

RESEARCH ARTICLE

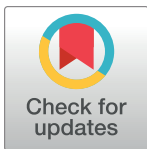
Quantifying the role of genome size and repeat content in adaptive variation and the architecture of flowering time in *Amaranthus tuberculatus*

Julia M. Kreiner^{1,2†*}, Solomiya Hnatovska^{2,3†}, John R. Stinchcombe², Stephen I. Wright²

1 Department of Botany, Biodiversity Research Centre, University of British Columbia, **2** Department of Ecology & Evolutionary Biology, University of Toronto, **3** Department of Molecular Genetics, University of Toronto

† These authors share first authorship on this work.

* julia.kreiner@ubc.ca



OPEN ACCESS

Citation: Kreiner JM, Hnatovska S, Stinchcombe JR, Wright SI (2023) Quantifying the role of genome size and repeat content in adaptive variation and the architecture of flowering time in *Amaranthus tuberculatus*. PLoS Genet 19(12): e1010865. <https://doi.org/10.1371/journal.pgen.1010865>

Editor: Nathan M. Springer, University of Minnesota, UNITED STATES

Received: July 12, 2023

Accepted: December 11, 2023

Published: December 27, 2023

Copyright: © 2023 Kreiner et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The code and metadata in our paper has been archived online at Github (https://github.com/jkreinz/GenomeSizeEvo_PlosGen) and Zenodo ([10.5281/zenodo.10366486](https://zenodo.org/record/10366486)), while sequence data are available for download at SRA under the BioProject accession number PRJNA770655.

Funding: J.M.K. was funded by a Killam Postdoctoral Fellowship from the Killam Trusts and a Biodiversity Research Centre Bioinformatics

Abstract

Genome size variation, largely driven by repeat content, is poorly understood within and among populations, limiting our understanding of its significance for adaptation. Here we characterize intraspecific variation in genome size and repeat content across 186 individuals of *Amaranthus tuberculatus*, a ubiquitous native weed that shows flowering time adaptation to climate across its range and in response to agriculture. Sequence-based genome size estimates vary by up to 20% across individuals, consistent with the considerable variability in the abundance of transposable elements, unknown repeats, and rDNAs across individuals. The additive effect of this variation has important phenotypic consequences—individuals with more repeats, and thus larger genomes, show slower flowering times and growth rates. However, compared to newly-characterized gene copy number and polygenic nucleotide changes underlying variation in flowering time, we show that genome size is a marginal contributor. Differences in flowering time are reflected by genome size variation across sexes and marginally, habitats, while polygenic variation and a gene copy number variant within the ATP synthesis pathway show consistently stronger environmental clines than genome size. Repeat content nonetheless shows non-neutral distributions across the genome, and across latitudinal and environmental gradients, demonstrating the numerous governing processes that in turn influence quantitative genetic variation for phenotypes key to plant adaptation.

Author summary

The remarkable and seemingly inconsequential variation in genome size across species has long been an enigma in evolutionary biology. Calling this viewpoint into question, correlations between genome size variation and traits linked to fitness are increasingly

Postdoctoral Fellowship from the University of British Columbia. S.I.W. was supported by the NSERC RGPIN-2020-05850 and a Canada research chair. J.R.S. was supported by the NSERC RGPIN-2022-04366 and the Swedish Collegium for Advanced Study. The funders had no role in 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.

uncovered. While this suggests that DNA content itself may be a source of adaptive genetic variation, repeat elements that propagate at the cost of the host are known to largely mediate this variation and may thus limit adaptive potential. Here we look to disentangle these multi-level dynamics, describing repeat dynamics across the genome and among individuals across diverse collections of a widespread agricultural weed, linking repeat content to genome size variation, and characterizing the relative importance of its phenotypic consequences. In *Amaranthus tuberculatus*, we find non-neutral repeat distributions across individuals across the range, and while this repeat variation underlies both variation in genome size and flowering time, we show that it makes a relatively minor contribution to variation in a fitness-related trait across the landscape compared to monogenic and polygenic features. Together, this work broadens our perspective on the complex selective dynamics that govern intraspecific variation in genome size and traits key to fitness in plants.

Introduction

The genome was colloquially viewed as a blueprint containing information to encode the phenotype; however, genome size and composition are also quantitative traits in themselves [1,2]. Early views on genome size evolution led to the prediction that more complex organisms, with a wider array of cell and tissue types, would have more complex genetic encoding—more genes and larger genomes. Yet, investigations of variation in genome size revealed a puzzling lack of correlation with perceptions of organismal complexity [3–5]; genome size was not in fact a clear predictor of ‘information content’. Variation in the amount of repetitive DNA sequences is now widely recognized as an important resolution to this paradox. In flowering plants, the primary driver of variation in repetitive content is transposable elements (TEs), which proliferate by creating copies of themselves in new sites across the genome, and whose proportional content is known to range between more than 85% of the genome in maize and as little or less than 20% in *Arabidopsis thaliana* [6–8]. Nevertheless, the role of repetitive sequence in shaping genomic and phenotypic variation across populations remains unknown for most species. Here, we examine the relative roles of repetitive sequence, gene copy number, and single nucleotide polymorphism (SNP) variation in the genetic basis, and geographic divergence, of a key life history trait in *Amaranthus tuberculatus*, an important agricultural weed.

A classic model explaining TE abundance variation involves a balance between the rate of transposition increasing TE insertions and the force of negative selection removing them due to their deleterious effects [9]. These effects stem from the ability of TEs to insert into and near genes, to cause ectopic recombination, and to affect transcription of nearby genes through the spread of epigenetic silencing [10–15]. On the one hand, if effective population size and thus the magnitude of drift and efficacy of selection differs among populations, population-level TE abundances may differ [16,17] and covary with population structure. Similarly, in genomic regions of low recombination (e.g., sex-determining regions), the efficacy of selection is reduced, leading to another source of variation in TE content across the genome and between the sexes [18], which has recently been implicated in sex differences in survival [19,20]. On the other hand, TE abundance may differ among individuals and populations due to variation in transposition, as a consequence of TEs differentially evolving to evade host silencing mechanisms [21] or due to local physiological responses to environmental stress. Biotic and abiotic stressors, such as temperature, irradiance, nutrient starvation and fungal pathogens, are known to induce higher rates of transposition of certain TEs [22–25]. Finally, TEs may also be a source of local adaptation at multiple scales, by influencing the expression, copy number,

and mutational landscape of particular focal genes (reviewed in [26,27]) and/or by generating variation in genome size that in turn influences quantitative traits [2].

While the evolutionary drivers of genome size variation are still debated [28], genome size is increasingly considered a potentially important contributor to plant adaptation [29]. Larger genomes may not only mean a larger adaptive mutational target size (especially for non-genic regulatory variation; discussed in [29]) but may also have direct developmental consequences: larger genomes are often associated with larger cells, slower cell division, slower organismal growth rates, and ultimately a longer time to maturity and reproduction [30–33]. While the nature and direction of causality between genome size and these genomic properties has been difficult to uncouple [34], variation in genome size is hypothesized to be under selection as a source of variation for local adaptation via life history rate timing and is supported by recent observations of intraspecific variation in corn (*Zea mays*) [2,35]. At high altitudes, where faster growth and time to flowering assure reproduction before the early end of the season, maize individuals not only harbour fewer TEs, heterochromatic knobs, and smaller genomes compared to lower altitude plants but also show a faster rate of cell production and earlier flowering time [2]. The association between genome size and elevation described in [2] remained significant even after controlling for genome-wide relatedness, consistent with a model of selection acting on genome size through its effects on flowering time (or highly correlated traits). Taken together, variation in TE abundance across populations may be mediated not only by the balance between transposition rate and selection on individual elements, but also by spatially fluctuating selection on repeat abundance through its effect on life history traits. While these results provide compelling support for the role of repetitive DNA and genome size in life history trait adaptation, the relative importance of this variation versus other genetic sources remains poorly understood.

To further evaluate how and why repeats vary in abundance within species and the relative contributions of these different sources of variation to key life history traits, we turned to the prevalent agricultural weed, *A. tuberculatus* (common waterhemp). The annual plant *A. tuberculatus* is dioecious and wind-pollinated, with a range centered around the Mississippi river in the United States [36]. The species consists of two varieties, which were historically isolated to the northeast (var. *tuberculatus*) and southwest (var. *rudis*) of midwestern USA and southeastern Canada. Key to the dynamics of adaptation in the species is variability in life history traits (e.g. flowering time, growth rates) across geographic clines, habitats (natural and agricultural), and by varietal ancestry [37]. The species has evolved rapidly to industrial agriculture over the last six decades, driven by extreme selective pressures on genome-wide standing genetic variation drawn preferentially from southwestern var. *rudis* ancestry in contrast with ancestry from the northeastern var. *tuberculatus* [38]. However, the contribution of variation in repeat content and genome size to these adaptive dynamics has yet to be explored.

Drawing on previously published high coverage (~28x) Illumina short-read genomic data from 187 individuals spanning natural and agricultural environments from Kansas to Ohio, accompanied by phenotypic measurements from a quantitative genetic common garden experiment, we address the following questions: (1) What is the distribution of repeat content across the genome and among individuals of *A. tuberculatus*? (2) Which repeat types are the predominant drivers of genome size variation and what are the consequences for key life history phenotypes? and (3) What is the relative importance of genome size compared to other sources of genetic variation for encoding flowering time and responding to environment-mediated selection? To do so, we combined sequence-based genome size estimates, base pair abundances for 16 repeat and TE classes across individuals, and a novel characterization of the genetic architecture of flowering time. Our findings demonstrate the complexity of selective forces that govern variation in repeat abundance, genome size, and life history, and that interact to determine local adaptation and sex differences in this ubiquitous species.

Results

Transposable element variation across the genome

We first set out to characterize TE composition across the *A. tuberculatus* female reference genome [39]. TEs make up 62.6% or ~417.8 Mb of the 668.5 Mb *A. tuberculatus* genome (Fig 1A). In total we annotated 888,765 TEs, including two taxonomic repeat orders of cut-and-paste DNA transposons (366,445 terminal inverted repeat elements [TIRs] and 182,448

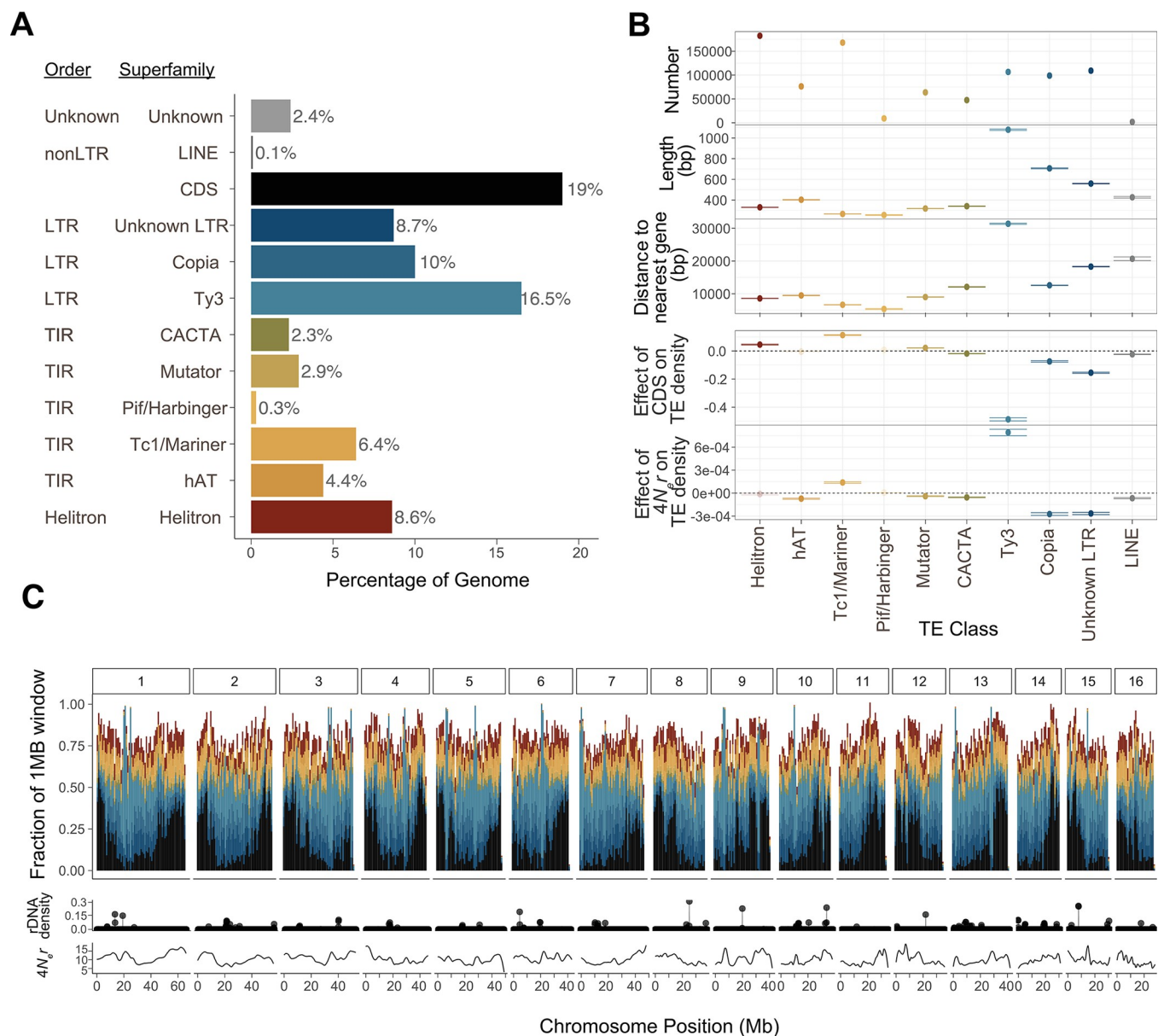


Fig 1. Variation in TE composition across the female *Amaranthus tuberculatus* reference genome by order and superfamily. A) The fraction of the female *A. tuberculatus* reference genome composed of different TE orders and superfamilies. B) Statistical summaries by TE superfamily, illustrating the differences in the number, size, and distance to genes across the reference genome. The bottom two rows represent the effect of coding sequence (CDS) density and $4N_e r$ (the effective recombination rate) on TE density as inferred from a multiple regression for each TE superfamily, for which all opaque lines are significant at $p < 0.05$. Horizontal bars represent standard error of the estimate. C) The distribution of TE superfamilies and coding sequence content across the 16 chromosomes (top; colour codes from A and B), relative to the 100 kb window density of rDNAs (middle) and means of the 100 kb population scaled recombination rate ($4N_e r$; bottom).

<https://doi.org/10.1371/journal.pgen.1010865.g001>

Helitrons) and two orders of copy-and-paste retrotransposons (317,533 long terminal repeat elements [LTRs] and 1841 non-LTRs/LINEs). Mean lengths ranged from 316 and 329 bps for annotated Helitrons and TIR elements to 781 bps for LTR elements, suggesting that many of these are non-autonomous and fragmented TEs, consistent with our knowledge of TE composition in plant genomes [40]. The greatest proportion of the genome was represented by LTR elements (including unknown LTRs, Copias, and Ty3 elements) in part due to their larger size, as is typical in plants (reviewed in [41]). Helitrons, in comparison, were nearly twice in number as individual LTR element families but are at least half their size on average.

While TE superfamilies vary in size and number within *A. tuberculatus*, the distribution of TEs also varies across the genome in relation to genic content (coding sequence; CDS) and the population recombination rate ($4N_e r$) (the two of which are positively correlated in *A. tuberculatus*), likely reflecting differences in activity, transposition biases, and the strength of negative selection [18] (Figs 1C, S1 & S2). LTR elements were the largest and typically found furthest from genes, while TIRs (e.g., PiF/Harbinger, Tc1/Mariners) tended to be the smallest and closest to genes (Fig 1B). When we tested how well the population recombination rate and CDS density predicted the composition of TE superfamilies in 100kb windows in a multiple regression framework, we found every possible combination of the direction and strength of these predictors across TE superfamilies (Fig 1B). Variability in LTRs across the genome was the most consistently explained by the strong negative effect of CDS (Copias: $F = 240$, $p < 10^{-15}$; Ty3s: $F = 1823$, $p < 10^{-15}$, Unknown LTRs: $F = 1558$, $p < 10^{-15}$), while typically exhibiting negative correlations with population recombination rate (Copias: $F = 238$, $p < 10^{-15}$; Unknown LTRs: $F = 340$, $p < 10^{-15}$; but negative for Ty3: $F = 356$, $p < 10^{-15}$; (Fig 1B). The correlations of population recombination rate and CDS were much more variable across TIR superfamilies (S1 Fig), while variability in Helitrons is not explained by recombination rate but is positively correlated with CDS ($F = 211.8262$, $p < 10^{-15}$; Fig 1B). Clearly, the *A. tuberculatus* genome reflects a complex genomic landscape of TE diversity as is seen in other systems (e.g. [8]).

We also annotated simple repeats and rDNA genes. rDNAs, repetitive genes functioning in ribosome production [42], made up 4.1% of the genome and were distributed across all 16 main chromosome scaffolds (Fig 1C). A total of 2968 5S rDNA genes, 55 28S rDNA genes and 52 18S rDNA genes were annotated. Simple, low complexity, and unknown repeats comprise 7.0%, 1.5%, and 11.1% of the genome, respectively.

Repeat variation across individuals

Looking beyond a single reference genome, we next investigated the extent of intraspecific variation in TE and repeat class abundance in *A. tuberculatus* and thus the potential for it to contribute meaningfully to adaptive evolution. We estimated the abundance of each repeat class and TE superfamily (see methods; as delimited in Fig 1A) within the genomes of 187 individuals. The median bp composition of repeat classes across individuals showed approximately the same rank order as annotated in our reference genome: Ty3s (86.8 MB), Copias (59.0 MB), Helitrons (50.1 MB), and unknown LTRs (49.3 MB) show the greatest mean bp contribution across individuals, while 5s rDNA (1.24 Mb), low complexity repeats (0.86 Mb) and non-LTR LINEs show the least (0.58 Mb) (S3 & S4 Figs). We next quantified both the variance and the coefficient of variation (CV, standard deviation scaled by the mean) in the bp composition of repeat types across individuals (Fig 2A). The three rDNA subunits had the highest CV, more than double that of nearly all other repeat classes. By contrast, TEs tended to show the lowest coefficient of variation among individuals; there was little variability in abundance of TE classes (as measured by CV), although the absolute variance in abundance was high given the large number of these elements (particularly Ty3, Copia, and Unknown LTRs) (Fig 2A).

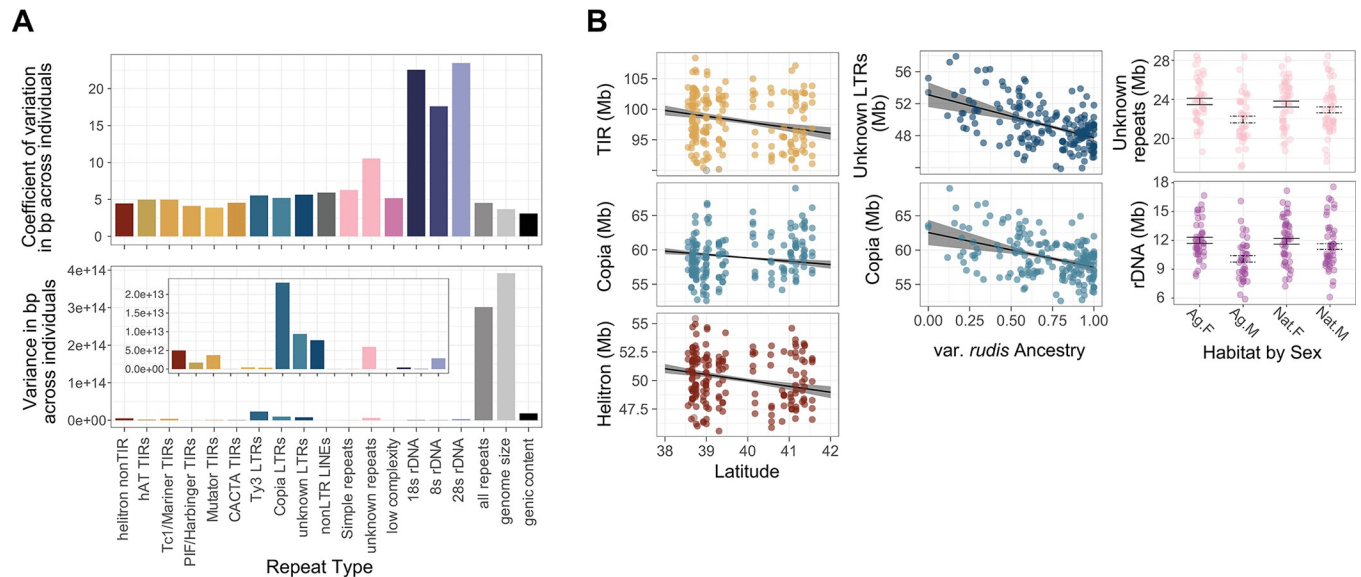


Fig 2. Variation in repeat abundance across individuals. **A)** The coefficient of variation (top) and variance (bottom) in bp amount of an individual's genome composed of a given repeat class. **B)** The relationship between repeat class abundance with latitude (left side), var. *rudis* ancestry (middle; based on the proportion of an individual's genome composed of var. *rudis*, as opposed to var. *tuberculatus*), and the interaction between habitat (Ag: agricultural site; Nat: natural site) and sex (M: male, F: female; right side). Points represent raw data, while regression lines and error bars represent the least squares mean from a mixed effect model that accounts for relatedness. Relationships shown for a subset of significant ($p < 0.05$) predictors.

<https://doi.org/10.1371/journal.pgen.1010865.g002>

Landscape and organismal predictors of repeat content

We inferred landscape and organismal level predictors of each TE superfamily and repeat class abundance using linear mixed models that included the relatedness matrix (i.e., population structure) as a random effect (as in [2,43], with the matrix visualized in **S5 Figs**; see **methods Eq 1** for mixed model). In doing so, we found evidence of latitudinal clines in repeat content (where population collections span from 38 to 41 degrees) that exceeded expectations from neutral population structure (structure visualized in **S3 and S4 Figs**). The abundance of several elements declined substantially with latitude, including TIR (slope = -942124, $\chi^2 = 5.41$, $p = 0.020$) [Tc1: $F = 7.44$, $p = 0.006$; PiF: $F = 6.92$, $p = 0.009$; hAT: $F = 4.57$, $p = 0.032$; Mutator: $F = 5.22$, $p = 0.022$], Helitron (slope = -516999, $F = 5.90$, $p = 0.015$), and Copia elements (slope = -488909, $F = 3.89$, $p = 0.049$) (**Fig 2B**). Because *A. tuberculatus* varietal ancestry varies along this same axis, ancestry proportion (based on structure inference, as in [37]) was explicitly included as a fixed effect predictor in addition to the relatedness matrix, suggesting that our observed latitudinal clines in repeat abundance are not simply the result of either of these time-scales of evolutionary history. Ancestry did, however, significantly predict two repeat types, Copia elements (slope = -5061330, $F = 3.94$, $p = 0.047$) and unknown LTRs (slope = -5369193, $F = 6.41$, $p = 0.011$) (**Fig 2B**), suggesting that a history of isolation and/or geographic variation in climate (or highly correlated [a]biotic forces) play a role in mediating these repeat abundances across the range.

We next investigated whether *A. tuberculatus*' recent colonization of agricultural habitats was associated with the abundance of particular repeat classes. No repeat class showed evidence for an effect of collection habitat (natural or agricultural). However, we did detect a significant sex by habitat interaction for two repeat classes (**Fig 2B**), where differences in repeat abundance between sexes depended on whether the comparison is made within natural or agricultural habitats. For repeat classes that were significantly less abundant in males, rDNAs (sex effect: $\chi^2 = 14.92$, $p = 0.0001$) and unknown repeats (sex effect: $\chi^2 = 12.74$, $p = 0.0004$), the

difference between sexes was greater in agricultural compared to natural habitats (sex by habitat effect; rDNAs: $\chi^2 = 4.29$, $p = 0.038$; unknown repeats: $\chi^2 = 3.28$, $p = 0.07$) (Fig 2B). While multiple test corrections for models with distinct dependent variables is debated, FDR correction within each repeat class model showed a maximum FDR of 19%, suggesting that 3/17 significant discoveries of TE content predictors may potentially be false positives (S3 Table).

Variation in genome size and its phenotypic consequences

To understand the cumulative effects of this variation in repeat content, we next quantified genome size variation across individuals. We took both k-mer [44] and read-depth based approaches to do so, and with the two measures being largely concordant ($r = 0.92$; S6 Fig), here we present results from read-depth based coverage (see methods). There is substantial genome size variation across the 186 individuals in this study, with the largest estimate being ~20% larger than the smallest estimate (min Mb = 543.2, max Mb = 650, mean Mb = 594; Figs 3 and S7), about double the variability seen in *A. thaliana* (up to 10% [1]).

We leveraged phenotypes measured in a common garden experiment in these same samples [37] to test whether genome size correlates with key life history traits. Focusing on traits related to growth and the timing of key life history transitions, we modeled the effects of genome size on growth rate (the increase in plant height between the 4–6 leaf stage and flowering) and time to flowering. Because sequenced individuals spanned multiple treatments in the common garden experiment [37], here we used family-mean phenotypes as measured in the control treatment. With individuals collected from a broad sample across habitats (natural and agricultural), latitudes, longitudes, varietal ancestries, and with separate sexes in the species, we included all these factors along with genome size as fixed variables in this model of flowering time, and the relatedness matrix as a random effect (methods Eq 2). Furthermore, with variability in the degree of sexual dimorphism across environments having been identified in this and other wind pollinated species [45], we tested for sex by habitat interactions.

We found that genome size is a significant predictor of flowering time ($\chi^2 = 12.0$, $p = 0.0005$), with every additional 10 MB predicted to delay flowering by 0.8 days (Fig 3). In our collections, that corresponded to an 8.4-day difference in flowering time between samples with the smallest and largest sampled genome. Genome size explained 6.5% of the variation in flowering time in this mixed effect model. The only predictor that explained more variation than genome size in this model of flowering time was sex (partial $r^2 = 16.6\%$, $\chi^2 = 34.6$, $p = 4.1 \times 10^{-9}$), with males flowering 8.7 days [SE = 1.48 days] earlier than females. Flowering time also showed a strong sex-by-habitat interaction ($\chi^2 = 9.78$, $p = 0.0017$) reflecting greater sexual dimorphism in flowering time in agricultural compared to natural habitats (Fig 3). We also found a main effect of habitat ($\chi^2 = 9.9$, $p = 0.0017$), longitude ($\chi^2 = 6.15$, $p = 0.013$), and latitude ($\chi^2 = 5.7$, $p = 0.0166$) (Fig 3A; as described in [37]). In a model of growth rate, genome size is negatively related to growth rate ($\chi^2 = 5.56$, $p = 0.0183$; Fig 3), explaining 3.1% of the variation, similar to the effect of sex (partial $r^2 = 3.3\%$, $\chi^2 = 5.89$, $p = 0.015$). We also find a habitat-by-sex interaction effect ($\chi^2 = 4.95$, $p = 0.026$) and a main effect of sex ($\chi^2 = 5.89$, $p = 0.015$) (Fig 3). Taken together, these results support the hypothesis of genome size playing a role in determining key life history traits in *A. tuberculatus*.

While flowering time, growth rates, and individual TE classes show strong and significant latitudinal and environmental variation, genome size itself does not show strong differentiation among habitats. When correcting population structure using the relatedness matrix (methods Eq 2), only sex ($\chi^2 = 5.0797$, $p = 0.024$) and marginally, sex by habitat ($\chi^2 = 3.3736$, $p = 0.06625$) significantly explain variation in genome size, where males tend to have 11.3 Mb smaller genomes than females, in agricultural but not natural habitats. Therefore, variation in

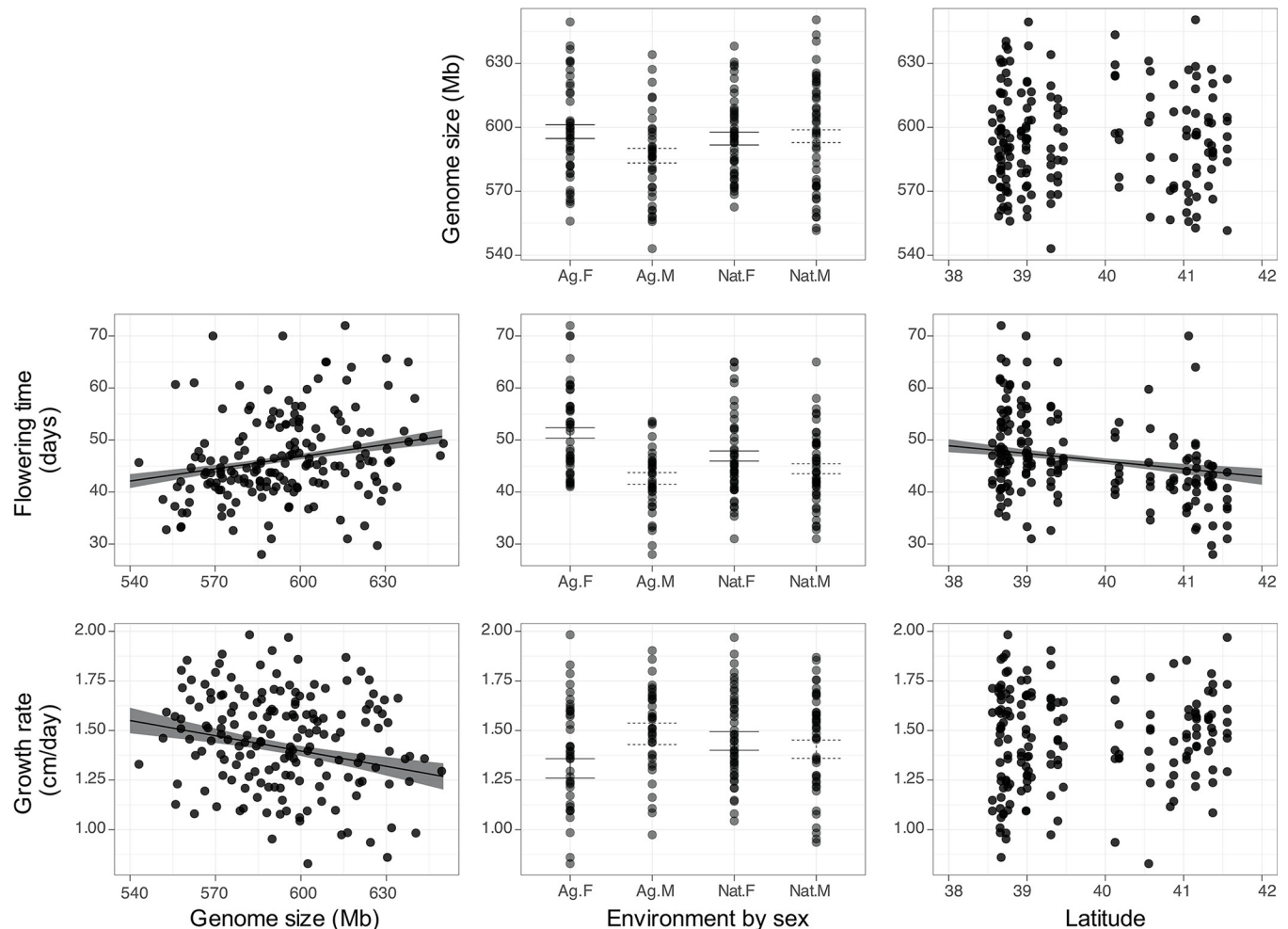


Fig 3. Genome size predicts flowering time (middle) and growth rate (bottom) in *A. tuberculatus*, genetic and morphological traits that also differ by latitude and habitat in a sex specific manner. Points show raw data, while regression lines and error bars depict least squares mean estimates from linear mixed modeling of genome size, flowering time, and growth rate. Ag/Nat.F or .M refer to male or female values in each habitat. Trend lines are shown for all significant relationships.

<https://doi.org/10.1371/journal.pgen.1010865.g003>

overall genome size has apparently not meaningfully responded to environmental selection through flowering time.

The relative importance of quantitative genome size variation, oligogenic, and polygenic features to flowering time evolution

We next characterized the contribution of large-effect and polygenic variants to these fitness related traits, ranging from copy number variation to SNPs. While genome-wide associations for SNPs and gene copy number variation with growth rate yielded no significant SNPs after Bonferroni or any level of FDR correction, flowering time showed multiple types of genetic associations.

A GWA of flowering time using gene-level copy number variation as predictors identifies 1,680/30,771 genes with a significant effect at the FDR $q < 0.10$ level and 34/30,771 genes at the Bonferroni $p < 0.10$ level. These 34 genes include FLOWERING LOCUS D on Scaffold 11, with the broader set of genes being significantly enriched for the ATP synthesis PANTHER pathway

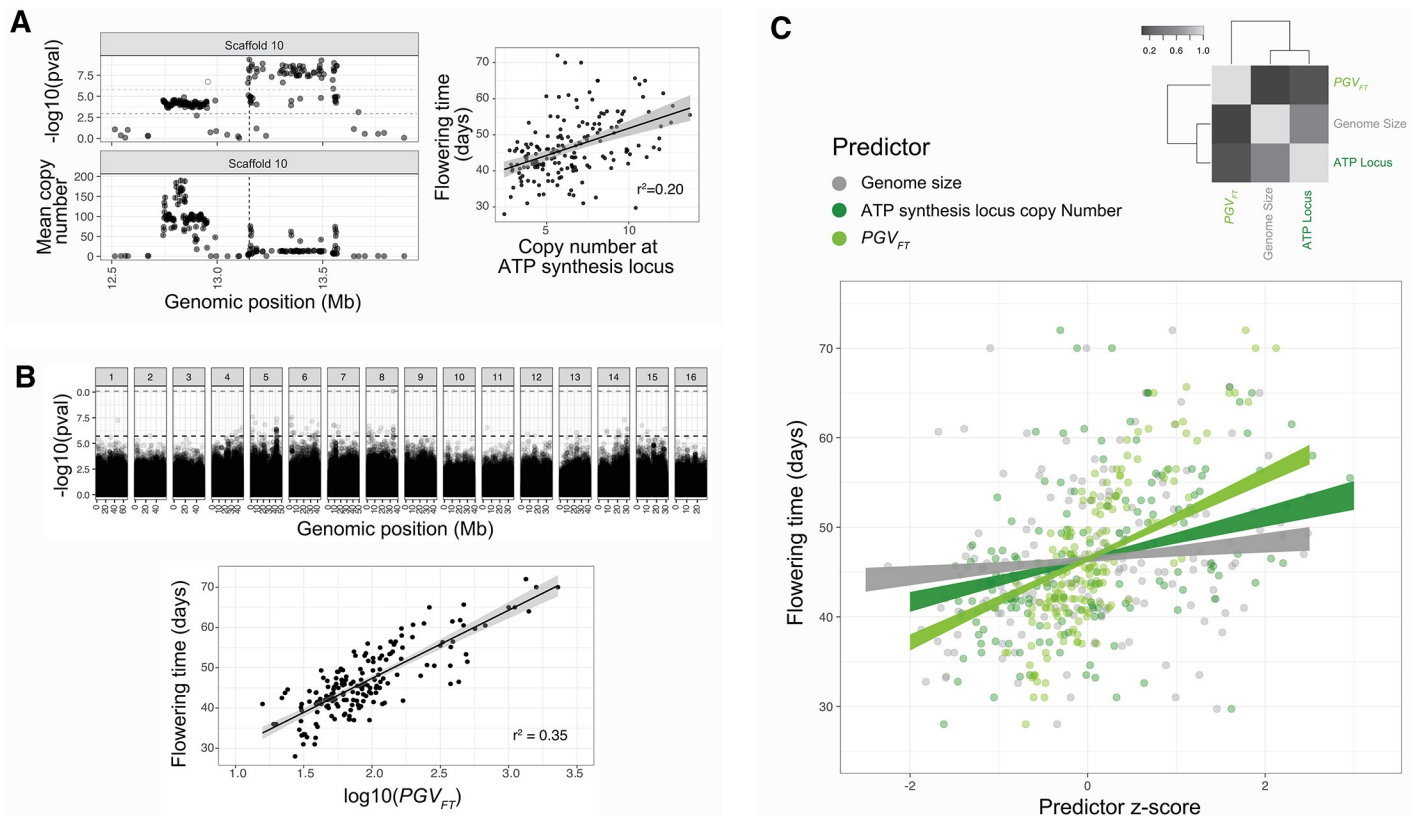


Fig 4. The genetic architecture of flowering time in *A. tuberculatus* and the relative importance of associated genetic features. A) The association of a copy number variant in the ATP synthesis pathway with flowering time (vertical line denoting locus with the most significant association genome wide). B) The polygenic value of individual flowering time (PGV_{FT}) based on 97 SNPs that pass a 10% FDR correction from a genome-wide association correcting for population structure (lower black horizontal dashed line, Bonferroni threshold also shown above). Black line in the bottom plot represents the linear regression fit between flowering time and PGV_{FT} . C) A mixed effect model for flowering time while controlling for relatedness demonstrates the relative importance of associated genomic features, from the polygenic value in B) and copy number variation at the ATP synthesis locus in A) to genome size variation. Correlation structure (Pearson's r) of fixed effect predictors also illustrated in the top right of C).

<https://doi.org/10.1371/journal.pgen.1010865.g004>

(Bonferroni p -value = 4.6×10^{-3}), GO molecular functions in NADH dehydrogenase (Bonferroni p -value = 6.1×10^{-6}) and proton-transporting ATP synthase activity (Bonferroni p -value = 6.9×10^{-4}), as well as numerous other related GO biological functions (S1 Table). This signal of enrichment appears to be predominantly driven by one large-effect locus on Scaffold 10 (Fig 4A), a region in which a cluster of 7 NADH-ubiquinone oxidoreductases (two *ND1*, *ND4L*, *ND5*, *ND6*, two *ND2*), one NADH dehydrogenase (*NAD7*), 4 ATP synthases (*ATP6*, *ATP4*, *atpA*, *atpB*), and one Cytochrome b6-f complex subunit 5 (*petG*) map. A regression of flowering time on the gene with the strongest statistical association in this region (and genome-wide) reveals that 20% of the variation in flowering time can be explained by copy number variation at this locus, which varies from ~2 to 14 copies (Fig 4A, right side).

In contrast, SNP-level associations with flowering time appear to reflect a more dispersed, polygenic architecture. A GWA of flowering time using SNPs as predictors while controlling for the relatedness matrix identifies 97 loci in 73 genes across the genome passing a 10% FDR correction. We therefore calculated the polygenic value [46,47] for flowering time [$PGV_{FT} = 2 \sum_{l=1}^L \alpha p$], where L equals the 97 loci with an FDR $q < 0.1$, α the effect size of a locus on flowering time, and p an individual's allele frequency at that locus (genotype). A bivariate correlation of the log transformed individual PGV_{FT} with flowering time shows that

the PGV_{FT} explains 35% of the variation in flowering time in the environmental conditions in which it was measured (Fig 4B). Since latitudinal and longitudinal clines in flowering time covary with ancestry [37], potentially driving high rates of false-negatives when controlling for population structure, we also compared a SNP-level GWAS without such a control. As expected, this GWAS has many more SNPs passing the 10% FDR correction ($n = 3549$; S8 Fig), and while several such hits may be false positives, they collectively show significant enrichment for numerous biological processes, such as post-embryonic plant morphogenesis with several genes having known functions in flower development and flowering time (see methods, S8 Fig).

Finally, we tested the relative importance and independence of these scaled polygenic and oligogenic predictors compared to genome size in a mixed effects model of flowering time that includes the relatedness matrix as a random effect (methods Eq 3). This model explains 88% of the variation in family-mean flowering time, 26% of which is attributed to the fixed-effect terms and 62% of which can be attributed to the genome-wide relatedness. Genome size remains a marginally significant explanatory variable ($\chi^2 = 3.24$, $p = 0.071$) explaining 2% of the variation in flowering time but is considerably less important compared to copy number variation at the ATP synthesis locus (partial $r^2 = 11\%$, $\chi^2 = 21.16$, $p = 9.6 \times 10^{-6}$), which in turn is less important than PGV_{FT} (partial $r^2 = 40\%$, $\chi^2 = 127.15$, $p < 10^{-15}$) (Fig 4C).

These polygenic and oligogenic architectures show stronger patterns of spatial differentiation than genome size, likely reflecting their relative contribution to the response to selection through flowering time. Copy number variation at the ATP synthesis locus demonstrates significant variability among sexes ($\chi^2 = 22.59$, $p = 2.01 \times 10^{-6}$), across natural and agricultural habitats ($\chi^2 = 9.27$, $p = 0.0023$), and among sexes depending on habitat type ($\chi^2 = 8.83$, $p = 0.0029$), exceeding neutral expectations. Similarly, $\log_{10}(PGV_{FT})$ shows a strong sex effect ($\chi^2 = 14.26$, $p = 0.0002$), latitude effect ($\chi^2 = 9.81$, $p = 0.0017$), marginally, a longitude effect ($\chi^2 = 3.246$, $p = 0.072$) and a sex by environment effect ($\chi^2 = 2.9$, $p = 0.084$).

Discussion

We report marked intraspecific variability in repeat content and genome size that is associated with flowering time variation in *A. tuberculatus*. Individuals with more repeats and larger genomes tend to show slower growth rates and time to flowering, especially in models that do not incorporate the effects of polygenic SNPs and gene copy number. We leveraged past whole genome sequencing and common garden phenotype data for nearly 200 individuals to show that this quantitative variation in genome size complements polygenic and copy number variation for flowering time and is independent from the effects of locus-specific TE copy number on flowering time (S8 Fig, methods). When newly identified polygenic variation and a photosynthesis-related copy number variant is modeled jointly along genome size, genome size is only a marginally significant, modest predictor of common-garden-measured flowering time across our collections.

Phenotypic and latitudinal associations with repeat content and/or genome size have been found across several systems, from maize [2,35,48] to *Drosophila* [49]. Mechanistically, cells with larger genomes are thought to take longer to undergo cell division and thus development, and this is supported by associations with cell size [33,50], cell production rate [2], stomatal density [33], flowering time [2], development time [51], growth form [52], and scaling laws [53]. We therefore predicted that earlier flowering waterhemp plants at higher latitudes, and in natural, not agricultural, habitats, might have smaller genomes and less repeat content. While we find evidence for an effect of genome size on life history traits (both flowering time and growth rate), genome size does not vary across latitude and habitat beyond neutral

expectations from genome-wide relatedness, except for differences across habitat types that are mediated by differences across sexes. This suggests that the relatively minor contribution of genome size to the genetic basis of flowering time has led to a lack of clear adaptive differentiation in genome size in this study. In part, the relative lack of signal of flowering time selection shaping genome size variation may result from the lack of large effect genome-size alleles in *A. tuberculatus*, such as heterochromatic knobs in maize [2]. Nonetheless, particular repeat classes do show clinal, habitat, and sex-biased variation that exceed neutral expectations, highlighting the numerous forms of selection that are likely affecting the abundances of individual element families in *A. tuberculatus*.

Previous work in *A. tuberculatus* has revealed the sundry of genetic mechanisms underlying agricultural adaptation, from standing genetic variation and de novo mutational origins to gene flow [37,38,54]. Here we showed the genetic architecture of flowering time appears to be nearly as multifaceted. In addition to a modest role for genome size, we describe two genetic features underlying genetic variation for flowering time. We find copy number variation at a cluster of genes in the ATP synthesis pathway that is associated with variation in flowering time, adding to the increasingly recognized importance of structural variation in flowering time evolution [55–57]. That copy number variation for genes in the photosynthetic pathway predicts flowering time supports the notion that modifying photosynthesis can impact developmental rate and plant yield (e.g., [58–60]), providing some of the first evidence for such a link in natural populations. We also quantified polygenic variation for flowering time encoded by single nucleotide polymorphisms at 97 associated loci across the genome, while controlling for population structure. These SNPs mapped to genes with molecular and biological functions varying from stigma and gynoecium development (*HEC1*), meristem and flower development (*PCN*, *DOT2*; [61]), gibberellic acid mediated anther and seed development (*SPL*; [62]), to DNA methylation and post transcriptional gene silencing (*AGO4*, *ROS11*; the former protein family having been implicated in mediating flowering time by modifying the expression of *FT* [63]) (S2 Table). A parallel analysis without controlling for population structure explains an additional 9% of the variation in flowering time, further suggesting differences in the architecture of flowering time among populations. Gene copy number at the ATP locus, polygenic variation for flowering time, and genome size all showed differentiation among habitats and sexes (although only an interaction effect between sex and habitat was evident for genome size), suggesting that strong selection for high performing, earlier flowering males [64–67] in agricultural habitats has been key to shaping the distribution of genome-wide variation of assorted complexity across the landscape.

While genome size does not show substantial variation across the range, the composition of repeat content does, suggesting the possibility of stabilizing selection on genome size with repeats competing for limited space in the genome [68–70]. The processes governing latitudinal clines in TIRs, Helitrons, and Copia beyond neutral expectations are unclear, but could be mediated by differences in transposition rate and/or selection at the host or repeat level [71] and may depend on their distribution across the genome. TE families varied tremendously in their correlations with gene density and the population-scaled recombination rate ($4N_e r$). Recombination rate has been shown to vary with temperature [72–74] and thus latitude, implying that geographic co-evolution of repeat content and recombination rate [18] may occur to differential extents across repeat types based on their distribution across the genome. In *A. tuberculatus*, LTR density is strongly governed by CDS density, consistent with their enrichment in gene-poor centromeric regions across species [75] and a stronger selection for removal and/or an insertion preference away from genes in low recombination rate regions [18]. In contrast, TIR superfamilies are more often positively associated with CDS and more typically occurring in high recombination rate regions, which may drive their differential

associations with latitude. Furthermore, methylation has been shown to covary with latitude and climate variables [76], with demethylation having a potentially adaptive role at low temperatures [77,78], suggesting that individuals with a higher content of active TE families may in part be due to differences in host silencing across environments. Overall, these results demonstrate that repeat associations with latitude and habitat are mediated by factors other than simply through their contributions to genome size.

TEs not only showed broadscale clines across the range, but across habitats (natural versus agricultural) depending on sex, mimicking the pattern seen for genome size. We expect that linked selection played an important role in generating this pattern. Recent work has shown that the *A. tuberculatus* chromosome that contains the male sex determining region harbors a fragment of the *Flowering Time* and *Heading date 3a* genes [79] in addition to the evidence we provide here of copy number and polygenic variation for flowering time being differentiated across sexes. Because the sex-determining region represents a large region of low recombination (the sex bias in genome size here suggesting up to 11 Mb of which may be absent in males), one possibility is that agricultural selection on a Y-haplotype that contained an early flowering variant of these genes could have by chance driven lower repeat content to higher frequency in such environments.

Taken together, the processes generating and governing variation for growth related life history traits are more complex than typically assumed. Flowering time is influenced in part by genome size—which, in the absence of polyploidy events and large-scale heterochromatic knobs, may predominantly reflect the balance between transposition and host removal for individual TE families, with some role for stabilizing selection on total genome size. We show this balance varies across numerous genomic, organismal, and environmental axes, and will thus require diverse range-wide collections (e.g. [80,81]) along with experimental quantification of transposition rates (e.g. [82]) and fitness effects to fully disentangle. Furthermore, quantification of individual element frequencies using long-read population sequencing will enable the characterization of element insertion frequencies, to assess the potential role of individual TE insertions in rapid adaptation to novel environments.

Methods

We used the female reference genome as described in Kreiner et al., (2019). Briefly, the *A. tuberculatus* reference genome was sequenced and assembled from an individual female plant from an agricultural habitat. The resulting 2514 contigs were scaffolded onto a chromosome resolved reference genome of a closely related species, *A. hypochondriacus* [83], creating a reference with 16 pseudochromosomes [54]. The accompanying gene annotation (described in [54]) was also used.

For all analyses, we used the 187 samples which were previously sequenced and analyzed in [37,38]. Briefly, each sample comes from intra-population crosses that were performed to control for maternal phenotypic effects, with no two samples having the same maternal or paternal genotype. The collections originated from 17 paired agricultural and natural populations, all of which were located within 25 km of each other, and spanned three degrees of latitude and 12 degrees of longitude [37]. Out of 200 samples from unique maternal lines, 187 were successfully sequenced with short-reads (2 x 150) on the Illumina NovaSeq 6000 platform, with sequencing depths ranging in coverage between 20–35X [37]. Due to a high error rate (>5%) five of the samples were excluded from the study, leaving 182 samples for the estimation and analysis of repeat content.

Phenotypic data for 176 of the sequenced samples, that among other traits included flowering time and vertical growth rate, was collected in a common garden experiment [37]. For

each maternal line, 30 replicates were grown, one sibling in each of the three treatments (water supplemented, control, and soy competition), replicated across 10 experimental blocks in the common garden experiment [37]. For this study, we only used phenotypic data collected in the control treatment [37]. For each maternal lineage, flowering times were averaged across the ten siblings in the different experimental blocks. The proportions of var. *rudis* ancestry was estimated with the Faststructure algorithm using SNP genetic information, also described in [37,84]. The relatedness matrix from genome-wide SNPs across all individuals was computed in gemma [85] using the centered genotype matrix algorithm (S5 Fig).

We took two approaches for estimating genome size. One was a kmer based approach, where we first counted 21-mers using the program KMC [86], and then fitted a Bayesian model of genome size to the histogram of 21-mers using Genomescope2.0 [44]. We tested this program using multiple upper-limit thresholds of coverage for kmer counts, which balances potential contamination from high copy organellar sequence and the inclusion of nuclear repeats with syntenicity to these regions. We found that default settings (coverage threshold of 10,000) led to a consistent underestimate of genome size relative to expectations from flow cytometry (~650 Mb; [47]). In comparison, increasing the coverage threshold to 100,000 kmer counts led to overall increases in genome size. However, this approach was unable to converge on an estimate of genome size for 29/182 samples. Our second approach was to use mapped read-depth of coverage to estimate genome size. We did so by calculating coverage of each base in the genome for a particular sample using samtools depth, estimating the cumulative sum of this coverage, and scaling that coverage by the median coverage of genic regions (which represent on average, single copy regions) to get an estimate of genome size. This read-depth based estimate was highly concordant with the kmer 100k threshold approach ($r = 0.92$, compared to an $r = 0.76$ with the default kmer settings approach; S6 Fig), with a mean and variance in genome size across individuals even closer to the prior from flow cytometry (594Mb, ranging from 543Mb to 650Mb). Therefore, we used read-depth based genome size estimates for the analyses in this paper, while all the results are qualitatively and quantitatively concordant regardless of method of estimation.

Non-overlapping TE annotation

EDTA (Extensive de-novo TE Annotator)-v1.9.7 was used to detect TEs in the reference genome and produce both a TE library and a non-overlapping annotation [87]. For TE library curation, the EDTA pipeline combines a number of high-performing TE finding programs and filters their outputs to produce a comprehensive and non-redundant TE library. LTRs are identified and filtered based on structural features by a combination of LTR_HARVEST_parallel, LTR_FINDER_parallel and LTR_retriever [87–89]. Helitrons were detected by Helitron-Scanner-v1.1 which uses a two-layered local combinational variable (LCV) algorithm [90]. TIRs are detected by machine learning with the TIR-Learner2.5 program [91]. Following a number of filtering steps, the EDTA program reduces interlibrary redundancy between LTR-RTs, Helitrons and TIR elements, combines them into one library and clusters the TEs into families based on a modified 80-80-95 Wicker rule, resulting in one representative sequence per family. The resulting library of representative TEs is used to mask the genome with RepeatMasker-4.1.1 and the remaining unmasked regions are searched by RepeatModeler-2.0.1 for missed TEs, including SINEs and LINEs [92,93]. Redundancy between the RepeatModeler library and the EDTA library is then removed and the two libraries are merged to create the final EDTA TE library. Furthermore, any repeats that are detected but not identified as TEs by the EDTA program, such as simple repeats, are listed as repeat regions, which we refer to as unknown repeats in this study.

We then used the EDTA annotation function to annotate the *A. tuberculatus* reference genome. The EDTA annotation function combines the high confidence structure-based annotations produced by structure-based programs in EDTA with homology-based annotations produced by RepeatMasker-4.1.1, using the EDTA library. Additionally, the annotation resolved overlapping regions using the following priorities: structure-based annotation > homology-based annotation, longer TE>shorter TE> nested inner TE> nested outer TE [87].

Ribosomal RNA annotation

RNAmmmer -1.2 was used to annotate eukaryotic rDNA [94]. RNAmmmer annotates rDNA using hmm's that have captured structural features of rDNA from multiple alignments of rDNA database sequences across different species. The output included subunits: 28S, 18S, and 8S. The 8S subunit is most likely synonymous to the 5S subunit, as confirmed by doing a blast search of the identified 8S subunit, finding it aligns with >95% identity to 5S subunits in other plant species. For this reason, we refer to 8S rDNA as 5S rDNA in this study.

Simple sequence annotation

To annotate simple sequences, low complexity regions and tandem repeats, we ran RepeatMasker on the reference genome with default parameters [92].

Estimating repeat abundances

To estimate the abundances of TEs, rDNA, simple repeats and low complexity regions, we first created a nonoverlapping annotation file combining the TE annotation, rDNA annotation and the simple repeats annotation. The following order of priorities was used to resolve the overlapping sequences of repeats: rDNA>known TE>simple repeats and low complexity regions>unknown repeats. Any sequences that were left with fewer than or equal to 20 base pairs in length were removed, as they are more likely to be false positives. The non-overlapping annotation was then used to estimate both the copy number of individual TEs annotated in the reference for each individual, and the additive bp contribution of repeat classes to each individual's genome. To do so, we calculated the mean coverage of reads mapping within each repeat using mosdepth [95] and scaled it by the median gene coverage genome wide to get an estimate of diploid copy number at each annotated region. We then multiplied the repeat element copy number by the length of that repeat, and summed within superfamilies and orders to get the bp repeat abundance for each, within each individual.

Statistical analyses of genome size and TE abundances

The ggplot2 package was used for visualization in most figures [96], in combination with the plot_grid function in cowplot [97] to create panels of multiple figures. The size and number of TEs was visualized directly from the non-overlapping annotation based on counts and bp ranges of each superfamily. Distance to the nearest gene was calculated using bedtools closest [98], finding the gene closest to each repeat element, reporting the distance with option -d. The fraction of each genomic window composed of different repeat types was calculated with the bedops command 'bedmap' [99], using the options--echo--count--bases-uniq-f. Population recombination rate was estimated with LDhat [100] in [54].

The variance and coefficient of variation of repeat abundance were both calculated in R, the latter by dividing the standard deviation by the mean of the repeat abundance distribution and multiplying by 100. We explored the predictors of individual repeat class abundances across populations. For each of the 16 repeat and TE classes (e.g. Copia LTRs, Ty3 LTRs, helitrons, 5S

rDNAs, simple repeats, etc.), as well as rDNAs, LTRs, and TIRs as a whole. To do so, we implemented a mixed effect model with the function `relmatLmer` from the package `lme4qtl` [101], which includes the relatedness matrix as a random effect, in a model that included the fixed effect terms:

$$\text{Repeat class} = \text{Long} + \text{Lat} + \text{Sex*Env} + \text{Ancestry} + \text{Residual Error} + (1|\text{Relatedness}) \quad [\text{Eq 1}]$$

where the $(1|)$ notation indicates a random effect after the bar operator, and the $*$ indicates fitting both main effects and the interaction effect for those terms.

All models were evaluated with an anova using type III sums of squares. Since these models are identical in structure, but differ in their response term, whether to do stringent FDR correction is debated. Regardless, we performed a within model FDR correction of p-values, and compared the FDR across the 17 predictors that initially pass a p-value < 0.05 threshold. Using this framework, we find the maximum FDR is 19% (i.e., $\sim 3/17$ discoveries are likely false positives incurred by interpreting them as significant; S3 Table).

To test whether flowering time and growth rate can be explained by variation in genome size, we implemented an additional model using `relmatLmer`:

$$\text{Phenotypic trait} = \text{Genome Size} + \text{Long} + \text{Lat} + \text{Sex*Env} + \text{Ancestry} + \text{Residual Error} + (1|\text{Relatedness}) \quad [\text{Eq 2}]$$

where the $(1|)$ and $*$ notation are as described previously. For all models, percent of model variance explained by each term in the model was calculated with `r2beta` [102] and `r2` for random versus fixed effects with the function `r.squaredGLMM` [103].

Characterization of other architectures underlying flowering time

We implemented two genome-wide association (GWA) approaches to better understanding the genetic architecture of flowering time in *A. tuberculatus*. Firstly, we investigated the extent that copy number variation is associated with flowering time through a GWA, using the family mean flowering time (and growth rate, to no avail) measured in the control environment from the common garden experiment in [37]. To estimate copy number variation, we used an approach somewhat paralleling TE abundance. First, we used `mosdepth` [95] to get the mean read coverage within each gene for each individual, and then scaled it by an individual's median genic coverage genome-wide to get an estimate of gene copy number. While for TE abundance estimates, we summed the scaled coverage (copy number) multiplied by element length across all elements of a given repeat class, for this gene level analysis, we used the gene-level estimate of copy number across individuals for further input into a GWA for flowering time.

The positive correlation of genome size and flowering time as previously described could also reflect the greater opportunity for particular phenotype affecting insertion event rather than a direct effect of repeat abundance on flowering time. To explore this hypothesis, we conducted a GWA of flowering time for repeat (as opposed to gene) copy number, following the same approach as above (stopping before summing across distinct annotated features). The associations across the genome greatly mirrored that of the gene level analysis (S7 Fig) with only one clear peak on Scaffold 10 at the same ATP synthesis locus. While TEs in the region may have causally impacted flowering time, another likely explanation is that TEs were amplified along with numerous genes in this region, hitchhiking along with the phenotypic effects of the host genes. By jointly modeling the effect of copy number variation at this locus along with genome size on flowering time, we further distinguish these alternatives. In both cases, we implemented a p-value FDR correction in R using the base function “p-adjust” and further investigated genes and repeats passing the 10% FDR threshold.

We next ran a GWA on high quality filtered SNPs from [37]. We did so using gemma [85] after using plink to convert from a vcf to binary file format, and using the gemma generated relatedness matrix as a covariate. With the 97 SNPs found across the genome that pass the 10% FDR threshold, we calculated each individual's polygenic value [46,47] for flowering time. The polygenic value was calculated as:

$$PGV_{FT} = 2 \sum_{l=1}^L \alpha p$$

where L equals the 97 loci with an FDR $q < 0.1$, α the effect size of the minor allele on flowering time from the GWA, and p an individual's genotype (equal to 0.5 for heterozygotes and 1 for homozygous alternates). Our calculation of polygenic values for flowering time was based on genotypes at just the 97 SNPs genome-wide that passed a 10% FDR correction rather than genome-wide SNPs, in an effort to circumvent issues with uncontrolled population structure [104–106].

Since ancestry varies along the axes of climate and length of growing the season in *A. tuberculatus*, we also tested whether controlling for population structure confounds a significant amount of genetic variation for flowering time by performing a GWA without such a control. Interestingly, several peaks pop out in this GWAS that are not present in the population structure corrected GWAS. The major peak on scaffold 10 corresponds with *ARF5/ARF19*, on scaffold 13 with *PME61*, and on scaffold 5 with *HEC*, and the 3454 genes that pass a 10% FDR correction show significant enrichment for numerous GO biological functions, including several of which are related to flowering time (e.g. reproductive shoot development, post-embryonic morphogenesis). The model of flowering time by genetic predictor (Eq 3) with the PGV_{FT} based on the uncorrected GWAS explains an additional 9% of the variation in flowering time, implying differences among populations collinear with differences in the architecture of flowering time among populations.

To compare the relative importance of genome size to other genomic predictors of flowering, we implemented an additional mixed effect models while accounting for relatedness with relmatLmer:

$$\text{Flowering Time} = \text{Genome Size} + \text{ATP Locus Copy Number} + PGV_{FT} + \text{Residual Error} + (1|\text{Relatedness}) \quad [\text{Eq 3}]$$

where the (1|) and * notation are as described previously. Since genome size, copy number of the ATP locus, and PGV_{FT} are all on very different scales, we transformed these predictors to z-values. Since the PGV_{FT} showed a non-normal distribution, we did this rescaling after making each individual's estimate > 0 (adding the minimum absolute value across individuals to each observation) and log10 transformation.

Finally, we asked whether each of these genomic predictors of flowering time show differentiation within and among populations, based on geographic, sex, habitat, and historical predictors. We therefore implemented Eq 1, but using the raw genome size estimates, raw ATP locus copy number estimates, and the $\log_{10}(PGV_{FT} + \min(\text{abs}(PGV_{FT})))$ as dependent variables.

Supporting information

S1 Fig. Bivariate correlations of TE density by superfamily with coding sequence density across the genome in 100kb windows.

(PDF)

S2 Fig. Bivariate correlations of TE density by superfamily with effective recombination rate ($4N_e r$) across the genome in 100kb windows.

(PDF)

S3 Fig. Repeat composition within and across individuals, with individuals oriented from south to north from left to right.

(PDF)

S4 Fig. Mean repeat composition across populations of *A. tuberculatus*. Base map from the Natural Earth public domain map dataset (<https://www.naturalearthdata.com/>).

(PDF)

S5 Fig. The relatedness matrix based on centered genotypes as computed in plink. Colors on the left represent population groupings ordered by longitude, with the most eastern populations in darker colours.

(PDF)

S6 Fig. Concordance of kmer and read-depth based estimates of genome size. Top figure shows the relationship between kmer estimates of genome size depending on kmer coverage threshold (default = 10,000 versus 100,000). Solid black line represents the 1:1 expectation.

(PDF)

S7 Fig. The distribution of mapped read-depth inferred genome sizes in this study, where the median estimated genome size is indicated by the vertical dashed line.

(PDF)

S8 Fig. Results from a genome-wide association of flowering time with genotypes, without controlling for population structure. Horizontal line represents a FDR 10% cutoff. SNPs above this line are enriched for GO biological processes listed above.

(PDF)

S9 Fig. Copy number associations with flowering time across the genome, for A) genes, and B) repeats. Horizontal dashed line indicates a 5% false discovery rate threshold.

(PDF)

S1 Table. Biological GO Enrichment results after Bonferroni correction from the gene-level copy number GWAS.

(XLSX)

S2 Table. Significant SNPs from a GWA with flowering time, for those within *A. tuberculatus* genes that have an orthologous match in *Arabidopsis thaliana*.

(XLSX)

S3 Table. Significance of associations between predictors and repeat classes for those with p-values < 0.05. Note that the maximum FDR for these relationships is 19%, implying that 3 of these 17 discoveries may be false positives.

(XLSX)

Acknowledgments

Thanks to the Whitlock lab, Sally Otto, Tyler Kent, and three anonymous reviewers for useful feedback on the manuscript.

Author Contributions

Conceptualization: Julia M. Kreiner, Solomiya Hnatovska, John R. Stinchcombe, Stephen I. Wright.

Formal analysis: Julia M. Kreiner, Solomiya Hnatovska.

Funding acquisition: Julia M. Kreiner, John R. Stinchcombe, Stephen I. Wright.

Investigation: Julia M. Kreiner, Solomiya Hnatovska.

Methodology: Julia M. Kreiner, Solomiya Hnatovska, Stephen I. Wright.

Project administration: Julia M. Kreiner, Solomiya Hnatovska, John R. Stinchcombe, Stephen I. Wright.

Supervision: Julia M. Kreiner, John R. Stinchcombe, Stephen I. Wright.

Visualization: Julia M. Kreiner, Solomiya Hnatovska.

Writing – original draft: Julia M. Kreiner, Solomiya Hnatovska, Stephen I. Wright.

Writing – review & editing: Julia M. Kreiner, Solomiya Hnatovska, John R. Stinchcombe, Stephen I. Wright.

References

- Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat Genet.* 2013; 45: 884–890. <https://doi.org/10.1038/ng.2678> PMID: 23793030
- Bilinski P, Albert PS, Berg JJ, Birchler JA, Grote MN, Lorant A, et al. Parallel altitudinal clines reveal trends in adaptive evolution of genome size in *Zea mays*. *PLoS Genet.* 2018; 14: e1007162. <https://doi.org/10.1371/journal.pgen.1007162> PMID: 29746459
- Mirsky AE, Ris H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J Gen Physiol.* 1951; 34: 451–462. <https://doi.org/10.1085/jgp.34.4.451> PMID: 14824511
- Swift HH. The desoxyribose nucleic acid content of animal nuclei. *Physiol Zool.* 1950; 23: 169–198. <https://doi.org/10.1086/physzool.23.3.30152074> PMID: 15440320
- Hahn MW, Wray GA. The g-value paradox. *Evol Dev.* 2002; 4: 73–75. <https://doi.org/10.1046/j.1525-142x.2002.01069.x> PMID: 12004964
- Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 2009; 5: e1000734. <https://doi.org/10.1371/journal.pgen.1000734> PMID: 19956538
- Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 2000; 408: 796–815. <https://doi.org/10.1038/35048692> PMID: 11130711
- Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. The genomic ecosystem of transposable elements in maize. *PLoS Genet.* 2021; 17: e1009768. <https://doi.org/10.1371/journal.pgen.1009768> PMID: 34648488
- Charlesworth B, Charlesworth D. The population dynamics of transposable elements. *Genet Res.* 1983; 42: 1–27.
- Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet.* 2008; 9: 397–405. <https://doi.org/10.1038/nrg2337> PMID: 18368054
- Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 2009; 19: 1419–1428. <https://doi.org/10.1101/gr.091678.109> PMID: 19478138
- Bonchev G, Parisod C. Transposable elements and microevolutionary changes in natural populations. *Mol Ecol Resour.* 2013; 13: 765–775. <https://doi.org/10.1111/1755-0998.12133> PMID: 23795753
- Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends Biochem Sci.* 2016; 41: 324–337. <https://doi.org/10.1016/j.tibs.2015.12.008> PMID: 26810602
- Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. *Biochim Biophys Acta Gene Regul Mech.* 2017; 1860: 157–165. <https://doi.org/10.1016/j.bbagr.2016.05.010> PMID: 27235540
- McCue AD, Panda K, Nuthikattu S, Choudury SG, Thomas EN, Slotkin RK. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *EMBO J.* 2015; 34: 20–35. <https://doi.org/10.15252/embj.201489499> PMID: 25388951
- Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH. Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. *Genome Biol Evol.* 2016; 8: 2964–2978. <https://doi.org/10.1093/gbe/evw208> PMID: 27566762

17. Lynch M, Conery JS. The origins of genome complexity. *Science*. 2003; 302: 1401–1404. <https://doi.org/10.1126/science.1089370> PMID: 14631042
18. Kent TV, Uzunović J, Wright SI. Coevolution between transposable elements and recombination. *Philos Trans R Soc Lond B Biol Sci*. 2017;372. <https://doi.org/10.1098/rstb.2016.0458> PMID: 29109221
19. Nguyen AH, Bachtrog D. Toxic Y chromosome: Increased repeat expression and age-associated heterochromatin loss in male *Drosophila* with a young Y chromosome. *PLoS Genet*. 2021; 17: e1009438. <https://doi.org/10.1371/journal.pgen.1009438> PMID: 33886541
20. Marais GAB, Gaillard J-M, Vieira C, Ploton I, Sanlaville D, Gueyffier F, et al. Sex gap in aging and longevity: can sex chromosomes play a role? *Biol Sex Differ*. 2018; 9: 33. <https://doi.org/10.1186/s13293-018-0181-y> PMID: 30016998
21. Ågren JA, Wright SI. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? *Chromosome Res*. 2011; 19: 777–786. <https://doi.org/10.1007/s10577-011-9229-0> PMID: 21850458
22. Bui QT, Grandbastien M-A. LTR Retrotransposons as Controlling Elements of Genome Response to Stress? In: Grandbastien M-A, Casacuberta JM, editors. *Plant Transposable Elements: Impact on Genome Structure and Function*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. pp. 273–296.
23. Negi P, Rai AN, Suprasanna P. Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response. *Front Plant Sci*. 2016; 7: 1448. <https://doi.org/10.3389/fpls.2016.01448> PMID: 27777577
24. Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D. Stress-Driven Transposable Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. *Mol Biol Evol*. 2020; 37: 221–239. <https://doi.org/10.1093/molbev/msz216> PMID: 31553475
25. Wos G, Choudhury RR, Kolář F, Parisod C. Transcriptional activity of transposable elements along an elevational gradient in *Arabidopsis arenosa*. *Mob DNA*. 2021; 12: 7. <https://doi.org/10.1186/s13100-021-00236-0> PMID: 33639991
26. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. *Mol Ecol*. 2013; 22: 1503–1517. <https://doi.org/10.1111/mec.12170> PMID: 23293987
27. Dubin MJ, Mittelsten Scheid O, Becker C. Transposons: a blessing curse. *Curr Opin Plant Biol*. 2018; 42: 23–29. <https://doi.org/10.1016/j.pbi.2018.01.003> PMID: 29453028
28. Blommaert J. Genome size evolution: towards new model systems for old questions. *Proc Biol Sci*. 2020; 287: 20201441. <https://doi.org/10.1098/rspb.2020.1441> PMID: 32842932
29. Mei W, Stetter MG, Gates DJ, Stitzer MC, Ross-Ibarra J. Adaptation in plant genomes: Bigger is different. *Am J Bot*. 2018; 105: 16–19. <https://doi.org/10.1002/ajb2.1002> PMID: 29532920
30. Baetcke KP, Sparrow AH, Nauman CH, Schwemmer SS. The relationship of DNA content to nuclear and chromosome volumes and to radiosensitivity (LD50). *Proc Natl Acad Sci U S A*. 1967; 58: 533–540. <https://doi.org/10.1073/pnas.58.2.533> PMID: 5233456
31. Cavalier-Smith T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J Cell Sci*. 1978; 34: 247–278. <https://doi.org/10.1242/jcs.34.1.247> PMID: 372199
32. Beaulieu JM, Moles AT, Leitch IJ, Bennett MD, Dickie JB, Knight CA. Correlated evolution of genome size and seed mass. *New Phytol*. 2007; 173: 422–437. <https://doi.org/10.1111/j.1469-8137.2006.01919.x> PMID: 17204088
33. Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol*. 2008; 179: 975–986. <https://doi.org/10.1111/j.1469-8137.2008.02528.x> PMID: 18564303
34. Gregory TR. Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biol Rev Camb Philos Soc*. 2007; 76: 65–101.
35. Díez CM, Gaut BS, Meca E, Scheinvar E, Montes-Hernandez S, Eguarte LE, et al. Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytol*. 2013; 199: 264–276. <https://doi.org/10.1111/nph.12247> PMID: 23550586
36. Waselkov KE, Olsen KM. Population genetics and origin of the native North American agricultural weed waterhemp (*Amaranthus tuberculatus*; *Amaranthaceae*). *Am J Bot*. 2014; 101: 1726–1736. <https://doi.org/10.3732/ajb.1400064> PMID: 25091000
37. Kreiner JM, Caballero A, Wright SI, Stinchcombe JR. Selective ancestral sorting and de novo evolution in the agricultural invasion of *Amaranthus tuberculatus*. *Evolution*. 2021. <https://doi.org/10.1111/evo.14404> PMID: 34806764
38. Kreiner JM, Latorre SM, Burbano HA, Stinchcombe JR, Otto SP, Weigel D, et al. Rapid weed adaptation and range expansion in response to agriculture over the past two centuries. *Science*. 2022; 378: 1079–1085. <https://doi.org/10.1126/science.abo7293> PMID: 36480621

39. Kreiner JM, Giacomini DA, Bemm F, Waithaka B, Regalado J, Lanz C, et al. Multiple modes of convergent adaptation in the spread of glyphosate-resistant *Amaranthus tuberculatus*. *Proc Natl Acad Sci U S A*. 2019; 116: 21076–21084.
40. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007; 8: 973–982. <https://doi.org/10.1038/nrg2165> PMID: 17984973
41. Galindo-González L, Mhiri C, Deyholos MK, Grandbastien M-A. LTR-retrotransposons in plants: Engines of evolution. *Gene*. 2017; 626: 14–25. <https://doi.org/10.1016/j.gene.2017.04.051> PMID: 28476688
42. Sáez-Vásquez J, Delseny M. Ribosome Biogenesis in Plants: From Functional 45S Ribosomal DNA Organization to Ribosome Assembly Factors. *Plant Cell*. 2019; 31: 1945–1967. <https://doi.org/10.1105/tpc.18.00874> PMID: 31239391
43. Kang M, Wang J, Huang H. Nitrogen limitation as a driver of genome size evolution in a group of karst plants. *Sci Rep*. 2015; 5: 11636.
44. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020; 11: 1432. <https://doi.org/10.1038/s41467-020-14998-3> PMID: 32188846
45. Puixeu G, Pickup M, Field DL, Barrett SCH. Variation in sexual dimorphism in a wind-pollinated plant: the influence of geographical context and life-cycle dynamics. *New Phytol*. 2019; 224: 1108–1120. <https://doi.org/10.1111/nph.16050> PMID: 31291691
46. Fisher RA. *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. OUP Oxford; 1930.
47. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. *PLoS Genet*. 2014; 10: e1004412. <https://doi.org/10.1371/journal.pgen.1004412> PMID: 25102153
48. Poggio L, Rosato M, Chiavarino AM, Naranjo CA. Genome Size and Environmental Correlations in Maize (*Zea mays* ssp. *mays*, Poaceae). *Ann Bot*. 1998; 82: 107–115.
49. Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, et al. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol Ecol*. 2019; 28: 1506–1522. <https://doi.org/10.1111/mec.14963> PMID: 30506554
50. Gregory TR. The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. *Blood Cells Mol Dis*. 2001; 27: 830–843. <https://doi.org/10.1006/bcmd.2001.0457> PMID: 11783946
51. Stelzer C-P, Pichler M, Hatheuer A. Linking genome size variation to population phenotypic variation within the rotifer, *Brachionus asplanchnoidis*. *Commun Biol*. 2021; 4: 596.
52. Trávníček P, Čertner M, Ponert J, Chumová Z, Jersáková J, Suda J. Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol*. 2019; 224: 1642–1656. <https://doi.org/10.1111/nph.15996> PMID: 31215648
53. Símová I, Herben T. Geometrical constraints in the scaling relationships between genome size, cell size and cell cycle length in herbaceous plants. *Proc Biol Sci*. 2012; 279: 867–875. <https://doi.org/10.1098/rspb.2011.1284> PMID: 21881135
54. Kreiner JM, Giacomini D, Bemm F, Waithaka B. Multiple modes of convergent adaptation in the spread of glyphosate-resistant *Amaranthus tuberculatus*. *BioRxiv*. 2018. Available: <https://www.biorxiv.org/content/https://doi.org/10.1101/498519v1.abstract>
55. Monnahan PJ, Kelly JK. The Genomic Architecture of Flowering Time Varies Across Space and Time in *Mimulus guttatus*. *Genetics*. 2017; 206: 1621–1635. <https://doi.org/10.1534/genetics.117.201483> PMID: 28455350
56. Todesco M, Owens GL, Bercovich N, Légaré J-S, Soudi S, Burge DO, et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *bioRxiv*. 2019. p. 790279. <https://doi.org/10.1101/790279>
57. Battlay P, Wilson J, Bieker VC, Lee C, Prapas D, Petersen B, et al. Large haploblocks underlie rapid adaptation in the invasive weed *Ambrosia artemisiifolia*. *Nat Commun*. 2023; 14: 1717. <https://doi.org/10.1038/s41467-023-37303-4> PMID: 36973251
58. Slattery RA, VanLoocke A, Bernacchi CJ, Zhu X-G, Ort DR. Photosynthesis, Light Use Efficiency, and Yield of Reduced-Chlorophyll Soybean Mutants in Field Conditions. *Front Plant Sci*. 2017; 8: 549. <https://doi.org/10.3389/fpls.2017.00549> PMID: 28458677
59. Simkin AJ, López-Calcano PE, Raines CA. Feeding the world: improving photosynthetic efficiency for sustainable crop production. *J Exp Bot*. 2019; 70: 1119–1140. <https://doi.org/10.1093/jxb/ery445> PMID: 30772919

60. Simkin AJ, Lopez-Calcano PE, Davey PA, Headland LR, Lawson T, Timm S, et al. Simultaneous stimulation of sedoheptulose 1,7-bisphosphatase, fructose 1,6-bisphosphate aldolase and the photorepiratory glycine decarboxylase-H protein increases CO₂ assimilation, vegetative biomass and seed yield in *Arabidopsis*. *Plant Biotechnol J*. 2017; 15: 805–816. <https://doi.org/10.1111/pbi.12676> PMID: 27936496
61. Casson SA, Topping JF, Lindsey K. MERISTEM-DEFECTIVE, an RS domain protein, is required for the correct meristem patterning and function in *Arabidopsis*. *Plant J*. 2009; 57: 857–869. <https://doi.org/10.1111/j.1365-313X.2008.03738.x> PMID: 19000164
62. Zhang Y, Schwarz S, Saedler H, Huijser P. SPL8, a local regulator in a subset of gibberellin-mediated developmental processes in *Arabidopsis*. *Plant Mol Biol*. 2007; 63: 429–439. <https://doi.org/10.1007/s11103-006-9099-6> PMID: 17093870
63. Xu C, Fang X, Lu T, Dean C. Antagonistic cotranscriptional regulation through ARGONAUTE1 and the THO/TREX complex orchestrates FLC transcriptional output. *Proc Natl Acad Sci U S A*. 2021; 118. <https://doi.org/10.1073/pnas.2113757118> PMID: 34789567
64. Lawrence CW. Genetic studies on wild populations of *Melandrium* III. *Heredity*. 1964; 19: 1–19.
65. Webb CJ. Flowering periods in the gynodioecious species *Gingidia decipiens* (Umbelliferae). *N Z J Bot*. 1976; 14: 207–210.
66. Purrington CB, Schmitt J. Consequences of sexually dimorphic timing of emergence and flowering in *Silene latifolia*. *J Ecol*. 1998; 86: 397–404.
67. Aljiboury AA, Friedman J. Mating and fitness consequences of variation in male allocation in a wind-pollinated plant. *Evolution*. 2022; 76: 1762–1775. <https://doi.org/10.1111/evo.14544> PMID: 35765717
68. Beaulieu JM. The right stuff: evidence for an “optimal” genome size in a wild grass population. *The New phytologist*. 2010. pp. 883–885. <https://doi.org/10.1111/j.1469-8137.2010.03412.x> PMID: 20707851
69. Šmarda P, Horová L, Bureš P, Hralová I, Marková M. Stabilizing selection on genome size in a population of *Festuca pallens* under conditions of intensive intraspecific competition. *New Phytol*. 2010; 187: 1195–1204. <https://doi.org/10.1111/j.1469-8137.2010.03335.x> PMID: 20561203
70. Hjelmén CE, Parrott JJ, Srivastav SP, McGuane AS, Ellis LL, Stewart AD, et al. Effect of Phenotype Selection on Genome Size Variation in Two Species of Diptera. *Genes*. 2020; 11. <https://doi.org/10.3390/genes11020218> PMID: 32093067
71. Ågren JA. Selfish genetic elements and the gene’s-eye view of evolution. *Curr Zool*. 2016; 62: 659–665. <https://doi.org/10.1093/cz/zow102> PMID: 29491953
72. Lloyd A, Morgan C, H Franklin FC, Bomblies K. Plasticity of Meiotic Recombination Rates in Response to Temperature in *Arabidopsis*. *Genetics*. 2018; 208: 1409–1420. <https://doi.org/10.1534/genetics.117.300588> PMID: 29496746
73. Weitz AP, Dukic M, Zeitler L, Bomblies K. Male meiotic recombination rate varies with seasonal temperature fluctuations in wild populations of autotetraploid *Arabidopsis arenosa*. *Mol Ecol*. 2021; 30: 4630–4641. <https://doi.org/10.1111/mec.16084> PMID: 34273213
74. Kohl KP, Singh ND. Experimental evolution across different thermal regimes yields genetic divergence in recombination fraction but no divergence in temperature associated plastic recombination. *Evolution*. 2018; 72: 989–999. <https://doi.org/10.1111/evo.13454> PMID: 29468654
75. Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view. *Chromosoma*. 2014; 123: 313–325. <https://doi.org/10.1007/s00412-014-0462-0> PMID: 24763964
76. Galanti D, Ramos-Cruz D, Nunn A, Rodríguez-Arévalo I, Scheepens JF, Becker C, et al. Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense*. *PLoS Genet*. 2022; 18: e1010452. <https://doi.org/10.1371/journal.pgen.1010452> PMID: 36223399
77. Xie HJ, Li H, Liu D, Dai WM, He JY, Lin S, et al. ICE1 demethylation drives the range expansion of a plant invader through cold tolerance divergence. *Mol Ecol*. 2015; 24: 835–850. <https://doi.org/10.1111/mec.13067> PMID: 25581031
78. Xie H, Sun Y, Cheng B, Xue S, Cheng D, Liu L, et al. Variation in ICE1 Methylation Primarily Determines Phenotypic Variation in Freezing Tolerance in *Arabidopsis thaliana*. *Plant Cell Physiol*. 2019; 60: 152–165. <https://doi.org/10.1093/pcp/pcy197> PMID: 30295898
79. Raiyemo DA, Bobadilla LK, Tranel PJ. Genomic profiling of dioecious *Amaranthus* species provides novel insights into species relatedness and sex genes. *BMC Biol*. 2023; 21: 1–18.
80. Contreras-Garrido A, Galanti D, Movilli A, Becker C, Bossdorf O, Drost H-G, et al. Transposon dynamics in the emerging oilseed crop *Thlaspi arvense*. *bioRxiv*. 2023. p. 2023.05.24.542068. <https://doi.org/10.1101/2023.05.24.542068>

81. Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, Scott A, et al. Cycles of satellite and transposon evolution in Arabidopsis centromeres. *Nature*. 2023. <https://doi.org/10.1038/s41586-023-06062-z> PMID: 37198485
82. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui M-A, Guy J, et al. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat Commun*. 2019; 10: 3421. <https://doi.org/10.1038/s41467-019-11385-5> PMID: 31366887
83. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biol*. 2017; 15: 74. <https://doi.org/10.1186/s12915-017-0412-4> PMID: 28854926
84. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*. 2014; 197: 573–589. <https://doi.org/10.1534/genetics.114.164350> PMID: 24700103
85. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012; 44: 821–824. <https://doi.org/10.1038/ng.2310> PMID: 22706312
86. Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*. 2017. pp. 2759–2761. <https://doi.org/10.1093/bioinformatics/btx304> PMID: 28472236
87. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019; 20: 275. <https://doi.org/10.1186/s13059-019-1905-y> PMID: 31843001
88. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. 2008; 9: 18. <https://doi.org/10.1186/1471-2105-9-18> PMID: 18194517
89. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol*. 2018; 176: 1410–1422. <https://doi.org/10.1104/pp.17.01310> PMID: 29233850
90. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc Natl Acad Sci U S A*. 2014; 111: 10263–10268. <https://doi.org/10.1073/pnas.1410068111> PMID: 24982153
91. Su W, Gu X, Peterson T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol Plant*. 2019; 12: 447–460. <https://doi.org/10.1016/j.molp.2019.02.008> PMID: 30802553
92. Smit AFA, Hubley R & Green P. RepeatMasker Open-4.0. 2013–2015. Available: <http://www.repeatmasker.org>
93. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020; 117: 9451–9457. <https://doi.org/10.1073/pnas.1921046117> PMID: 32300014
94. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007; 35: 3100–3108. <https://doi.org/10.1093/nar/gkm160> PMID: 17452365
95. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018; 34: 867–868. <https://doi.org/10.1093/bioinformatics/btx699> PMID: 29096012
96. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media; 2009.
97. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2 (2020). R package version 1.1. 1. 2021.
98. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; 26: 841–842. <https://doi.org/10.1093/bioinformatics/btq033> PMID: 20110278
99. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics*. 2012; 28: 1919–1920. <https://doi.org/10.1093/bioinformatics/bts277> PMID: 22576172
100. McVean G, Auton A. LDhat 2.1: a package for the population genetic analysis of recombination. Department of Statistics, Oxford, OX1 3TG, UK. 2007. Available: <https://pdfs.semanticscholar.org/37c0/71c2bcd115bcce4bb2e4eb89a7dbb861c136.pdf>
101. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martinez-Perez A, Aschard H, Soria JM. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*. 2018; 19: 68. <https://doi.org/10.1186/s12859-018-2057-x> PMID: 29486711
102. Jaeger B. r2glmm: Computes R squared for mixed (multilevel) models. R package version 01.
103. MuMIn BK. Multi-Model Inference. R Package Version 1. 43. 15. 2019. 2020.

104. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019;8. <https://doi.org/10.7554/eLife.39702> PMID: 30895926
105. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*. 2019;8. <https://doi.org/10.7554/eLife.39725> PMID: 30895923
106. Barton N, Hermisson J, Nordborg M. Why structure matters. *eLife*. 2019. <https://doi.org/10.7554/eLife.45380> PMID: 30895925