Supplemental data

Materials and Methods

Dataset assembly and analysis of repeat variability

The following TRF (Benson 1999) parameters were used: matching weight 2, mismatching penalty 5, indel penalty 5, match probability 0.8, indel probability 0.1, score ≥40 and maximum period 500. For each repeat found, its genomic coordinates and characteristics (such as unit size, number of units, repeat purity, base composition, consensus sequence) were stored in a MySQL database. We then investigated if the found repeats vary between the three available high-coverage *S. cerevisiae* genomes, namely strains S288C, RM11-1A (version 1; RM11-1A Sequencing Project. Broad Institute of Harvard and MIT) and YJM789 (version 2, *S. cerevisiae* YJM789 Genome Project, Stanford Genome Technology Center). We used Mercator (Dewey and Pachter 2006) and MAVID (Bray and Pachter 2004) to perform whole genome multiple alignment of the three sequences. Using default parameters, more than 95 % of S288C, YJM789 and RM11-1A genomes were covered by the alignment, 11,653,400 nt. (95.8%), 11,371,262 nt. (95.1%) and 11,366,133 nt. (96.8%), respectively. S288C genomic TR positions were subsequently mapped onto this whole genome alignment. The alignment was sliced at the edges of each TR positions plus 25 nt. on each side. To determine if the same basic repeat is conserved between the three strains, the consensus pattern was subsequently re-aligned onto each sub-sequence using a Wrap-around dynamic programming algorithm (TrlocalS from USC Sequence Alignment Package).

At this moment, there are only 3 high-coverage/high-quality genome sequences of *S. cerevisiae* strains available. Low-coverage sequences obtained with today's single-molecule sequencing techniques do not yield accurate sequencing of tandem repeats. The use of (only) three strains to assess repeat variability could lead to a high "false negative" rate, *i.e.* variable repeats that are categorized as non-variable because they do not vary between the three strains used in our analysis. To estimate this false negative rate, we calculated the number of variable repeats if all possible combinations of only 2 genome sequences are used and compared the number to that obtained by comparing all 3 genomes. This analysis yielded a rather small false negative rate of 5.5%.

The human dataset was assembled from a whole genome alignment between *Homo sapiens* (Hg18, NCBI build 36.1), *Pan troglodytes* (panTRo2, Build 2 version 1) and *Macacca mulata* (rheMac2, preliminary assembly, UCSC genome browser). The plant dataset was built from the alignment between two *Arabidopsis thaliana* accessions: Columbia full genome (version January 22 2004) and Landsberg erecta traces (Jander et al. 2002), using Blastz. The insects dataset was build from the whole genome alignment of 3 closely related species from the *Drosophila melanogaster* group (*D. melanogaster*, *D. sechellia* and *D. simulans*), downloaded from http://www.biostat.wisc.edu/~cdewey/fly_CAF1/. The bacterial *Neisseria meningitides* dataset was built from two different strains: Z2491 (Refseq NC_003116) and MC58 (Refseq NC_003112). The *Mycobacterium tuberculosis* dataset was also built from two different strains: H37Rv (Refseq NC_000962) and CDC1551 (Refseq NC_002755).

Bacterial genomes were aligned with Mercator and MAVID. The same procedure as described above was used to determine variable and non-variable repeats from these species. This resulted in 392,753 conserved repeats in the human dataset (188,769 variable and 203,984 non-variable), 29,974 conserved repeats in the plant dataset (4,524 variable and 25,450 non-variable), 15,314 conserved repeats in the drosophila dataset (1,764 variable and 13,550 non-variable), 461 conserved repeats in the *Neisseria meningitides* dataset (51 variable and 410 non-variable) and 2,856 conserved repeats in the *Mycobacterium tuberculosis* dataset (114 variable and 2,742 non-variable).

Model development and selection

All models were trained on a balanced training dataset comprising 320 of all naturally occurring repeats in the *S. cerevisiae* genome (training dataset). To select the most relevant repeat characteristics for inclusion in the final model, we applied a forward variable selection procedure using LS-SVMs with an RBF kernel. The selection criterion we used was the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) performance on the remaining 2423 repeats in the *S. cerevisiae* genome (validation dataset). Starting with the single variable that resulted in the highest AUC performance on the validation set, we iteratively added a single variable that contributed the most to optimizing this criterion, until no further increase in the AUC performance on the validation dataset was obtained. The model parameters, i.e., the regularization parameter γ and the kernel parameter σ, were tuned by optimizing the '10-fold cross-validation' performance (generalization performance) on the 320 repeats in the yeast training dataset. This procedure selected the following repeat characteristics in the order mentioned: 'number of units', 'unit length', and 'purity'. The normalization procedure in LS-SVMlab labeled all three variables as continuous. Therefore, each variable was normalized (zero mean and unit standard deviation). The optimal values for γ and σ^2 obtained after tuning were 2154.4 and 12.9 respectively, which resulted in an optimal 10-fold cross-validation performance of 92.66%. The rather small value for σ^2 suggests that the final model was able to capture a substantial amount of nonlinearity without overfitting the training data since the generalization performance was optimized. The AUC performance of this model was 98.71% for the training set and 96.20% for the validation set.

Model benchmarking

The performances of our model (SERV) and other existing methods were tested on the different datasets described above. For each method, the numbers of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) was calculated. We also computed the sensitivity (SN), specificity (SP) and Matthew's correlation coefficient (MCC) based on these values, using the following formulas.

$$SN = \frac{TP}{(TP + FN)}$$

$$SP = \frac{TN}{(TN + FP)}$$

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FN).(TN+FP).(TP+FP).(TN+FN)}}$$

ROC curves and the corresponding AUC performances performances were calculated as described previously (Hanley and McNeil 1982). Significance of differences between the AUC performances of the different models was calculated as described by Hanley and McNeil (1983).

Experimental validation of model

To insert CA repeats in the *URA3* gene, we first PCR-amplified the Hygromycin B (HYGB) resistance marker cassette from plasmid pAG34 using primers 609 and 610. This generates a product containing the HYGB marker flanked by 35 CA repeats on either end, which are in turn flanked by a 20-mer priming sequence. These priming sequences correspond to 20 nucleotides upstream and downstream of the *URA3* START codon, respectively. In a second round of PCR amplification, the former PCR product is used as a template, using primers 612 and 613. This second PCR generates the long homologous ends that are necessary to target the construct to the genomic *URA3* gene. Transformants were selected on YPD plates containing 200 μg ml^{-1} hygromycin B (Sigma Aldrich). After 3 days incubation at 30ºC, growing colonies were replica-plated onto SC plates containing 1 g l^{-1} 5-fluoroorotic acid (5-FOA, Toronto Research Chemicals Inc.) to select for Ura^- cells (Boeke et al. 1987). Correct insertion of the construct in the *URA3* gene was confirmed by PCR using primers 607 and 608. Using the flanking CA repeats, the HYGB marker was subsequently looped out by plating the cells onto SC-Ura plates. This procedure selects for HYGB loopouts that retained a number of repeats that does not alter the *URA3* reading frame. The number of repeats in these strains was subsequently determined by PCR using primers 754 and 755. The PCR products were visualized on a 2.5 % agarose gel with a 50 nt. size marker (Invitrogen), and the number of repeats was determined from the length of this PCR product. Other repeats with different repeat purity or repeat unit length were inserted in essentially the same way, except for the use of other primer pairs for the first PCR reaction (primers 738 to 753, see Supplemental Table 4).

Mutation rates were measured as described earlier (Verstrepen et al. 2005). Each switching rate was corrected for the fact that on average, only two thirds of mutation events lead to an out-of-frame mutation, and only one third of all events leads to an in-frame mutation. Changes in repeat numbers were confirmed by PCR. All experiments were repeated at least 3 times and the median number of colonies was used to calculate the mutation rate.

Cited References

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Boeke, J.D., J. Trueheart, G. Natsoulis, and G.R. Fink. 1987. 5-Fluoroorotic Acid as a Selective Agent in Yeast Molecular-Genetics. *Methods in Enzymology* **154:** 164-175.

Bray, N. and L. Pachter. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* **14:** 693-699.

Dewey, C. and L. Pachter. 2006. Mercator: Multiple whole-genome orthology map construction.

Hanley, J.A. and B.J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143:** 29-36.

Hanley, J.A. and B.J. McNeil. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148:** 839-843.

Jander, G., S.R. Norris, S.D. Rounsley, D.F. Bush, I.M. Levin, and R.L. Last. 2002. Arabidopsis map-based cloning in the post-genome era. *Plant Physiol* **129:** 440-450.

Verstrepen, K.J., A. Jansen, F. Lewitter, and G.R. Fink. 2005. Intragenic tandem repeats generate functional variability. *Nat Genet* **37:** 986-990.