern plants building a neighbor joining tree of all 123 samples and confirmed that the majority fell within an almost-

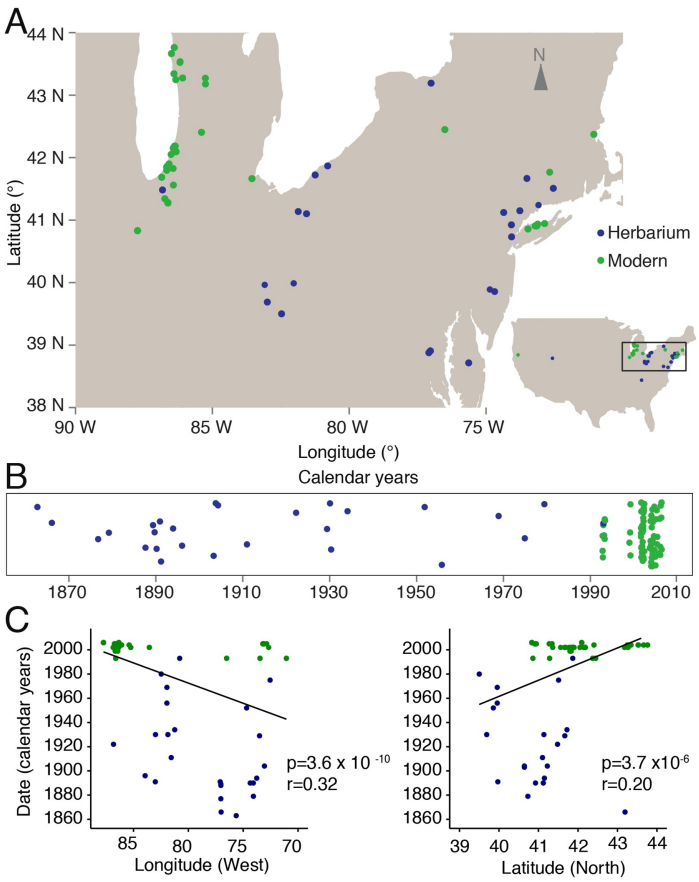

Fig 1. Geographic location and temporal distribution of HPG1 samples. (A) Sampling locations of herbarium (blue) and modern individuals (green). **(B)** Temporal distribution of samples (random vertical jitter for visualization purposes). **(C)** Linear regression of latitude and longitude as a function of collection year (p-value of the slope and Pearson correlation coefficient are indicated).

https://doi.org/10.1371/journal.pgen.1007155.g001

identical clade, the HPG1 (Fig 2A) [23]. Because any degree of introgression from other non-HPG1 lineages would confound the discovery of new mutations downstream, we removed all divergent samples and built a neighbour joining tree (n = 103 samples), which revealed that the HPG1 samples were very similar to each other, with very little within-population structure (Fig 2B). A parsimony network was used to detect recombinant genomes within this HPG1 clade (Fig 2C), which led us to remove three potential intra-lineage recombinants. Repeating the parsimony network cleared all previously inferred reticulations due to recombinations (Fig 2D). After such stringent filtering, we kept 27 of the 35 herbarium samples, and 73 of the 87 modern samples (S1 Table). These constitute a set of non-admixed, non-recombined and quasi-identical HPG1 individuals.

Pairs of HPG1 herbarium genomes differed by 28–207 SNPs genome-wide, pairs of HPG1 modern genomes by 2–259 SNPs, and pairs of historic-modern HPG1 genomes by 56–244 SNPs. That is, whole-genome identity was at least 99.9997% in any pairwise comparison. Of the approximately five to six thousand segregating SNPs in the HPG1 population, the vast majority, about 95% (Supplementary Text 3), have not been reported outside of this lineage [21]. Importantly, the density of SNPs along the genome was low and evenly distributed (typically fewer than 20 SNPs / 100 kb) with no peaks of much higher frequency, which makes us confident that chunks of introgressions from other lineages do not exist in this

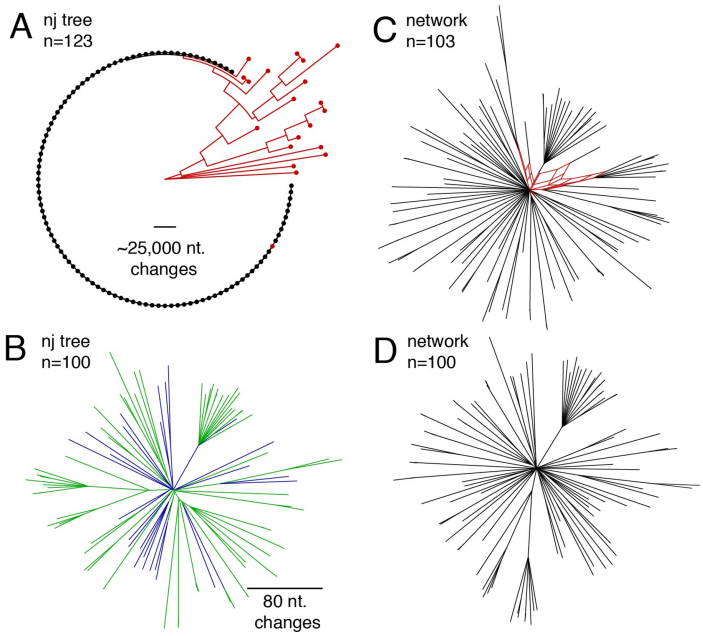

Fig 2. Relationship among herbarium and modern samples. (A) Neighbor joining tree with all 123 samples (dots) and rooted with the most distant sample. The black clade of almost-identical samples is the HPG1 lineage. Scale line shows the equivalent branch length of over 25,000 nucleotide changes. **(B)** Neighbor joining tree only with the HPG1 black clade from (A). Colors represent herbarium (blue) and modern individuals (green). Scale line shows the equivalent branch length of 80 nucleotide changes. Note that no outgroup was included. **(C, D)** Network of samples using the parsimony splits algorithm, before **(C)** and after **(D)** removing three intra-HPG1 recombinants (in red). Note that the network algorithm returns in (D) a network devoid of any reticulation, which indicates absence of intra-haplogroup recombination.

https://doi.org/10.1371/journal.pgen.1007155.g002

putatively pure HPG1 set (Fig 4). For comparison, random pairs of *A. thaliana* accessions from the native range or pairs of non-HPG1 typically differ by about 500 SNPs / 100 kb [21] (see scale in Fig 2A).

There were no SNPs in mitochondrial nor chloroplast genomes, which already suggested a recent common origin, and genome-wide nuclear diversity ($\pi = 0.000002$, $\theta_W = 0.00001$, with 5,013 full informative segregating sites) was two orders of magnitude lower than in the native range of the species ($\theta_W = 0.007$) [21] (S1 Table) (Supplementary Text 6). The population recombination parameter was also four orders of magnitude lower ($4N_e r = \rho = 3.0\text{x}10^{-6}$ cM bp^{-1}) than in the native range ($\rho = 7.5\text{x}10^{-2}$ cM bp^{-1}) [26] (Supplementary Text 6). While recombination occurs in every generation, regardless of self-fertilization or outcrossing, it is only observable after outcrossing between genetically non-identical individuals. We must stress that because *A. thaliana* can outcross at rates of several percent per generation [23,27], but because the HPG1 population is genetically so homogeneous, we are mostly "blind" to the consequences of outcrossing in this special case. The lack of "observable recombination" in the genome is important, as it allows for the use of straightforward phylogenetic methods to calculate a mutation rate. The enrichment of low frequency variants in the site frequency spectrum (Tajima's $D = -2.84$; species mean = -2.04, [21]) and low levels of polymorphism are consistent with a recent bottleneck followed by population expansion, which usually generates star-like phylogenies (Figs 2 and 3). The obvious explanation is that the strong bottleneck corresponds to a colonization founder event, likely by few closely related individuals or perhaps even a single plant.

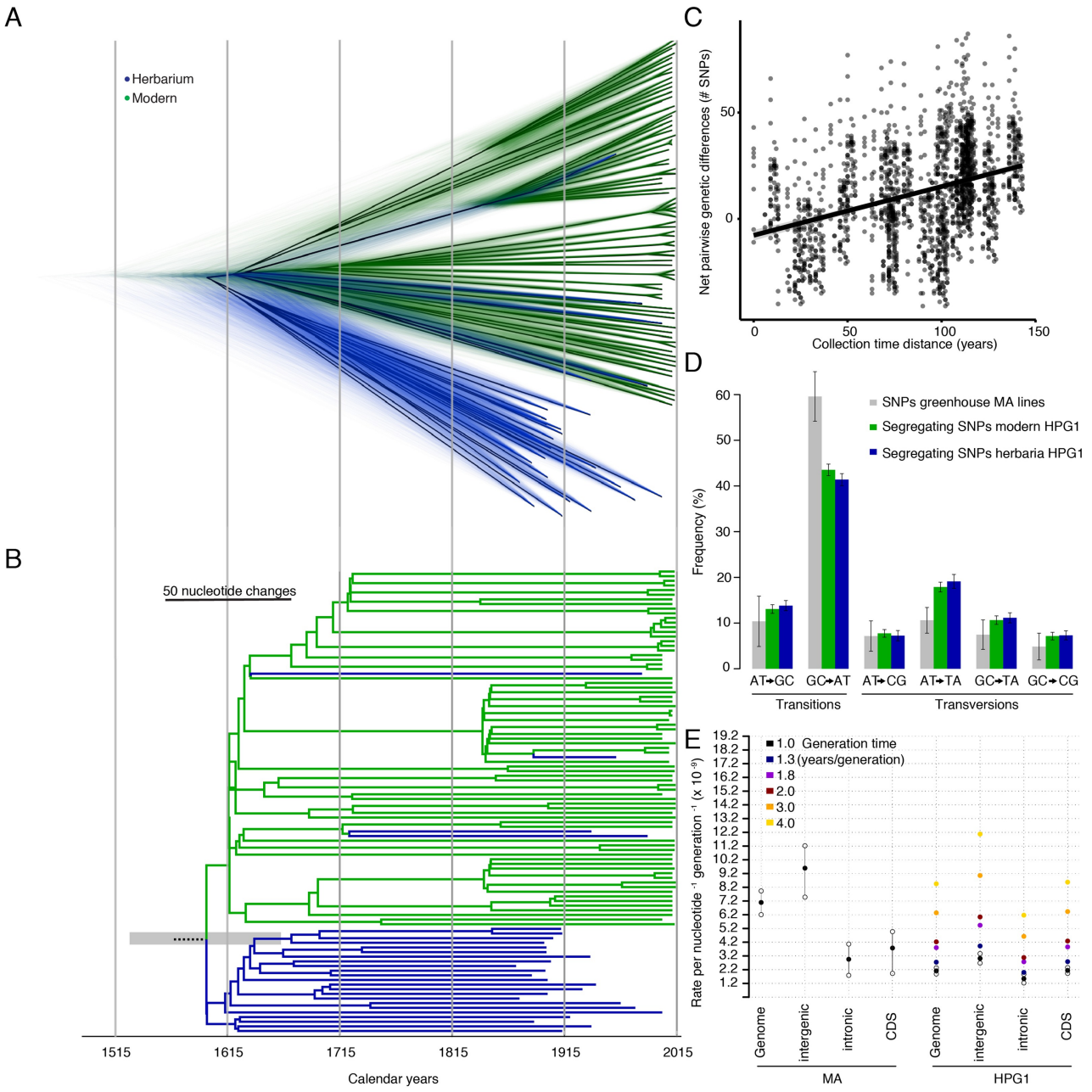

Fig 3. Substitution rates. (A) Bayesian phylogenetic analyses employing tip-calibration. A total of 10,000 trees were superimposed as transparent lines, and the most common topology was plotted solidly. Tree branches were calibrated with their corresponding collection dates. **(B)** Maximum Clade Credibility (MCC) tree summarizing the trees in (A). Note the scale line shows the equivalent branch length of 50 nucleotide changes. The grey transparent bar indicates the 95% Highest Posterior Probability of the root date. **(C)** Regression between pairwise net genetic and time distances. The slope of the linear regression line corresponds to the genome substitution rate per year. **(D)** Substitution spectra in HPG1 samples, compared to greenhouse-grown mutation accumulation (MA) lines. **(E)** Comparison of genome-wide, intergenic, intronic, and genic substitution rates in HPG1 and mutation rates in greenhouse-grown MA lines. Substitution rates for HPG1 were re-scaled to a per generation basis assuming different generation times. Confidence intervals in HPG1 substitution rates were obtained from 95% confidence intervals of the slope from 1,000 bootstraps (S4 Table for actual values).

https://doi.org/10.1371/journal.pgen.1007155.g003

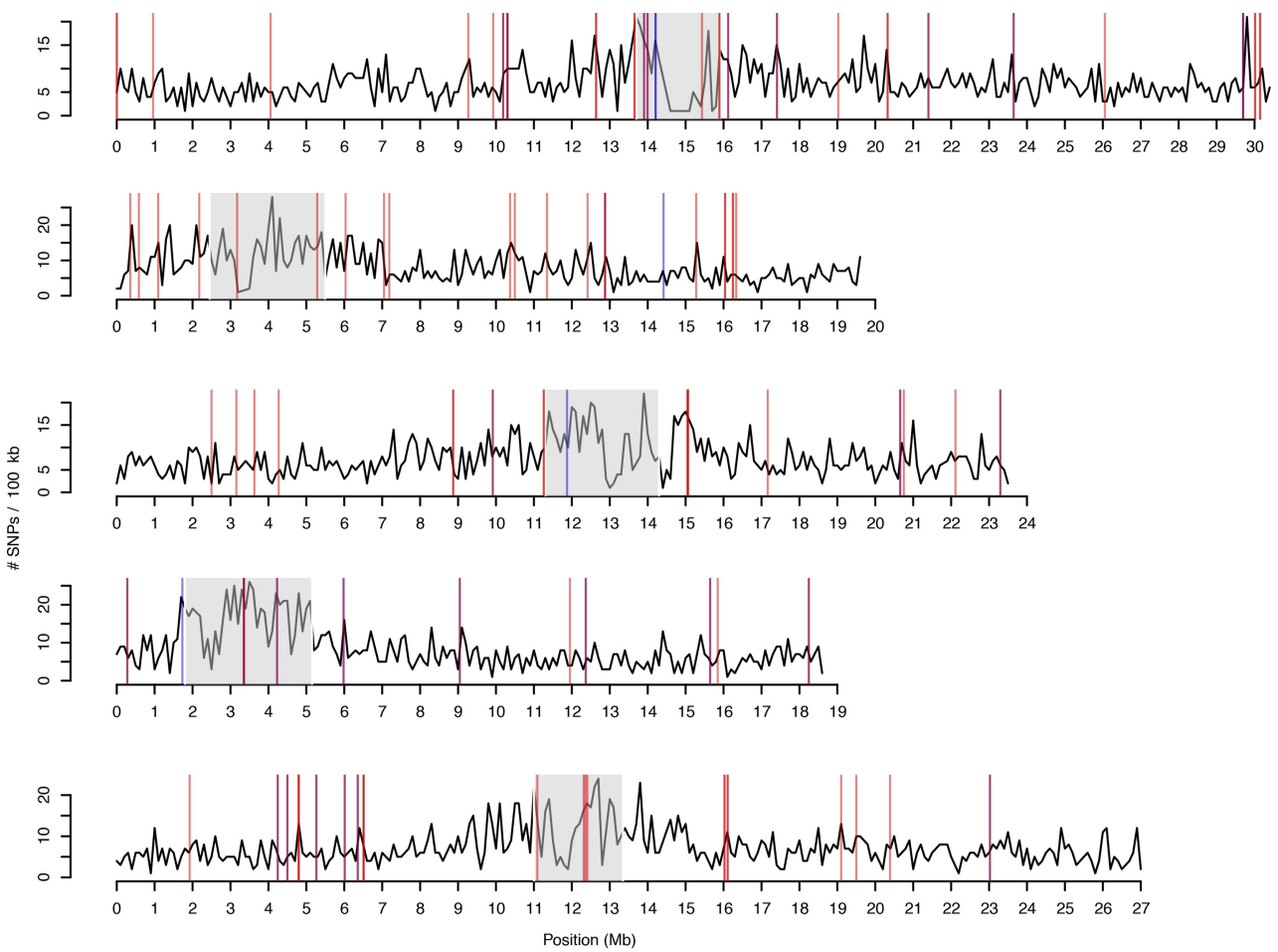

Fig 4. Density of SNPs along all chromosomes and location of GWAS hits. Black line shows number of SNPs per 100 kb window. Centromere locations are indicated by grey shading. Vertical lines indicate SNPs associated with root phenotypes (red) and climatic variables (blue) (Table 1 and S5 Table).

https://doi.org/10.1371/journal.pgen.1007155.g004

Altogether these patterns indicate that the collection of HPG1 plants we investigated constitute a quasi-clonal and quasi-identical set of individual genomes, mostly devoid of observable recombination and population structure, and thus eminently suited for the study of naturally arising *de novo* mutations.

The genome-wide substitution rate

It is important to distinguish between the *mutation rate*, which is the rate at which genomes change due to DNA damage, faulty repair, gene conversion and replication errors, and *substitution rate*, which is the rate at which mutations survive and accumulate under the influence of demographic processes and natural selection [28,29]. Under neutral evolution, mutation and substitution rates should be equal [29]. The simple evolutionary history of the HPG1 population enables direct estimates of substitution rates, and the comparison of theses between different genome annotations, as well as with mutation rates from controlled conditions experiments, could reveal the role played by both demographic and selective forces.

To estimate the substitution rate in the HPG1 lineage, we used distance- and phylogeny-based methods that take advantage of the known collection dates (Supplementary Text 7). The

distance method is independent of recombination and has been previously applied to viruses [30] and humans [31]. The substitution rate is calculated from correlation between differences in collection time in historic-modern sample pairs, and the number of nucleotide differences between those pairs relative to a reference (Fig 3C), scaled to the size of the genome accessible to Illumina sequencing. This method resulted in an estimated rate of $2.11x10^{-9}$ substitutions $site^{-1}$ $year^{-1}$ (95% bootstrap Confidence Interval [CI]: 1.88–$2.33x10^{-9}$) using rigorous SNP calling quality thresholds. Relaxing the thresholds for base calling and minimum genotyped rate affects both the number of called SNPs and the length of the interrogated reference sequence [32]. These largely cancelled each other out, and the adjusted estimates were relatively stable, between 2.1–$3.2x10^{-9}$ substitutions $site^{-1}$ $year^{-1}$ (S3 Table, Supplementary Text 3).

The second method, a Bayesian phylogenetic approach, uses the collection years for tip-calibration and assumes a relaxed molecular clock. It summarizes thousands of plausible coalescent trees, and it has been extensively used to calculate evolutionary rates in various organisms [33–35]. This method yielded a substitution rate of $4.0x10^{-9}$, with confidence ranges overlapping the above estimates (95% Highest Posterior Probability Density [HPPD]: 3.2–$4.7x10^{-9}$).

Based on the similar results obtained with two very different methods, we can confidently say that the substitution rate in the wild populations of HPG1 is between 2 and 5 $x10^{-9}$ $site^{-1}$ $year^{-1}$.

To date the colonization of N. America by HPG1 *A. thaliana* and to improve the description of intra-HPG1 relationships compared to that from a NJ tree, we further used a Bayesian phylogeny. At first sight, the 73 modern samples appeared separated from the herbarium samples (Fig 3B), but the superimposition of thousands of possible trees showed that the apparent separation of samples was less clear near the root (Fig 3A). Long terminal branches reflected that the majority of the variants are singletons, typical of populations that expand after bottlenecks.

The mean estimate of the last common HPG1 ancestor, the average tree root, was the year 1597 (HPPD 95%: 1519–1660) (Fig 3A and 3B), and an alternative non-phylogenetic method gave a similar estimate, 1625. Both estimates are older than a previously suggested date in the 19^{th} century, using a laboratory mutation rate estimate and having no information from herbarium samples [25]. Because HPG1 appears to have been the most abundant lineage in N. America since the 1860s, we believe it could have been one of the first, if not the first *A. thaliana* colonizer that could establish itself in N. America. If that is true, the time of coalescence of the HPG1 diversity could be close to the time of HPG1 introduction to N. America. During the colonial period, many European immigrants settled on the East coast, consistent with N. American *A. thaliana* lineages being genetically closest to British and coastal West European populations [21]. Coincidently, the oldest herbarium samples (12 out of the 27) were HPG1 and came from the East Coast, and we found a significant correlation between collection date and both latitude and longitude (Fig 1C). This could indicate that after the colonization they moved from the East Coast to the Midwest—the other main area of the distribution that experienced an agricultural expansion in the 19^{th} century [36]. Still, these conclusions need to be treated with caution, since regardless of the robustness of the results and our attempts to sample evenly from available collections, there could be unknown biases in the 19^{th} century herbaria.

Mutation spectra across genome annotations

Although for dating divergence events a substitution rate expressed in years is ideal, in order to compare substitution and mutation rates, both need to be expressed per generation. While *A. thaliana* is an annual plant, seed bank dynamics generate a delay of average generation time

at the population scale. A comprehensive study of multiple *A. thaliana* populations in Scandinavia found that dormant seeds could wait for longer than a year in the seed bank, generating overlapping generations and an delayed average generation time of 1.3 years [37] with a notable variance across populations. Multiplication by the mean generation time led to an adjusted rate of 2.7×10^{-9} substitutions site^{-1} generation^{-1} (95% CI 2.4–3.0×10^{-9}) (Fig 3E). To be able to compare this rate with a reference, we also re-sequenced mutation accumulation (MA) lines in the Col-0 reference background grown under controlled conditions in the greenhouse that had been analyzed before with less advanced short read sequencing technology [38]. From the new re-sequencing data, we obtained an updated rate of 7.1×10^{-9} mutations site^{-1} generation^{-1} (95% CI 6.3–7.9×10^{-9}) (S2 and S3 Tables Supplementary Text 4 and 7). This mutation rate is two- to three-fold higher than the per-generation substitution rate estimate in the wild, but within the same order of magnitude. The same holds for rates in different genome annotations, i.e. genic, intronic and intergenic regions, but the confidence intervals overlapped in many cases (S3 Table).

Differences in per-generation rates between laboratory and wild populations could stem from both methodological as well as biological causes. For instance, if the true average generation time was actually over 3 years / generation, the differences would cancel out (Fig 3E). Limitations in mapping structural variation in non-reference samples could lower the substitution rate, which may explain why we calculated an atypically low substitution rate in regions with transposable elements (see Supplementary Text 7.2.1). Environmentally-driven effects that are not yet well understood, such as variable methylation status of cytosines, account for much of the variation in local substitution rates [39], and could increase or decrease the rate (see Supplementary Text 7.2.3, S4 Fig).

An alternative evolutionary explanation to the aforementioned laboratory and wild populations' rates differences is that purifying selection in the wild would slow down the accumulation of mutations by removing deleterious mutations (Fig 3E). This has been observed before and is one of the accepted causes of the discrepancy between the so called long- and short-term substitution rates in a range of organisms [40].

In order to provide evidence for negative purifying selection acting in the wild, we performed three types of analyses involving comparisons across genomic annotations within the HPG1 dataset. Firstly, by calculating contingency tables and computing a Fisher's exact test, we compared the deviation of expected and observed SNPs between coding regions (more likely under purifying selection), with intergenic regions, intronic regions, and all non-coding regions of genome. All three pairwise comparisons showed a depletion of coding SNPs and an enrichment of intergenic, intronic and non-coding SNPs (odds ratio>2, $p<10^{-16}$). An obvious explanation is that in genome annotations where a mutation is more likely to be deleterious, i.e. coding regions, the number of observed variants should be lower due to selection having removed them from the population before we could sequence them.

Secondly, we studied the Site Frequency Spectrum (SFS) of genetic variants. The rationale was that because purifying natural selection is more efficient at removing intermediate-frequency variants, variants that tend to be deleterious or slightly deleterious should be found at lower frequency than those that only suffer neutral drift [41]. We built contingency tables of coding, intergenic, intronic and non-coding variants segregating above and below the conventional frequency cutoff of 5% to separate low- and intermediate-frequency variants [42]. We found that SNPs in coding regions were more likely to be at low frequency than those in intergenic (odds ratio = 2.34, $p = 3.09 \times 10^{-11}$), intronic (odds ratio = 1.48, p = 0.02), and all non-coding regions (odds ratio = 2.05, $p = 1.29 \times 10^{-8}$). We carried out the same analysis using nonsynonymous and synonymous SNPs, which are easily interpretable in terms of

the selection regimes under which they evolve. We did not find an enrichment (p = 0.67), perhaps due to an insufficient number of testable mutations (S3 Table).

Thirdly, to verify that the full frequency spectrum of coding SNPs was shifted to lower frequencies (i.e. the results were not dependent on the arbitrary 5% frequency cutoff), we used the nonparametric Kolmogorov-Smirnov test for two samples. We found that the cumulative distribution of the site frequency spectrum (CD_{SFS}) of coding regions is above (i.e., the frequency distribution is overall skewed to lower values) both the intergenic CD_{SFS} ($p = 3.25x10^{-6}$) and the non-coding regions CD_{SFS} (p = 0.001), but not the intronic CD_{SFS} (p = 0.60) (S5 Fig). As in our previous analysis, the comparison between the nonsynonymous and synonymous CD_{SFS} yielded, likely for similar reasons, no differences (p = 0.53).

All in all, these results support that purifying selection is a force shaping to some degree the diversity across the HPG1 genome and might therefore as well contribute to the differences between HPG1 and MA rates.

Potentially advantageous *de novo* mutations

Finally, having discovered over 5,000 *de novo* mutations in the HPG1 lineage, we wondered whether there is any evidence for an adaptive role of these *de novo* mutations in the colonization of N. America by HPG1. We noted that some new mutations had risen to intermediate or even high frequencies in the HPG1 samples. This might have been the consequence of drift from stochastic demographic processes, or it could have been caused by positive natural selection. To find direct evidence for the latter, we grew the modern accessions in a common garden and studied phenotypes of known importance in ecology of invasions [43], namely flowering time and root traits (see Supplementary Text 8). Using linear mixed models, we calculated the proportion of variance explained (also called narrow sense heritability, h^2) with a kinship matrix of all SNPs that had become common (>5%, n = 391). We found significant heritable variation for multiple traits including the growth rate in length ($h^2 = 0.64$) and the average root gravitropic direction ($h^2 = 0.54$). As in our study mutations are the main source of genetic variants, these mutations—or mutations linked to them—should be responsible for significant quantitative variation in several traits (S4 Table, Supplementary Text 10). The existence of mutation-driven phenotypic variation at least indicates that natural selection could have acted upon such phenotypic variation.

Although linkage disequilibrium (LD) among SNPs is high, the fact that HPG1 genomes differ in very few SNPs greatly reduces the list of candidate loci that might generate the observed phenotypic variation (S7 Fig) [44]. With this reasoning in mind and understanding the limitations imposed by LD, we carried out a genome-wide association (GWA) analysis and found 79 SNPs associated with one or more root traits, mostly growth and directionality (Fig 4). Twelve SNPs were in coding regions and seven resulted in nonsynonymous changes—some producing non-conservative amino-acid changes and thus likely to affect protein structure and/or function (Table 1, based on transition scores from [45]). Due to the aforementioned LD, in some cases the results of associations could not be confidently assigned to a specific SNP and thus we report the number of other associated mutations with $r^2 > 0.5$ (Table 1, S7 Fig). We note that linked genetic variation that has gone undetected (e.g., structural variation) could be causal rather than the identified SNPs. For some cases, however, we were able to pinpoint clear candidates that were not in LD with other SNPs and whose functional annotation had a strong connection to the phenotype (Table 1, S7 Fig). For example, one SNP associated with root gravitropism was not linked to any other SNP hit and it was found at 40% frequency (top 3% percentile). This SNP produces a cysteine to tryptophan change in AT5G19330, which is involved in abscisic acid response, strongly expressed in

Table 1. Genic SNPs associated with different traits. For nonsynonymous SNPs, the amino acid change and the Grantham score (ranging from 0 to 215), which measures the physico-chemical properties of the amino acids, are reported. All SNPs in the table were significant ($p < 0.05$) after raw p-values were corrected by an empirical p-value distribution from a permutation procedure. * highlights those that also passed a double Bonferroni threshold, correcting by number of SNPs and number of phenotypes ($p < 0.0001$). LD corresponds to how many other SNP hits are in high linkage ($r^2 > 0.5$). S5 Table contains information on all significant SNPs and S4 Table for details on phenotypes and climatic variables.

Trait†	Location (chr-bp)	Gene	Anno–tation	Protein	aa change	LD	Bonf.
G	1–958,948	AT1G03810	nonsyn	Oligonucleotide binding	A>P, 27	53	
D	1–13,994,958	AT1G36933	transposon	Copia		49	
S	1–20,324,050	AT1G54440	intronic	RRP6-LIKE 1		11	*
D	1–23,648,407	AT1G63740	nonsyn	TIR-NLR family	Y>S, 144	46	
G	2–358,395	AT2G01820	syn	RLK family		43	*
G	2–585,918	AT2G02220	syn	PSKR1		42	*
G	2–6,034,545	AT2G14247	syn	Expressed protein		38	*
G	2–7,047,529	AT2G16270	nonsyn	Unknown protein	P>A, 27	37	*
G	2–7,186,220	AT2G16580	intronic	SAUR8		36	*
G	2–10,495,275	AT2G24680	intronic	B3 family		34	*
G	2–12,415,084	AT2G28900	intronic	OEP16		32	
S	2–16,039,488	AT2G38290	3' UTR	AMT2		8	*
S	2–16,247,290	AT2G38910	nonsyn	CPK20	A>G, 60	7	*
G	2–16,333,662	AT2G39160	nonsyn	Unknown protein	A>G, 60	29	
G	3–2,500,258	AT3G07830	syn	PGA3		28	*
G	3–3,629,794	AT3G11530	intronic	VPS55		26	*
G	3–4,269,626	AT3G13229	5' UTR	DUF868 domain		25	*
D	3–11,873,293	AT3G30219	transposon	Gypsy		0	
G & D	4–4,228,138	AT4G07440	transposon	Oligonucleotide binding		19	
G & D	4–9,046,942	AT4G15960	nonsyn	Alpha/beta-hydrolase	A>Q, 24	18	
G & D	4–15,646,341	AT4G32410	syn	ANY1		15	
G	4–15,845,001	AT4G32840	3' UTR	PFK6		14	
D	5–4,245,213	AT5G13260	syn	Unknown protein		12	
D	5–4,500,202	AT5G13950	nonsyn	Unknown protein	A>G, 60	11	
G	5–4,797,923	AT5G14830	transposon	Retrotransposon		10	
G	5–6,508,329	AT5G19330	nonsyn	ARIA	C>W, 215	0	
G	5–11,090,365	AT5G29037	transposon	Gypsy		4	
G	5–12,312,975	AT5G32630	pseudogene	–		3	
G	5–12,358,159	AT5G32825	transposon	CACTA		2	
S	5–16,024,197	AT5G40020	intronic	Thaumatin superfamily		2	*

†Traits with significant associations were root gravitropism (G), size (S), or low summer precipitation.

https://doi.org/10.1371/journal.pgen.1007155.t001

growing roots, and confers salt tolerance when overexpressed [46]. Another nonsynonymous SNP associated with root growth is located in AT2G38910, which encodes a calcium-dependent kinase that is a factor regulating root hydraulic conductivity and phytohormone response *in vitro* [47,48].

Nineteen other SNPs were associated with climate variables after correction for latitude and longitude (www.worldclim.org, S4 Table), and generally tended to coincide with top root-associated SNPs (odds ratio = 3.9, Fisher's Exact test p = 0.002; Fig 4, and S5 Table). Specifically, this means that alleles increasing root length and gravitropic growth were present in areas with lower precipitation, and *vice versa* (Pearson's correlation r = 0.85, p = 0.003). This indicates that phenotypic variation generated by mutations coincides with environmental (and not

geographic) gradients along the colonized areas. Compared to other mutations with matched allele frequencies, root-associated mutations are first found in older herbarium samples nearer to Lake Michigan (S6 Fig), the area in N. America that seems to be most densely populated by *A. thaliana* [21]. A more densely spaced time series of samples would be needed to confirm the older age of specific mutations, but our observation could be explained by spatially varying selection across N. America, which may maintain antagonistic pleiotropic mutations for longer time than neutral mutations. The association of putatively adaptive mutation with climate variables could also be explained by such a phenomenon. Nevertheless, to confirm hypotheses of local adaptation by *de novo* mutations, it will be necessary to grow collections of divergent HPG1 individuals in multiple contrasting N. American locations over several years. Ideally, one would revive historical specimens to compare their performance to modern populations [49]. All in all, our results are compatible with natural positive selection having already acted on root morphology variation that was generated by *de novo* mutations in this colonizing lineage.

Conclusions

In summary, we have exploited whole-genome information from historic and contemporary collections of a herbaceous plant to empirically characterize evolutionary forces during a recent colonization. With this natural time series experiment we could directly estimate the nuclear substitution rate in wild *A. thaliana* populations—a parameter difficult to characterize experimentally [9]. This allowed us to date the colonization time and spread of HPG1 in N. America. We provide evidence that purifying selection has already changed the site frequency spectrum in the course of just a few centuries. Finally, we discovered that a small number of *de novo* mutations that rose to intermediate frequency can together explain quantitative variation in root traits across environments. This strengthens the hypothesis that some *de novo* variation could have had an adaptive value during the colonization and expansion process, a hypothesis that has been put forward as one of the possible solutions to the genetic paradox of invasion in plants [17]. This process might be more relevant in self-fertilizing plants, which typically have less diversity than outcrossing ones [50], but have higher growth rates [43] and account for the majority of successful plant colonizers [5]. While *A. thaliana* HPG1 is not an invasive, harmful species, it can teach us about fundamental evolutionary processes behind successful colonizations and adaptation to new environments. Our work should encourage others to search for similar natural experiments and to unlock the potential of herbarium specimens to study "evolution in action".

Methods

Sample collection and DNA sequencing

Modern *A. thaliana* accessions were from the collection described by Platt and colleagues [23], who identified HPG1 candidates based on 149 genome-wide SNPs (S1 Table, S1 Text). Herbarium specimens were directly sampled by Max Planck colleagues Jane Devos and Gautam Shirsekar, or sent to us by collection curators from various herbaria (S1 Table, S1 Text). Among the substantial number of specimens in the herbaria of the University of Connecticut, the Chicago Field Museum and the New York Botanical Garden, we selected herbarium specimens spaced in time so there was at least one sample per decade starting from the oldest record (1863). The differences in geographic biases of herbarium and modern collections are difficult to know [2], thus we did choose both historic and modern samples that were as regularly distributed in space as possible, and sample overlapping locations wherever possible. DNA from herbarium specimens was extracted as described [51] in a clean room facility at the University

of Tübingen. Two sequencing libraries with sample-specific barcodes were prepared following established protocols, with and without repair of deaminated sites using uracil-DNA glycosylase and endonuclease VIII (refs. [52–54]) (S2 Text). The reads of repaired libraries are available at https://www.ebi.ac.uk/ena/data/view/PRJEB24619. We also investigated patterns of DNA fragmentation and damage typical of ancient DNA [24] (S2 Text). DNA from modern individuals was extracted from pools of eight siblings using the DNeasy plant mini kit (Qiagen, Hilgendorf, Germany). Genomic DNA libraries were prepared using the TruSeq DNA Sample or TruSeq Nano DNA sample prep kits (Illumina, San Diego, CA), and sequenced on Illumina HiSeq 2000, HiSeq 2500 or MiSeq instruments. Paired-end reads from modern samples were trimmed and quality filtered before mapping using the SHORE pipeline v0.9.0 [25,55]. Because ancient DNA fragments are short (S1 Fig) we merged forward and reverse reads for herbarium samples after trimming, requiring a minimum of 11 bp overlap [51], and treated the resulting as single-end reads. Reads were mapped with GenomeMapper v0.4.5s [56] against an HPG1 pseudo-reference genome [25], and against the Col-0 reference genome, and SNPs were called with SHORE for the HPG1 pseudo-reference genome mappings [25,57] using different thresholds (Supplementary Text 3). Average coverage depth, number of covered genome positions, and number of SNPs identified per accession relative to HPG1 are reported in S1 Table. We also re-sequenced the genomes of twelve Col-0 MA lines [57,58] (S2 Table) (Supplementary text 4) to recalculate and update the laboratory mutation rate from Ossowski et al. [38] with the newer sequencing technologies.

Phylogenetic methods and genome-wide statistics

We used the Pegas, Ape and Adegenet packages in R [59–61] to manipulate and visualize the genetic distances of all samples as well as the HPG1 subset (Supplementary Text 7). We constructed parsimony networks using SplitsTree v.4.12.3 [62], with confidence values calculated with 1,000 bootstrap iterations. We built Maximum Clade Credibility Trees using the Bayesian phylogenetic tools implemented in BEAST v.1.8 [63] (see below).

Transforming the variant sites into a FASTA format, we estimated genetic diversity as Watterson´s θ [64] and nucleotide diversity π, and the difference between these two statistics as Tajimas's *D* [65] using DnaSP v5 [66]. Then we re-scaled the estimates using the sequencing-accessible genome sizes (S3 Table). We estimated pairwise linkage disequilibrium (LD) between all possible combinations of informative sites, ignoring singletons, by computing r^2, *D* and *D*' statistics using DnaSP v5 [66]. For the modern individuals, we calculated the recombination parameter rho ($4N_e r$) also using DnaSP v5 [66].

Substitution and mutation rate analyses

Similarly as in Fu et al. [67], we used genome-wide nuclear SNPs to calculate pairwise "net" genetic distances using the equation $D'_{ij} = D_{ic}\text{-}D_{jc}$, where D'_{ij} is the net distance between a modern sample *i* and a herbarium sample *j*; D_{ic} the distance between the modern sample *i* and the reference genome *c*; and D_{jc} is the distance between a modern sample (j) and the reference genome (c). We calculated a pairwise time distance in years between the collection times, *T*'ij, and calculated the linear regression: $D' = a+bT'$. The slope coefficient *b* describes the number of substitution changes per year. We used either all SNPs or subsets of SNPs at different annotations (genic, intergenic etc.) appropriately scaled by accessible genome length. Because the points used to calculate the regression are non-independent, a bootstrap has been recommended to overcome to a certain extent the anti-conservative confidence intervals [30] (Supplementary Text 7 and S3 Fig).

To fully account for the non-independence of points, we need to work with phylogenies. The Bayesian phylogenetics approach we used is implemented in BEAST v1.8 [63] and is called tip-calibration, and calculates a substitution rate along the phylogeny. Our analysis optimized simultaneously and in an iterative fashion using a Monte Carlo Markov Chain (MCMC) a tree topology, branch length, substitution rate, and a demographic Skygrid model (Supplementary Text 7). The demographic model is a Bayesian nonparametric one that is optimized for multiple loci and that allows for complex demographic trajectories by estimating population sizes in time bins across the tree based on the number of coalescent—branching—events per bin [68]. We also performed a second analysis run using a fixed prior for substitution rate of $3x10^{-9}$ substitutions $site^{-1}$ $year^{-1}$ based on our previous net distance estimate to confirm that the MCMC had the same parameter convergence, e.g. tree topology, as in the first "estimate-all-parameters" run.

Having a substitution rate per year we can estimate the time to the most common recent ancestor L solving $d = 2L \times \mu$ where d is the average pairwise genetic distance between our samples and μ is the calculated substitution rate from the distance method. This yielded 363 years, which subtracted to the average collection date of the samples, produced a point estimate of 1615. We compare this estimate with the inferred phylogeny root from the BEAST analysis.

Inference of genome-wide selection

We separately analyzed sequences at different annotations, since as they might be under different selection regimes (i.e. evolutionary constraints). We computed, using the HPG1 dataset, one-tailed Fisher's exact test using the base stats package in R [69] on contingency tables of the total number of base pairs against the number of SNPs, and those separated by positions being annotated as a coding against non-coding (intergenic, intronic, all other noncoding). The test returned whether coding regions have a lower number of SNPs than other reference annotation (intronic, interenic, all non-coding regions), as expected by the total number of positions in the genome annotated as such. We also constructed contingency tables to test whether SNPs annotated as coding compared to those annotated as non-coding were more likely to be found at low (<5%) or intermediate (5≥%) frequency.

Finally, we calculated the unfolded Site Frequency Spectrum (SFS) based on the order of appearance of genetic variants in the herbarium dataset. We then used the Kolmogorov–Smirnov two-samples test and 10,000 bootstrap resampling using the R package Matching v. 4.9–2 (ref. [70]) to calculate whether the frequency spectrum was lower for coding SNPs than for other SNPs. Additionally, we also repeated these analyses comparing nonsynonymous and synonymous mutations instead of coding and non-coding regions.

Association analysis

We collected flowering, seed and root morphology phenotypes for 63 accessions (Supplementary Text 8). For associations with climate parameters, we followed a similar rationale as previously described [71]. We extracted information from the bioclim database (http://www.worldclim.org/bioclim) at a 2.5 degrees resolution raster and intersected it with geographic locations of HPG1 samples (n = 100). We performed association analyses under several models and p-value corrections using the R package GeneABEL [72] (Supplementary Text 8.2). To calculate the variance of the trait explained by all genetic variants, we used a linear mixed model: $y = Zu + \varepsilon$; where y is the phenotype or climate variable, Z is the design matrix of genome identities, u is the random genome background effect informed by the kinship matrix and distributed as MVN $(0, \sigma_g A)$, and ε is the random error term. The ratio of σ_g / σ_T is commonly called narrow sense heritability, "chip" heritability, or proportion of variance explained by genotype

[73]. Only SNPs with MAF>5% (n = 391) were used to build a kinship or relationship matrix *A*. Note that the differences between any two genotypes were of the order of one or few dozens of SNPs. While this approach is appropriate to calculate a chip heritability, it would not be very useful to detect significant SNP, as the random factor accumulates all the available variation (S4 Table). We therefore run a regular GWA model without kinship matrix: $y = Xb + \varepsilon$; where X corresponds to the genotype states at a given SNP, and b is the fixed phenotypic effect of the SNP. To evaluate significance, we generated a p-value empirical null distribution based on running such model over 1,000 permuted datasets, which lead to conservative associations (S7 Fig, Data Appendix S1). The p-values from running the association in the real data that were below the 5% tail in the empirical distribution could be considered significant. However, we also established a conservative "double" Bonferroni correction, where the significant threshold was lowered to 0.01% (= 5% / [number of SNPs + number of phenotypes tested]). All significant SNPs are shown in S5 Table, and a subset in Table 1. Although many phenotypic traits did not have significant SNPs, we show all the QQ plots in the S2 Text.

Supporting information

S1 Text. Extended materials and methods, and supporting analyses.
(PDF)

S2 Text. For each trait employed in association analyses, we report the histogram distribution and the QQ plot of p-values to ensure that no trait departs exaggeratedly from the normal distribution, and that no inflation of p-values is observed (when lambda $\leq$ 1, there is no inflation of false positives).
(PDF)

S1 Table. Sample information. (Abbreviation H* indicates herbarium samples that cluster with the modern HPG1 clade rather than the historic HPG1 clade in Fig 3, highlighted as a star in the map from Fig 1. Abbreviations of herbarium collections or seed sources: UCONN = University of Connecticut Herbarium; CFM = Chicago Field Museum; NY = New York Botanical Garden; ABRC = Arabidopsis Biological Resources Center; OSU = Ohio State University).
(XLSX)

S2 Table. Sample information for Col-0 mutation accumulation lines. Information about each Mutation Accumulation (MA) line and their number of SNPs at different annotations. Also the total number of SNPs, average number of mutations and total bp covered in the genome per annotation are reported.
(XLSX)

S3 Table. Mutation rate estimates for different annotations in HPG1 and mutation accumulation lines. Mutation rates from MA lines are compared to HPG1 substitution rates from the dataset of 32_15 quality filter and complete information (see SOM) (Abbreviations: stat, descriptive statistic; bp, base pairs; lower and upper, lower and upper 95% CI; Nonsyn. and Syn., nonsynonymous and synonymous sites; UTR, untranslated region sites; HPG1 adj., substitution rate of HPG1 adjusted by a mean generation time of 1.3 years).
(XLSX)

S4 Table. Description of phenotypic and climatic variables for association mapping analyses. Mean and standard deviation (s.d.) across accessions for each phenotypic and climatic variables. Broad sense heritabilities (H2) were calculated from between line and within line (between replicate) variance in ANOVA. P-value corresponds to F test. Narrow sense

heritabilities (h2) were calculated employing linear mixed models and kinship matrix from mean accession values. P-values correspond to Likelihood Ratio test.
(XLSX)

S5 Table. SNP hits from association analyses and several descriptors. SNP hits significant at the 5% level after permutation correction are shown. Additionally, if raw p-values pass a double Bonferroni threshold of 0.01% are marked with a "tick". (Abbreviations: nonsyn. and syn., nonsynonymous and synonymous changes; regular one-letter abbreviation was used for amino acid changes).
(XLSX)

S1 Fig. Ancient-DNA characteristics of unrepaired herbarium libraries. (A) Fraction of *A. thaliana* DNA in sample. **(B)** Median length of merged reads. **(C)** Fraction of cytosine to thymine (C-to-T) substitutions at first base (5' end). **(D)** Relative enrichment of purines (adenine and guanine) at 5' end breaking points. Position -1 is compared with position -5 (negative numbers indicate genomic context before upstream reads' 5' end).
(PDF)

S2 Fig. Separation between HPG1 and other North American lineages. (A) Neighbor-joining tree built using Illumina-based SNP calls at the 149 genotyping markers originally used to identify HPG1 candidates. HPG1 accessions are shown in black, whereas other North American lineages are depicted in red (see explanation below for four HPG1-like accessions). **(B)** Neighbor-joining tree based on genome-wide SNPs. Accessions colored as in (A). Note that three accessions originally classified as HPG1 based on 149 SNPs (A) are placed outside this clade. A further accession (BRR7) within the HPG1 main branch was a recombinant removed from the analysis.
(PDF)

S3 Fig. Substitution spectrum and rates. (A) Site frequency spectrum for all transitions and transversions. **(B)** Distributions of "net" pairwise genetic distances between historic and modern samples used to calculate mutation rates per genomic annotation (from quality 32_15 and complete information per site). UTRs were excluded because of the small number of SNPs. **(C)** Mutation rates calculated for different genomic annotations and quality thresholds (32_32, 32_15, 24_24) and missing values (NA50: maximum 50% missing data per SNP; COMPL: missing data 0%). Mean and 95% confidence intervals are shown.
(PDF)

S4 Fig. Relationship between methylation and substitutions. (A, B) Fraction of methylation of cytosines in HPG1 pseudo-reference[7] at intergenic (A) or coding regions (B). **(C, D)** Fraction of methylation of cytosines in Col-0 reference genome(5) at intergenic (C) or coding regions (D). In each of the four comparisons, a grey histogram represents distribution of methylation of 1,000 random sets of invariant cytosines. Lines represent average methylation degree at those sites in HPG1 that changed from cytosine to thymine (red). We differentiate those substitutions that are shared—fixed—across all individuals (light red) or whose allele are present at an intermediate—segregating—frequency (dark red). Likewise, average methylation is shown for sites that changed from cytosine to adenine (blue) that that are fixed (light blue) or segregating (dark blue). The fact that the average methylation is higher in new substitutions than in invariant positions supports a connection between methylation and mutability of sites.
(PDF)

S5 Fig. Comparison of site frequency spectra across genomic annotations. Cumulative empirical distribution, at different genomic annotations, of the unfolded Site Frequency

Spectrum of SNPs oriented based on the order of appearance of alleles in the herbarium genomes. Note the steep slope at low frequency indicating large numbers of such variants.
(PDF)

S6 Fig. Spatial and temporal emergence of root-associated mutations. (A) Age distribution of derived SNPs with a significant trait association (the herbarium sample in which they were first recorded) (red), compared with genome-wide SNPs with at least 5% minor allele frequency (grey), or without frequency cutoff (black). **(B)** Spatial centroid of all samples carrying a derived allele. Since it is an average location, centroids can be in a body of water. Ten random draws of 50 SNPs for each category were used to produce the density lines in (A) and points in (B).
(PDF)

S7 Fig. Linkage disequilibrium of significant SNPs. (A-F) Linkage disequilibrium between SNPs with significant trait associations. Histogram of genetic distances **(A)** between samples when evaluating only coding regions at 5% minimum allele frequency. Linkage disequilibrium between SNP hits measured as r^2 **(B)** and D' **(C)**. Three significant SNPs were further studied to exemplify the power of association analyses with HPG1. For each, phenotypic differences between accessions that differ in the focal SNP and that are otherwise virtually genetically identical are compared both with all pairs of accessions and with pairs of accessions completely identical for coding regions. Below each violin plot is the histogram of linkage disequilibrium of the focal SNP with all other SNP hits. The three focal SNPs evaluated are located in AT5G19330 **(D)**, AT1G54440 **(E)** and AT2G16580 **(F)**.
(PDF)

Acknowledgments

For providing and retrieving herbarium specimens, we thank R. Capers, J. Devos, G. Shirsekar, M. S. Dossmann, J. Freudenstein, C. M. Herring, C. Niezgoda, C. A. McCormick, J. Peter and M. Thines. We thank X. Zhao and I. Henderson for recombination estimates, C. Lanz for sequencing support, C. Goeschl, B. Zierfuss and B. Wohlrab for help with root analyses, and P. Lang, D. Seymour, and D. Koenig for thorough proofreading and comments on the manuscript. We thank R. Colautti for useful comments on the theoretical framing of the manuscript, M. Nordborg for discussions and pointing us to the work of A.R. Templeton, K. Pruefer for input on data analysis, and D. Tautz, T. Mackay and the Weigel and Burbano labs for comments on the manuscript.

Author Contributions

Conceptualization: Moises Exposito-Alonso, Claude Becker, Joy Bergelson, Johannes Krause, Hernán A. Burbano, Detlef Weigel.

Data curation: Moises Exposito-Alonso, Claude Becker.

Formal analysis: Moises Exposito-Alonso, Claude Becker, Jörg Hagmann, Dominik G. Grimm, Rob W. Ness, Hernán A. Burbano.

Funding acquisition: Wolfgang Busch, Joy Bergelson, Johannes Krause, Hernán A. Burbano, Detlef Weigel.

Investigation: Moises Exposito-Alonso, Claude Becker, Verena J. Schuenemann, Ella Reiter, Claudia Setzer, Radka Slovak, Benjamin Brachi, Jörg Hagmann, Dominik G. Grimm, Jiahui Chen, Rob W. Ness, Hernán A. Burbano.

Methodology: Moises Exposito-Alonso, Claude Becker, Verena J. Schuenemann, Ella Reiter, Claudia Setzer, Radka Slovak, Benjamin Brachi, Jörg Hagmann, Dominik G. Grimm, Jiahui Chen, Rob W. Ness, Hernán A. Burbano.

Supervision: Wolfgang Busch, Joy Bergelson, Rob W. Ness, Johannes Krause, Hernán A. Burbano, Detlef Weigel.

Validation: Moises Exposito-Alonso, Claude Becker.

Visualization: Moises Exposito-Alonso, Claude Becker.

Writing – original draft: Moises Exposito-Alonso.

Writing – review & editing: Moises Exposito-Alonso, Claude Becker, Joy Bergelson, Rob W. Ness, Johannes Krause, Hernán A. Burbano, Detlef Weigel.

References

1. Green RE, Shapiro B. Human evolution: turning back the clock. Curr Biol. 2013 Apr 8; 23(7):R286–8. https://doi.org/10.1016/j.cub.2013.02.050 PMID: 23578879
2. Crawford PHC, Hoagland BW. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? J Biogeogr. 2009; 36(4):651–61.
3. Colautti RI, Lau JA. Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. Mol Ecol. 2015 May; 24(9):1999–2017. https://doi.org/10.1111/mec.13162 PMID: 25891044
4. van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, et al. Global exchange and accumulation of non-native plants. Nature. 2015 Aug 19; 525(7567):100–3. https://doi.org/10.1038/nature14910 PMID: 26287466
5. Razanajatovo M, Maurel N, Dawson W, Essl F, Kreft H, Pergl J, et al. Plants capable of selfing are more likely to become naturalized. Nat Commun. 2016 Oct 31; 7:13313. https://doi.org/10.1038/ncomms13313 PMID: 27796365
6. Sax DF, Stachowicz JJ, Brown JH, Bruno JF, Dawson MN, Gaines SD, et al. Ecological and evolutionary insights from species invasions. Trends Ecol Evol. 2007 Sep; 22(9):465–71. https://doi.org/10.1016/j.tree.2007.06.009 PMID: 17640765
7. Gauze GF. The struggle for existence. Baltimore: The Williams & Wilkins company; 1934. 192–192 p.
8. Hardouin EA, Tautz D. Increased mitochondrial mutation frequency after an island colonization: positive selection or accumulation of slightly deleterious mutations? Biol Lett. 2013 Apr 23; 9(2):20121123. https://doi.org/10.1098/rsbl.2012.1123 PMID: 23389667
9. Halligan DL, Keightley PD. Spontaneous mutation accumulation studies in evolutionary genetics. Annu Rev Ecol Evol Syst. 2009; 40(1):151–72.
10. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010 Apr 30; 328(5978):636–9. https://doi.org/10.1126/science.1186802 PMID: 20220176
11. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 1987 Dec; 84(24):9054–8. PMID: 3480529
12. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 2012 Sep 11; 10(9):e1001388. https://doi.org/10.1371/journal.pbio.1001388 PMID: 22984349
13. Pennings PS, Hermisson J. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol Biol Evol. 2006 May 1; 23(5):1076–84. https://doi.org/10.1093/molbev/msj117 PMID: 16520336
14. Karasov T, Messer PW, Petrov DA. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genet. 2010 Jun; 6(6):e1000924. https://doi.org/10.1371/journal.pgen.1000924 PMID: 20585551
15. Rouco M, López-Rodas V, Flores-Moya A, Costas E. Evolutionary changes in growth rate and toxin production in the cyanobacterium Microcystis aeruginosa under a scenario of eutrophication and temperature increase. Microb Ecol. 2011 Aug; 62(2):265–73. https://doi.org/10.1007/s00248-011-9804-0 PMID: 21271244

16. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012 Sep 11; 13(10):745–53. https://doi.org/10.1038/nrg3295 PMID: 22965354

17. Estoup A, Ravigné V, Hufbauer R, Vitalis R, Gautier M, Facon B. Is There a Genetic Paradox of Biological Invasion? Annu Rev Ecol Evol Syst. 2016; 47(1):51–72.

18. Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends Ecol Evol. 2008 Jan; 23 (1):38–44. https://doi.org/10.1016/j.tree.2007.09.008 PMID: 18006185

19. Dlugosch KM, Parker IM. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol. 2008 Jan; 17(1):431–49. https://doi.org/10.1111/j.1365-294X.2007.03538.x PMID: 17908213

20. Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. Mol Ecol. 2015 May; 24(9):2095–111. https://doi.org/10.1111/mec.13183 PMID: 25846825

21. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell. 2016 Jun 9; 166:481–91. https://doi.org/10.1016/j.cell.2016.05.063 PMID: 27293186

22. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, et al. African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. Proceedings of the National Academy of Sciences [Internet]. 2017 May 4; http://www.pnas.org/content/early/2017/05/03/1616736114.abstract

23. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The scale of population structure in Arabidopsis thaliana. PLoS Genet. 2010 Feb; 6(2):e1000843. https://doi.org/10.1371/journal.pgen.1000843 PMID: 20169178

24. Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. Royal Society Open Science. 2016 Jun 1; 3(6):160239. https://doi.org/10.1098/rsos.160239 PMID: 27429780

25. Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, et al. Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS Genet. 2015; 11(1):e1004920–e1004920. https://doi.org/10.1371/journal.pgen.1004920 PMID: 25569172

26. Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, et al. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat Genet. 2013 Nov; 45(11):1327–36. https://doi.org/10.1038/ng.2766 PMID: 24056716

27. Bomblies K, Yant L, Laitinen R a., Kim S-T, Hollister JD, Warthmann N, et al. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet. 2010 Mar; 6(3):e1000890–e1000890. https://doi.org/10.1371/journal.pgen.1000890 PMID: 20361058

28. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013 Dec; 14 (12):827–39. https://doi.org/10.1038/nrg3564 PMID: 24166031

29. Kimura M. On the evolutionary adjustment of spontaneous mutation rates. Genet Res. 1967 Apr; 9 (01):23–23.

30. Drummond A, Pybus OG, Rambaut A. Inference of viral evolutionary rates from molecular sequences. Adv Parasitol. 2003; 54:331–58. PMID: 14711090

31. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014 Oct 23; 514(7523):445–9. https://doi.org/10.1038/nature13810 PMID: 25341783

32. Ness RW, Morgan AD, Colegrave N, Keightley PD. Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. Genetics. 2012 Dec; 192(4):1447–54. https://doi.org/10.1534/genetics.112.145078 PMID: 23051642

33. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007 Jan; 7:214–214. https://doi.org/10.1186/1471-2148-7-214 PMID: 17996036

34. Millar CD, Dodd A, Anderson J, Gibb GC, Ritchie PA, Baroni C, et al. Mutation and evolutionary rates in adélie penguins from the antarctic. PLoS Genet. 2008 Oct 3; 4(10):e1000209. https://doi.org/10.1371/journal.pgen.1000209 PMID: 18833304

35. Christin P-A, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. Molecular dating, evolutionary rates, and the age of the grasses. Syst Biol. 2014 Mar; 63(2):153–65. https://doi.org/10.1093/sysbio/syt072 PMID: 24287097

36. Klein Goldewijk K, Ramankutty N. Land cover change over the last three centuries due to human activities: The availability of new global data sets. GeoJournal. 2004; 61(4):335–44.

37. Falahati-Anbaran M, Lundemo S, Stenøien HK. Seed dispersal in time can counteract the effect of gene flow between natural populations of Arabidopsis thaliana. New Phytol. 2014 May; 202(3):1043–54. https://doi.org/10.1111/nph.12702 PMID: 24471774

38. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010 Jan 1; 327 (5961):92–4. https://doi.org/10.1126/science.1180677 PMID: 20044577

39. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011 Oct; 43(10):956–63. https://doi.org/10.1038/ng.911 PMID: 21874002

40. Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, et al. Time-dependent rates of molecular evolution. Mol Ecol. 2011 Aug; 20(15):3087–101. https://doi.org/10.1111/j.1365-294X.2011.05178.x PMID: 21740474

41. Charlesworth B, Charlesworth D. Elements of Evolutionary Genetics. Roberts and Company Publishers; 2010.

42. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012 Dec 27; 8(12):e1002822. https://doi.org/10.1371/journal.pcbi.1002822 PMID: 23300413

43. van Kleunen M, Dawson W, Maurel N. Characteristics of successful alien plants. Mol Ecol. 2015 May; 24(9):1954–68. https://doi.org/10.1111/mec.13013 PMID: 25421056

44. Templeton AR, Sing CF, Kessling A, Humphries S. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics. 1988 Dec; 120(4):1145–54. PMID: 3147219

45. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974 Sep 6; 185 (4154):862–4. PMID: 4843792

46. Kim S, Choi H-I, Ryu H-J, Park JH, Kim MD, Kim SY. ARIA, an Arabidopsis arm repeat protein interacting with a transcriptional regulator of abscisic acid-responsive gene expression, is a novel abscisic acid signaling component. Plant Physiol. 2004 Nov; 136(3):3639–48. https://doi.org/10.1104/pp.104.049189 PMID: 15516505

47. Li G, Boudsocq M, Hem S, Vialaret J, Rossignol M, Maurel C, et al. The calcium-dependent protein kinase CPK7 acts on root hydraulic conductivity. Plant Cell Environ. 2015 Jul; 38(7):1312–20. https://doi.org/10.1111/pce.12478 PMID: 25366820

48. Choi H-I, Park H-J, Park JH, Kim S, Im M-Y, Seo H-H, et al. Arabidopsis calcium-dependent protein kinase AtCPK32 interacts with ABF4, a transcriptional regulator of abscisic acid-responsive gene expression, and modulates its activity. Plant Physiol. 2005 Dec; 139(4):1750–61. https://doi.org/10.1104/pp.105.069757 PMID: 16299177

49. Franks SJ, Weis AE. A change in climate causes rapid evolution of multiple life-history traits and their interactions in an annual plant. J Evol Biol. 2008 Sep; 21(5):1321–34. https://doi.org/10.1111/j.1420-9101.2008.01566.x PMID: 18557796

50. Arunkumar R, Ness RW, Wright SI, Barrett SCH. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics. 2015 Mar; 199(3):817–29. https://doi.org/10.1534/genetics.114.172809 PMID: 25552275

51. Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, et al. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. eLife. 2013 May 28; 2:e00731. https://doi.org/10.7554/eLife.00731 PMID: 23741619

52. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010 Jun; 2010(6):db.prot5448.

53. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010 Apr; 38(6):e87. https://doi.org/10.1093/nar/gkp1163 PMID: 20028723

54. Kircher M. Analysis of High-Throughput Ancient DNA Sequencing Data. In: Shapiro B, Hofreiter M, editors. Ancient DNA. Humana Press; 2011. p. 197–228. (Methods in Molecular Biology).

55. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res. 2008 Dec; 18(12):2024–33. https://doi.org/10.1101/gr.080200.108 PMID: 18818371

56. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, et al. Simultaneous alignment of short reads against multiple genomes. Genome Biol. 2009 Sep 17; 10(9):R98. https://doi.org/10.1186/gb-2009-10-9-r98 PMID: 19761611

57. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature. 2011 Dec 8; 480(7376):245–9. https://doi.org/10.1038/nature10555 PMID: 22057020

58. Shaw RG, Byers DL, Darmo E. Spontaneous mutational effects on reproductive traits of arabidopsis thaliana. Genetics. 2000 May; 155(1):369–78. PMID: 10790410

59. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008 Jun 1; 24(11):1403–5. https://doi.org/10.1093/bioinformatics/btn129 PMID: 18397895

60. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004 Jan 22; 20(2):289–90. PMID: 14734327

61. Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. 2010 Feb 1; 26(3):419–20. https://doi.org/10.1093/bioinformatics/btp696 PMID: 20080509

62. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006 Feb; 23(2):254–67. https://doi.org/10.1093/molbev/msj030 PMID: 16221896

63. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012 Aug; 29(8):1969–73. https://doi.org/10.1093/molbev/mss075 PMID: 22367748

64. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975 Apr; 7(2):256–76. PMID: 1145509

65. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989 Nov 1; 123(3):585–95. PMID: 2513255

66. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009 Jun 1; 25(11):1451–2. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325

67. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol. 2013 Apr; 23(7):553–9. https://doi.org/10.1016/j.cub.2013.02.044 PMID: 23523248

68. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard M a. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2012 Mar; 30(3):713–24. https://doi.org/10.1093/molbev/mss265 PMID: 23180580

69. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. https://www.R-project.org/

70. Sekhon JS. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R [Internet]. Vol. 42, Journal of Statistical Software. 2011. p. 1–52. http://www.jstatsoft.org/v42/i07/

71. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, et al. Adaptation to climate across the Arabidopsis thaliana genome. Science. 2011 Oct; 334(6052):83–6. https://doi.org/10.1126/science.1209244 PMID: 21980108

72. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007 May 15; 23(10):1294–6. https://doi.org/10.1093/bioinformatics/btm108 PMID: 17384015

73. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jun 20; 42(7):565–9. https://doi.org/10.1038/ng.608 PMID: 20562875