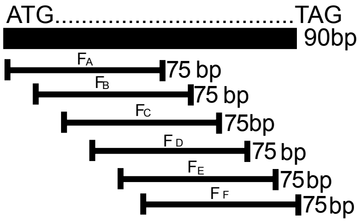
SUPPLEMENT A-1

Determination of CI and CI thresholds

Define sORF in intergenic regions



Calculate Coding Index (CI):

$$CI = \frac{P(coding|F_A) + P(coding|F_B) \dots \ +P(coding|F_F)}{7}$$

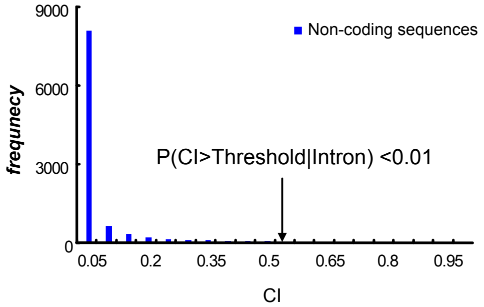
Define coding sequences

If (CI value of an sORF > the threshold), the the sORF is a coding sequence

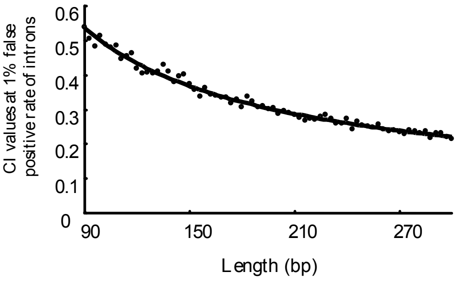
Generate random sequences for coding and non-coding sequences with different length

Estimate Coding Index (CI)

Determine CI threshold based on intron random seq at a particular length



Define CI thresholds



SUPPLEMENT A-2

Determination of CI thresholds based on simulation studies of CI values of NCDS-like random sequences

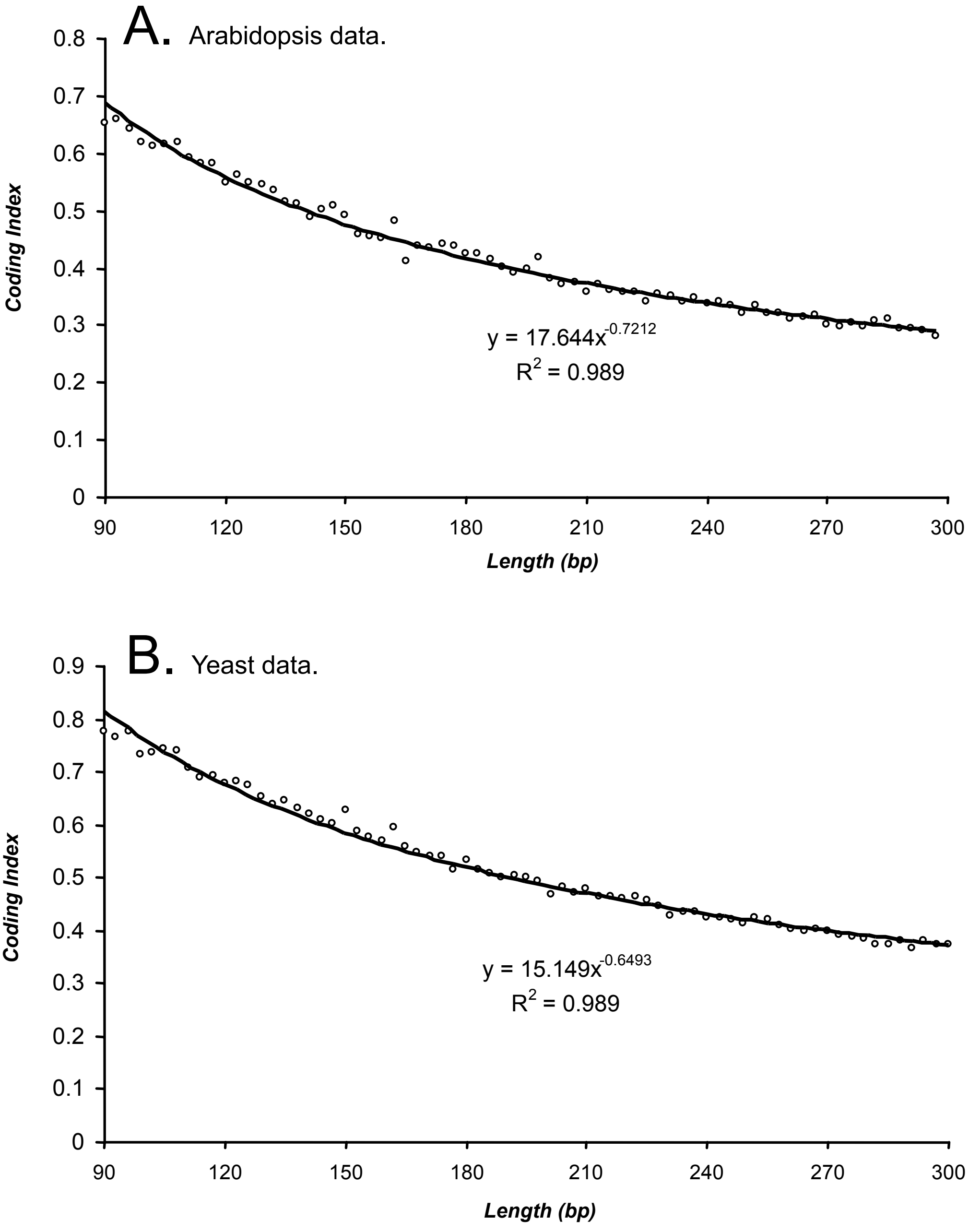
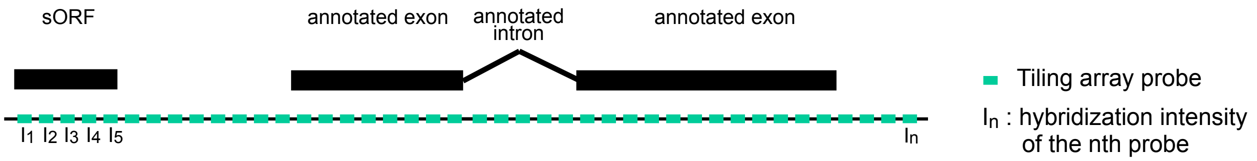


Figure legend: The CI thresholds for (A) Arabidopsis thaliana and for (B) yeast. Each circle represents the CI threshold value of a particular size class at 1% false positive rate. The false positive rates were determined based on the distributions of the CI values of 100,000 random NCDS-like sequences for each size class (ranging from 90-300 nucleotides with 3 nt increment). The CI threshold values are fitted with the power. The equation for the fit and correlation coefficients are shown.

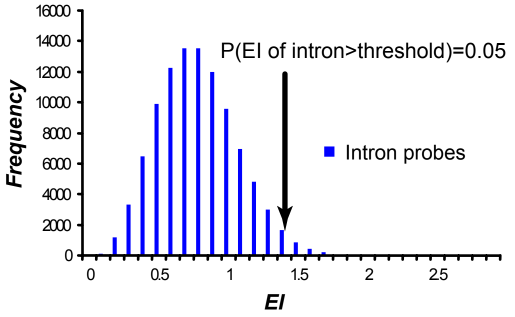
SUPPLEMENT A-3

Determination of EI and EI thresholds



Determine EI threshold for a sequence with a given number of probes based on random intron probes

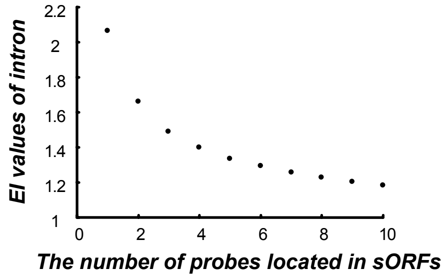
1. Randomly pick 10000 intron probes
2. Generate EI value distribution
3. Define the EI value > 95% of intron probes, this is the threshold.



Calculate Expression Index (CI):

$$EI(sORF)=\frac{I_1+I_2+I_3+I_4+I_5}{\ }$$

Define EI threshold fit



Define transcription

If (EI value of an sORF > the threshold), then the sORF is regarded as transcribed.

SUPPLEMENT A-4

Determination of EI thresholds based on a given number of probes based on random intron probes

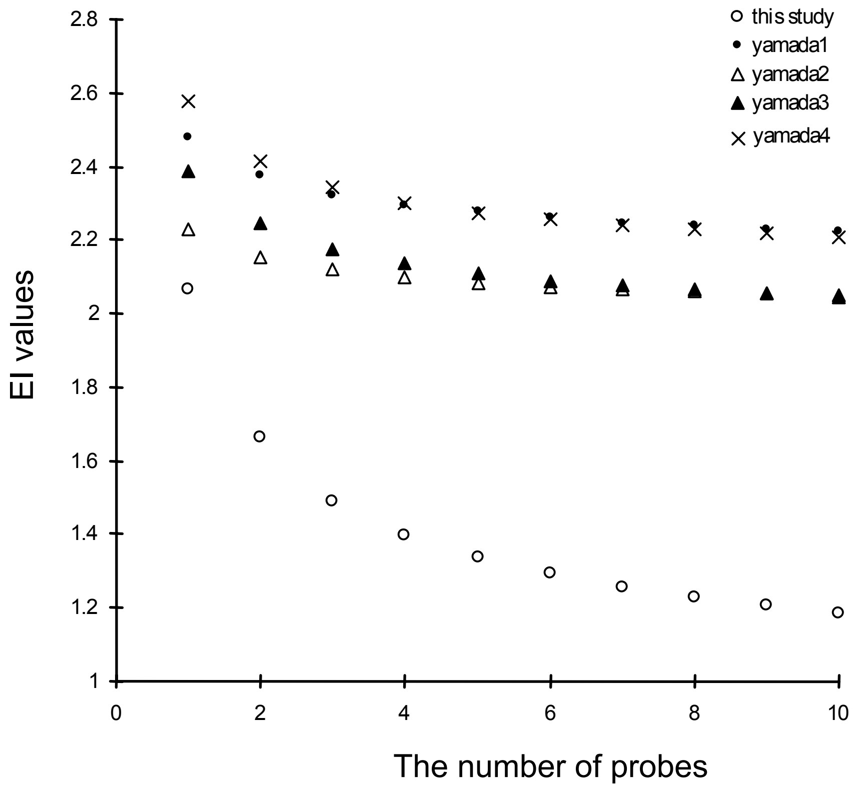


Figure legend: Each point represents the EI threshold value of a given number of probe(s) at 5% false positive rate. The false positive rates were determined based on the EI value distributions of 100,000 randomly sampled intron probes.