

1 **Short tandem repeats with massive variation and functional consequences**
2 **across strains of *Arabidopsis thaliana*.**

3
4 Maximilian O. Press¹, Rajiv C. McCoy¹, Ashley N. Hall², Joshua Akey¹, and Christine
5 Queitsch^{1*}

6
7 1: University of Washington Department of Genome Sciences, Seattle, WA

8 2: University of Washington Department of Molecular and Cellular Biology, Seattle, WA

9 *: to whom correspondence should be addressed: queitsch@uw.edu

10

11 Keywords: Short tandem repeat, microsatellite, quantitative genetics, selection,
12 *Arabidopsis thaliana*.

1 **Abstract**

2 Short tandem repeat (STR) mutations may be responsible for more than half of the
3 mutations in eukaryotic coding DNA, yet STR variation is rarely examined as a
4 contributor to complex traits. We assess the scope of this contribution across a
5 collection of 96 strains of *Arabidopsis thaliana* by massively parallel STR genotyping.
6 95% of examined STRs are polymorphic, and the median STR has six alleles. Modest
7 STR expansions are found in most strains, some of which have evident functional
8 effects. For instance, three of six intronic STR expansions are associated with intron
9 retention. We infer selective constraint on STRs, and find the strongest signatures of
10 purifying selection on coding STRs. Lastly, we detect dozens of novel STR-phenotype
11 associations that could not be detected with SNPs, and validate some experimentally.
12 Our results demonstrate that STRs comprise a large unascertained reservoir of
13 functionally relevant genomic variation.

14

15 **Introduction**

16 Mutation rates vary within a genome by several orders of magnitude¹, from $\sim 10^{-8}$ - 10^{-9} for
17 substitutions to 10^{-3} - 10^{-4} on average for short tandem repeats (STRs)²⁻⁴. STR mutations
18 occur through addition or subtraction of repeat units. Given the prevalence of STR loci
19 in eukaryotic genomes, we would expect more *de novo* STR mutations than single
20 nucleotide substitutions in the human genome per generation⁴, even in coding regions⁵.
21 Thus, while the overall mutation rate is under strong control from natural selection,
22 some loci experience many more mutations than others⁶. The existence of such highly
23 mutable loci contradicts the commonly-used infinite sites model⁷, which assumes that
24 no locus mutates more than once in a population; it further poses complications for
25 quantitative genetic models that assume contributions from many independent loci⁸.

26

27 In spite of the large effect that STRs can have on complex traits and diseases in model
28 organisms and humans⁹⁻¹¹, their variation is largely not considered in genotype-
29 phenotype association studies due to technical obstacles. However, STR genotyping
30 methods of sufficient accuracy, throughput, and cost-effectiveness to ascertain STR

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 alleles at high throughput have recently become available^{12–14}. Studies leveraging these
2 methods suggest considerable contributions of STRs to heritable phenotypic
3 variation^{12,15}.

4
5 From an evolutionary perspective, the high mutation rate and clear phenotypic effects of
6 STRs are speculated to provide readily accessible evolutionary paths for rapid
7 adaptation^{16–18}. On larger time scales, the presence of STRs with high mutation rates in
8 coding regions appears to be maintained by selection^{19–21}. These observations
9 independently argue for important functional roles of STRs.

10

11 In the present study, we apply massively parallel STR genotyping to a diverse panel of
12 well-characterized *A. thaliana* strains. We use these data to generate and test
13 hypotheses about the functional effects of STR variation, combining observations of
14 gene disruption by STR expansion, inferences about STR conservation, and phenotypic
15 association analyses with follow-up experiments. Based on our results, we argue that
16 STRs must be included in any comprehensive account of phenotypically relevant
17 genomic variation.

18

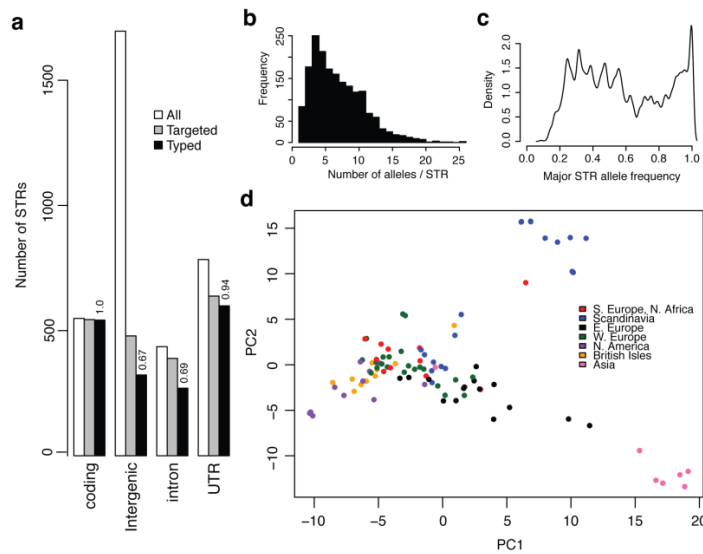
19 **Results**

20 *STR genotyping reveals complex allele frequency spectra.*

21 We targeted 2,050 STR loci for genotyping with molecular inversion probes (MIPs)¹²
22 across a core collection of 96 *A. thaliana* strains (Methods). These loci were all less
23 than 200bp in size and had nucleotide purity of at least 90%, encompassing nearly all
24 gene-associated STRs and ~40% of intergenic STRs (Fig. 1a). We used comparisons
25 with the Col-0 reference genome, PCR analysis of selected STRs, and dideoxy
26 sequencing to estimate that MIP STR genotype calls were ~95% accurate, and
27 inaccurate calls were generally only one to two units away from the correct copy
28 number (Supplementary Note, Supplementary Fig. 1-4, Supplementary Table 1).

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*



1
2 **Figure 1. STRs in *A. thaliana* show a complex allele frequency distribution and geographic**
3 **differentiation.** (a): Distribution and ascertainment of STR loci. “All”: all STRs matching the
4 definition of STRs for this study, *e.g.* ≤ 180 bp length in TAIR10, $\geq 89\%$ purity in TAIR10, 2-10
5 bp nucleotide motif. “Targeted”: the 2050 STRs targeted for MIP capture. “Typed”: STRs
6 successfully genotyped in the Col-0 genome in a MIPSTR assay. Numbers above bars indicate
7 the proportion of targeted STRs in the relevant category that were successfully genotyped. (b):
8 The distribution of allele counts across all genotyped STRs. (c): The distribution of major allele
9 frequencies (frequency of the most frequent allele at each locus) across genotyped STRs. (d):
10 Principal component analysis (PCA) reveals substantial geographic structure according to STR
11 variation. PC1 and PC2 correspond, respectively, to 5.2% and 4.0% of total STR allele variance.
12

13 95% of STRs were polymorphic. Most STRs were highly multiallelic (Fig. 1b; mean=6.4
14 alleles, median=6 alleles), and this variation was mostly unascertained by the 1001
15 Genomes resource for *A. thaliana* (Supplementary Fig. 3a). Coding STRs were only
16 slightly less polymorphic (mean=4.5 alleles, median=4 alleles), though whether this
17 difference is due to selection or to mutation rate variation is unclear. 45% of STRs had a
18 major allele with frequency < 0.5 , confusing the familiar concepts of major and minor
19 alleles, which have provided a common framework for detecting genotype associations
20 (Fig. 1c). Specifically, the Col-0 reference strain carries the major STR allele at only
21 48% of STR loci. Moreover, rarefaction analysis implied that more STR alleles at these
22 loci are expected with further sampling of *A. thaliana* strains (Supplementary Fig. 4c).

23 Principal component analysis of STR variation revealed genetic structure corresponding
24 to Eurasian geography (Fig. 1d, Supplementary Fig. 5), consistent with previous

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 observations that population structure is associated with geography in *A. thaliana*^{22,23}.

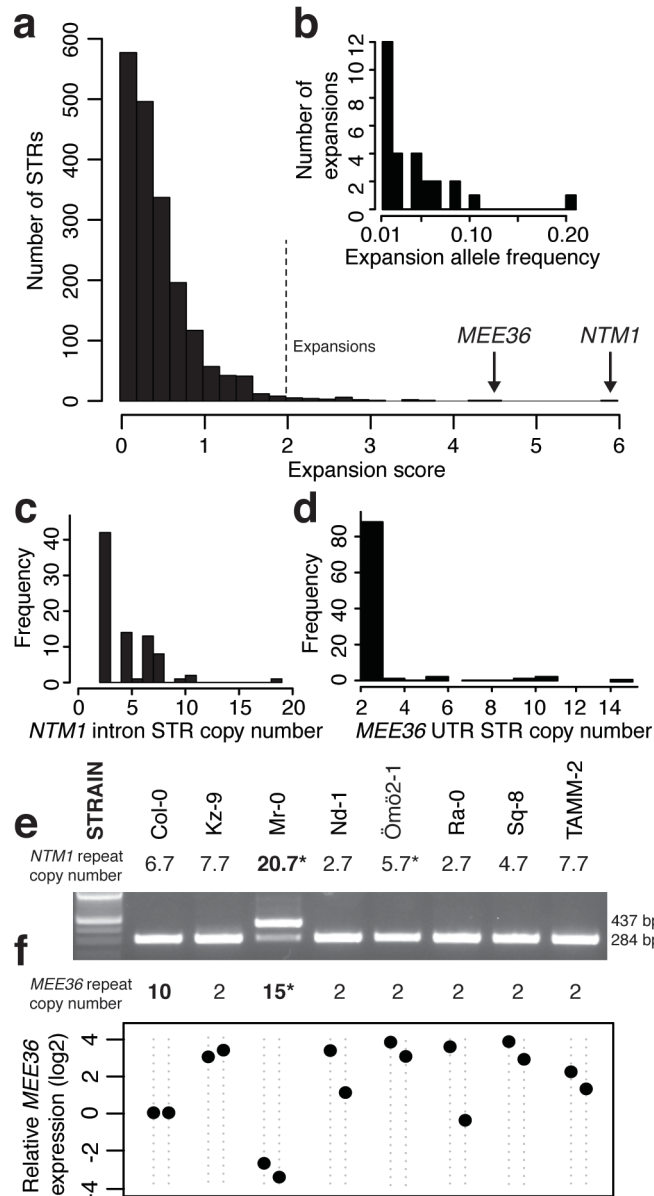
2 This result, which agrees with previous observations from a much larger set of genome-
3 wide single nucleotide polymorphism markers, shows that a comparatively small panel
4 of STRs suffices to capture detailed population structure in *A. thaliana*. Overall, we find
5 that STRs are extremely allele-rich, and their variation reflects expected population
6 structure.

7 *Novel STR expansions are associated with splice disruptions.*

8 We next examined the frequency and functional consequences of STR expansions in *A.*
9 *thaliana*. STR expansions, extreme high-copy-number variants of comparatively short
10 STRs, are widely recognized as contributing to human disease²⁴ and other
11 phenotypes²⁵. While large (>150 bp) expansions are difficult to infer, we detected
12 modest STR expansions with a simple heuristic comparing the longest allele observed
13 at each locus to the median allele (Fig. 2a). We found expansions in 64 of 96 *A. thaliana*
14 strains, each carrying at least one expanded STR allele from one of 28 expansion-prone
15 STRs (9 coding, 6 intronic, 8 UTR, 5 intergenic). Most expansions were found in
16 multiple strains (Fig. 2b), although expansion frequency was likely underestimated due
17 to a higher rate of missing data at these loci. STR expansions causing human disease
18 can be as small as 25 copies²⁴, whereas we detected expansions up to ~50 copies.

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*



1
2 **Figure 2. Inferring and assessing the functional effects of modest STR expansions.** (a): The
3 distribution of expansion scores across STRs, where the expansion score is computed as
4 $(\max(\text{STR length}) - \text{median}(\text{STR length})) / \text{median STR length}$. We called any STRs with a score
5 greater than 2 a modest expansion (indicated). (b): Distribution of allele frequencies of the 28
6 expanded STR alleles. (c, d): Distribution of STR copy number of the intronic STR in the *NTM1*
7 gene and the 3' UTR STR in the *MEE36* gene. (e): RT-PCR demonstrates intron retention in
8 *NTM1* mRNA in the Mr-0 strain, which carries the STR expansion, yielding an aberrant 437-bp
9 product. (f): *MEE36* transcript abundances measured by qRT-PCR and normalized relative to
10 *UBC21* transcript levels. For each strain, two independent biological replicates are shown as
11 points. Transcript levels are expressed relative to Col-0 levels (set to 1). *: STR genotype
12 corrected by dideoxy sequencing. Strains and order are the same between (e) and (f).

13 We assayed the effects of STR expansions on expression of associated genes. The

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 most dramatic expansions (with large relative copy number increase) affected an
2 intronic STR in the *NTM1* gene (Fig. 2c; five other expansions also resided in introns)
3 and a STR in the 3' UTR of the *MEE36* gene (Fig. 2d). These genes, respectively, have
4 roles in cell proliferation and embryonic development. Intronic STR mutations can
5 disrupt splicing, causing altered gene function²⁶ and human disease²⁷. Therefore, we
6 assayed effects on splicing of all six expanded intronic STRs. In three cases, the
7 expanded allele was associated with partial or full retention of its intron (Supplementary
8 Fig. 6, Supplementary Note), including the major *NTM1* splice form in the Mr-0 strain
9 (Fig. 2e). We confirmed intron retentions by dideoxy sequencing of cDNA
10 (Supplementary Fig. 7a). Retention of the *NTM1* intron is predicted to lead to a
11 nonsense mutation truncating most of the NTM1 protein (Supplementary Fig. 7b). For
12 the other two retentions, more complex and STR allele-specific mRNA species were
13 formed (Supplementary Fig. 6, Supplemental Text). The *MEE36* STR expansion alleles
14 were associated with dramatically reduced *MEE36* transcript levels (Fig. 2f), possibly
15 due to the STR expansion altering transcript processing²⁸. These examples emphasize
16 the potential for previously unascertained STR variation to modify gene function.
17 Moreover, considering STR allele frequency distributions, as we did here by focusing on
18 outliers, enables predictions about STR functional effects.

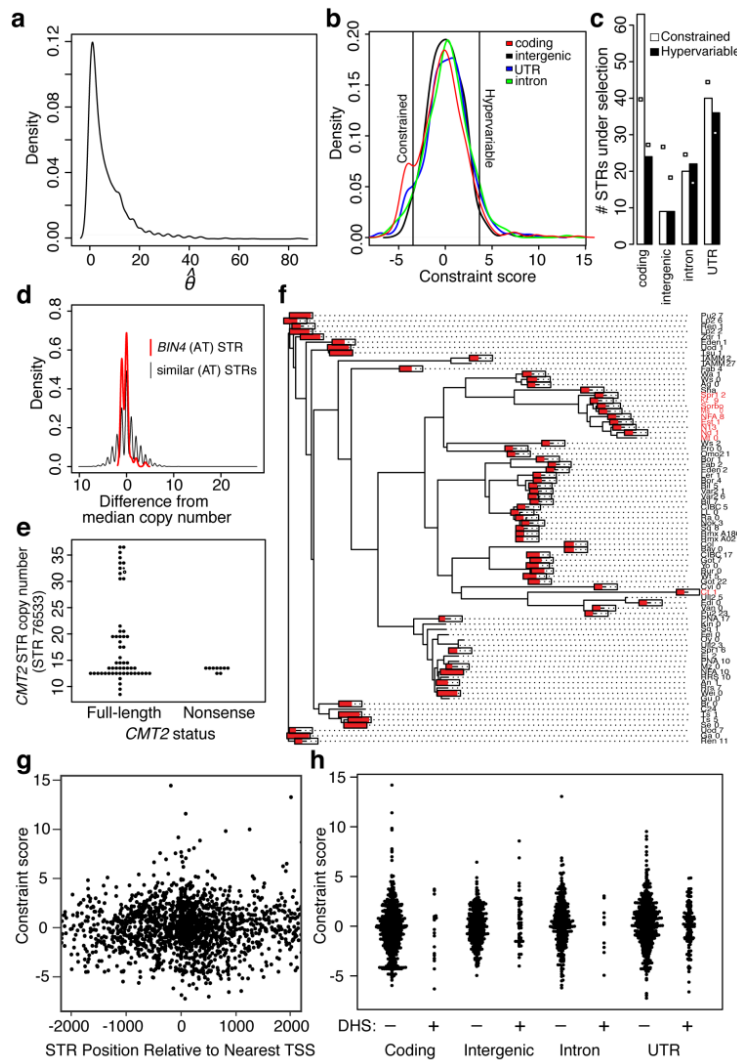
19 *Signatures of functional constraint on STR variation.*

20 Using our observed STR allele frequency distributions, we next attempted to infer
21 selective processes acting on STRs. Previous models for evaluating functional
22 constraint on STRs are few²⁹, though there is consensus that selection affects STR
23 variation^{17,19,29,30}. Naively, we would expect that coding STRs should show increased
24 constraint (lower variation). Consistent with this expectation, we observed that most
25 invariant STRs are coding (53 of 84 invariant STRs genotyped across at least 70
26 strains; odds ratio = 4.5, $p = 5 \times 10^{-11}$, Fisher's Exact Test). However, methods of inferring
27 selection by allele counting are confounded by population structure and by mutation
28 rate, which varies widely across STRs in this (Fig. 3a) and other studies^{3,31}.

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 Therefore, to account for mutation rate and population structure, we used support vector
 2 regression (SVR) to predict STR variability across these 96 strains, using well-
 3 established correlates of STR variability (*e.g.* STR unit number and STR purity; Online
 4 Methods)^{3,32,33}. Selection was defined as deviation from expected variation of a neutral
 5 STR among strains. We trained SVRs on the set of intergenic STRs, which should
 6 experience minimal selection relative to STRs associated with genes (Supplementary
 7 Fig. 8-10). We used bootstrap aggregation of SVR models to compute a putative
 8 constraint score for each STR by comparing its observed variability to the expected
 9 distribution from bootstrapped SVR models (Fig. 3b, Supplementary Note,
 10 Supplementary File 1).



11
 12 **Figure 3. Detecting functionally constrained STRs.** (a): The distribution of $\hat{\theta}$ (estimated

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 mutation rate⁷) across all genotyped STR loci. (b): Distribution of “selection scores” across all
2 STRs, separated by locus category. Vertical lines indicate 2.5% and 97.5% quantiles of the
3 distribution of intergenic STRs, which are used as thresholds for putative constraint and
4 hypervariability respectively. (c): Constrained or hypervariable STRs separated by locus
5 category. White boxes indicate the expected numbers for each bar, based on number of STRs in
6 each locus category and number of STRs under different types of selection. (d): *BIN4* intron STR
7 is constrained relative to similar STRs. Allele frequency spectra are normalized by subtracting
8 the median copy number (9 for the *BIN4* STR). All pure STRs with TA/AT motifs and a median
9 copy number between 7 and 12 are included in the “similar STRs” distribution. (e): Lack of
10 association between near-expansion *CMT2* STR alleles and previously described nonsense
11 mutations. (f): Neighbor-joining tree of a 10KB region of *A. thaliana* chromosome 4
12 encompassing the *CMT2* gene across 81 strains with available data. Tip labels in red carry an
13 adaptive nonsense mutation early in the first exon of *CMT2*, as noted previously⁴¹. Red bars
14 drawn on tips of the tree indicate the length of the *CMT2* intronic STR (as a proportion of its
15 maximum length, 36.5 units). The bars are omitted for tips with missing STR data. (g):
16 Constraint score from (b) plotted with respect to nearest TSS. (h): Constraint score from (b)
17 plotted with respect STR annotations.

18 According to constraint scores, 132 STRs were less variable than expected, suggesting
19 purifying selection on these loci (Fig. 3c). Among these, 63 STRs were coding (OR =
20 2.4, $p = 3.7 \times 10^{-6}$, Fisher’s Exact Test); this enrichment of coding STRs among
21 constrained STRs agrees with our naïve analysis of invariant STRs above. Examples of
22 constrained coding STRs included STRs encoding homologous polylysines in three
23 different histone H2B proteins. Generally, coding STRs showing purifying selection
24 encoded roughly half as many polyserines and twice as many acidic homopolymers as
25 expected from the *A. thaliana* proteome (Supplementary Table 3)³⁴. Although many
26 more coding STRs are probably functionally constrained, our power to detect such
27 constraints is limited by the size of the dataset. We also observed high conservation of
28 some STRs in non-coding regions, though this is less interpretable given the ambiguous
29 relationship between sequence conservation and regulatory function in *A. thaliana*³⁵.
30 The most constrained intronic STR, in the *BIN4* gene, which is required for
31 endoreduplication and normal development³⁶, shows a restricted allele frequency
32 spectrum compared to similar STRs (Fig. 3d).

33 Hypervariable coding STRs (showing more alleles than expected) were too few for
34 statistical arguments, but nonetheless showed several notable patterns (Supplementary

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 Note, Supplementary Fig. 11). For example, 3/24 hypervariable coding STRs encoded
2 polyserines in F-box proteins, suggesting that STR variation may serve as a mechanism
3 of diversification in this protein family, which shows dramatically increased family size
4 and sequence divergence in some plant lineages^{37,38}. One non-coding hypervariable
5 STR was associated with the *Chromomethylase 2* (*CMT2*) gene, which is under
6 selection in *A. thaliana*³⁹. Specifically, *CMT2* nonsense mutations in some populations
7 are associated with temperature seasonality. We considered whether the extreme
8 *CMT2* STR alleles might be associated with these nonsense mutations. Instead, these
9 extreme alleles exclusively occurred in strains with full-length *CMT2* (Fig. 3e). Strains
10 with the common *CMT2* nonsense mutation form a tight clade in the *CMT2* sequence
11 tree, whereas the *CMT2* STR length fluctuates rapidly throughout the tree and appears
12 to converge on longer alleles independently in different clades (Fig. 3f). These
13 convergent changes are consistent with a model in which the *CMT2* STR is a target of
14 selection.

15 We further assessed whether STR conservation can be attributed to *cis*-regulatory
16 function. The abundance of STRs in eukaryotic promoters⁴⁰, and their associations with
17 gene expression (in *cis*)^{15,41}, have suggested that STRs affect transcription, possibly by
18 altering nucleosome positioning⁴¹. We examined whether STRs near transcription start
19 sites (TSSs) showed signatures of functional constraint (Fig. 3g). We found little
20 evidence for reduced STR variation near TSSs, suggesting that *cis*-regulatory effects do
21 not generally constrain STR variation. Moreover, we found no relationship between
22 putative constraint and whether STRs reside in accessible chromatin sites marking
23 regulatory DNA (Fig. 3h).

24 *STRs yield numerous novel genotype-phenotype associations.*

25 We next addressed whether STR genotypes contribute new information for explaining
26 phenotypic variation. It is often assumed that STR phenotype associations should be
27 captured through linkage to common single nucleotide polymorphisms (SNPs)⁴². This
28 assumption is inconsistent with available data. Human STR data^{42,43} and simulation

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

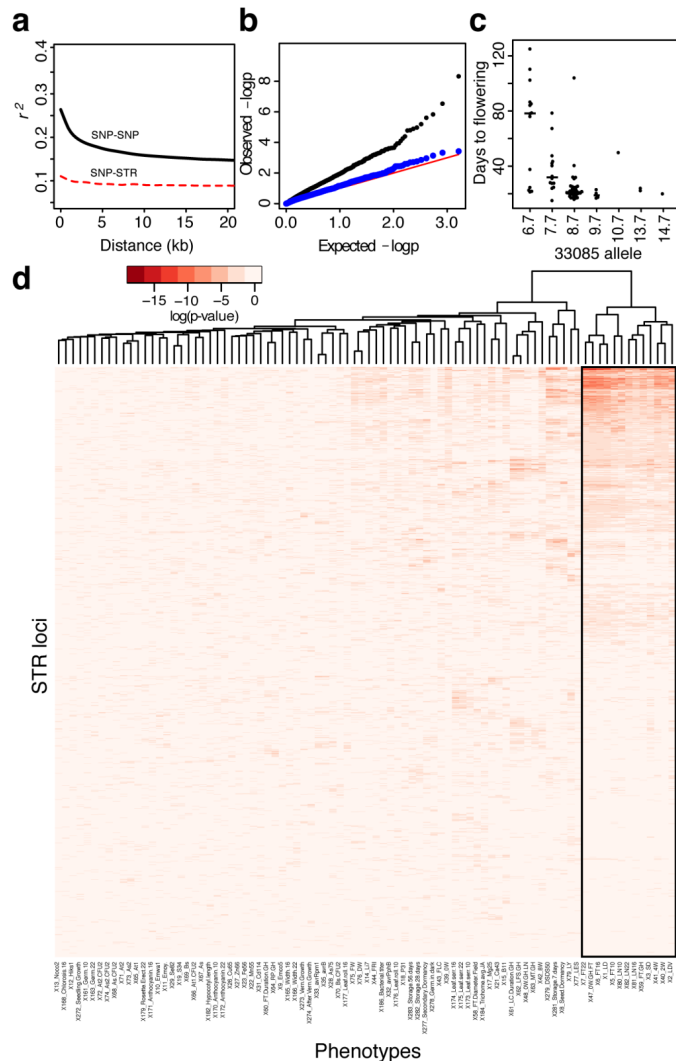
1 studies⁴⁴ indicate decreased linkage disequilibrium (LD) between STRs and SNPs,
2 making it uncertain whether SNP markers can meaningfully tag STR genotypes. Indeed,
3 we found little evidence of LD between STR and SNPs in *A. thaliana* (Fig. 4a),
4 Moreover, the observed LD around STR loci declined with increasing STR allele
5 number (Supplementary Fig. 12), consistent with an expected higher mutation rate at
6 multiallelic loci^{7,44}. This result suggests that STR-phenotype associations need to be
7 directly tested rather than relying on linkage to SNPs. *A. thaliana* offers extensive high-
8 quality phenotype data for our inbred strains, which have been previously used for SNP-
9 based association studies⁴⁵. We tested each polymorphic STR for associations across
10 the 96 strains with each of 105 published phenotypes. For the subset of 32 strains for
11 which RNA-seq data were available⁴⁶, we also tested for associations between STR
12 genotypes and expression of genes within 1MB (*i.e.*, eQTLs; Supplementary Note).
13 Although our power to detect eQTLs was limited by sample size, we detected 12
14 associations. The strongest association was between a STR residing in long noncoding
15 RNA gene *AT4G07030* and expression of the nearby stress-responsive gene *AtCPL1*
16 (Supplementary Fig 13, Supplementary Table 4).

17 We next focused on organismal phenotypes. Certain STRs showed associations with
18 multiple phenotypes, and flowering time phenotypes were particularly correlated with
19 one another (Fig. 4c,d). Similar to these patterns, SNPs have also shown associations
20 with multiple phenotypes, and correlated flowering time phenotypes are among the
21 strongest associations in the same strains⁴⁵. As in previous association studies using
22 STRs¹⁵, some inflation was apparent in test p-values compared to expectations, though
23 the same tests using permuted STR genotypes showed negligible inflation (Fig. 4b,
24 Supplementary Fig 14). Negligible inflation with permuted genotypes has been used
25 previously to exclude confounding from population structure¹⁵, which we will also
26 presume here (Supplementary Note). We found 137 associations between 64 STRs and
27 25 phenotypes at stringent genome-wide significance levels (Methods; Supplementary
28 Table 5). Given the low LD observed between STRs and other variants, STRs are likely
29 to be the causal variants, rather than merely tagging them. Our analysis found plausible

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

- 1 candidate genes, such as *COL9*, which acts in flowering time pathways and contains a
- 2 flowering time-associated STR, and *RABA4B*, which acts in the salicylic acid defense
- 3 response⁴⁷ and contains a STR associated with lesion formation.



- 4
- 5 **Figure 4. Diverse associations of STRs with quantitative phenotypes.** (a): Multi-allelic LD⁶⁴
- 6 estimates for STR and SNP loci. Lowess lines for each category are plotted. All values of $r^2 < 0.05$
- 7 are omitted from lowess calculation for visualization purposes. (b): Quantile-quantile plot of p-
- 8 values from tests of association between STRs and germination rate after 28 days of storage.
- 9 (c): An example association between an STR (33085) and a phenotype (flowering time in long
- 10 days after 4 weeks vernalization) in *A. thaliana* strains. Median of each distribution is indicated
- 11 by a bar proportional in width to the number of observations. (d): Heatmap showing pairwise
- 12 associations between STRs and phenotypes, summarized by the p-value from a linear mixed
- 13 model, fitting STR allele as a fixed effect and kinship as a random effect. Both rows and columns
- 14 are clustered, though the row dendrogram was omitted for clarity. STRs with genotype

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 information in fewer than 25 strains are not displayed. Flowering time phenotypes are boxed in
2 black.

3

4 We evaluated whether these associations might have been found using SNP-based
5 analyses, and whether STR effects are large enough to be meaningful. We found that
6 STR effects on phenotype are largely not accounted for by nearby SNP variation.

7 Considering the strongest association for each STR, only 18 of the 64 STRs were near
8 potentially confounding SNP variants, and most associations (14/18) were robust to
9 adjustment for nearby SNP genotypes⁴⁵ (Supplementary Table 6, Supplementary Note).

10 One notable exception was an STR closely linked to a well-known deletion of the *RPS5*
11 gene in a hypervariable region of chromosome 1⁴⁸ that causes resistance to bacterial
12 infection⁴⁹. *RPS5* status is under balancing selection in *A. thaliana*⁴⁸. In this case, the
13 association and the linkage are apparently strong enough (the STR is ~4KB upstream of
14 the deletion) that this STR tags *RPS5*'s effect on infection. The observation of a STR
15 tagging a hypervariable region leads us to speculate that STR variation holds
16 information about genomic regions with complex mutational histories.

17 To assess the STR contribution to the variance of a specific trait, we performed a naïve
18 variance decomposition of the long-day flowering phenotype into SNP and STR
19 components, as represented by the loci showing associations with this trait. Our results
20 suggested that STRs potentially contribute as much or more variance than SNPs to this
21 phenotype (Supplementary Note, Supplementary Table 7). Estimated effect sizes for
22 STR variants on this phenotype were similar to those of large-effect SNP variants⁴⁵
23 (Supplementary Fig 15).

24 Finally, we used mutant analysis to evaluate the two strongest flowering time
25 associations, a coding STR in *AGL65* and an intronic STR in the uncharacterized gene
26 *AT4G01390*; neither locus had been associated with flowering time phenotypes. We
27 found that disruptions of both STR-associated genes had modest early flowering
28 phenotypes (by ~2 days and ~1 rosette leaf, $p < 0.05$ for each in linear mixed models;
29 Supplementary Fig 16), supporting the robustness of our STR-phenotype associations.

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 Taken together, our study suggests that STRs contribute substantially to phenotypic
2 variation.

3 **Discussion**

4 Our results imply that STRs contribute substantially to trait heritability in *A. thaliana*. Far
5 from being “junk DNA”, STRs are apparently constrained by functional requirements,
6 STR variation can disrupt gene function, and is associated with phenotypic variation.
7 Considering that STR variation is represented only poorly through linkage to nearby
8 SNPs, STRs likely contribute to the missing heritability of complex traits.

9 STRs are particularly relevant to the phenotypic variance due to *de novo* mutations
10 (V_m). Estimates of V_m from model organisms are on the order of 1%, but may vary
11 substantially from trait to trait⁵⁰. STRs are good candidates for a substantial proportion
12 of this variance, given their high mutation rate, residence in functional regions, and their
13 functional constraint observed here. In previous work⁵¹, we showed apparent copy
14 number conservation of a STR in spite of a high mutation rate. In this case, deviation
15 from the conserved copy number produced aberrant phenotypes. Our observation that
16 constrained STRs are common suggests that STRs are a likely source of deleterious *de*
17 *novo* mutations which are removed by selection.

18 The extent to which STRs affect phenotype is only partially captured in this study.
19 Specifically, we assayed two STRs shown to cause phenotypic variation in prior
20 transgenic *A. thaliana* studies^{25,52}, but these STR loci did not show strong signatures of
21 phenotypic association or of selection. This lack of ascertainment suggests that many
22 more functionally important STRs exist in *A. thaliana* than we can detect with the
23 analyses presented here. For example, the polyQ-encoding STR in the *ELF3* gene
24 causes dramatic variation in developmental phenotypes⁵², yet we find no statistical
25 associations between this locus and phenotype across our 96 strains. In this case the
26 lack of phenotype association is expected; *ELF3* STR alleles interact epistatically with
27 several loci⁵³ and associations are thus difficult to detect. Indeed, we have argued that
28 STRs are more likely than less mutable classes of genomic variation to exhibit

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 epistasis⁹. In consequence, we expect that the associations described in the present
2 study are an underestimate of STR effects on phenotype. Moreover, our data are
3 constrained by MIP technology, which limits the size and composition of STR alleles
4 that we can ascertain (Figure 1a, Supplementary Figures 1-2).

5 Considering next a mechanistic perspective, the association we observe between
6 intronic STR expansions and splice disruptions may be an important mechanism by
7 which STRs contribute to phenotypic variation. In humans, unascertained diversity of
8 splice forms contributes substantially to disease⁵⁴, and this diversity is larger than
9 commonly appreciated⁵⁵. We demonstrate that this mechanism is common at least for
10 expansions; future work should evaluate how tolerant introns are to different
11 magnitudes of STR variation, as these effects on protein function may prove to be both
12 large and of relatively high frequency. The phenotypic contributions of loci with high
13 mutation rate remain underappreciated, specifically in cases where such loci are difficult
14 to ascertain with high-throughput sequencing. The results presented here argue that
15 STRs are likely to play a substantial role in phenotypic variation and heritability.
16 Accounting for the heterogeneity of different classes of genomic variation, and
17 specifically variation in mutation rate, will advance our understanding of the genotype-
18 phenotype map and the trajectory of molecular evolution.

19 **METHODS**

20 **ONLINE METHODS**

21 **Probe design**

22 We used TRF⁵⁶ (parameters: matching weight 2, mismatching penalty 5, indel penalty 5,
23 match probability 0.8, indel probability 0.1, score ≥ 40 and maximum period 10) to
24 identify STRs in the TAIR8 build of the *Arabidopsis thaliana* genome, identifying 7826
25 putative STR loci under 200bp (Supplementary File 2). We restricted further analysis to
26 the 2409 loci with repeat purity $\geq 89\%$. We chose 2307 STRs from among these,
27 prioritizing STRs in coding regions, introns, or untranslated regions (UTRs), higher STR
28 unit purity, and expected variability (VARscore³²). We designed⁵⁷ molecular inversion

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 probes (MIPs) targeting these STR loci in 180bp capture regions with 8bp degenerate
2 tags in the common MIP backbone. For this purpose, we converted STR coordinates
3 to the TAIR10 build and used the TAIR10 build as a reference genome. We used single
4 nucleotide variants (SNVs) in 10 diverse *Arabidopsis thaliana* strains⁵⁸ to avoid
5 polymorphic sites in designing MIP targeting arms. We filtered out MIPs predicted to
6 behave poorly (MIPGEN logistic score < 0.7), discarded MIPs targeting duplicate
7 regions, and substituted MIPs designed around SNVs as appropriate. We attempted to
8 re-design filtered MIPs with 200bp capture regions using otherwise identical criteria.
9 This yielded a final set of 2050 STR-targeting MIP probes (Supplementary File 3).

10

11 **MIP and library preparation**

12 These 2050 probes were ordered from Integrated DNA Technologies as desalted DNAs
13 at the 0.2 picomole scale and resuspended in Tris-EDTA pH 8.0 (TE) to a concentration
14 of 2 μ M and stored at 4° C. We pooled and diluted probes to a final stock
15 concentration of 1 nM. We phosphorylated probes as described previously⁵⁹. We
16 performed DNA preparation from whole aerial tissue of adult *A. thaliana* plants. We
17 prepared MIP libraries essentially as described previously^{12,59} using 100 ng *A. thaliana*
18 genomic DNA for each of 96 *A. thaliana* strains.

19

20 **Sequencing**

21 We sequenced pooled capture libraries essentially as previously described¹² on
22 NextSeq and MiSeq instruments collecting a 250bp forward read sequencing the
23 ligation arm and captured target sequence, an 8bp index read for library
24 demultiplexing, and a 50bp reverse read sequencing the extension arm and
25 degenerate tag for single-molecule deconvolution. In each run, 10% of the sequenced
26 library pool consisted of high-complexity whole-genome library to increase sequence
27 complexity. For statistics and further details of data acquired for each library see
28 Supplementary Tables 8 and 9.

29

30 **STR annotation**

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 We annotated STRs according to Araport11⁶⁰, classifying all STRs as coding, intronic,
2 intergenic, or UTR-localized, and indicating whether each STR overlapped with
3 transposable element sequence. To identify regulatory DNA, we used the union of
4 seven distinct DNaseI-seq experiments⁶¹ covering pooled or isolated tissue types. For
5 additional details, see the Supplementary Note.

6

7 **Sequence analysis**

8 Sequences were demultiplexed and output into FASTQ format using BCL2FASTQ
9 v2.17 (Illumina, San Diego). We performed genotype calling essentially as described
10 previously¹², with certain modifications (Supplementary Note, Supplementary Table 1).
11 Note that our *A. thaliana* strains are inbred, and more stringent filters and data
12 processing would be necessary to account for heterozygosity. For information about
13 comparison with the Bur-0 genome, see the Supplementary Note. Updated scripts
14 implementing the MIPSTR analysis pipeline used in this study are available at
15 https://osf.io/mv2at/?view_only=d51e180ac6324d2c92028b2bad1aef67.

16

17 **Statistical analysis and data processing**

18 We performed all statistical analysis and data exploration using R v3.2.1⁶². For plant
19 experiments, we fit mixed-effects models using Gaussian (flowering) or binomial
20 generalized linear models using experiment and position as random effects and
21 genotype as a fixed effect.

22

23 **STR expansion inference**

24 We inferred STR expansions where the maximum copy number of an STR is at least
25 three times larger than the median copy number of that STR. Various alleles of STR
26 expansions were inspected manually in BAM files. Selected cases were dideoxy-
27 sequenced and analyzed as described.

28

29 **Plant material and growth conditions**

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 Plants were grown on Sunshine soil #4 under long days (16h light : 8h dark) at 22°C
2 under cool-white fluorescent light. T-DNA insertion mutants were obtained from
3 ABRC⁶³ (Supplementary Table 11). For flowering time experiments, plants were grown
4 in 36-pot or 72-pot flats; days to flowering (DTF) and rosette leaf number at flowering
5 (RLN) were recorded when inflorescences were 1cm high. Results are combined
6 across at least three experiments.

7

8 **Gene expression and splicing analysis**

9 We grew bulk seedlings of indicated strains on soil for 10 days, harvested at Zeitgeber
10 time 12 (ZT12), froze samples immediately in liquid nitrogen, and stored samples at -
11 80°C until further processing. We extracted RNA from plant tissue using the SV RNA
12 Isolation kit (including DNase step; Promega, Madison, WI), and subsequently treated it
13 with a second DNase treatment using the Turbo DNA-free kit (Ambion, Carlsbad, CA).
14 We performed cDNA synthesis on ~500ng RNA for each sample with oligo-dT adaptors
15 using the RevertAid kit (ThermoFisher, Carlsbad, CA). We performed PCR analysis of
16 cDNA with indicated primers (Supplementary Table 10) and ~25 ng cDNA with the
17 following protocol: denaturation at 95° 5 minutes, then 30 cycles of 95° 30 seconds,
18 55° 30 seconds, 72° 90 seconds, ending with a final extension step for 5 minutes at
19 72°. We gel-purified and sequenced electrophoretically distinguishable splice variants
20 associated with STR expansions. Each RT-PCR experiment was performed at least
21 twice with different biological replicates.

22

23 **Population genetic analyses**

24 For PCA, STRs with missing data across the 96 strains were omitted, leaving 987 STRs
25 with allele calls for every strain. $\hat{\theta}$ was estimated using the approximation $\hat{\theta} = \frac{1}{8\bar{X}^2} -$
26 0.5 ,⁷ where \bar{X} is the average frequency of all STR alleles at a locus. We computed
27 multiallelic linkage disequilibrium estimates for SNP-SNP and SNP-STR locus pairs
28 using MCLD⁶⁴. We downloaded array SNP data for the same lines (TAIR9 coordinates)
29 from [http://bergelson.uchicago.edu/wp-](http://bergelson.uchicago.edu/wp-content/uploads/2015/04/call_method_75.tar.gz)
30 [content/uploads/2015/04/call_method_75.tar.gz](http://bergelson.uchicago.edu/wp-content/uploads/2015/04/call_method_75.tar.gz)⁶⁵. For each locus, both SNP and STR,

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 we computed linkage disequilibrium scores with 150 surrounding loci. To facilitate
2 comparison, we computed lowess estimates of linkage only for those locus pairs in the
3 plotted distance window in each case, and only for locus pairs with $r^2 > 0.05$.

4

5 **Inference of conservation**

6 STRs typed across 70 *A. thaliana* strains or fewer were dropped from this analysis, as
7 the estimates of their variability were unlikely to be accurate, leaving 1825 STRs. We
8 measured STR variation as the base-10 logarithm of the standard deviation of STR
9 copy number (Supplementary Note). We used bootstrap aggregation (“bagging”) to
10 describe a distribution of predictions as follows. An ensemble of 1000 support vector
11 regression (SVR, fit using the *ksvm()* function in the kernlab package⁶⁶) models was
12 used to predict expected neutral variation of each STR as quantified by each measure
13 (Supplementary Note). We used this distribution of bootstrapped predictions for
14 intergenic STRs to compute putative conservation scores (Z-scores) for each STR.
15 Scores below the 2.5% ($Z < -3.46$) and above the 97.5% ($Z > 3.65$) quantiles of
16 intergenic STRs were considered to be putatively constrained and hypervariable
17 respectively.

18

19 **eQTL inference**

20 We downloaded normalized transcriptome data for *A. thaliana* strains from NCBI GEO
21 GSE80744⁴⁶. We used the Matrix eQTL package⁶⁷ to detect associations, fitting also 10
22 principal components from SNP genotypes to correct for population structure.
23 Following precedent¹⁵, we fitted additive models assuming that STR effects on
24 expression would be a function of STR copy number. We used Kruskal-Wallis rank-
25 sum tests to test the null hypothesis of no association following correction.

26

27 **Genotype-phenotype associations**

28 We downloaded phenotype data from [https://github.com/Gregor-Mendel-](https://github.com/Gregor-Mendel-Institute/atpolydb/blob/master/miscellaneous_data/phenotype_published_raw.tsv)
29 [Institute/atpolydb/blob/master/miscellaneous_data/phenotype_published_raw.tsv](https://github.com/Gregor-Mendel-Institute/atpolydb/blob/master/miscellaneous_data/phenotype_published_raw.tsv). We
30 followed precedent⁴⁵ in log-transforming certain phenotypes. In all analyses we treated

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

1 STRs as factorial variables (to avoid linearity assumptions) in a linear mixed-effect
2 model analysis to fit STR allele effects on phenotype as fixed effects while modeling
3 the identity-by-state kinship matrix between strains (computed from SNP data) as a
4 correlation structure for strain random effects on phenotype. We performed this
5 modeling using the *lmeKin()* function from the *coxme* R package⁶⁸. We repeated every
6 analysis using permuted STR genotypes as a negative control to evaluate p-value
7 inflation, and discarded traits showing such inflation. We used $p < 10^{-6}$ as a genome-
8 wide significance threshold commensurate to the size of the *A. thaliana* genome and
9 the data at hand. For flowering time phenotypes we used a more stringent $p < 10^{-10}$
10 threshold, as these phenotypes showed somewhat shifted p-value distributions (which
11 were nonetheless inconsistent with inflation, according to negative controls). We
12 identified potentially confounding SNP associations using the
13 <https://gwas.gmi.oeaw.ac.at/#/study/1/phenotypes> resource, using the criterion that a
14 SNP association must have a $p \leq 10^{-4}$ and be within roughly 100KB of the STR to be
15 considered. We fit models including SNPs as fixed effects as before, and performed
16 model selection using AICc⁶⁹. Additional details about association analyses are in the
17 Supplementary Note.

18

19 **Data and code availability**

20 MIP sequencing data are available in FASTQ format at the Sequence Read Archive,
21 under project number PRJNA388228. Analysis scripts are provided at
22 https://osf.io/5jm2c/?view_only=324129c85b3448a8bd6086263345c7b0, along with
23 data sufficient to reproduce analyses.

24

25 **ACKNOWLEDGMENTS**

26 We thank Keisha Carlson, Alberto Rivera, and members of the Queitsch lab for
27 technical assistance and important conversations. We thank Evan Boyle, Choli Lee,
28 Matthew Snyder, and Jay Shendure for assistance and advice concerning MIP design
29 and use. We thank the Dunham Lab and the Fields Lab for access to and help with
30 sequencing instruments. We thank UW Genome Sciences Information Technology for

Press *et al.* (2017)

Massive variation of STRs in *A. thaliana*

- 1 high-performance computing resources. This work was supported in part by NIH New
- 2 Innovator Award DP2OD008371 to CQ.

1 **References**

- 2 1. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role
3 of de novo mutations in health and disease. *Genome Biol.* **17**, 241 (2016).
- 4 2. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat.*
5 *Genet.* **44**, 1161–5 (2012).
- 6 3. Gymrek, M., Willems, T., Erlich, Y. & Reich, D. E. A framework to interpret short tandem
7 repeat variation in humans. *bioRxiv* 92734 (2016). doi:10.1101/092734
- 8 4. Willems, T. *et al.* Population-Scale Sequencing Data Enable Precise Estimates of Y-STR
9 Mutation Rates. *Am. J. Hum. Genet.* **98**, 919–933 (2016).
- 10 5. Payseur, B. A., Jing, P. & Haasl, R. J. A Genomic Portrait of Human Microsatellite Variation.
11 *Mol. Biol. Evol.* **28**, 303–312 (2011).
- 12 6. Harpak, A., Bhaskar, A. & Pritchard, J. K. Mutation Rate Variation is a Primary Determinant
13 of the Distribution of Allele Frequencies in Humans. *PLOS Genet.* **12**, e1006489 (2016).
- 14 7. Haasl, R. J. & Payseur, B. A. The Number of Alleles at a Microsatellite Defines the Allele
15 Frequency Spectrum and Facilitates Fast Accurate Estimation of θ . *Mol. Biol. Evol.* **27**, 2702–
16 2715 (2010).
- 17 8. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height.
18 *Nat. Genet.* **42**, 565–9 (2010).
- 19 9. Press, M. O., Carlson, K. D. & Queitsch, C. The overdue promise of short tandem repeat
20 variation for heritability. *Trends Genet.* **30**, 504–512 (2014).
- 21 10. Fondon, J. W., Hammock, E. A. D., Hannan, A. J. & King, D. G. Simple sequence
22 repeats: genetic modulators of brain function and behavior. *Trends Neurosci.* **31**, 328–34
23 (2008).

Press *et al.* (2017) Massive variation of STRs in *A. thaliana*

- 1 11. Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and
2 candidates for ‘missing heritability’. *Trends Genet.* **26**, 59–65 (2010).
- 3 12. Carlson, K. D. *et al.* MIPSTR: A method for multiplex genotyping of germline and
4 somatic STR variation across many individuals. *Genome Res.* **125**, 750–761 (2015).
- 5 13. Willems, T., Zielinski, D., Gordon, A., Gymrek, M. & Erlich, Y. Genome-wide profiling
6 of heritable and de novo STR variations. *bioRxiv* 77727 (2016). doi:10.1101/077727
- 7 14. Highnam, G. *et al.* Accurate human microsatellite genotypes from high-throughput
8 resequencing data using informed error profiles. *Nucleic Acids Res.* **41**, e32 (2013).
- 9 15. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression
10 variation in humans. *Nat. Genet.* **48**, 22–29 (2015).
- 11 16. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly
12 mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
- 13 17. King, D. G. Indirect selection of implicit mutation protocols. *Ann. N. Y. Acad. Sci.* **1267**,
14 45–52 (2012).
- 15 18. Gemayel, R., Vincens, M. D., Legendre, M. & Verstrepen, K. J. Variable tandem repeats
16 accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–77
17 (2010).
- 18 19. Mularoni, L., Ledda, A., Toll-Riera, M. & Albà, M. M. Natural selection drives the
19 accumulation of amino acid tandem repeats in human proteins. *Genome Res.* **20**, 745–54
20 (2010).
- 21 20. Yu, F. *et al.* Positive selection of a pre-expansion CAG repeat of the human SCA2 gene.
22 *PLoS Genet.* **1**, e41 (2005).

Press *et al.* (2017) Massive variation of STRs in *A. thaliana*

- 1 21. Sawaya, S. M., Lennon, D., Buschiazzo, E., Gemmell, N. & Minin, V. N. Measuring
2 microsatellite conservation in mammalian evolution with a phylogenetic birth-death model.
3 *Genome Biol. Evol.* **4**, 636–47 (2012).
- 4 22. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**,
5 e196 (2005).
- 6 23. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in
7 *Arabidopsis thaliana*. *Cell* **0**, (2016).
- 8 24. Usdin, K. The biological effects of simple tandem repeats: Lessons from the repeat
9 expansion diseases. *Genome Res.* **18**, 1011–1019 (2008).
- 10 25. Sureshkumar, S. *et al.* A genetic defect caused by a triplet repeat expansion in
11 *Arabidopsis thaliana*. *Science* **323**, 1060–3 (2009).
- 12 26. Li, Y.-C., Korol, A. B., Fahima, T. & Nevo, E. Microsatellites Within Genes: Structure,
13 Function, and Evolution. *Mol. Biol. Evol.* **21**, 991–1007 (2004).
- 14 27. Ranum, L. P. W. & Day, J. W. Dominantly inherited, non-coding microsatellite
15 expansion disorders. *Curr. Opin. Genet. Dev.* **12**, 266–271 (2002).
- 16 28. Jackson, R. J. Cytoplasmic regulation of mRNA function: The importance of the 3'
17 untranslated region. *Cell* **74**, 9–14 (1993).
- 18 29. Haasl, R. J. & Payseur, B. A. Microsatellites as targets of natural selection. *Mol. Biol.*
19 *Evol.* **30**, 285–98 (2013).
- 20 30. Huntley, M. A. & Clark, A. G. Evolutionary Analysis of Amino Acid Repeats across the
21 Genomes of 12 *Drosophila* Species. *Mol. Biol. Evol.* **24**, 2598–2609 (2007).

Press *et al.* (2017) Massive variation of STRs in *A. thaliana*

- 1 31. Schlötterer, C., Kauer, M. & Dieringer, D. Allele Excess at Neutrally Evolving
2 Microsatellites and the Implications for Tests of Neutrality. *Proc. Biol. Sci.* **271**, 869–874
3 (2004).
- 4 32. Legendre, M., Pochet, N., Pak, T. & Verstrepen, K. J. Sequence-based estimation of
5 minisatellite and microsatellite repeat variability. *Genome Res.* **17**, 1787–96 (2007).
- 6 33. Eckert, K. A. & Hile, S. E. Every microsatellite is different: Intrinsic DNA features
7 dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinog.*
8 **48**, 379–88 (2009).
- 9 34. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. & Gentles, A. J. Amino acid runs in
10 eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 333–8
11 (2002).
- 12 35. Alexandre, C. M. *et al.* Regulatory DNA in *A. thaliana* can tolerate high levels of
13 sequence divergence. *bioRxiv* 104323 (2017). doi:10.1101/104323
- 14 36. Breuer, C. *et al.* BIN4, a Novel Component of the Plant DNA Topoisomerase VI
15 Complex, Is Required for Endoreduplication in Arabidopsis. *Plant Cell* **19**, 3655–3668
16 (2007).
- 17 37. Xu, G., Ma, H., Nei, M. & Kong, H. Evolution of F-box genes in plants: Different modes
18 of sequence divergence and their relationships with functional diversification. *Proc. Natl.*
19 *Acad. Sci.* **106**, 835–840 (2009).
- 20 38. Clark, R. M. *et al.* Common Sequence Polymorphisms Shaping Genetic Diversity in
21 Arabidopsis thaliana. *Science* **317**, 338–342 (2007).
- 22 39. Shen, X. *et al.* Natural CMT2 Variation Is Associated With Genome-Wide Methylation
23 Changes and Temperature Seasonality. *PLOS Genet.* **10**, e1004842 (2014).

Press *et al.* (2017) Massive variation of STRs in *A. thaliana*

- 1 40. Sawaya, S. *et al.* Microsatellite tandem repeats are abundant in human promoters and are
2 associated with regulatory elements. *PloS One* **8**, e54710 (2013).
- 3 41. Vences, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable
4 tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–6 (2009).
- 5 42. Payseur, B. A., Place, M. & Weber, J. L. Linkage disequilibrium between STRPs and
6 SNPs across the human genome. *Am. J. Hum. Genet.* **82**, 1039–50 (2008).
- 7 43. Willems, T. F., Gymrek, M., Highnam, G., Mittelman, D. & Erlich, Y. The landscape of
8 human STR variation. *Genome Res.* gr.177774.114– (2014). doi:10.1101/gr.177774.114
- 9 44. Sawaya, S., Jones, M. & Keller, M. *Linkage disequilibrium between single nucleotide*
10 *polymorphisms and hypermutable loci*. 20909 (Cold Spring Harbor Labs Journals, 2015).
- 11 45. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis
12 thaliana inbred lines. *Nature* **465**, 627–31 (2010).
- 13 46. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana
14 Accessions. *Cell* **166**, 492–505 (2016).
- 15 47. Antignani, V. *et al.* Recruitment of PLANT U-BOX13 and the PI4Kβ1/β2
16 Phosphatidylinositol-4 Kinases by the Small GTPase RabA4B Plays Important Roles during
17 Salicylic Acid-Mediated Plant Defense Signaling in Arabidopsis. *Plant Cell* **27**, 243–261
18 (2015).
- 19 48. Tian, D., Araki, H., Stahl, E., Bergelson, J. & Kreitman, M. Signature of balancing
20 selection in Arabidopsis. *Proc. Natl. Acad. Sci.* **99**, 11525–11530 (2002).
- 21 49. Karasov, T. L. *et al.* The long-term maintenance of a resistance polymorphism through
22 diffuse interactions. *Nature* **512**, 436–440 (2014).
- 23 50. Lynch, M. The rate of polygenic mutation. *Genet. Res.* **51**, 137–148 (1988).

Press *et al.* (2017) Massive variation of STRs in *A. thaliana*

- 1 51. Rival, P. *et al.* The conserved PFT1 tandem repeat is crucial for proper flowering in
2 *Arabidopsis thaliana*. *Genetics* **198**, 747–754 (2014).
- 3 52. Undurraga, S. F. *et al.* Background-dependent effects of polyglutamine variation in the
4 *Arabidopsis thaliana* gene ELF3. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19363–19367 (2012).
- 5 53. Press, M. O. & Queitsch, C. Variability in a Short Tandem Repeat Mediates Complex
6 Epistatic Interactions in *Arabidopsis thaliana*. *Genetics* genetics.116.193359 (2016).
7 doi:10.1534/genetics.116.193359
- 8 54. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with
9 transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
- 10 55. Nellore, A. *et al.* Human splicing diversity and the extent of unannotated splice junctions
11 across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266
12 (2016).
- 13 56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
14 *Res.* **27**, 573–580 (1999).
- 15 57. Boyle, E. A., O’Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: optimized
16 modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*
17 **30**, 2670–2672 (2014).
- 18 58. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*.
19 *Nature* **477**, 419–23 (2011).
- 20 59. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O’Roak, B. J. & Shendure, J. Single
21 molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency
22 variation. *Genome Res.* **23**, 843–54 (2013).

Press *et al.* (2017) Massive variation of STRs in *A. thaliana*

- 1 60. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana
2 reference genome. *Plant J. Cell Mol. Biol.* **89**, 789–804 (2017).
- 3 61. Sullivan, A. M. *et al.* Mapping and Dynamics of Regulatory DNA and Transcription
4 Factor Networks in *A. thaliana*. *Cell Rep.* **8**, 2015–30 (2014).
- 5 62. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation
6 for Statistical Computing, 2016).
- 7 63. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of Arabidopsis thaliana.
8 *Science* **301**, 653–7 (2003).
- 9 64. Zaykin, D. V., Pudovkin, A. & Weir, B. S. Correlation-Based Inference for Linkage
10 Disequilibrium With Multiple Alleles. *Genetics* **180**, 533–545 (2008).
- 11 65. Horton, M. W. *et al.* Genome-wide patterns of genetic variation in worldwide
12 Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
- 13 66. Karatzoglou, A., Smola, A. & Hornik, K. *kernlab: Kernel-Based Machine Learning Lab*.
14 (2016).
- 15 67. Shabalín, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
16 *Bioinformatics* **28**, 1353–1358 (2012).
- 17 68. Therneau, T. M. *coxme: Mixed Effects Cox Models*. (2015).
- 18 69. Hurvich, C. M. & Tsai, C.-L. Regression and time series model selection in small
19 samples. *Biometrika* **76**, 297–307 (1989).

20