SUPPLEMENTAL NOTE

MIPSTR optimization and protocol updates. We found that MIP captures were most robust using a previously reported MIP capture protocol (Hiatt et al. 2013). Additionally, we implemented changes in the computational pipeline that led to improvements. Specifically, we altered the synthetic references used for alignment to use fractional repeat numbers found in the reference rather than integer unit counts, as the fractions tend to be shared across different strains, and more than half of repeats are found in fractional copy number (*e.g.* CAGCAGCA, where the unit is CAG, would have copy number 2.66). Therefore, alignments could be subject to artificially induced gaps if integer copy numbers are used in the synthetic reference. Additionally, we found that the requirement that different MIP capture events (tag-defined read groups) be present at read counts greater than one was unnecessary, as indeed 91% of capture events for Col-0 STRs are represented in a single read with the current protocol. We relaxed this requirement to at least four reads representing at least two capture events.

MIPSTR parameters. The MIPSTR technique (Carlson et al. 2015) was originally developed for high stringency to compensate for anticipated high levels of technical error. However, many parameters of the MIPSTR computational pipeline were never evaluated quantitatively for their effect on performance. Specifically, it was unclear whether the sensitivity/specificity tradeoff for including reads in the analysis was optimized. Using data for the *Arabidopsis thaliana* reference strain Columbia (Col), we performed a scan across BWA algorithms, parameters, and alignment score thresholds that yielded the highest agreement with this high-quality dideoxy-based reference genome. For both maximal exact match (MEM) and Smith-Waterman algorithms, we varied gap open costs across values [0,1,2,4,6 (default)] and gap extend costs across values [0,1,2,4]. We found that MEM, open 2, extend 2 showed the best performance in the context of other default parameters. We then tested the mismatch cost at values [0,1,2,4], finding that overall accuracy was best at mismatch cost 1. Finally, we evaluated different alignment score cutoffs for filtering reads, across values [145,150,155,160]. We found that a mismatch cost of 1 and alignment score threshold

of 150 gave the largest overall number of correct genotypes, with some tolerance to deviation in any parameter (Supplemental Table S1).

Accuracy of the method. A dropout of ~15% of loci was largely attributable to inefficient MIP capture of abundant intergenic TA/AT dinucleotide STRs (Supplemental Fig. S1, S2). Comparing our genotype calls to the high-quality Columbia (Col-0) reference genome, we found that MIPSTR genotype calls were 96% accurate (Supplemental Fig. S3b). In contrast, a comparison of MIPSTR calls with a reference-guided draft assembly of the Bur-0 strain's genome(Gan et al. 2011) indicated that less than 1% of STR variation was accounted for in this assembly (Supplemental Fig. S3c). As third-generation sequencing technologies become cost-appropriate for population analyses, however, STR calls in new genome assemblies are likely to become substantially better(Chaisson et al. 2015). Dideoxy sequence analysis of a small STR set in four diverse strains showed an accuracy of 95% (Supplemental Fig. S3d). Technical reproducibility of genotype calls was extremely high (Supplemental Fig. S3e). Across the 96 *A. thaliana* strains, the average genotyping rate was 80% with a subset of dinucleotides accounting for most dropout (Supplemental Fig. S4a, S4b). Rarefaction analysis implied that many more STR alleles are expected with further sampling of *A. thaliana* strains (Supplemental Fig. S4c).

STR annotation. We used Araport11 (Cheng et al. 2017) to annotate STRs with overlapping genomic features as follows. We chose to require at least 50% overlap with any given feature, and as such features we considered genes and transposons. We used BEDTOOLS (Quinlan and Hall 2010) to find all such features for each STR. Given this overlap, we annotated each STR such that it had only one annotation among coding/intron/UTR/intergenic using the following logic, because Araport did not provide unambiguous mappings. We annotated as 'intergenic' all STRs which did not overlap with a gene, or which overlapped with a pseudogene, or a transposable element gene, or a large segment (>200bp) annotated as transposable element. (Of the 158 STRs falling in such transposable element regions, only 46 overlapped with a TE gene). This

left only STRs overlapping with genes. We annotated as 'coding' all genic STRs overlapping with the annotation 'CDS' from any transcript of any gene. We annotated as 'intronic' all genic STRs which had the annotation 'protein' but did not overlap with a CDS. We annotated as 'UTR' all genes overlapping with 5'/3' UTRs or other transcript regions. This meant that some ambiguous annotations, for instance noncoding RNA genes, were annotated as UTRs (which is technically true). Spot checking indicated that this approach gave reasonably accurate results; we found no inconsistencies in extensive manual data checks. We further used BEDTOOLS to compare STR loci to seven different DNaseI-seq experiments in *A. thaliana* strain Col-0 (Sullivan et al. 2014). These included: whole 7 day seedlings; root hair cells only; whole root; root non-hair cells only; seed coat at 4 days post anthesis; opened flowers only; non-opened flowers only. We downloaded these datasets from http://plantregulome.org/ on March 13th, 2017. We recorded the number of DNaseI hypersensitive sites (DHS) overlapping at least 50% with each STR. For purposes of analysis, we considered all STRs overlapping with a DHS in at least one tissue to have a DHS.

Bur-0 genome comparisons. To compare our STR genotype calls to a draft genome of the Bur-0 strain (Gan et al. 2011), we considered all indels of this draft genome relative to the Col-0 reference that overlapped with MIP-targeted STR regions. This yielded a short list of variants, which we manually compared to our own STR variant calls (Supplemental Figure S3c).

Analysis of expanded STRs. To detect modestly expanded STRs, we used a simple metric to determine, for each STR, whether its allele size distribution showed evidence of substantial positive outliers. This was accomplished by estimating the median of the allele distribution, and computing the relative distance of the maximum allele size from the median allele size. STRs that showed expansions had higher rates of missing data (20% vs. 15% missing for comparable STRs; $p = 0.0009$, U-test). To understand the functional consequences of these modest expansions, we assayed gene expression in selected diverse *A. thaliana* strains representing various intronic expansions and

controls. We designed primers flanking STR-residing introns, along with primers bridging other exon-exon junctions where possible. We investigated potential splicing irregularities by reverse-transcription PCR (RT-PCR) and agarose gel electrophoresis, followed by gel purification and dideoxy sequencing of relevant RT-PCR amplicons (Supplemental Figure S6).

In three of six cases, for genes *NTM1*, *AT1G30540* and *AT1G24145*, we found that intronic expansions were associated with splicing irregularities (Supplemental Figure S6a,b). We did not formally establish a causal relationship between splice defects and STR expansions. In all three cases of splice defects, there are additional mutations in the intron. In one case, *AT1G24145*, there is a splice-site-disrupting mutation in Kz-9 explaining the partial intron retention (but not the whole intron retentions in other strains, Supplemental Figure S6a); in the case of *AT1G30540*, there is a ~600bp insertion in the intron. As there is no reliable way to establish the ordering of these mutations, these STR expansions may result from relaxed selection on the introns in question. Nonetheless, these associations suggest that STR expansions may be a sign of such splicing irregularities, as in each case all relevant polymorphisms in intron sequences were absent from the 1001 Genomes resource (Alonso-Blanco et al. 2016). It is further noteworthy that normally-spliced transcripts appear to be present in all cases at some level, though we cannot judge relative quantities of normal vs. aberrantly spliced products. We additionally measured transcript abundance of selected expansion-associated genes relative to control gene *UBC21*, with mixed results. In contrast to *MEE36* (Figure 2f), we found no evidence of effects of another expansion, in the 5' UTR of the *CaM1* gene, on expression of *CaM1* by qRT-PCR (data not shown).

Inference of selection on STRs. We used a bootstrap-aggregation (bagging) procedure as described in the main text and Methods to learn an ensemble model for predicting the expected degree of allelic variation in each STR. We estimated STR variability estimated using two complementary measures: 1) the base-10 logarithm of the standard deviation of an STR's unit copy number across all alleles (log(SD)), and 2) the Shannon entropy of the allele distribution for each STR (computed using the infotheo package

(Meyer 2014)). Predictions about variation based on each measure were performed, but log(SD) was ultimately used for all inference as following a more tractable distribution of STR variation (Supplemental Fig. S8). Features used to predict each STR's expected neutral variation included STR purity (in Col-0), median unit copy number, STR unit length, STR GC content, and GC content of the MIP-captured region. Features discarded from analysis included MIP targeting and ligation arm GC content, chromosome, and chromosomal position. 1000 bootstrapped SVR models, each trained on a randomly sampled 25% of intergenic STRs, were then used to make predictions about the variability of each STR, yielding a distribution of 1000 predictions for the variability of each STR. While we cannot necessarily expect that all intergenic STRs are free of selective influence, they should at the least be strongly enriched for neutrally evolving STRs relative to other STRs, and they thus represent a reasonable approximation to a neutral model for STR variability. We compared two different approaches to predicting variability of the entire set of STRs from intergenic STR variation, support vector regression (SVR) and simple linear regression with cross-validation (Supplemental Figure S9a, S9b). We found that while both methods performed similarly on gene-associated STRs, SVR models showed somewhat better fit to intergenic STRs, specifically demonstrating a more or less linear fit between predictions and true values. However, improved fit could have been due to overfitting, so we used bootstrap aggregation (bagging) to ensure that predictions were robust and to describe a distribution of predictions (Supplemental Figure S9c). These bagging results were similar to cross-validation alone, though predictions on intergenic STRs showed slightly worse fit, consistent with ameliorated overfitting. Notably, we replicate our past observation (with a much smaller dataset) that coding STRs showed a rather lower correlation with expected variability than noncoding STRs (Carlson et al. 2015) (Supplemental Figure S9a, S9b, S9c). These collected results suggest that intergenic STRs are a reasonably good model for neutral STR variation. Finally, there was no substantial evidence for a correlation of bias in inferences of selection made from this model with the response variable (Supplemental Figure S9d). Predictions were not substantially different when the entire STR set was used for bagging (instead of

intergenic STRs alone; Supplemental Figure S9e), arguing that intergenic STRs alone are fairly representative as a training set. Using all STRs for bagging also yielded qualitatively similar inferences about constraint; for example, coding STRs still show elevated signatures of constraint compared to other STR classes (Supplemental Figure S9f). Also, this argues that overfitting did not meaningfully affect predictions on genic STRs made from a model fit to only intergenic STRs. The most informative features of the SVR model (fit to intergenic STRs) were the median STR copy number across strains, GC content of the MIP capture region, and the size of the STR unit (Supplemental Figure S10). We make a series of observations about constrained and hypervariable STRs in the main text and at Supplemental Figure S11 and Supplemental Table S2.

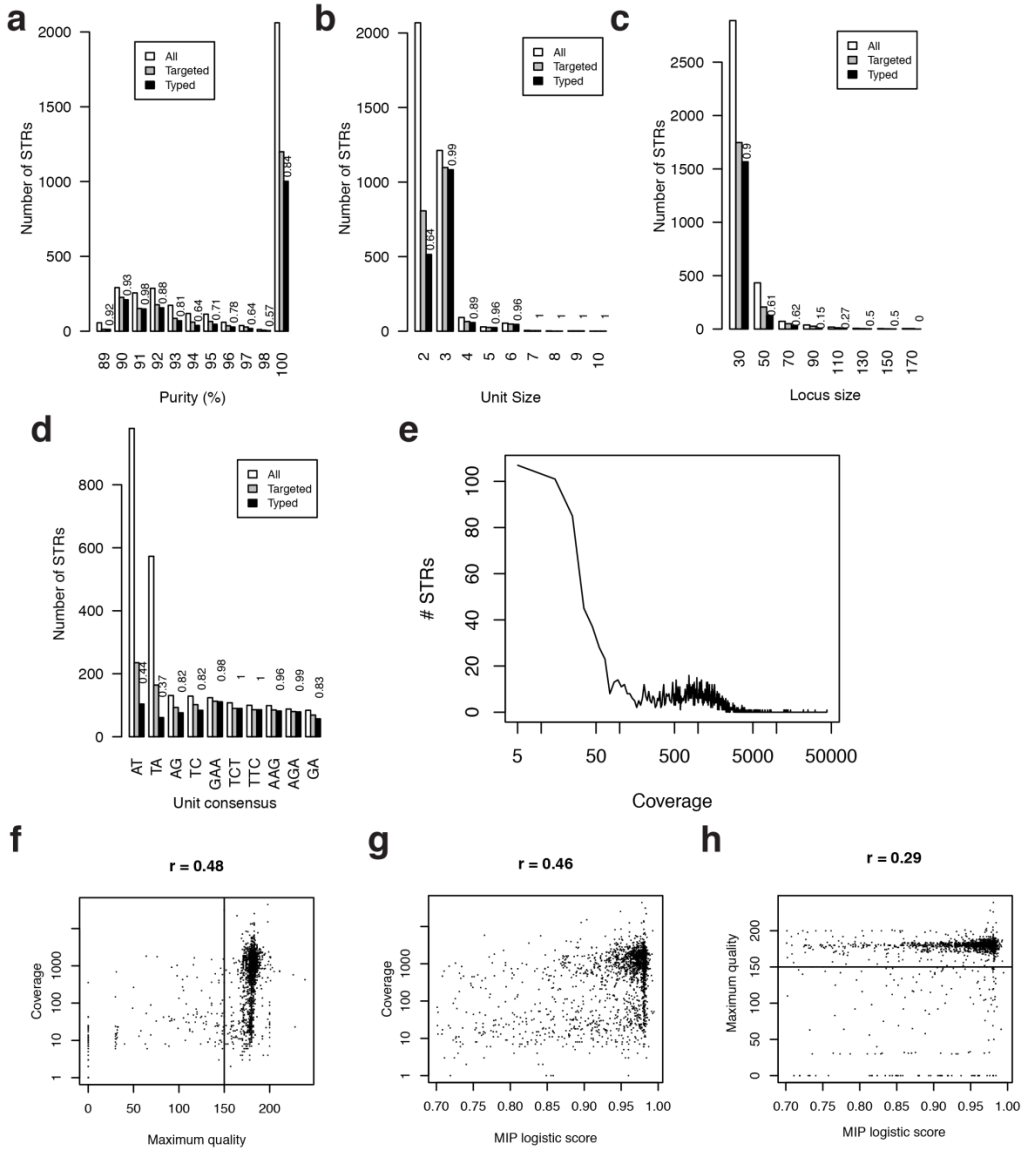
In the main text, we note that 3/24 coding STRs found to be hypervariable relative to SVR predictions encoded polyserines in F-box proteins. Further, two such polyserines were at the extreme N-terminus of the protein. While STRs prefer the N and C-terminal regions of coding sequences (Huntley and Clark 2007), there is modest evidence that this tendency is more marked in hypervariable coding STRs than those under purifying selection (odds ratio = 2.8, p = 0.050, Fisher's Exact Test). Potentially these regions of proteins are under relaxed selection, yielding opportunities for functional adaptation. Across both coding and noncoding STRs, STR motifs generally matched background patterns (Supplemental Figure S11).

Detection of STR-phenotype associations and effect size estimation. There is not a well-established framework for detecting associations between STR genotypes and phenotype, and thus such analyses have tended to operate on an *ad hoc* basis (Mackay et al. 2012; Carlson et al. 2015; Press et al. 2014; Gymrek et al. 2015). We chose a linear mixed model as a relatively conservative analysis, which allowed us to account for population structure to some extent, and also maintained a certain degree of parsimony. However, we chose to model STR alleles as categorical variables, because there is no generalizable functional reason why STR allele length should necessarily

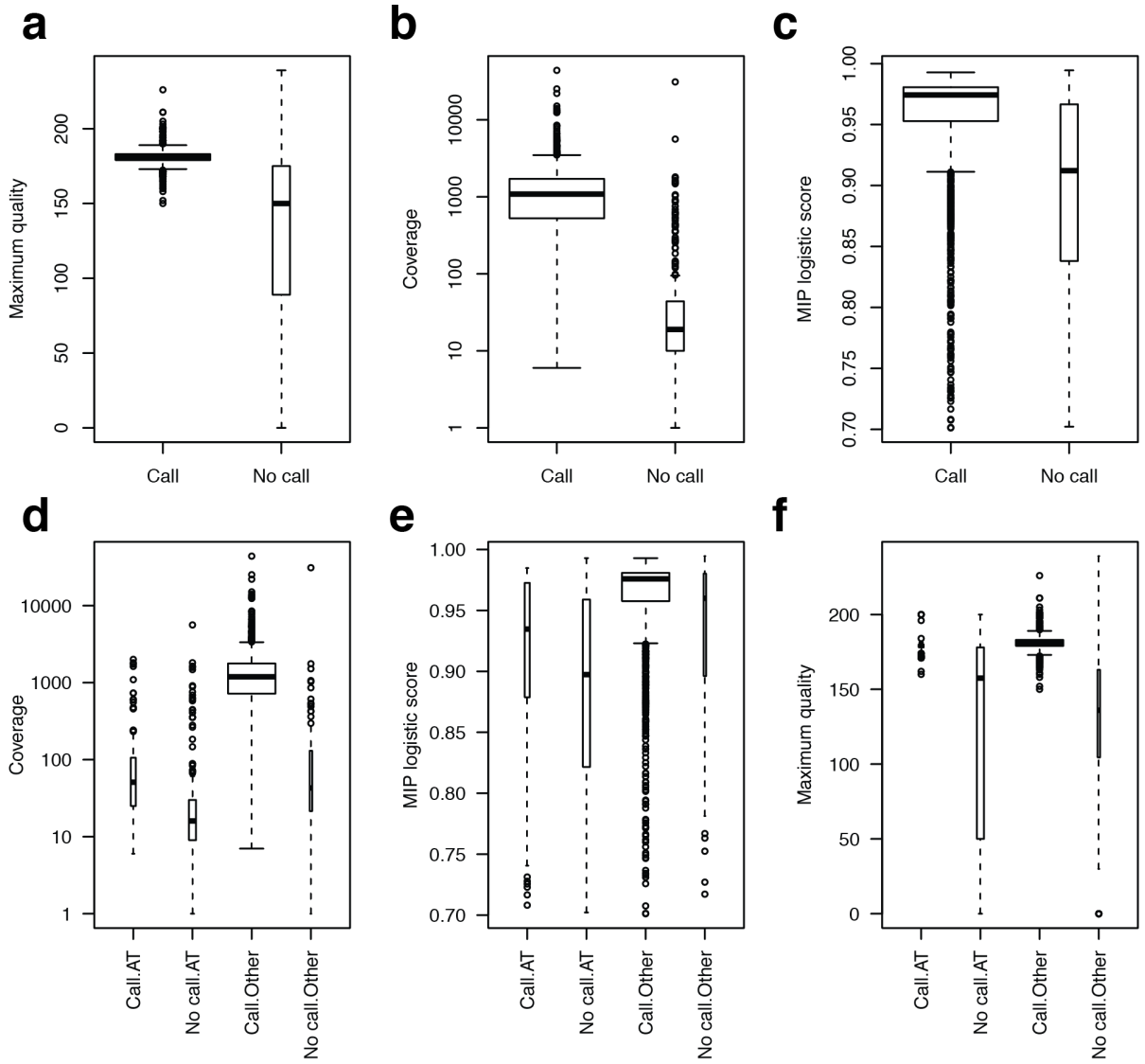
show a linear association with phenotype (and there are known counterexamples in which it does not) (Press et al. 2014). While it may reduce power somewhat, this approach allows us to make fewer assumptions about the nature of the genotype-phenotype map. There is some evidence that the overall approach shows inflation for some, but not all, phenotypes (Supplemental Figure S13). However, this apparent inflation disappears when genotypes are randomized, indicating that the model itself is not obviously mis-specified with respect to e.g. population structure. Notably, this pattern has been observed previously in STR-phenotype association tests (Gymrek et al. 2015). In contrast, simple tests that do not correct for population structure give clearly inflated p-value distributions for STR-phenotype associations (data not shown). It could be that STRs are a more sensitive readout of population structure than the traditional SNPs of genome-wide association studies, which cannot be accounted for by identity-by-state summary statistics. This topic deserves further study in the future, as analyses of STR data are likely to become routine with the influx of new technologies. Nonetheless, for the purposes of this study, follow-up validation experiments (Supplemental Fig S16) indicate that we do detect meaningful STR-phenotype associations with our tests, even if these tests are merely heuristic.

We additionally performed an ANOVA on residual variance in the phenotype days to flowering, long days, after adjusting for the random effect of population structure (Supplemental Table S7). Using this admittedly naïve technique, we find that the sums of squares associated with STR loci are larger than those associated with SNP loci. This observation is rather surprising, given that the SNP loci are generally well-known and well-described contributors to flowering variation (Atwell et al. 2010), whereas the STRs are much less well understood. Furthermore, in the context of the Type II ANOVA (where each factor's contribution is estimated only after fitting all other factors), none of the loci are nominally significant at the 5% level. We therefore treat this result with caution, but nonetheless consider that it may be informative, pending validation in other systems and better functional characterization of the STR loci.

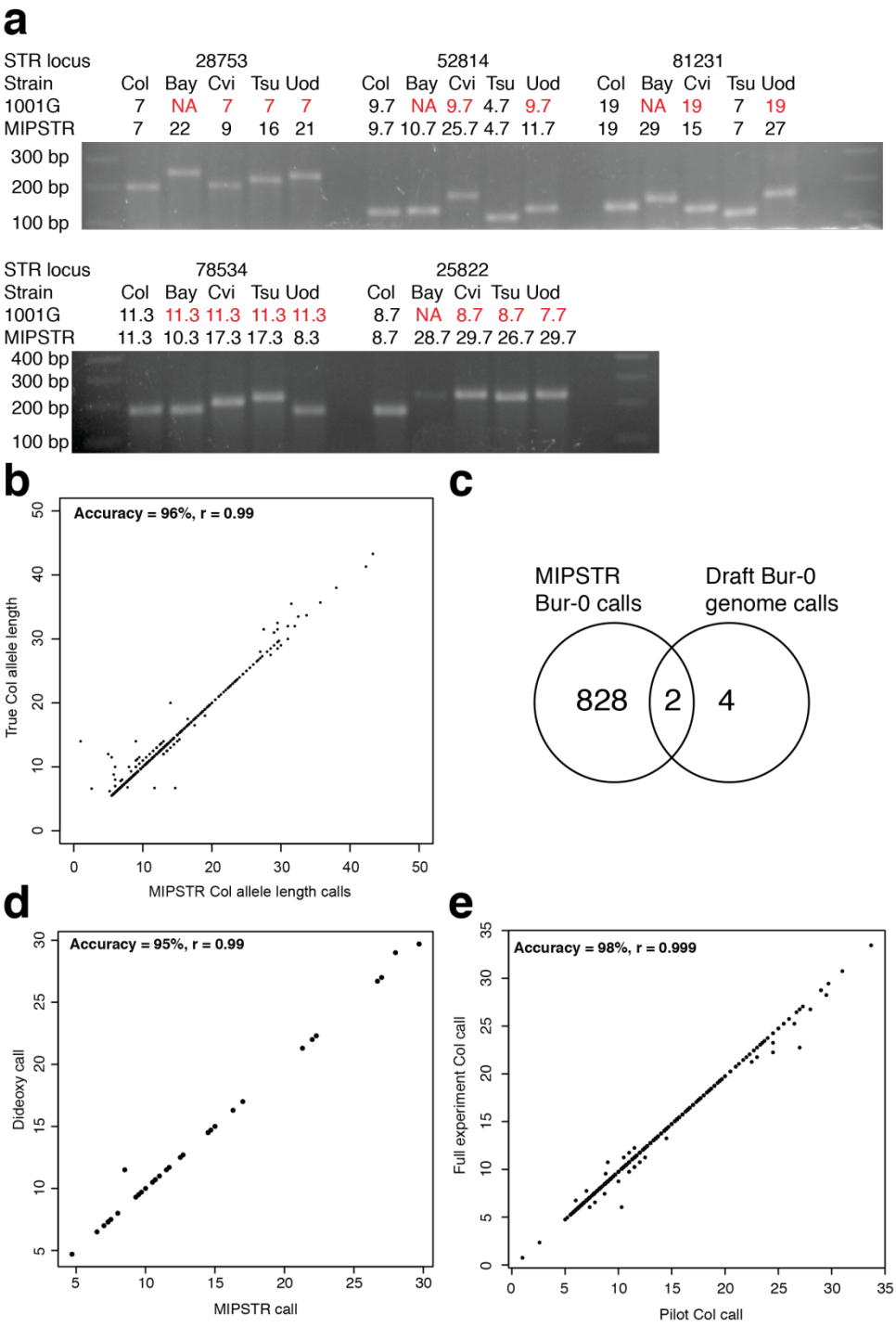
SUPPORTING FIGURES



Supplemental Fig. S1. Ascertainment of STRs is related to various STR qualities, sequencing coverage, and MIP features. Distribution and ascertainment of STRs by: (a): STR purity, (b): consensus unit size, (c): overall STR locus size, and (d): unit consensus sequence. For (a-d): "All": all STRs which match the definition of target S TRs (in the reference genome) for this study, e.g. <=180 bp length in TAIR10, >=89% purity in TAIR10, 2-10 bp nucleotide motif. "Targeted": the 2046 STRs which were targeted by an explicitly designed MIP. "Typed": STRs successfully genotyped in the Col-0 reference genome in a MIPSTR assay. Numbers above bars indicate the proportion of targeted STRs in the relevant category that were successfully genotyped. (e): coverage distribution across all targeted STRs (note log scale on X axis). (f-h): relationships of coverage, maximum read alignment score ("quality") at each locus, and MIPGEN(Boyle et al. 2014) locus logistic score. For (f-h): solid lines indicate the alignment score threshold used; Pearson correlation of the plotted quantities is displayed above each plot.

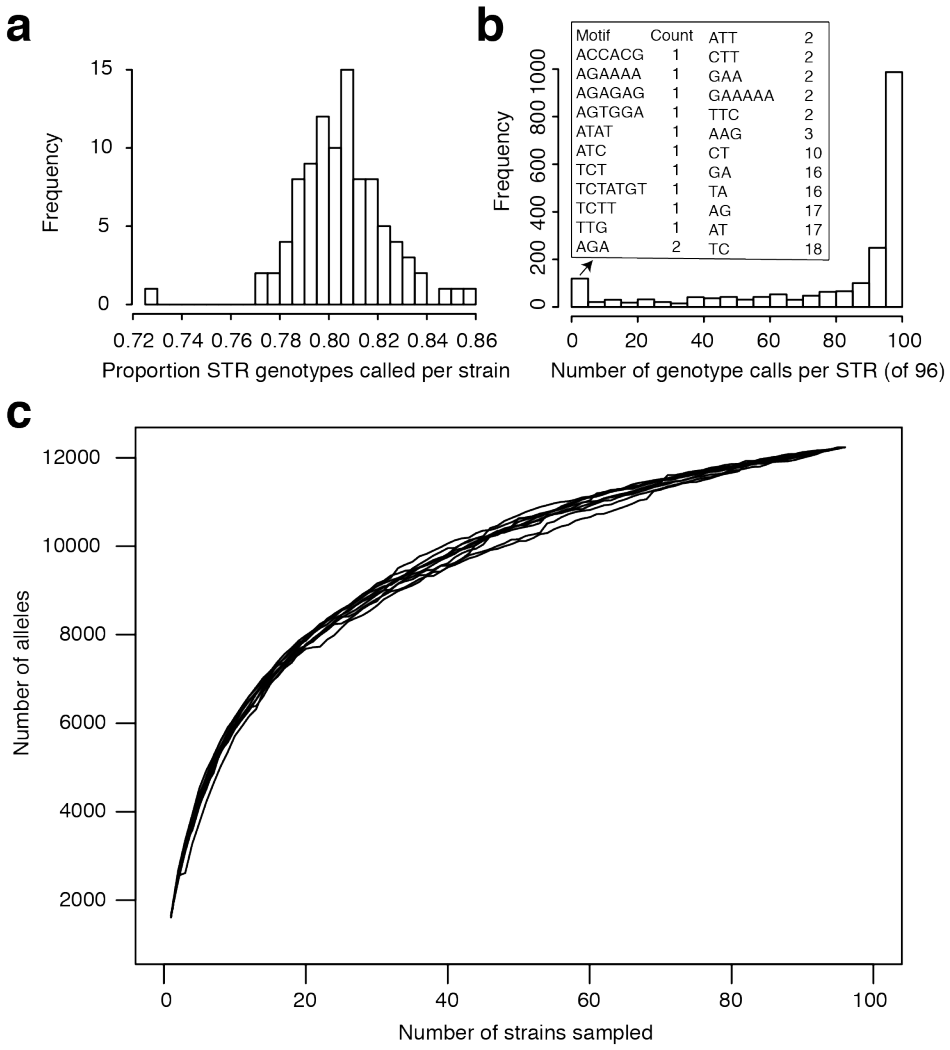


Supplemental Fig. S2. **Coverage differences explain MIP dropout and are associated with $[AT]_n$ dinucleotide STR loci.** (a-c): differences between STR loci with genotype calls and loci without calls for various technical quantities. (d-f): Similar to (a-c), but breaking down also by whether the targeted STR locus is an $[AT]_n$ dinucleotide. In all boxplots, widths are scaled to the number of observations in each category.

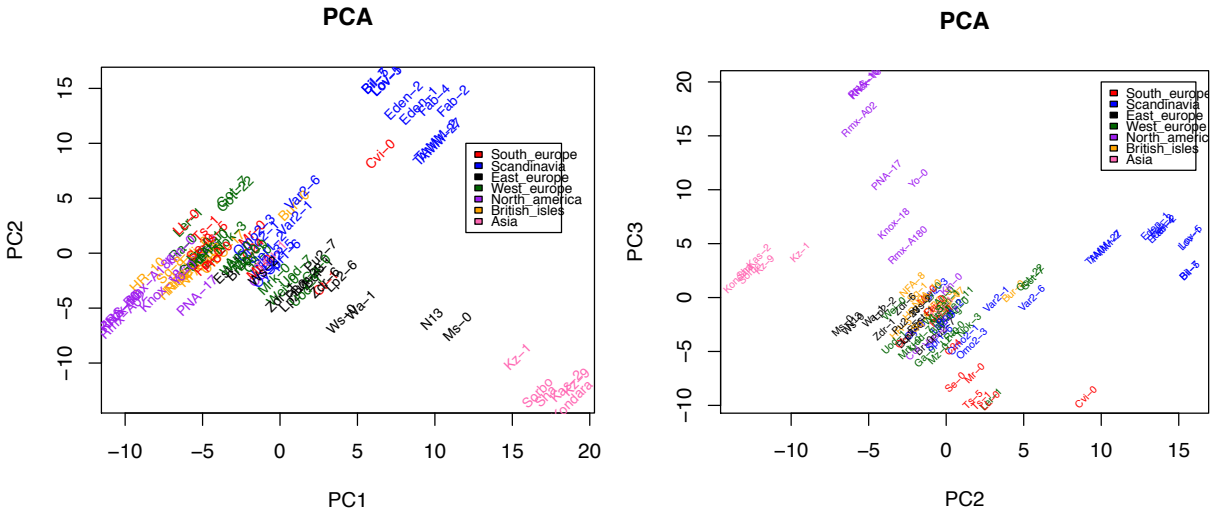


Supplemental Fig. S3. MIPSTR method makes accurate, reproducible calls in diverse *Arabidopsis thaliana* strains. (A): MIPSTR allele calls at five STR loci agree with amplified fragment length polymorphism data and reveal reference bias for STRs in the 1001 Genomes resource database (1001G) across five diverse *A. thaliana* strains. NA: no genotype call; red text indicates incorrect 1001G allele calls. (B, D, E): Comparison of MIPSTR calls with other data sources or between experiments. Accuracy = percentage of alleles called in both data sources (B: TAIR10 reference genome, D: dideoxy sequencing of diverse strains, E: different MIPSTR

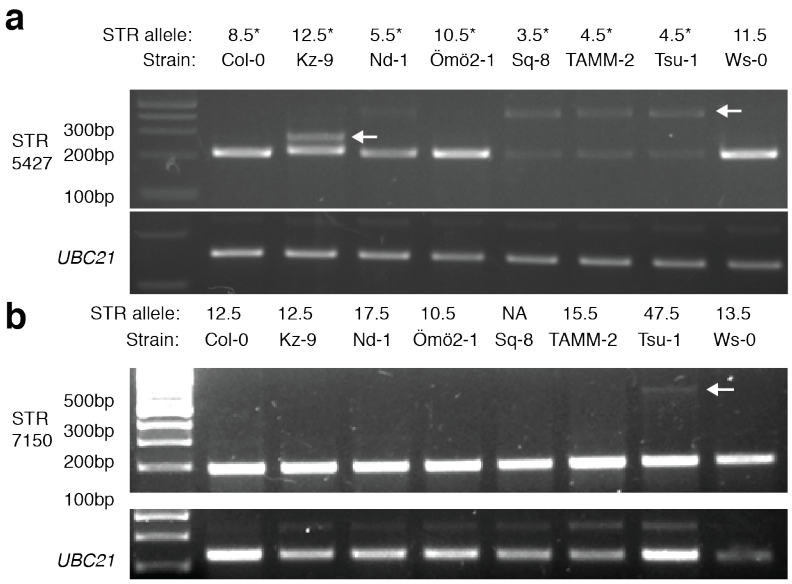
experiments) that agree; r = Pearson correlation coefficient for allele length between data sources. (C): Comparison of STR polymorphisms (relative to Col) in the Bur-0 strain in a draft genome assembly (Cao et al. 2011) and from MIPSTR (this study). The four variants called in the draft assembly but not by MIPSTR are marked as missing genotypes by MIPSTR, not reference genotypes. (D): STRs in strains Col, Bay-0, Cvi-0, Tsu-1, Uod-7 genotyped by dideoxy sequencing, n=37 total after omitting missing genotypes. (E): Technical variation in MIPSTR calls between independently generated libraries of a small pilot experiment at low plexity (5 libraries) on a MiSeq instrument and the full experiment at high plexity (96 libraries) on a NextSeq instrument.



Supplemental Fig. S4. Ascertainment of STRs by MIPSTR. (A): Number of STR genotypes called per strain is consistent across strains. (B): Number of calls per STR shows most STRs are highly ascertained. Inset table represents the the STR motif composition of STRs in the 0-5 bin of the histogram. (C): Rarefaction analysis indicates a failure to saturate new STR alleles in the current sample of strains, indicating further diversity. Five different rarefaction samples are shown as curves.



Supplemental Fig. S5. Principal component analysis of STR variation reveals demographic history. (A): Same as Figure 1C, but with strain names. (B): PC2 plotted against PC3 reveals additional structure.



Supplemental Fig. S6. Additional molecular phenotypes associated with STR modest expansions. (a, b): STR-associated aberrant splicing detected by RT-PCR on cDNAs from indicated strains. White arrows indicate aberrant splice forms (whole or partial intron retention). For each sample, separate reactions with primers targeting *UBC21* transcripts were performed alongside as a control. One of two to three similar biological replicates is shown. (a): Splicing of the STR-bearing intron of *AT1G24145*. *: Indicated STR allele is inferred from dideoxy sequencing, having a missing genotype from MIPSTR. (b): Splicing of the STR-bearing intron of *AT1G30540*. High molecular weight of the aberrant form is due to an apparent large (non-STR) insertion in the intron, indicated by dideoxy sequencing and PCR from genomic DNA.

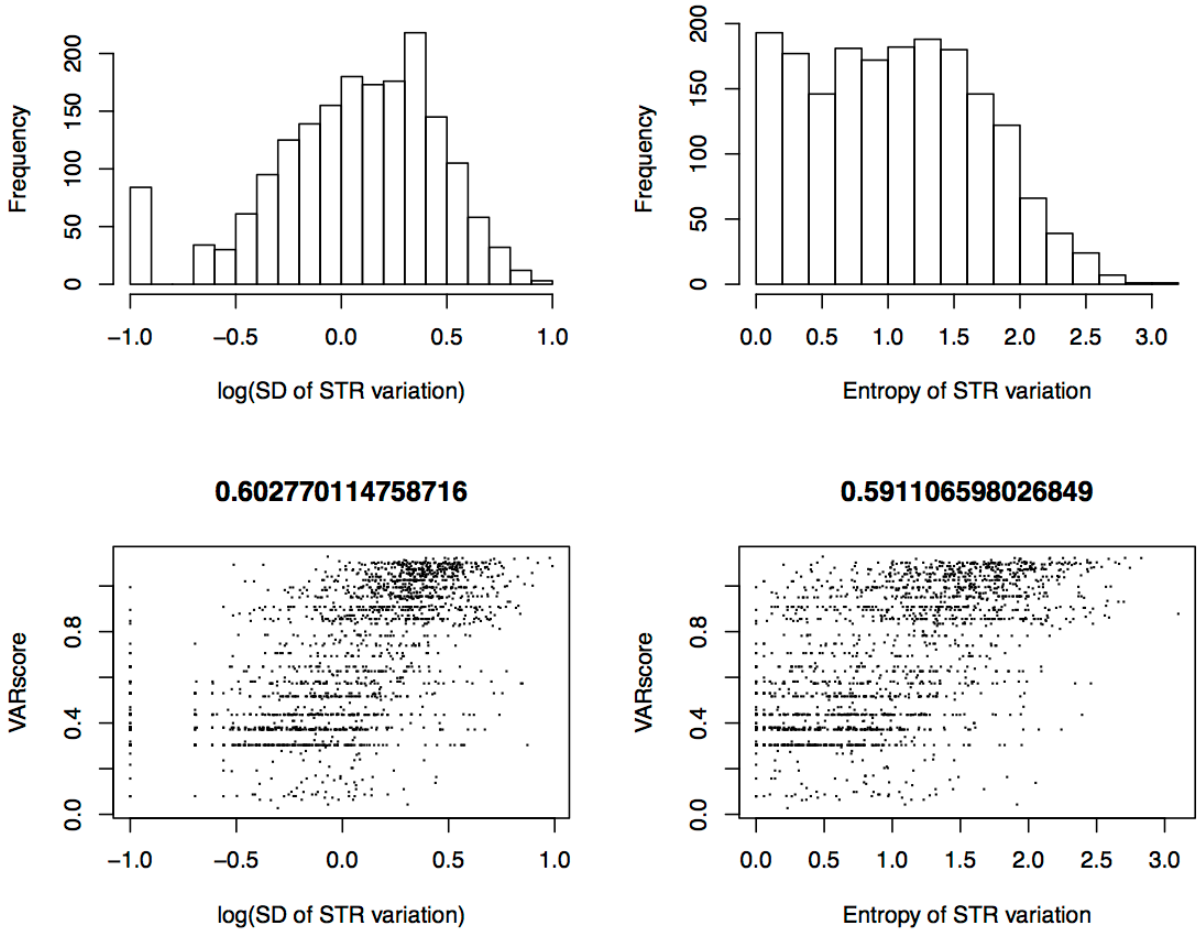
a

```
TAIR10_gDNA GAGTCACAAACCCAGCAAGGATTTGCCTAgtaagtttattatgaatgatc
Mr-0_cDNA   GANTCACAAACCCAGCAAGGATTTGCCTAgtaagtttattatgaatgatc
TAIR10_cDNA GAGTCACAAACCCAGCAAGGATTTGCCTA----------------------

TAIR10_gDNA ttgatctgtctctctttgtatcatcat---cgatcatttctaaatttggg
Mr-0_cDNA   ttgatctgtctctctttgtatcatcatcatcgatcatttctaaatttggg
TAIR10_cDNA ---------------------------------------------------

TAIR10_gDNA ttctt---tctgttgttgttgttgttgtt----------------------
Mr-0_cDNA   ttcttcttcttgttgttgttgttgttgttgttgttgttgttgttgttgtt
TAIR10_cDNA ---------------------------------------------------

TAIR10_gDNA ------------gctgctgcagAGTTAGCTAGGATCCAATCGGAAGATCT
Mr-0_cDNA   gttgttgttgttgttgttgcagAGTTAGCTAGGATCCAATCGGAAGATCT
TAIR10_cDNA ----------------------AGTTAGCTAGGATCCAATCGGAAGATCT

TAIR10_gDNA TGAATGGTATTTCTTCTCTCCGATTGAGTACACGAACCCGAATAAGATGA
Mr-0_cDNA   TGAATGGTATTTCTTCTCTCCGATTGAGTACACGAACCCGAATCAGATGA
TAIR10_cDNA TGAATGGTATTTCTTCTCTCCGATTGAGTACACGAACCCGAATAAGATGA

TAIR10_gDNA AAATGAAGAGGACGACAGGTTCTGGGTTTTGGAAACCTACTGGTGTTGAT
Mr-0_cDNA   AAATGAAGAGGACGACAGGTTCTGGGTTTTGGAAACCTACTGGTGTTGAT
TAIR10_cDNA AAATGAAGAGGACGACAGGTTCTGGGTTTTGGAAACCTACTGGTGTTGAT

TAIR10_gDNA CGGGAAATTAGGGATAAAAGAGGAAATGGTGTTGTGATAGGGATTAAGAA
Mr-0_cDNA   CGGGAAATTAGGGATAAAAGAGGAAATGGTGTTGTGATCGGGATTAAGAA
TAIR10_cDNA CGGGAAATTAGGGATAAAAGAGGAAATGGTGTTGTGATAGGGATTAAGAA

TAIR10_gDNA GACGCTTGTGTACCATGAAGGTAAGAGTCCTCATGGAGTTAGAACTCCTT
Mr-0_cDNA   GACGCTTGTGTACCATGAAGGTAAGAGTCCTCGTGGAGTTAGAACTCCTT
TAIR10_cDNA GACGCTTGTGTACCATGAAGGTAAGAGTCCTCATGGAGTTAGAACTCCTT

TAIR10_gDNA GGGTT
Mr-0_cDNA   GGGTT
TAIR10_cDNA GGGTT
```

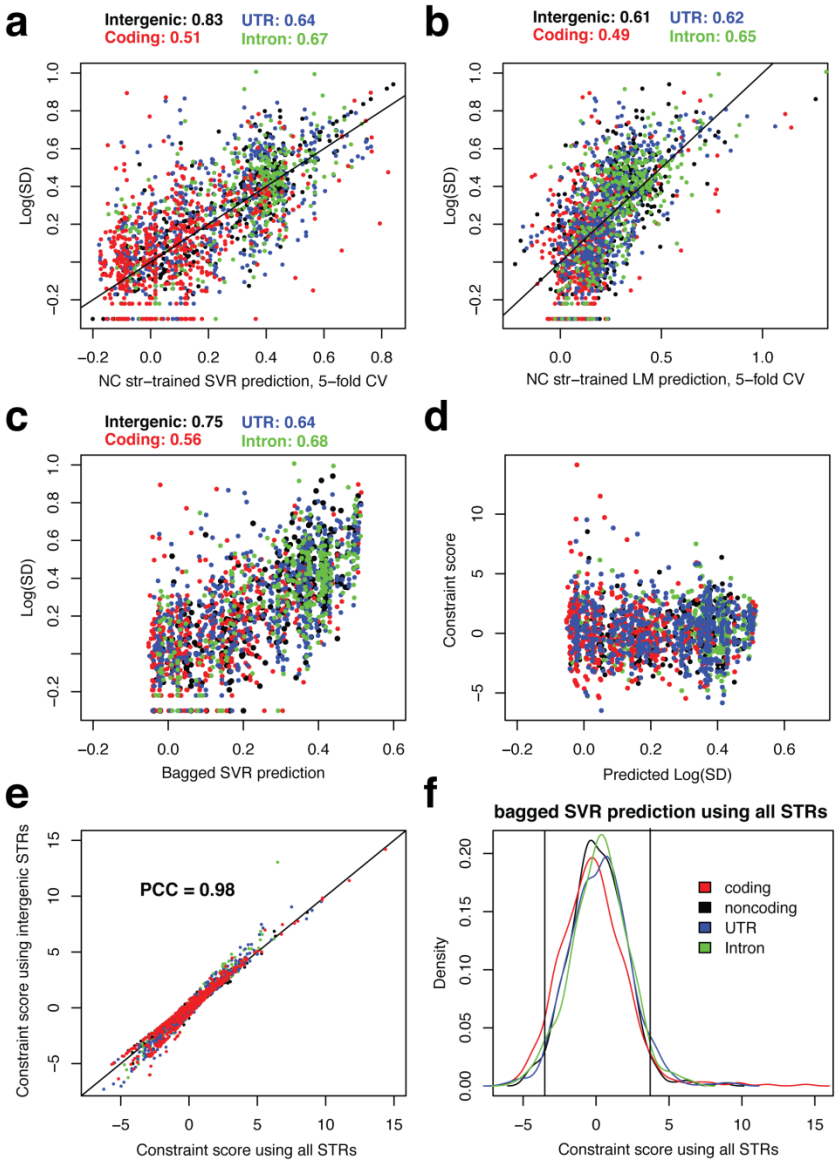
b

```
NTM1_full     MMKGLIGYRFSPTGEEVINHYLKNKLLGKYWLVDEAISEINILSHKPSKD
NTM1_Mr0_cDNA MMKGLIGYRFSPTGEEVINHYLKNKLLGKYWLVDEAISEINILSHKPSKD

NTM1_full     LPKLARIQSEDLEWYFFSPIEYTNPNKMKMKRTTGSGFWKPTGVDREIRD
NTM1_nosplice LPSKFIMNDLDLSLFVSSSSIISKFGFFFLLLLLLLLLLLLLLLLLLLQ

NTM1_full     KRGNGVVIGIKKTLVYHEGKSPHGVRTPWVMHEYHITCLPHHKRKYVVCQ
NTM1_nosplice SX-------------------------------------------------

NTM1_full     VKYKGEAAEISYEPSPSLVSDSHTVIAITGEPEPELQVEQPGKENLLGMS
NTM1_nosplice ---------------------------------------------------

NTM1_full     VDDLIEPMNQQEEPQGPHLAPNDDEFIRGLRHVDRGTVEYLFANEENMDG
NTM1_nosplice ---------------------------------------------------

NTM1_full     LSMNDLRIPMIVQQEDLSEWEGFNADTFFSDNNNNYNLNVHHQLTPYGDG
NTM1_nosplice ---------------------------------------------------

NTM1_full     YLNAFSGYNEGNPPDHELVMQENRNDHMPRKPVTGTIDYSSDSGSDAGSI
NTM1_nosplice ---------------------------------------------------

NTM1_full     STTSYQGTSSPNISVGSSSRHLSSCSSTDSCKDLQTCTDPSIISREIREL
NTM1_nosplice ---------------------------------------------------

NTM1_full     TQEVKQEIPRAVDAPMNNESSLVKTEKKGLFIVEDAMERNRKKPRFIYLM
NTM1_nosplice ---------------------------------------------------

NTM1_full     KMIIGNIISVLLPVKRLIPVKKLX
NTM1_nosplice ------------------------
```

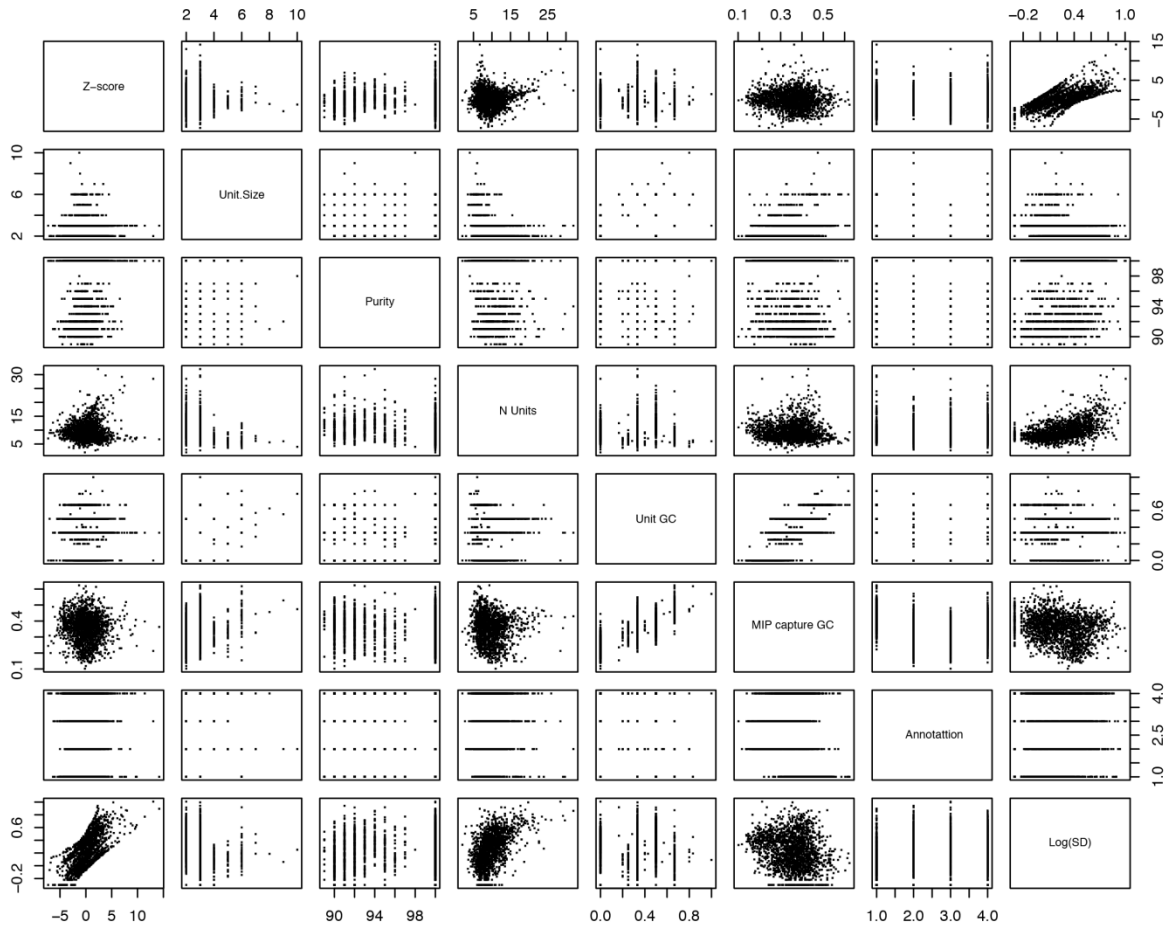
Supplemental Fig. S7. *NTM1* STR-associated intron retention in the Mr-0 strain is predicted to lead to a nonsense mutation. (a): Alignment of the TAIR10 reference gDNA and cDNA sequences in *NTM1* and dideoxy sequencing results of the region around the STR-containing intron from the Mr-0 *A. thaliana* strain. (b): *In silico* translation of the *NTM1* dideoxy sequencing results, with and without splicing the intron (and assuming no further variation).



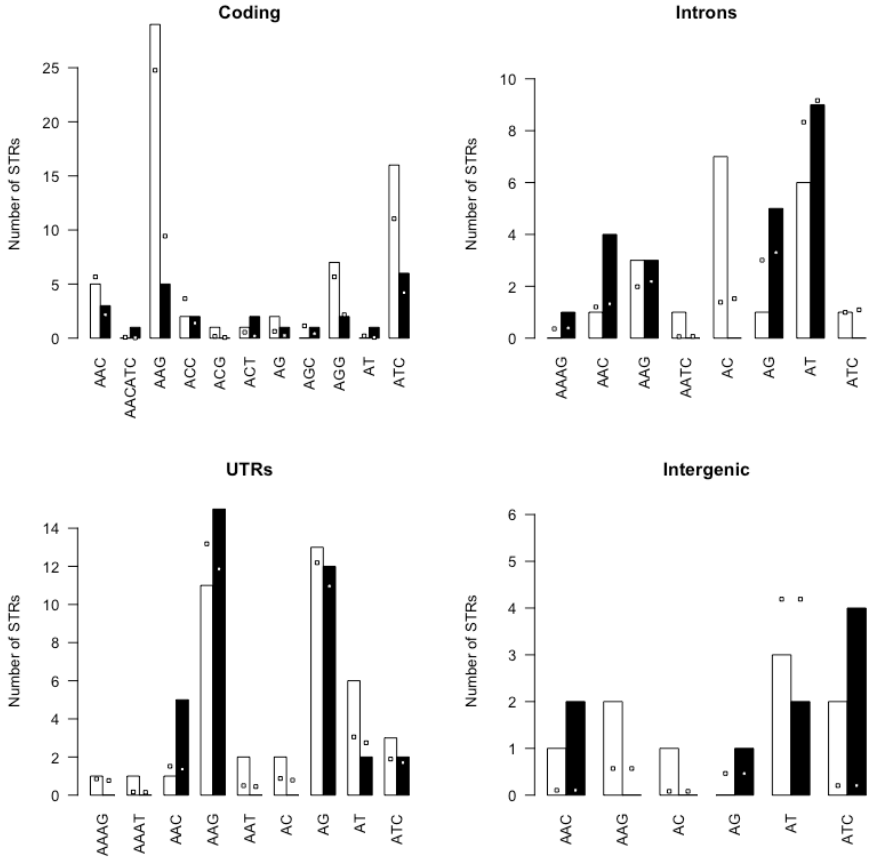
Supplemental Fig. S8. Different summaries of STR allelic variation are correlated with existing reference-sequence-based predictors of STR variability. Bottom panels: value above plot is the Pearson correlation between the two variables.



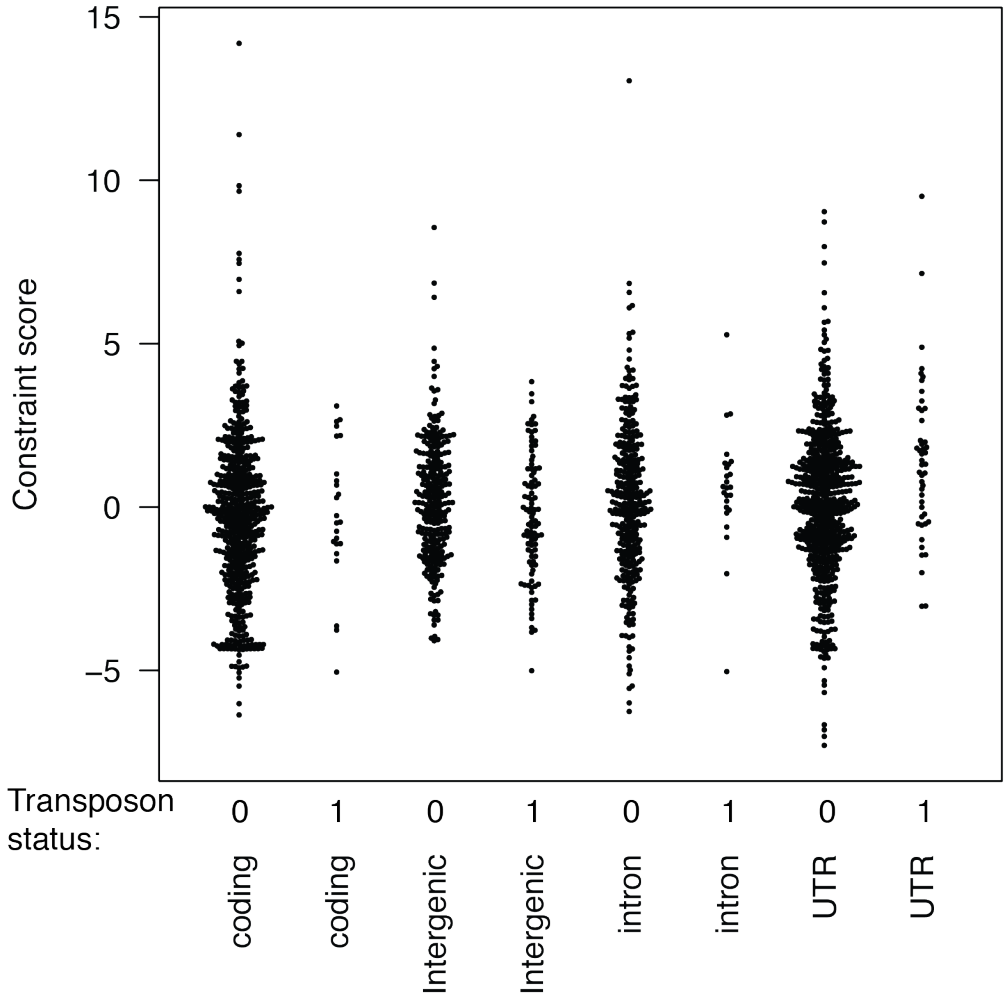
Supplemental Fig. S9. Performance of various models in predicting STR variation across 96 *A. thaliana* strains. (A): performance of support vector regression (SVR) when fit to all STRs with 5-fold cross-validation, stratified by genomic location. (B): performance of simple linear regression prediction when fit to all STRs with 5-fold cross-validation. (C): performance of bootstrap aggregation of SVR models ("bagging") fit to intergenic STRs only. (D): Little evidence for bias of bagged SVR predictions across the range of the predictor. In all panels, black points are intergenic STRs, red points are coding STRs, blue points are UTR STRs, and green points are intronic STRs. Values above each plot represent the Pearson correlation coefficient between the prediction and the true value in different STR classes. (E): Comparison of constraint scores derived from bagged SVR models trained on only intergenic STRs (Y axis) or all STRs (X axis). PCC: Pearson correlation coefficient. (F): Distribution of constraint scores across STR categories (indicated by color), using a constraint score derived from bagged SVR models trained on all STRs.



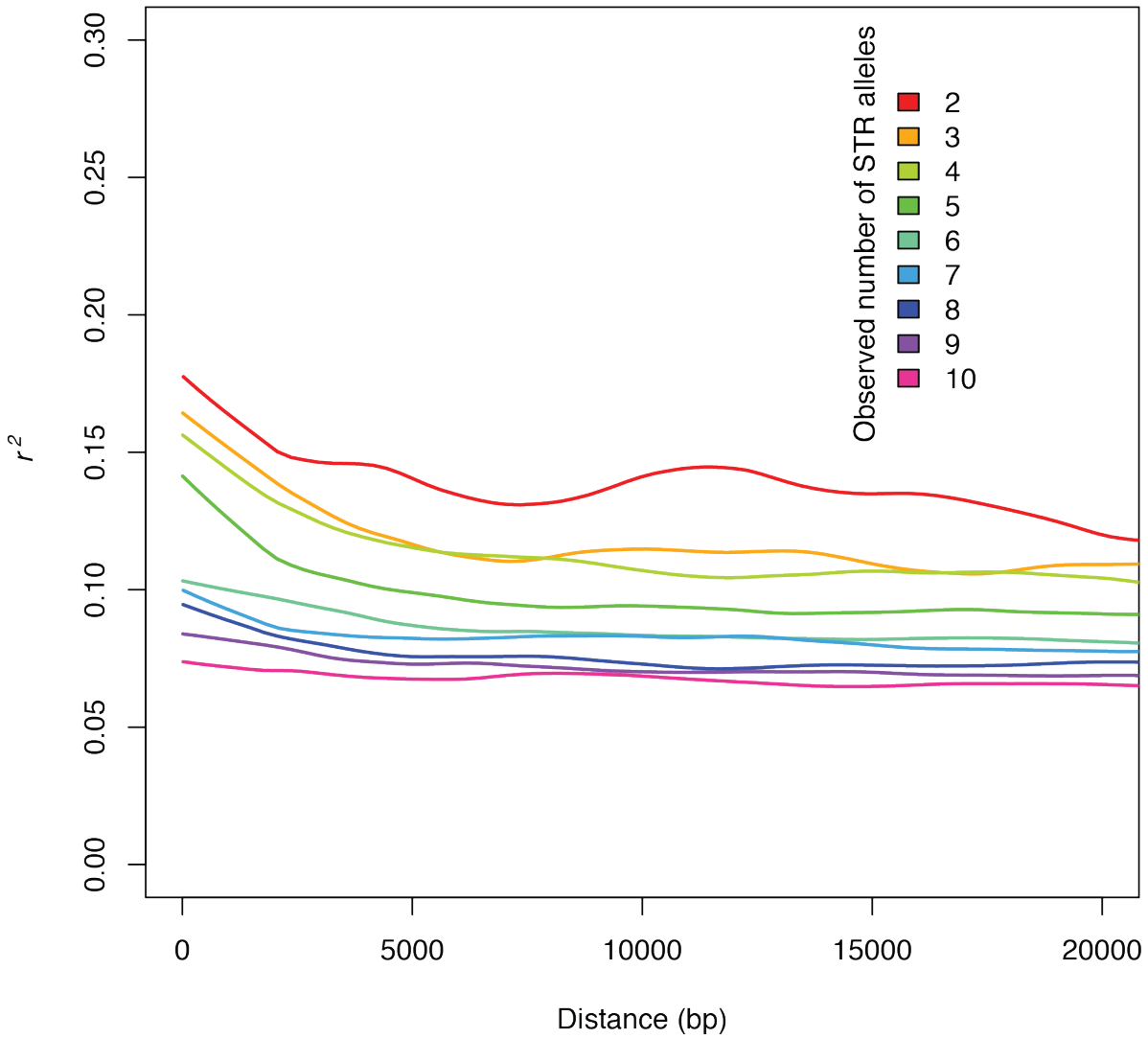
Supplemental Fig. S10. Relationships between SVR model features, response variable (Log(SD)), conservation score (Z-score), and STR annotations. Each feature of the final bagged SVR model is plotted against each other feature across all STRs, the response variable (Log(SD)), STR annotation, and the final conservation score prediction.



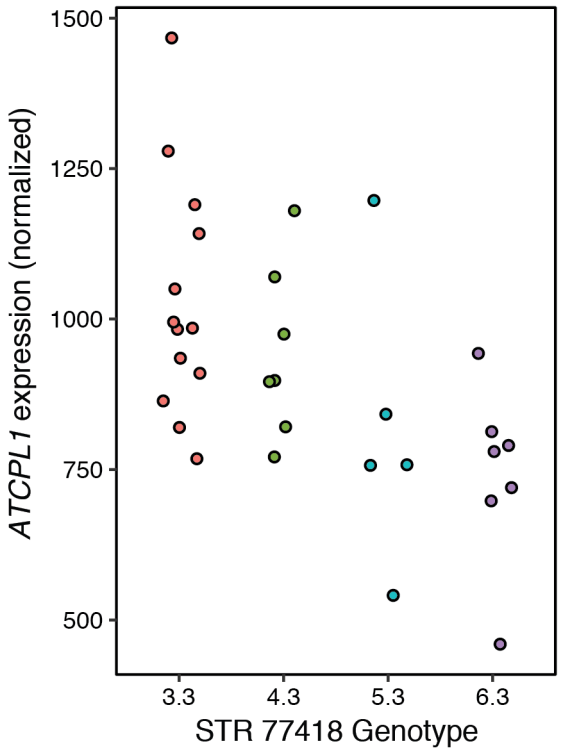
Supplemental Fig. S11. Number of STRs showing non-neutral variation, stratified by locus category and unit motif. White box-shaped points indicate the expected numbers for each bar, based on number of STRs in each locus category and number of STRs under different types of non-neutral behavior (conserved: white, hypervariable: black). STR motifs are collapsed into equivalent motif categories by reverse complementation and motif frame shifting (e.g. CTG = GCT = AGC).



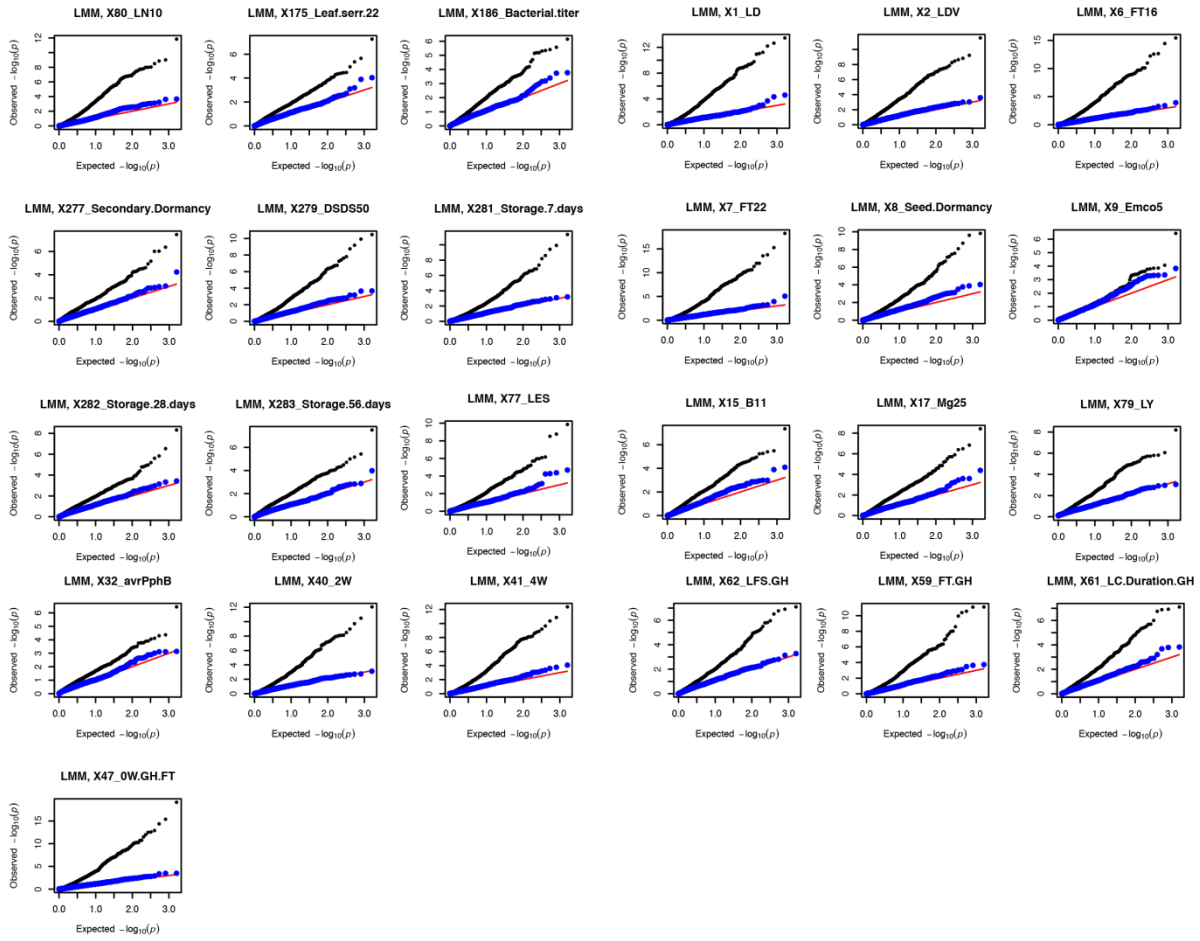
Supplemental Figure S12. Little effect of transposable element annotation on STR conservation. TE: Transposable element, Not TE: not a transposable element. Note that high overlap (>200bp) with TEs leads to classification as intergenic (e.g. noncoding).



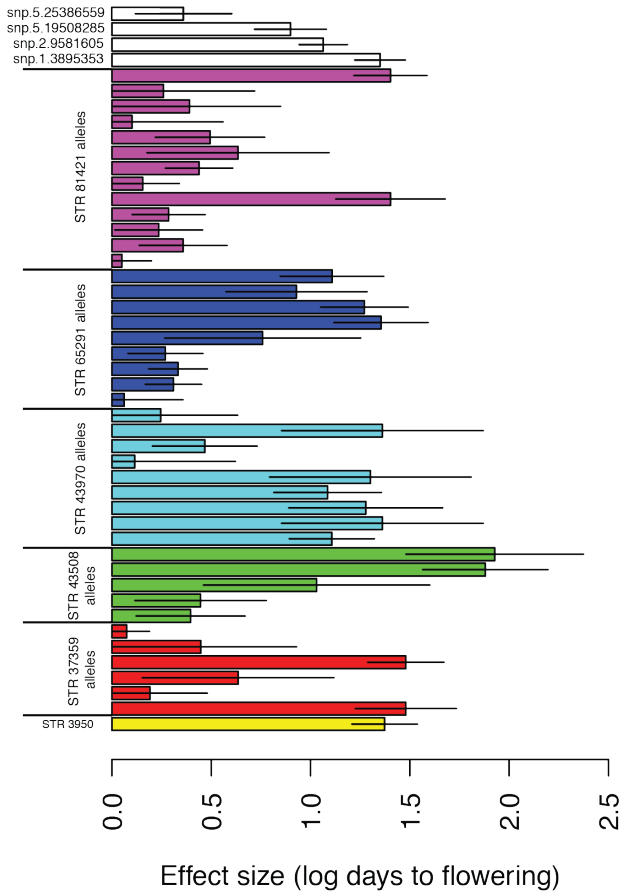
Supplemental Fig. S13. Linkage disequilibrium is dependent on the number of alleles at an STR locus. Lines of different colors represent lowess estimates of the relationship between distance between LD (measured by r^2) and genomic distance for STR/SNP locus pairs of a certain STR allele count. For visualization purposes, r^2 values lower than 0.05 were omitted from lowess estimation. Note that scales are different from Fig. 6c.



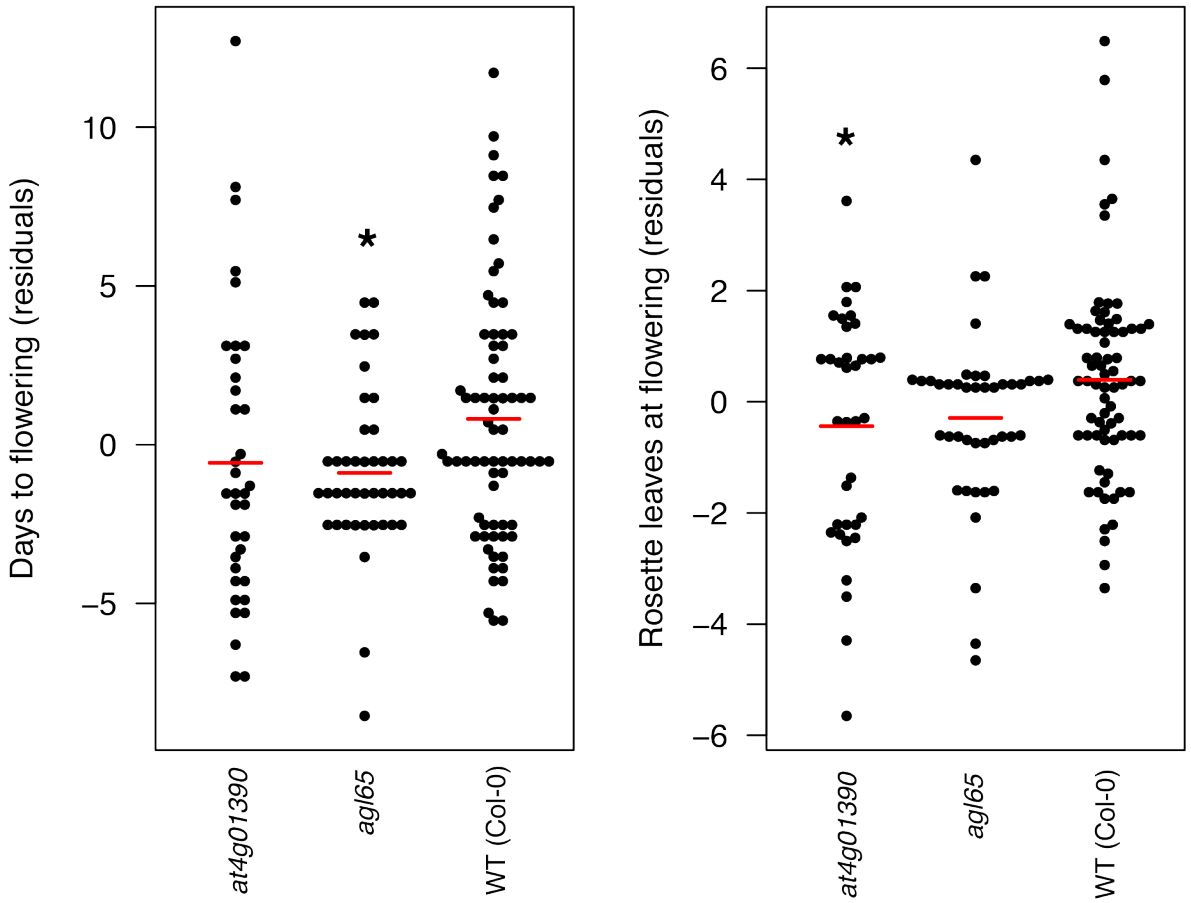
Supplemental Fig. S14. A strong STR-mediated eQTL. For a full list of associations, see Supplemental Table S4.



Supplemental Fig. S15. Q-Q plots of p-values for association tests between STR genotype and phenotypes showing associations at $p < 10^{-6}$ or $p < 10^{-10}$ (flowering traits), while correcting for population structure in a linear mixed model. Black points indicate results from true genotypes, blue points are identical but with genotypes permuted (negative control tests). Red line indicates expected p-value distribution under the null hypothesis.



Supplemental Fig. S16. Relative effect sizes for the phenotype days to flowering in long days for associated STR and SNP loci. Fixed effects from a linear mixed model also fitting kinship among strains as a random effect. For a variance decomposition of this phenotype with respect to these loci, see Supplemental Table S8.



Supplemental Fig. S17. Validation of flowering time associations using insertion mutants. Residuals are estimated from a linear mixed model fitting experiment and growth tray position as random effects. Red horizontal bars indicate mean of residuals within each genotype. *: $p < 0.05$ of fixed effect for genotype on phenotype. Mutants are insertional T-DNA mutants in the Col-0 background.

Supplemental Tables

Supplemental Table S1. Performance of various BWA parameter sets on Col-0 (algorithm, open, extend, mismatch, alignment score threshold) by various metrics (proportion STRs typed, proportion STRs correct, correlation with correct alleles, overall accuracy).
Included in Supplemental XLSX file.

Supplemental Table S2. STRs showing modest expansions (expansion score >=2) in at least one strain.

ID	Number of units	Purity	Motif	Locus size	annotation	chr	start	stop	gene	Expansion Score
65400	6.7	100	CAA	20.1	intron	Chr4	672378	672397	AT4G01540	5.92592593
47883	10	100	AT	20	UTR	Chr3	5587549	5587568	AT3G16440	4.5
89068	10	100	TA	20	Intergenic	Chr5	9760426	9760445	AT5TE35475	4.33333333
2539	6.7	100	AGA	20.1	coding	Chr1	4261712	4261731	AT1G12490	3.73134328
5427	8.5	90	TATA	34	intron	Chr1	8541165	8541198	AT1G24145	3.6
19084	7.3	100	CTT	21.9	UTR	Chr1	21181668	21181689	AT1G56540	3.42465753
15746	7.3	100	AGA	21.9	coding	Chr1	17436847	17436868	AT1G47510	3.15068493
25619	9	91	TCT	27	UTR	Chr1	30082339	30082365	AT1G79970	3
7358	8.7	91	TTC	26.1	intron	Chr1	11087597	11087622	AT1G31070	2.98850575
414	13	94	GAA	39	UTR	Chr1	641971	642008	AT1G08815	2.77777778
95706	7	100	TCA	21	UTR	Chr5	15004737	15004757	AT5G37780	2.75
41915	11	100	AT	22	intron	Chr2	15267011	15267032	AT2G36390	2.7
84110	6.7	100	AGG	20.1	UTR	Chr5	3433738	3433757	AT5G10880	2.68656716
45985	10	100	TC	20	Intergenic	Chr3	2321443	2321462	NA	2.66666667
77933	8	100	TGA	24	UTR	Chr4	12336368	12336391	AT4G23680	2.625
32842	11.3	100	GAA	33.9	Intergenic	Chr2	5490658	5490691	NA	2.56637168
7150	12.5	100	AG	25	intron	Chr1	10817051	10817075	AT1G30540	2.51851852
100082	6.7	94	CTCCGG	40.2	coding	Chr5	19584200	19584239	AT5G48320	2.45614035
35201	8	100	ACC	24	coding	Chr2	7039826	7039849	AT2G16250	2.375
5792	14.5	92	AG	29	intron	Chr1	9023890	9023918	AT1G26100	2.35294118
33137	8.7	91	TCT	26.1	coding	Chr2	5736639	5736664	AT2G13770	2.29885058
97752	8.3	95	CAC	24.9	coding	Chr5	16915919	16915944	AT5G42310	2.28915663
74375	6	100	CAAGAA	36	UTR	Chr4	7654240	7654275	AT4G13160	2.2
103981	8.3	100	ATG	24.9	Intergenic	Chr5	25146813	25146837	NA	2.1686747
81607	7	100	AGA	21	UTR	Chr4	18243655	18243675	AT4G39170	2.16666667
43048	12.3	94	TTC	36.9	UTR	Chr2	16925770	16925806	AT2G40520	2.15053763
96928	9	91	CAT	27	Intergenic	Chr5	16009433	16009459	NA	2.11111111
47029	18	94	CT	36	UTR	Chr3	4136084	4136119	AT3G12955	2

Supplemental Table S3. Frequency of coding STR amino acid composition compared to *A. thaliana* proteome (Karlin et al. 2002).

Amino Acid	Proteome frequency[1]	count	Purifying frequency[1]	Standard error	Fold enrichment
A	0.063	0	0	0	0
C	0.001	0	0	0	0
D	0.084	9	0.142857143	0.044086671	1.700680272
E	0.158	18	0.285714286	0.056915648	1.808318264
F	0.012	1	0.015873016	0.015746536	1.322751323
G	0.11	5	0.079365079	0.034055572	0.721500722
H	0.027	2	0.031746032	0.022088646	1.175778954
I	0.001	1	0.015873016	0.015746536	15.87301587
K	0.072	7	0.111111111	0.039594258	1.543209877
L	0.052	1	0.015873016	0.015746536	0.305250305
M	0.004	1	0.015873016	0.015746536	3.968253968
N	0.05	4	0.063492063	0.030721695	1.26984127
P	0.104	1	0.015873016	0.015746536	0.152625153
Q	0.058	1	0.015873016	0.015746536	0.273672687
R	0.021	0	0	0	NA
S	0.334	8	0.126984127	0.041948411	0.380191997
T	0.035	1	0.015873016	0.015746536	0.453514739
V	0.011	1	0.015873016	0.015746536	1.443001443
W	0	0	0	0	NA
Y	0	0	0	0	NA

[1]: Proportion of proteins with an amino acid run with at least one of the indicated type (does not sum to 1).

Supplemental Table S4. STR-expression associations in *cis* (FDR < 0.5 for 1MB STRs, 0.1 for others).

STR	gene	beta	t-stat	p-value	FDR
STRs within 1MB of affected genes					
77418	AT4G21670	-0.726600648	-6.505786818	2.43E-06	0.202737303
80659	AT4G33120	0.862123883	6.291530503	3.84E-06	0.202737303
100562	AT5G48090	-1.378327481	-5.956905693	7.96E-06	0.251368899
77418	AT4G21860	0.772661033	5.78392224	1.17E-05	0.251368899
83137	AT5G06660	-0.904906529	-5.75807398	1.24E-05	0.251368899
99144	AT5G44565	-1.766134236	-5.672178594	1.50E-05	0.251368899
99144	AT5G47770	2.021168016	5.624181318	1.67E-05	0.251368899
85241	AT5G18860	0.898198603	5.545078054	1.99E-05	0.262642511
43830	AT2G43950	-1.909967903	-5.226353678	4.09E-05	0.450427951
1840	AT1G11350	-1.60739376	-5.140770441	4.98E-05	0.450427951
82895	AT5G03770	-2.568724952	-5.130318095	5.10E-05	0.450427951
21447	AT1G62640	-0.608653253	-5.09972821	5.47E-05	0.450427951
104448	AT5G62540	-0.989527092	-5.093510359	5.55E-05	0.450427951
39206	AT2G28130	-2.029925926	-5.050530478	6.12E-05	0.461566352
46518	AT3G10640	0.914437602	5.004009256	6.81E-05	0.479364022
84539	AT5G10946	1.513238091	4.973511699	7.30E-05	0.482050245
STRs within 100KB of affected genes					
STR_77418	AT4G21670	-0.730803	-6.827481	9.47E-07	0.06755988
STRs within 10KB of affected genes					
STR_21123	AT1G63880	1.2084067	6.161354	4.11E-06	0.03389828
STR_83137	AT5G06660	-0.9154243	-5.694793	1.19E-05	0.04893082
STR_8109	AT1G32583	-0.8341916	-5.286204	3.06E-05	0.08414067

STR_43830	AT2G43950	-1.883862	-5.147784	4.23E-05	0.08729684
STRs within 1KB of affected genes					
STR_83137	AT5G06660	-0.9154243	-5.694793	1.19E-05	0.02214102
STR_41951	AT2G36550	-0.5829935	-4.765462	1.04E-04	0.06673268
STR_73034	AT4G10310	-0.9215156	-4.589755	1.59E-04	0.06673268
STR_44826	AT3G01790	0.3455633	4.548121	1.75E-04	0.06673268
STR_1354	AT1G07025	-0.2752847	-4.440702	2.27E-04	0.06673268
STR_77555	AT4G22280	2.1427839	4.428316	2.33E-04	0.06673268
STR_44903	AT3G02390	1.3386234	4.368212	2.69E-04	0.06673268
STR_79636	AT4G30910	0.2047619	4.343241	2.86E-04	0.06673268
STR_83557	AT5G08020	-1.4256175	-4.280532	3.32E-04	0.06891407
STR_44048	AT2G45170	-0.4441011	-4.1577	4.46E-04	0.07590795
STR_63630	AT3G58020	1.0055077	4.156214	4.47E-04	0.07590795

Supplemental Table S5. STR-phenotype associations and with relevant information.
Included in Supplemental XLSX file.

Supplemental Table S6. STR-phenotype associations are robust to adjustment for SNPs.
Included in Supplemental XLSX file.

Supplemental Table S7. Analysis of variance in the phenotype of days to flowering (long days) with respect to associated SNP and STR loci. Sums of squares represent marginal contributions ("type II").

Locus	Sum of Squares	DF	F-value	p-value
snp.1.3895353	1153.6	1	2.353	0.1316
snp.2.9581605	508.0	1	1.0362	0.3138
snp.5.19508285	191.7	1	0.391	0.5347
snp.5.25386559	39.5	1	0.0805	0.7778
STR 37359	1149.9	6	0.3909	0.8813
STR 3950	15.4	1	0.0314	0.8602
STR 43058	1338.2	5	0.5459	0.7406
STR 43970	3336.3	7	0.9721	0.4622
STR 65291	6233.0	8	1.5892	0.1531
STR 81421	4313.7	12	0.7332	0.7123
Residuals	23532.8	48	-	-

Supplemental Table S8. Sequencing libraries used in this study.
Included in Supplemental XLSX file.

Supplemental Table S9. Details of sequencing runs used in this study.

Sequencing run	instrument, kit	demultiplexed reads	libraries	notes
Pilot	MiSeq v2, 300 cycle	15008516	5	
Full experiment	NextSeq 500 v2, 300 cycle High Output	137132721	96	flow cell malfunction, lost ~60% of reads in demultiplexing

Supplemental Table S10. Primers used in this study.
Included in Supplemental XLSX file.

Supplemental Table S11. *A. thaliana* mutants used in this study.

ID	gene	alias	note	reference
SALK_113736	*AT4G01390*	*at4g101390*	homozygote isolated from segregants	this study
CS66521	*AGL65*	*agl65-1*	from SALK_009651	Adamczyk and Fernandez (2009) *Plant Phys.*

Supplemental References

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **0**. http://www.cell.com/article/S0092867416306675/abstract (Accessed June 27, 2016).

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**: 627–31.

Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J. 2014. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**: 2670–2672.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics* **43**: 956–63.

Carlson KD, Sudmant PH, Press MO, Eichler EE, Shendure J, Queitsch C. 2015. MIPSTR: A method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Research* **125**: 750–761.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608–611.

Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J* **89**: 789–804.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**: 419–23.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, et al. 2015. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics* **48**: 22–29.

Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. 2013. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome research* **23**: 843–54.

Huntley MA, Clark AG. 2007. Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Mol Biol Evol* **24**: 2598–2609.

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 333–8.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–8.

Meyer PE. 2014. *infotheo: Information-Theoretic Measures*. https://cran.r-project.org/web/packages/infotheo/index.html (Accessed January 11, 2017).

Press MO, Carlson KD, Queitsch C. 2014. The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics* **30**: 504–512.

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

Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al. 2014. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. *Cell Reports* **8**: 2015–30.