Supplementary information

Shuji 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, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)\,; x_i \in \{A, T, G, C\}$ has random variables of genome state $\boldsymbol{S} = (S_1, \ldots, S_n)\,; S_i \in \{0, 1, \ldots, 25\}$. We let $\boldsymbol{s} = (s_1, \ldots, s_n)$ be a random number vector of $\boldsymbol{S}$. Variable S_i 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 $\Lambda : S \to \{0, 1\}$ as

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

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

$$K = (d_{ij}) = 2\boldsymbol{y}(\boldsymbol{s})\boldsymbol{y}(\boldsymbol{s})' - \mathbf{1}\mathbf{1}', \tag{2}$$

where $\mathbf{1}$ is $n \times 1$ vector of size n whose all elements are 1.

Suppose that there are m probes $Pb_1, \ldots, Pb_m$ in the n interval. Here we limited the probe whose values of more than three replicates in at least one condition exceed $e^{3.754}$which is the lowest 1% value of exon expression in the training data (RIKEN *Arabidopshis* 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 Pb_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 $\boldsymbol{f}_k = (f_k^1, \ldots, f_k^h)$ be a vector of tags for each experiment at probe k. Pearson's correlation coefficient γ_{kl} between expressions values of Pb_k and Pb_l is given by the following equation:

$$\gamma_{kl} = \frac{\sum_{i=1}^{h} \left(f_k^i - \bar{\boldsymbol{f}}_k\right)\left(f_l^i - \bar{\boldsymbol{f}}_l\right)}{\sqrt{\sum_{i=1}^{h} \left(f_k^i - \bar{\boldsymbol{f}}_k\right)^2}\sqrt{\sum_{i=1}^{h} \left(f_l^i - \bar{\boldsymbol{f}}_l\right)^2}}, \tag{3}$$

where $\bar{\boldsymbol{f}}$ is sample mean of $\boldsymbol{f}$. 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 θ and translated the correlation matrix R to expanded matrix $C_\theta(R) = (c_{ij}), i, j = 1, \cdots, n$ of size $n \times n$, where

$$c_{ij} = \begin{cases} 2I_\theta(\gamma_{kl}) - 1 & \text{if } a_k \leq i \leq b_k, a_l \leq j \leq b_l \text{ and } k \neq l, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$$I_\theta(z) = \begin{cases} 1 & \text{if } z > \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

[*]to whom correspondence should be addressed

Let,

$$n_{K_E} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij} \neq 0 \bigcap d_{ij}=1} \tag{6}$$

$$n_{K_O} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij} \neq 0 \bigcap d_{ij}=-1} \tag{7}$$

$$n_{C_E,K_E} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij}=1 \bigcap d_{ij}=1} \tag{8}$$

$$n_{C_O,K_O} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{c_{ij}=-1 \bigcap d_{ij}=-1}, \tag{9}$$

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

$$PCS(C,K) = \frac{en_{C_E,K_E} + n_{C_O,K_O}}{en_{K_E} + n_{K_O}}, \tag{10}$$

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

$$CMS(\boldsymbol{s}, R, \theta) = n \log PCS(C_\theta(R), K(\boldsymbol{s})) \tag{11}$$

If $\boldsymbol{s}$ 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_j^E and r_k^I 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 Gth Gaussian mixture distribution, and their probabilistic density functions p_E and p_I are described as follows:

$$p_E(r) = \sum_{i=1}^{G} g_i^E \psi_i^E(r), \;\; p_I(r) = \sum_{i=1}^{G} g_i^I \psi_i^I(r), \tag{12}$$

where ψ_i^E and ψ_i^I are probabilistic density functions of a normal distribution with mean μ_{E_i} and variance $\sigma_{E_i}^2$ and μ_{I_i} and $\sigma_{I_i}^2$, respectively. Weights of functions g_i^E and g_i^I satisfy the following:

$$\sum_{i=1}^{G} g_i^E = 1, \;\; \sum_{i=1}^{G} g_i^I = 1. \tag{13}$$

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

$$ES(\boldsymbol{s}) = \sum_{j=1}^{N_E} \log p_E(r_j^E), \;\; IS(\boldsymbol{s}) = \sum_{j=1}^{N_I} \log p_I(r_j^I). \tag{14}$$

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

$$MTS(\boldsymbol{s}, \boldsymbol{x}) = \sum_{i=1}^{n} \log P(S_i = s_i, x_i | S_{i-1} = s_{i-1}, x_{i-1}). \tag{15}$$

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

$$P(\boldsymbol{S} = \boldsymbol{s} | \boldsymbol{x}, R, \theta) = \frac{\exp\left\{\alpha\, MTS(\boldsymbol{s}, \boldsymbol{x}) + \beta\, CMS(\boldsymbol{s}, R, \theta) + \xi IS(\boldsymbol{s}) + ES(\boldsymbol{s})\right\}}{Z(\boldsymbol{S})}, \tag{16}$$

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

$$\begin{aligned} \widehat{\boldsymbol{s}} &= \underset{\boldsymbol{s}}{\operatorname{argmax}}\, P(\boldsymbol{S} = \boldsymbol{s} | \boldsymbol{x}, R, \theta) & (17) \\ &= \underset{\boldsymbol{s}}{\operatorname{argmax}} \left\{\alpha\, MTS(\boldsymbol{s}, \boldsymbol{x}) + \beta\, CMS(\boldsymbol{s}, R, \theta) + \xi IS(\boldsymbol{s}) + ES(\boldsymbol{s})\right\}. & (18) \end{aligned}$$

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 θ^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:

$$\begin{aligned} n_I &= \sum_{k=1}^{w} I_{\theta_I}\left(\frac{1}{w}\sum_{l=1}^{w}(1 - I_\theta(\gamma_{kl}))\right), &(19)\\ n_E &= w - n_I, &(20)\\ a_k &= 1 - I_{\theta_I}\left(\frac{1}{w}\sum_{l=1}^{w}(1 - I_\theta(\gamma_{kl}))\right), &(21)\\ \boldsymbol{a} &= (a_1, \dots, a_k)', &(22)\end{aligned}$$

where θ_I is an occupation threshold whether the probe is not exon (Table S6). If the occupation of exon n_E/w exceeds threshold θ_E (Table S6), we can calculate:

$$\begin{aligned} V_t &= (v_{kl}) = \boldsymbol{a}\boldsymbol{a}', &(23)\\ Y_t &= (y_{kl}) = (I_\theta(\gamma_{kl})), &(24)\\ n_{\mathrm{E}v} &= \sum_{k=1}^{W}\sum_{l=k+1}^{W} 1_{v_{kl}=1 \cap y_{kl}=1}, &(25)\\ n_{\mathrm{Eall}} &= \sum_{k=1}^{W}\sum_{l=k+1}^{W} 1_{v_{kl}=1}, &(26)\\ n_{\mathrm{I}v} &= \sum_{k=1}^{W}\sum_{l=k+1}^{W} 1_{v_{kl}=0 \cap y_{kl}=0}, &(27)\\ n_{\mathrm{Iall}} &= \sum_{k=1}^{W}\sum_{l=k+1}^{W} 1_{v_{kl}=0}. &(28)\end{aligned}$$

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

$$F(V_t, Y_t) = \frac{n_{\mathrm{E}v}}{n_{\mathrm{Eall}}} \times \frac{n_{\mathrm{I}v}}{n_{\mathrm{Iall}}}. \qquad (29)$$

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

$$F(V_t, Y_t) = \frac{n_{\mathrm{E}v}}{n_{\mathrm{Eall}}}. \qquad (30)$$

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

$$\widehat{t} = \underset{t}{\mathrm{argmax}}\, F(V_t, Y_t). \qquad (31)$$

The probe $\widehat{t_s}$ from where the prediction of a transcript structure starts is given by the following equation:

$$\widehat{t_s} = \underset{t_s \in \{\widehat{t}, \dots, \widehat{t}+w-1\}}{\mathrm{argmax}} \max(j - i);\ \left(i, j \mid \prod_{k=i}^{j} I_\theta(\gamma_{kt_s}) = 1;\ t_s - w + 1 \leq i \leq t_s \leq j \leq t_s + w - 1\right). \qquad (32)$$

Consequently, the starting position i^0 is set to

$$i^0 = \left[\frac{a_{\widehat{t_s}} + b_{\widehat{t_s}}}{2}\right], \qquad (33)$$

where $a_{\widehat{t_s}}, b_{\widehat{t_s}}$ are start and end positions of probe $\widehat{t_s}$ respectively and operator [ ] is a floor function.

From position i^0, we expanded region size n by DP and estimate structure $\boldsymbol{s}$. The expansion is carried in two directions. One is a direction to 3' end from i^0 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 $P(\boldsymbol{S} = \boldsymbol{s}|\boldsymbol{x}, R, \theta)$ with $P(\boldsymbol{S} = \boldsymbol{s})$. Firstly, we consider i^0 is exon i.e. $\widehat{s}_0 = 4$ (Figure S2). Then, we calculated

$$\widetilde{s}_2^k = \operatorname*{argmax}_{s \in \{0,\ldots,25\}} P(S_2 = k, S_1 = s, S_0 = \widehat{s}_0). \tag{34}$$

From the obtained $\widetilde{s}_2^k$, we calculated

$$\begin{aligned} \widetilde{s}_3^k &= \operatorname*{argmax}_{s \in \{0,\ldots,25\}} P(S_3 = k, S_2 = s, S_1 = \widetilde{s}_2^s, S_0 = \widehat{s}_0). && (35) \\ \widetilde{s}_4^k &= \operatorname*{argmax}_{s \in \{0,\ldots,25\}} P(S_4 = k, S_3 = s, S_2 = \widehat{s}_2 = \widetilde{s}_3^s, S_1 = \widetilde{s}_2^{\widehat{s}_2}, S_0 = \widehat{s}_0). && (36) \end{aligned}$$

Consequently, i-th values are given by

$$\widetilde{s}_i^k = \operatorname*{argmax}_{s \in \{0,\ldots,25\}} P(S_i = k, S_{i-1} = s, S_{i-2} = \widehat{s}_{i-2} = \widetilde{s}_{i-1}^s, \cdots, S_1 = \widetilde{s}_2^{\widehat{s_2}}, S_0 = \widehat{s}_0). \tag{37}$$

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

$$\begin{aligned} \widehat{s}_i &= 25, \\ \widehat{s}_{i-1} &= \widetilde{s}_i^{\widehat{s}_i} \\ &\vdots \\ \widehat{s}_1 &= \widetilde{s}_2^{\widehat{s}_2} \\ \widehat{s}_0 &= 4. \end{aligned} \tag{38}$$

We can obtain structure $\widehat{\boldsymbol{s}}_i$ 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:

$$\operatorname*{argmax}_{s \in \{0,\ldots,25\}} \widetilde{s}_i^k = 25, \quad \text{for} \quad i = j, j-1, \ldots, j - T_e + 1. \tag{39}$$

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:

$$AS(\boldsymbol{s}) = \alpha MTS(\boldsymbol{s}, \boldsymbol{x}) + \beta CMS(\boldsymbol{s}, R, \theta) + \xi IS(\boldsymbol{s}) + ES(\boldsymbol{s}) + B. \tag{40}$$

Then, an optimal back-scanning start point $\widehat{i}$ is given by

$$\widehat{i} = \operatorname*{argmax}_{0 < i \leq j} \frac{AS(\widehat{\boldsymbol{s}}_i)}{i}. \tag{41}$$

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

A direct calculation of $P(\boldsymbol{S}_i)$ is difficult. However, comparing $P(S_i = k, S_{i-1} = s, \widehat{\boldsymbol{s}}_{i-2})$ and $P(S_i = k, S_{i-1} = s', \widehat{\boldsymbol{s}}_{i-2})$; $\widehat{\boldsymbol{s}}_{i-2} = (\widehat{s}_{i-2}, \widehat{s}_{i-3}, \ldots, \widehat{s}_0)$ 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, θ_E, θ_I and θ_g, 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 θ. First, we set the initial threshold $\theta^0 = 0.22$. Let θ^z be estimated threshold of z-th iteration. After the detection of start window $\widehat{t} - \widehat{t} + W - 1$ and getting V_t and Y_t from θ^z, the threshold is re-estimated as

$$\theta^z = \frac{\mathrm{med}(\gamma_{kl} \mid v_{kl} = 1 \cap y_{kl} = 1) + \mathrm{med}(\gamma_{kl} \mid v_{kl} = 0 \ \cap \ y_{kl} = 0)}{2}. \tag{42}$$

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

$$\theta^{z+1} = \frac{\mathrm{med}\left(c_{ij}^{-1} \mid c_{ij} \neq 0 \ \cap \ \Lambda(s_i^z) \cdot \Lambda(s_j^z) = 1\right) + \mathrm{med}\left(c_{ij}^{-1} \mid c_{ij} \neq 0 \ \cap \ \Lambda(s_i^z) \cdot \Lambda(s_j^z) = 0\right)}{2}, \tag{43}$$

where c_{ij}^{-1} is positional correlation γ_{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

$$PCS(C_{\theta^{z+1}}(R), K(\boldsymbol{s}^{z+1})) < PCS(C_{\theta^z}(R), K(\boldsymbol{s}^z)). \tag{44}$$

Then, we began to predict a new transcript structure in the genome, excluding the already estimated region. Remained 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 $\boldsymbol{s}$ and $\boldsymbol{v}_k = (v_k^1, \ldots, v_k^m)$ be expression value at k-th experiment. We assumed that expreesion value $\boldsymbol{v}_k$ are modeled as

$$\boldsymbol{v}_k = A\boldsymbol{f}_k + \boldsymbol{u}_k, \tag{45}$$

where $m \times q$ matrix $A = (a_{ij})$ is called a factor loading matrix. Vectors $\boldsymbol{f}_k = (f_1, \ldots, f_q)$ and $\boldsymbol{u}_k = (u_1, \ldots, u_m)$ 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 $\boldsymbol{a}_2 = (a_{12}, a_{22}, \ldots, a_{m2})$ construct a specific structure apart from the first factor loadings $\boldsymbol{a}_1$, 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 $\boldsymbol{v}$. Then, matrix A is reestimated by a maximum likelihood estimation. Finally, the obtained matrix A is obliquely rotated by a criterion of the 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 $\boldsymbol{l} = (l_1, \ldots, l_m)$ be a vector of center positions of probes. Define subsets $\omega_1, \omega_2 \subset \{1, \ldots, m\}$ as

$$\omega_1 = \{i \mid a_{i1} > \theta_f\}, \tag{46}$$

$$\omega_2 = \{i \mid a_{i2} > \theta_f\}. \tag{47}$$

Then, we calculated the sample mean and variance of probe positions for ω_1 and ω_2 as

$$m_1 = \frac{1}{\#\omega_1} \sum_{i \in \omega_1} l_i, \tag{48}$$

$$m_2 = \frac{1}{\#\omega_2} \sum_{i \in \omega_2} l_i, \tag{49}$$

$$\sigma_1^2 = \frac{1}{\#\omega_1} \sum_{i \in \omega_1} (l_i - m_1)^2, \tag{50}$$

$$\sigma_2^2 = \frac{1}{\#\omega_2} \sum_{i \in \omega_2} (l_i - m_2)^2, \tag{51}$$

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

$$\#\omega_2 \geq M, \tag{52}$$

$$|m_1 - m_2| > L, \tag{53}$$

$$\frac{\sigma_2/\#\omega_2}{\sigma_1/\#\omega_1} < \theta_l. \tag{54}$$

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:

$$
\begin{aligned}
i_1 &= \underset{i \in \{1,\ldots,m\}}{\operatorname{argmax}} \left(\sum_{j=1}^{i} a_{j1} - \sum_{j=i+1}^{m} a_{j1} \right), && (55)\\
i_2 &= \min\left(i \mid i > i_1, a_{i2} > \theta_f\right), && (56)\\
l_d &= (l_{i_1} + l_{i_2})/2. && (57)
\end{aligned}
$$

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

$$
\begin{aligned}
i_1 &= \underset{i \in \{1,\ldots,m\}}{\operatorname{argmax}} \left(\sum_{j=i+1}^{m} a_{j1} - \sum_{j=1}^{i} a_{j1} \right), && (58)\\
i_2 &= \max\left(i \mid i < i_1, a_{i2} > \theta_f\right), && (59)\\
l_d &= (l_{i_1} + l_{i_2})/2. && (60)
\end{aligned}
$$

Consequently, we restart over the prediction of transcript on 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 $i = 2, \ldots, q$
    $c = 0.$
    $a_{\max} = \min\left(1,\ \max_{k=1,\ldots,m} a_{ki}\right).$
    for $j = 1, \ldots, m$
        $n_j = 0.2 + \{a_{\max} - \min(1,\ a_{ji})\}^2$
        if $\underset{l \in \{1,\ldots,q\}}{\operatorname{argmax}}\ a_{jl} = i \bigcap a_{ji} > 0.45$
            $c = c + 1.\ C^i_c = \{j\}.\ B^i_c = j.\ D^i_c = n_j$
        end if
    end for $j$
    t = c
    while $t > 1$
        $d_{\min} = \infty$
        for $\{(j,k) \mid j < k\ ;\ j,k = 1,\ldots,c\}$
            if $\frac{\sum_{l \in C^i_j} \sum_{s \in C^i_k} \gamma_{ls}}{\#C^i_j \#C^i_k} < 0.4$
                continue
            end if
```

```
      $B_{\mathrm{tmp}} = \frac{D_j^i B_k^i + D_k^i B_j^i}{D_j^i + D_k^i}$
      $d = \sum_{l \in C_j^i} n_l \mid l - B_{\mathrm{tmp}} \mid + \sum_{s \in C_k^i} n_s \mid s - B_{\mathrm{tmp}} \mid$
      if $\mod (\#C_j^i + \#C_k^i,\ 2) = 1$
        $d = \frac{4d}{\left(\#C_j^i + \#C_k^i - 1\right)\left(\#C_j^i + \#C_k^i + 1\right)}$
      else
        $d = \frac{4d}{\left(\#C_j^i + \#C_k^i - 1\right)\left(\#C_j^i + \#C_k^i + 1\right) + 1}$
      end if
      if $d < d_{\min} \bigcap \#C_c^j (d - D_c^j) \leq 1.0 \bigcap \#C_c^k (d - D_c^k) \leq 1.0$
        $d_{\min} = d.\ B_{\mathrm{opt}} = B_{\mathrm{tmp}}.$ set combine pair as $(\widetilde{j}, \widetilde{k})$.
      end if
    end for $(j, k)$
    if $d_{\min} < 1.0$
      $C_{\widetilde{j}}^i = \{l \mid l \in C_{\widetilde{j}}^i \bigcup C_{\widetilde{k}}^i\}.\ B_{\widetilde{j}}^i = B_{\mathrm{opt}}.\ D_{\widetilde{j}}^i = d_{\min}.$
      for $s = \widetilde{k}, \ldots, c - 1$
        $C_s^i = C_{s+1}^i.\ B_s^i = B_{s+1}^i.\ D_s = D_{s+1}^i.$
      end for $s$
      $c = c - 1$
    end if
    t = t-1
  loop
  for $j = 1 \ldots c$
    if $\#C_j^i \geq 3$
      output $C_j^i$
    end if
  end for $j$
end for $i$
```

In the algorithm, operation $\mid l - B_{tmp} \mid$ 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 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 θ_g 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, Kodzius *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” software (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 “tblastn” 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

```
>OMAT1P011320.F
CNNNNNNNNNNNNAGAGTAACGNGGAGTGNGGCNANGNNTTTNTTNNNNTTGNNCTTNTTGCTACNNGATGACAGCNCACTACCTTTACANNATTGGGTCCTCTC
NTCTTCTTGNTGAGCTTATATGTACATTGTCTAGCCTCCACGCTCTGTAAGCTAGTGACCCNNGCNATCCTCNTTGATTTATGNCGTGATTGGGAATCAGAGACTC
CATGGAGAGCCGGTATATACTCGTGTTTTCGACNNATTCCNCAATGACTTGGACTTCGNACTACNACTATGACAGCTCACTACCTTTACAAGATTCGGGGCTCTCA
TCTACCTTTAATGAGCTTATATGTACATTGTCTAGCCTCGACGCTCTGTAAGCTAGTGACCCAAGCAATCCTCAATGATTTATGGCGTGATTGGGAATCAGAGACT
CCATGGAGAGCCGGTATATACTCGTGTTTTCGANNNAATCCACAATGAC

>OMAT1P011320.R
CNNNTNNANNGNGCNCTCCATGGNGNCTCTGATTCCCAATCACGCCATAAATCATTGAGGATTGCTTGGGTCACTAGCTTACAGAGCGTCGAGGCTAGACAATGT
ACATATAAGCTCATTAAAGGTAGATGAGAGCCCCGAATCTTGTAAAGGTAGTGAGCTGTCATAGTTGTAGTACGAAGTCCAAGTGAAAGATACATGGCTCCAATC
CTCGTTTACTCTACTCCACAGAGACTCCATAAAATNNAACTGATTGGAATTCGTTTAAAGCTAAAATGCCGACTCACAACACAAACAAAAAGGAATAAAATAATG
CTTCAGATTATAACCATGAAAAGAAAACAGCCAGAACCATTGTACTTTTGGGTGAACCACTCGGATCAGATTCCAAATCGCTGTTATAGAGAAAGAGATCCCAAA
CTGACCGTTTACTCTACTCCACAGAGACTCCATAAAATCGAAACTGATTGGA

>OMAT1P012900.F
NNNNNNNTTGGTTTCTTCNTGTTTAGTTTTTGTTTTCCTTCCTCTCTGACAGATCTGAACCTTTTTTTTTTCNNCGGTTTCNGATCTAAAATTTCNGATCTAAATCTA
CAATTTCTAAATTTTTCGTCTTTGGATTTTGCTGCTTGTTTTTGTCCCTTTTATGNGCAGCACCAGTTCGGGCCGTTCGAACCTCGGNAGAGGCGAGTTCTTTCACC
GCCTTCAGAAGTAGATTTCTCTCCTTTGAAGATCTCNATCTCCTTTCTTCTTCTTATGACCATCATCATTTGGGTCGCGGGGTGTTNGTTNGTTCACCGCCGTTNGTG
GNAAAGAGAGNCCTGGTTTTGTTGGTTTTGTTTTCCTTAGTGTGAATCTGTTTGTTTTTATGGTGATTTAGACTTTGTAATTTTNATTTTTNATTTTTNNAGTTTGNA
TTCNGAGAAGATTGGTNCANANGCGNGNATTTGNATGTTTGAGTTTATTTCCTCTTTAATTTTTCTTAANTAGANTTTTNGTTTCTAGGAGTCTACATGTACTCGCC
GGCTNATGNNNGGGGANAAGGGTT

>OMAT1P012900.R
NNNNNNNNNNGNAGNNTCCTANAAACAAAATCTAATTAAGAAAAATTAAAGAGGAAATAAACTCAAACATACAAATACACGCATCTGAACCAATCTTCTCGGA
ATCCAAACTCAAAAAATCAAAAATAAAATTACAAAGTCTAAATCACCATAAAAACAAACAGATTCACACTAAGGAAAACAAAACCAACAAAACCAGGACTCTCT
```

```
TTACCACAAACGGCGGTGAACAAACAAACACCCCGCGACCCAAATGATGATGGTCATAAGAAGAAGAAAGGAGATCGAGATCTTCAAAGGAGAGAAATCTACT
TCTGAAGGCGGTGAAAGAACTCGCCTCTGCCGAGGTTCGAACGGCCAGAACTGGTGCTGCACATAAAAGGGACAAAAACAAGCAGCAAAATCCAAAGACGAAA
AATTTAGAAATTGTAGATTTAGATCTGAAATTTTAGATCTGAAACCGTTGGAAAAAAAAANGGTTCNNATCTGTCAGAGAGGAAGGAAAACAAAANCTAAACAT
GAAGAAACCAAGAAATTTCGGTTCTGNNNAAAAAAAAAAGCC

>OMAT3P106080.F
NNNNNTNNNNNCTTNNNNNGNCTTATAAGNTTCACATTATACATTCTTCATCATTGCCGGGGTTNNNGAGGCTTCANTGGCAAAAAAAAAACCANNNNGNNGAG
ATCGNNTTANAACTGCTTNNATCTGCAGNCGGACGCGTNNGNCGNTGGTNNANNNANNNNNNNGTCTGGGGGTGGNATTCNGGTCGNGGATANNGTCGAAGTT
GAATTGATAAACGATGATGCTATCAAACATGATTANTACTCTCTCGGANGTTTTTTTTTTTTCTTTCTTTCTTGGGGGGGGTGTGTCGAGACGCCTNNNTGNAAAA
AAAAAAACACACCCGCGNAGTGTCTTACACAAATATATATAAANGGGCCTCCACTATGCGNATCTNTGTATCGANTTGGCGTGTTCTTGACNNTGACACTCCCAA
GCGTAGATCCAGANCNNGAGTGNNATCTCGATATTGCTATCGACCNAGATATNNCCGGAGNAANNGGCCGAGACCGGGGCGGCGGTGCAA

>OMAT3P106080.R
CGNNNNNNNNTTTTTTTTTNNNNNNNNNGTTTTTTTNANACCCTAGATGATGAGNATNGGTGNNGCNTGNCTTTTTGTTGTNNNAAATTTGGGTTTTTTTTTCTCCCT
TGTNNGNTTTTNNNAACNCTCTTATGTCATTTTNNNGCCCCTGTGNTAGTNNCNTTNTNACCTCTTCCTCNCATGCTGTTATNACTCANTCATCCCTTGTAAGCTCC
ATTGATGCACCCATGTTAGTATTACGCAANNCNAAAGAAAAATTTTTTTTTCCCGCGNTTGAAAAAAAGTGTGTTTTTTTTTTTTTNANNCACCCAAAAAGTGTGA
TATATTTGCCCCCCCGTTTTTATANNACAAAAANCTTCCACTCTGNTTATCCAGGNTTNNGNANGGGAATGGCTCAGCTGTGGCCCTATNCATTCNCGTGGNTAGA
TATGCAAGCTACTTGTCAAGCAATGNACCCCTGTNNNATAANGCA

>OMAT3P108090.F
NNNNNNNNGNANCCCCNNTATCAAAACGATGTCGTTTCCATTACTTAAGAAAGATTGAAAATTCAGAAACCAGACTTCGTCCATAGATTATTTACAGAGATCAAA
AAGATTTAAGCGTGATCAAGATTGTCTAGATGATTTATTCAAATGAAAATAAGAGAGAAAAAAGAGAGAGATTATTAGGGTTTCCAGGAAATGTTTGTATAGAG
GATGATGATAAAATCGGAGCAACATCTTCTGCTCTTAATGTTTTTCTTCATTTGTGTCTTTTTTTCGATTAATTTTATTTTTATTTTTGTGTTCTATTTGAGTTCCTAA
TTTCGTAGTGAAAACTCGACCACATTTTTCTCTTTTGNNNNNNNGTTTGAT

>OMAT3P108090.R
GNNNNTCGANTTTTCACTACGAAATTAGGAACTCAAATAGAACACAAAAATAAAAATAAAATTAATCGAAAAAAAGACACAAATGAAGAAAAACATTAAGAGC
AGAAGATGTTGCTCCGATTTTATCATCATCCTCTATACAAACATTTCCTGGAAACCCTAATAATCTCTCTCTTTTTTCTCTCTTATTTTCATTTGAATAAATCATCTA
GACAATCTTGATCACGCTTAAATCTTTTTGATCTCTGTAAATAATCTATGGACGAAGTCTGGTTTCTGAATTTTCAATCTTTCTTAAGTAATGGAAACGACATCGTT
TTTGATATTTGGGGGTTTACCTCTGAAAGAATCAATAGAAGGNGGGTGGTGGG

>OMAT3P109670.F
NNNNTCTCNNTGTTGTTAGCTCTTCTTCTTTCTACACCTATAACCACAAATGTTTTAGGGATAAGCTCTTCTTTTTCCCCTCAATCTCTTCTTGGTCTGATTTTTTATT
GTGTGCTCACTAAGCTCTTCCTATCACCAAACTCACGGCTAGATTCACTTATTTTCTTATGGTTGGCAAATTACAAGCTTCATCGGACGAAAAGATTGGTCATCATT
GTCGCCGTCGCGGGTTCGTTGTCACCGTCGGCGTCCCGTCTGTCACCGTCGCCGTGGCTCTTTGTCGCCGTTAACCCTTGTCAAAAACCCTAAAATTTTAATGGGTT
GAGCCTTGTGAATTCGGGTTGGGCCTTGTAAATTTTATAATTGGGTTTGTAAATTTGTTAATGTATTTTGATGGNNNNAAANGTTTGGGCANTTGTCAAAAACCCT
AAAATTTTAATGGGTTGAGCCTTGTGAATTCGGGTTGGGCCTTGTAAATTTTATAATTGGGTTTGTAAATTTGTTAATGTATTTTGATGGTGTATAAAGTTTGG

>OMAT3P109670.R
TTNNNAANTTTACAACCCAANTTATAAAATTTACAAGGCCCAACCCGAATTCACAAGGCTCAACCCATTAAAATTTTAGGGTTTTTGACAAGGGTTAACGGCGAC
AAAGAGCCACGGCGACGGTGACAGACGGGACGCCGACGGTGACAACGAACCCGCGACGGCGACAATGATGACCAATCTTTTCGTCCGATGAAGCTTGTAATTTG
CCAACCATAAGAAAATAAGTGAATCTAGCCGTGAGTTTGGTGATAGGAAGAGCTTAGTGAGCACACAATAAAAAATCAGACCAAGAAGAGATTGAGGGGAAAA
AGAAGAGCTTATCCCTAAAACATTTGTGGTTATAGGTGTAGAAAGAAGAAGAGCTAACAACAATTGATGAAGAAAAAAGAAAATCANNNNTNAAGGTTGGAGA
NNAGAGATTGAGGGGAAAAAAGAAGAGCTTATCCCTAAAACATTTGTGGTTATAGGTGTAGAAAGAAGAAGAGCTAACAACAATTGATGAAGAAAAAAGAAAA
TCAGGGATTAGGTTGGAGA

>OMAT4P003550.F
NGNNNNNNNGGGNTTGATTACTTGCAACTAGACTAGAGTATCGTACCTTAAGAACATATCAAGCTTCATTTACAGCCATTGGATCAGGTGTATTCAATCTTGCGC
ACTCAAACACCAAGACATTCCATATCTCGACCCCAAAGCCTTCAATGCACTCCAACAAAGAGATTCCTTTAAATCAATGAAGAACACGTCCTTTAGGAGCTTCTA
CATGGACCAGAGGCTTCTCTCACATGGAAATATCAAGAAGATTTCGAGATATAAGGAGTCAATCATATTTCCTTATTCGGCCAAGATTCAAGCAATTAAGCCTAA
CGGCTATAATATCTTGTGCATCTTNAATTTTGTGCACAAGATATTATAGCCGTNAGGTATAATTTGGATCTTGGNCGAATAAGAAATNNGNTTGANCCTTAANNNC
TGNAAACTNNTGNAATTTCCNNGGAAAAAANCCCCGGNNNNGNNNNACCCCNNNNNGGGGGTTNTTTTTTANTTNAANAGAANCNNTTTNNGGGGGNNNNNAA
NTTTNGGNNAAAAAAAAAAAANTNNNNNTTGTNNNCCCCAAAAAAAAAANCCCCCCCCNGGGGGGAAAANNNNNNTTTTTTTTNNGNGNAAANCCNNNNNTTTNNN
GGNAAAAANNCNCCCCCCCCNNNGGGGGGGNNAAAAAA

>OMAT4P003550.R
GNNNNNNNGNCGTTAGNNTTNNTGCTTGAATCTTGGCCGAATAAGGAAATATGATTGACTCCTTATATCTCGAAATCTTCTTGATATTTCCATGTGAGAGAAGCCT
CTGGTCCATGTAGAAGCTCCTAAAGGACGTGTTCTTCATTGATTTAAAGGAATCTCTTTGTTGGAGTGCATTGAAGGCTTTGGGGTCGAGATATGGAATGTCTTGG
TGTTTGAGTGCGCAAGATTGAATACACCTGATCCAATGGCTGTAAATGAAGCTTGATATGTTCTTAAGGTACGATACTCTAGTCTAGTTTGCAAGTAATCAAACCC
```

```
GTCAGCTTCCAGATAANNNAAAGTGTAGGAANNGGTTTAANTGCTGAGNTCTTGCGGAAAAGGGGGAAAAAAAGGAAACCNNTTTANCCCNAANCTCTCNGAN
CTCTCNCNTTTTAAAAGGAAACACGCCTCTGGTCCAAAAAAACCNCAAAACACGTTTTCTTAGGGGATATAAAGAATCCCTCTGTGGGGGCTTTAANCTTTTTGG
GAAAAAAAAAATAANCCATTTTTTTCCCCCACACAACAAAACCAAGCNGAANGGTCCCGAAAAAGAGTTTTTTTGNGGGGAGAAAACACCCCTTTTCGAAAGGA
AAAAACAGAGCAAGGGGGGGGGGGGGGGGNAAAAAAA

>OMAT4P101380.F
NNNNNNNNNNNNNTNNNGNNCTGNGNNTTTCTTTTAGCTTTTGGCCAGATTGTTTTTGGTTCAAATTATGCTTTCTTTGTCTCGATTTCCTTATTTTCGGTCTGATT
AGACTCTTTTGCAATTCTATTTGAGTACTGGAATCGTAATTCCTTGAATCCCAGGCGTTTAAAGAGCTCGATTTGTGTTCCAAGTTTATCAATTTCAATTTTTAAGA
ACAAATTGGAGTTAGGGGTTAAGGTTTTTTTTTNGTGACTTTGGATTGAATTTATGCAATTGTGGCTATTTCTTTAACCGATCTATGGAATCGAGTTTGTTTCTCTTC
GTTTATAGGGGTTGCTGAAAGCTCNTCCCAATTTAAATGA

>OMAT4P101380.R
CGNNNNNNNNNNNNNNNNNGNANNNNNACTCGATTCCATANGNATCGGTTAAAGAAATAGCCACAATTGCATAAATTCAATCCAAAGTCACAAAAAAAAAAACCT
TAACCCCTAACTCCAATTTGTTCTTAAAAATTGAAATTGATAAACTTGGAACACAAATCGAGCTCTTTAAACGCCTGGGATTCAAGGAATTACGATTCCAGTACTC
AAATAGAATTGCAAAAGAGTCTAATCAGACCGAAAATAAGGAAATCGAGACAAAGAAAGCATAATTTTGAACCAAAAACAATCTGGCCAAAAGCTAAAAGAAA
CACCAGAGCTATATGGGGAATTTCTAGAACTTAGCNNNNCCCCCCCGCTA

>OMAT4P111870.F
NNNTNNNNTATTTCANGGTGGTGGCATGGAATCTTCAGCTGATGGAATCCGGCGAATTATGTGGAAGAAGCTCTTGGTGTTGATTTGATTAAGCCCGACCGACGG
AGCAGAGGTGGATCGTGGATCAGCCGGCAACGTCAGTTTCAACATCTCTTCATCTCACGTTGGACACGTGTTGGCAACGTGGCACTTTTTTTGTCTTAGCCTTATC
TTTAGAATTGGGTGGGTTAAGCTGTGCTTTAAATTTTAATCAAAGTNNNNAATGGTGAGGG

>OMAT4P111870.R
NTTTANNCNCNGNNTTNNCCACCCAATTCTAAAGATAAGGCTAAGACAAAAAAAGTGCCACGTTGCCAACACGTGTCCAACGTGAGATGAAGAGATGTTGAAAC
TGACGTTGCCGGCTGATCCACGATCCACCTCTGCTCCGTCGGTCGGGCTTAATCAAATCAACACCAAGAGCTTCTTCCACATAATTCGCCGGATTCCATCAGCTGA
AGATTCCATGCCACCACCTTGAAATAAATTCAGAGATCTCTCAATNNNNNCTTCAAAGGAAA

>OMAT5P004400.F
NNNTTTGNNNTCTTTGTATAGTTTTCATTTTTGAAGGTCACAGAAAGCTCATTTTGATCTTTTGCAATGCTGATTCCTTCTACCTTAATATCCATTTCCAGTTTTTAA
ATACTTGACCAGCCTCCACATTCACATGTACAGCAGAGTAATTTCTGCAATCTCATCAGCTAGCATCGCAGGCAATATAATGTTTTTAAATTTCTGTAAATATAAT
TTTCTCTCAGTGTCCTTTCCACGTTGTTNNNNNNNNNGAATC

>OMAT5P004400.R
NNNNNGNNNTGNNAGNAATTATATTTACNGAAATTTAAAAACATTATATTGCCTGCGATGCTAGCTGATGAGATTGCAGAAATTACTCTGCTGTACATGTGAATG
TGGAGGCTGGTCAAGTATTTAAAAACTGGAAATGGATATTAAGGTAGAAGGAATCAGCATTGCAAAAGATCAAAATGAGCTTTCTGTGACCTTTCAAAAATTGA
AAACTATACAAAGATTGTCAAAACTAATAGTGACTGTNNNNTGGNNAGAANTT

>OMAT5P005810.F
GGNGCACCTTGGCCCTTGNCTGAGAANCAGCTTTTTCTCATATTCAGTTTTTCTAAATTTGTTGTTTTAAAAAAAAAAATTGGTTTGTTTTTTTGGCTGTNCTATGA
TGATTATATTGCTTTAACCNCCCTGAAAAGATGGGCTTGAGAATGTGATGNGGTNCTNCTTGATTTAANCCAAAGAAAGAGTTCNTGAGCTATATGTATTGCTTGA
TGACNCTTTGATCTAAAATACTTGAAGGGGATTTGTTTCCCTTTGNGTTTTAGATCAAGGAAGAGATCAGNGTAGAGCNCTTGTNCGANATTTTTCTTAGTTTAAA
TCTTGAAGAACATTATCATTTTCNCAAGCACNCAATGGNGGAATCATAAA

>OMAT5P005810.R
NATGTTCTTCAGATTTAAACTAAGAAAAATATCGAACAAGAGCTCTACACTGATCTCTTCCTTGATCTAAAACACAAAGGGAAACAAATCCCCTTCAAGTATTTTA
GATCAAAGTGTCATCAAGCAATACATATAGCTCAAGAACTCTTTCTTTGGATTAAATCAAGAAGAACCACATCACATTCTCAAGCCCATCTTTTCAGTGTGGTTAA
AGCAATATAATCATCATAGAACAGCCAAAAAAACAAACCAATTTTTTTTTTTAAAACAACAAATTTAGAAAAACTGAATATGAGAAAAAGCTGCTTCTCAGACAA
GGGCCANGGTTGTCCNCATGNGTGAGCTTGAAAAAACGNGAANAA

>OMAT5P008250.F
NNNNNNNTCATCGNNTNNNGNNNTNNNNTCCTTAGCCACGTCCTTTTCCATGTCCATATTCTATTGTTTGATCATCTGTCACAGGTTGAAGTCTCTCTTTTTGTTTT
GAGTGTTTTTATCATGTTCATAAATTCATGGTATCAATCAAGAATCTCACGGTCACATTCTCAAGCAGAGAAGCTAAACATTGGTTGATTAAACCATAGAGTGACG
CTCTTATTTCATTACATGGTAGGGGCTTCTATCTGATGATGTTCATAGTGATTCTTCTTTTAATTTTGTTTGGTTTTGACGAAATACATAATAAGTGTCTCCCAAGAT
GTTTGTGAGTGTTTTTTTTTCTTATCCTTTAGTCCAAATTTTTCATAACTTAAGATCCATCGAATTTTGATCTTGAATATGTATGGTTTTGAGTGAGATGTTTACCAT
CGAATACTCTTTTATTCCTAGATTCATGGATTGTAGGTTTATATATTTACAAGCGTCGACTCTTTTATGTCANGGTTTACATGATATTCTGGAAAACAATTATGTTT
GTACGTTTTCTATTCATTCTCNNNNTTNNCCCAGGTAA

>OMAT5P008250.R
CNNNNNNTTGNTTCNTGANNNGGTTCCNGAATATCATGTAAACCTTGACATAAAAGAGTCGACGCTTGTAAATATATAAACCTACAATCCATGAATCTAGGAATA
```

AAAGAGTATTCGATGGTAAACATCTCACTCAAAACCATACATATTCAAGATCAAAATTCGATGGATCTTAAGTTATGAAAAATTTGGACTAAAGGATAAGAAAA
AAAAACACTCACAAACATCTTGGGAGACACTTATTATGTATTTCGTCAAAACCAAACAAAATTAAAAGAAGAATCACTATGAACATCATCAGATAGAAGCCCCT
ACCATGTAATGAAATAAGAGCGTCACTCTATGGTTTAATCAACCAATGTTTAGCTTCTCTGCTTGAGAATGTGACCGTGAGATTCTTGATTGATACCATGAATTTA
TGAACATGATAAAAACACTCAAAACAAAAAGAGAGACTTCAACCTGTGACAGATGATCAAACAATAGAATATGGACATGGAAAAGGACGTGGCTAAGGACGAT
ATAGCCTGAGAGATGACGATTAGAGTTCCACTGNNCNNNAAANCTGGA

>OMAT5P009330.F
AGNNNNNNNCTTCNNNNNNNNGGNNNTGACTGGTTTTTATCTTTTTTCCAGTCATCGGATTATTATCGATTGCTTTACTTCTTTTCCATGTTTTCTTTCAATAATTTCT
AAAAATCTGAGTTGGTAATTAGTTTTAAGTGCTAGCTATTTAAAATTAATGTCTCTGGCTTTTGGTTTTAGAAACCTCGATTTGCAAGAAATCGTTCAAACNATTA
CATCTGGGTTTGAAAGATTGAAGGATAATCCGATGCTTATTAGGATTCTCTTAAGCCGTCTATTTCTTTGGGAGCTCCATCCATATCTATTCTGGTGGTAANNAAG
GATTATTC

>OMAT5P009330.R
GNNNNNNNNNNNGNAGCTCCCAAAGNNNTAGACGGCTTAAGAGAATCCTAATAAGCATCGGATTATCCTTCANTCTTTNNAACCCAGATGTAATCGTTTGAACG
ATTTCTTGCAAATCGAGGTTTCTAAAACCAAAAGCCAGAGACATTAATTTTAAATAGCTAGCACTTAAAACTAATTACCAACTCAGATTTTTAGAAATTATTGAAA
GAAAACATGGAAAAGAAGTAAAGCAATCGATAATAATCCGATTGACTGGAAAAAAGATAAAAACCAGTCAACACCACTTATGGAAGCTGGATTTGCTCGACATA
AACCGGTTATCTCCACTATTGAAGCTGAAANNNNNNGGGCAAGAGANNNNNNNNNNNNNGCNNNNNNNNNNNNCNNNNANAGTCNGCTAAGGAATAANNNNNN
NNNA

>OMAT5P108720.F
CAACGTTTCGGAAGAGAGGCGTACAAGGAACCATGTGCGTATTCACATGAAGCGGCCTTTCCTTCAGCCGGATCAGAATCACACGTTGACGCAACTCAATTCATA
ACTCGTAAGCAGTCTATAGAAGATATATTCCCTTCCGTTTGGTCCCCTTCGACCAACTCCATCAGTGCTTCTCAATTTTAGCATTCTTCTTCTTCTTTTGTGGGTTGT
CTCTGTTTTCGGTTAAATGTTCCGATTTTTTTGTTATGATATGGTTTNAACTTTGCAAAA

>OMAT5P108720.R
TCGGACATTTAACCGAAAACAGAGACAACCCACAAAAGAAGAAGAAGAATGCTAAAATTGAGAAGCACTGATGGAGTTGGTCGAAGGGGACCAAACGGAAGGG
AATATATCTTCTATAGACTGCTTACGAGTTATGAATTGAGTTGCGTCAACGTGTGATTCTGATCCGGCTGAAGGAAAGGCCGCTTCATGTGAATACGCACATGGTT
CCTTGTACGCCTCTCTTTCCGAAACGTTGGCTGCAACTTCGATGGAATTGTATTTGA

>OMAT5P111020.F
GGATGATCGTCGAACTCANGATTTTTGGCTACTTTCGGCCTTGAGAANAGAAGGGNTACCGCCCAGCNGNATTGGGTTGCTTTTTCAGGATNCGATTTCGTCGATG
TGCCTAGGTCGGAGACAACTANCTTTCCGGGAGATCTGTTGTGGGTTGGNGCATCGNANGNATGNNNTTNATGATNAAAACCGGTCTCTTTCGCTGAAGATTCCC
GATCTTCGTTGGTGACCACGAAACGGTAGCTCTTGAGCCCGCCGGATTTAAGNTCCCGGTCCTCTGANANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNGA

>OMAT5P111020.R
GGGCTAGAGCTCCGTTTCGTGGTCNCCANGCGAANCATNGGGAATCTCCTNNGNANGNGACCGGTTTTCATCTTNNAANCCATCCCTCCGATGCACCAACCCNCA
NNAGTATCTCCCGGANAGCTAGTTGCTCNGACTAGGCNCATCGACGAAATCGGATCCTGAAAAAGCAACCCANTCCACTGGGCGGNAACCCTTCTCTTNCAAACC
GAAAGTAGCCAAAAAACCTGAGTTTCGACGATTCAATCCGGCTCGATGGATGNTAAATTGAGGCANCGTTCTNCTCTTAGNNNATCTNANATNCANTTCTGGTNT
CGTNCNGGTCGGCAAATGATACGGTATANGCTNANNTTCCTNGGGGAAAGGAGAGACGNCTCCGGNTCTNNGGCTCCGNCTCTNTGTCCTGCTGANANTCAAGG
CGANATNTNTNGCTNCNNCCNTCNNNGGANNGTTCGCTNGATCAGCAGCCNTCCNTCCAGNGNNGCTGNGNTCNCNTACTGATNNNTNNTNNCNTGCNNAGANN
TTCNNCGGATCNNTT

FIGURE LEGENDS

Figure S1

Calculation procedure of Positional Correlation matrix Score (PCS). $\mid E \times E \mid$ represents the number of exon position - exon position pairs. $\mid E \times E > \theta \mid$ indicates the number of exon position - exon position pairs with correlations greater than θ. Symbol ! 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
Stress treatments	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	-
Organ conditions	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. Detailed 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
	ACTGGATTCTGGAGCGTGGT		ACAAGTGGTGTGCACATTGG	
OMAT1P011320	1	Plus	11616183 - 11617412	Positive
	CCAATCAGTTCGATTTTATGGAG		CATTGTGGATTATGTCGAAAACA	
OMAT1P012900	1	Plus	17298015 - 17298701	Positive
	AGTGACTTTTTCAGCACCAGAAC		TCAAACCTTCCAAAGACATAAGC	
OMAT1P022650	1	Plus	28894180 - 28895571	Negative
	TGTGAATACTGAGGGCTATTTTTCT		ATCATATCTCCATCGCTGCAA	
OMAT3P106080	3	Minus	6244443 - 6244353,6244121..6244021	Positive
	GCGTAATACTAACATGGGTGCAT		AATCATGTTTGATAGCATCATCG	
OMAT3P108090	3	Minus	8901356 - 8900627	Positive
	CCCACCACCACCCTTCTAT		GATCAAACACAAACTCAAAAGAGA	
OMAT3P109670	3	Minus	12562546 - 12562294,12562171 - 12561898	Positive
	CTCCAACCTAATCCCTGATTTTC		GCCCAAACTTATTACACCATCAA	
OMAT4P003550	4	Plus	7846953 - 7847399	Positive
	TCCTACACTTGCCTTATCTGGAA		GTGCACAAAATAAAGATGCACAA	
OMAT4P101380	4	Minus	2507889 - 2506855	Positive
	AGTAGCGGTGGGCTAGCTAAGT		CATTTAAATCGATGGAGCTTTCA	
OMAT4P111870	4	Minus	18152411 - 18151821	Positive
	TTCCTTTGAAGACTAGATTGAGAGA		TCCCTCACCATTTACAACTTTGA	
OMAT5P004400	5	Plus	4213180 - 4213660,4214127 - 4214164	Positive
	AATTTCTCCCAAGCTACAGTCAC		AGGATTCAAACAGGAAACAACG	
OMAT5P005810	5	Plus	5448601 - 5449651	Positive
	ATTCACGTTTTTCAAGCTCACTC		ATGATTCCACATATGAGTGCTTG	
OMAT5P008250	5	Plus	8221332 - 8222128	Positive
	CCAGATTCAGAACAAGTGGAACT		TTACCTGGGAATTCATGAGAATG	
OMAT5P009330	5	Plus	9819455 - 9820157	Positive
	CTCTTGCCCGGTATCTTTCAG		CGAATAATCTTTGTTTACCACCA	
OMAT5P108720	5	Minus	14005742 - 14005315	Positive
	CAAATACAATCATACGAAGTTGCAG		TTTGCAAAGTAAAAGCCATATCA	
OMAT5P111020	5	Minus	17981089 - 17980575	Positive
	GCCTCAATTTAGCATCCATCG		CAGAGGACCGGAGCCTTA	

Table S6. Estimated parameters for ARTADE2.

	α	β	ξ	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

e	W	θ_0	θ_E	θ_I	θ_g
3.6	500	0.22	0.6	0.85	0.7

Table S7. Parameters for the factor analysis.

θ_f	M	L	θ_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**.

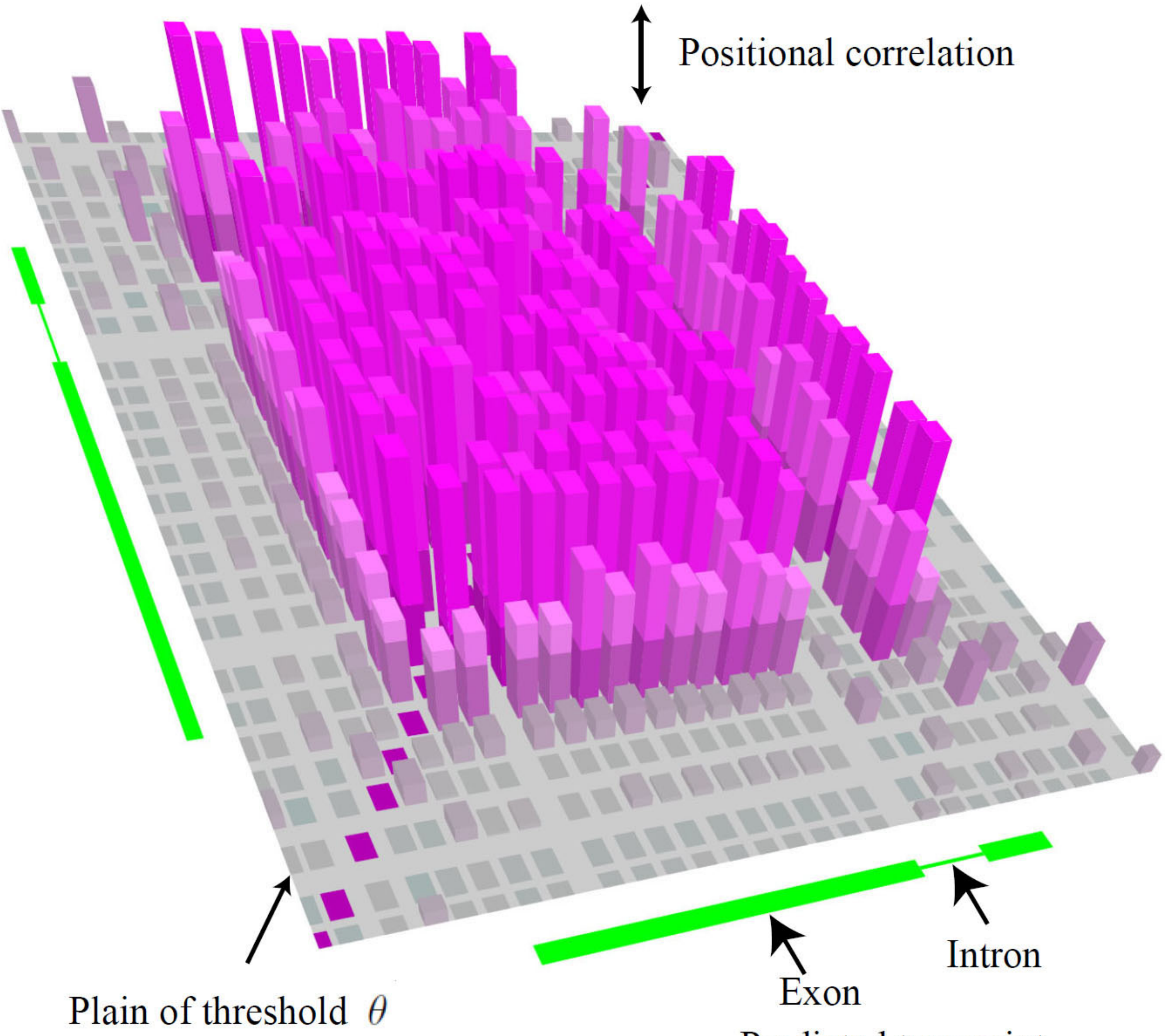
Reiland, S. *et al*. (2009) Large-scale *Arabidopsis* phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant. Physiol.*, **150**, 889–903.

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.



$$PCS = \frac{e \mid E \times E > \theta \mid + \mid !(E \times E) \leq \theta \mid}{e \mid E \times E \mid + \mid !(E \times E) \mid}$$

Figure S1

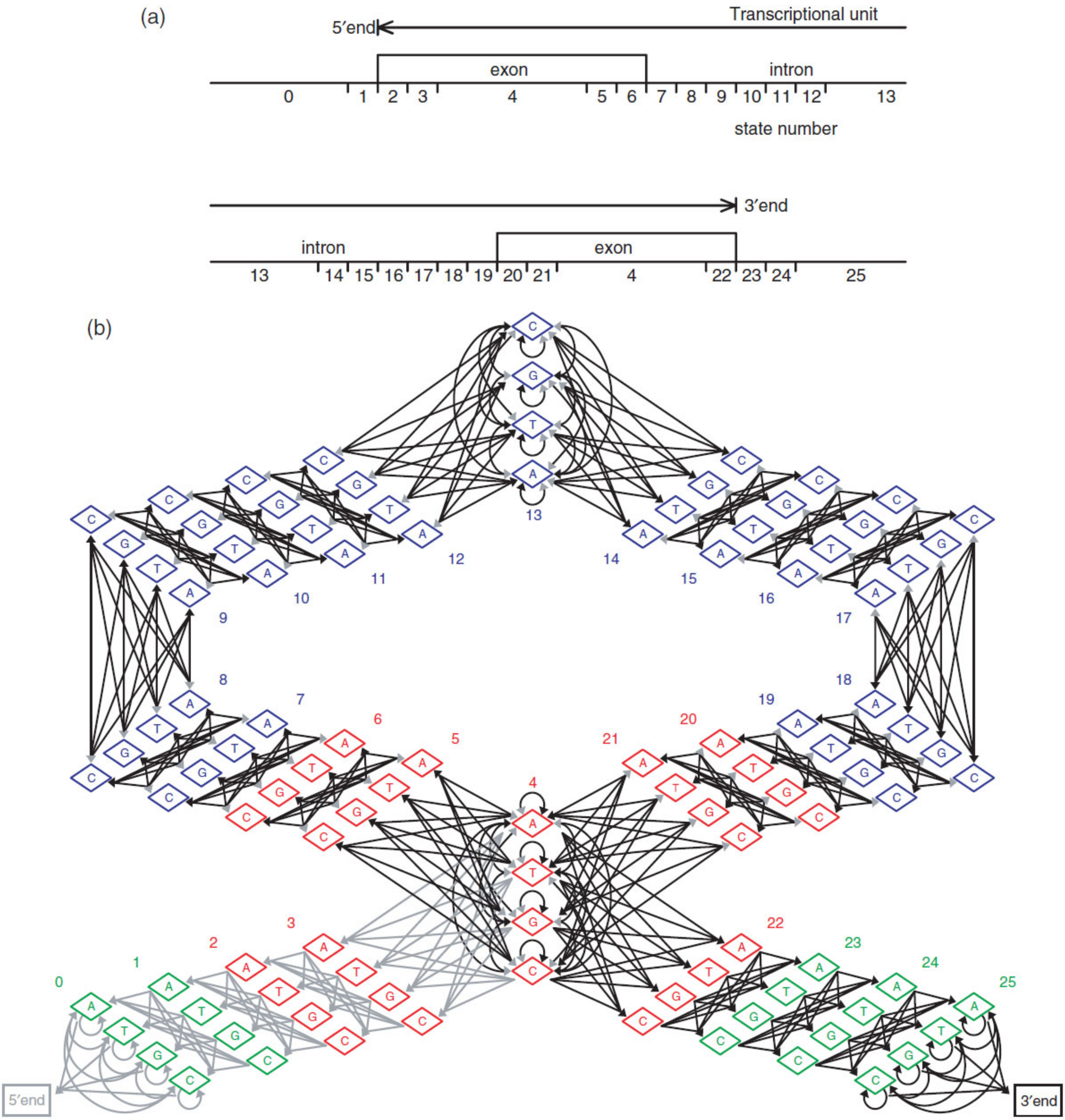
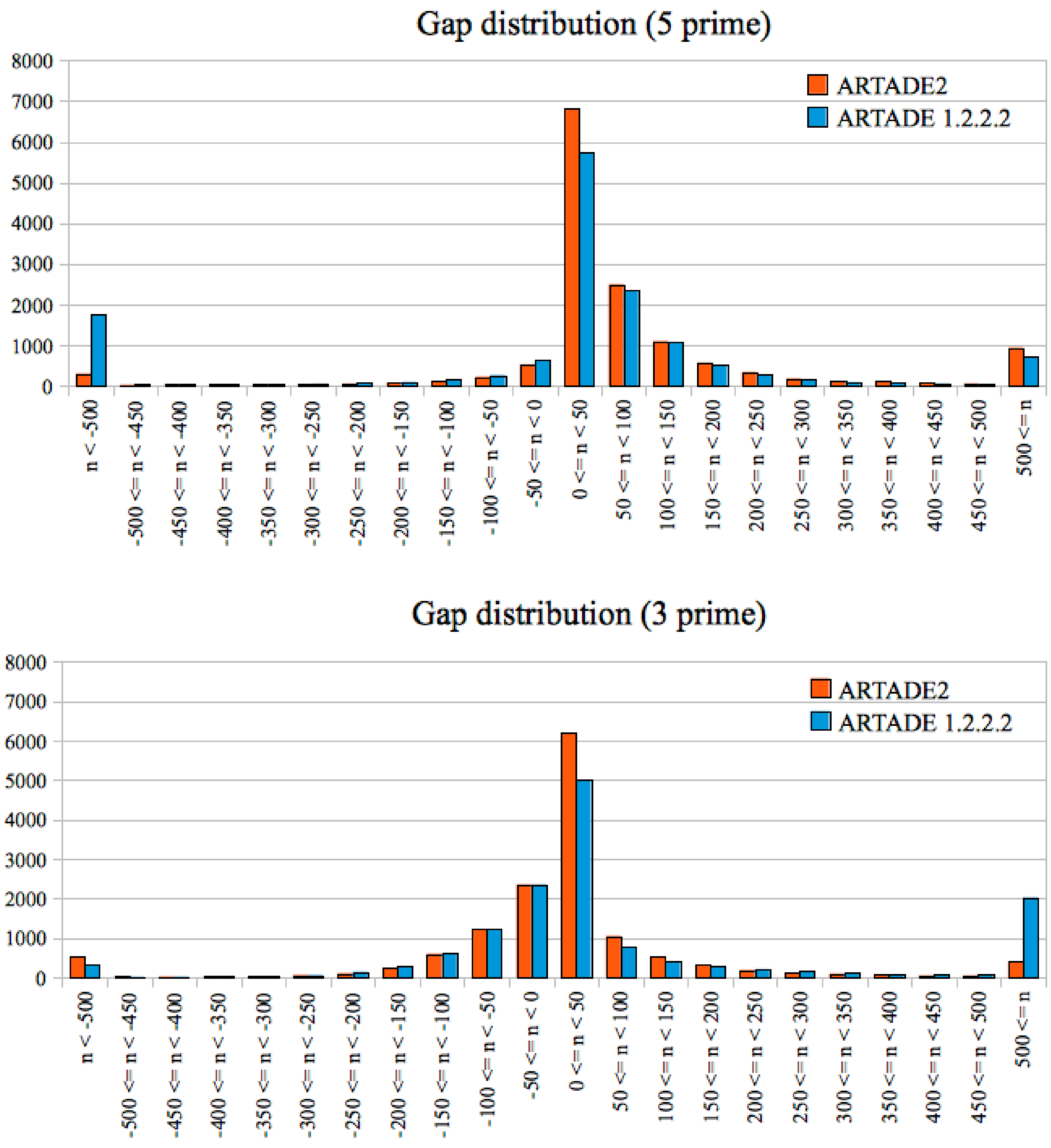

(a)
Transcriptional unit
5′end
exon
intron
0
1
2
3
4
5
6
7
8
9
10
11
12
13
state number
3′end
intron
exon
13
14
15
16
17
18
19
20
21
4
22
23
24
25
(b)
13
12
11
10
9
8
7
14
15
16
17
18
19
6
5
4
21
20
3
2
1
0
22
23
24
25
C
G
T
A
5′end
3′end

Figure S2



Gap distribution (5 prime)

Gap distribution (3 prime)

Figure S3

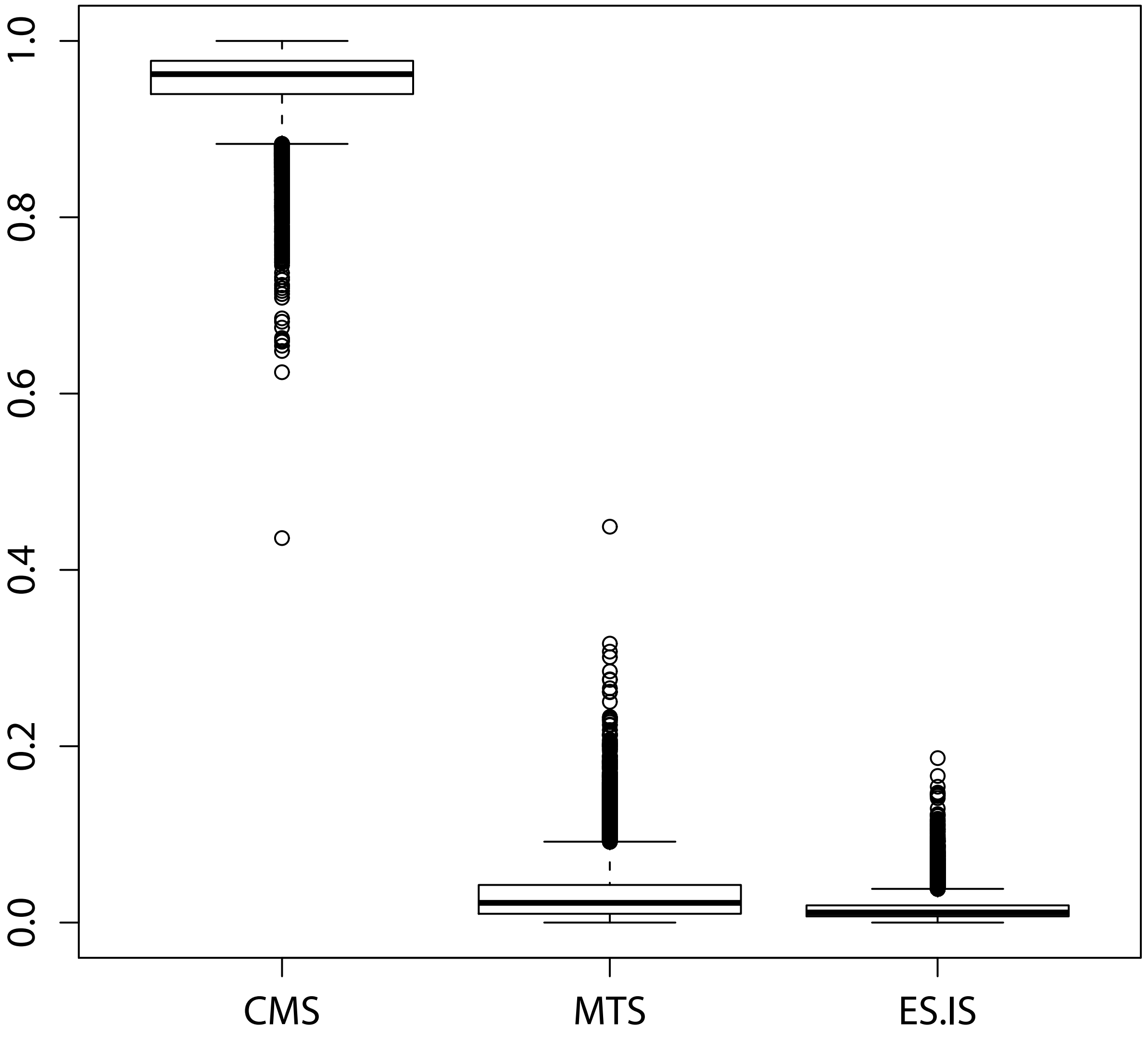


Figure S4

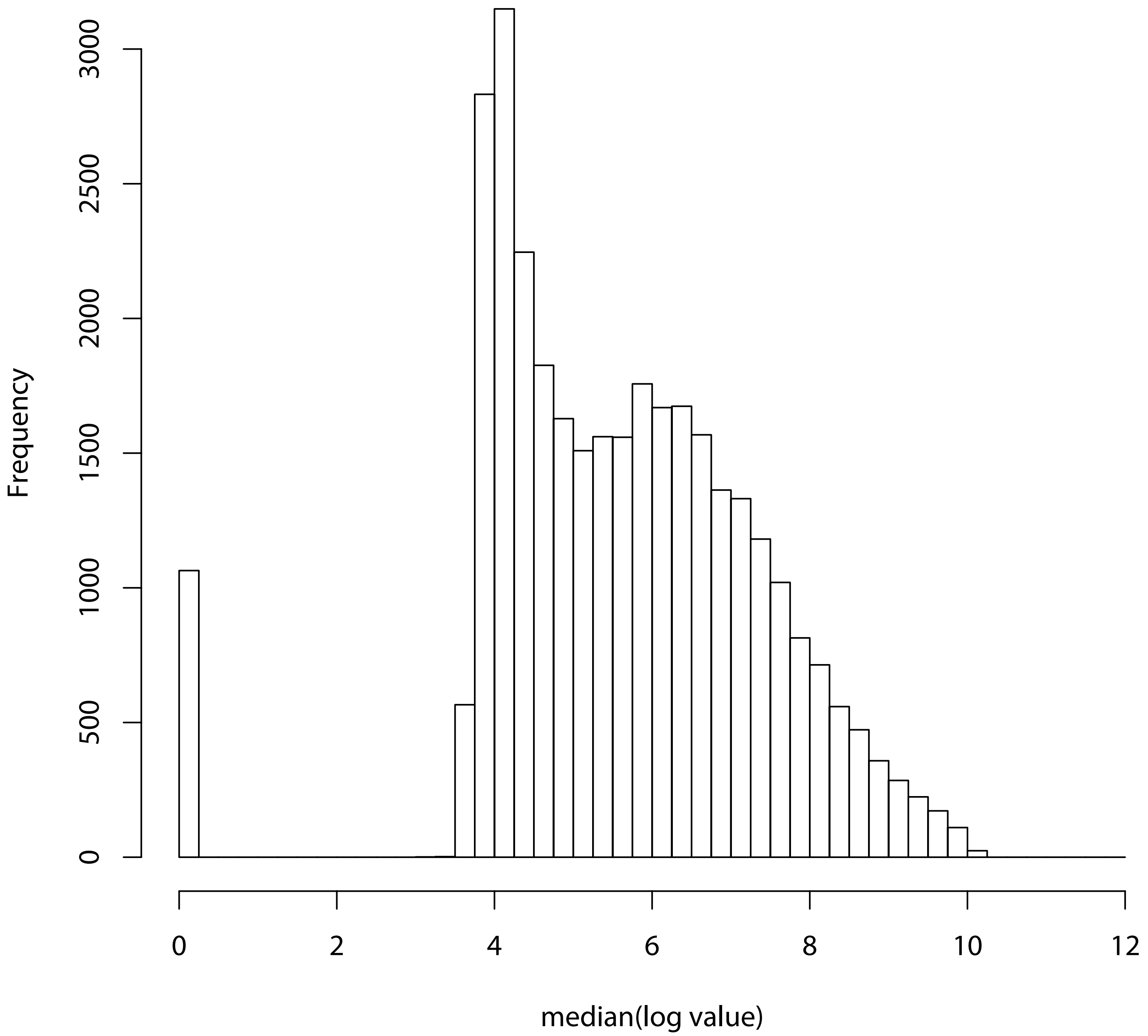

Frequency
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
12
median(log value)

Figure S5

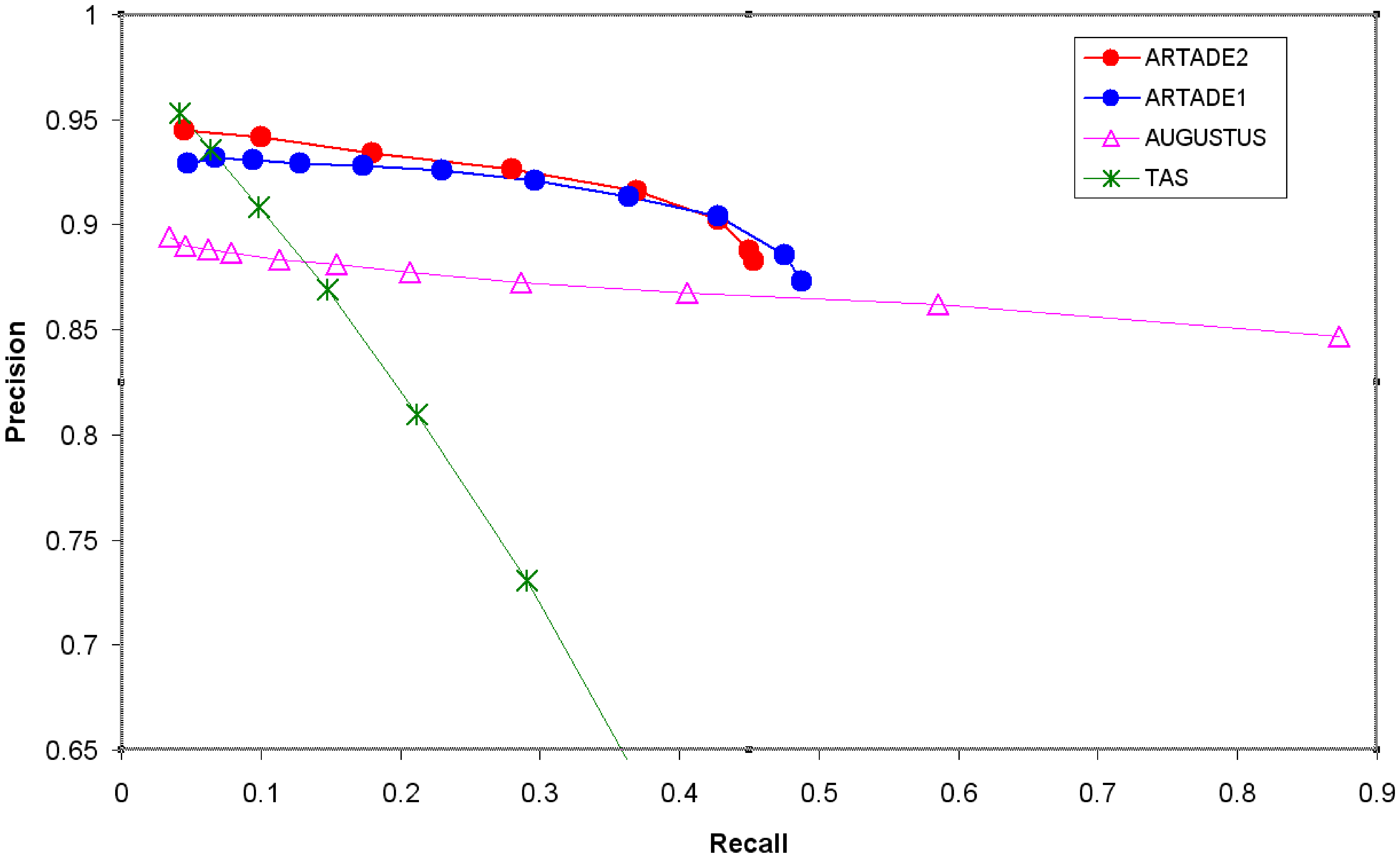

1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
Precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Recall
ARTADE2
ARTADE1
AUGUSTUS
TAS

Figure S6

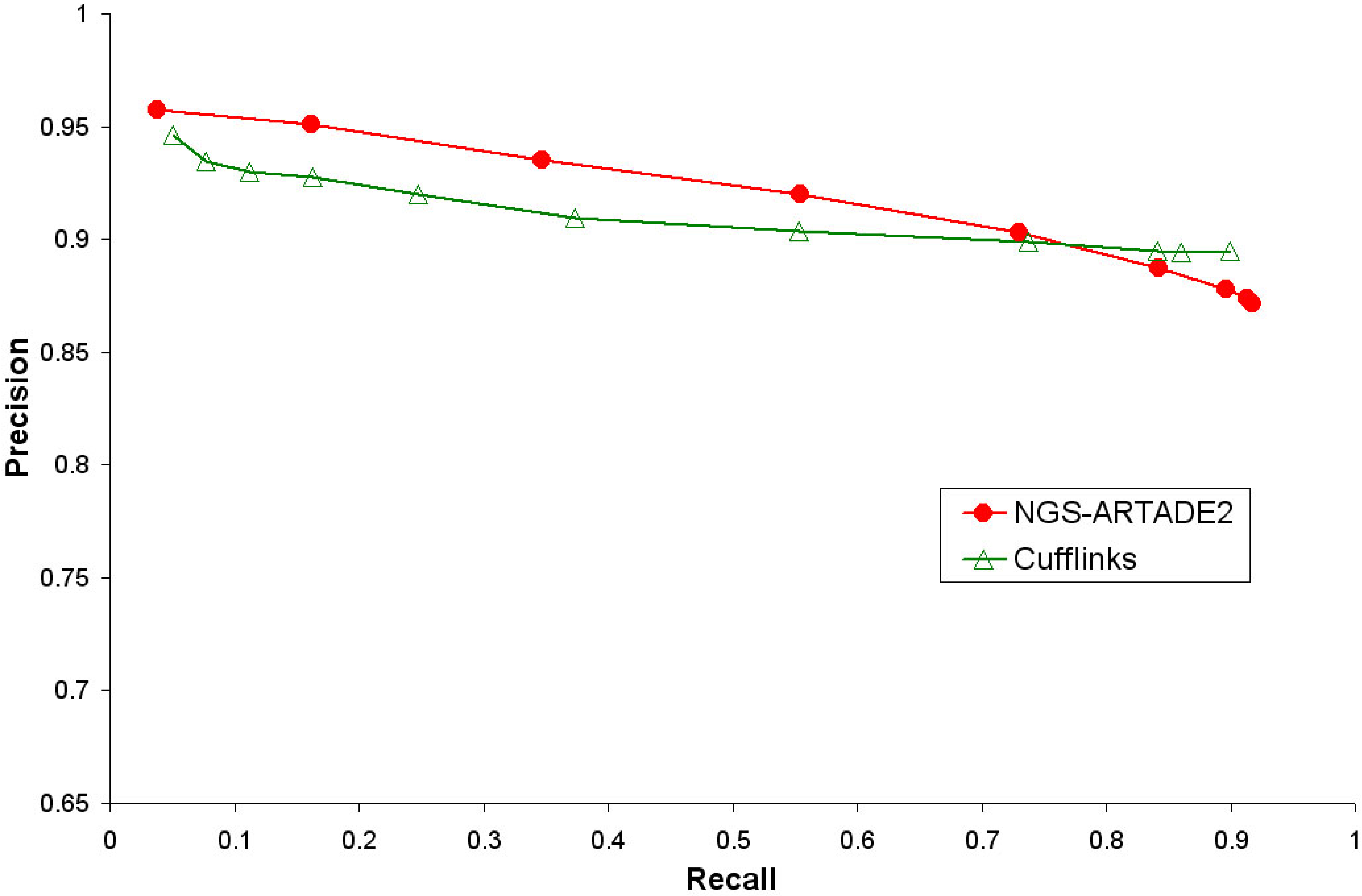

1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
Precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Recall
NGS-ARTADE2
Cufflinks

Figure S7

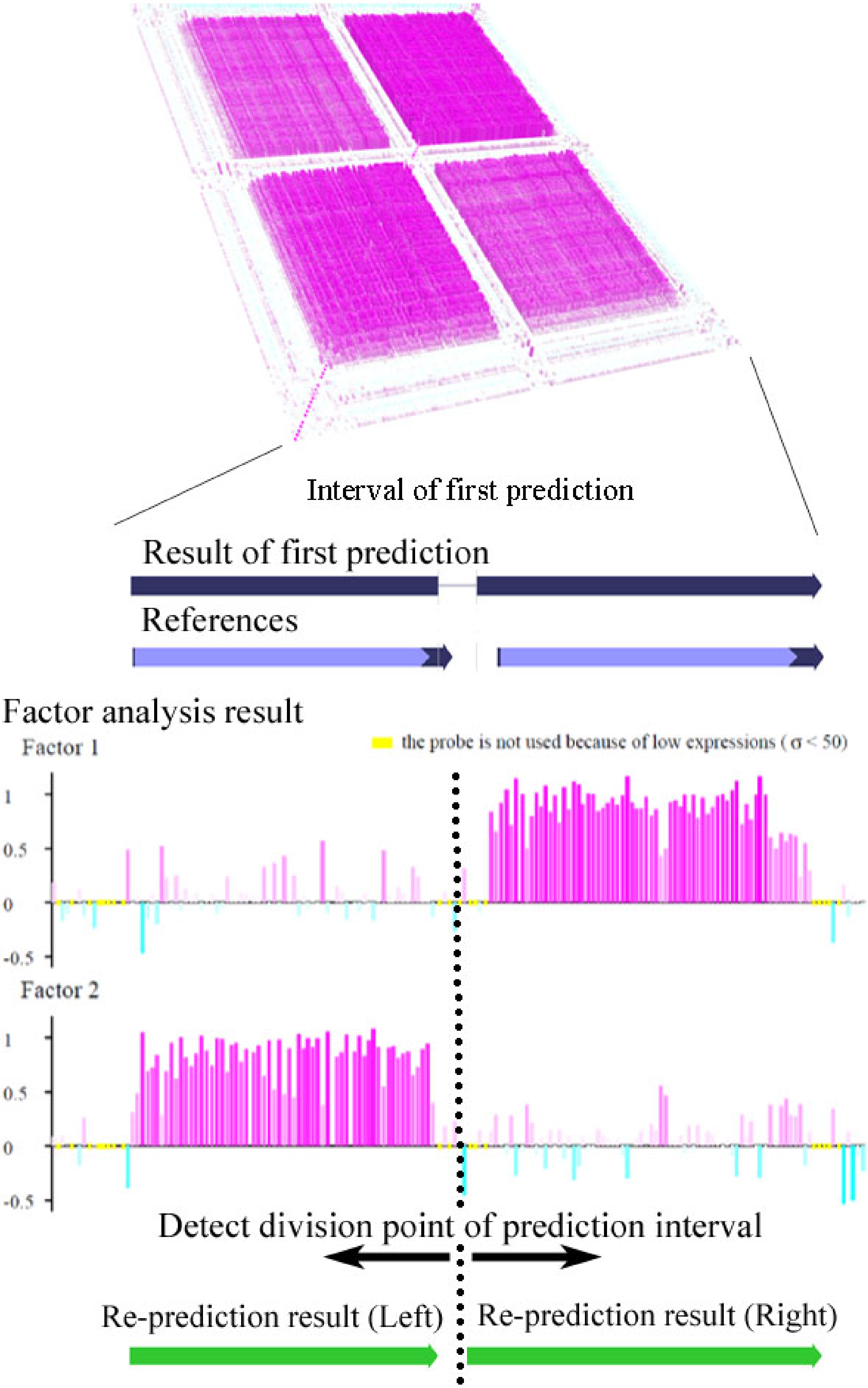

Interval of first prediction
Result of first prediction
References
Factor analysis result
Factor 1
🟨 the probe is not used because of low expressions (σ < 50)
1
0.5
0
-0.5
Factor 2
1
0.5
0
-0.5
Detect division point of prediction interval
Re-prediction result (Left)
Re-prediction result (Right)

Figure S8

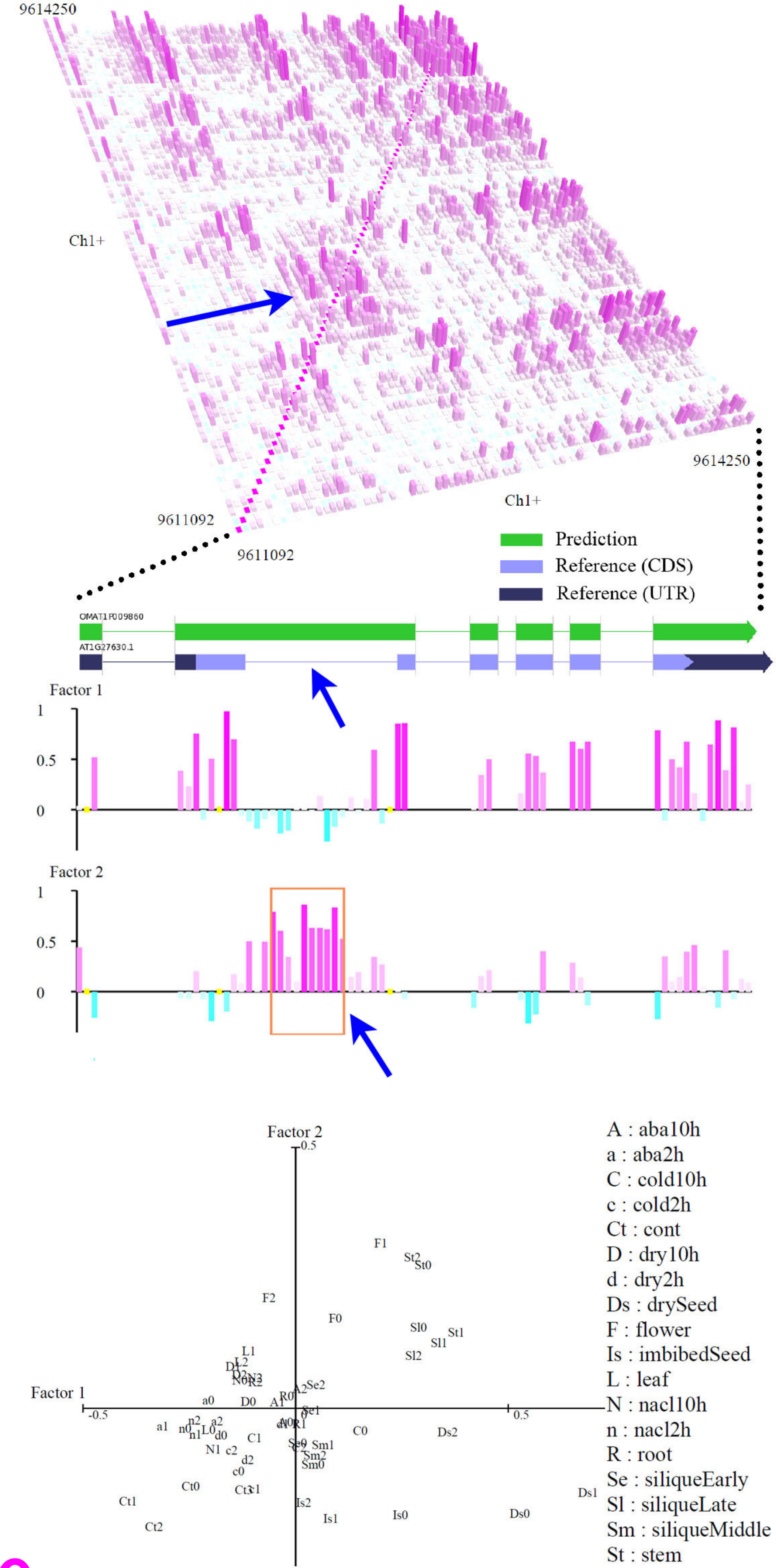

9614250
Ch1+
9611092
9611092
Ch1+
9614250
Prediction
Reference (CDS)
Reference (UTR)
OMAT1P009860
AT1G27630.1
Factor 1
1
0.5
0
Factor 2
1
0.5
0
Factor 2
0.5
Factor 1
-0.5
0.5
F1
St2
St0
F2
F0
Sl0
St1
Sl1
Sl2
L1
L2
D1
D2
N3
N0
R2
a0
D0
A1
R0
A2
Se2
Se1
a1
n0
n2
n1
L0
a2
d0
A0
R1
C1
N1
c2
Se0
Sm1
Sm2
Sm0
C0
Ds2
d2
c0
Ct0
Ct3
c1
Ct1
Is2
Is1
Is0
Ds0
Ds1
Ct2
A : aba10h
a : aba2h
C : cold10h
c : cold2h
Ct : cont
D : dry10h
d : dry2h
Ds : drySeed
F : flower
Is : imbibedSeed
L : leaf
N : nacl10h
n : nacl2h
R : root
Se : siliqueEarly
Sl : siliqueLate
Sm : siliqueMiddle
St : stem

Figure S9

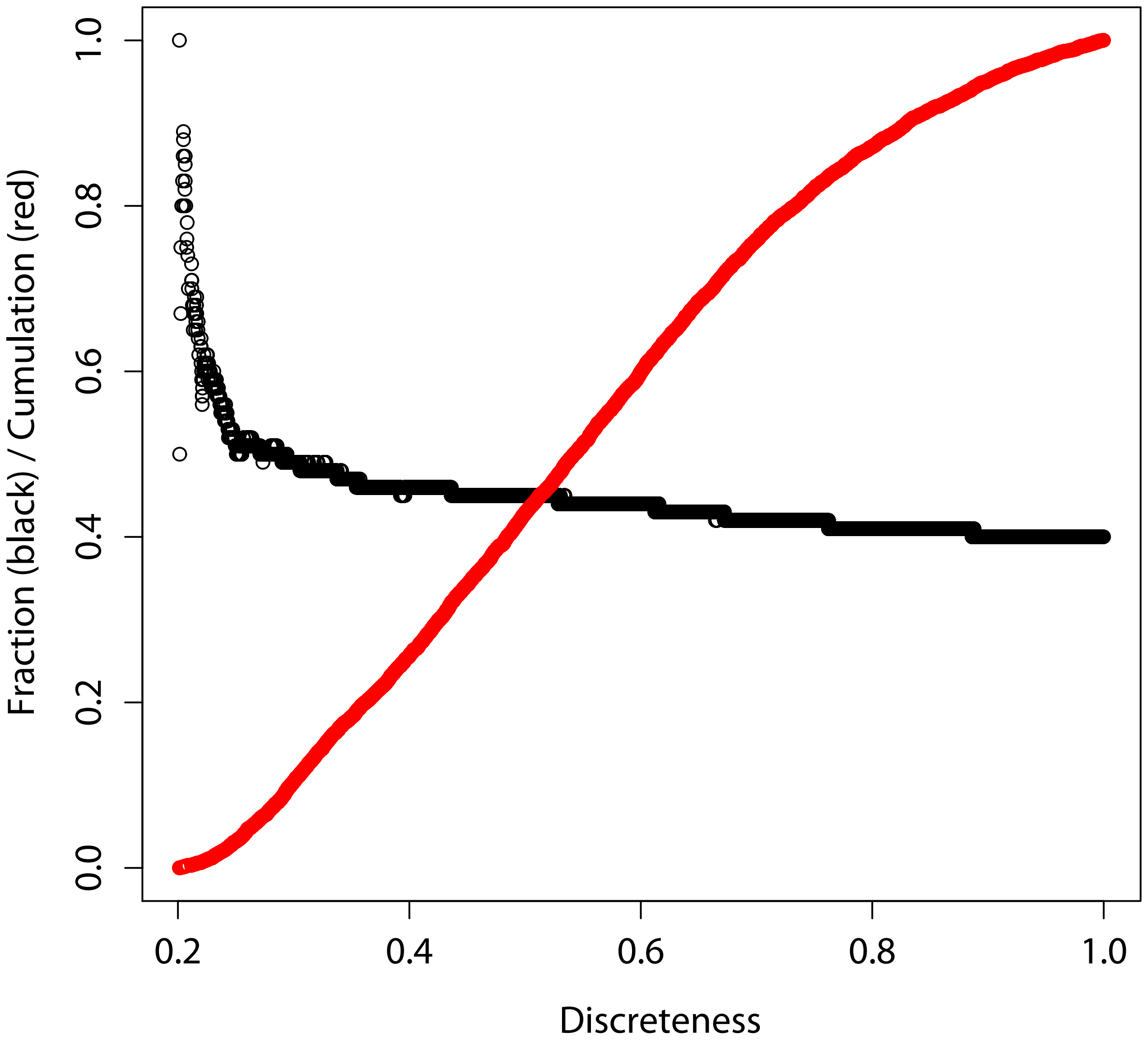

Fraction (black) / Cumulation (red)
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
Discreteness

Figure S10