Supplementary Material: Repeat or not repeat? - Statistical validation of tandem repeat prediction in genomic sequences

Elke Schaper[1,2,3] , Andrey V. Kajava [4], Alain Hauser[5], Maria Anisimova[1,2]

[1]Computer Science Department, ETH Zürich, Universitätsstrasse 6, CH-8092 Zürich, Switzerland, [2]Swiss Institute of Bioinformatics, Quartier Sorge - Batiment Genopode, CH-1015 Lausanne, Switzerland, [3]Institute of Integrative Biology, ETH Zürich, Universitätsstrasse 16, CH-8092 Zürich, Switzerland, [4]Centre de Recherches de Biochimie Macromolculaire, CNRS, University of Montpellier 1 and 2, Montpellier, France, and [5]Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland

SUPPLEMENTARY METHODS 1: SEQUENCE SIMULATION

Negative sequence set

For DNA sequences the character frequencies were taken from the Ensembl 64 assembly of the human genome:

$\pi_A = 0.292$, $\pi_C = 0.208$, $\pi_T = 0.292$, $\pi_G = 0.208$

For AA sequences the character frequencies were taken from the Ensembl 64 assembly of the human proteome:

$\pi_A = 0.069$, $\pi_R = 0.057$, $\pi_N = 0.036$, $\pi_D = 0.048$,
$\pi_C = 0.022$, $\pi_E = 0.072$, $\pi_Q = 0.048$, $\pi_G = 0.066$,
$\pi_H = 0.026$, $\pi_I = 0.043$, $\pi_L = 0.098$, $\pi_K = 0.057$,
$\pi_M = 0.022$, $\pi_F = 0.036$, $\pi_P = 0.064$, $\pi_S = 0.084$,
$\pi_T = 0.054$, $\pi_W = 0.012$, $\pi_Y = 0.026$, $\pi_V = 0.060$

Positive sequence set

Nucleic sequence was simulated on the TN93 model. The applied parameters in the notation introduced by Yang (1) are $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\beta = 0.7$, as well as equal nucleic acid frequencies $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$.

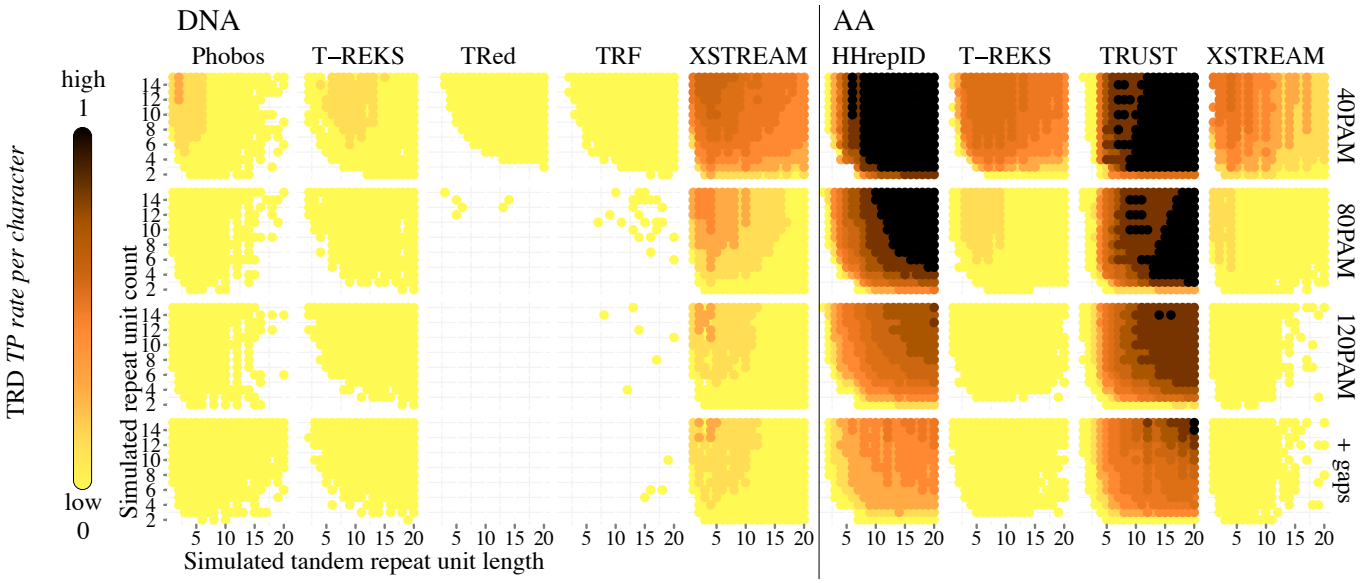
Amino acid sequence was simulated on the LG model (2). Gaps were inserted with exponentially distributed waiting times and Zipfian distributed indel length with exponent 1.821 as proposed by Chang & Benner (3). Indels placement can be anywhere outside the sequence. The maximal indel length was set to 50.

For birth-death trees, the birth rate equals the death rate.

The simulations for the positive sequence set were conducted using the ALF package for artificial simulation of genomic events (4).

REFERENCES

1. Yang, Z. (2006) Computational Molecular Evolution, Oxford University Press, Oxford oxford series in ecology and evolution edition.
2. Le, S. Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix.. Molecular biology and evolution,**25**(7), 1307–1320.
3. Chang, M. S. S. and Benner, S. A. (2004) Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments.. Journal of molecular biology,**341**(2), 617–631.
4. Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. (2012) ALF–a simulation framework for genome evolution.. Molecular biology and evolution,**29**(4), 1115–1123.

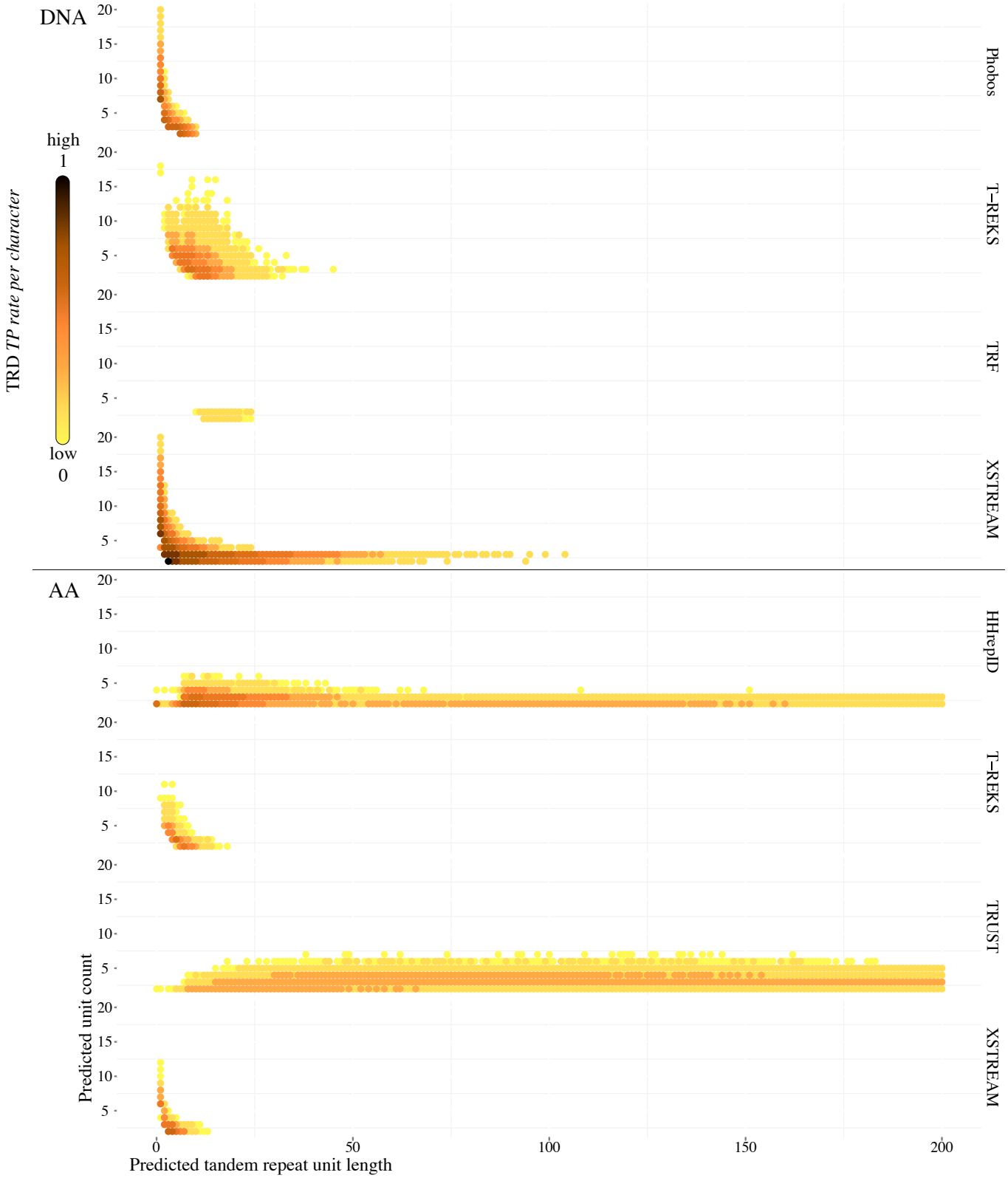


Supplementary Figure 1. True positive TR prediction on simulated DNA and amino acid sequence data for seven commonly used TRDs. Shown is the detection coverage, or TP rate per character as a function of the TR unit length (≤ 20) and the TR unit count (≤ 15). Each test set consisted of 1,000 simulated TRs. For sequence simulation, the TN93 model with equal nucleic frequencies (DNA) and the LG model (AA), respectively, were applied to ultrametric star trees. Indel events are simulated by a symmetric birth-death process with Zipfian distributed length ≤ 50 chars and an average of 0.02 indel events per site. Results are shown for three different TR divergences (40, 80 and 120 in PAM units) for nongappy TRs and additionally for gappy highly diverged TRs (120 PAM). The sequence data is the same used for Fig. 3**b-c** in the main manuscript. The detections patterns for theTP rate per character and the TP rate per repeat shown in Fig. 3**b** show high similarity. The TP rate per character multiplied with the TP rate per repeat in order to calculate the TP rate per character over all true TRs operates on the same scale as the FP rate per character and thus it is possible to compare the absolute values of these two measures.

Supplementary Table 1. False positive and false negative prediction rates for most commonly used tandem repeat detectors (TRDs).

		Negative sequence set.				Positive sequence set.											
						Short, recent TRs				Long, recent TRs				Long, diverged TRs			
		FP rate per sequence		FP rate per character		TP rate per TR		TP rate per character		TP rate per TR		TP rate per character		TP rate per TR		TP rate per character	
		default	filtered	default	filtered	default	filtered	default	filtered	default	filtered	default	filtered	default	filtered	default	filtered
DNA	Phobos	0.889	0.889	0.021	0.021	0.482	0.48	0.207	0.207	0.047	0.000	0.000	0.000	0.037	0.001	0.000	0.000
	T-REKS	0.419	0.398	0.018	0.017	0.037	0.037	0.032	0.032	0.033	0.03	0.024	0.024	0.002	0.001	0.001	0.001
	TRed	0.000	0.000	0.000	0.000	0.013	0.013	0.012	0.012	0.005	0.005	0.005	0.005	0.000	0.000	0.000	0.000
	TRF	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.009	0.009	0.007	0.007	0.000	0.000	0.000	0.000
	XSTREAM	1.000	1.000	0.472	0.414	0.949	0.926	0.631	0.619	0.828	0.294	0.224	0.224	0.792	0.043	0.026	0.026
AA	HHrepID	0.821	0.681	0.156	0.113	0.066	0.064	0.042	0.041	0.997	0.997	0.957	0.957	0.490	0.410	0.297	0.295
	T-REKS	0.104	0.098	0.002	0.002	0.053	0.053	0.035	0.035	0.277	0.277	0.238	0.238	0.001	0.000	0.000	0.000
	TRUST	0.497	0.494	0.269	0.269	0.001	0.001	0.001	0.001	1.000	1.000	0.989	0.988	0.661	0.633	0.566	0.562
	XSTREAM	0.392	0.391	0.004	0.004	0.853	0.850	0.492	0.492	0.152	0.142	0.114	0.114	0.020	0.000	0.000	0.000

FP and TP rates of TR detection on simulated negative and positive data sets. The parameters simulated short TRs were $l = 2$, $n = 15$ and for long TRs $l = 15$ and $n = 3$. Recent TRs have an average evolutionary distance of 40 PAM, diverged TRs of 120 PAM, respectively. Results are shown before and after filtering according to a 1% significance level on the model-based LRT score as a function of l and n. The FP rate of TR prediction per character was calculated as $\frac{x_{FP}}{x}$, where x_{FP} is the number of characters falsely predicted to belong to a TR in a sequence of x characters. This statistic operates on the same scale as the TP rate per character multiplied with the TP rate per repeat resulting in TP rate per character over all true TRs. Thus, the absolute values of these two measures can be compared. For each combination of l and n, a tandem repeat could theoretically begin on almost any of the 1000 characters within a sequence on the negative sequence set. Due to this multiple testing, the FP rates per sequenceare expected to be much higher than the fixed FP rate per repeat of 1%.



Supplementary Figure 2. False positive TR prediction on simulated DNA and amino acid sequence data for seven commonly used TRDs. Shown is the logarithmic FP rate per repeat as a function of the TR unit length (≤ 200) and the TR unit count (≤ 20). The test set consisted of 200,000 sequences of length 1,000, simulated by drawing random 3-mers from the human genome and proteome from Ensembl archive 64. The data corresponds to Fig. 3**a** in the main manuscript, shown here for a wider range and in higher resolution.