Additional file 1: Supplementary Method.

A quantitative sequence-based prediction of the TATA-binding protein (TBP) binding affinity for the human gene promoter

The initializing data are the 90-bp DNA sequence \{s_{-90}, s_i, s_{-1}\} immediately upstream of the transcription start site (TSS, s_0) (where \(s_i \in \{a, c, g, t\}\)).

We used the linear approximation of the three-step molecular mechanism of TBP’s binding to the [−70; −20] region of the eukaryotic gene promoters—e.g.: (i) TBP slides along DNA ↔ (ii) TBP stops at a potential TBP-binding site ↔ the DNA helix bends to the 90° angle and stabilizes the local TBP-promoter complex—as follows:

\[-\ln(K_D) = 10.9 - 0.2 \left\{ \ln(K_{\text{SLIDE}}) + \ln(K_{\text{STOP}}) + \ln(K_{\text{BEND}}) \right\}, \tag{1} \]

where 10.9 (ln units) is nonspecific TBP-DNA affinity \((10^{-5} \text{ M})\), 0.2 is the stoichiometric coefficient, and \(K_{\text{STOP}}\) is our heuristic estimate of the equilibrium constant of the second step of the TBP stops at a TBP-binding site (the maximal score value of Bucher’s position-weight matrix, the commonly accepted criterion of the canonical form of a TBP-binding site [146]); \(K_{\text{SLIDE}}\) is our heuristic estimate of the equilibrium constant of the first step of the TBP sliding along DNA; we estimated its value empirically as

\[-\ln(K_{\text{SLIDE}}) = \text{MEAN}_\text{15bp} \{0.8[\text{TA}]_{35\text{HALF}} - 3.4 \text{MGW}_\text{CENTER} - 35.1\}, \]

where \([\text{TA}]_{35\text{HALF}}\) is the frequency of dinucleotide TA within the 3’ half of the sequence being analyzed; \(\text{MGW}_\text{CENTER}\) is the arithmetical mean width of the mutator groove of the DNA helix [147]; 0.8, −3.4, and −35.1 are linear regression coefficients taken from our original experimental data [148].

In Eq. (1), \(K_{\text{BEND}}\) is our heuristic estimate of the equilibrium constant at the third step of DNA helix bending; we estimated its value empirically as

\[-\ln(K_{\text{BEND}}) = \text{MEAN}_{\text{TATA-box}} \{0.9[\text{TA, AA, TG, AG}]_{\text{FLANK}} + 2.5[\text{TA, TC, TG}]_{\text{CENTER}} + 14.4\}, \]

where 0.9, 2.5, and 14.4 are linear regression coefficients calculated from our original experimental data [149]; \(\text{MEAN}_{\text{TATA-box}}\) is the arithmetical mean value for both DNA strands of the TBP-binding site at the position of the maximal score value of Bucher’s position-weight matrix [146].

Using all the 78 possible nucleotide substitutions, \(s_{i+j} \rightarrow \xi_j\), at each j-th position \((-13 \leq j \leq 12; 3 \times 26)\) within the 26-bp DNA window centered by i-th position of the promoter DNA under study, we estimated heuristically the standard deviation of the \(-\ln(K_D)\) estimates (Eq. 1), namely:

\[
\delta = [(\Sigma_{1 \leq i \leq 26} \Sigma_{\xi \in \{a,c,g,t\}} |\ln(K_D(\{s_{i-13} \ldots s_{i+j} \xi s_{i+j+1} \ldots s_{i+12}\})/K_D(\{s_{i-13} \ldots s_{i+j} s_{i+j+1} \ldots s_{i+12}\}))|/78)^2 \right]^{1/2}. \tag{2}
\]

Thus, the preliminary result of the DNA sequence analysis is the maximal value of \(-\ln(K_D) \pm \delta\) among all the possible estimates of TBP’s binding affinity for the DNA fragment of 26-bp in length, \(\{s_{i-13} \ldots s_i \ldots s_{i+12}\}\) at the i-th position in-between −70 and −20 for both DNA chains (where \(K_D\) is the equilibrium dissociation constant expressed in moles per liter; M).

Applying Eqs. (1–2) to the cases of two mutor and ancestral alleles of a given gene, \((-\ln(K_D^{(\text{mut})}) \pm \delta_{(\text{mut})})\) and \((-\ln(K_D^{(\text{wt})}) \pm \delta_{(\text{wt})})\), we calculated Fisher’s Z-score such as

\[
Z = \text{abs}[(\ln(K_D^{(\text{mut})}/K_D^{(\text{wt})})]/[\delta_{(\text{mut})}^2 + \delta_{(\text{wt})}^2]^{1/2}.
\]

The statistical package R [150] transformed this Z-score value into the \(p\) value of the probability rate of acceptance of the hypothesis “\(H_0: \ -\ln(K_D^{(\text{mut})}) \neq -\ln(K_D^{(\text{wt})})\)” (where \(\alpha = 1 - p\) is the statistical significance level). At this statistically significant level \(\alpha < 0.05\) (i.e., at \(p > 0.95\)), we made the final decision:

**IF** \{INEQUALITY “\(-\ln(K_D^{(\text{mut})}) > -\ln(K_D^{(\text{wt})})\)” is statistically significant\},

**THEN** \{DECISION is “there is excessive expression of the mutor allele of a given gene versus the ancestral allele”\};

**ELSE** \{IF \{INEQUALITY “\(-\ln(K_D^{(\text{mut})}) < -\ln(K_D^{(\text{wt})})\)” is statistically significant\},

**THEN** \{DECISION is “there is lower expression of the mutor allele of this gene versus the ancestral allele”\},

**OTHERWISE** \{DECISION is “alteration of the expression of this gene is insignificant”\}. 
