Evolutionary adaptability linked to length variation in short genic tandem repeats

William B Reinar[1*], Jonfinn B Knutsen[1*], Sissel Jentoft[1], Ole K Tørresen[1], Melinka A Butenko[1] and Kjetill S Jakobsen[1†]

[1]University of Oslo, Department of Biosciences, Norway

*These authors contributed equally to this work

†Corresponding author

Abstract

There is increasing evidence that short tandem repeats (STRs) – mutational hotspots present in genes and in intergenic regions throughout most genomes – may influence gene and protein function and consequently affect the phenotype of an organism. However, the overall importance of STRs and their standing genetic variation within a population, e.g. if and how they facilitate evolutionary change and local adaptation, is still debated. Through genome-wide characterization of STRs in over a thousand wild *Arabidopsis thaliana* accessions we demonstrate that STRs display significant variation in length across the species' geographical distribution. We find that length variants are correlated with environmental conditions, key adaptive phenotypic traits as well as gene expression levels. Further, we show that coding STRs are overrepresented in putative protein interaction sites. Taken together, our results suggest that these hypervariable loci play a major role in facilitating adaptation in plants, and due to the ubiquitous presence of STRs throughout the tree of life, similar roles in other organisms are likely.

Introduction

Short tandem repeats (STRs), often defined as units of 1-6 base pairs repeated in tandem, are present in genes and in intergenic regions throughout most genomes. STR mutations arising due to replication slippage either lead to a reduction or an increase in the number of repeated units, and have estimated mutation rates of 1×10^{-4} to 1×10^{-3} per locus[1], which are orders of magnitude greater than point mutations. Length variations can have dramatic phenotypic effects, mostly known from studies of human disease[2]. For instance, the number of repeated CAG units in a STR within the coding part of the *huntingtin* gene correlates with the age of onset of Huntington's disease[3]. However, such dramatic consequences are likely exceptions. Studies of STRs in other organisms have shown that length variations in certain STRs can modulate gene and protein function without inducing deleterious effects[4,5]. It has been proposed that STRs associated with certain genes may serve as 'tuning knobs' that facilitate adaptation[6]. Support for this is found in studies of selected STRs in wheat and barley populations, suggesting a link between STR diversity, and ecological factors, such as drought[7–9]. Nevertheless, the overall contributory effect of STR variation to local adaptation has not yet been addressed using genome-wide data from a larger collection of samples representing the global distribution of a species.

The small weed *Arabidopsis thaliana* has proven to be a unique model system to study and understand mechanisms underlying plant development, stress response, plasticity and adaptive responses in confined environments. Additionally, due to recent whole genome sequencing initiatives and transcriptome studies, *A. thaliana* has rapidly emerged as a powerful ecological and evolutionary study system that provides insight into population genomics, intra-specific genome evolution and how wild populations respond to biotic and abiotic conditions in their local environments[10–12]. Despite the fast-growing numbers of such studies, the biological importance of genetic variation caused by length variation in STRs has been more or less overlooked. One exception is a recent study by Press et al.[13], where close to 2000 STR loci was

explored among 97 *A. thaliana* strains and was found to have functional importance. For instance, a STR length expansion in the 3'UTR of *MEE36* caused retention of its intron and was associated with reduced expression levels. Hence, there is a need to fully characterize the length variation of STRs in different *A. thaliana* populations at the whole-genome level. If STRs provide adaptive benefits, correlations between STR length variants and the environment associated with the plants geographical origin should be evident. To explore this avenue, we analyzed whole-genome data from more than a thousand *A. thaliana* natural specimens (accessions), sequenced by the 1001 Genomes Consortium[10]. The investigated accessions have been collected from a variety of habitats – from the accession *Cvi* collected in Cape Verde, to *Strand-1* collected in Northern Norway – and thus have adapted to very different local biotic and abiotic conditions. Hence, the sample set provides an unprecedented opportunity to study STR variation in light of local adaptation. To our knowledge, we here present the first population-scale whole-genome STR profiling of a non-human organism.

Our results demonstrate that more than half of mononucleotide, dinucleotide and trinucleotide STR loci identified in wild *A. thaliana* accessions differed in length throughout its geographical distribution, and that the length variation in most cases was significantly correlated to specific environmental conditions. Of these, almost one third of the variants are located in the vicinity or within genes. Strikingly, close to 80 % of STRs located within protein coding sequence had length variation significantly associated with bioclimatic variables. We show that the coding STRs tend to overlap with putative protein binding sites, indicating a functional role for STRs in protein-protein and protein-DNA interactions. Moreover, we found that STR lengths co-varied with gene expression levels, and with variation in key adaptive phenotypic traits, such as the timing of flowering. Taken together, our results suggest that particular STR length variants provide advantages under certain biotic or abiotic conditions, and thus play a major role in facilitating adaptation. In a wider perspective, this first whole-genome population STR-profiling of a non-human organism provides a framework for similar analyses of adaptation-driving mechanisms in other organisms.

Results

Extensive STRs length variation in wild A. thaliana accessions

In order to characterize the simple repeats in *A. thaliana* we investigated the variation in 18835 mono-, di- and tri-STRs in 1041 *A. thaliana wild* accessions representing to a large extent its distribution in the Northern Hemisphere (Figure 1a). The alignment of Illumina short reads from the accessions to the *A. thaliana* reference genome (the accession Columbia-0) revealed massive variation, which is still a likely underestimation, as the sequencing reads probably were too short to capture the longest variants. Furthermore, insufficient sequencing coverage for some accessions made it impossible to variant call all STR loci in all accessions, and some accessions did not have reads compatible with the alignment tool. Nevertheless, 11665 STR loci (65.1 %) were determined to be variable, based on the frequency of the major allele (MAF ≤ 0.9, number of accessions variant called ≥ 25). The percentage of variable STRs was strongly affected by the genomic context and whether the STR was a mono-STR, di-STR or tri-STR (Figure 1d). In general, mono- and di-STRs displayed a high number of STR variants, with median allele numbers of 10 and 11, respectively. The number of tri-STR alleles was found to be much lower, with a median of 4 (Figure 1c). Compared with SNPs, the vastly different allele frequency spectra of STRs is illustrative of the added genetic diversity one should take into account to grasp the complete genetic variation in a population (Figure 1b). Notably, the major allele frequency distribution for tri-STRs displays a long tail, as a considerable amount of tri-STR loci (n = 2138) was deemed to be variable in the total population, of which 758 were located in coding DNA. In introns, promoters, 3'UTRs and intergenic regions, variable STRs make up the largest group, whereas non-variable STRs are primarily found in 5'UTRs and coding DNA sequence. Further, we found that tri-STRs present in coding sequence encode amino acid homopolymer tracts of mostly glutamic acids (E) and serines (S), while high fractions of variable STRs were found in phenylalanine (F), aspargine (N) and glutamine (Q) tracts, with 52 %, 48 % and 38 % variable loci (Figure 1e). Interestingly, it has been shown that proteins with certain

homopolymer tracts, such as Q, E and F tracts, have higher than average connectivity in protein networks, indicating that such sites are functional mediators of interactions, either interacting directly or by facilitating the interactions between domains[14].

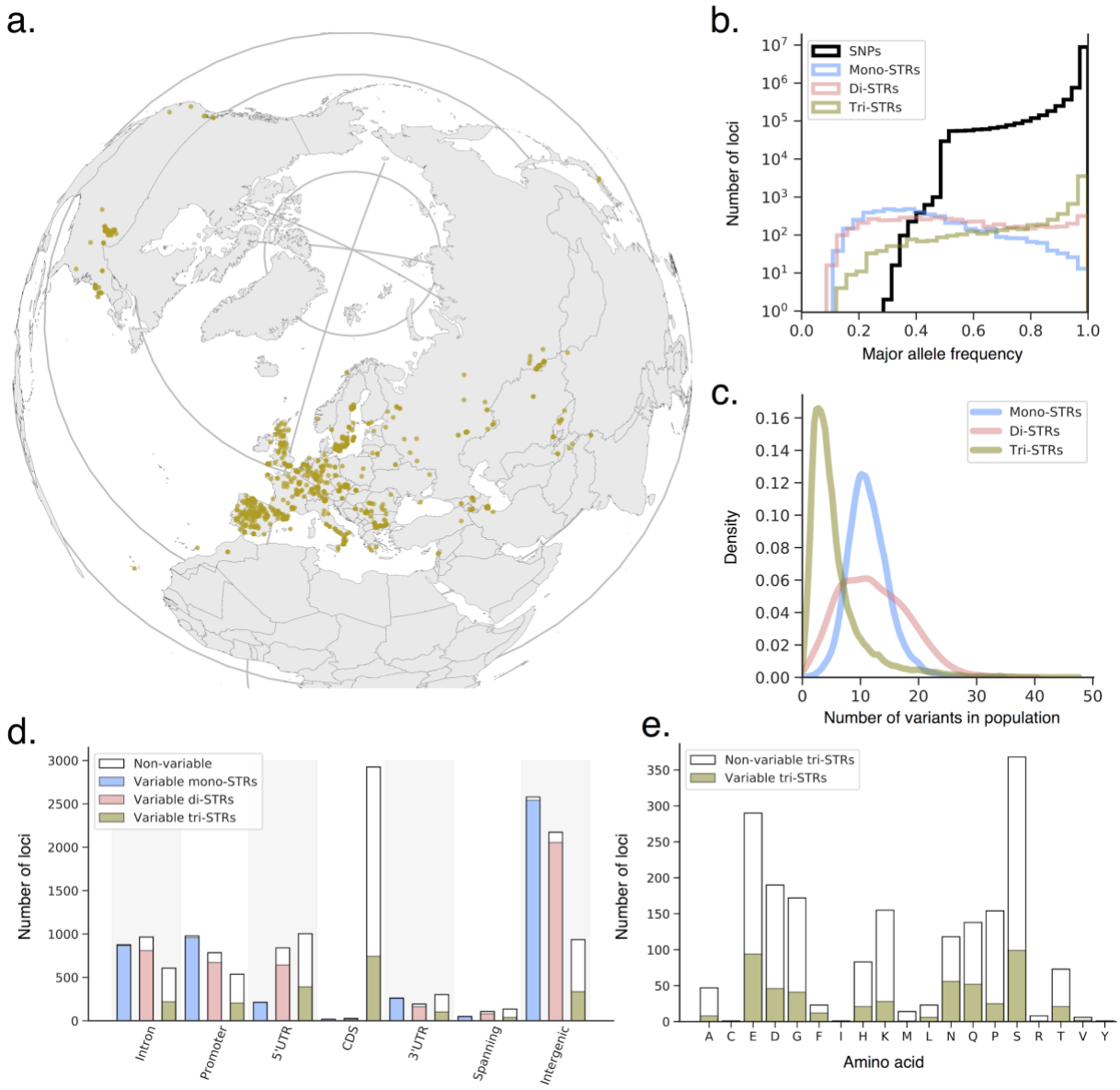


Figure 1. Mono-, di- and tri-STR length variation in the *A. thaliana* population. **a.** The distribution of sampled accessions. **b.** Distribution plot showing the major allele frequencies for SNPs and STRs. Note that the y-axis is $\log_{10}$ scaled. **c.** Kernel density estimate plot showing the number of variants for mono-, di-, and trinucleotide STRs. **d.** Barplot showing the number of variable and non-variable STRs in their genomic context. We defined variable STR loci as those with a major allele frequency less than 0.9 in the surveyed population. **e.** Barplot showing the number of variable and non-variable trinucleotide STR loci grouped by the homopolymer tract of amino acids that they encode.

Most STR loci are correlated with environmental conditions

As a next step, we addressed associations between STR variants and the markedly different environments where the accessions have been sampled. This was investigated by testing if the combined length of both alleles of a STR loci (STR dosage) among different *A. thaliana* accessions was linearly correlated with the environmental variables representing their local habitat (Supplementary Figure 1). For this, we used linear mixed-effect models with population structure as a random effect, as described in the Methods section. The environmental variables consisted of 19 derivative measures of temperature, solar radiation, precipitation and humidity. After Benjamini-Hochberg correction for multiple testing, 64.3 % of the loci yielded significant associations with at least one of the 19 environmental variables. To evaluate the test results, we plotted the theoretical *p* value distribution, the *p* value distribution of a control where the STR lengths were shuffled among accessions, and the observed *p* value distribution. The resulting quantile-quantile (QQ) plot shows that there is strong *p* value inflation in the observed data compared to the control and the theoretical *p* values, and thus statistical support for rejecting the null hypothesis of no association (Figure 2a). The QQ plots for each environmental variable show that 'temperature seasonality' – a measure of fluctuation in temperature within a year – has the strongest deviation between the observed *p* value distribution and the expected distribution (Supplementary Figure 2). To test if genes with habitat-associated STRs were enriched in particular functions, we performed gene ontology (GO) enrichment analysis and KEGG pathway enrichment analysis. DNA binding, protein binding and protein kinase activity were found to be significantly enriched molecular functions. Of biological processes, the genes tend to be associated with plant development and hormonal responses (Figure 2b). Notably, over 150 of the genes are involved in biosynthesis of secondary metabolites, compounds that mainly serve as defense against biotic stress, such as predation.

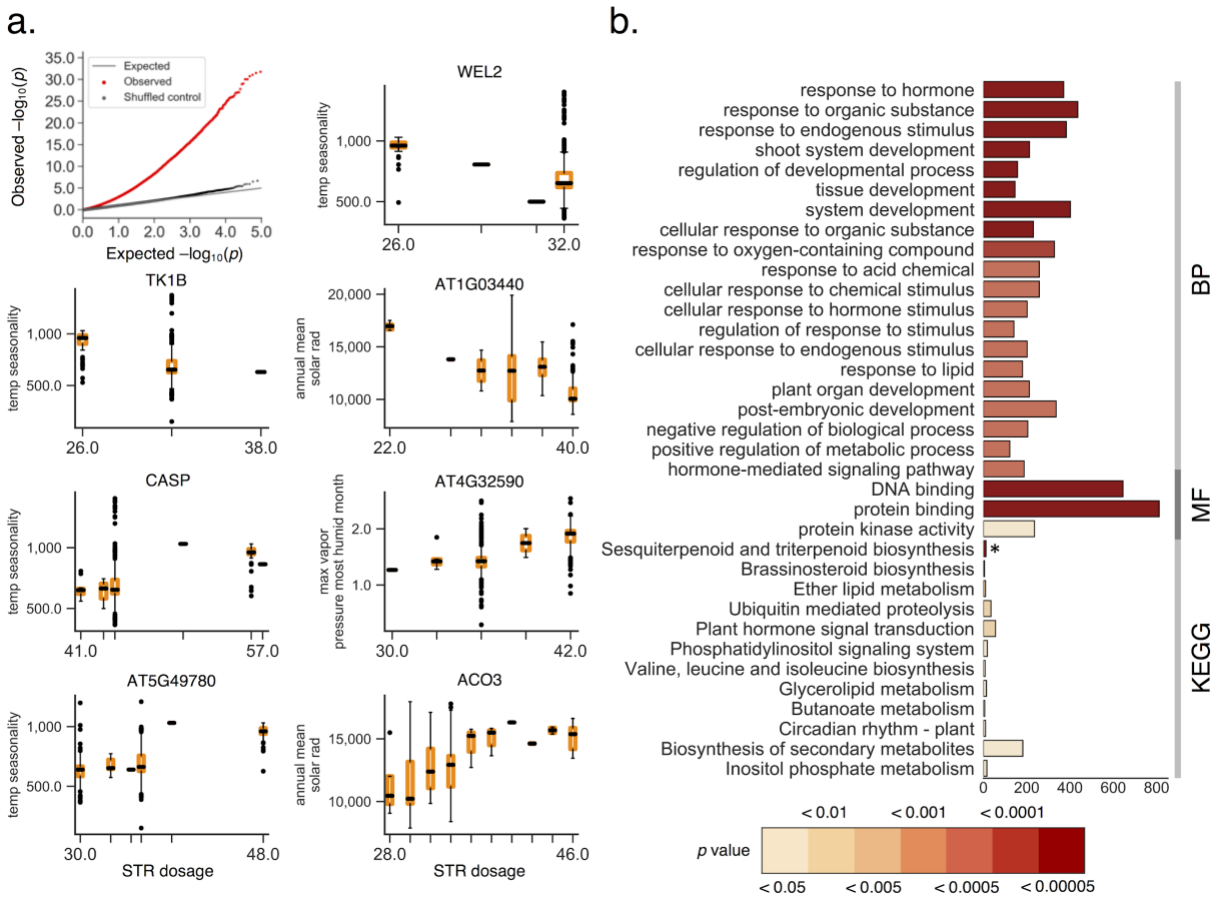


Figure 2. STR length variation is linked to habitat. **a.** Top left: Quantile-quantile (QQ) plot of genome-wide association tests modelling the length of a STR as a response to differences in habitat. The red line show observed *p* values, the gray dots are the *p* values of shuffled controls and the line shows the expected *p* values distribution based on the number of tests we performed. The boxplots show associations between STR dosage and environmental variables for the top eight STR-habitat associations of genic STRs. WEL2 is involved in chloroplast movement to optimize photosynthesis, TK1B is a thymidine kinase, AT1G03440 is a Leucine-rich repeat (LRR) family protein, CASP may be involved in intra-Golgi transport, AT4G32590 is a ferredoxin-like protein, AT5G49780 is a Leucine-rich repeat (LRR) family protein, ACO3 is a mitochondrial aconitate hydratase and contributes to oxidative stress tolerance. In the y-axis, temperature seasonality is given as the standard deviation of yearly C° * 100, solar radiation is in MJ m^{-2} and vapor pressure is in kPa. **b.** Gene Ontology (GO) and KEGG pathway enrichment of genes with habitat associated STRs. BP: Biological Process, MF: Molecular Function. The asterisk denote that 'Sesquiterpenoid and triterpenoid biosynthesis' was the only KEGG pathway remaining statistically significant after correction for multiple testing. The x-axis indicate the number of genes linked to the respective GO term or KEGG pathway. For biological processes, the top 20 enriched terms sorted by *p* value is shown. To avoid the most general terms, only BP terms linked to less than 500 genes are shown. Redundant MF terms are not shown. The full list of terms is available in the Supplementary Data.

Protein coding STRs overlaps with predicted functional sites

To further investigate the extent of biological connectivity among genes with habitat-associated STRs in coding sequence, we tested for enrichment in physical interactions between the proteins these genes encode. We found that the proteins tend to structurally interact more with each other than what would be expected by chance, when compared with the known *A. thaliana* interactome (enrichment test performed *via* STRING v10.5 webservice, *p* value: $1.05.e^{-14}$). Surprisingly, almost 80 % of STRs coding for amino acids had length variation that correlated with habitat. To explore if such STRs could have a functional effect, we investigated where they are located in relation to protein structure. We found that the distribution of STR positions in proteins resembled a right-skewed inverted bell curve, indicating a selective location in protein termini (Supplementary Figure 3). Further, we detected a mild overrepresentation of STRs in signal and transit peptide sequences, as annotated by the UniProtKB/Swiss-Prot consortium (Supplementary Figure 3). In domains, we found STRs to be underrepresented, though with a notable presence in protein kinase domains (Supplementary Figure 3). Tri-STRs encode amino acid homopolymers, which have been reported to accumulate in disordered protein regions that do not form a stable structure[15]. As disordered regions can play an important role in protein interactions[16,17], we tested for enrichment of STRs in predicted protein-protein, protein-DNA and protein-RNA interaction sites (PPIs, DPIs and RPIs) within disordered regions. Two of the PPI prediction tools we used, DisoRDPbind and ANCHOR, detect unstable regions that are likely to be stabilized if bound to a globular protein. The third tool, MoRFpred, detects shorter stretches likely to be stabilized by binding to a protein partner. Although disagreeing on the degree of statistical enrichment, DisoRDPbind and ANCHOR agreed on 214 PPI sites that had an overlapping STR, and MoRFpred found 90 such sites (Figure 3b). One of the sites was found in the protein ELF3, where STR length variation has previously been shown to impact the connectivity with interaction partners[18], leaving support to our approach. STRs were statistically enriched in predicted DNA-protein interaction sites, but not in RNA-protein binding sites. In addition, we focused on disordered flexible linkers (DFLs), as such regions can fine-tune interactions or the distance between protein domains[17,19]. The results of DFLpred revealed a

1.47 fold enrichment of STRs in DFLs. Taken together, these results support that length variation in coding STRs can tune protein function by altering protein-protein and protein-DNA interactions, or affect the structure of flexible linkers within proteins. To further explore the extent of protein-protein interactions among proteins with STR-encoded homopolymer tracts where length variation were associated with habitat, we used geneMANIA to find the largest coherent network of these proteins and interaction partners (Figure 3a). The interaction network consists of numerous clusters of proteins, often involved in the same processes, such as chromatin remodelling, developmental regulation and biotic stress. We propose that coding STR length variation affecting interactions within such clusters could lead to changes on the phenotypic level that under certain environmental conditions provide selectional advantages.

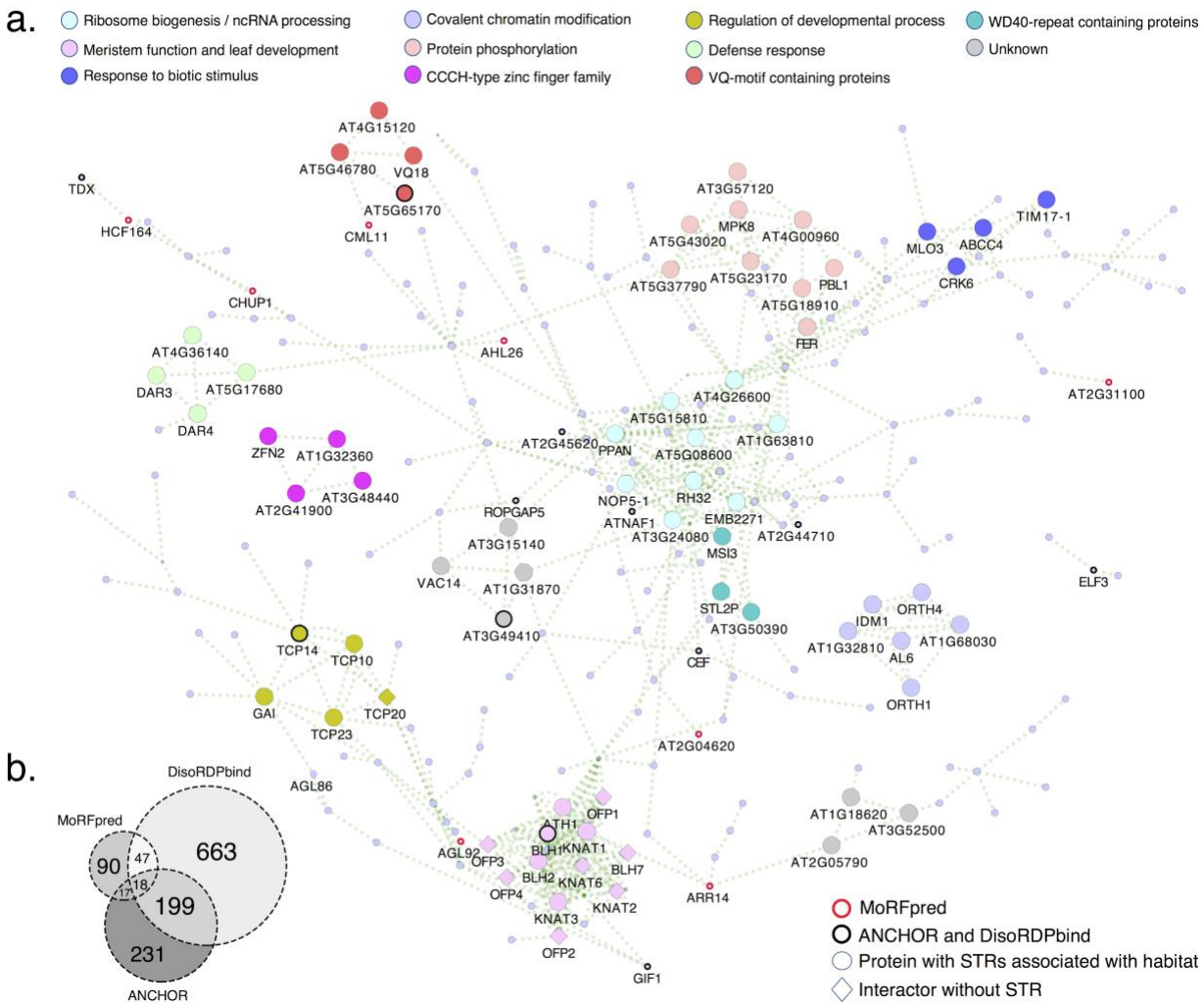


Figure 3. Interaction network of proteins with coding STR length variation associated with habitat. **a.** The largest protein-protein interaction network of proteins encoded by genes with STRs in coding sequence (circles), that are correlated with habitat. Close interaction partners, not necessarily with a STR, have a diamond shape. Proteins with a STR overlapping a predicted protein-protein interaction site, as agreed upon by the prediction tools DisoRDPbind and ANCHOR have a thick outline, or a red outline if predicted by MoRFpred. Densely connected subnetworks, as reported by the MCODE clustering algorithm, are colored and have increased sizes. The top legend gives information on biological processes or other shared characteristics of the clusters, based on manual assertion *via* ThaleMine. **b.** VENN diagram showing the number of overlaps between STRs and protein-protein interaction sites using three different prediction tools.

Hundreds of STRs associated with gene expression

To further address functional aspects of STR variation, we tested if differences in STR lengths influence gene expression. For this, we used RNA-sequencing data from rosette leaf tissue gathered by Kawakatus et al.[11]. STRs were required to be no more than 500 bp distant from the target gene, which should capture most genic STRs. As for the tests of covariation with environmental variables, the deviation between observed *p* value distribution and expected *p* values given no association was high, evident from the QQ plot (Figure 4a). 399 cis-STRs had length variants co-varying with gene expression values (henceforth termed eSTRs). Most eSTRs were localized inside the transcript, i.e. in CDS, introns and UTRs (Figure 3b). There was no statistically significant bias in the direction of effect between STR dosage and gene expression (χ^2 test: p = 0.073). This result supports a role for STRs in regulation of gene expression by length variation, as shown in humans by Gymrek et al.[20] through analyses of human lymphoblast transcriptomes[20].

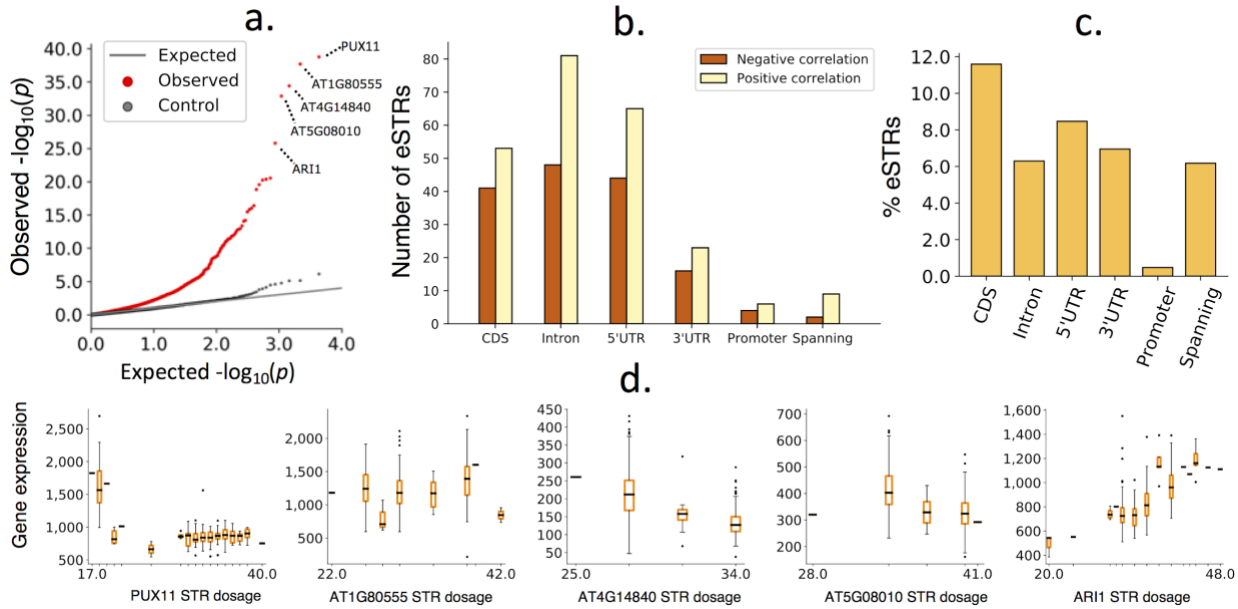


Figure 4. Associations between STRs and gene expression. **a.** QQ plot of eQTL results with 500 bp as maximum distance between STR and target gene. **b.** Direction of effect as predicted by linear models, grouped by genomic context. **c.** The percentage of significant associations for each STR category, grouped by genomic context. **d**. Boxplots showing associations between outliers labeled in a. PUX11 likely functions in the ubiquitin conjugation pathway, AT1G80555 is a isocitrate/isopropylmalate dehydrogenase, AT4G14840 and AT5G08010 are uncharacterized, and ARI1 might act as a E3 ubiquitin ligase.

STR lengths covary with phenotypic traits

There is a high degree of phenotypic variation among the *A. thaliana* accessions, and we hypothesized that some of this variation can be attributed to STR length variation. Utilizing experimental data deposited in a public database of *A. thaliana* phenotypes (https://arapheno.1001genomes.org), we regressed 225 measurements of phenotypic traits on STR lengths. The phenotypic traits can be broadly categorized as measurements of metabolite concentration, ion concentration in various tissues, flowering time (FT), developmental traits and disease resistance. Strikingly, we found that STR length variation within 1202 genes was correlated with measurements of when the accessions flower. QQ plots with the results of genome-wide associations tests with flowering time and two other traits are shown in Figure 5a. An example of a FT-associated STR loci includes a tri-STR present in the coding region of *AGL65*, a gene known to be required for pollen maturation[21] (Figure 5a-b, middle). Examples of associations with other traits include a correlation between the length of a STR in the 5'UTR of *WAX2*, a protein involved in cuticle wax production[22], and the number of aphids feeding on the plant (Figure 5a-b, right). Note that the latter association did not pass correction for multiple testing. Another example is the gene *IP5P7*, which functions in salt stress tolerance[23,24] and has intronic STR length variation correlated with the sodium concentration in leaves (Figure 5a-b, left). Taken together, these results support that STR length variation can facilitate adaptation to microhabitats, such as when to flower, how to cope with differing salt concentrations, and the presence of herbivores.

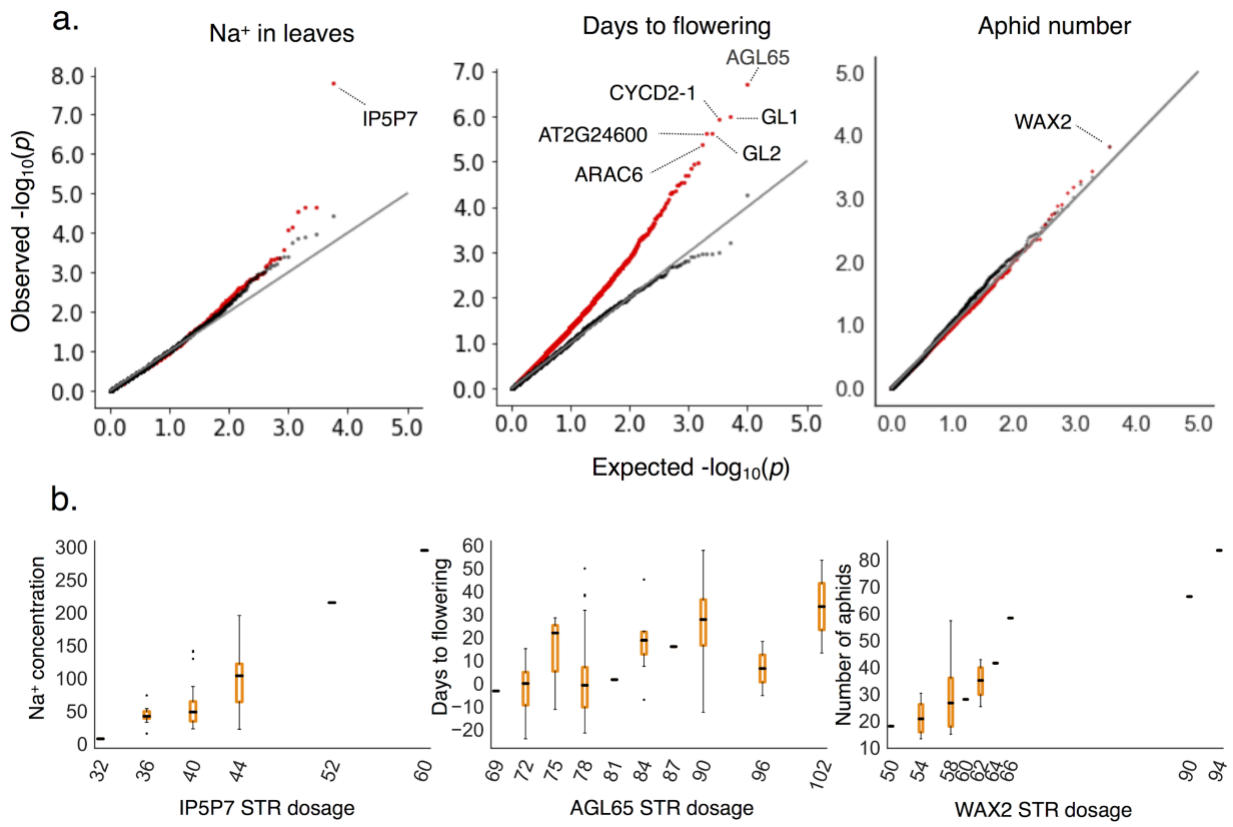


Figure 5. STRs and phenotypic variation. **a.** Example quantile-quantile (QQ) plots of *p* values resulting from association test with three different traits (of 225 performed in total). The top genes, based on the *p* value of the associated STR, are highlighted. IP5P7 functions in salt stress tolerance[23,24], AGL65 is known to be involved in pollen maturation[21], CYCD2-1 is a cell cycle protein[25], GL1 and GL2 are involved in trichome formation[26,27], AT2G24600 is an ankyrin repeat family protein of unknown function, ARAC6 may be involved in cell polarity control[28] and WAX2 is involved in cuticle wax production[22]. **b.** Boxplots showing the associations between STR dosage in the highlighted genes from the QQ plots and the tested traits.

Discussion

There is accumulating evidence that STRs length variability may impact gene regulation and protein function and thereby affect phenotypic variation. In turn, this is likely to provide rapid adaptation under specific selection regimes. Published examples on STR variation in association with specific genes range from adaptive phase variation in bacteria[29], control of flocculation in yeast[4], maintenance of the circadian rhythm in response to temperature in *Drosophila*[30], to swift evolution of limb and skull morphology in canids[31]. Here, we have used a genome-wide approach to investigate the extent of the evolutionary importance of STRs in a wild growing specimens of *A. thaliana* across its global distribution in the northern hemisphere. Our results demonstrate that genome-wide STR length variation throughout the genome is profoundly linked to the environmental conditions where the accessions grow. By utilizing the wealth of knowledge resulting from decades of functional studies with *A. thaliana* as a model plant, we find that the STRs tend to sit in genes involved in developmental regulation, defense against pathogens or plant hormone pathways, all crucial features for adapting to microhabitats.

In a broader perspective, the nature of STR length variation somewhat resemble that of epigenetic changes (epimutations), in that both STR length variations and epimutations occur at a rapid mode and are reversible. Both STR variation and epimutations may be involved in phenotypic plasticity-associated phenomena as well as local adaptation. However, the evolutionary impact of epigenetics is highly debated[32,33]. In contrast to epimutations, STR mutations acting at the DNA level are inheritably stable, and as our results demonstrate, might affect protein interactions as well as gene regulation. Thus, the functional impacts of STR variations by far exceed the regulatory effects of epigenetics. Based on our findings, we reason that rapidly mutating STRs – located in functionally important sites – dramatically expand possible phenotypic outcomes of protein-protein and protein-gene networks, including gene regulation, of which selection can act upon. Given the high STR length mutation rate, an important implication is that adaptation to novel habitats or to a fluctuating environment can

happen faster than what point mutations (SNPs) alone can permit. Another layer of complexity is added by STR-driven gene expression, as presented here and in other studies[13,20], mechanisms of which include altering binding of transcription factors, affecting RNA stability/translation, remodelling of chromatin and/or affecting DNA methylation patterns[34,35,36]. The recent report that STR expansions can affect gene splicing[13] may add an additional mechanism to how length variation potentially shapes the phenotypic landscape. Taken together, these results and the strikingly different allele frequency spectrum STRs display compared to that of SNPs, leads us to argue that to fully grasp the genetic variation in a population that can fuel adaptive evolution, it is paramount to incorporate STR length variation. Furthermore, our results support a shift in the current view of STRs as neutrally evolving markers that occasionally lead to disease. Contrary to this view, the link between STRs and local habitat, adaptive phenotypic traits and gene expression, coupled with the extensive overlap with predicted functional sites, suggests that STR length variation is a potent evolutionary force. As STRs are intrinsic features of genomes, we find it likely that these mutational hotspots drive adaptation not only in small weeds, but across the Tree of Life.

Methods

STR variant calling

We used lobSTR[37] (v. 3.0.2) to align the raw reads produced by the 1001 genome consortium[10] to the Arabidopsis TAIR10 (The Arabidopsis Information Resource 10) reference genome[38]. Tandem Repeats Finder[39] (TRF) was run on the reference genome with standard parameters, except that 'score' was set to 30. We limited our analysis to mono- di- and trinucleotide STRs. The TRF output (.dat) reported the genomic coordinates of each STR. Note that STRs reported by TRF might be interrupted. Increasing the 'score' parameter leads to higher purity in detected STRs, but will increase the length required to score the sequence as a STR. We did not want to rule out shorter repeats, as STRs with fewer units can be biologically interesting as well. As input to lobSTR we used all the raw reads submitted by the 1001 genome consortium to the SRA archive[6], extracting the FASTA sequences from the .sra files using the SRA toolkit (v. 2.7.0, fastq-dump, with parameters: --split-files, --fasta 0). We created a custom index of TAIR10, using lobSTR's Python (v.2.7.10) (Python Software Foundation, https://www.python.org/) scripts for this purpose (lobstr_index.py and GetSTRInfo.py). Consequently, reads from the accessions were aligned to the loci reported by TRF. The resulting BAM files were sorted using samtools (v. 1.3.1) and merged using bamtools (v. 2.3.0). The merged BAM files served as input for lobSTR's program 'allelotype', applying the PCR stutter noise model that follows the lobSTR distribution, which suits Illumina read data. The variant calling format (.vcf) file produced by 'allelotype' were indexed using tabix (v. 0.2.6), sorted using VCFtools[40] (v. 0.1.11) and annotated using BCFtools (part of SAMtools[41]) (v. 1.3). The annotation assigned a gene name to the row in the VCF file if the repeat was found within a gene (relying on the TAIR10 gene annotation GFF). We noted that STRs overlapping other STRs can lead to cases where 'allelotype' cannot determine the origin of the detected variation.

Data set filtering and subsetting

To minimize false variants that can arise due to sequencing errors, we only kept length variants supported by at least five reads. STR loci were divided into variable (major allele frequency, or MAF ≤ 0.9) and non-variable (MAF ≥ 0.9) STRs. Here we only included STRs with calls from at least 25 accessions, as this sample size should produce accurate MAF estimates of STRs[42]. For the association analyses, we did not require the sample size for each STR call to be ≥ 25, but to reduce the number of tests and thus increase power to detect true associations, we only analysed STRs with a MAF ≤ 0.9. We converted the TAIR10_GFF3_genes.gff retrieved from TAIR to a BED file and used pybedtools[43] (a wrapper of BEDTools[44]) to find overlaps between STRs and the features annotated in the GFF file. We required that the whole STR was contained within the feature. STRs not present in features covered in the GFF was defined as intergenic. STRs spanning multiple features (for instance, both a intergenic region and a 5'UTR) were designated as "Spanning". We defined STRs within 500 bp upstream of a gene as in being in that gene's promoter, based on estimates of *A. thaliana* promoter lengths from Korcuc et al.[45].

Protein feature analysis

We retrieved all reviewed *A. thaliana* protein data from UniProtKB/SwissProt[46] (downloaded the 7th of August 2017) and tested for overlap between TRF output run on coding DNA sequence from TAIR and the features in the GFF. To investigate whether or not STRs were enriched in intrinsically disordered regions (IDRs), a feature not assessed by UniProtKB/SwissProt curators, we downloaded TAIR10 proteome disorder predictions from d2p2[47] and tested for overlap between STRs and predicted IDRs. We used DisoRDPbind[48] to predict IDR interaction sites with regard to proteins, DNA and RNA. Intersections between STR loci and interaction sites were quantified using pybedtools. In addition to DisoRDPbind, we used ANCHOR[49] and MoRFpred to predict protein-protein interaction. ANCHOR detects unstable sites that are predicted to be stable when bound to globular partner. MoRFpred predicts short binding regions located within longer IDRs that bind protein partners *via* disorder-to-order transitions. For predicting disordered flexible linkers we used DFLpred[50]. We conducted a permutation test for each set of overlaps to identify if the number of overlaps

were significantly different from what is expected with random STR positioning. The 'randomstats'-module of pybedtools were used for this purpose and generates an estimated *p* value based on the chosen number of permutations (with n = 1000, the lower estimated *p* value limit is 0.001).

Population structure inference

In order to obtain the best group designation to correct for population structure, we conducted SNP-based and STR-based principal component analysis. In the STR-based PCA we used 1971 loci, where at least 70 % of accessions were variant called. Missing values were given the mean STR length. In the SNP-based PCA we used 50 000 random SNPs from the SNPeff analysis performed by the 1001 Genomes Consortium[10]. We assessed how well the STR-based PCA recapitulated the SNP-based PCA by visual examination (Supplementary Figure 5). We found that five groups, Italy/Balkan/Caucasus, Spain, Central Europe, Western Europe and 'admixed' formed a common cluster in the STR-based PCA, and merged these into one group. Our final six groups were composed of the merged group, Germany, Asia, South Sweden, North Sweden and relicts.

Habitat modelling

Prior to the modelling we removed reads from accessions that is currently in a verification pipeline due to possible mix-ups during sampling[51]. We mined environmental data (derivative measures of temperature, precipitation, solar radiation and humidity) from Worldclim 2.0 (www.worldclim.org; bioclimatic variables at 2.5 arc-minutes resolution[52]), based on each accessions coordinates (with ~10 km precision). For each STR loci, we tested how well the environmental variables predict a change in STR length. To do this, we built linear mixed effect (LME) models with the combined STR length of both alleles for each STR locus (STR dosage) modeled by the values of each predictor variable (of which there were 19, after removing the most correlated variables). Tests were conducted using the R package 'nlme' (v. 3.1, default parameters)[53].

Gene expression modelling

We used normalized RNA-seq profiling data of the rosette leaves from 727 *A. thaliana* accessions[11]. Of these, we had produced STR variant calling data from 665. Using MatrixEQTL[54], we modelled expression using additive linear models, where we tested the significance of STR dosage on gene expression. We accounted for population structure by including the previously defined groups as a covariate.

Gene Ontology and protein networks

We used ThaleMine[55] to test for enrichment of GO terms. For the protein-protein interaction enrichment test we used STRING[56]. For drawing protein networks we used Cytoscape[57] with the GeneMANIA[58] add-on and loaded a gene list including all genes with a coding STR associated to habitat. We looked for at most 20 additional interaction partners for each gene based on automatic weighting. To detect dense clusters within the network we used the MCODE[58,59] Cytoscape add-on.

Phenotypes

We gathered phenotype data for 225 different experiments from the AraPheno database (https://arapheno.1001genomes.org[60]). For measurements of days until flowering at 10°C and 16°C, data was available from all the accessions sequenced by the 1001 Genomes Project. For other phenotypes the number of accessions with phenotype measurements ranged from 198 to 57 accessions. We tested the effect of STR dosage at each STR locus as a predictor of phenotype using linear mixed effect models. The previously defined groups were used as a random effect, to control for population structure. Tests were conducted using the R package 'nlme' (v. 3.1, default parameters)[53].

Code and data availability

Scripts used in this study is available from: https://github.com/uio-cels/TandemRepeats/tree/master/scripts

Zipped file with data analyzed and generated throughout the study is available from: DOI: 10.6084/m9.figshare.6188711

Contributions

W.B.R. and J.B.K. contributed equally to this work.

K.S.J. and M.A.B. conceived the project. O.K.S. and S.J. gave technical support and conceptual advice. W.B.R. and J.B.K. designed and carried out all analyses. W.B.R. and J.B.K. wrote the manuscript with input from all authors.

Affiliations

Department of Biosciences, University of Oslo, Oslo, Norway

William B Reinar, Jonfinn B Knutsen, Sissel Jentoft, Ole K Tørresen, Melinka A Butenko & Kjetill S Jakobsen

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was funded by grants from the Research Council of Norway (RCN grant 251076) to K.S.J. Most computational work was performed on the Abel Supercomputing Cluster (Norwegian Metacenter for High-Performance Computing (NOTUR) and the University of Oslo), operated by the Research Computing Services group at USIT, the University of Oslo IT Department. We thank The 1001 Genomes Project for making available the sequencing data used in this study. The authors would also like to thank the following people for their

contributions: Kjetil Voie, Aliaksandr Hubin, Danny Hitchcock, Lex Nederbragt (University of Oslo) and Rüdiger Simon (Heinrich Heine Universität Dusseldorf).

References

1. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44,** 1161–1165 (2012).
2. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447,** 932–940 (2007).
3. Stine, O. C. *et al.* Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. *Hum. Mol. Genet.* **2,** 1547–1549 (1993).
4. Gemayel, R. *et al.* Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. *Mol. Cell* **59,** 615–627 (2015).
5. Undurraga, S. F. *et al.* Background-dependent effects of polyglutamine variation in the Arabidopsis thaliana gene ELF3. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 19363–19367 (2012).
6. King, D. G., Soller, M. & Kashi, Y. Evolutionary tuning knobs. *Endeavour* **21,** 36–40 (1997).
7. Turpeinen, T., Tenhola, T., Manninen, O., Nevo, E. & Nissilä, E. Microsatellite diversity associated with ecological factors in Hordeum spontaneum populations in Israel. *Mol. Ecol.* **10,** 1577–1591 (2001).
8. Li, Y. C. *et al.* Edaphic microsatellite DNA divergence in wild emmer wheat, Triticum dicoccoides , at a microsite: Tabigha, Israel. *Theor. Appl. Genet.* **101,** 1029–1038 (2000).
9. Li, Y.-C. *et al.* Natural selection causing microsatellite divergence in wild emmer wheat at the ecologically variable microsite at Ammiad, Israel. *Theor. Appl. Genet.* **100,** 985–999 (2000).
10. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166,** 481–491 (2016).
11. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166,** 492–505 (2016).
12. Exposito-Alonso, M. *et al.* Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. *Nat Ecol Evol* **2,** 352–358 (2018).

13. Press, M. O., McCoy, R. C., Hall, A. N., Akey, J. M. & Queitsch, C. Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana. Preprint at *https://www.biorxiv.org/content/early/2017/10/21/145128* (2017).

14. Pelassa, I. & Fiumara, F. Differential Occurrence of Interactions and Interaction Domains in Proteins Containing Homopolymeric Amino Acid Repeats. *Front. Genet.* **6,** (2015).

15. Simon, M. & Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.* **10,** R59 (2009).

16. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16,** 18–29 (2015).

17. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83,** 553–584 (2014).

18. Press, M. O. & Queitsch, C. Variability in a Short Tandem Repeat Mediates Complex Epistatic Interactions in Arabidopsis thaliana. *Genetics* **205,** 455–464 (2017).

19. Piai, A. *et al.* Just a Flexible Linker? The Structural and Dynamic Properties of CBP-ID4 Revealed by NMR Spectroscopy. *Biophys. J.* **110,** 372–381 (2016).

20. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. (2015). doi:10.1101/017459

21. Adamczyk, B. J. & Fernandez, D. E. MIKC* MADS Domain Heterodimers Are Required for Pollen Maturation and Tube Growth in Arabidopsis. *Plant Physiol.* **149,** 1713–1723 (2009).

22. Chen, X., Goodwin, S. M., Boroff, V. L., Liu, X. & Jenks, M. A. Cloning and characterization of the WAX2 gene of Arabidopsis involved in cuticle membrane and wax production. *Plant Cell* **15,** 1170–1185 (2003).

23. Kaye, Y. *et al.* Inositol polyphosphate 5-phosphatase7 regulates the production of reactive oxygen species and salt tolerance in Arabidopsis. *Plant Physiol.* **157,** 229–241 (2011).

24. Golani, Y. *et al.* Inositol polyphosphate phosphatidylinositol 5-phosphatase9 (At5ptase9) controls plant salt tolerance by regulating endocytosis. *Mol. Plant* **6,** 1781–1794 (2013).

25. Cockcroft, C. E., den Boer, B. G., Healy, J. M. & Murray, J. A. Cyclin D control of growth rate in plants. *Nature*

405, 575–579 (2000).

26. Perazza, D., Vachon, G. & Herzog, M. Gibberellins promote trichome formation by Up-regulating GLABROUS1 in arabidopsis. *Plant Physiol.* **117,** 375–383 (1998).
27. Ohashi, Y., Oka, A., Ruberti, I., Morelli, G. & Aoyama, T. Entopically additive expression of GLABRA2 alters the frequency and spacing of trichome initiation. *Plant J.* **29,** 359–369 (2002).
28. Kost, B. *et al.* Rac Homologues and Compartmentalized Phosphatidylinositol 4, 5-Bisphosphate Act in a Common Pathway to Regulate Polar Pollen Tube Growth. *J. Cell Biol.* **145,** 317–330 (1999).
29. Zhou, K., Aertsen, A. & Michiels, C. W. The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol. Rev.* **38,** 119–141 (2014).
30. Sawyer, L. A. Natural Variation in a Drosophila Clock Gene and Temperature Compensation. *Science* **278,** 2117–2120 (1997).
31. Fondon, J. W., 3rd & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 18058–18063 (2004).
32. Burggren, W. Epigenetic Inheritance and Its Role in Evolutionary Biology: Re-Evaluation and New Perspectives. *Biology* **5,** (2016).
33. Heard, E. & Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157,** 95–109 (2014).
34. Echeverria, G. V. & Cooper, T. A. RNA-binding proteins in microsatellite expansion disorders: mediators of RNA toxicity. *Brain Res.* **1462,** 100–111 (2012).
35. Galvão, R., Mendes-Soares, L., Câmara, J., Jaco, I. & Carmo-Fonseca, M. Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. *Brain Res. Bull.* **56,** 191–201 (2001).
36. Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30,** 315–320 (2002).
37. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22,** 1154–1162 (2012).
38. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.

Nucleic Acids Res. **40,** D1202–10 (2012).

39. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
40. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
42. Hale, M. L., Burg, T. M. & Steeves, T. E. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS One* **7,** e45170 (2012).
43. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27,** 3423–3424 (2011).
44. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–34 (2014).
45. Korkuc, P., Schippers, J. H. M. & Walther, D. Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol.* **164,** 181–200 (2014).
46. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. in *Plant Bioinformatics* 89–112 (2007).
47. Oates, M. E. *et al.* D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.* **41,** D508–16 (2013).
48. Peng, Z. & Kurgan, L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.* **43,** e121 (2015).
49. Mészáros, B., Simon, I. & Dosztányi, Z. Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Comput. Biol.* **5,** e1000376 (2009).
50. Meng, F. & Kurgan, L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* **32,** i341–i350 (2016).
51. Pisupati, R. *et al.* Verification of Arabidopsis stock collections using SNPmatch - an algorithm for genotyping high-plexed samples. *bioRxiv* 109520 (2017). doi:10.1101/109520
52. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37,** 4302–4315 (2017).

53. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. (2017).

54. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).

55. Krishnakumar, V. *et al.* ThaleMine: A Warehouse for Arabidopsis Data Integration and Discovery. *Plant Cell Physiol.* **58,** e4 (2017).

56. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–52 (2015).

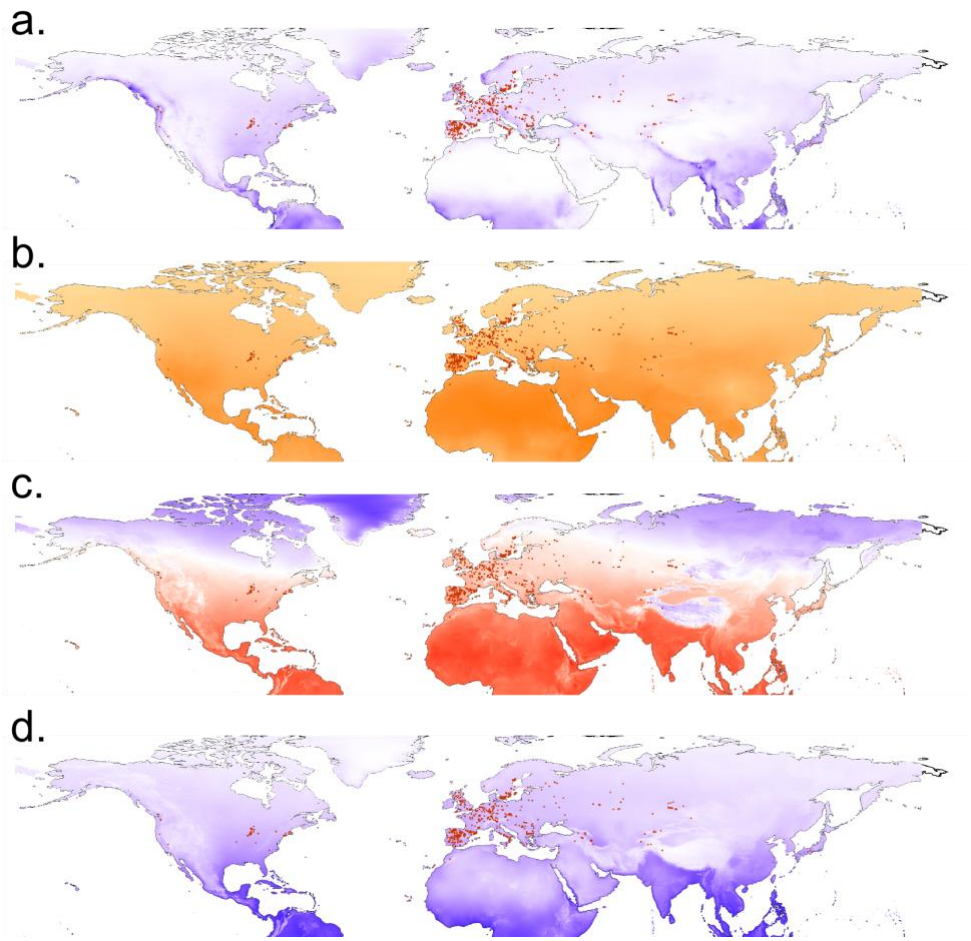
57. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

58. Montojo, J., Zuberi, K., Rodriguez, H., Bader, G. D. & Morris, Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. *F1000Res.* (2014). doi:10.12688/f1000research.4572.1

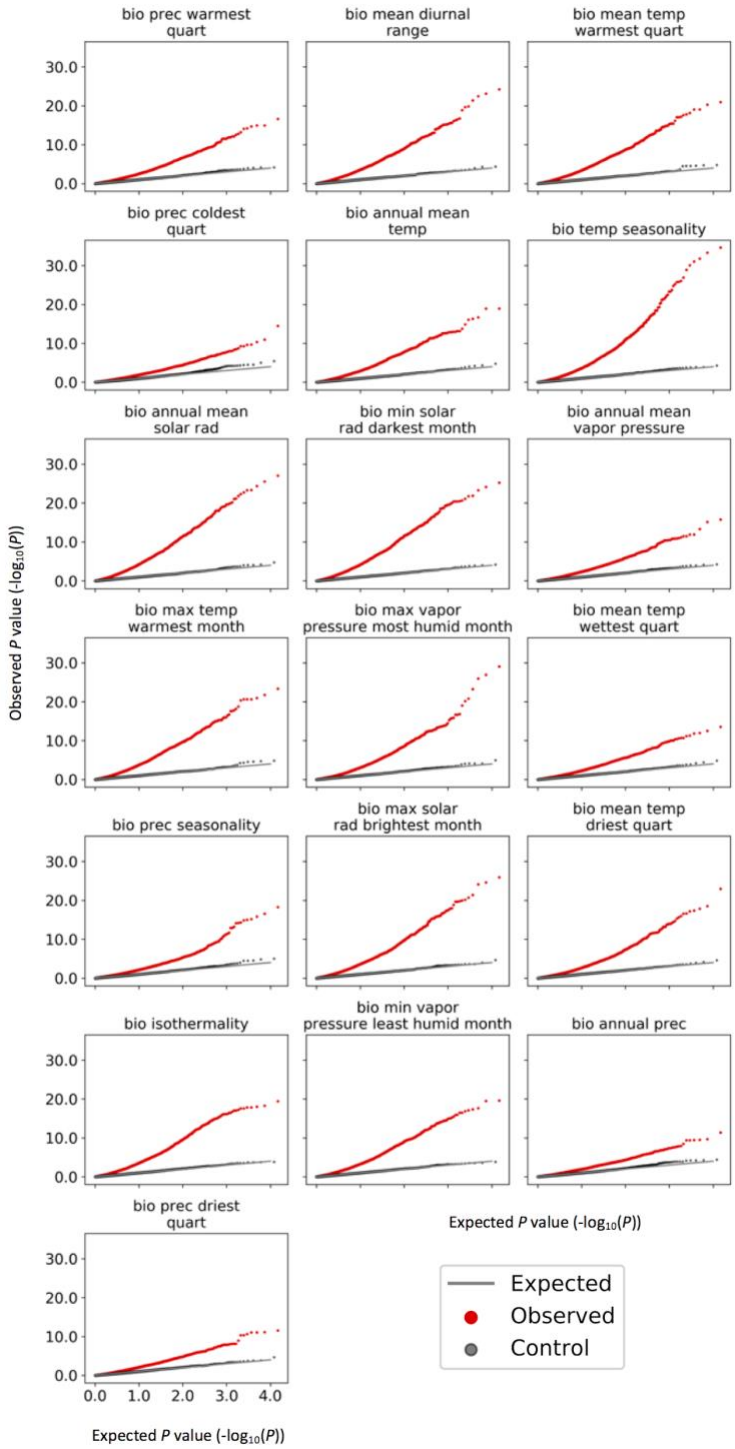
59. Bader, G. D. & Hogue, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4,** 2 (2003).

60. Seren, Ü. *et al.* AraPheno: a public database for Arabidopsis thaliana phenotypes. *Nucleic Acids Res.* **45,** D1054–D1059 (2017).

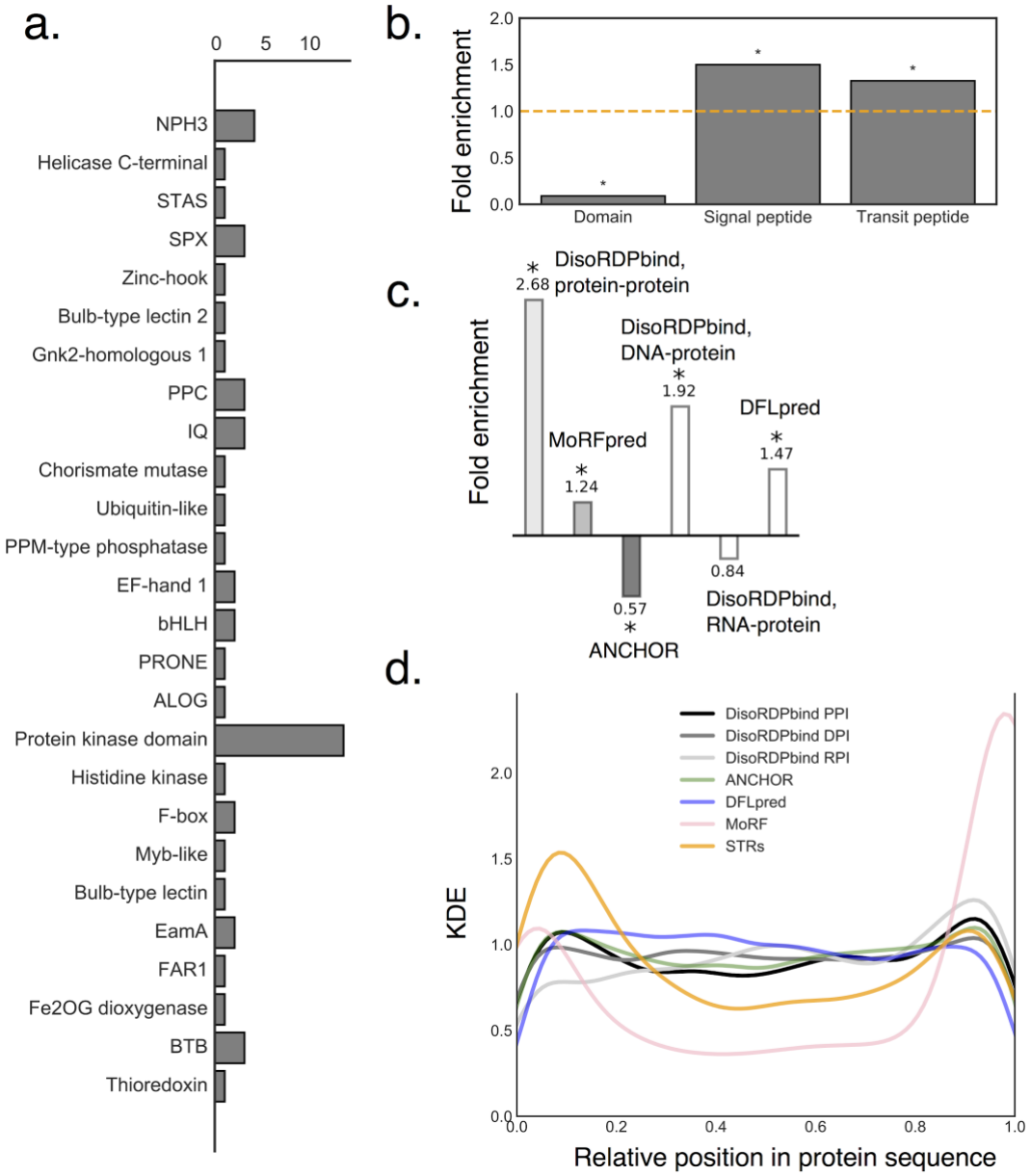
Supplementary Figures



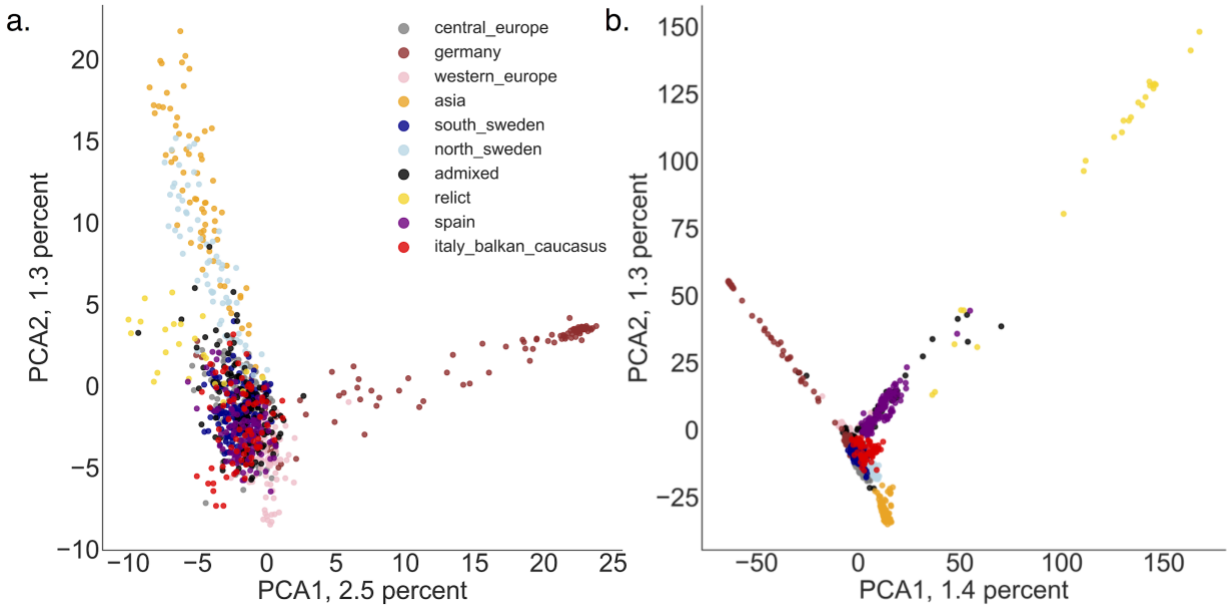
Supplementary Figure 1: Maps showing the distribution of sampled accessions (red dots) and variation in the northern hemisphere for four of the 19 environmental variables. Darker colors indicate higher values. Each variable is calculated from monthly measurements. **a.** Annual mean precipitation (mm). **b.** Annual mean solar radiation (kJ m^{-2} day^{-1}). **c.** Annual mean temperature (°C). **d.** Annual mean vapor pressure (kPa).



Supplementary Figure 2. QQ plots showing the *p* values of association tests between STR length and 19 environmental variables.



Supplementary Figure 3. The distribution of amino acid-encoded STRs in proteins. **a.** The number of STRs present in domains. **b.** Enrichment scores (observed number of overlaps divided by the median of randomized shuffled controls) of STRs in protein-protein, protein-DNA and protein-RNA interaction sites as well as disordered flexible linkers (DFL) and intrinsically disordered regions. **c.** Fold enrichment scores for STRs in domains, and signal and transit peptides. Test with a *p* value ≤ 0.001 are denoted with an asterix. **d.** Location of STRs, putative disordered protein-protein, DNA-protein and RNA-protein interaction sites (PPIs, DPIs and RPIs), ANCHOR sites, DFLpred in relation to protein sequence.



Supplementary Figure 4. Principal Component Analysis of *Arabidopsis thaliana* accessions. **a**. STR-based PCA based on 1971 loci. **b**. SNP-based PCA based on 50,000 random SNPs.