Supplementary information

Shoji Kawaguchi 1, Kei Iida 1, Erimi Harada 1, Kousuke Hanada 1, 2, Akihiro Matsui 2, Masanori Okamoto 2, 3, Kazuo Shinozaki 2, Motoaki Seki 2, and Tetsuro Toyoda 1*

1 Bioinformatics and Systems Engineering division, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan.
2 RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan.
3 Current Address: Center for Plant Cell Biology, Department of Botany and Plant Sciences, University of California, Riverside, 3119A IIGB, Riverside, CA 92521, USA.

ARTADE2 MATHEMATICS

Transcript structure model based on multiple tiling arrays

A base sequence of size \( n \), \( x = (x_1, x_2, \ldots, x_n) \); \( x_i \in \{A, T, G, C\} \) has random variables of genome state \( S = (S_1, \ldots, S_n) \); \( S_i \in \{0, 1, \ldots, 25\} \). We let \( s = (s_1, \ldots, s_n) \) be a random number vector of \( S \). Variable \( S \) takes the value whether the state is the start, end, or interval of an exon, intron, or outer. Toyoda and Shinozaki (2005) defined the transition diagram of these states (Figure S2). Then, we defined a map \( \Lambda : S \rightarrow \{0, 1\} \) as

\[
\Lambda(s) = \begin{cases} 
1 & \text{if } s \text{ is exon}, \\
0 & \text{otherwise}.
\end{cases}
\]

and set \( y(s) = (\Lambda(s_1), \ldots, \Lambda(s_n))' \), where ' indicates a transpose. Then, we obtained the exon-intron matrix \( K \) from \( y \) as follows:

\[
K = (d_{ij}) = 2y(s)y(s)' - 1',
\]

where \( 1 \) is \( n \times 1 \) vector of size \( n \) whose all elements are 1.

Suppose that there are \( m \) probes \( P_{b_1}, \ldots, P_{b_m} \) in the \( n \) interval. Here we limited the probe whose values of more than three replicates in at least one condition exceed \( e^{0.754} \), which is the lowest 1% value of exon expression in the training data (RIKEN Arabidopsis full length cDNA (RAFL) mapped on chromosome 1 plus strand). We let \( [a_k, b_k] \) \((1 \leq a_k \leq b_k \leq n, b_{k-1} < a_k) \) be the right and left end positions of probe \( P_{b_k} \) in the \( n \) interval. We observed the probe expression in several experiments under certain conditions. Then, we let \( h \) be the number of all experiments and \( f_k = (f_{k1}, \ldots, f_{kn}) \) be a vector of tags for each experiment at probe \( k \). Pearson's correlation coefficient \( \gamma_{kl} \) between expressions values of \( P_{b_k} \) and \( P_{b_l} \) is given by the following equation:

\[
\gamma_{kl} = \frac{\sum_{i=1}^{n} (f_{ki} - \bar{f}_k)(f_{li} - \bar{f}_l)}{\sqrt{\sum_{i=1}^{n} (f_{ki} - \bar{f}_k)^2} \sqrt{\sum_{i=1}^{n} (f_{li} - \bar{f}_l)^2}}.
\]

where \( \bar{f} \) is sample mean of \( f \). We then obtained a correlation matrix of size \( m \times m \), \( R = (\gamma_{kl}), k, l = 1, \ldots, m \). Here we defined a threshold parameter \( \theta \) and translated the correlation matrix \( R \) to expanded matrix \( C_{\theta}(R) = (c_{ij}), i, j = 1, \ldots, n \) of size \( n \times n \), where

\[
c_{ij} = \begin{cases} 
2I_{\theta} (\gamma_{kl}) - 1 & \text{if } a_k \leq i \leq b_k, a_l \leq j \leq b_l \text{ and } k \neq l, \\
0 & \text{otherwise}.
\end{cases}
\]

\[
I_{\theta}(z) = \begin{cases} 
1 & \text{if } z > \theta, \\
0 & \text{otherwise}.
\end{cases}
\]

* to whom correspondence should be addressed
Let,

\[
\begin{align*}
{n_{KE}} &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij}\neq 0 \land d_{ij}=1} \\
{n_{KO}} &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij}\neq 0 \land d_{ij}=-1} \\
{n_{CE,KE}} &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij}=1 \land d_{ij}=1} \\
{n_{CO,KO}} &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij}=-1 \land d_{ij}=-1},
\end{align*}
\]

and define the Positional Correlation matrix Score (PCS) between \( C \) and \( K \), \( PCS(C, K) \) as

\[
PCS(C, K) = \frac{en_{CE,KE} + n_{CO,KO}}{en_{KE} + n_{KO}},
\]

where coefficient \( e \) indicates a weight parameter of the exon. We set \( PCS(C, K) \) to 0 exceptionally, if \( n_{KE} = 0 \). From the PCS, we define a Correlation Matrix Score (CMS) as follows:

\[
CMS(s, R, \theta) = n \log PCS(C \theta(R), K(s))
\]

If \( s \) is given, we can calculate the number of exons and introns and the size of the transcript. Let \( N_E \) and \( N_I \) be the number of exons and introns, respectively. Then, we let \( r^E_j \) and \( r^I_j \) be the lengths of the exon \( j (= 1, \ldots, N_E) \) and intron \( k (= 1, \ldots, N_I) \), respectively. Here we assumed that the exon and intron lengths follow the \( G \)-th Gaussian mixture distribution, and their probabilistic density functions \( p_E \) and \( p_I \) are described as follows:

\[
p_E(r) = \sum_{i=1}^{G} g_i^E \psi_i^E(r), \quad p_I(r) = \sum_{i=1}^{G} g_i^I \psi_i^I(r),
\]

where \( \psi_i^E \) and \( \psi_i^I \) are probabilistic density functions of a normal distribution with mean \( \mu_i \) and variance \( \sigma_i^2 \), and \( \mu_i \) and \( \sigma_i^2 \), respectively. Weights of functions \( g_i^E \) and \( g_i^I \) satisfy the following:

\[
\sum_{i=1}^{G} g_i^E = 1, \quad \sum_{i=1}^{G} g_i^I = 1.
\]

By using these probabilistic densities, scores of exon and intron lengths \( ES \) and \( IS \) are defined as follows:

\[
ES(s) = \sum_{j=1}^{N_E} \log p_E(r^E_j), \quad IS(s) = \sum_{j=1}^{N_I} \log p_I(r^I_j).
\]

The third score Markov Transition Score (MTS) is given by probability of nucleotides sequence under the assumption of Markov process as:

\[
MTS(s, x) = \sum_{i=1}^{n} \log P(S_i = s_i, x_i|S_{i-1} = s_{i-1}, x_{i-1}).
\]

We assumed that the genome structure follows a logistic model based on the four scores. Then, the conditional occurrence probability of state \( S \) is given by the following equation:

\[
P(S = s|x, R, \theta) = \frac{\exp \{ \alpha MTS(s, x) + \beta CMS(s, R, \theta) + \xi IS(s) + ES(s) \}}{Z(S)},
\]

where \( Z(S) \) is the normalized constant. Therefore, the genome structure \( \hat{s} \) is obtained by the decision rule of

\[
\hat{s} = \arg\max_s P(S = s|x, R, \theta) = \arg\max_s \{ \alpha MTS(s, x) + \beta CMS(s, R, \theta) + \xi IS(s) + ES(s) \}.
\]

The scores of equation (18) is comparable to scores of ARTADE proposed by Toyoda and Shinozaki (2005). However, the new model uses novel score based on the correlation matrix instead of expression value of tiling array probes whose expressions over the threshold and scores.
of exon and intron lengths are changed to Gaussian mixture distributions from log-normal distribution. Moreover, coefficients of weight for each score are set.

A structure is obtained by the maximization of equation (16). As a practical measure, we must determine the position of the structure and its size \( n \) from the genome in advance. We therefore determined a start point of the estimation from the correlation matrix of tiling arrays and estimated a structure with expanding probability space by dynamic programming (DP).

We used window size \( W \) and an initial threshold \( \theta^0 \) (Table S6). Let \( R_t \) be a correlation matrix of probes size \( w \times w \) which are included in the \( t \) and \( t + W - 1 \) positions. Here, we restrict \( t \) where \( w \geq 10 \). From \( R_t \), we can calculate the following values:

\[
n_{It} = \sum_{k=1}^{w} I_{\theta t} \left( \frac{1}{w} \sum_{l=1}^{w} (1 - I_{\theta t} (\gamma_{kl})) \right),
\]

\[
n_{Et} = w - n_{It},
\]

\[
a_k = 1 - I_{\theta t} \left( \frac{1}{w} \sum_{l=1}^{w} (1 - I_{\theta t} (\gamma_{kl})) \right),
\]

\[
a = (a_1, \ldots, a_k)',
\]

where \( \theta_t \) is an occupation threshold whether the probe is not exon (Table S6). If the occupation of exon \( n_{Et}/w \) exceeds threshold \( \theta_{Et} \) (Table S6), we can calculate:

\[
V_t = (v_{kl}) = a a',
\]

\[
Y_t = (y_{kl}) = (I_{\theta t} (\gamma_{kl})),
\]

\[
n_{Et} = \sum_{k=1}^{w} \sum_{l=1}^{w} 1_{\gamma_{kl} = 1},
\]

\[
n_{Eall} = \sum_{k=1}^{w} \sum_{l=1}^{w} 1_{\gamma_{kl} = 1},
\]

\[
n_{I} = \sum_{k=1}^{w} \sum_{l=1}^{w} 1_{\gamma_{kl} = 1},
\]

\[
n_{Iall} = \sum_{k=1}^{w} \sum_{l=1}^{w} 1_{\gamma_{kl} = 1}.
\]

Then, an adjacent value between \( V_t \) and \( Y_t \), \( F(V_t, Y_t) \) is given by the following equation:

\[
F(V_t, Y_t) = \frac{n_{Et}}{n_{Eall}} \times \frac{n_{I}}{n_{Iall}}.
\]

In the case of \( n_{Iall} = 0 \), we replaced \( F(V_t, Y_t) \) by

\[
F(V_t, Y_t) = \frac{n_{Et}}{n_{Eall}}.
\]

If max\( F(V_t, Y_t) < \theta_g \), we finish the prediction in this interval. Otherwise we determine the position \( \hat{t} \) where \( F(V_t, Y_t) \) is maximal as

\[
\hat{t} = \arg\max_t F(V_t, Y_t).
\]

The probe \( \hat{t} \), from where the prediction of a transcript structure starts is given by the following equation:

\[
\hat{t} = \arg\max_{\hat{t}} \max_{j-i} \left( \sum_{k=1}^{j} I_{\theta t} (\gamma_{kl}) = 1; t_s - w + 1 \leq i \leq j \leq t_s + w - 1 \right).
\]

Consequently, the starting position \( \hat{t}^0 \) is set to

\[
\hat{t}^0 = \left[ \frac{a_{\hat{t}} + b_{\hat{t}}}{2} \right],
\]

where \( a_{\hat{t}}, b_{\hat{t}} \) are start and end positions of probe \( \hat{t} \), respectively and operator \( \lfloor \rfloor \) is a floor function.
From position \(i\), we expanded region size \(n\) by DP and estimate structure \(s\). The expansion is carried in two directions. One is a direction to 3' end from \(i\) and another is direction to 5' end. We differently estimate parameters for the two expansions (Table S6). Here we note about the case of the expansion to 3' end. To simplify the equation, replace \(P(S = s|x, R, \theta)\) with \(P(S = s)\). Firstly, we consider \(i = 3'\) is exon i.e. \(\tilde{s}_0 = 4\) (Figure S2). Then, we calculated

\[
\tilde{s}_2^2 = \arg\max_{s \in \{0, \ldots, 25\}} P(S_2 = k, S_1 = s, S_0 = \tilde{s}_0).
\]

(34)

From the obtained \(\tilde{s}_2^2\), we calculated

\[
\tilde{s}_3^k = \arg\max_{s \in \{0, \ldots, 25\}} P(S_3 = k, S_2 = s, S_1 = \tilde{s}_2^k, S_0 = \tilde{s}_0),
\]

(35)

\[
\tilde{s}_4^k = \arg\max_{s \in \{0, \ldots, 25\}} P(S_4 = k, S_3 = s, S_2 = \tilde{s}_2^k, S_1 = \tilde{s}_3^k, S_0 = \tilde{s}_0).
\]

(36)

Consequently, \(i\)-th values are given by

\[
s_i^k = \arg\max_{s \in \{0, \ldots, 25\}} P(S_i = k, S_{i-1} = s, S_{i-2} = \tilde{s}_{i-2}, \cdots, S_1 = \tilde{s}_2^k, S_0 = \tilde{s}_0).
\]

(37)

The structure from the position \(i, \tilde{s}_i = (\tilde{s}_0, \tilde{s}_1, \ldots, \tilde{s}_i)\) is given by the back-scanning as

\[
\begin{align*}
\tilde{s}_i &= 25, \\
\tilde{s}_{i-1} &= \tilde{s}_i^k \\
& \vdots \\
\tilde{s}_1 &= \tilde{s}_2^k \\
\tilde{s}_0 &= 4.
\end{align*}
\]

(38)

We can obtain structure \(\tilde{s}_i\) by the DP matching. However, we do not know the optimal back-scanning start point \(i\) because the expansion is infinitely continued and \(P\) expands its probabilistic space with increasing the structure size. For the problem, we stopped the expansion if terminated states are continually estimated \(T_e\) times (Table S6). The stop point \(j\) can be formulated as follows:

\[
\arg\max_{s \in \{0, \ldots, 25\}} s_i^k = 25, \quad \text{for} \quad i = j, j-1, \ldots, j - T_e + 1.
\]

(39)

The optimal back-scanning start point may not correspond to \(j\). Some points may be candidates for optimal back-scanning start point. Here, scores used in the logistic model are considered to increase with rising structure size \(n\) at order \(o(n)\) if we take experiments of scores. Therefore, we define a new score \(AS\) for comparing structures of different sizes with bias factor \(B\) (TableS6) as follows:

\[
AS(s) = \alpha MT S(s, x) + \beta CM S(s, R, \theta) + \xi I S(s) + ES(s) + B.
\]

(40)

Then, an optimal back-scanning start point \(\tilde{i}\) is given by

\[
\tilde{i} = \arg\max_{0 < i \leq j} \frac{AS(\tilde{s}_i)}{i}.
\]

(41)

We here restrict comparing points to \(\{i\} \) where states were consecutive estimated to 25 more than \(Q\) times (Table S6) i.e. \(\tilde{s}_i = \tilde{s}_{i-1} = \cdots = \tilde{s}_{i-Q+1} = 25\).

A direct calculation of \(P(S_i)\) is difficult. However, comparing \(P(S_i = k, S_{i-1} = s, \tilde{s}_{i-2})\) and \(P(S_i = k, S_{i-1} = s', \tilde{s}_{i-2}); \tilde{s}_{i-2} = (\tilde{s}_{i-2}, \tilde{s}_{i-3}, \ldots, \tilde{s}_0)\) is simple because we must only calculate a score variation in the right formula of equation (18). The expansion to 5' end is also executed by replacing the back-scanning start, 25 with 0.

**Optimization and parameters estimation**

The proposed transcript structure model has many unknown parameters. We must therefore estimate parameters using known gene structures and optimizing parameters iteratively in the structure prediction. For parameter \(W, \theta_E, \theta_F\) and \(\theta_R\), we arbitrary set values. We used 2,813 RIKEN *Arabidopsis* full length cDNA (RAFL) mapped on chromosome 1 plus strand for training data of the method. First, means and variances of Gaussian mixture distribution of exon and intron lengths were estimated by an expectation-maximization (EM) algorithm (Dempster et al. (1977)) using under 5000-bp length exons and under 1000bp length introns. We fix function number \(G\) to 10 in equation (12).
The prediction is iteratively optimized by alternately estimating the structure and a correlation threshold $\theta$. First, we set the initial threshold $\theta_0 = 0.22$. Let $\theta^*$ be estimated threshold of $\varepsilon$-th iteration. After the detection of start window $\ell - l + W - 1$ and getting $\hat{V}_i$ and $Y_{\ell}$ from $\theta^*$, the threshold is re-estimated as

$$
\theta^* = \frac{\text{med}(\gamma_{kt} | v_{kt} = 1 \cap y_{kt} = 1) + \text{med}(\gamma_{kt} | v_{kt} = 0 \cap y_{kt} = 0)}{2}.
$$

Using re-estimated $\theta^*$, the structure $s^*$ is predicted by maximizing $P(s)$ and searching the optimal back-scanning start point. Here, we restrict the size of $s^*$ where $s^*$ is exon, intron or intergenic within 100 base distance from 3' end and 5' end. Then, $\theta^*$ is updated to

$$
\theta^{*+1} = \frac{\text{med} (c_{ij}^{-1} | c_{ij} \neq 0 \cap \Lambda(s_j^*) \cdot \Lambda(s_j^*) = 1)}{2} + \frac{\text{med} (c_{ij}^{-1} | c_{ij} \neq 0 \cap \Lambda(s_j^*) \cdot \Lambda(s_j^*) = 0)}{2},
$$

where $c_{ij}^{-1}$ is positional correlation $\gamma_{kl}$ at $a_k \leq i \leq b_k$ and $a_l \leq j \leq b_l$, and $c_{ij}^{-1}$ is 0 if $\gamma_{kl} < 0$. The structure and parameter estimation is continued until

$$
PCS(C_1(R), K(s^{*+1})) < PCS(C_0(R), K(s^{*})).
$$

Then, we began to predict a new transcript structure in the genome, excluding the already estimated region. Remaining parameters were set so that the prediction accuracy of 2,813 RAFL gene models is maximal. In the parameter estimation, we did not embed factor analysis described in the next section. Table S6 shows estimated parameters.

**Use of factor analysis to remove concatenating of different transcripts**

The transcript predicted with ARTADE2 may not be consummate in some cases if several transcripts located continuously in the genome have highly correlated expression. The positional correlations in the region appear to have one transcription. The problem is overcome through a factor analysis of the correlation matrix of tiling array probes. Let $Pb_1, \ldots, Pb_m$ be probes within predicted structure of $s$ and $v_k = (v_{1k}, \ldots, v_{mk})$ be expression value at $k$-th experiment. We assumed that expression value $v_k$ are modeled as

$$
v_k = Af_k + u_k,
$$

where $m \times q$ matrix $A = (a_{ij})$ is called a factor loading matrix. Vectors $f_k = (f_1, \ldots, f_q)$ and $u_k = (u_1, \ldots, u_m)$ are called a common factor vector and a unique factor vector, respectively, and are not correlated with each other. We assume that number of factor $q$ is 2. Then the factor loading matrix is estimated with the factor analyzing method. If the estimated second factor loadings $a_{2} = (a_{12}, a_{22}, \ldots, a_{m2})$ construct a specific structure apart from the first factor loadings $a_{1}$, the structure at where second factor-loadings becomes high may differ from that of the first principle model.

A was set as the first and second eigenvectors of $v$. Then, matrix $A$ is reestimated by a maximum likelihood estimation. Finally, the obtained matrix $A$ is obliquely rotated by a criterion of Promax method through an orthogonal rotation of the Varimax method. We select the Promax method because the method can create different factors even if factors correlate mutually. Using estimated matrix $A$, we check the possibility for existence of multiple or spatial structures.

Let $n$ be the size of predicted interval of ARTADE2 and $l = (l_1, \ldots, l_m)$ be a vector of center positions of probes. Define subsets $\omega_1, \omega_2 \subset \{1, \ldots, m\}$ as

$$
\omega_1 = \{i \mid a_{ii} > \theta_f\},
$$

$$
\omega_2 = \{i \mid a_{1i} > \theta_f\}.
$$

Then, we calculated the sample mean and variance of probe positions for $\omega_1$ and $\omega_2$ as

$$
m_1 = \frac{1}{\#\omega_1} \sum_{i \in \omega_1} l_i,
$$

$$
m_2 = \frac{1}{\#\omega_2} \sum_{i \in \omega_2} l_i,
$$

$$
\sigma_1^2 = \frac{1}{\#\omega_1} \sum_{i \in \omega_1} (l_i - m_1)^2,
$$

$$
\sigma_2^2 = \frac{1}{\#\omega_2} \sum_{i \in \omega_2} (l_i - m_2)^2,
$$

where $\#\omega$ means number of elements of set $\omega$. We considered the estimated $s$ as multiple structures, if it is satisfied that

$$
\#\omega_2 \geq M,
$$

$$
|m_1 - m_2| > L,
$$

$$
\frac{\sigma_2}{\#\omega_2} < \theta_t.
$$
Therefore, we must divide the estimation region for the multiple structures. In the case of $m_1 < m_2$, the division point $l_d$ is settled as follows:

$$i_1 = \arg\max_{i \in \{1, \ldots, m\}} \left(\sum_{j=1}^{i} a_j - \sum_{j=i+1}^{m} a_j\right),$$

(55)

$$i_2 = \min \{i \mid i > i_1, a_{i2} > \theta_f\},$$

(56)

$$l_d = (l_{i1} + l_{i2})/2.$$  

(57)

If $m_2 \leq m_1$, the estimation is reversed as:

$$i_1 = \arg\max_{i \in \{1, \ldots, m\}} \left(\sum_{j=i+1}^{m} a_j - \sum_{j=1}^{i} a_j\right),$$

(58)

$$i_2 = \max \{i \mid i < i_1, a_{i2} > \theta_f\},$$

(59)

$$l_d = (l_{i1} + l_{i2})/2.$$  

(60)

Consequently, we restart over the prediction of transcription both $(l_d + 1, n)$ and $(1, l_d)$ regions. Parameters of the factor analysis are also adjusted to maximize prediction accuracies of training data set by 2,813 RAFL on chromosome 1 plus strand. Table S7 lists the parameters of the factor analysis.

**FACTOR ANALYSIS FOR DETECTION OF REGIONS HAVING ALTERNATIVE ISOFORMS**

The factor analysis can also be applied to detect regions altered by selecting of transcription start or termination sites whose patterns differ among different conditions. Predicted transcript structure is factorized by promax method. Here, we use only probes which are included in exon regions and for which standard deviation of the expression value has over 50.0 for the factor analysis. Set probe id $1 \ldots m$ to these probes in order of genome position. Factor number $q$ is estimated by using Minimum Average Partial method (Velicer et al. (2000)). However, we restrict the maximum factor number to 5 because only less than 0.02% of annotated gene loci (The Arabidopsis Information Resource (TAIR), ver.9) have more than 6 kinds of alternative gene models.

We detect specifically expressed regions from the obtained $m \times q$ factor loading matrix $A = (a_{ij})$ of positional correlations in these probes. If the estimated factor number $q$ is larger than 2, we adapt the following algorithm.

**Clustering of regions with high factor loadings**

Calculate positional correlation $(\gamma_{ij})$ of every probe pairs from 1 to $m$.

for $i = 2, \ldots, q$

$\gamma = 0.0$

$\gamma_{\max} = \min \left(1, \max_{k=1, \ldots, m} \gamma_{ik}\right)$.

for $j = 1, \ldots, m$

$n_j = 0.2 + \left(\gamma_{\max} - \min(1, \gamma_{ij})\right)^2$

if $\arg\max_{l \in \{1, \ldots, q\}} a_{jl} = i \wedge a_{ji} > 0.45$

$t = t + 1$

$C_{ij} = \{j\}$

$B_t = j$

$D_t = n_j$

end if

end for $j$

$t = c$

while $t > 1$

$d_{\min} = \infty$

for $\{(j, k) \mid j < k ; j, k = 1, \ldots, c\}$

$\frac{\sum_{l C_j} \sum_{s C_k} \gamma_{ls}}{\#C_j \#C_k} < 0.4$

continue

end if
In the algorithm, operation $|l - B_{tmp}|$ means taking absolute value and $\# C$ means number of elements in cluster $C$. The clustering algorithm detects region of high factor loadings (over 0.4) with high density. Value $D$ of cluster $C$ is named discreteness.

**EXPANSION OF ARTADE2 FOR MRNA-SEQ DATA**

In this section, we expand the ARTADE2 method to adapt to mRNA-Seq data observed by a Next Generation Sequencer. Unlike tiling arrays, mRNA-Seq data is not fixed position in position on the genome. Therefore, we defined 10-bases-grids on the genome sequences, and then counted mRNA-Seq tags within each grid as a pre-process for applying ARTADE2. We used all mRNA-Seq tags although probes of low values were eliminated in tiling array study. Note that we diminished or spitted a side of a grid if the grid had positions on where no mRNA-Seq data is not fixed in position on the genome. Therefore, we defined 10-bases-grids on the genome sequences, and then counted mRNA-Seq data is not fixed in position on the genome. Therefore, we defined 10-bases-grids on the genome sequences, and then counted mRNA-Seq tags within each grid as a pre-process for applying ARTADE2. We used all mRNA-Seq tags although probes of low values were eliminated in tiling array study. Note that we diminished or spitted a side of a grid if the grid had positions on where no mRNA-Seq data have expressions. For the expansion, we set $c_{ij}$ as $-1$ in equation (4) if position $i$ or $j$ does not have tags in all conditions. Parameters of ARTADE2 are slightly changed so that $B$ of expansion to 5’end (Table S6) is 0.0 from $-50.0$ and $M$ (Table S7) is 35 from 10. We also reduce $\theta_5$ to 0.3 (Section: Transcript structure model based on multiple tiling arrays, Table S6) to increase the recall of exons of TAIR9 gene models by just about same recall of Cufflinks.

**COMPARING PREDICTED GENE MODELS WITH “OMIC” DATA SETS**

For assessing novel genes found with ARTADE2, we used several public data sets of “omic” analyses results. We focused on transcriptome, degradome, and proteome data. For transcriptome data, we used high-throughput sequencing results of mRNA-seq samples and small RNA samples (NCBI SRA accession numbers: SRX002554, SRX002508, Lister et al. (2008)). We also used our own cap analysis of gene expression (CAGE, Kodrius et al. (2006)) tags for RNA samples of untreated plants or plants subjected to drought conditions or ABA treatment (GEO accession numbers: GSE9646, GSE15700, GSE26074). For degradome data, we used analysis of 5’ end tag sequences of uncapped RNAs derived by a method called parallel analysis of RNA ends (PARE, German et al. (2009)). We used the degradome data.
set with an SRA accession number of SRP000713 (German et al. (2009)). We mapped small RNA, CAGE, and degradome tags to the *Arabidopsis thaliana* genome sequence with “soap” (Li et al. (2009)). We used “tophat” (Trapnell et al. (2009)) to map RNA-seq tags to the genome because we sought to obtain coverage information for known or novel exon-exon junctions by RNA-seq tags. We used soap and tophat software with parameters that permit two base mismatches against the genome sequences.

For mass spectrometry outputs for proteomes (mass), we used data sets with EBI PRIDE accessions numbers of 3321-3354 (Baerenfaller et al. (2008)), 8743-8750 (Grobei et al. (2009)), 9164-9176 (Reiland et al. (2009)), 9886-9893 (Piques et al. (2009)), and 10068. We also used sets published at proteomics.ucsd.edu (Castellana et al. (2008)) for proteome data. We mapped the peptide sequences to the annotated protein sequences with BLAST (Altschul et al. (1997)), and then the results were translated to the location on the genome sequence. We also searched responsible loci for the peptide sequences with BLAST in “blastn” mode if a peptide sequence was not mapped to the annotated proteins. After we described all -omics data by the relationship with genomic locations, we counted RNA or peptide tags located within transcript regions predicted by ARTADE2. In the case that a single RNA or peptide was mapped to multiple (≥n) locations, we counted 1/n RNA or peptide tags for each locus. The results of counting RNA or peptide tags for each ARTADE2 transcripts are described in supplemental Table S3.

### RT-PCR ASSAYS FOR DETECTING NOVEL GENE CANDIDATES

We performed RT-PCR assays to confirm the existence of novel gene candidates. We described the IDs of tested gene candidates, sequences for gene-specific primers, and brief results of the experiments (Fig. 6 in main paper and Table S5). These PCR primers were designed using the Primer 3 program (Rozen and Skaletsky (2000)). We used RNA samples from “Control” and “Dry 2h” conditions that were prepared in the same way as the RNA samples used for tiling arrays. *Arabidopsis thaliana* (ecotype Columbia) seeds were sterilized and stored for 3 days at 4°C. Plants were grown in plastic dishes on MS base medium under long-day conditions (16-hours light/8-hours dark) for 2 weeks at 22°C. For drought stress treatments, plants were removed from the medium and left for 2-hours on plastic dishes at 22°C.

Total RNAs from drought stress-treated (Dry 2h) and untreated (Control) whole plants were isolated and subjected to Deoxyribonuclease I (Invitrogen) treatments to remove genomic DNA. For RT-PCR using gene-specific primers, 0.5µg of total RNAs were used to generate the first-strand cDNAs with reverse transcriptase (SuperScript III Reverse Transcriptase, Invitrogen), and Ribonuclease H (Takara) treatments were then performed to remove template RNAs. After PCR (Ex Taq, Takara) with reverse transcripts as DNA templates, agarose gel electrophoresis was performed to separate the PCR products. After the PCR assays, electrophoresis and sequencing of the RT-PCR products were performed. Furthermore, the mapping positions of the products were analyzed. We tested the expression of the locus, chromosome 2, position 7565124-7565520, as a negative control, which contained an intergenic region between AT2G17390 and AT2G17410. We confirmed that there were no RT-PCR products from both strands of its locus.

### CDNA SEQUENCES OF POSITIVE RESULTS

**RT-PCR ASSAYS FOR DETECTING NOVEL GENE CANDIDATES**

We performed RT-PCR assays to confirm the existence of novel gene candidates. We described the IDs of tested gene candidates, sequences for gene-specific primers, and brief results of the experiments (Fig. 6 in main paper and Table S5). These PCR primers were designed using the Primer 3 program (Rozen and Skaletsky (2000)). We used RNA samples from “Control” and “Dry 2h” conditions that were prepared in the same way as the RNA samples used for tiling arrays. *Arabidopsis thaliana* (ecotype Columbia) seeds were sterilized and stored for 3 days at 4°C. Plants were grown in plastic dishes on MS base medium under long-day conditions (16-hours light/8-hours dark) for 2 weeks at 22°C. For drought stress treatments, plants were removed from the medium and left for 2-hours on plastic dishes at 22°C.

Total RNAs from drought stress-treated (Dry 2h) and untreated (Control) whole plants were isolated and subjected to Deoxyribonuclease I (Invitrogen) treatments to remove genomic DNA. For RT-PCR using gene-specific primers, 0.5µg of total RNAs were used to generate the first-strand cDNAs with reverse transcriptase (SuperScript III Reverse Transcriptase, Invitrogen), and Ribonuclease H (Takara) treatments were then performed to remove template RNAs. After PCR (Ex Taq, Takara) with reverse transcripts as DNA templates, agarose gel electrophoresis was performed to separate the PCR products. After the PCR assays, electrophoresis and sequencing of the RT-PCR products were performed. Furthermore, the mapping positions of the products were analyzed. We tested the expression of the locus, chromosome 2, position 7565124-7565520, as a negative control, which contained an intergenic region between AT2G17390 and AT2G17410. We confirmed that there were no RT-PCR products from both strands of its locus.

### CDNA SEQUENCES OF POSITIVE RESULTS

**RT-PCR ASSAYS FOR DETECTING NOVEL GENE CANDIDATES**

We performed RT-PCR assays to confirm the existence of novel gene candidates. We described the IDs of tested gene candidates, sequences for gene-specific primers, and brief results of the experiments (Fig. 6 in main paper and Table S5). These PCR primers were designed using the Primer 3 program (Rozen and Skaletsky (2000)). We used RNA samples from “Control” and “Dry 2h” conditions that were prepared in the same way as the RNA samples used for tiling arrays. *Arabidopsis thaliana* (ecotype Columbia) seeds were sterilized and stored for 3 days at 4°C. Plants were grown in plastic dishes on MS base medium under long-day conditions (16-hours light/8-hours dark) for 2 weeks at 22°C. For drought stress treatments, plants were removed from the medium and left for 2-hours on plastic dishes at 22°C.

Total RNAs from drought stress-treated (Dry 2h) and untreated (Control) whole plants were isolated and subjected to Deoxyribonuclease I (Invitrogen) treatments to remove genomic DNA. For RT-PCR using gene-specific primers, 0.5µg of total RNAs were used to generate the first-strand cDNAs with reverse transcriptase (SuperScript III Reverse Transcriptase, Invitrogen), and Ribonuclease H (Takara) treatments were then performed to remove template RNAs. After PCR (Ex Taq, Takara) with reverse transcripts as DNA templates, agarose gel electrophoresis was performed to separate the PCR products. After the PCR assays, electrophoresis and sequencing of the RT-PCR products were performed. Furthermore, the mapping positions of the products were analyzed. We tested the expression of the locus, chromosome 2, position 7565124-7565520, as a negative control, which contained an intergenic region between AT2G17390 and AT2G17410. We confirmed that there were no RT-PCR products from both strands of its locus.

### CDNA SEQUENCES OF POSITIVE RESULTS
FIGURE LEGENDS

Figure S1
Calculation procedure of Positional Correlation matrix Score (PCS). \(| E \times E |\) represents the number of exon position - exon position pairs. \(| E \times E > \theta |\) indicates the number of exon position - exon position pairs with correlations greater than \(\theta\). Symbol \(\lnot\) is negative operator. The coverage region used in the calculation includes intergenic regions within 100 base distance from 5' and 3' ends. Pairs on the same probes are not counted. Coefficient \(e\) is previously learned by training data. (Table S6).

Figure S2
Gene state model of ARTADE1 and ARTADE2. We assumed that sequences of genome were translated under the assumption of Markov chains.

Figure S3
A graph for prediction performance of transcription start (5 prime) and termination (3 prime) sites (TSS/TTS) estimations. The gap is calculated as distance on the genomic positions from TSS/TTS point of reference gene model to the point of the predicted gene model. The distance is calculated according to transcriptional direction. We compared the TSS/TTS on 14,239 reference gene models which had overlaps with predicted gene models within both of ARTADE1 and ARTADE2 results.

Figure S4
A box plot of relative importance (RI) of each score (equation (18)). Scores are calculated as differences between predicted and null models. The score of null model calculated under assumption that the predicted region has no transcript structure (All states in the region are estimated to 0 or 25 in Figure S1). Sum of relative importance of scores for each models are standardized to 1. Correlation Matrix Score (CMS) occupies entire of RI in almost case. Therefore, ARTADE2 predicts transcript structure to fit positional correlations. RI of MTS tends to high if the number of probes existing in the predicted region is few.

Figure S5
Histogram of maximal expression values for 33,239 TAIR9 representative gene models. At the first, we calculate gene expression values for each gene for each condition, which are defined with median values of tiling array probes located with in the exon regions of the gene. Then, maximal expression value is determined as a maximum value among all conditions. Value 0 (= log(1)) means that the gene has no probes in the exon region. A histogram of TAIR9 genes contains two distributions. One seems to be a distribution of not expressed genes (left peak) and another is a distribution of genes which is expressed at least one condition (right peak). Based on this distribution, we set a threshold; maximal expression values \(e^7\) for defining expressed genes. A set of these expressed genes are used on assessing performance of transcript reconstruction with ARTADE2 and other methods.

Figure S6
Precision and recall plots in comparison between predicted gene models and all TAIR9 gene models including genes which may not be expressed under any conditions. ARTADE2 had best precisions in all methods. However, AUGUSTUS had a high recall rate in comparison with whole references, because ARTADE2 has no predicting power for not expressed genes. See Fig. 5 in main paper for the precision and recall plot with expressed TAIR9 genes.

Figure S7
Precision and Recall curves of exons for NGS-ARTADE2 and Cufflinks models which overlap with highly expressed (over \(e^6\) of Fig. 7 in main paper) gene models of TAIR9 gene models with the current mRNA-seq data set. The curves are transited according to PCS decreasing for NGS-ARTADE2 and decreasing of tag-coverage over gene models for (Cufflinks). The precision and recall is calculated in single nucleotide resolution. We allowed that a single reference gene model is covered by multiple predicted gene models. With this rule, the two curves are almost the same. However ARTADE2 showed better performance than Cufflinks for reconstructing full-length transcripts, shown with Fig. 7 and Fig. 8 in main paper.

Figure S8
An example that single gene model in an initial ARTADE2 prediction is split into two transcripts by factor analysis. We found several genomic regions which generated transcripts with highly co-expression and close genomic locations. In such situations, ARTADE2 may wrongly merge these transcripts into one model. To solve this problem, all predicted transcripts are tested and split with factor analysis. If factors are considerably different on left and right sides separated by a certain point, the model is split and ARTADE2 is performed again to re-predict the gene model in each separated region.
Figure S9
An example for detecting alternatively spliced regions from an ARTADE2 model; OMAT1P009860 were not annotated in the reference. The cluster created in second factor corresponds to differences between known gene structures and predicted one. Score plotting shows that both factors are expressed in most organs and conditions including flower and stem. However in detail, we can find that expression of the second factor is low in dry-seeds and imbibed-seeds.

Figure S10
The black curve shows the transition of fraction of factor analysis result regions having overlaps with known alternative splicing or alternative TSS/TTS according to decreasing of discreteness. The red curve shows cumulative frequency distribution for the discreteness values.
# TABLES

## Table S1. Data specifications.

| Name                  | # of experiments | Tiling arrays | mRNA-Seq |
|-----------------------|------------------|---------------|----------|
| Control               | 4                | 4             |          |
| ABA 10h               | 3                | -             | -        |
| ABA 2h                | 3                | -             | -        |
| Cold 10h              | 3                | -             | -        |
| Cold 2h               | 3                | -             | -        |
| Dry 10h               | 3                | 2             | -        |
| Dry 2h                | 3                | 2             | -        |
| NaCl 10h              | 3                | -             | -        |
| NaCl 2h               | 3                | -             | -        |
| Dry seed              | 3                | 2             | -        |
| Flower                | 3                | 2             | -        |
| Imbibed seed          | 3                | -             | -        |
| Leaf                  | 3                | 2             | -        |
| Root                  | 3                | 2             | -        |
| Silique early         | 3                | -             | -        |
| Silique middle        | 3                | -             | -        |
| Silique late          | 3                | -             | -        |
| Stem                  | 3                | -             | -        |

Total 55 16

GEO accession numbers: GSE9646, GSE15700, GSE26074. Details of RNA sample preparation were described previously (Matsui et al. (2008); Okamoto et al. (2010)).

## Table S2. Tag counts and normalization for the study of Next Generation Sequencer.

| Sample name | Experiment ID | # of Mapped Reads | Total nucleotides | Multiplier for normalization |
|-------------|---------------|-------------------|-------------------|----------------------------|
| Control     | 1             | 5,154,978         | 257,748,900       | 3.88                       |
| Control     | 2             | 6,189,453         | 309,472,650       | 3.23                       |
| Control     | 3             | 32,799,384        | 1,639,969,200     | 0.61                       |
| Control     | 4             | 34,683,809        | 1,734,190,450     | 0.58                       |
| Dry 10h     | 1             | 11,040,574        | 552,028,700       | 1.81                       |
| Dry 10h     | 2             | 29,890,234        | 1,494,511,700     | 0.67                       |
| Dry 2h      | 1             | 7,212,499         | 360,624,950       | 2.77                       |
| Dry 2h      | 2             | 33,844,091        | 1,692,204,550     | 0.59                       |
| Dry seed    | 1             | 6,821,706         | 341,085,300       | 2.93                       |
| Dry seed    | 2             | 15,952,876        | 797,643,800       | 1.25                       |
| Flower      | 1             | 6,578,896         | 328,944,800       | 3.04                       |
| Flower      | 2             | 29,186,579        | 1,459,328,950     | 0.69                       |
| Leaf        | 1             | 8,033,855         | 401,692,750       | 2.49                       |
| Leaf        | 2             | 28,140,034        | 1,407,001,700     | 0.71                       |
| Root        | 1             | 6,663,504         | 333,175,200       | 3.00                       |
| Root        | 2             | 25,354,391        | 1,267,719,550     | 0.79                       |
Table S3. Prediction table of ARTADE1.2.2.2.

| Method            | Number of match genes | 5' end prediction | 3' end prediction | Structure match rate |
|-------------------|-----------------------|--------------------|--------------------|----------------------|
| Control           | 8,298                 | 84.98%             | 83.26%             | 79.44%               |
| ABA10h            | 8,822                 | 85.14%             | 84.12%             | 80.04%               |
| ABA 2h            | 8,504                 | 85.67%             | 83.60%             | 79.78%               |
| Cold 10h          | 8,277                 | 85.57%             | 83.87%             | 79.72%               |
| Cold 2h           | 9,300                 | 85.17%             | 83.33%             | 79.62%               |
| Dry 10h           | 8,190                 | 85.85%             | 82.37%             | 79.55%               |
| Dry 2h            | 8,914                 | 85.35%             | 83.21%             | 79.62%               |
| Nacl 10h          | 8,463                 | 84.44%             | 82.57%             | 79.27%               |
| Nacl 2h           | 9,002                 | 85.30%             | 82.67%             | 79.53%               |
| Dry seed          | 6,369                 | 85.13%             | 84.02%             | 79.44%               |
| Flower            | 9,975                 | 84.47%             | 84.15%             | 79.58%               |
| Imbibed seed      | 7,948                 | 85.12%             | 84.20%             | 80.10%               |
| Leaf              | 8,820                 | 85.59%             | 83.99%             | 79.68%               |
| Root              | 10,007                | 85.34%             | 84.70%             | 79.88%               |
| Silique early     | 8,955                 | 84.92%             | 84.19%             | 79.56%               |
| Silique middle    | 8,631                 | 84.86%             | 84.32%             | 79.26%               |
| Silique late      | 7,179                 | 84.64%             | 83.45%             | 78.67%               |
| Stem              | 9,596                 | 84.66%             | 84.33%             | 79.63%               |
Table S4. Verification table.

| Support combination                  | Known | Novel |
|--------------------------------------|-------|-------|
| RNA-seq, PARE, small RNA, mass, CAGE | 2,789 | 5     |
| RNA-seq, PARE, small RNA, mass       | 351   | 2     |
| RNA-seq, PARE, small RNA, CAGE       | 402   | 52    |
| RNA-seq, PARE, small RNA             | 121   | 29    |
| RNA-seq, PARE, mass, CAGE            | 7,677 | 4     |
| RNA-seq, PARE, mass                  | 1,338 | 10    |
| RNA-seq, PARE, CAGE                  | 1,347 | 83    |
| RNA-seq, PARE                        | 489   | 58    |
| RNA-seq, small RNA, mass, CAGE       | 0     | 1     |
| RNA-seq, small RNA, mass             | 0     | 1     |
| RNA-seq, small RNA, CAGE             | 0     | 2     |
| RNA-seq, small RNA                   | 1     | 3     |
| RNA-seq, mass, CAGE                  | 8     | 0     |
| RNA-seq, mass                        | 12    | 0     |
| RNA-seq, CAGE                        | 7     | 11    |
| RNA-seq                              | 11    | 17    |
| PARE, small RNA, mass, CAGE          | 48    | 3     |
| PARE, small RNA, mass                | 62    | 4     |
| PARE, small RNA, CAGE                | 22    | 22    |
| PARE, small RNA                      | 46    | 54    |
| PARE, mass, CAGE                     | 307   | 16    |
| PARE, mass                           | 334   | 18    |
| PARE, CAGE                           | 189   | 149   |
| PARE                                 | 252   | 348   |
| small RNA, mass, CAGE                | 8     | 0     |
| small RNA, mass                      | 8     | 1     |
| small RNA, CAGE                      | 5     | 12    |
| small RNA                            | 10    | 40    |
| mass, CAGE                           | 47    | 5     |
| mass                                 | 78    | 18    |
| CAGE                                 | 41    | 121   |
| No evidence                           | 92    | 400   |
| **Sum**                              | 16,102| 1,489 |
## Table S5. RT-PCR confirmation table for novel gene candidates.

| ID          | Chromosome | Direction | Position (Exons) | Result   | F primer                  | R primer                  |
|-------------|------------|-----------|------------------|----------|---------------------------|---------------------------|
| OMAT1P004260| 1          | Plus      | 4140408 - 4140632 | Negative | ACTGGATCTGGAAGCTTGTT     | ACAAGTGTTTGACACTATTGG    |
| OMAT1P011320| 1          | Plus      | 1161683 - 1161742 | Positive | CAACTCAGTTGATTTTAAAGAAG | CATTTGAGATTGTAACCTAAAA   |
| OMAT1P012900| 1          | Plus      | 17298015 - 17298701 | Positive | AGTGACTTTTTTACGACCAAAAC | TCAAACCTTTCAAGACAAAGGC   |
| OMAT1P022650| 3          | Minus     | 6244443 - 62444121 | Negative | TGTTAATATGAGGCATATTTCCT | ATCAATATCTCAACCGCTGCAA   |
| OMAT3P106080| 3          | Minus     | 8900356 - 8900627 | Positive | GGCTAATCTACAACTGGTTGCA  | AATCAAGTTGTGAAACCATCG    |
| OMAT3P108090| 3          | Minus     | 15262546 - 15262711 | Positive | GCCACCACACACCTTCAT      | GATCAACACACCTCAAAAAGAGA  |
| OMAT3P110670| 3          | Minus     | 7846953 - 7847399 | Positive | TTCCAACCTAATCCCTGATTTC  | GCCCCAACTTATACACACTCAA   |
| OMAT4P003550| 4          | Plus      | 2507889 - 2506855 | Positive | ATGAGCGGTTGGCATGCTAAGT  | CATTTAATCTGAGTTCTTTCA    |
| OMAT4P101380| 4          | Minus     | 18152411 - 18151821 | Positive | TTTTGAAAGAGCTATGAGAGA   | TTTCACACACTACATTTTGA     |
| OMAT5P004400| 5          | Plus      | 4213180 - 42131660 | Positive | AAATTTCTGCAATACCAGTAC   | AGTGATCAAACAGGAACAAGC    |
| OMAT5P005810| 5          | Plus      | 5448601 - 5449651 | Positive | ATTCAGTTTTTTCAGACCTCAC  | AIGATCCACATAGAGTTCTTG    |
| OMAT5P008250| 5          | Minus     | 8221332 - 8222128 | Positive | CCAGATTCAGAACAAGTGGAGACT | TATCCTGGGAATTTCACTGAAAAGT |
| OMAT5P009330| 5          | Plus      | 9819455 - 9820157 | Positive | CTTGTGCTCGTGTATTCTCAG   | CGAAATCTTCTTTGTTACACCA   |
| OMAT5P108720| 5          | Minus     | 17981089 - 17980575 | Positive | CAAATACATACAGCGGTGCGAC  | TTTGCAAGTAAAAAGCGCCTACA  |
| OMAT5P111020| 5          | Minus     | 14005315          | Positive | GCCCTAATTTAGCAATCAGTTG  | CATAGGACCGAGCTTAA         |

## Table S6. Estimated parameters for ARTADE2.

| $\alpha$ | $\beta$ | $\xi$ | $B$ | $Q$ | $T_e$ |
|----------|---------|------|-----|-----|------|
| Expansion to 3'end | 1.9 | 0.20 | 1.25 | 0.0 | 30 | 1000 |
| Expansion to 5'end | 3.7 | 0.29 | 1.25 | -50.0 | 80 | 1400 |

| $\epsilon$ | $W$ | $\theta_0$ | $\theta_F$ | $\theta_I$ | $\theta_g$ |
|-----------|-----|------------|-----------|----------|----------|
| 3.6 | 500 | 0.22 | 0.6 | 0.85 | 0.7 |
Table S7. Parameters for the factor analysis.

| $\theta_f$ | $M$ | $L$ | $\theta_l$ |
|------------|-----|-----|------------|
| 0.4        | 10  | 1000| 3.0        |
REFERENCES

Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic. Acids. Res.*, **25**, 3389–3402.

Baerenfaller, K. et al. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, **320**, 938–941.

Castellana, N.E. et al. (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci. USA*, **105**, 21034–21038.

Dempster, A.P. et al. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statist. Soc. Ser. B.*, **39**, 1–38.

German, M.A. et al. (2009) Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat. Protoc.*, **4**, 356–362.

Grobei, M.A. et al. (2009) Deterministic protein inference for shotgun proteomics data provides new insights into *Arabidopsis* pollen development and function. *Genome Res.*, **19**, 1786–1800.

Kodzius, R. et al. (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, **3**, 211–222.

Li, R. et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Lister, R. et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, **133**, 523–536.

Matsui, A. et al. (2008) *Arabidopsis* transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol.*, **49**, 1135–1149.

Okamoto, M. et al. (2010) Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of *Arabidopsis* using tiling arrays. *Plant J.*, **62**, 39–51.

Piques, M. et al. (2009) Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in *Arabidopsis*. *Mol. Syst. Biol.*, **5**.

Reiland, S. et al. (2009) Large-scale *Arabidopsis* phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant Physiol.*, **150**, 889–903.

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.

Toyoda, T. and Shinozaki, K. (2005) Tiling array-driven elucidation of transcriptional structures based on maximum-likelihood and Markov models. *Plant J.*, **43**, 611–621.

Trapnell, C. et al. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Velicer, W.F. et al. (2000) Determining the number of components: A review and evaluation of alternative procedures. *Problems and solutions in human assessment*, 41–71.
Figure S1

\[ PCS = \frac{e | E \times E > \theta | + |!(E \times E) \leq \theta |}{e | E \times E | + |!(E \times E) |} \]
Figure S2
Figure S3
Figure S4
Figure S5
Figure S8
Figure S10

Discreteness

Fraction (black) / Cumulation (red)