Appendix A. Hybridization Protocol

Total RNA was isolated from 21 day-old inflorescences and 1 μg RNA was used for cDNA synthesis using oligo-dT-T7 primers (MessageAmp©II aRNA Kit, Ambion). After T7-based linear wamplification, 8 μg of the resulting RNA was converted into cDNA, fragmentized and labeled using the WT Double-Stranded cDNA Synthesis Kit and the WT Double-Stranded DNA Terminal Labeling Kit (Affymetrix) according to manufacturer's instructions. These dsDNA targets were hybridized to *Arabidopsis* TilingR arrays (Affymetrix) that were washed and stained using the GeneChip©Fluidics Station 450 (Affymetrix) and scanned using the GeneChip©Scanner 3000 (Affymetrix) according to manufacturer's instructions.

Appendix B. Intensity distribution for *A. thaliana* tiling arrays

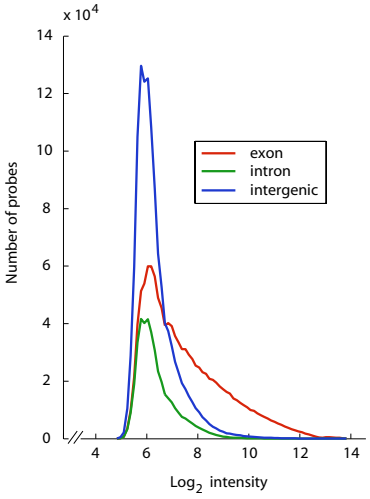


Figure 8. Intensity distributions for probes mapped to exons, introns and intergenic regions, respectively. The mapping to gene models is based on TAIR's genome annotation version 7. Probes (partly) mapping to more than one sequence type were excluded.

Appendix C. Training the HMSVM for transcript identification

Given n training examples $(\boldsymbol{x}^{(i)}, \boldsymbol{\sigma}^{(i)})$, $i = 1, \ldots, n$, we enforce a large margin of separation between the correct path $\boldsymbol{\sigma}^{(i)}$ and *any* other wrong[c] path $\overline{\boldsymbol{\sigma}} \neq \boldsymbol{\sigma}^{(i)}$ by solving the

[c]We consider a path wrong, when its loss is greater than zero.

following quadratic optimization problem:

$$\min_{\boldsymbol{\theta},\,\boldsymbol{\xi}\geq\mathbf{0}} \quad \frac{1}{n}\sum_{i=1}^{n}\xi^{(i)} + C\sum_{k\in\mathcal{S}}\sum_{l=1}^{L-1}(\theta_{k,l}-\theta_{k,l+1})^2 \qquad \text{subject to}$$

$$F_\theta(\boldsymbol{\chi},\boldsymbol{\sigma}^{(i)}) - F_\theta(\boldsymbol{\chi}^{(i)},\overline{\boldsymbol{\sigma}}) \;\geq\; \ell(\boldsymbol{\sigma}^{(i)},\overline{\boldsymbol{\sigma}}) - \xi^{(i)} \quad \forall \ell(\overline{\boldsymbol{\sigma}},\boldsymbol{\sigma}^{(i)}) > 0, i = 1,\ldots,n,$$

where C is the regularization parameter and $\ell(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ the loss function that quantifies the difference between the two state sequences. The description of how the above optimization problem can be solved, goes beyond the scope of this contribution and the reader is referred to previous work on this problem[1,19,13].

Appendix D. Position-specific sequence effects

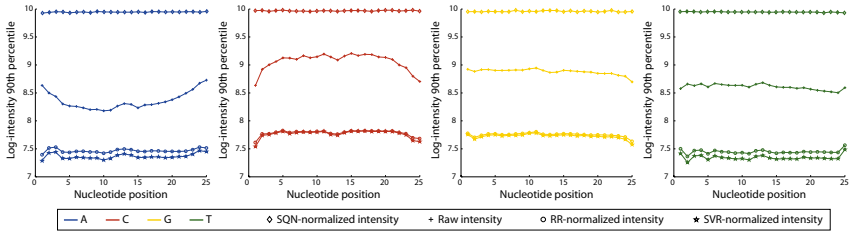


Figure 9. Position-specific quantile plots for raw and normalized intensities (see inset). The effect of each nucleotide is displayed on one panel; the position within the probe sequence is indicated on the x-axis and the 90th intensity percentile on the y-axis.

Position-specific mono-nucleotide effects can be graphically displayed with so-called quantile plots.[15] Such plots are shown in Figure 9 for the 90th intensity percentile which is expected to be enriched for measurements of highly expressed transcripts.

Appendix E. Intensity depends on distance from 3' end

In addition to probe sequence effects we found another major cause of intensity deviations from ideal constant transcript intensities. Intensity is generally higher near the 3' transcript end. The most likely explanation for this is a bias in the T7-based linear amplification which starts at the polyA tail attached to the 3' transcript end. Such a bias is expected if amplification terminates with a certain probability before the 5' transcript end is reached. The steep increase in intensity from the 3' end is possibly an artifact of the genome annotation for which a problem with unrealistically long UTRs has been known.

A correction for this bias can be simultaneously learned with sequence normalization by providing SVR or RR with additional features. With features describing the distance from the polyA tail, a piecewise-linear function can be estimated to quantify the distance

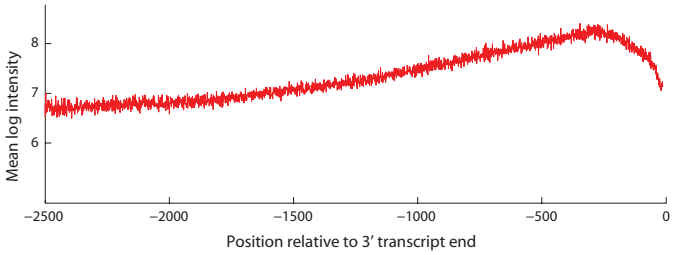


Figure 10. Hybridization intensity on average decreases with increasing distance from the 3' transcript end. This is most likely an artifact of T7-based linear amplification.

effect on deviation from the transcript intensity. However, such features can only be computed if transcripts are known *a priori* (e.g. for splice form detection), and hence this extension is of limited use for the identification of novel transcripts. Nevertheless, we conducted experiments to quantify the 3' amplification bias relative to probe sequence effects. (cf. Appendix E). When distance features are given for training and testing, we obtained $T_1 = 0.44, T_2 = 0.33$ for SVR and $T_1 = 0.46, T_2 = 0.31$ for RR–a relative improvement of $\approx 25\%$.