Transcriptional regulation plays an important role in orches-
Arabidopsis
of reconstructed patterns on a schematic
online.
Bioinformatics
Supplementary data are available at
Supplementary information:
Contact:

process, and to identify transcriptional regulators that coordinate
wide gene expression data; the output of this transcriptional
sequencing technologies have allowed biologists to capture genome-
development [reviewed in Iyer-Pascuzzi and Benfey (2008)
trating a host of biological processes, particularly during

results: Given these partial observations, we present a novel
method for reconstructing integrated high-resolution spatiotemporal
data. Our method is based on a new iterative algorithm for finding
approximate roots to systems of bilinear equations.
availability: Source code for solving bilinear equations is
available at http://math.berkeley.edu/~dustin/bilinear/. Visualizations
of reconstructed patterns on a schematic Arabidopsis root are
contact: dustin@math.berkeley.edu
Supplementary information: Supplementary data are available at
Bioinformatics online.

1 INTRODUCTION
Transcriptional regulation plays an important role in orches-
trating a host of biological processes, particularly during
development [reviewed in Iyer-Pascuzzi and Benfey (2008)
and Levine and Davidson (2005)]. Advances in microarray and
sequencing technologies have allowed biologists to capture genome-
wide gene expression data; the output of this transcriptional
regulation. This expression data can then be used to identify
genes whose expression is correlated with a particular biological
process, and to identify transcriptional regulators that coordinate

the expression of groups of genes that are important for the same
biological process.
The identification of such genes and transcriptional regulators
is complicated by the complex heterogeneous mixture of cell
types and developmental stages that comprise each organ of an
organism. Expression patterns that are found only in a subset
of cell types within an organ will be diluted and may not be
detectable in the collection of expression patterns obtained from
RNA isolated from samples of an entire organ. Therefore, techniques
have been developed to enrich samples for specific cell types or
developmental stages, especially for studies in plants (Busch and
Lohmann, 2007). In the model plant, Arabidopsis thaliana, several
features of the root organ reduce its developmental complexity
and facilitate analysis. Specifically, most root cell types are found
within concentric cylinders moving from the outside of the root
to the inside of the root (Supplementary Fig. 1). These cell-type
layers display rotational symmetry, thus, simplifying the spatial
features of development. This feature has been exploited in the
development of a cell type enrichment method. This enrichment
method uses green fluorescent protein (GFP)-marked transgenic
lines and fluorescently activated cell sorting (FACS) to collect cell-
type-enriched samples and has allowed for the identification of
cell-type-specific expression patterns (Birnbaum et al., 2003, 2005).
Using this technique, high-resolution expression data have been
obtained for nearly all cell types in the Arabidopsis root (herein
called the marker-line dataset) (Brady et al., 2007, Jiao et al., 2009).
Another feature that makes the Arabidopsis root a tractable
developmental model is that cell types are constrained in files along
the root’s longitudinal axis and most of these cells are produced
from a stem cell population found at the apex of the root. This
feature allows a cell’s developmental timeline to be represented by
its position along the length of the root. To obtain a developmental
time-series expression dataset individual Arabidopsis roots were
sectioned into 13 pieces, each piece representing a developmental
time point (herein called the longitudinal dataset) (Brady et al.,
2007). Each of these sections, however, contains a mixture of cell
types, and the microarray expression values obtained are therefore
the average of the expression levels over multiple cell types present
at these specific developmental time points.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
While the 19 fluorescently marked lines in Brady et al. (2007) cover expression in nearly all cell types, they do not comprehensively mark all developmental stages of these cell types. Also, the procambium cell type was not measured, as a fluorescent marker-line which marks that cell type did not exist at the time. However, expression from the longitudinal dataset, does contain averaged expression of all cell types, and may be used to infer the missing cell-type data.

Previous studies have looked at separating expression data from Arabidopsis roots in which cell type and developmental stage are independent (Brady et al., 2007; Chaudhuri et al., 2008). However, neither method takes all data into account when reconstructing expression. In Brady et al. (2007), only high-relative gene expression is considered, and in Chaudhuri et al. (2008), no attempt is made to infer expression for cells not covered by any marker-line.

In this work, we formulate a model for expression levels in Arabidopsis roots in which cell type and developmental stage are independent sources of variation. The microarray data specifying overall expression levels for certain mixtures of cells lead to an overconstrained system of bilinear equations. Moreover, due to the nature of the problem, we are exclusively interested in positive real solutions. We present a new method for finding non-negative real approximate solutions to bilinear equations, based on the techniques of expectation-maximization (EM; Pachter and Sturmfels, 2005), approximate solutions to bilinear equations, based on the techniques of expectation-maximization (EM; Pachter and Sturmfels, 2005, approximation of expectation-maximization (EM; Pachter and Sturmfels, 2005) from likelihood maximization in statistics. Earlier work has used EM to find non-negative matrix factorizations (Lee and Seung, 2001), and our method is a generalization of that work.

We applied our method to estimate spatiotemporal subregion expression patterns for 20/872 Arabidopsis transcripts. These patterns have identified gene expression in cell types and developmental stages which were previously unknown. Visualizations of these patterns on a schematic Arabidopsis root are available at http://www.arexdb.org/.

### 2 METHODS

#### 2.1 Expression data

Our method uses the normalized expression data collected in Brady et al. (2007). Expression levels were measured across 13 longitudinal sections in a single root (longitudinal dataset) and across 19 different markers (marker-line dataset). For simplicity, the J2501 line was removed from further analysis as it is redundant with the WOODEN-LEG marker-line. The APL marker-line was also removed, as it contains domains of expression marked by both the S32 and SUC2 marker-lines and adds no extra information. The remaining 17 markers covering 14 cell types are listed in the second column of Table 1.

Due to computational constraints, the original normalization of this data was performed for the longitudinal and marker-line datasets independently (Brady et al., 2007). In order to account for differences caused by these separate normalization procedures, we adjusted the marker-line data by a global factor of 0.92. This factor was calculated by comparing the expression values of ubiquitous, evenly expressed probe sets between the two datasets. We assume that by comparing these probe sets, any true expression differences due to cell type and longitudinal section specificity should be minimal and thus any differences in expression level is a byproduct of the separate normalization procedures. A set of 43 probesets were identified which were expressed ubiquitously (above a normalized value of 1.0 in all samples) and whose expression did not vary significantly among samples within a dataset (ratio of min/max expression within a dataset is at most 0.5).

The scaling factor necessary to make the mean expression within the marker-line dataset equal to the mean expression within the longitudinal dataset was calculated for each probe set in this set. The median value of these 43 scaling factors was 0.92, which was used as the global adjustment factor (Supplementary Table 1).

#### 2.2 Model

To model the transcript expression level of an individual cell, we assume that the effects of its cell type and its section on its expression level are independent of each other. More precisely, we assume that the transcript expression level of a cell of type $i$ in section $j$ is equal to the product $x_i y_j$, where $x_i$ depends only on the section and $y_j$ depends only on the cell type.

In other words, for each transcript, there is an idealized profile of expression over different cell types, and an idealized profile of expression over different sections. Within a given section, our assumption is that the transcript expression level varies proportionally to its cell-type profile, and within a given cell type, proportionally to its longitudinal profile.

Although our model’s assumption of independence between cell-type and longitudinal section is a simplification, we believe it is appropriate for two reasons. First, we expect that in most cases, a transcript controlling development will have a single temporal pattern in the cell type or cell types in which it is active and have negligible expression elsewhere. This profile is consistent with our model by taking $y_j$ to be either high or almost zero depending on whether or not the transcript is expressed in that cell type. Second, the input longitudinal and marker-line datasets correspond roughly to independent measurements of the temporal and cell-type profiles of expression level. Thus, fitting independent temporal and cell-type profiles is less speculative than a more complicated model would be in the absence of more detailed data.

Each microarray sample in the two datasets (described in Section 2.1) is composed of a distinct mixture of cell types and sections. Within each sample, the measured transcript expression level is a convex linear combination of

### Table 1. The 14 cell types in the Arabidopsis root and the 17 marker-lines which mark them (Brady et al., 2007)

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Marker-lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiescent center</td>
<td>AGL42, RM1000, SCR5</td>
</tr>
<tr>
<td>Columella</td>
<td>PENT11</td>
</tr>
<tr>
<td>Lateral root cap</td>
<td>LRC</td>
</tr>
<tr>
<td>Hair cell</td>
<td>CORL9 (8–13)</td>
</tr>
<tr>
<td>Non-hair cell</td>
<td>GL2</td>
</tr>
<tr>
<td>Cortex</td>
<td>J0571, CORTEX (7–13)</td>
</tr>
<tr>
<td>Endodermis</td>
<td>J0571, SCR5</td>
</tr>
<tr>
<td>Xylem pole pericycle</td>
<td>WOL (2–9), J0121 (9–13), J2661 (13)</td>
</tr>
<tr>
<td>Phloem pole pericycle</td>
<td>WOL (2–9), S17 (8–13), J2661 (13)</td>
</tr>
<tr>
<td>Phloem</td>
<td>S32, WOL (2–9)</td>
</tr>
<tr>
<td>Phloem companion cells</td>
<td>SUC2 (10–19), WOL (2–9)</td>
</tr>
<tr>
<td>Xylem</td>
<td>S4 (2–7), S18 (8–13), WOL (2–9)</td>
</tr>
<tr>
<td>Lateral root primordia</td>
<td>RM1000</td>
</tr>
<tr>
<td>Procambium</td>
<td>WOL (2–9)</td>
</tr>
</tbody>
</table>

For markers that only mark the cell type in some of the sections, these sections are indicated by the range in parentheses.
Table 2. The cell-count matrix gives the number of cells in each spatiotemporal subregion

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>24</th>
<th>51</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>12</td>
<td>152</td>
<td>24</td>
<td>48</td>
<td>12</td>
<td>12</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>28</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>280</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>210</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>210</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>210</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>45</td>
<td>20</td>
<td>20</td>
<td>25</td>
</tr>
</tbody>
</table>

The 13 rows correspond to longitudinal sections 1 through 13. From left to right, the 14 columns correspond to the following spatiotemporal subregions: quiescent center, columella, lateral root cap, hair cell, non-hair cell, cortex, endodermis, xylem pole pericycle, phloem pole pericycle, phloem, phloem companion cells, xylem, lateral root primordia, and procambium.

The expression levels of its constituent cells. Under the above independence assumption, the longitudinal measurements give us a system of 13 equations,

\[ \sum_{j=1}^{14} a_{kj} y_j = b_k \quad \text{for } k = 1, \ldots, 13, \]

where \( a_k \) and \( y_j \) are the model parameters for the 13 sections and 14 cell types, respectively, \( a_{kj} \) is the proportion of the \( j \)-th cell type in the \( k \)-th section, and \( b_k \) is the measured expression level in the \( k \)-th longitudinal section. The coefficients \( a_{kj} \) form a \( 13 \times 14 \) matrix, where the \( k \)-th row is proportional to the \( k \)-th row of the cell-count matrix (Table 2), but rescaled to sum to 1.

In addition, the 17 marker-line measurements give the additional equations,

\[ \sum_{j=1}^{14} a_{kj} y_j = b_k \quad \text{for } k = 14, \ldots, 30, \]

where the \( b_k \)s are the measured expression level of the 17 markers, indexed with \( k \) from 14 to 30, and \( a_{kj} \) is the proportion of the cells in the \( k \)-th marker-line which come from the \( j \)-th section and the \( j \)-th cell type. For each marker-line \( k \), we denote by \( a_{kj} \) the matrix which has \( a_{kj} \) in the \( i \)-th row and \( j \)-th column. The matrix \( a_{kj} \) is zero except for those spatiotemporal subregions marked by that marker as indicated in Table 1. The non-zero entries of \( a_{kj} \) are proportional to the corresponding entries of the cell matrix, but rescaled to sum to 1.

In order to combine (1) and (2) in a uniform system of equations, we define \( a_{kj} = a_{kj}^* \) for \( k \leq 13 \) and \( a_{kj} = 0 \) for \( k \geq 14 \). Then (1) and (2) together give

\[ \sum_{i=1}^{15} \sum_{j=1}^{14} a_{ij} x_{ij} y_j = b_k \quad \text{for } k = 1, \ldots, 30 \]

2.3 Cell matrix

As described in the previous section, the coefficients \( a_{ij} \) in our model depend on the number of cells in each spatiotemporal subregion. These cell number estimates were generated by visual inspection of successive optical cross-sections of Arabidopsis roots along the longitudinal axis using confocal laser scanning microscopy. For the xylem, phloem and procambium cell types, cell counts were obtained from earlier experiments (Bonke et al., 2003; Mahonen et al., 2000). What follows is a detailed description of this visual and literature analysis. These results are also summarized in Table 2.

Longitudinal section 1 encompasses 2 tiers of 12 columella cells and 3 tiers of lateral root cap cells (15, 18 and 18 moving up from the tip).

Longitudinal section 2 contains 1 tier of 12 columella cells and 6 tiers of lateral root cap cells (20, 20, 28, 28, 28 and 28 moving up from the tip). For all other cell types in longitudinal section 2, three tiers of cells are present. Eight trichoblast (hair cell precursor) cells and 16 atrichoblast (non-hair cell precursor) cells are present circumferentially throughout the root, resulting in 24 and 48 cells, respectively, in the hair cell and non-hair cell precursor file in longitudinal section 2. Throughout the root, eight cortex and eight endodermis cells are present circumferentially. However, in longitudinal section 2, the cortex/endodermis initial is undergoing asymmetric periclinal divisions to produce the cortex and endodermis cell files, so we consider there to be approximately 0.5 cells of the cortex and endodermis type, resulting in 12 cells of each type in longitudinal section 2. When the Arabidopsis root is 7 days old, each longitudinal section from 3–13 contains approximately five cells of each type along the root’s longitudinal axis.

In longitudinal section 2, the tangential and periclinal divisions that give rise to phloem cell files do not occur, but do occur in longitudinal section 3 (Bonke et al., 2003). Three cells are present in the main xylem axis in the first tier of cells, four cells in the second tier and five cells in the third tier (Mahonen et al., 2000). Eight procambial cells are present in the first cell tier, 12 procambial cells in the second tier and 18 cells in the third tier resulting in 28 procambial cells in longitudinal section 2 (Mahonen et al., 2000). For all the sections, xylem pole pericycle cells are the two cells that flank the xylem axis on either end, and phloem pole pericycle cells are considered the intervening cells. Four pericycle cells can be identified as flanking xylem cells in all three tiers of cells present in longitudinal section 2 (Mahonen et al., 2000). Seven intervening phloem pole pericycle cells can be found in tier one, and eight intervening cells can be identified in the third tier (Mahonen et al., 2000), resulting in 22 procambial cells in longitudinal section 2.

In a 7-day-old root, each of the longitudinal sections 3–13 contains approximately five tiers of cells. In longitudinal section 3, columella cells can no longer be identified, and 10 tiers of lateral root cap cells exist containing 28 cells each. In sections 4–6, a lateral root cap cell is twice the length and half the width of an epidermal cell. Eighty-four cells were identified in each tier, and two and a half tiers of cells each for longitudinal sections 4–6 resulting in 210 cells for each longitudinal section. Secondary cell growth does not occur in the developmental stages sampled, therefore, this number remains fixed throughout all developmental stages. In longitudinal section 12, a non-emerged lateral root is hypothesized to be present based on microarray expression data (Brady et al., 2007). This lateral root is estimated to be approximately 130 cells, or one tier of cells in longitudinal section 2.
2.4 Solving bilinear equations

In this section, we present our method for solving the system of bilinear equations given in (3). More generally, we have a system

\[ f(x, y) = \sum_{k=1}^{\ell} a_{ij}(s, t) y_k = b_k \quad \text{for } k = 1, \ldots, \ell \]

(4)

In our application, we have \( n = 13, m = 14 \) and \( \ell = 30 \). Unlike other numerical methods for solving systems of polynomial equations, our algorithm has the advantage that it finds only non-negative, real solutions. Moreover, even in systems where there are no exact solutions, as will generally be the case in an overconstrained system of equations, our method will find approximate solutions. A more detailed, technical mathematical study of the method will be available in a forthcoming paper by D.A.C.

Our method is based on the EM (Pachter and Sturmfels, 2005, Section 1.3) and IPF (Darroch and Ratcliff, 1972) algorithms used for maximum likelihood estimation in statistics. These are iterative algorithms which be available in a forthcoming paper by D.A.C. However, relaxing these conditions does not change the convergence proof. Both the EM and IPF algorithms are local searches which may converge only to local minima, and so, for each transcript, we ran our algorithm 20 different times starting from 20 different randomly chosen starting points. For every transcript in our data, all 20 runs of the algorithm converged to the same solution, up to a small tolerance. We therefore believe that in almost all cases we have found a global, and not merely local, minimum to the modified Kullback–Leibler divergence. Most likely, this consistency is a consequence of the particular coefficients of our equations, and in general there may be multiple local minima.

2.5 Computational validation methodology

In order to validate our method, we simulated expression profiles according to various models and tested our method’s ability to reconstruct the underlying parameters. First, we simulated data according to the same independence model defined in Section 2. The underlying spatiotemporal subregion expression levels were sampled from a log-normal distribution with SD 0.5. The simulated measurements \( b_k \) were computed from these subregion levels according to our model of the Arabidopsis root in (3). Finally, multiplicative error was added, distributed according to a log-normal distribution with SD 0.03 to simulate measurement noise. This procedure created expression data with varying but comparable expression levels, which we will call the ‘uniform’ dataset. However, since we are particularly interested in genes for which the expression levels are not uniform, we also produced simulations with the expression level for a given section or cell type raised by a factor of 10, which we will call the ‘elevated’ dataset. In this dataset, we only measured the error for the same section or cell type which was elevated. These simulations measure our ability to detect a dominant expression pattern.

In addition, we designed simulations that test the robustness of the algorithm to failures of the bilinear model for root expression levels. For each section and cell type, we simulated data in which the expression levels for cells in that section or cell type did not follow the bilinear model, and call these the ‘section’ and ‘cell type’ datasets, respectively. Instead, the expression levels in the given section or cell type were chosen independently according to a log-normal distribution with SD 0.5 x \( \sqrt{2} \). The factor of \( \sqrt{2} \) was introduced because the product of two log-normally distributed numbers with SD 0.5 is distributed log-normally with SD 0.5 x \( \sqrt{2} \).

The predictions were compared with the true expression levels across the spatiotemporal subregions within each section and each cell type. For each section and each cell type, the expression levels in its spatiotemporal subregions were averaged; ignoring these combinations which are not physically present in the root (i.e. those whose entry in Table 2 is 0). The difference between the predicted and true average expressions was computed as a proportion of the true average expression. We then computed the root mean square of the proportional error over 500 simulations.

2.6 Visualization of predicted expression patterns

Predicted expression values were colored according to an Arabidopsis root template (Supplementary Fig. 1). The green channel of each cell was set according to a linear mapping between the expression range shown in the template [1, 10] or [1, 5] to the range [0, 255]. Expression values above or below that range are given values of 255 or 0, respectively. The mapping is also shown to the right of the false color image in the form of a gradient key. Phloem cells by longitudinal section are visualized separately on the right-hand side of the root as they are physically occluded by other cells.
in the left-hand side representation. The minimum and maximum range of expression value visualized can also be adjusted by the user.

2.7 In vivo validation methodology

To validate predicted expression values, we used transgenic Arabidopsis thaliana lines containing transcriptional GFP fusions in the Columbia ecotype (Lee et al., 2006). For each gene being validated, six plants from at least two insertion lines previously described as expressing GFP were characterized. All plants were grown vertically on 1× Murashige and Skoog salt mixture, 1% sucrose and 2.3 mM 2-(N-morpholino)ethane sulfonic acid (pH 5.7) in 1% agar. Seedlings were prepared for microscopy at 5 days of age. Confocal images were obtained using a 25× water-immersion lens on a Zeiss LSM-510 confocal laser-scanning microscope using the 488 nm laser for excitation. Roots were stained with 10 μg/mL propidium iodide for 0.5-2 min and mounted in water. GFP was rendered in green and propidium iodide in red. Images were saved in TIFF format. Images were manually stitched together in Adobe Photoshop CS2 using the Photomerge command. The black background surrounding the root was modified to ensure uniformity across figures. No other image enhancement was performed.

3 RESULTS

3.1 Computational validation

The root mean square percentage errors in the reconstruction of each parameter are shown in Table 3. In the first two columns, where the data were generated according to the bilinear model, the error rate is generally no greater than the simulated measurement error. In most cases, elevated expression led to a lower error rate. In particular, reconstruction of expression in procambium was much more accurate in the elevated dataset.

The last two columns show that the algorithm is robust to violations of the bilinear model. Also, the predicted expression level in each cell type is generally not greatly affected by the failure of the model in other cell types, and similarly with sections.

3.2 In vivo validation

To determine whether our algorithm is able to accurately resolve spatiotemporal subregion-level transcript expression values, it would be ideal to compare the predictions to measured microarray expression values of the same spatiotemporal subregion. However, due to technical constraints, it is not possible to measure mRNA expression to such a degree of specificity and thus we cannot validate the estimates directly. Instead, we validated the method by visually comparing the predicted pattern of expression to patterns obtained from transcriptional GFP fusions using laser scanning confocal microscopy, as described in Lee et al. (2006).

For each gene validated, a false-colored root image was generated by coloring each spatiotemporal subregion of an annotated Arabidopsis root template (Supplementary Fig. 1) according to the expression level in that subregion as predicted by our method. This false-colored image was then visually compared against the actual pattern of fluorescence observed in plants expressing a transcriptional GFP fusion specific for the promoter of the gene. These transcriptional GFP fusions contain up to 3 kb of regulatory sequence upstream of the translational start site of the respective gene. In many cases, this sequence is sufficient to recapitulate endogenous mRNA expression patterns as defined by cell-type resolution microarray data (Lee et al., 2006). This comparative method of validation allows us to assess the accuracy of spatiotemporal subregion expression predictions in an efficient and technically feasible way.

As a benchmark validation test, a set of three transcriptional fusions which were used to obtain some of the marker-line dataset were examined: S18(At5g12870), S46(At3g25710) and S32(At2g18380). These fusions were originally selected for the use in profiling because they exhibited enriched cell-type expression as observed by laser scanning confocal microscopy and subsequently confirmed in the microarray expression data. The expression predictions from our method accurately recapitulated the observed pattern of all three benchmark genes (Fig. 1 and data not shown).

To assess the novel predictive ability of our method to reconstruct in vivo expression patterns given missing data, we selected transcriptional fusion genes for which our method predicts expression in cell types or spatiotemporal subregions that were not marked by fluorescent marker-lines in the original dataset. At least two lines per transcriptional fusion were monitored. With respect to an unmarked cell type, our method predicted that AT4G37940 was highly expressed in the columella and developing procambium.

---

Table 3. Root mean square percentage error rates in the reconstruction of simulated data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Error rate</th>
<th>Uniform</th>
<th>Elevated</th>
<th>Cell type</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>2.7</td>
<td>3.3</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>3.4</td>
<td>5.7</td>
<td>7.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 3</td>
<td>3.3</td>
<td>5.8</td>
<td>7.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 4</td>
<td>3.2</td>
<td>5.3</td>
<td>6.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 5</td>
<td>3.1</td>
<td>5.3</td>
<td>6.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 6</td>
<td>3.3</td>
<td>5.3</td>
<td>6.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 7</td>
<td>3.1</td>
<td>3.7</td>
<td>5.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 8</td>
<td>3.0</td>
<td>3.6</td>
<td>4.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 9</td>
<td>3.0</td>
<td>3.6</td>
<td>4.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 10</td>
<td>2.7</td>
<td>3.5</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 11</td>
<td>2.9</td>
<td>3.4</td>
<td>4.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 12</td>
<td>3.3</td>
<td>4.4</td>
<td>5.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 13</td>
<td>2.4</td>
<td>3.6</td>
<td>5.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quiescent center</td>
<td>3.0</td>
<td>3.0</td>
<td>3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columella</td>
<td>3.1</td>
<td>4.9</td>
<td>4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral root cap</td>
<td>2.6</td>
<td>3.6</td>
<td>3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hair cell</td>
<td>3.4</td>
<td>9.1</td>
<td>4.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-hair cell</td>
<td>3.0</td>
<td>3.1</td>
<td>3.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cortex</td>
<td>2.9</td>
<td>6.9</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endodermis</td>
<td>2.8</td>
<td>3.5</td>
<td>3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xylem pole pericycle</td>
<td>3.3</td>
<td>10.8</td>
<td>4.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phloem pole pericycle</td>
<td>3.0</td>
<td>9.4</td>
<td>4.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phloem</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phloem css</td>
<td>3.3</td>
<td>11.7</td>
<td>4.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xylem</td>
<td>2.2</td>
<td>2.5</td>
<td>2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral root primordia</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procambium</td>
<td>8.3</td>
<td>12.7</td>
<td>12.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first column is under a model of comparable but varying expression levels across all sections and cell types. The second type is the error rate when that section or cell type has its expression level raised by a factor of 10. In the third and fourth columns, the error rate is generally no greater than the error rate when that section or cell type has its expression level raised by a factor of 10. The third and fourth columns show models in which the bilinear assumption is violated in one of the sections or one of the cell types, respectively. In all cases, 3% measurement error has been added to the expression level.
Imaging of a transcriptional fusion of this gene confirmed this expression (Fig. 2).

To determine if our method could correctly differentiate expression in a specific developmental stage of a cell type, we selected AT5G43040 for further analysis. The collection of marker-lines used to generate the original dataset included a marker for all developmental stages of non-hair cells, composed of their precursors (atrichoblasts) and fully developed non-hair cells. However, the marker-line used for hair cells only marks mature hair cells, and not their precursors (trichoblasts). Our method predicts AT5G43040 expression throughout the epidermis—in mature hair cell, trichoblast, mature non-hair cell and atrichoblast cell files—with higher expression predicted in non-hair cells than in hair cells. This differential expression was validated using the AT5G43040 transcriptional fusion (Supplementary Fig. 2) demonstrating that our method is not only able to identify expression in a developmental stage of a cell type not marked by the marker-line data, but also to accurately differentiate relative levels of a transcript. However, it should be noted that expression in the transcriptional fusion did not fully corroborate the expression predicted by our algorithm—specifically, expression was found in the lateral root cap which was not predicted by our algorithm.

4 DISCUSSION

We have shown that spatiotemporal patterns of gene expression in the Arabidopsis root can be reconstructed using information from the marker-line and longitudinal datasets. Current experimental techniques are limited in their ability to rapidly and accurately microdissect organs into all component cell types at all developmental stages. Our computational technique helps to overcome these limitations. We fully integrate the marker-line and longitudinal datasets into a comprehensive expression pattern, across both space and time. In particular, this method has enabled the identification of Arabidopsis root procambium and trichoblast-specific genes, which have been previously the experimentally intractable cell types.
Our high-resolution expression patterns will allow us to better understand the regulatory logic that controls developmental processes of the *Arabidopsis* root. These transcriptional regulatory networks are key to understanding developmental processes and environmental responses. With only a portion of these genes and fewer cell types, high-resolution spatiotemporal data has been used to identify transcriptional regulatory modules (Brady et al., 2007). Our more accurate and complete dataset will allow a more comprehensive discovery of regulatory networks across additional cell types.

Moreover, we expect that our algorithm and the model which underlies it are applicable to time course experiments on other heterogeneous cell mixtures. Measurements in multicellular organisms are taken from complex cell mixtures of organs, tissues, heterogeneous cell lines or cancerous samples. When precise histological characterization of these samples can estimate underlying cell-type composition, our method can be used to reconstruct the underlying cell-type-specific gene expression patterns or any other type of quantitative data, such as high-throughput protein abundance measurements. Theoretically, this algorithm can be applied to identify missing data in any experimental system that captures data in two or more dimensions which are assumed to be independent of one another.

**ACKNOWLEDGEMENTS**

The authors would like to thank Wolfgang Busch, Anjali Iyer-Pascuzzi, Terri Long, Anne Shiu and Rossangela Sozzani for critical review of the manuscript.

**Funding:** Defense Advanced Research Projects Agency project ‘Microstates to Macrodynamics: A New Mathematics of Biology’; the National Science Foundation project AT2010 (to S.M.B., D.A.O. and P.N.B.); the National Sciences and Engineering Research Council (to S.M.B.).

**REFERENCES**


