

Supplemental Information

Title: Natural diversity of telomere length distributions across 100 *Saccharomyces cerevisiae* strains

Authors: Clotilde Garrido^{1,2}, Cintia Gómez-Muñoz^{1,2}, Etienne Kornobis^{3,4}, Nicolas Agier^{1,2}, Oana Iliaia^{1,2}, Gilles Fischer^{1,2,*}, Zhou Xu^{1,2,*}

Affiliations:

¹ Sorbonne Université, CNRS, Laboratory of Computational, Quantitative and Synthetic Biology, CQSB, F-75005 Paris, France.

² Sorbonne Université, CNRS, Inserm, Institut de Biologie Paris-Seine, IBPS, F-75005 Paris, France.

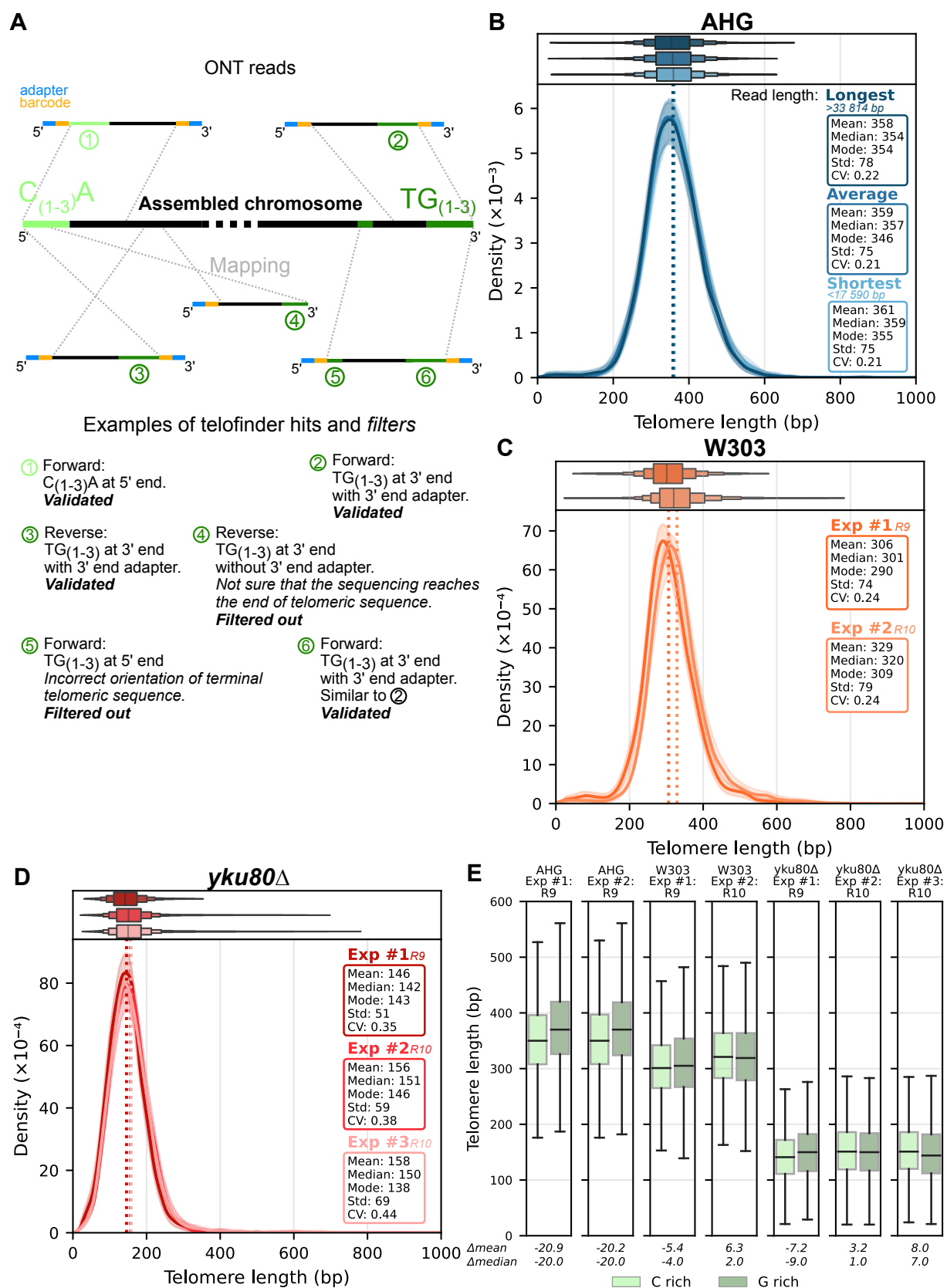
³ Institut Pasteur, Université Paris Cité, Plate-forme Technologique Biomics, F-75015 Paris, France.

⁴ Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France.

Correspondence:

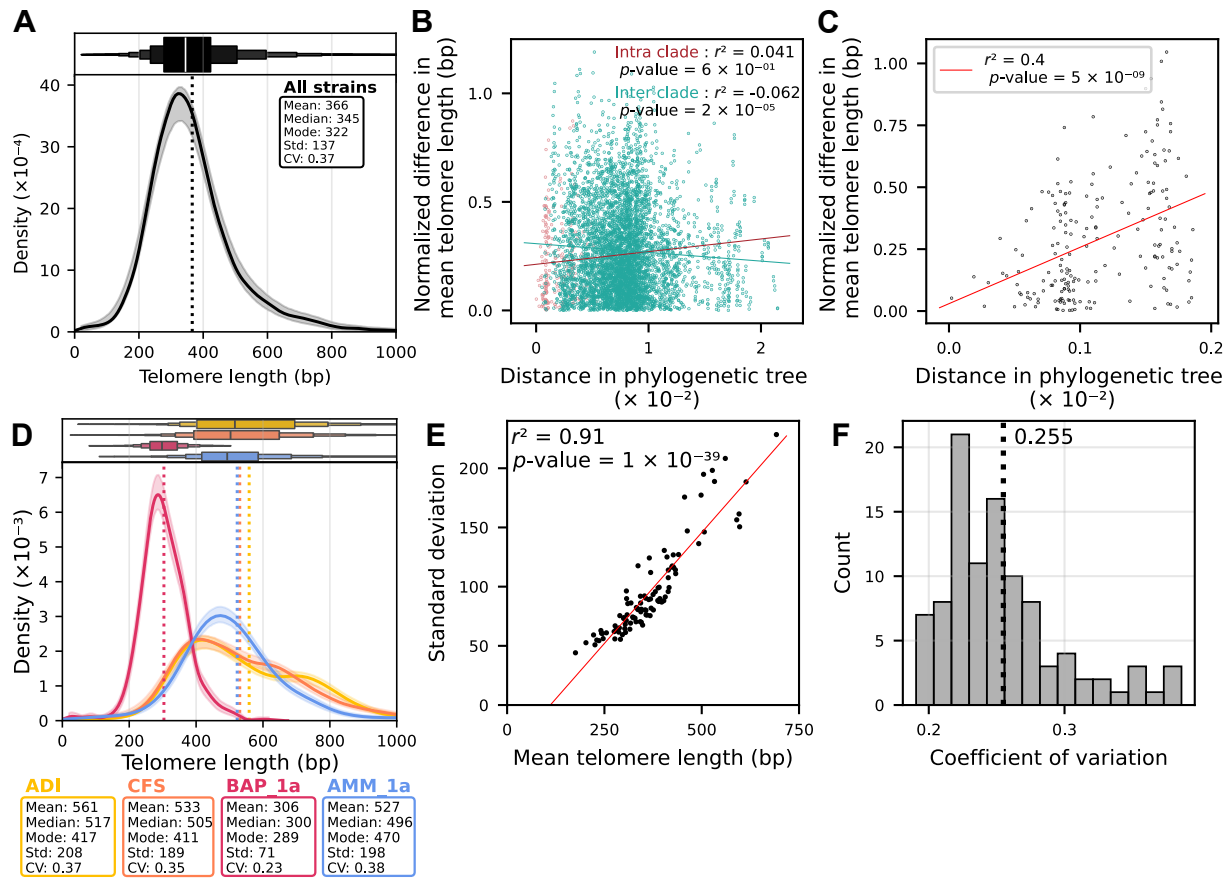
gilles.fischer@sorbonne-universite.fr

zhou.xu@sorbonne-universite.fr

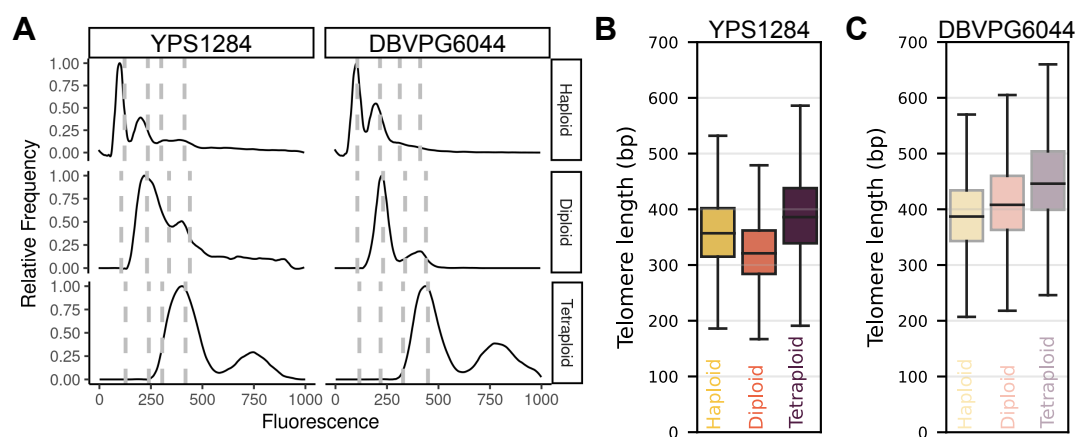


Supplemental Figure 1. Robust TL distribution measurement from Nanopore reads. (A) Filters used to eliminate telomere-like sequences with the wrong orientation and internal telomeric sequences, and to ensure the read reached the end the molecule. Examples numbered 1 to 6 illustrate validated or filtered out reads. (B) TL distributions computed from equally

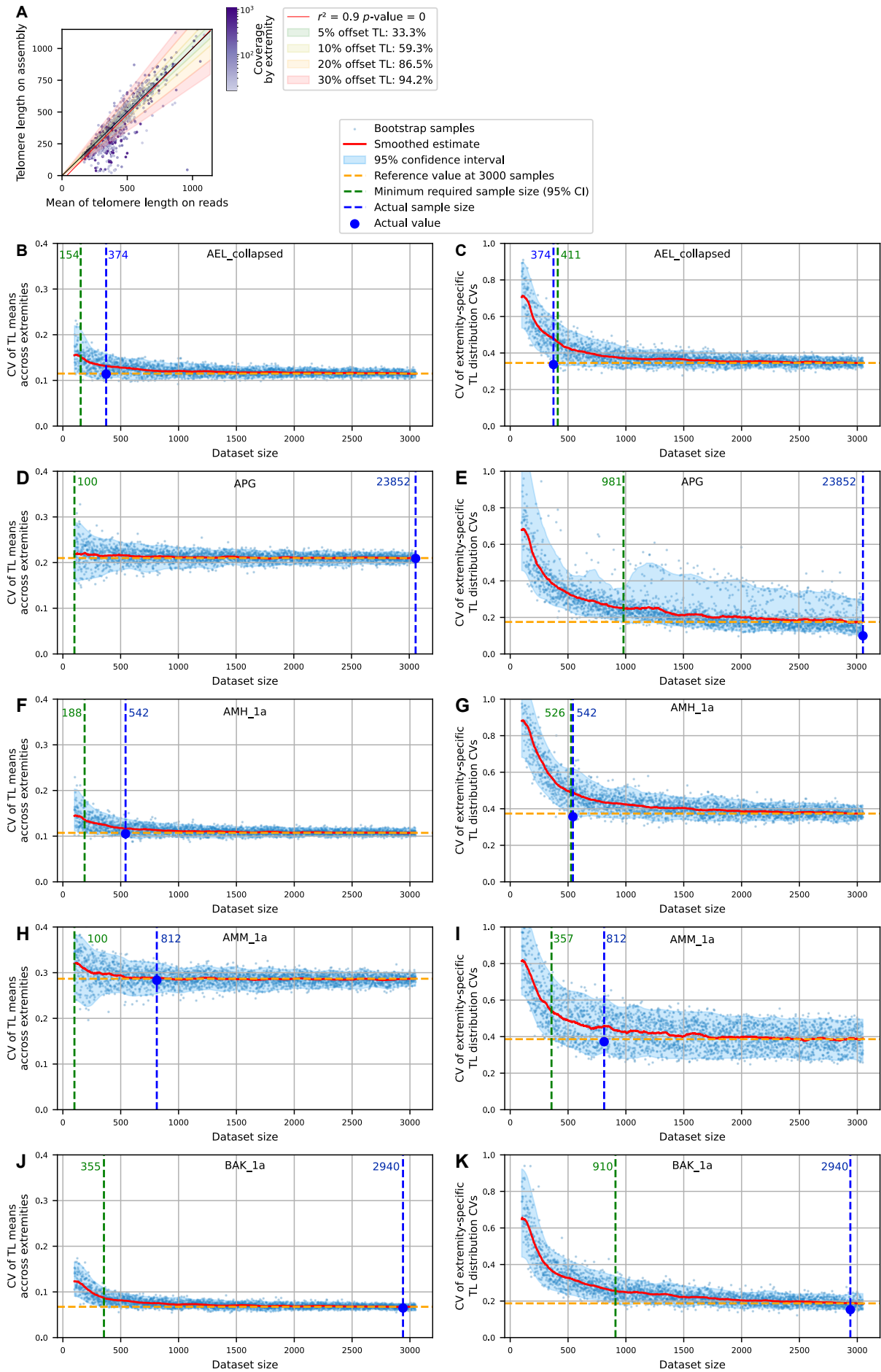
sized subsets of reads of different lengths (“Longest” >34 kb, “Average” between 18 and 34 kb, and “Shortest” <18 kb), using the AHG reads from (O'Donnell et al., 2023). (C) Independent biological replicates of W303's TL distribution. (D) Independent biological replicates of the *yku80Δ* mutant's TL distribution. (E) Comparison between TL distributions derived from the C-rich or G-rich strands for the indicated experiments. (C-E) The version of the ONT flowcell technology used, “R9” (for R9.4.1) or “R10” (for R10.4.1), is indicated.



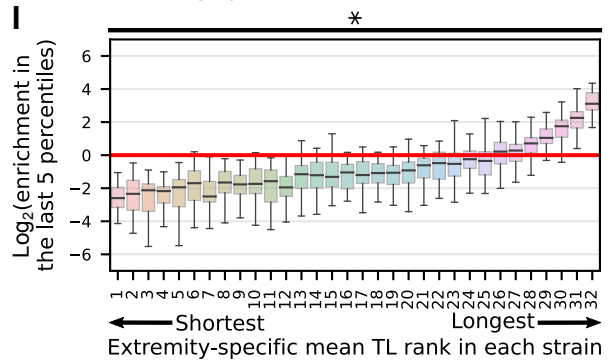
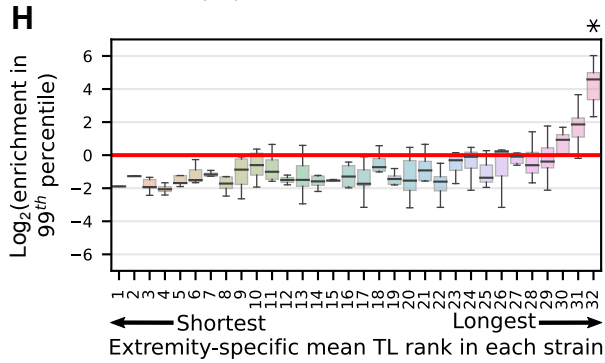
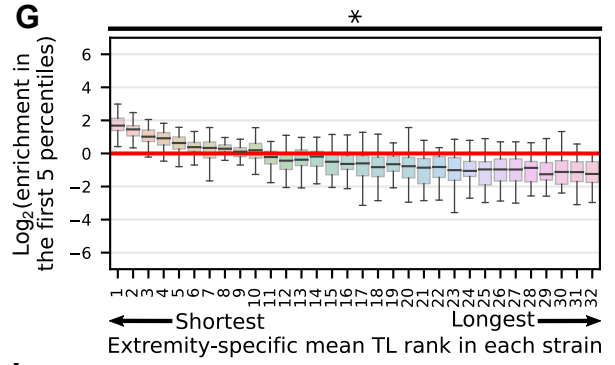
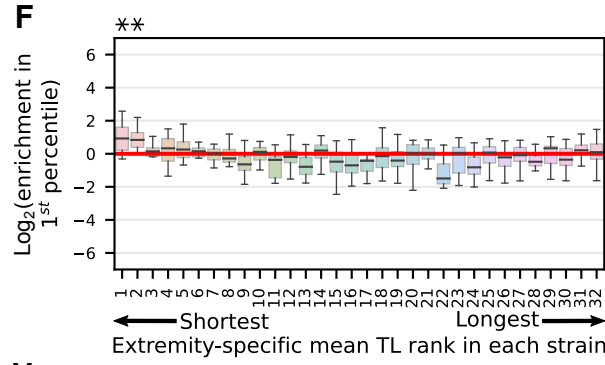
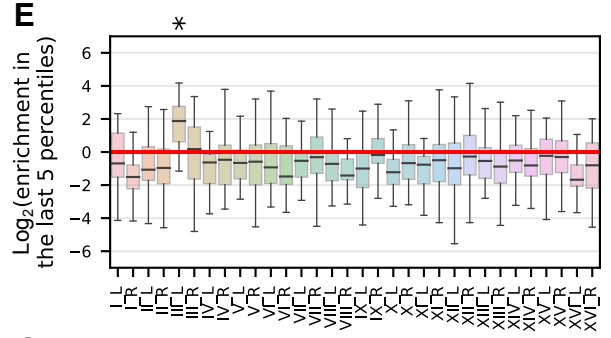
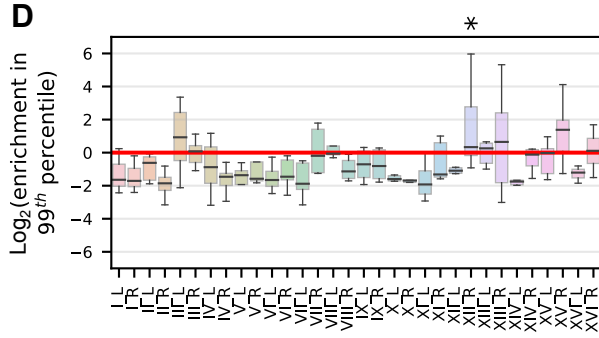
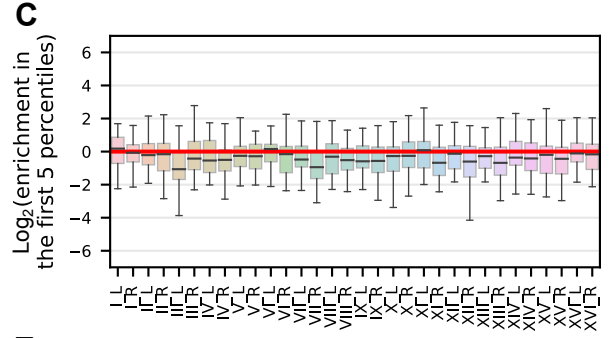
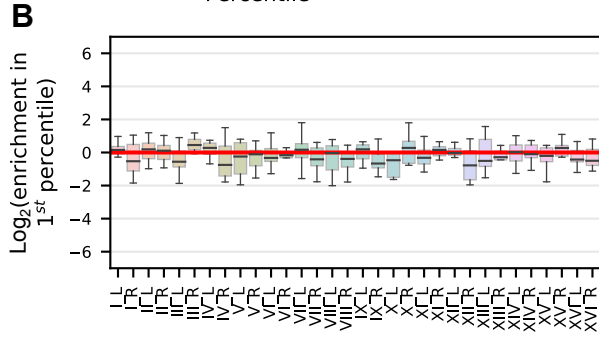
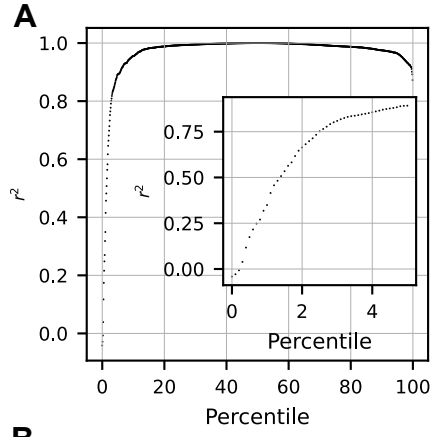
Supplemental Figure 2. Characterization of TL distribution metrics in natural strains. (A) Representative overview of TL distribution in the species computed from a balanced random sample of telomeric reads from each strain. (B) Correlation between phylogenetic distance and the normalized difference in mean TL (difference between the 2 TL means divided by the average of the 2 means). Each dot represents a pair of strains. A pair within the same clade or between 2 clades is shown in red or cyan, respectively. Pearson's correlation (r^2) and associated p -values are shown. (C) Same as (B), with a focus on the 4% closest pairs of strains. (D) TL distributions for strains ADI, CFS, BAP_1a and AMM_1a. (E) Correlation between mean TL and standard deviation across strains. (F) Distribution of the CVs of all strains.



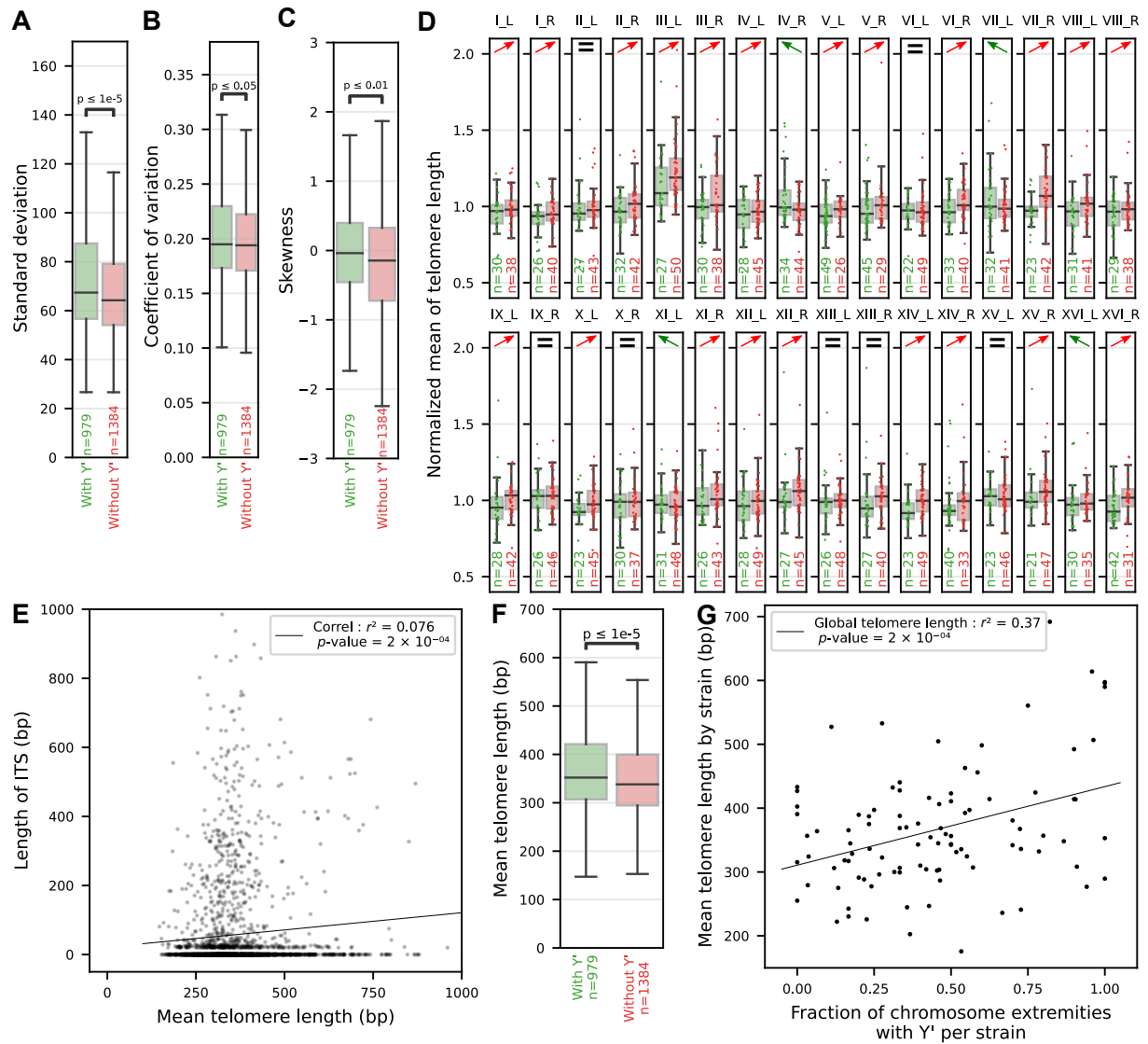
Supplemental Figure 3. Tetraploids exhibit longer telomeres than isogenic diploids and haploids. (A) Propidium iodide staining and flow cytometry analysis of DNA content of the indicated haploid, diploid, and tetraploid strains. Vertical dotted lines represent 1C, 2C, 3C, and 4C DNA contents based on positive control strains used in the same flow cytometry experiment. (B) Boxplots of global TL distributions of isogenic haploid, diploid, and tetraploid strains from the same West African DBVPG6044 strain background. (C) Same as (B) with the North American YPS128 strain background.



Supplemental Figure 4. Comparison between assembly-derived and read-derived extremity-specific TL values, and sample size requirements. (A) Extremity-specific TL values derived from genome assemblies from (O'Donnell et al., 2023) and means of extremity-specific TL distributions computed from reads are plotted. Each dot represents the telomere at one extremity and its read coverage is indicated by its color. Shaded areas of different colors display 5-30% offsets between the two values and the percentage in the legend indicate the fraction of affected extremities. (B, D, F, H, J) Bootstrap analyses of the CV of extremity-specific TL means as a function of the size of the dataset (number of telomeric reads). Each blue dot corresponds to one bootstrap simulation. The average is represented by the red line and the shaded blue area corresponds to 95% of the bootstrap values. The horizontal yellow dotted line shows the value for a dataset size of 3000. The green dotted line indicates the threshold sample size, corresponding to the point where the blue area diverges from the yellow line. The blue dotted line shows the actual dataset size. We determine that the sample size is adequate if the green dotted line is on the left of the blue dotted line. (C, E, G, I, K) Same as (B, D, F, H, J) but for the CV of the extremity-specific TL distribution CVs. (B) and (C) show the example of a strain that does not meet the criteria for the CV of the extremity-specific TL distribution CVs. (D-K) shows this analysis for the 4 strains displayed in Fig. 4A.



Supplemental Figure 5. Contributions of specific extremities to the shortest or longest telomeres. (A) Coefficient of determination r^2 between the median TL and the n^{th} percentile (x-axis). Inset shows a zoom on the first 5 percentile. (B) Enrichment of specific chromosome extremities in the set of individual telomeres shorter than the 1st percentile in a strain-specific TL distribution. The boxplot represents the enrichment values from all strains in Log₂ scale. (C) Same as (B) for the 5th percentile. (D) and (E), same as (B) and (C) but for the set of individual telomeres longer than the 99th (respectively, 95th) percentile. (F) Enrichment of the chromosome extremities ranked for each strain according to their mean extremity-specific TL in the set of individual telomeres shorter than the 1st percentile. (G) Same as (F) for the 5th percentile. (H) and (I), same as (G) and (F) but for the set of individual telomeres longer than the 99th (respectively, 95th) percentile. (B-I) In these panels, an ANOVA test was performed to detect differences in means. The following p -values were obtained, respectively: 0.28, 0.069, 0.034, 1.8×10^{-15} , 0.0065, 9.8×10^{-129} , 2.2×10^{-17} , 9.6×10^{-177} . * indicates extremities that are significantly different from at least two others based on p -values from Tukey's HSD tests with a 0.05 threshold.



Supplemental Figure 6. Correlation between the presence of a Y' element and TL. (A) Standard deviations of the extremity-specific TL distributions, clustered depending on the presence or absence of Y' element in the corresponding subtelomere. The number of chromosome extremities in each group is indicated by "n". The p -value from a Student's t -test is indicated. (B) Same as (A) but for the coefficient of variation of the extremity-specific TL distribution. (C) Same as (A) but for the skewness. (D) Distributions of the normalized extremity-specific TL means shown for each extremity across strains, clustered depending on the presence or absence of Y' element in the corresponding subtelomere. Red and green arrow indicate the direction of the differences in TL means, while "=" indicates identical means within a margin of 1%. (E) Plot representing the total length of ITSs at a chromosome extremity against the mean TL at the same extremity, across all strains, showing no correlation. (F) Distributions of extremity-specific TL means across the extremities of all strains, clustered according to the presence or absence of the Y' element in the corresponding subtelomere. (G) Correlation between the global TL means and the fraction of extremities with Y' elements in the corresponding strain.

Supplemental Table S1. List of strains and accessions.

Summary table of all strains used in this work. Relevant sequencing statistics and TL distribution metrics are indicated.

Supplemental Code. Codes used in this work to measure telomere length from ONT sequencing data.