bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Genome size and repeat content contribute to a complex architecture of flowering time in *Amaranthus tuberculatus*

Julia M. Kreiner [1,2, †,*], Solomiya Hnatovska[2,3, †], John R. Stinchcombe[2], Stephen I. Wright[2]

[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
[†]Co-first author
*Corresponding author email: julia.kreiner@ubc.ca

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. K-mer based genome size estimates vary by up to 20% across individuals, with transposable elements, unknown repeats, and rDNAs being the primary contributors to this variability. 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. Compared to newly-characterized gene copy number and polygenic nucleotide changes underlying variation in flowering time, we show that genome size remains a modest but significant contributor to the genetic basis of flowering time. Differences in flowering time across sexes and habitats are not mirrored by genome size variation, but rather polygenic variation and a gene copy number variant within the ATP synthesis pathway. Repeat content nonetheless shows non-neutral distributions across the genome, and across latitudinal and environmental gradients, reflecting numerous governing processes that in turn influence quantitative genetic variation for phenotypes key to plant adaptation.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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 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, characterizing 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 relative 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 traditionally 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 as much as 85% of the genome in maize and as little as 20% in *Arabidopsis thaliana [6,7]*. Nevertheless, the role of repetitive sequence in shaping genomic and phenotypic variation across populations remains unknown for most species.

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 [8]. 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 [9–14]. On the one hand, if effective population size and thus the

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

magnitude of drift and efficacy of selection differs among populations, population-level TE abundances may differ [15,16] and covary with population structure. Similarly, in genomic regions of low recombination (e.g., sex-determining regions), the efficacy of selection is reduced, driving another source of variation in TE content across the genome and between the sexes [17], which has recently been implicated in sex differences in survival [18,19]. 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 [20] 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 [21–24]. 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 [25,26]) 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 [27], genome size is increasingly considered an important contributor to plant adaptation [28]. Larger genomes may not only mean a larger adaptive mutational target size (especially for non-genic regulatory variation; discussed in [28]) 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 [29–32]. While the nature and direction of causality between genome size and these genomic properties has been difficult to uncouple [33], 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)* [2,34]. 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.

To further evaluate how and why repeats vary in abundance within species, we turned to the prevalent agricultural weed, *Amaranthus tuberculatus* (common waterhemp). *A. tuberculatus* is an annual, dioecious, wind-pollinated plant, whose range is centered around the Mississippi river in the United States [35]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

agricultural), and by varietal ancestry [36]. 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 [37]. 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) 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) How important is genome size in encoding flowering time and responding to environment mediated selection? To do so, we combined k-mer-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 [38]. TEs make up 66.28% or ~443 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 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. 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 [39]). 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 density; CDS) and the population recombination rate ($4N_er$) (the two of which are positively correlated in *A. tuberculatus*), likely reflecting differences in activity, transposition biases, and the strength of negative selection [17] (**Fig 1C, Sup Figs 1 & 2**). LTR elements were the largest and typically

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

found furthest from genes, while TIRs (e.g., PiF/Harbingers, 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, 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. [40]).

We also annotated simple repeats and rDNA genes. rDNAs, repetitive genes functioning in ribosome production [41], made up 4.1% of the genome and were distributed across all 16 main chromosome scaffolds. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

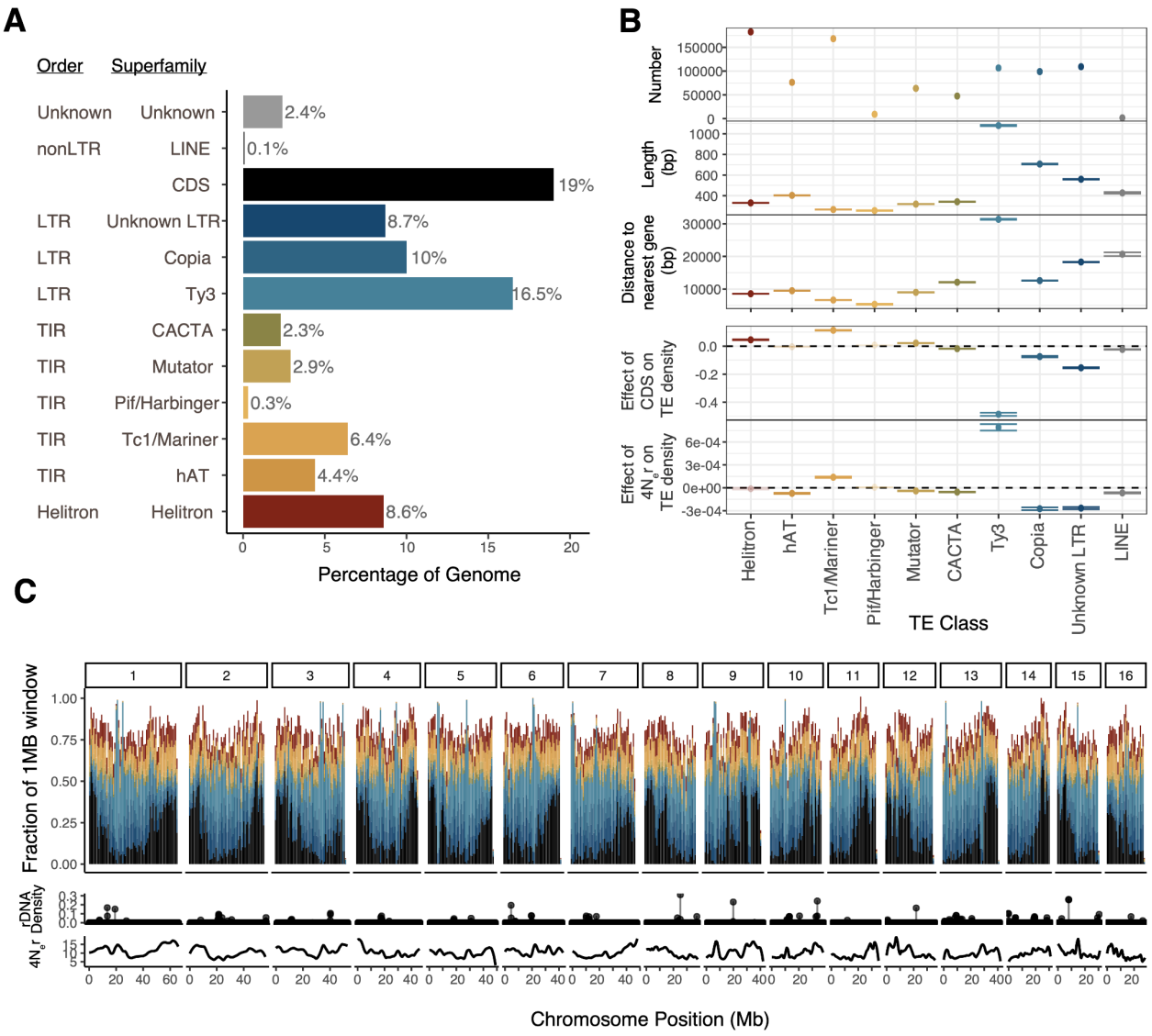


Fig 1. Variation in TE composition across the female *Amaranthus tuberculatus* reference genome by order and superfamily. A) The fraction of the female *A. tuberculatus* reference genome composed of different TE orders and superfamilies. **B)** Statistical summaries by TE superfamily, illustrating the differences in the number, size, and distance to genes across the reference genome. The bottom two rows represent the effect of coding sequence density (CDS) and $4N_er$ 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 ($4N_er$; bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Repeat variation across individuals

Looking beyond a single reference genome, we 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) **(Sup Fig 3 & 4)**. 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 (**Figure 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 proportionally 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) (**Figure 2A).**

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 [2,42], with the matrix visualized in **Sup Fig 5;** see methods **eq. 4** for mixed model). In doing so, we found widespread evidence of latitudinal clines in repeat content that exceeded expectations from neutral population structure (visualized in **Sup Fig 1)**. The abundance of several elements declined substantially with latitude, including TIR elements (B=-942124, χ^2 =5.41, p = 0.020) [Tc1: F= 7.44 , p = 0.006; PiF: F=6.92, p= 0.009; hAT: F=4.57, 0.032; Mutator (F=5.22, p=0.022)], Helitrons (B=-516999, F=5.90, p=0.015), and Copia elements (B=-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 [36]) was explicitly included as a fixed effect predictor in addition to the relatedness matrix, suggesting that such latitudinal clines are not simply the result of the different timescales of evolutionary history. Ancestry did, however, significantly predict two repeat types, Copia elements (B=-5061330, F=3.94, p=0.047) and unknown LTRs (B=-5369193, F=6.41, p=0.011) (**Fig 2B)**, suggesting that a history of reproductive 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

an effect of current 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: χ^2=14.92, p=0.0001) and unknown repeats (sex effect: χ^2=12.74, p=0.0004), the difference between sexes was greater in agricultural compared to natural habitats (sex by habitat effect; rDNAs: χ^2=4.29, p=0.038; unknown repeats: χ^2=3.28, p=0.07) **(Fig 2B)**.

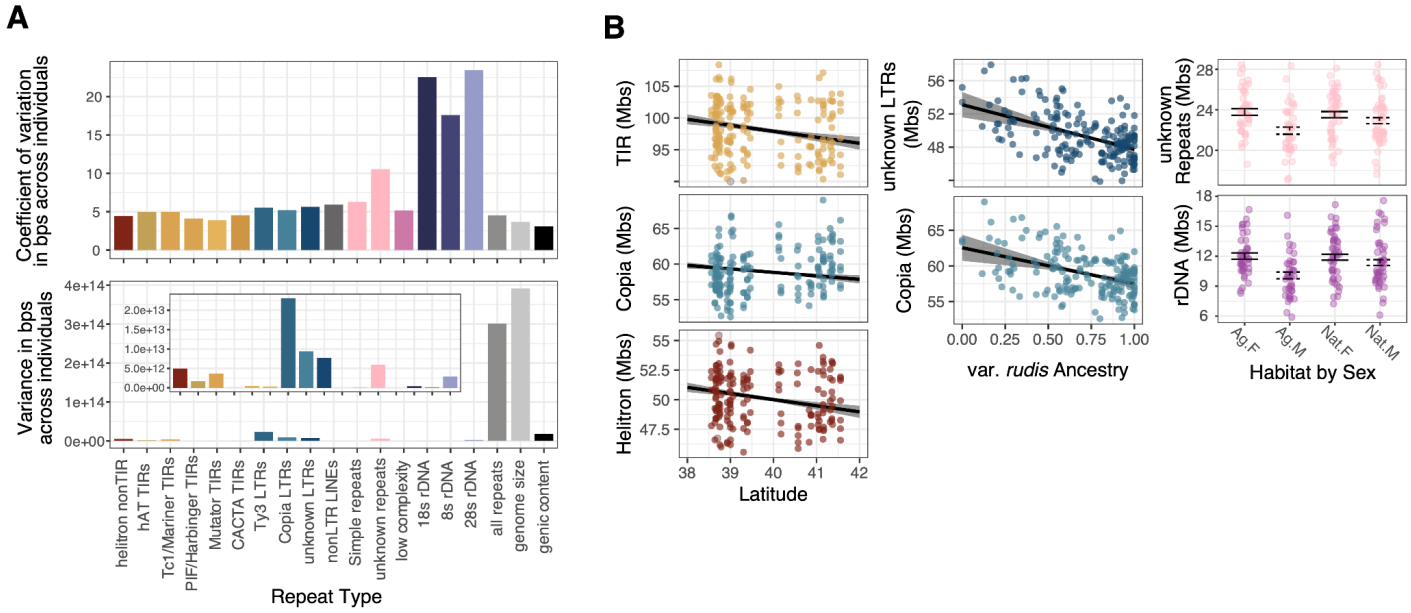


Fig 2. Variation in repeat abundance across individuals. A) The coefficient of variation (top) and variance (bottom) in bp amount of an individual's genome composed of a given repeat class. **B)** The relationship between repeat classes with latitude (left), var. *rudis* ancestry (middle), and the interaction between habitat (Ag: agricultural site; Nat: natural site) and sex (right). Points represent raw data, while regression lines and error bars represent the least squares mean from a mixed effect model that accounts for relatedness. Only significant relationships shown.

Variation in genome size

To understand the cumulative effects of this variation in repeat content, we next quantified genome size variation across individuals. Based on k-mer inference [43], there is substantial genome size variation across 186 individuals in this study, with the largest estimate being ~20% larger than the smallest estimate (min Mb = 486.2, max Mb = 580.5, mean Mb = 538.8; **Sup Fig 6, Fig 3**), about double the variability seen in *A. thaliana* (up to 10% [1]). K-mer-based estimates are known to underestimate genome size and especially so for species with a past history of whole genome duplication [44,45], and indeed the average genome size estimate from these

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

inferences falls outside of the confidence interval of genome size inferred from flow cytometry analyses of 3 individuals (95% CI 621.4–729.8 [46]). Nevertheless, our estimates of genome size correlate strongly with total repeat content in bps ($r=0.83$, $p < 10^{-15}$) and, as expected, more weakly with gene content in bps ($r=0.17$, $p = 0.024$). The estimated coefficient of variation we observe for genome size is lower than any individual repeat type **(Fig 2A)**, likely reflecting the uncoupling of different repeat class abundances within and across individuals.

To test which repeat types contribute most to variation in genome size, we modeled genome size as a function of the different repeat types and genic content (methods eq. 1). When comparing the major categories of repeats including TEs, rDNA, simple repeats, low complexity regions and other unknown repeats as independent variables in such a model, TEs ($F=195.85$, $p < 10^{-15}$), rDNAs ($F=19.37$, $p=1.9 \times 10^{-15}$), and unknown repeats ($F=34.97$, $p=1.8 \times 10^{-8}$) significantly explain genome size variation among individuals. The full model explains 72% of the variance in estimated genome size, with TEs alone explaining 32.3% of the variation in genome size, rDNAs 3.2%, and unknown repeats 5.8%.

The phenotypic consequences of genome size variation

We leveraged phenotypes measured in a common garden experiment in these same samples [36] 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 [36], 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 such 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 differences in sexual dimorphism across environments having been identified in other wind pollinated species [47], we tested for sex by habitat interactions.

We found that genome size is a significant predictor of flowering time ($F =7.70$, $p=0.006$), with every additional 10 MB predicted to delay flowering by 0.7 days (Figure 3C), in a mixed effect model. In our collections, that corresponded to approximately a week difference in flowering time between samples with the smallest and largest sampled genome. Genome size explained 4.4% of the variation in flowering time in this model. Overall, sex explained the most variation in this model of flowering time (partial $r^2 = 20\%$, $F=41.83$, $p =1.05 \times 10^{-11}$) with males flowering 9.55 days [SE = 1.47 days] earlier than females. The second strongest predictor of flowering time was a sex by habitat interaction ($F = 13.52$, $p=0.0002$) reflecting greater sexual dimorphism in

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

flowering time in agricultural compared to natural habitats (**Fig 3B**). We also found a main effect of habitat (F=11.29, p=0.0007), longitude (F = 7.69, p=0.0056), and latitude (F = 5.482, p=0.0192) (Fig 3A; as described in [36]). In a model of growth rate, genome size is positively related to growth rate and the most significant predictor (F=4.57, p=0.034; **Fig 3C**), explaining 2.8% of the variation in growth rate, similar to the effect of sex (partial r^2 =2.7%, F= 4.42, p=0.037). We also find a marginal habitat by sex interaction effect (F=3.85, p = 0.052) (**Fig 3B**). 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 significant latitudinal and environmental variation, we do not find any significant correlation between genome size and environment (latitude: $\chi2$ = 2.6267, p = 0.1051; habitat: $\chi2$ = 0.0012, p= 0.9728) or sex ($\chi2$ =0.0004, p = 0.9843), when accounting for neutral population structure through a relatedness matrix (methods eq. 3). Total genome size has apparently not responded directly to selection through flowering time, but instead varies as a product of underlying repeat dynamics.

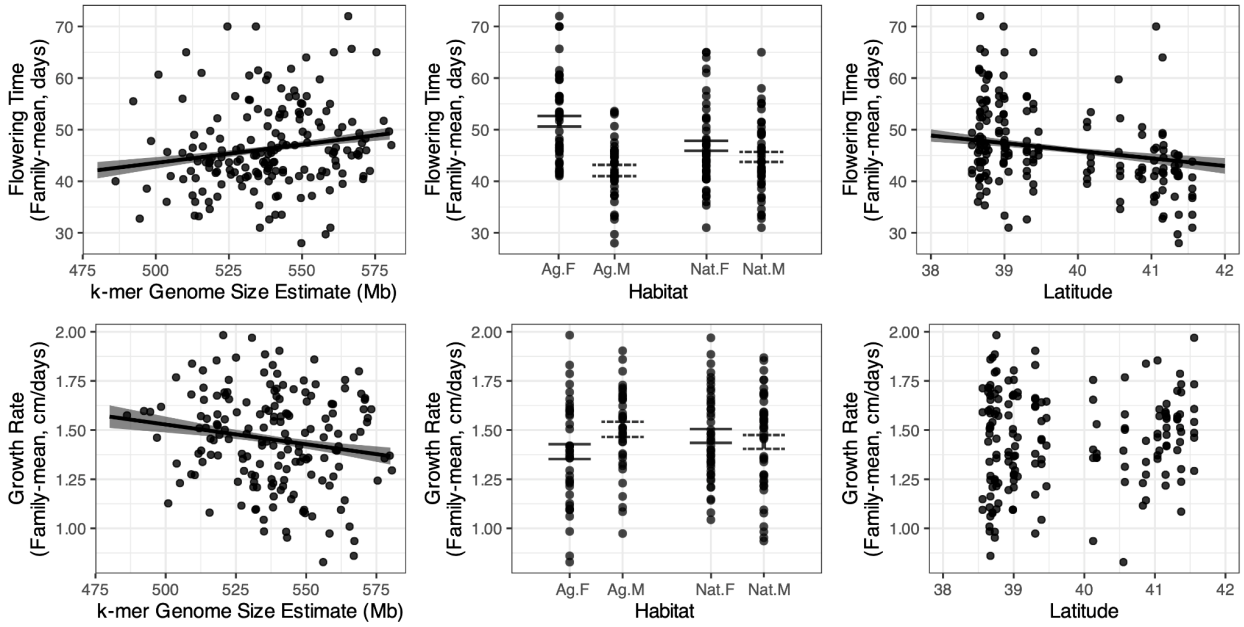


Fig 3. Genome size predicts flowering time (top) and growth rate (bottom) in *A. tuberculatus*, traits that also differ by habitat in a sex specific manner. Points show raw data, while regression lines and error bars depict least squares mean estimates from linear mixed modeling of flowering time and growth rate. Trend lines are shown for all significant relationships.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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 genic 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<0.05 level and 34/30771 genes at the Bonferroni p<0.05 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 (Bonferroni p-value=4.6 x 10^{-3}), GO molecular functions in NADH dehydrogenase (Bonferroni p-value = 6.1 x 10^{-6}) and proton-transporting ATP synthase activity (Bonferroni p-value = 6.9 x 10^{-4}), as well as numerous other related GO biological functions (**Sup Table 1)**. 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, ATP MI25, 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 **(Figure 4A)**.

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 96 loci in 73 genes across the genome passing a 10% FDR correction. We therefore calculated the polygenic value [48,49] for flowering time [$PGV_{FT} = 2\sum_{l=1}^{L} \alpha\, p$], where L equals the 96 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 individual PGV_{FT} with flowering time shows that the PGV_{FT} explains 44% of the variation in flowering time in the environmental conditions in which it was measured (**Figure 4B)**.

The positive correlation of genome size and flowering time as previously described could also reflect the greater opportunity for particular phenotype affecting insertion events 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. The associations across the genome greatly mirrored that of the gene level analysis (Sup Fig 7) with only one clear peak

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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.

Finally, we tested the relative importance and independence of these 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. This model explains 95% of the variation in family-mean flowering time, 24% of which is attributed to the fixed-effect terms and 71% of which can be attributed to the genome-wide relatedness. Genome size remains a significant explanatory variable (χ^2 =4.10, p = 0.043) explaining 2.3% of the variation in flowering time but is considerably less important compared to copy number variation at the ATP synthesis locus (partial r^2 = 12.8%, χ^2 =25.62, p = 4.2 x 10^{-7}), which in turn is less important than PGV_{FT} (partial r^2 = 44.9%, χ^2 =162.6, p < 10^{-15}) (**Fig 4C**).

In contrast to genome size, these polygenic and oligogenic architectures show patterns of spatial differentiation reflecting selection on flowering time. Copy number variation at the ATP synthesis locus demonstrates significant variability among sexes (χ^2= 22.59, p = 2.01 x 10^{-6}), across natural and agricultural habitats (χ^2= 9.27, p = 0.0023), and among sexes depending on habitat type (χ^2= 8.83, p = 0.0029), exceeding neutral expectations (methods eq. 3). Similarly, PGV_{FT} shows a strong sex effect (χ^2=9.13, p =0.0025), a habitat effect (χ^2=4.16, p = 0.041), and marginally, a longitude effect (χ^2=3.06, p = 0.08).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

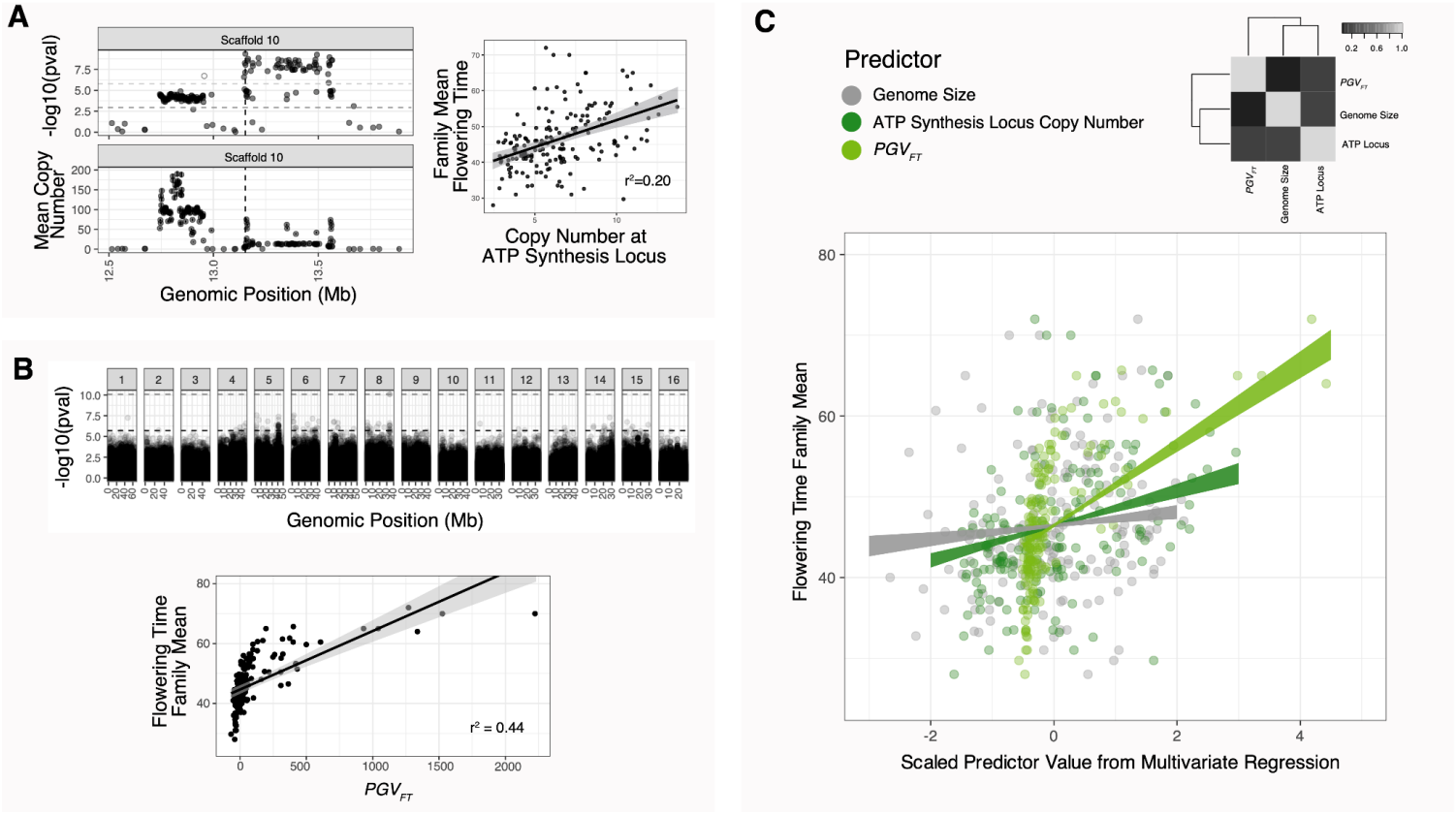


Figure 4. **The genetic architecture of flowering time in *A. tuberculatus* and the relative importance of associated genetic features. A)** The association of a copy number variant in the ATP synthesis pathway with flowering time (vertical line denoting locus with the most significant association genome wide). **B)** The polygenic value of individual flowering time (PGV_{FT}) based on 96 SNPs that pass a 10% FDR correction from a genome-wide association (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 of fixed effect predictors also illustrated in the top right of C).

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

genomes tend to show slower growth rates and time to flowering. 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. In comparison to newly identified polygenic variation and a photosynthesis-related large effect gene copy number variation, genome size remains a significant but 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 [2,34,50] to *Drosophila* [51]. 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 [32,52], cell production rate [2], stomatal density [32], flowering time [2], development time [53], growth form [54], and scaling laws [55]. 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 important 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. This suggests that genome size predominantly impacts variation for flowering time within populations in *A. tuberculatus*, although this may also simply result from the minor contribution of genome size to flowering time. In part, the 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 variation and differences across sexes, highlighting the numerous processes 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 origins to and gene flow [36,37,56]. Here we show the genetic architecture of flowering time appears to nearly as multifaceted. In addition to a 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 [57–59]. 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., [60–62]), providing some of the first evidence for such a link in natural populations. We also quantify polygenic variation for flowering time encoded by single nucleotide polymorphisms at 96 associated loci across the genome. These SNPs mapped to genes with molecular and biological functions varying from stigma and gynodiecium development (*HEC1*), meristem and flower development (*PCN, DOT2; [63]*), gibberlic acid mediated anther and seed development (*SPL8; [64]*), to DNA methylation and post transcriptional gene silencing (*AGO4,*

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

ROS1; the former protein family having been implicated in mediating flowering time by modifying the expression of *FT* [65]) (**Sup Table 2**). Unlike genome size, both gene copy number and polygenic variation for flowering time showed differentiation among habitats and sexes, suggesting that strong selection for high performing, earlier flowering males [66–69] in agricultural habitats has been key to shaping the distribution of this genetic variation across the landscape.

While genome size does not show substantial variation among populations, the composition of repeat content does, suggesting the possibility of stabilizing selection on genome size with repeats competing for limited space in the genome [70–72]. The processes governing non-neutral 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 [73]) 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 ($4N_er$). Recombination rate has been shown to vary with temperature [74–76] and thus latitude, implying that geographic co-evolution of repeat content and recombination rate [17] 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 [77] and a stronger selection for removal and/or an insertion preference away from genes in low recombination rate regions [17]. 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 [78], with demethylation having a potentially adaptive role at low temperatures [79,80], 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. We expect that linked selection played a key role in this finding. Recent work has shown that the *A. tuberculatus* chromosome that contains the male sex determining region harbours a fragment of the *Flowering Time* and *Heading date 3a* [81] 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, one possibility is that agricultural selection on a 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 by genome size—

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

436 which, in the absence of polyploidy events and large-scale heterochromatic knobs, may
437 predominantly reflect the balance between transposition and host removal for individual TE
438 families, with some role for stabilizing selection on total genome size. We show this balance
439 varies across numerous genomic, organismal, and environmental axes, and will thus require
440 diverse range-wide collections (e.g. [82,83]) along with experimental quantification of
441 transposition rates (e.g. [84]) and fitness effects to fully disentangle. Furthermore, quantification
442 of individual element frequencies using long-read population sequencing will enable the
443 characterization of element insertion frequencies, to assess the potential role of individual TE
444 insertions in rapid adaptation to agricultural environments.

445

446 Methods

447 We used the female reference genome as described in Kreiner et al., (2019). Briefly, the *A.*
448 *tuberculatus* reference genome was sequenced and assembled from an individual female plant
449 from an agricultural habitat. The resulting 2514 contigs were scaffolded onto a chromosome
450 resolved reference genome of a closely related species, *A. hypochondriacus [85]*, creating a
451 reference with 16 pseudochromosomes [56]. The accompanying gene annotation (described in
452 [56]) was also used.

453 For all analyses, we used the 182 samples which were previously sequenced and analyzed in
454 [36,37]. Briefly, each sample comes from intra-population crosses that were performed to control
455 for maternal phenotypic effects, with no two samples having the same maternal or paternal
456 genotype. The collections originated from 17 paired agricultural and natural populations, all of
457 which were located within 25 km of each other, and spanned three degrees of latitude and 12
458 degrees of longitude [36]. Out of 200 samples from unique maternal lines, 187 were successfully
459 sequenced with short-read Illumina sequencing depths ranging in coverage between 20-35 X
460 [36]. Due to a high error rate (>5%) in five of the samples were excluded from the study, leaving
461 182 samples for the estimation and analysis of repeat content. Phenotypic data for 176 of the
462 sequenced samples, that among other traits, included flowering time and vertical growth rate was
463 collected in a common garden experiment [36]. For each maternal line, 30 replicates were
464 grown, one sibling in each of the three treatments (water supplemented, control, and soy
465 competition), replicated across 10 experimental blocks in the common garden experiment [36].
466 For this study, we only used phenotypic data collected in the control treatment [36]. For each
467 maternal lineage, flowering times were averaged across the ten siblings in the different
468 experimental blocks. The proportions of var. *rudis* ancestry were estimated with the Faststructure
469 algorithm using SNP genetic information, also described in [36,86]. The relatedness matrix from
470 genome-wide SNPs across all individuals was computed in gemma [87] using the centered
471 genotype matrix algorithm (**Sup Fig 5)**. Genome size was estimated by first counting 21-mers
472 using the program KMC [88], and then by fitting a Bayesian model to the histogram of 21-mers

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

using Genomescope2.0 [43].

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 [89]. 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_retreiver [89–91]. Helitrons were detected by HelitronScanner-v1.1 which uses a two-layered local combinational variable (LCV) algorithm [92]. TIRs are detected by machine learning with the TIR-Learner2.5 program [93]. Following a number of filtering steps, the EDTA program reduces interlibrary redundancy between LTR-RTs, Helitrons and TIR elements, combines then 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 [94,95]. 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 I 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 [89].

Ribosomal RNA annotation

RNAmmer -1.2 was used to annotate eukaryotic rDNA [96]. RNAmmer 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

To annotate simple sequences, low complexity regions and tandem repeats, we ran RepeatMasker on the reference genome with default parameters [94].

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 [97] and scaled it by the median genic 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 [98], in combination with the plot_grid function in cowplot [99] 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 [100], 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' [101], using the commands --echo --count --bases-uniq-f. Population recombination rate was estimated with LDhat [102] in [56].

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. To test which repeat types contribute most to variation in genome size, we implemented a model using "lme4" version 1.1-27.1 with genome size as the dependent variable and all the major categories of repeat types as the independent variables. Since genome sizes were estimated using a K-mer analysis, we weighed the samples by their fit in the K-mer estimated model:

[eq.1] Genome Size = TEs + rDNA + Simple repeats + Low complexity + Unknown repeats + individual read level error rate + Residual Error, weights=K-mer model fit

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Percent of model variance explained by each term was calculated by comparing the r^2 of the full model to a model reduced by each term. Note that predictors in this model were kept broad (i.e. TEs instead of Helitrons + LTRs + TIRs etc.) as a result of strong collinearity between repeat classes within individuals.

To test whether flowering time and growth rate can be explained by variation in genome size and which repeat abundances are driving this variation, we used the package lme4qtl that implemented a mixed effect model with the relatedness matrix as a random effect [103]. We also tested the relative importance of genome size on flowering time by including it in a model with two other genetic predictors of flowering time (copy number of an ATP synthesis pathway locus and the polygenic value for flowering time; **Eq 3**):

[eq.2] Phenotypic trait = Genome Size + Long + Lat + Sex * Env + Ancestry + Residual Error + (1|Relatedness)

[eq.3] Phenotypic trait = Genome Size + ATP Locus Copy Number + PGV_{FT} + Residual Error + (1|Relatedness)

where the (1|) notation indicates a random effect after the bar operator. Percent of model variance explained by each term in the model was calculated with r2beta [104] and r2 for random versus fixed effects with the function r.squaredGLMM [105].

Similarly, 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.), we implemented a mixed effect model with lme4qtl that controls for the relatedness matrix (as above), and that included the fixed effect terms:

[**eq.4**] Repeat class = Long + Lat + Sex * Env + Ancestry + Residual Error + (1|Relatedness)

All models were evaluated with an anova using type III sums of squares.

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 [36]. To estimate copy number variation, we used mosdepth [97] 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. We

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

used this gene by individual level information as the input for a GWA of gene copy number variation with flowering time. We applied the same coverage-based approach from the gene copy number GWA to implement a TE and repeat copy number variation GWA with flowering time. 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 [36]. We did so using gemma [87] 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 [48,49] 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 [106–108].

Acknowledgements

Thanks to the Whitlock lab, Sally Otto, and Tyler Kent for feedback on the manuscript. JK was funded by a Killam Postdoctoral Fellowship and a Biodiversity Research Centre Bioinformatics Postdoc, S.I.W. was supported by an NSERC Discovery Grant and a Canada research chair. J.R.S. was supported by an NSERC Discovery Grant.

References

1. 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.

2. 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.

3. Mirsky AE, Ris H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. J Gen Physiol. 1951;34: 451–462.

4. Swift HH. The desoxyribose nucleic acid content of animal nuclei. Physiol Zool. 1950;23: 169–198.

5. Hahn MW, Wray GA. The g-value paradox. Evol Dev. 2002;4: 73–75.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

6. 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.

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

8. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genet Res . 1983;42: 1–27.

9. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9: 397–405.

10. 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.

11. Bonchev G, Parisod C. Transposable elements and microevolutionary changes in natural populations. Mol Ecol Resour. 2013;13: 765–775.

12. 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.

13. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. Biochim Biophys Acta Gene Regul Mech. 2017;1860: 157–165.

14. 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.

15. 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.

16. Lynch M, Conery JS. The origins of genome complexity. Science. 2003;302: 1401–1404.

17. Kent TV, Uzunović J, Wright SI. Coevolution between transposable elements and recombination. Philos Trans R Soc Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0458

18. 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.

19. Marais GAB, Gaillard J-M, Vieira C, Plotton 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.

20. 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.

21. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

649 22. Negi P, Rai AN, Suprasanna P. Moving through the Stressed Genome: Emerging Regulatory Roles
650 for Transposons in Plant Stress Response. Front Plant Sci. 2016;7: 1448.

651 23. Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D. Stress-Driven Transposable
652 Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. Mol Biol Evol.
653 2020;37: 221–239.

654 24. Wos G, Choudhury RR, Kolář F, Parisod C. Transcriptional activity of transposable elements along
655 an elevational gradient in Arabidopsis arenosa. Mob DNA. 2021;12: 7.

656 25. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol
657 Ecol. 2013;22: 1503–1517.

658 26. Dubin MJ, Mittelsten Scheid O, Becker C. Transposons: a blessing curse. Curr Opin Plant Biol.
659 2018;42: 23–29.

660 27. Blommaert J. Genome size evolution: towards new model systems for old questions. Proc Biol Sci.
661 2020;287: 20201441.

662 28. Mei W, Stetter MG, Gates DJ, Stitzer MC, Ross-Ibarra J. Adaptation in plant genomes: Bigger is
663 different. Am J Bot. 2018;105: 16–19.

664 29. Baetcke KP, Sparrow AH, Nauman CH, Schwemmer SS. The relationship of DNA content to
665 nuclear and chromosome volumes and to radiosensitivity (LD50). Proc Natl Acad Sci U S A.
666 1967;58: 533–540.

667 30. Cavalier-Smith T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell
668 growth rate, and the solution of the DNA C-value paradox. J Cell Sci. 1978;34: 247–278.

669 31. Beaulieu JM, Moles AT, Leitch IJ, Bennett MD, Dickie JB, Knight CA. Correlated evolution of
670 genome size and seed mass. New Phytol. 2007;173: 422–437.

671 32. Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA. Genome size is a strong predictor of cell
672 size and stomatal density in angiosperms. New Phytol. 2008;179: 975–986.

673 33. Gregory TR. Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value
674 enigma. Biol Rev Camb Philos Soc. 2007;76: 65–101.

675 34. Díez CM, Gaut BS, Meca E, Scheinvar E, Montes-Hernandez S, Eguiarte LE, et al. Genome size
676 variation in wild and cultivated maize along altitudinal gradients. New Phytol. 2013;199: 264–276.

677 35. Waselkov KE, Olsen KM. Population genetics and origin of the native North American agricultural
678 weed waterhemp (Amaranthus tuberculatus; Amaranthaceae). Am J Bot. 2014;101: 1726–1736.

679 36. Kreiner JM, Caballero A, Wright SI, Stinchcombe JR. Selective ancestral sorting and de novo
680 evolution in the agricultural invasion of Amaranthus tuberculatus. Evolution. 2021.
681 doi:10.1111/evo.14404

682 37. Kreiner JM, Latorre SM, Burbano HA, Stinchcombe JR, Otto SP, Weigel D, et al. Rapid weed
683 adaptation and range expansion in response to agriculture over the past two centuries. Science.
684 2022;378: 1079–1085.

685 38. Kreiner JM, Giacomini DA, Bemm F, Waithaka B, Regalado J, Lanz C, et al. Multiple modes of

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

686 convergent adaptation in the spread of glyphosate-resistant *Amaranthus tuberculatus*. Proc Natl
687 Acad Sci U S A. 2019;116: 21076–21084.

688 39. Galindo-González L, Mhiri C, Deyholos MK, Grandbastien M-A. LTR-retrotransposons in plants:
689 Engines of evolution. Gene. 2017;626: 14–25.

690 40. Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. The genomic ecosystem of transposable
691 elements in maize. PLoS Genet. 2021;17: e1009768.

692 41. Sáez-Vásquez J, Delseny M. Ribosome Biogenesis in Plants: From Functional 45S Ribosomal DNA
693 Organization to Ribosome Assembly Factors. Plant Cell. 2019;31: 1945–1967.

694 42. Kang M, Wang J, Huang H. Nitrogen limitation as a driver of genome size evolution in a group of
695 karst plants. Sci Rep. 2015;5: 11636.

696 43. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free
697 profiling of polyploid genomes. Nat Commun. 2020;11: 1432.

698 44. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic
699 genome content. Philos Trans R Soc Lond B Biol Sci. 2015;370: 20140331.

700 45. Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. Measuring Genome Sizes Using Read-
701 Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). G3 .
702 2020;10: 3047–3060.

703 46. Stetter MG, Schmid KJ. Analysis of phylogenetic relationships and genome size evolution of the
704 Amaranthus genus using GBS indicates the ancestors of an ancient crop. Mol Phylogenet Evol.
705 2017;109: 80–92.

706 47. Puixeu G, Pickup M, Field DL, Barrett SCH. Variation in sexual dimorphism in a wind-pollinated
707 plant: the influence of geographical context and life-cycle dynamics. New Phytol. 2019;224: 1108–
708 1120.

709 48. Fisher RA. The Genetical Theory of Natural Selection: A Complete Variorum Edition. OUP Oxford;
710 1930.

711 49. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10:
712 e1004412.

713 50. Poggio L, Rosato M, Chiavarino AM, Naranjo CA. Genome Size and Environmental Correlations in
714 Maize (Zea mays ssp. mays, Poaceae). Ann Bot. 1998;82: 107–115.

715 51. Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, et al. Population-specific
716 dynamics and selection patterns of transposable element insertions in European natural populations.
717 Mol Ecol. 2019;28: 1506–1522.

718 52. Gregory TR. The bigger the C-value, the larger the cell: genome size and red blood cell size in
719 vertebrates. Blood Cells Mol Dis. 2001;27: 830–843.

720 53. Stelzer C-P, Pichler M, Hatheuer A. Linking genome size variation to population phenotypic
721 variation within the rotifer, Brachionus asplanchnoidis. Commun Biol. 2021;4: 596.

722 54. Trávníček P, Čertner M, Ponert J, Chumová Z, Jersáková J, Suda J. Diversity in genome size and GC

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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.

55. 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.

56. 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/10.1101/498519v1.abstract

57. Monnahan PJ, Kelly JK. The Genomic Architecture of Flowering Time Varies Across Space and Time in Mimulus guttatus. Genetics. 2017;206: 1621–1635.

58. 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. doi:10.1101/790279

59. 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.

60. 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.

61. Simkin AJ, López-Calcagno PE, Raines CA. Feeding the world: improving photosynthetic efficiency for sustainable crop production. J Exp Bot. 2019;70: 1119–1140.

62. Simkin AJ, Lopez-Calcagno PE, Davey PA, Headland LR, Lawson T, Timm S, et al. Simultaneous stimulation of sedoheptulose 1,7-bisphosphatase, fructose 1,6-bisphophate aldolase and the photorespiratory glycine decarboxylase-H protein increases CO2 assimilation, vegetative biomass and seed yield in Arabidopsis. Plant Biotechnol J. 2017;15: 805–816.

63. 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.

64. 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.

65. 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. doi:10.1073/pnas.2113757118

66. Lawrence CW. Genetic studies on wild populations of Melandrium III. Heredity . 1964;19: 1–19.

67. Webb CJ. Flowering periods in the gynodioecious species Gingidia decipiens (Umbelliferae). N Z J Bot. 1976;14: 207–210.

68. Purrington CB, Schmitt J. Consequences of sexually dimorphic timing of emergence and flowering in Silene latifolia. J Ecol. 1998;86: 397–404.

69. Aljiboury AA, Friedman J. Mating and fitness consequences of variation in male allocation in a wind-pollinated plant. Evolution . 2022;76: 1762–1775.

70. Beaulieu JM. The right stuff: evidence for an “optimal” genome size in a wild grass population. The

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

New phytologist. 2010. pp. 883–885.

71. 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.

72. Hjelmen 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. doi:10.3390/genes11020218

73. Ågren JA. Selfish genetic elements and the gene's-eye view of evolution. Curr Zool. 2016;62: 659–665.

74. 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.

75. 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.

76. 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.

77. Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view. Chromosoma. 2014;123: 313–325.

78. 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.

79. 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.

80. 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.

81. 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.

82. 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. doi:10.1101/2023.05.24.542068

83. 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. doi:10.1038/s41586-023-06062-z

84. 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.

85. 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)

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

798 chromosomes provide insights into genome evolution. BMC Biol. 2017;15: 74.

799 86. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in
800 large SNP data sets. Genetics. 2014;197: 573–589.

801 87. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat
802 Genet. 2012;44: 821–824.

803 88. Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics.
804 Bioinformatics. 2017. pp. 2759–2761. doi:10.1093/bioinformatics/btx304

805 89. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable
806 element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol.
807 2019;20: 275.

808 90. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo
809 detection of LTR retrotransposons. BMC Bioinformatics. 2008;9: 18.

810 91. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long
811 Terminal Repeat Retrotransposons. Plant Physiol. 2018;176: 1410–1422.

812 92. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of
813 Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A. 2014;111: 10263–10268.

814 93. Su W, Gu X, Peterson T. TIR-Learner, a New Ensemble Method for TIR Transposable Element
815 Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome.
816 Mol Plant. 2019;12: 447–460.

817 94. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015. Available:
818 http://www.repeatmasker.org

819 95. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for
820 automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A.
821 2020;117: 9451–9457.

822 96. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent
823 and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35: 3100–3108.

824 97. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes.
825 Bioinformatics. 2018;34: 867–868.

826 98. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media;
827 2009.

828 99. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2 (2020). R package
829 version 1.1. 1. 2021.

830 100. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
831 Bioinformatics. 2010;26: 841–842.

832 101. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-
833 performance genomic feature operations. Bioinformatics. 2012;28: 1919–1920.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

834 102. McVean G, Auton A. LDhat 2.1: a package for the population genetic analysis of recombination.
835 Department of Statistics, Oxford, OX1 3TG, UK. 2007. Available:
836 https://pdfs.semanticscholar.org/37c0/71c2bcd115bcce4bb2e4eb89a7dbb861c136.pdf

837 103. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martinez-Perez A, Aschard H, Soria JM. lme4qtl:
838 linear mixed models with flexible covariance structure for genetic studies of related individuals.
839 BMC Bioinformatics. 2018;19: 68.

840 104. Jaeger B. r2glmm: Computes R squared for mixed (multilevel) models. R package version 01.

841 105. MuMIn BK. Multi-Model Inference. R Package Version 1. 43. 15. 2019. 2020.

842 106. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic
843 adaptation on height is overestimated due to uncorrected stratification in genome-wide association
844 studies. Elife. 2019;8. doi:10.7554/eLife.39702

845 107. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced
846 signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8. doi:10.7554/eLife.39725

847 108. Barton N, Hermisson J, Nordborg M. Why structure matters. eLife. 2019.
848 doi:10.7554/eLife.45380

849

850

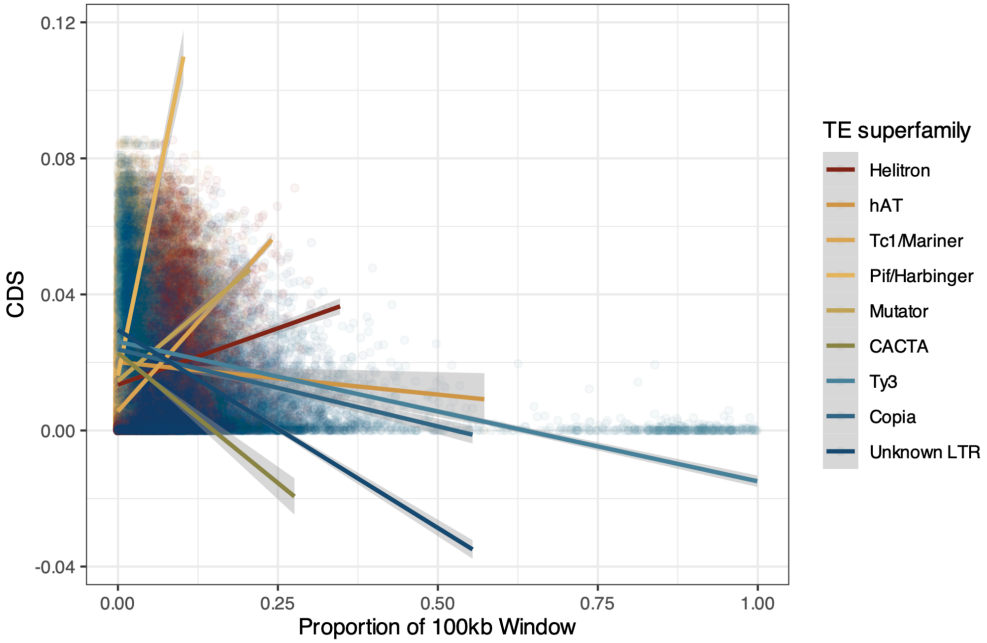
851

852

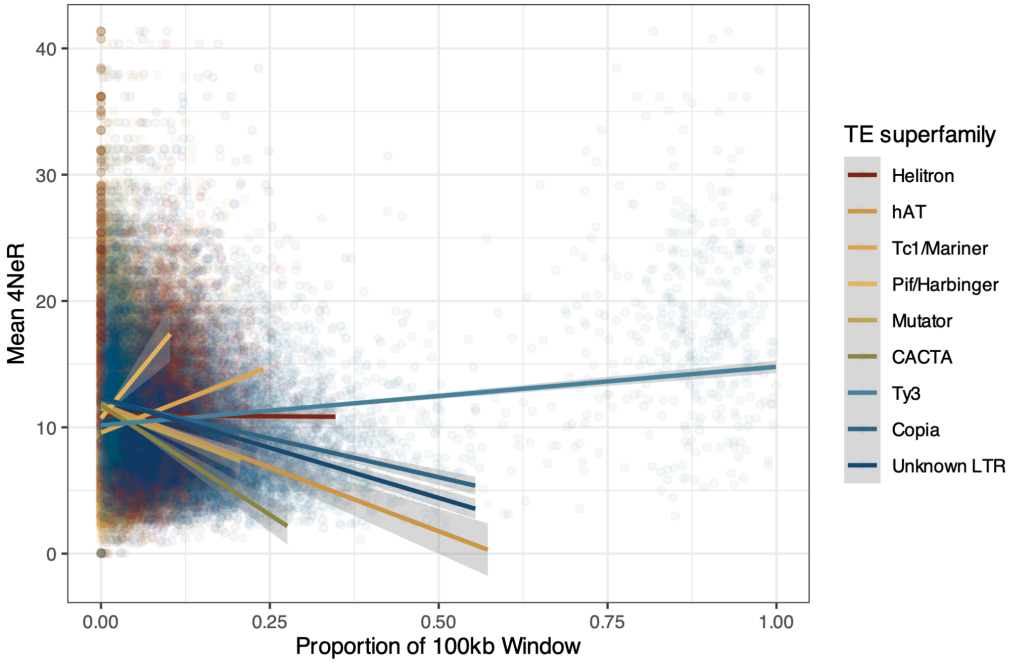
853

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Figures and Tables

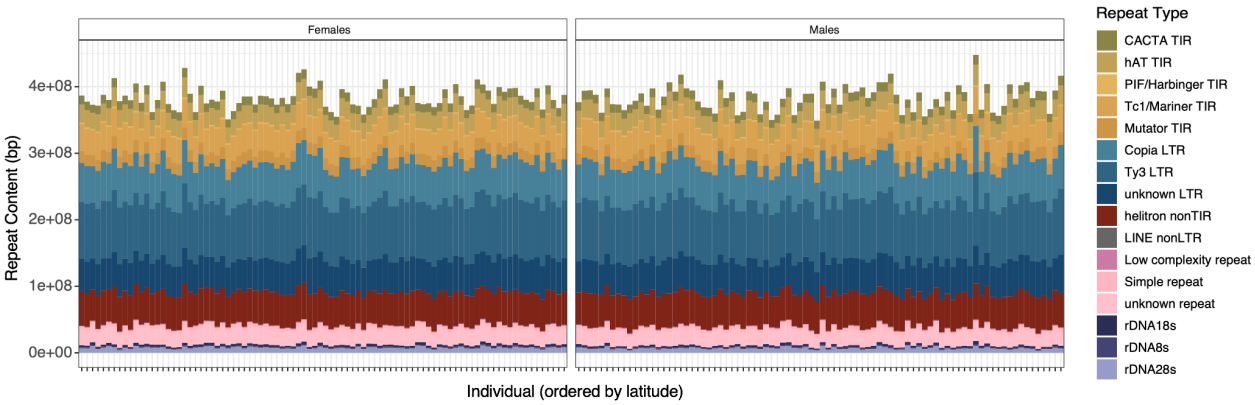


Sup Fig 1. Bivariate correlations of TE density across the genome and coding sequence density across the genome.

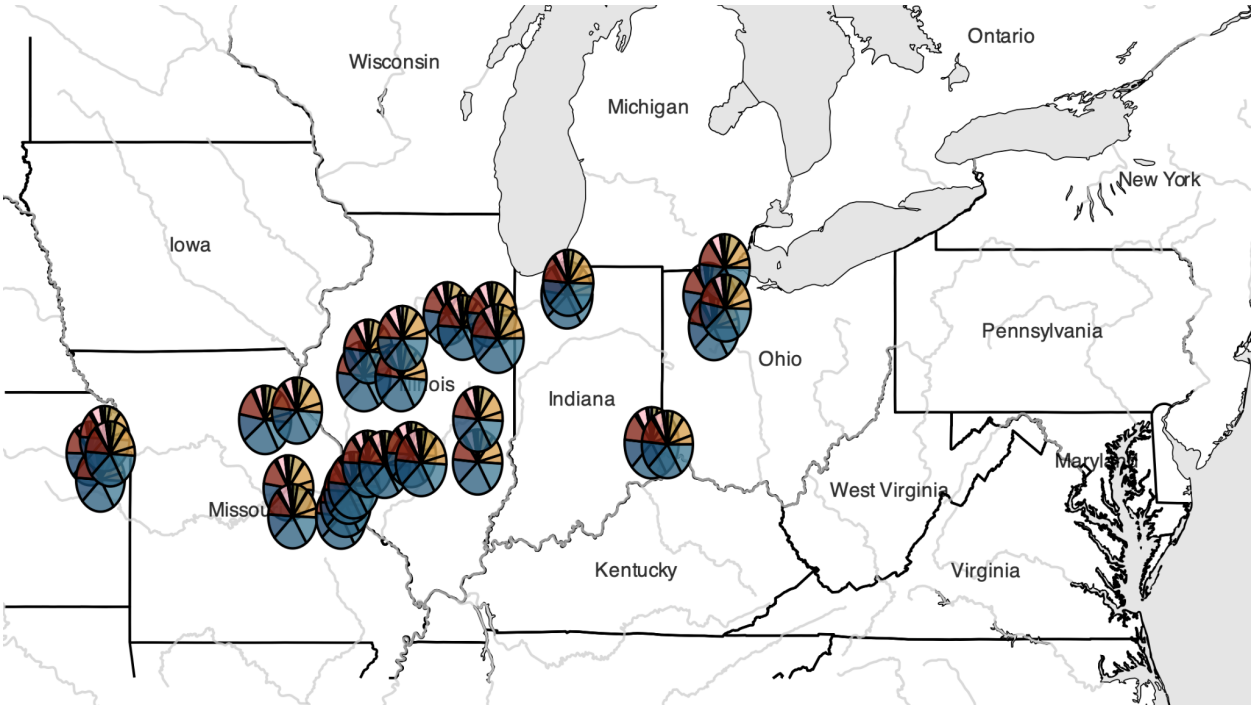


Sup Fig 2. Bivariate correlations of TE density across the genome and coding sequence density across the genome.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

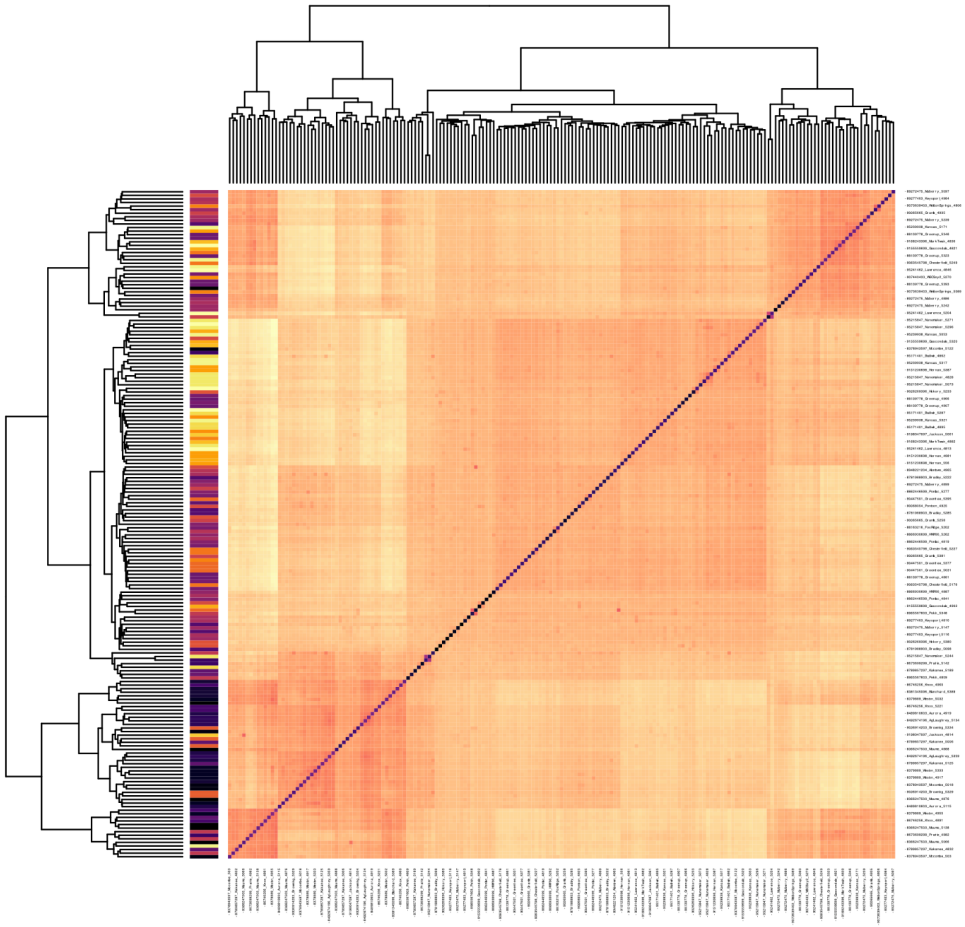


Sup Fig 3. Repeat composition within and across individuals, with individuals oriented from south to north from left to right.



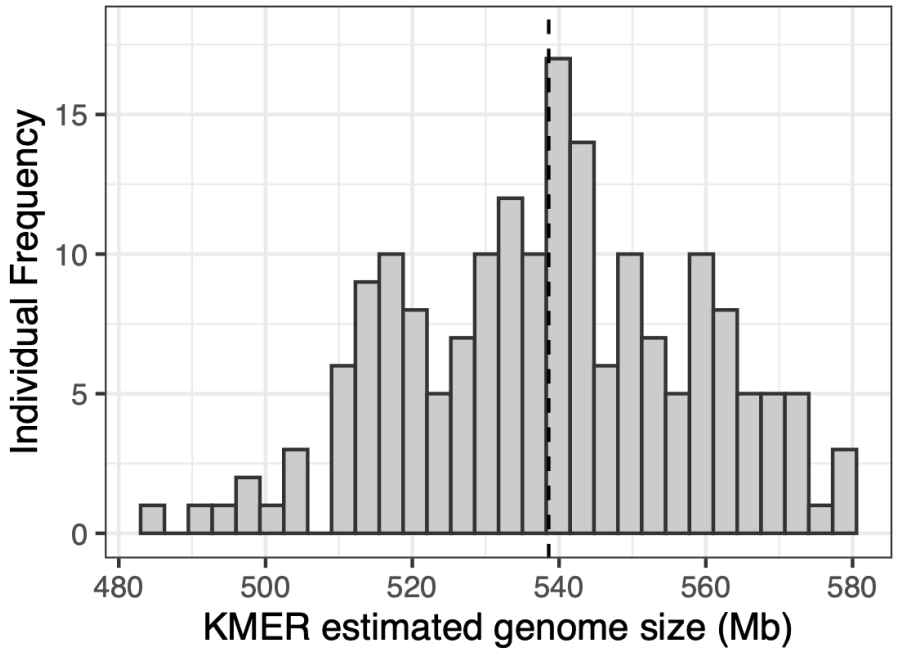
Sup Fig 4. Mean repeat composition across populations of *A. tuberculatus*

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



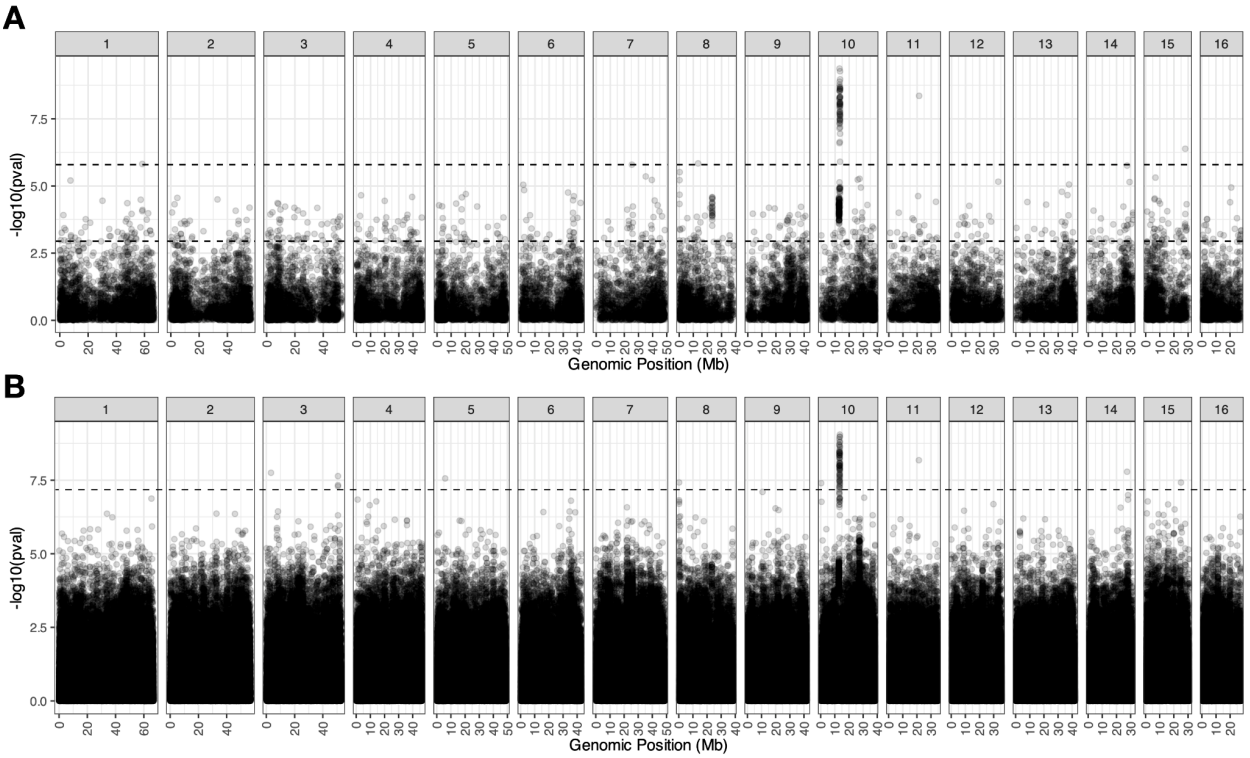
Sup Fig 5. The relatedness matrix based on centered genotypes as computed in plink. Colors on the left represent population groupings, ordered by longitude (with the most eastern populations in darker colours).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Sup Fig 6. The distribution of k-mer inferred genome sizes in this study, where the mean estimated genome size is indicated by the vertical dashed line.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Sup Fig 7. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 1. Biological GO Enrichment results after Bonferroni correction from the gene-level copy number GWAS.

GO biological process complete	#	expected	Fold Enrichment	bonferroni p-value
electron transport chain	15	3.51	4.27	0.0219
generation of precursor metabolites and energy	38	11.45	3.32	0.00000234
cellular metabolic process	371	270.23	1.37	1.55E-09
metabolic process	489	377.73	1.29	2.73E-11
cellular process	556	447.08	1.24	3.26E-11
cellular respiration	17	4.57	3.72	0.0291
energy derivation by oxidation of organic compounds	18	5.01	3.6	0.0249
photosynthesis	26	7.26	3.58	0.000286
translation	38	15.52	2.45	0.00383
macromolecule metabolic process	327	249.74	1.31	0.0000472
organic substance metabolic process	447	353.81	1.26	2.82E-07
organic substance biosynthetic process	182	131.88	1.38	0.0213
biosynthetic process	193	142.48	1.35	0.0321
peptide biosynthetic process	39	15.69	2.49	0.00203
peptide metabolic process	41	16.63	2.47	0.00124
organonitrogen compound metabolic process	243	186.24	1.3	0.0171
nitrogen compound metabolic process	342	265.11	1.29	0.0000788
amide metabolic process	47	20.81	2.26	0.002
amide biosynthetic process	42	17.71	2.37	0.0026
cellular nitrogen compound biosynthetic process	100	55.01	1.82	0.0000637
cellular nitrogen compound metabolic process	210	137.15	1.53	7.31E-07
cellular biosynthetic process	148	97.37	1.52	0.00138
organonitrogen compound biosynthetic process	90	52.52	1.71	0.00385
gene expression	88	43.97	2	0.00000652
primary metabolic process	378	292.63	1.29	0.00000377
nucleobase-containing compound biosynthetic process	41	19.09	2.15	0.0374
nucleobase-containing compound metabolic process	159	107.14	1.48	0.00174
heterocycle metabolic process	176	126.82	1.39	0.0241

Table 2. Significant SNPs from a GWA with flowering time, for those where their *A. tuberculatus* containing gene has an orthologous match in *Arabidopsis thaliana.*

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

946

Scaffold	**SNP pos**	**p-value**	***A. tuberculatus* gene**	***Arabidopsis thaliana orthologue***
Scaffold_1	47452652	0.03453878	AMATA_chromosomes_01864	AT2G27040.2 | Argonaute family protein
Scaffold_12	15625876	0.09085806	AMATA_chromosomes_24437	AT5G01360.1 | trichome birefringence-like protein _DUF828_
Scaffold_13	23324283	0.06911431	AMATA_chromosomes_18741	AT5G53370.1 | pectin methylesterase PCR fragment F
Scaffold_13	24404836	0.09844032	AMATA_chromosomes_18775	AT3G43720.1 | Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
Scaffold_14	20943076	0.06400042	AMATA_chromosomes_27334	AT3G01750.1 | Ankyrin repeat family protein
Scaffold_14	23436312	0.09085806	AMATA_chromosomes_27443	AT2G23820.1 | Metal-dependent phosphohydrolase
Scaffold_14	27350300	0.03453878	AMATA_chromosomes_27742	AT5G04490.1 | phytol kinase 1 VTE5
Scaffold_14	30780591	0.05856916	AMATA_chromosomes_28065	AT4G21440.1 | MYB-like 102
Scaffold_15	4576038	0.05001236	AMATA_chromosomes_28532	AT2G20360.1 | NAD_P_-binding Rossmann-fold superfamily protein
Scaffold_4	36359923	0.06856213	AMATA_chromosomes_13519	AT1G21840.1 | urease accessory protein F
Scaffold_4	43139171	0.06400042	AMATA_chromosomes_14013	AT3G14420.2 | Aldolase-type TIM barrel family protein
Scaffold_4	46422286	0.05856916	AMATA_chromosomes_14342	AT5G55480.1 | SHV3-like 1
Scaffold_5	2034930	0.03360651	AMATA_chromosomes_08437	AT4G07410.1 | Transducin family protein / WD-40 repeat family protein
Scaffold_5	2106017	0.05856916	AMATA_chromosomes_08443	AT1G10220.1 | ZCF37
Scaffold_5	5958416	0.04435092	AMATA_chromosomes_08761	AT1G35220.1 | FAM91 carboxy-terminus protein
Scaffold_5	12581943	0.07875264	AMATA_chromosomes_09007	AT1G21710.1 | 8-oxoguanine-DNA glycosylase 1
Scaffold_5	21471937	0.09085806	AMATA_chromosomes_09297	AT1G02065.2 | squamosa promoter binding protein-like 8
Scaffold_5	38378789	0.09554805	AMATA_chromosomes_09674	AT2G45770.1 | signal recognition particle receptor protein%2C chloroplast _FTSY_
Scaffold_5	41481985	0.04435092	AMATA_chromosomes_09826	AT4G09980.2 | Methyltransferase MT-A70 family protein
Scaffold_5	41548927	0.05856916	AMATA_chromosomes_09829	AT4G33760.1 | tRNA synthetase class II _D%2C K and N_ family protein
Scaffold_5	41887990	0.0640917	AMATA_chromosomes_09843	AT2G13560.1 | NAD-dependent malic enzyme 1
Scaffold_5	42268048	0.08479116	AMATA_chromosomes_09859	AT5G67060.1 | basic helix-loop-helix _bHLH_ DNA-binding superfamily protein
Scaffold_6	766284	0.05001236	AMATA_chromosomes_14401	AT5G55510.1 | Mitochondrial import inner membrane translocase subunit Tim17/Tim22/Tim23 family protein
Scaffold_6	2564440	0.07672023	AMATA_chromosomes_14502	AT1G79590.3 | syntaxin of plants 52
Scaffold_6	24370798	0.05856916	AMATA_chromosomes_15146	AT5G04850.1 | SNF7 family protein
Scaffold_6	31915973	0.09085806	AMATA_chromosomes_15418	AT1G01720.1 | NAC _No Apical Meristem_ domain transcriptional regulator superfamily protein
Scaffold_6	32038527	0.03810637	AMATA_chromosomes_15422	AT5G18310.1 | ubiquitin hydrolase
Scaffold_6	36230199	0.0835516	AMATA_chromosomes_15715	AT1G65880.1 | benzoyloxyglucosinolate 1
Scaffold_6	39036537	0.06870575	AMATA_chromosomes_15937	AT5G19630.1 | alpha/beta-Hydrolases superfamily protein
Scaffold_6	42027459	0.05001236	AMATA_chromosomes_16191	AT1G65230.1 | transmembrane protein%2C putative _DUF2358_
Scaffold_7	8282074	0.05001236	AMATA_chromosomes_10465	AT4G23740.2 | Leucine-rich repeat protein kinase family protein
Scaffold_7	9438776	0.05001236	AMATA_chromosomes_10493	AT4G33070.1 | Thiamine pyrophosphate dependent pyruvate decarboxylase family protein
Scaffold_8	2261279	0.0947496	AMATA_chromosomes_20085	AT5G03610.1 | GDSL-like Lipase/Acylhydrolase superfamily protein
Scaffold_8	2705630	0.05856916	AMATA_chromosomes_20122	AT2G36490.1 | demeter-like 1
Scaffold_8	5454320	0.06856213	AMATA_chromosomes_20370	AT5G43960.1 | Nuclear transport factor 2 _NTF2_ family protein with RNA binding _RRM-RBD-RNP motifs_ domain-containing protein
Scaffold_8	8852408	0.08401684	AMATA_chromosomes_20613	AT2G28690.1 | TOX high mobility group box protein%2C putative _DUF1635_
Scaffold_8	16102740	0.0640917	AMATA_chromosomes_21045	AT1G78915.3 | Tetratricopeptide repeat _TPR_-like superfamily protein
Scaffold_8	18341312	0.05856916	AMATA_chromosomes_21170	AT5G16780.2 | SART-1 family
Scaffold_8	34198010	0.05856916	AMATA_chromosomes_21584	AT5G22300.1 | nitrilase 4

947

948

949

950

951

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.13.548797; this version posted July 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

952
953