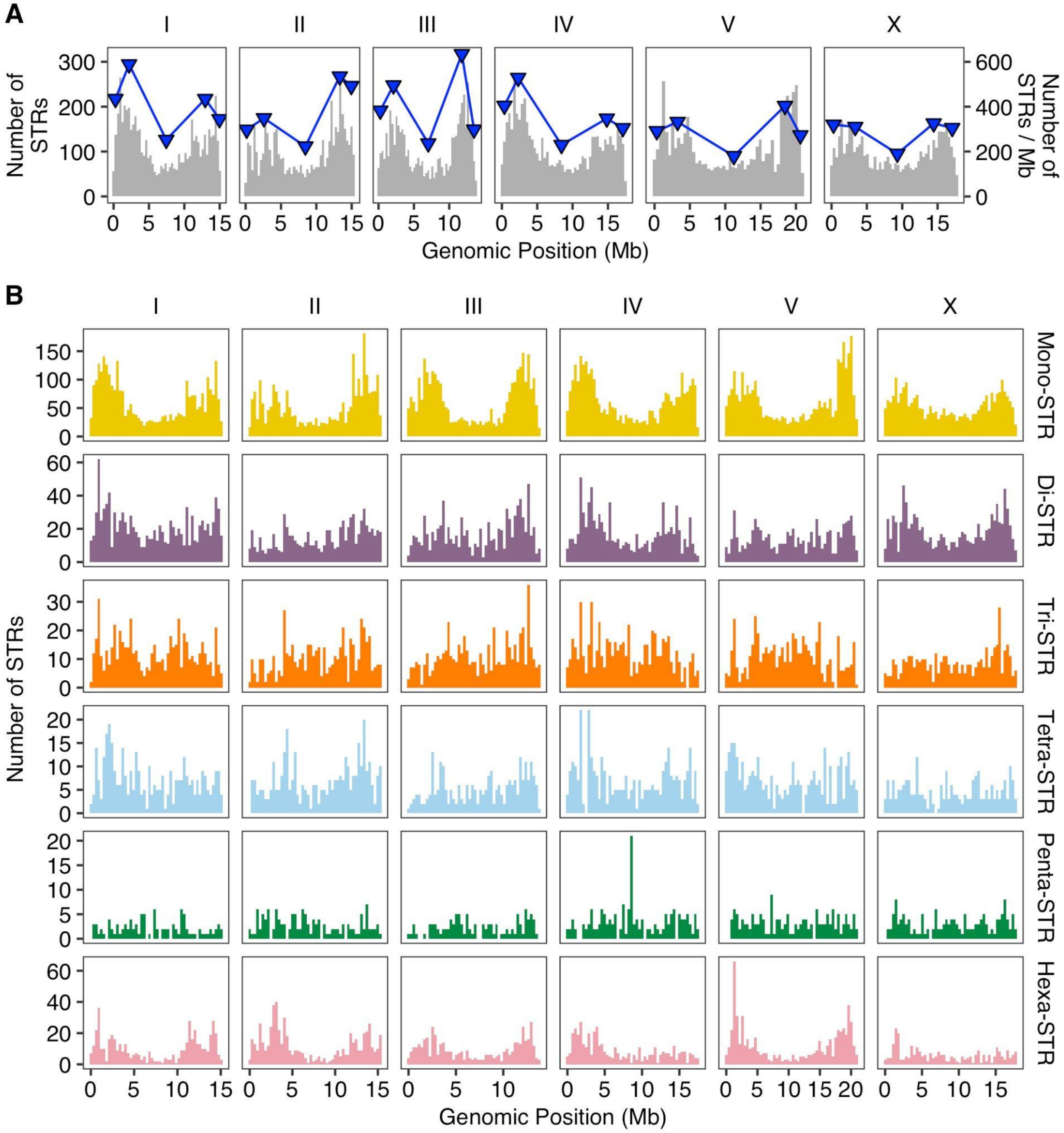
Supplemental Material

Natural variation in *C. elegans* short tandem repeats

Gaotian Zhang[1], Ye Wang[1,2], and Erik C. Andersen[1,*]

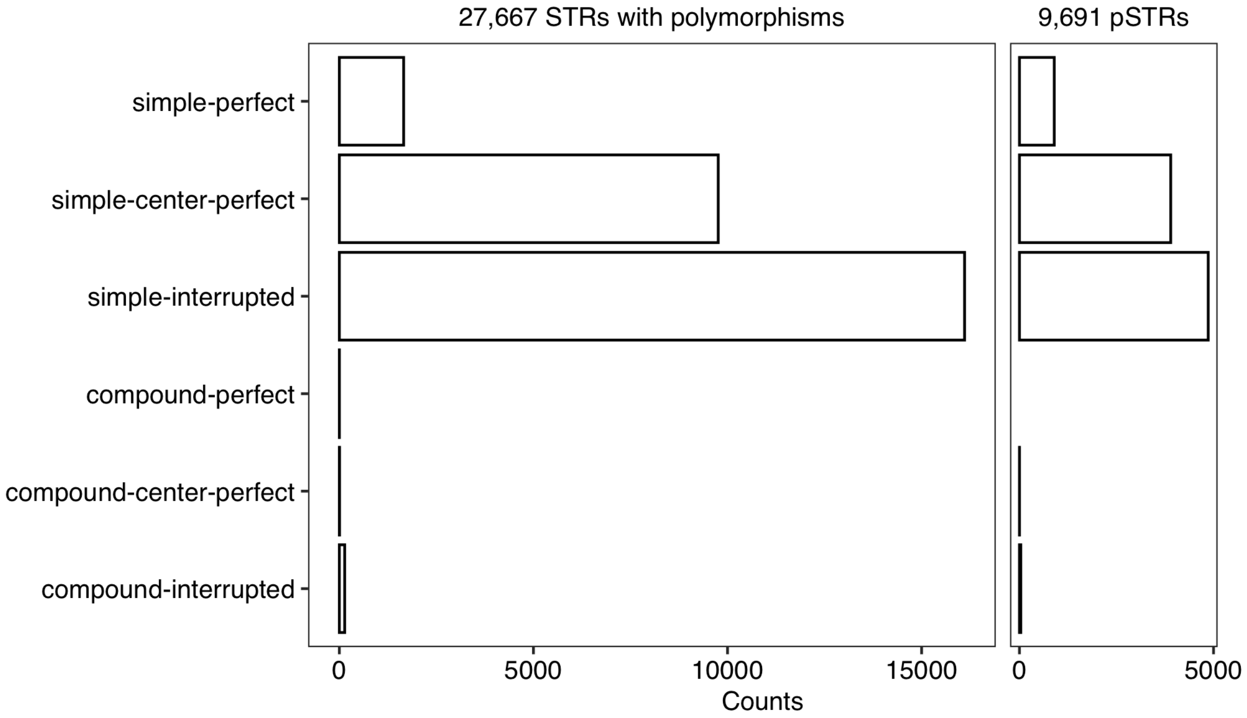
1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
2. Current address: Sichuan Key Laboratory of Conservation Biology on Endangered Wildlife, Chengdu Research Base of Giant Panda Breeding, Chengdu, Sichuan, P. R. China
[*]Corresponding author. E-mail: erik.andersen@gmail.com (E.C.A.)

TABLE OF CONTENTS



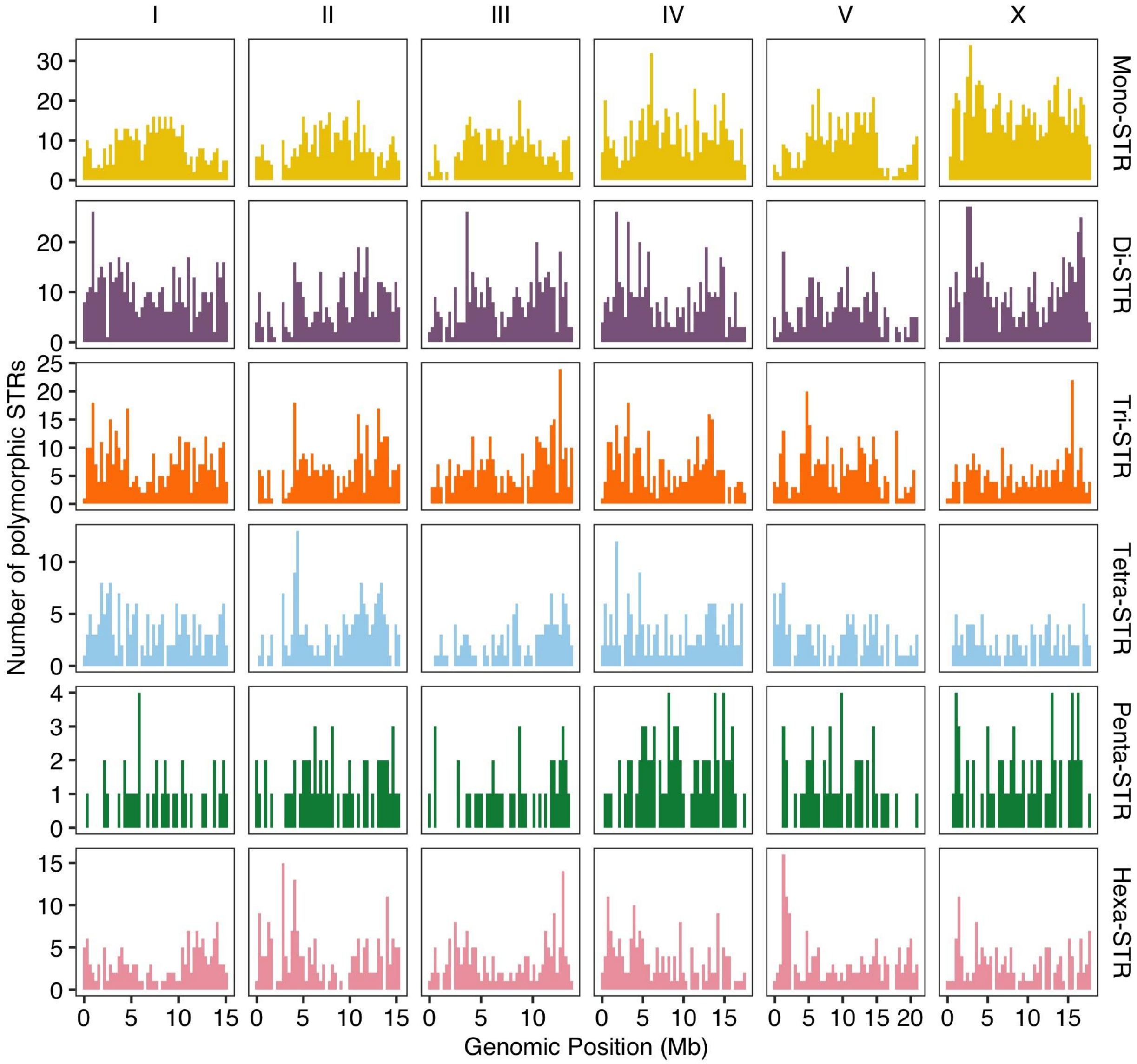
Supplemental Fig. S1:

The distribution of reference STRs across *C. elegans*. *(A)* The distribution of reference STRs in the *C. elegans* genome. Blue triangles represent the number of STRs per Mb in different genomic domains (tips, arms, and centers) (Rockman and Kruglyak 2009). *(B)* The distribution of reference STRs with different motif lengths in the *C. elegans* genome.



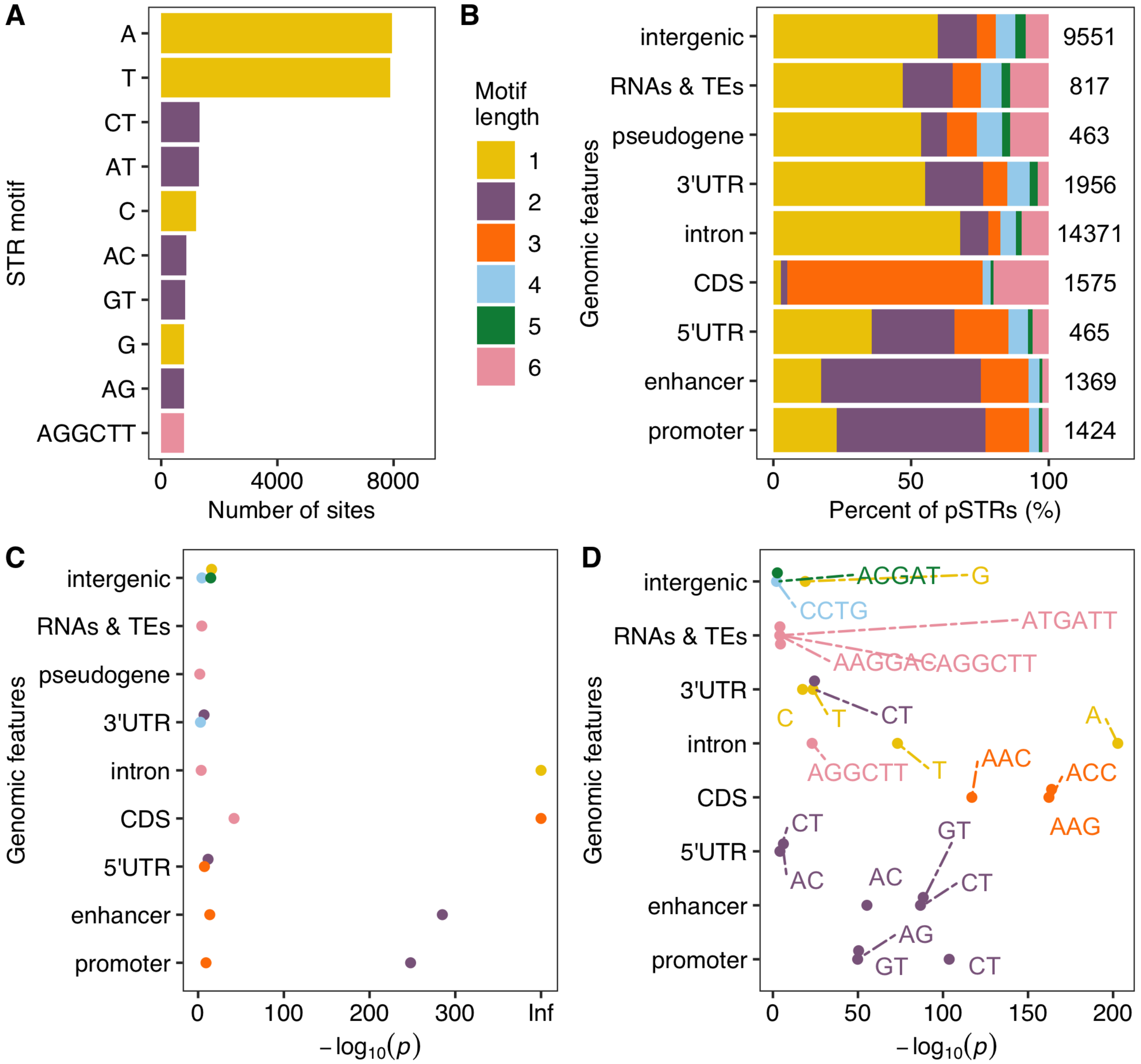
Supplemental Fig. S2:

STR categories by compositions. Counts (x-axis) of STRs in six different categories (y-axis) by compositions in 27,667 STRs with polymorphisms across 540 wild strains and 9,691 pSTRs with missing calls in less than 10% of all strains.



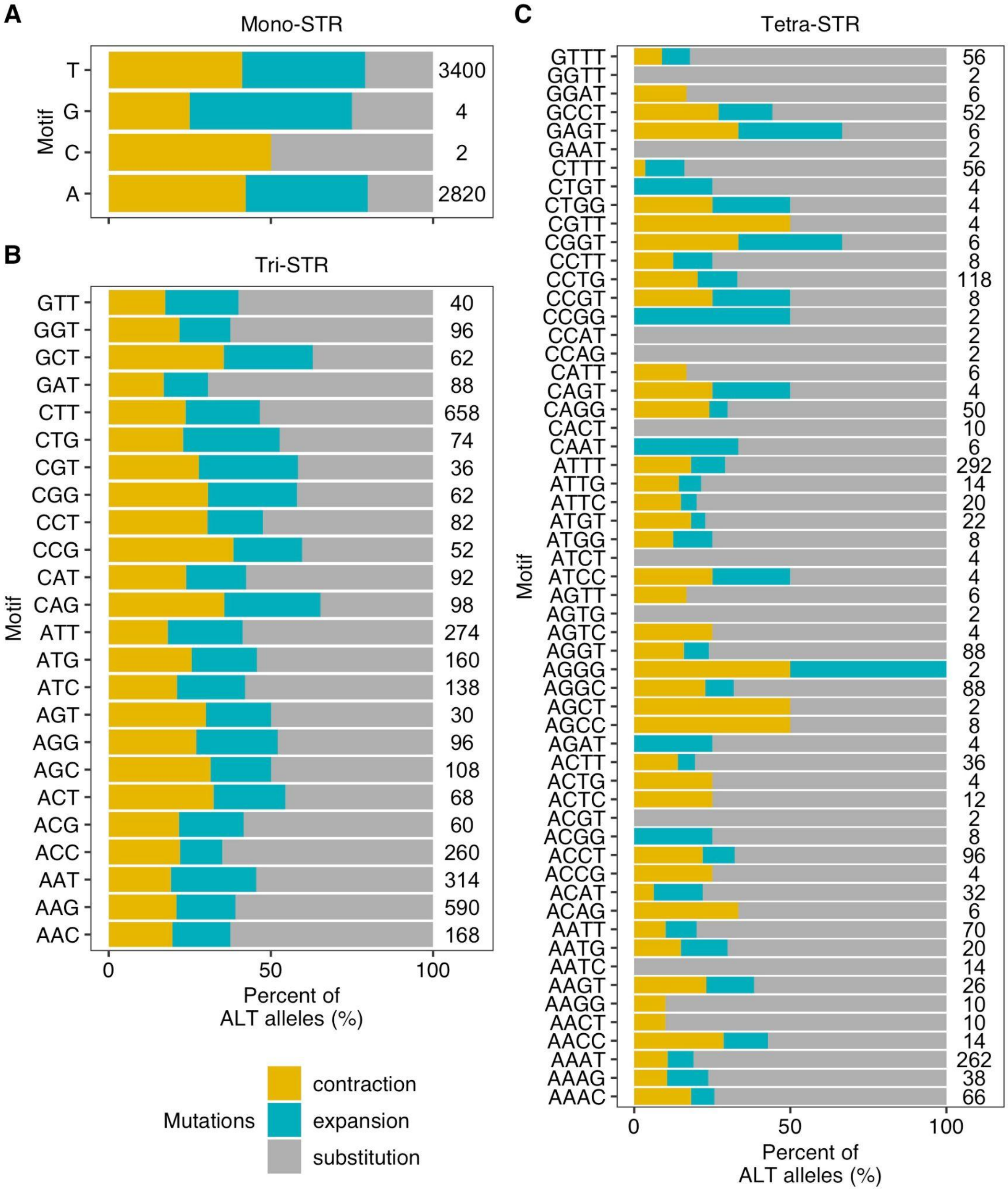
Supplemental Fig. S3:

The distribution of polymorphic STRs with different motif lengths.



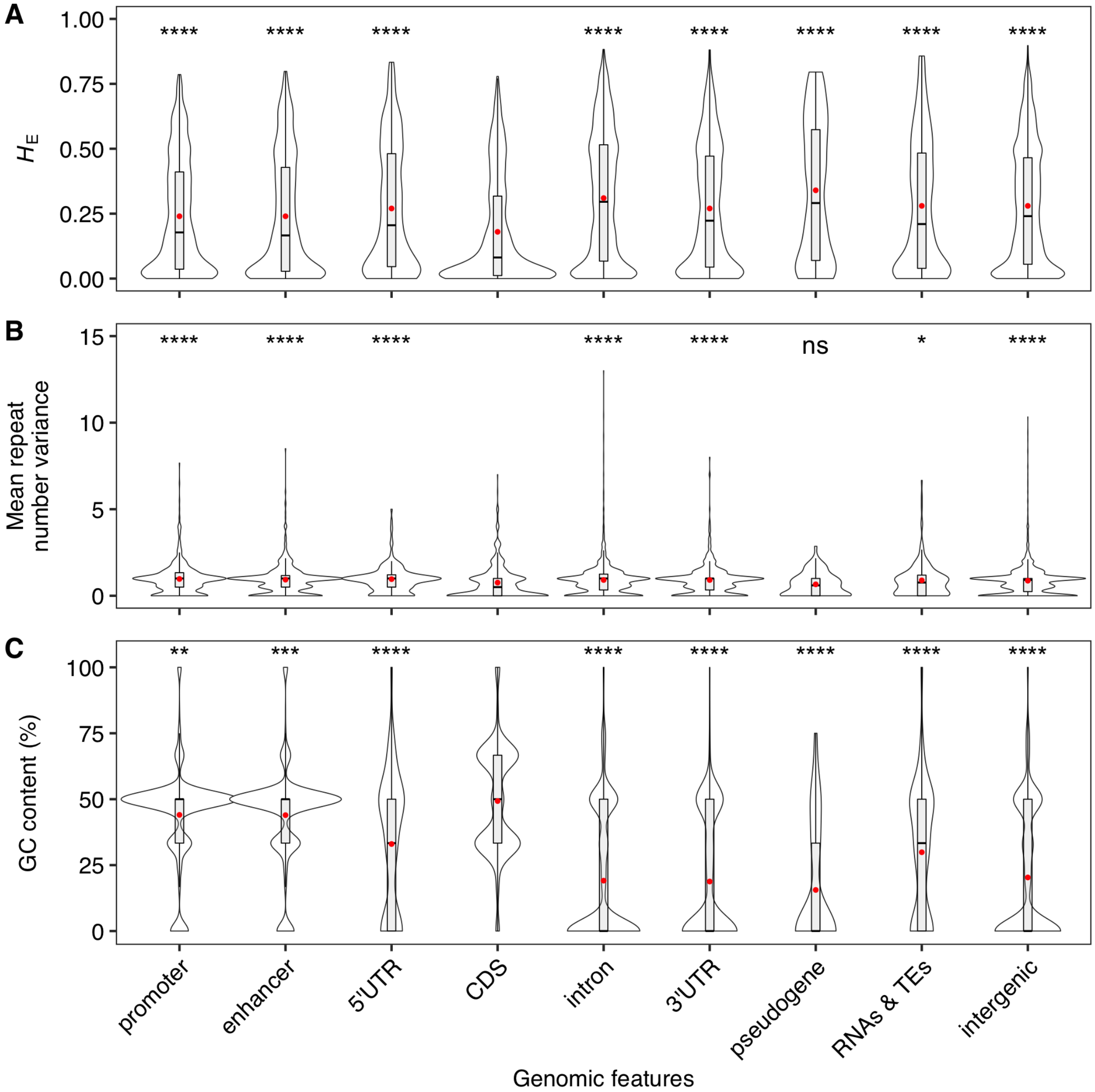
Supplemental Fig. S4:

Motifs and genomic features of reference STRs in *C. elegans*. *(A)* The top ten most frequent motif sequences in reference STRs are shown on the y-axis, and the number of those sits on the x-axis. *(B)* Percent of reference STRs with different motif lengths in each genomic feature. The total number of reference STRs in each genomic feature is indicated. *(C)* Enriched STRs with different motif lengths (colored as in (B)) in different genomic features are shown. *(D)* The top three most enriched STR motif sequences (labeled) in each genomic feature (if enriched motifs were found) are shown. Statistical significance (Supplemental Table S2) for enrichment tests was calculated using the one-side Fisher's exact test and was corrected for multiple comparisons (Bonferroni method).



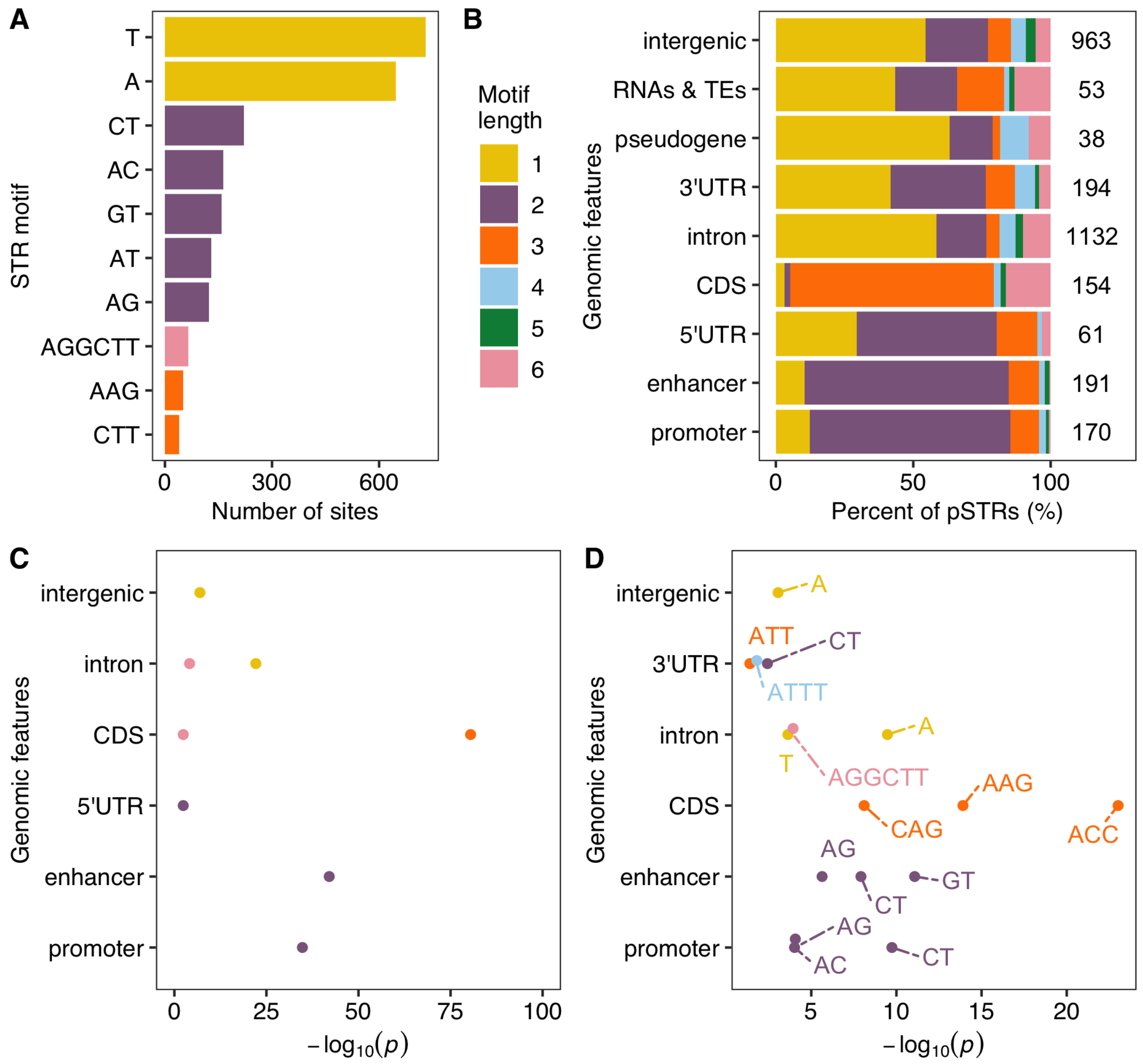
Supplemental Fig. S5:

Percent of alternative alleles in different STR mutations. Percent of alternative alleles showing contraction, expansion, or substitution in mono-STRs *(A)*, tri-STRs *(B)*, and tetra-STRs *(C)*. The total number of STRs with different motifs is indicated on the right.



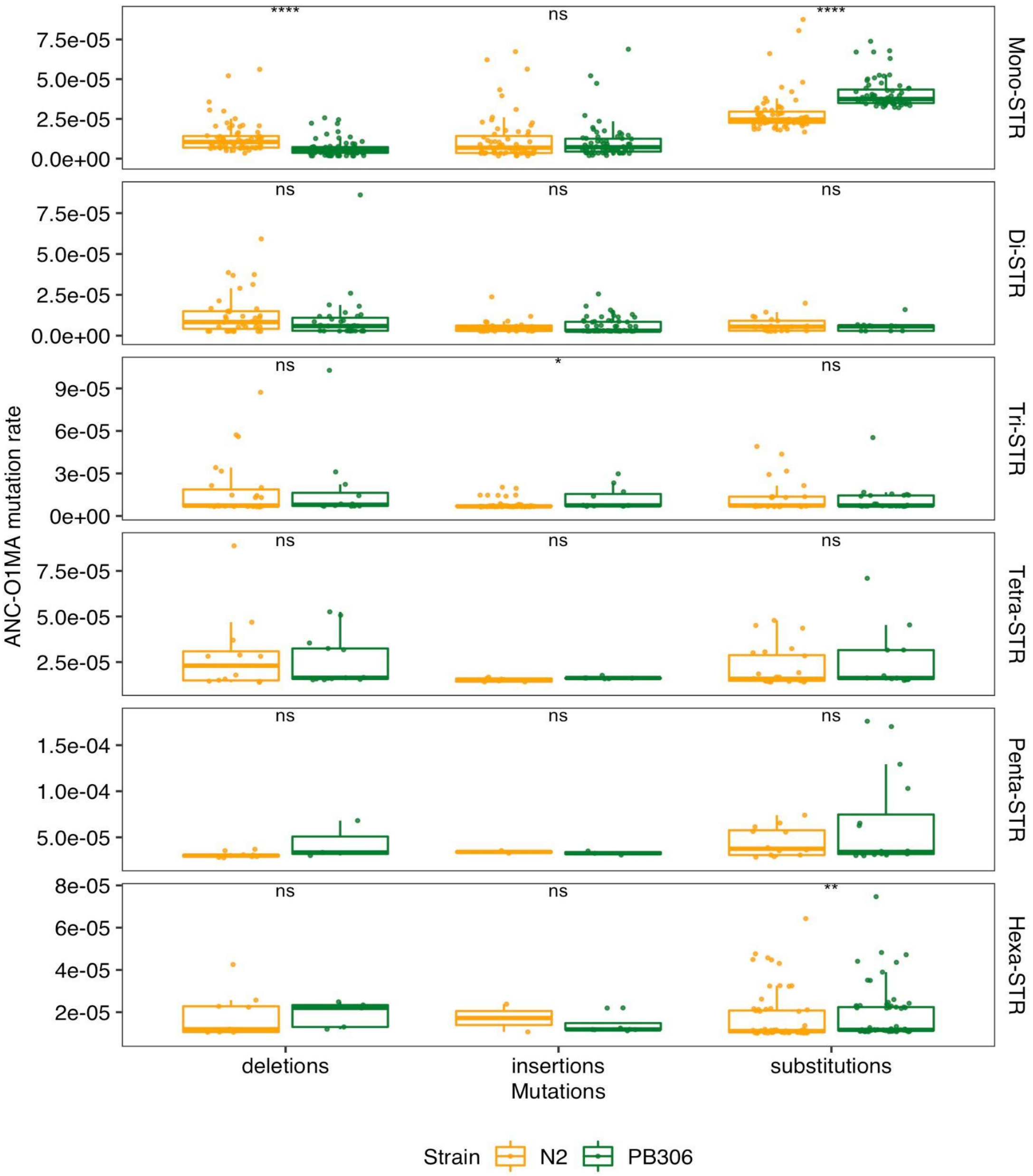
Supplemental Fig. S6:

Mutations of STRs in CDS regions are constrained. Comparisons of expected heterozygosity (H_E) *(A)*, mean repeat number variance of each STRs *(B)*, and motif GC content *(C)* between polymorphic STRs in CDS regions and other regions. Red dots indicate mean values of each estimate in each region. Statistical significance (Supplemental Table S2) was calculated using the two-sided Wilcoxon test and was corrected for multiple comparisons (Bonferroni method). Significance of each comparison is shown above (ns: adjusted $p > 0.05$; *: adjusted $p \leq 0.05$; **: adjusted $p \leq 0.01$; ***: adjusted $p \leq 0.001$; ****: adjusted $p \leq 0.0001$).



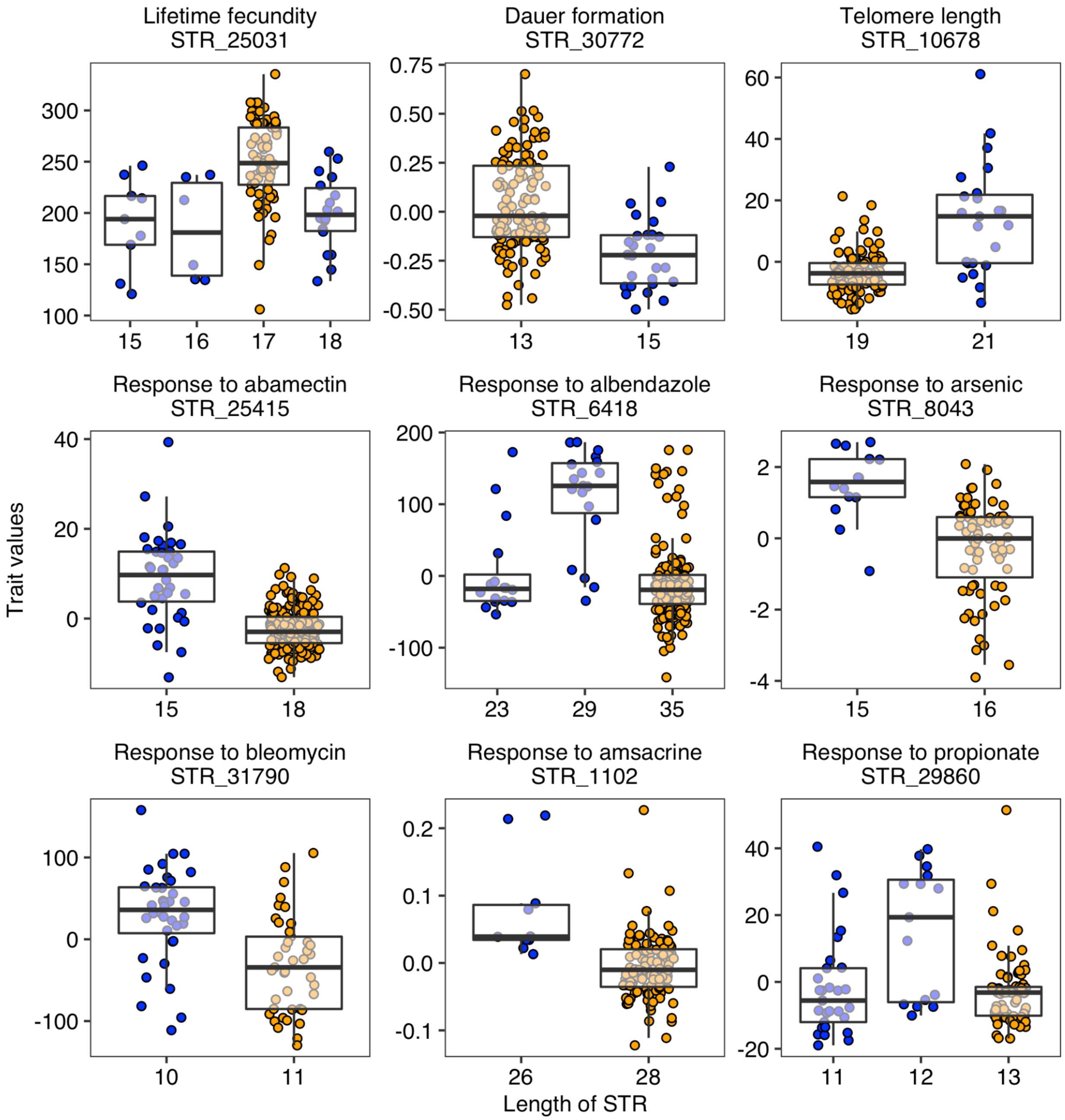
Supplemental Fig. S7:

Motifs and genomic features of polymorphic STRs in MA lines. *(A)* The top ten most frequent motif sequences in polymorphic STRs. *(B)* Percent of polymorphic STRs with different motif lengths in each genomic feature. The total number of polymorphic STRs in each genomic feature are indicated. *(C)* Enriched STRs with different motif lengths (colored as in (B)) in different genomic features are shown. *(D)* The top three most enriched STR motif sequences (labeled) in each genomic feature (if enriched motifs were found) are shown. Statistical significance (Supplemental Table S2) for enrichment tests was calculated using the one-side Fisher's exact test and was corrected for multiple comparisons (Bonferroni method).



Supplemental Fig. S8:

STR mutation rates between two O1MA lines. Comparison of STR mutation rates of deletions, insertions, and substitutions between O1MA lines derived from N2 (orange) and PB306 (green) using pSTRs of different motif lengths. Each dot represents the mutation rate between the ancestor strain (ANC) and one of O1MA lines (ANC-O1MA). Statistical significance (Supplemental Table S2) of difference comparisons were calculated using the two-sided Wilcoxon test and *p*-values were adjusted for multiple comparisons (Bonferroni method). Significance of each comparison is shown above each comparison pair (ns: adjusted $p > 0.05$; **: adjusted $p \leq 0.01$; ***: adjusted $p \leq 0.001$; ****: adjusted $p \leq 0.0001$).



Supplemental Fig. S9:

Effects of STR variation on phenotypic traits. Tukey box plots showing phenotypic variation of nine traits (indicated above each panel) between strains with different lengths of the STRs (indicated under the trait name) that showed the most significant association with the trait. Each point corresponds to a strain and is colored orange or blue for strains with the N2 reference allele or the alternative alleles, respectively.

Supplemental Table legends

Supplemental Table S1 (separate file)
Various features (genomic positions, motifs, expansion/contraction scores, etc.) of reference STRs and pSTRs in wild *C. elegans* strains.

Supplemental Table S2 (separate file)
Fisher's Exact test and Wilcoxon test statistics.

Supplemental Table S3 (separate file)
Allele frequency and expected heterozygosity of pSTRs in wild *C. elegans* strains. Expected heterozygosity of SNVs in swept strains.

Supplemental Table S4 (separate file)
Percentages of homozygous alternative alleles and heterozygous alleles in each wild strain.

Supplemental Table S5 (separate file)
PCA results using pSTRs and SNVs among wild *C. elegans* strains, respectively.

Supplemental Table S6 (separate file)
Features of pSTRs among MA lines.

Supplemental Table S7 (separate file)
Mutation rates of pSTRs among MA lines.

Supplemental Table S8 (separate file)
Statistics of Likelihood-ratio tests for association between STR length variation and phenotypic variation of 11 organismal traits.

Supplemental Table S9 (separate file)
Phenotypic values of nine organismal traits and lengths of STRs that showed the most significant association with each trait.

References

Rockman MV, Kruglyak L. 2009. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet* **5**: e1000419.