The Annals of Applied Statistics
2015, Vol. 9, No. 2, 950–968
DOI: 10.1214/15-AOAS820

BAYESIAN DETECTION OF EMBRYONIC GENE EXPRESSION ONSET IN *C. ELEGANS*[1]

BY JIE HU[*], ZHONGYING ZHAO[†], HARI KRISHNA YALAMANCHILI[‡], JUNWEN WANG[‡], KENNY YE[§] AND XIAODAN FAN[*]

Chinese University of Hong Kong[*], Hong Kong Baptist University[†], University of Hong Kong[‡] and Albert Einstein College of Medicine[§]

To study how a zygote develops into an embryo with different tissues, large-scale 4D confocal movies of *C. elegans* embryos have been produced recently by experimental biologists. However, the lack of principled statistical methods for the highly noisy data has hindered the comprehensive analysis of these data sets. We introduced a probabilistic change point model on the cell lineage tree to estimate the embryonic gene expression onset time. A Bayesian approach is used to fit the 4D confocal movies data to the model. Subsequent classification methods are used to decide a model selection threshold and further refine the expression onset time from the branch level to the specific cell time level. Extensive simulations have shown the high accuracy of our method. Its application on real data yields both previously known results and new findings.

1. Introduction. The process of how a single-cell zygote develops into an embryo with different tissues is still a fundamental but open problem in biology. Undoubtedly, gene expression dynamics plays a key role in this procedure. Understanding when and where a gene starts expression in the embryo, that is, the embryonic gene expression onset, is a crucial step for solving this puzzle.

Modern high throughput experimental techniques, such as microarray experiments and time-lapse confocal microscopy, can produce gene expression data with high spatial and temporal resolution, which is necessary for the study of embryogenesis. *C. elegans* is often used as the model organism for embryogenesis study due to its transparency and invariant cell lineage from zygote to adult [Sulston et al. (1983)]. Bao et al. (2006) and Murray et al. (2008) introduced a system to automatically analyze the continuous reporter gene expression in *C. elegans* with cellular resolution from zygote to embryo using the confocal laser microscope. With this automatic system, Murray et al. (2012) analyzed the expression patterns of 127 genes and provided a compendium of gene expression dynamics. Long et al. (2009) and Liu et al. (2009) also developed an analyzer to convert high-resolution

Received September 2014; revised January 2015.

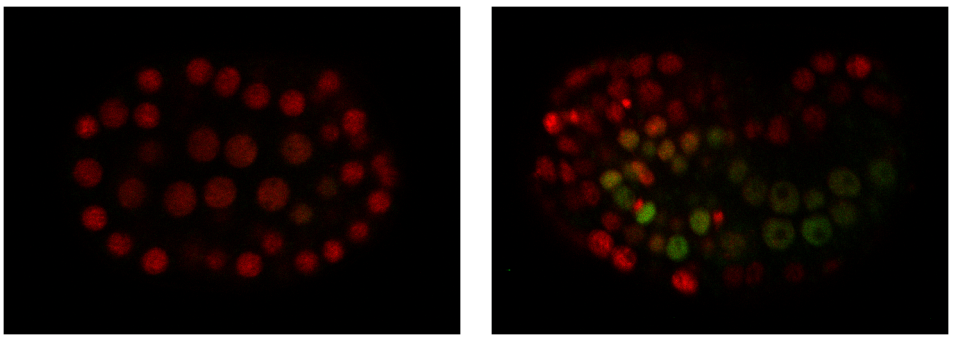
[1]Supported in part by a grant from the Research Grants Council of the Hong Kong SAR (Project no. CUHK 400913).

Key words and phrases. 4D confocal microscopy, embryonic onset, change point detection, Bayesian method.

confocal laser microscope images into data tables and then analyzed cell fate from gene expression profiles. Later Spencer et al. (2011) took advantage of a spatial and temporal map of *C. elegans* gene expression to provide a basis for establishing the roles of individual genes in cellular differentiation.

The aforementioned confocal microscopy on *C. elegans* embryogenesis is for tracing the expression of one specific target gene on an individual embryo. Due to strain differences (such as the insertion of green fluorescent protein DNA sequence into different locations of the *C. elegans* genome) and variability in experimental and environmental factors, even data sets for measuring the same gene show high quantitative variation, indicating considerable noise. Furthermore, the expression change on the embryonic cell lineage poses a change point problem on a binary tree, which is a nonlinear problem rarely studied by current literatures. The lack of principled statistical methods makes the comprehensive understanding of these data sets too crude to be convincing. For example, Murray et al. (2012) used an ad hoc threshold to report the expression onset among all the data sets, which ignored the variation among different runs of confocal microscopy. Here, we apply a Bayesian method for automatic detection of gene expression onset from the 4D confocal microscopy data by introducing experiment-specific background and signal distributions, which in turn can benefit downstream analysis, such as gene network inference based on such high spatial and temporal data [Yalamanchili et al. (2013)].

Our real data application is based on the data provided by Murray et al. (2012), which is downloadable from http://epic.gs.washington.edu/. Figure 1 shows the confocal fluorescent images of two stages of an embryo. The green fluorescent protein labels the expression product of the gene PHA-4, which appears in the



(A) 150-cell stage (B) 550-cell stage

FIG. 1. *Confocal fluorescent images of a* C. elegans *embryo*. (A) *The embryo at the* 150*-cell stage, with ubiquitous labeling of nuclei by red fluorescent protein mCherry*; (B) *The embryo at the* 550*-cell stage, with ubiquitous labeling of nuclei by red fluorescent protein mCherry and specific labeling of the gene expression product of PHA*-4 *by green fluorescent protein. The expression cells are in pharynx and intestine.*

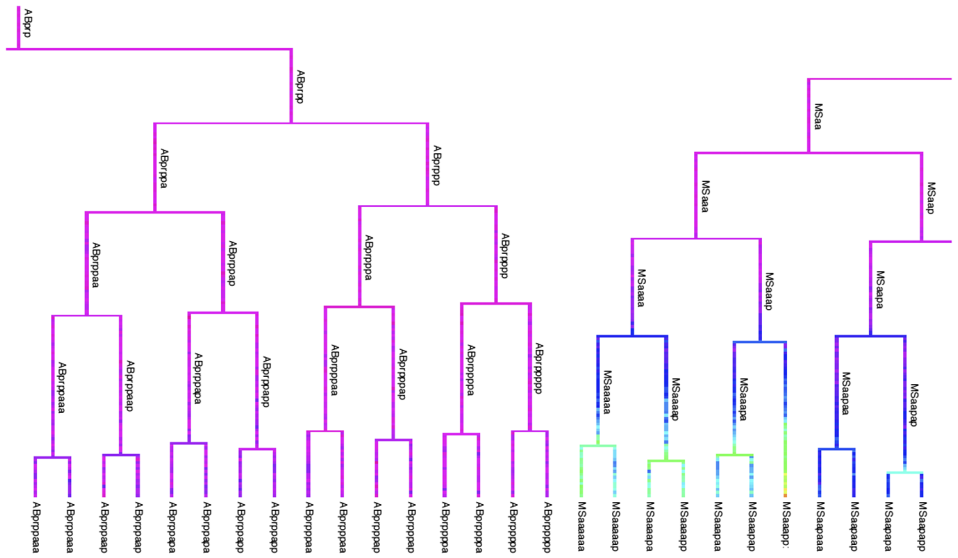


FIG. 2. *An example of data. The figure shows a part of a cell lineage tree for one data file, which represents the measured fluorescent intensity from one time-lapse confocal microscopy experiment on* C. elegans. *The cell lineage and cell nomenclature are from Sulston et al.* (1983).

550-cell stage but not in the 150-cell stage. Figure 2 shows a part of a cell lineage tree from one data file, which corresponds to one run of the confocal microscopy on one embryo. Each horizontal line represents a cell division event. Each vertical line represents a cell with the length proportional to its lifetime (i.e., how long a cell lived). The color at each point of the vertical line represents the measured fluorescent intensity at the corresponding time, which gradually increases with the color changing from purple to red as the rainbow color order. (In the later content, we use figures with gray scales to represent tree structures and measured fluorescent intensity gradually increases with the color changing from white to black.) The blue and green cells in Figure 2 form a cluster whose fluorescence level is significantly higher than the overall background, which indicates that they may express the target gene. Thus, estimating expression onset is actually a change point detection problem. However, methods used to detect change points in regular one-dimensional time series, such as Guralnik and Srivastava (1999), Picard (1985) and Perreault et al. (2000), cannot handle the tree structure in our case.

In Section 2 we present a four-step method to detect the onset time, where the key step is a Bayesian algorithm to fit a change point model to the tree data. We apply this method on both synthesized data and real data and show the estimation results in Section 3. Section 4 concludes the paper. Other details of the algorithm and the model diagnosis are provided in supplemental materials [Hu et al. (2015)].

2. Methods. We introduce the following four-step method for the Detection of Embryonic Gene Expression Onset (DEGEO), where the key step features a probabilistic change point model on the cell lineage tree and a full Bayesian approach to infer the cell where a gene starts expressing:

- *Step* 1: summarize the measured fluorescent intensity of each cell to a single cell score.
- *Step* 2: fit the tree of cell scores to a change-point-in-tree model in order to detect an expression branch, where a Markov chain Monte Carlo (MCMC) algorithm is used to estimate the change point and the other model parameters.
- *Step* 3: use Support Vector Regression (SVR) to decide when to stop detecting extra expression branches.
- *Step* 4: refine the onset detection by detecting the specific onset time on the reported expression branches.

2.1. *Experiment and data*. For each 4D confocal laser scanning microscope experiment performed by Murray et al. (2012), we have a data file containing a time series for each embryonic cell from its birth to its division or death. Each measurement is a fluorescent intensity at each time point (on average, one data value per minute) over the duration of the cell's life. We use the time series data of Column "blot" in the data files downloaded from http://epic.gs.washington.edu/, which has been normalized in order to reduce the influence of background noise. We represent this measured fluorescent intensity data of the ith cell at the jth time point by y_{ij}. Other details about the real data are provided in the Appendix.

2.2. *Assumptions*. During the embryogenesis process, once a cell initializes the expression of a gene, its descendants will inherit some of this gene's products and may also continue expressing this gene. Thus, a positive correlation between relatives is expected. Therefore, we make a transitivity assumption by assuming the following: if a cell expresses a gene, its child cells will also express the corresponding gene and the gene expression values of the two sibling cells are positively correlated. This assumption is justified by the data as shown in Supplement A.

Experimenters have two methods to mark an expressed gene, namely, promoter fusion and protein fusion [Murray et al. (2008)]. A special characteristic of the data from promoter fusion is that the fluorescent protein degrades much more slowly than that of protein fusion. Thus, once a cell initializes a gene's expression, the resulting fluorescent intensity will be inherited by its descendants and seldom decreases. If the child cells continue expressing this gene, the fluorescent intensity will increase due to the accumulation of the fluorescent protein. Since most experiments are based on promoter fusion, we assume a general nondecreasing trend for the fluorescent intensities along the paths from ancestors to descendants on an expression branch. Again, we use the data to justify this assumption as shown in Supplement A. This paper focuses on data from promoter fusion.

Furthermore, we assume that when a gene is not expressed in a cell, its cell score, which is defined in Section 2.3.1, follows a normal distribution with parameters μ and σ_1^2. And the histogram of two control files are listed in Supplement A.

2.3. *The DEGEO procedure.*

2.3.1. *Step* 1: *Summarize the time series of a cell into one cell score.* Due to the abnormal fluctuation of y_{ij} right before and after the cell division time, we truncate the first two and last two data points for all cells whose lifetimes are more than 8 time points (96.2% of the cells belong to this group). Cells with fewer data points are truncated less. The remaining data points are called the valid data. We define the cell score for each cell as

$$x_i = \frac{y_i^{(0.05)} + y_i^{(0.95)}}{2},$$

where $y_i^{(0.05)}$ and $y_i^{(0.95)}$ denote the 5% and 95% quantiles of the time series $\{y_{ij}\}$ of the ith cell, respectively. The cell score is designed in this way such that a true expression signal (which should last longer than 5% of the cell's lifetime) could be captured even if the expression lasts shorter than half of the cell's lifetime (in this case, taking median may not discover the expression). On the other time, rare outlier values (which should not occupy more than 5% of the cell's lifetime) can be filtered out from the cell score. In contrast, a median will miss short trends while a mean will be too sensitive to outliers. Thus, a 4D confocal movie data file is transformed to a tree of cell scores. The cell scores X, together with the lifetimes T and their family relationships, will be used in step 2 to detect the cells where the target gene start expressing.

2.3.2. *Step* 2: *Fit a change-point-in-tree model.* Let x_{i_1} and x_{i_2} be the cell scores of a pair of sibling cells while x_{i_0} indicates that of their mother cell. Let t_{i_1} and t_{i_2} be the lifetimes of the cells corresponding to x_{i_1} and x_{i_2}. M indicates the change point, that is, the cell where the target gene starts expressing. Therefore, all descendant cells of the cell M form a branch, which we call an expression branch. In the case that cells with close kinship are expression onsets simultaneously, the change point may be the most recent common ancestor cell of all expression cells. In this case, the expression branch may contain cells which have not expressed the target gene. For example, in Figure 5, the exact expression onset cells are Exxx, but our algorithm reports the cell E as change point in step 2. Nevertheless, our algorithm will refine it to Exxx in step 4.

Denote $A(x_i)$ as the set of all ancestor cells of the cell corresponding to the cell score x_i. If a cell corresponding to the cell score x_i is not within the expression branch, that is, $M \notin A(x_i)$, we assume its cell score is independent and identically distributed (i.i.d.) Gaussian noise. For a cell in the expression branch, its cell

score is assumed to be associated with its mother, its sibling and its lifetime. More specifically, the two kinds of cell scores are modeled by a change-point-in-tree model as follows:

$$
\begin{aligned}
&x_i | M \notin A(x_i) \sim N\big(\mu, \sigma_1^2\big),\\
&\begin{pmatrix} x_{i_1} \\ x_{i_2} \end{pmatrix} \Big| x_{i_0}, M \in A(x_{i_1}, x_{i_2})\\
&\qquad \sim N\left(\begin{pmatrix} x_{i_0} + \beta t_{i_1} \\ x_{i_0} + \beta t_{i_2} \end{pmatrix}, \begin{pmatrix} \sigma_2^2 & \rho\sigma_2^2 \\ \rho\sigma_2^2 & \sigma_2^2 \end{pmatrix}\right).
\end{aligned} \tag{1}
$$

The above change-point-in-tree model contains one unknown change point M and five unknown parameters. We will use a Bayesian approach to estimate them from a data file. To facilitate Bayesian computing, we use conjugate prior distributions for unknown parameters. Detail prior distributions are as follows:

$$
\begin{aligned}
\sigma_1^2 &\sim \Gamma^{-1}(g, h),\\
\sigma_2^2 &\sim \Gamma^{-1}(a, b),\\
\beta &\sim N(r, s),\\
\mu &\sim N(p, q),\\
\rho &\sim \text{Beta}(u, v),\\
M &\sim \text{Uniform over all cells in the candidate set.}
\end{aligned}
$$

Settings of the hyperparameters in the above prior distributions and the sensitivity analysis are listed in Supplement B. The candidate set contains all cells on the cell lineage tree which may initialize a gene's expression. Here we exclude boundary cells of the tree from the candidate set because an expression pattern changing at the boundary is either false positive or a signal too weak to be significantly different from the background. More specifically, considering the situation of the *C. elegans* embryo, only cells whose number of descendants is between 6 and 30 are put in the candidate set, while for a large expression branch, the DEGEO algorithm will divide it into several small expression branches and detect them one by one. The joint posterior distribution is as follows:

$$
\begin{aligned}
&f\big(\sigma_1^2, \sigma_2^2, \beta, \mu, \rho, M | X, T\big)\\
&\quad \propto f\big(\sigma_1^2, \sigma_2^2, \beta, \mu, \rho, M\big) \cdot f\big(X, T | \sigma_1^2, \sigma_2^2, \beta, \mu, \rho, M\big)\\
&\quad \propto \big(\sigma_1^2\big)^{-g-1} e^{-h/\sigma_1^2} \cdot \big(\sigma_2^2\big)^{-a-1} e^{-b/\sigma_2^2} \cdot e^{-(\beta-r)^2/(2s)}\\
&\qquad \times e^{-(\mu-p)^2/(2q)} \cdot \rho^{u-1}(1-\rho)^{v-1}\\
&\qquad \times \frac{1}{\sigma_1^{|N_M|}} e^{-(1/(2\sigma_1^2))\sum_{N_M}(x_i-\mu)^2} \cdot \frac{1}{(\sqrt{1-\rho^2}\sigma_2^2)^{|\overline{N}_M|}} e^{-J/(2(1-\rho^2)\sigma_2^2)}.
\end{aligned}
$$

The conditional posterior distributions of all parameters can then be deduced as follows:

$$\sigma_1^2|\sigma_2^2,\beta,\mu,\rho,M,X,T \sim \Gamma^{-1}\Big(g+\frac{|N_M|}{2}, h+\frac{1}{2}\sum_{N_M}(x_i-\mu)^2\Big),$$

$$\sigma_2^2|\sigma_1^2,\beta,\mu,\rho,M,X,T \sim \Gamma^{-1}\Big(a+|\overline{N}_M|, b+\frac{J}{2(1-\rho^2)}\Big),$$

$$\beta|\sigma_1^2,\sigma_2^2,\mu,\rho,M,X,T \sim N\Big(\frac{K}{1/s+(1/((1-\rho^2)\sigma_2^2))\sum_{\overline{N}_M}(t_{i_1}^2+t_{i_2}^2-2\rho t_{i_1}t_{i_2})}, \frac{1}{1/s+(1/((1-\rho^2)\sigma_2^2))\sum_{\overline{N}_M}(t_{i_1}^2+t_{i_2}^2-2\rho t_{i_1}t_{i_2})}\Big),$$

$$\mu|\sigma_1^2,\sigma_2^2,\beta,\rho,M,X,T \sim N\Big(\frac{p/q+\sum_{N_M}x_i/\sigma_1^2}{1/q+|N_M|/\sigma_1^2}, \frac{1}{1/q+|N_M|/\sigma_1^2}\Big),$$

$$\rho|\sigma_1^2,\sigma_2^2,\beta,\mu,M,X,T \propto \rho^{u-1}(1-\rho)^{v-1}\Big(\frac{1}{\sqrt{1-\rho^2\sigma_2^2}}\Big)^{|\overline{N}_M|}e^{-J/(2(1-\rho^2)\sigma_2^2)},$$

$$M|\sigma_1^2,\sigma_2^2,\beta,\mu,\rho,X,T \propto \Big(\frac{1}{\sigma_1}\Big)^{|N_M|}e^{-1/(2\sigma_1^2)\sum_{N_M}(x_i-\mu)^2}\Big(\frac{1}{\sqrt{1-\rho^2\sigma_2^2}}\Big)^{|\overline{N}_M|}e^{-J/(2(1-\rho^2)\sigma_2^2)},$$

where

$$K=\frac{1}{(1-\rho^2)\sigma_2^2}\sum_{\overline{N}_M}\big[(t_{i_1}-\rho t_{i_2})(x_{i_1}-x_{i_0})+(t_{i_2}-\rho t_{i_1})(x_{i_2}-x_{i_0})\big]+\frac{r}{s},$$

$$J=\sum_{\overline{N}_M}\big[(x_{i_1}-x_{i_0}-\beta t_{i_1})^2+(x_{i_2}-x_{i_0}-\beta t_{i_2})^2 - 2\rho(x_{i_1}-x_{i_0}-\beta t_{i_1})(x_{i_2}-x_{i_0}-\beta t_{i_2})\big],$$

N_M : set of cells inside the expression branch with change point M,

$\overline{N}_M$: set of cells outside the expression branch with change point M.

To fit the change-point-in-tree model in equation (1) to a tree of cell scores, we use an MCMC algorithm, which iteratively updates each parameter from its conditional posterior distribution until converging, as judged by the potential scale reduction factor (or $\hat{R}$) [Gelman and Rubin (1992)]. More specifically, an MCMC chain is said to have converged if $|\hat{R}-1|<0.2$ holds for all parameters. As shown

in Supplement D, this MCMC algorithm converges fast. The output of the algorithm is a sample from the converged joint posterior distribution of the change point and all parameters, from which we can get both the point estimates and the uncertainty measures of all parameters. More specifically, we regard the change point value with the highest posterior probability M^* as the MCMC detected branch, and the conditional posterior mean values (conditional on the reported $M = M^*$) of other parameters as fitted parameters.

2.3.3. *Step* 3: *Use SVR to classify an MCMC detected branch*. The above MCMC algorithm forces the fitting of a tree of cell scores to the model in equation (1), which assumes a single expression branch. Since a target gene may express in zero or multiple branches in the embryo, the detected expression branch may be false positive or a nonunique true positive. To deal with this issue, we feed some features of an MCMC detected branch to a trained SVR to further decide whether we shall report the corresponding branch as expressing. SVR is a version of Support Vector Machine and has been proposed by Harris et al. (1997). The used features are provided in Supplement C. The training of SVR is explained in Section 3.1.1.

If the trained SVR classifies the MCMC detected branch as expressing, we delete the branch from the tree and run the MCMC algorithm to fit the change-point-in-tree model again. This procedure is iterated until an MCMC detected branch is classified as nonexpression. That is, the trained SVR serves as a criterion to stop searching more expression branches from the tree of cell scores.

2.3.4. *Step* 4: *Find the detailed onset time within a cell*. Steps 2 and 3 report expression at the level of branches in the tree with cell scores x_i. Some cells in the SVR reported branches may not express the target gene. For the cell where the target gene initiates expressing, the detailed onset time may not be the birth time of the cell. In step 4, we will further detect expression onset cells and corresponding times, that is, which data point y_{ij}.

For each 4D confocal movie data file, we make use of the sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ of the valid data points except those belonging to expression branches detected at step 3, which provide the most accurate estimation of background noise in case the tree contains multiple expression branches. Valid data points which are greater than the 97.5% quantile of the noise distribution $N(\hat{\mu}, \hat{\sigma}^2)$ are regarded as extreme values. We then search along all paths from the change point M to all leaf cells inside the SVR reported branch to identify all expression segments which satisfy the following: (a) the time series segment (a block of neighboring data points) along the path contains at least 10 valid data points; and (b) at least 97.5% percent of the valid data points in the segment are extreme values. We define a valid data point as an expression onset if it is the earliest expression segment time point on a path from the change point M to a leaf cell inside a SVR reported branch, and define a valid data point as an expression end if it is the last one on a path.

3. Results and discussion. We use synthesized data, where the background truth is known, to train SVR and test the performance of our method and compare it with that of Murray et al. (2012). We also apply our method on a real data set.

3.1. *Synthesized data.* Three synthesized data sets are generated for simulation studies.

3.1.1. *Synthesized data set* 1. The first data set is synthesized to mimic the real data. To create a mimic tree of cell scores, we first randomly pick one well-annotated real tree of cell scores whose expression branches have been reliably labeled, then use the cell scores of its nonexpression cells as the background noise distribution to generate a whole tree of noise cell scores, and finally replace a random set of branches with real expression branches with the same branch structures. The above mimicking procedure is repeated to generate 120 trees of cell scores in synthesized data set 1. Each of the mimic trees of cell scores shares the same noise and expression cell score distribution as a real data file, and we know which cells are really expressing.

The MCMC algorithm in step 2 is run on each of the 120 trees. Once it converges, the detected branch is deleted and the MCMC algorithm is run again on the remaining tree, until the MCMC detected branch no longer contains any really expression cell. In 116 of the 120 trees, the MCMC algorithm precisely detects all true expression branches first before finally detecting a nonexpression branch as expressing. It shows that the MCMC algorithm alone can accurately detect expression branches from this mimic data set.

By repeatedly running the above MCMC algorithm on the 120 mimic trees, the detected branches contain many true expression branches (code as output $= 1$ for SVR) and some false expression branches (code as output $= 0$ for SVR). Since we know the true expression status of these MCMC detected branches, we use selected features (see Supplement C) of these branches as the training data set to fit a SVR classifier. Figure 3 shows the fitted output values of all branches in the training data set. The true expression branches and false expression ones can be fairly separated by a threshold for the SVR output value. As shown in Supplement C, the best threshold is 0.15 because its mean false classification rate is minimized in this training data set.

To test the accuracy of the trained SVR on independent data, we synthesize another set of 120 mimic trees and run the MCMC algorithm using the above same procedure. The trained SVR is used to classify the MCMC detected branches. Table 1 shows the results, where the false classifications are grouped by the number of true expression branches in the corresponding trees. Detailed figures are provided in Supplement C and the mean rate of false classifications is minimized when the threshold is 0.15, which agrees with the threshold from the training data. As we can see, the trained SVR performs very well on this test data.

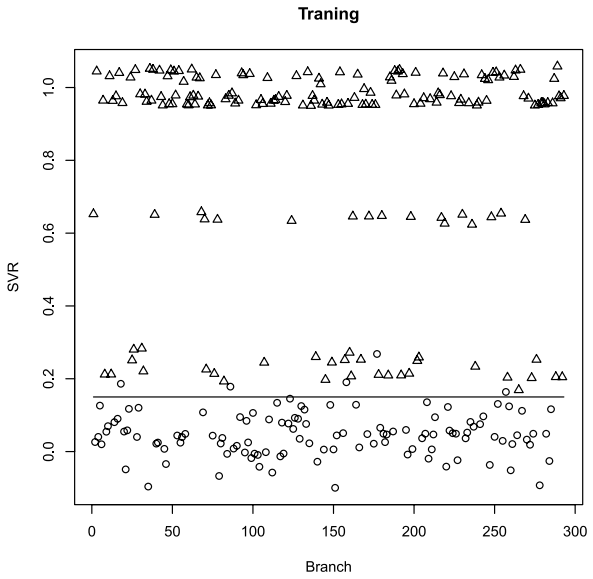


FIG. 3. *Fitted SVR output values of the training branches. Each point represents a branch, where the horizontal coordinate shows the branch index and the vertical coordinate shows the SVR output value. True expression branches are denoted as circles, while false ones are denoted as triangles. The horizontal line shows the classification boundary with SVR output threshold at* 0.15.

3.1.2. *Synthesized data set* 2. The second synthesized data set is generated from the model in equation (1), therefore, it fully satisfies all model assumptions. Using a true data file's lifetime and tree structure as template, we first randomly select 0 to 10 cells as the roots of expression branches, then generate true values of parameters by sampling from their prior distributions, and finally generate all cell scores according to equation (1). This procedure is repeated to generate 110 synthesized trees, with 10 trees for each of the 11 kinds of expression branch numbers.

TABLE 1
No. of misclassified branches when applying the trained SVR on MCMC detected branches from testing mimic trees

No. of true expression branches contained in corresponding trees	**None**	**One**	**Two**	**Three**	**Four**	**Overall**
No. of MCMC detected branches	30	56	81	80	40	287
SVR false positive	1	0	0	1	0	2
SVR false negative	0	0	0	0	0	0
SVR false classifications	1	0	0	1	0	2

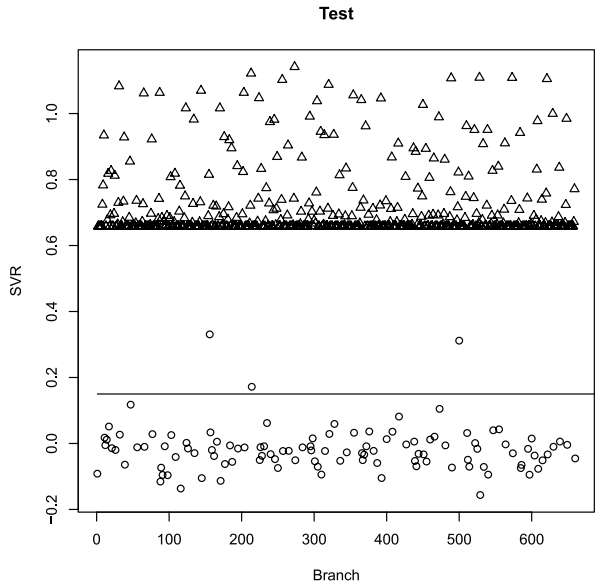


FIG. 4. *Predicted SVR output values for MCMC detected branches from the second synthesized data set. Triangles indicate true expression branches, while circles indicate false. The horizontal line shows the SVR threshold* 0.15.

For each of the synthesized trees, we run the MCMC algorithm and use the trained SVR from Section 3.1.1 to decide when to stop as described in Sections 2.3.2 and 2.3.3. When an MCMC detected branch is classified as nonexpression by the trained SVR, the tree is no further fitted by the MCMC algorithm. Figure 4 shows the SVR output values of all SVR reported branches (triangles above the horizontal threshold line) and the MCMC detected branches which are classified as nonexpression by the trained SVR (circles below the horizontal threshold line). It shows that all true expression branches are correctly reported. Only three false branches are reported by the trained SVR.

3.1.3. *Synthesized data set* 3. The third data set is used to compare the performance of the DEGEO algorithm and the method proposed by Murray et al. (2012) (denoted as APM) in detecting expression onset cells. The data set is synthesized by mimicking the trees of original data files. More specifically, we first pick one well-annotated real tree whose expression onset cells have been reliably labeled, then use the data points y_{ij} of its nonexpression branches as the background noise distribution to generate a whole tree of noise data points, and then insert the data points y_{ij} of a random set of real expression branches. The above procedure is repeated to generate 120 trees for the synthesized data set 3. Each of the mimic trees shares the same noise and expression distribution as a real data file, and we know which cells are expression onsets. Table 2 shows True Positive Rate (TPR), False

TABLE 2
Performance comparison between DEGEO and APM. Standard errors for APM proportions are approximately 0.034–0.054 *for TPR*, 0.001–0.002 *for FPR*, 0.006–0.022 *for PPV*

	TPR		FPR		PPV	
No. of expression branches	**DEGEO**	**APM**	**DEGEO**	**APM**	**DEGEO**	**APM**
0	–	–	0	0.058	–	0
1	1	0.500	0	0.048	1	0.040
2	1	0.528	0	0.047	1	0.086
3	1	0.602	0	0.051	1	0.140
4	1	0.567	0	0.035	1	0.206

Positive Rate (FPR) and Positive Predictive Value (PPV) of the two methods at the cell level. The estimated probabilities under DEGEO have much higher accuracies than those of Murray et al. (2012).

3.2. *Real data*. For the real promoter fusion data from Murray et al. (2012), the four-step DEGEO procedure in Section 2 is used to find all expression onset time points.

To evaluate the performance of our method on the real data, we compile a benchmark real data set by manually annotating expression branches on 20 real data files. For a comparison with the SVR stopping criterion on detecting the expression branches, we also test another intuitive stopping criterion based on the parameter β. More specifically, we stop further MCMC searching if the β of a new MCMC detected branch is less than one third of the mean values of β's of all previously detected branches. Table 3 compares the performances of the two stopping criteria on the benchmark real data set in terms of detecting expression branches. It shows that SVR is far better that the β-based stopping criterion, because it reports most of the true expression branches with an acceptable false negative rate. Detailed results on the benchmark set are provided in Supplement C.

We run DEGEO for each of the real promoter fusion data files from Murray et al. (2012). The detailed results for each data file are provided in Supplement F, which includes all SVR reported expression branches, all exact expression onset

TABLE 3
Performances of the 2 *stopping criteria on the benchmark real data set in terms of the number of wrongly or correctly reported branches*

Stopping criterion	**False positive**	**True positive**	**False negative**
β-based	31	143	69
SVR(0.15)	15	185	27

TABLE 4

Comparison of expression onset estimation with current literatures and Murray et al. (2008). The "Onset" columns list the embryonic stage (in terms of the number of cells at the onset time) and the cell [named according to the nomenclature in Sulston et al. (1983)] containing the expression onset. The "Expression" columns show which tissues are expressing the target gene. The x in cell names works as a wildcard character. The cited papers in the third column provide the source of the information

	Onset (cells stage)			Expression (cells)		
Gene	**Literature**	**DEGEO**	**ROM**	**Literature**	**DEGEO**	**ROM**
end-3	28	26–28 Ex	<200 Ex	Intestine [Maduroa et al. (2005)]	16 intestine	20 intestine
hlh-1	90+ C	133–140, 161–171 Cxpx	90 Cxpx	Muscle precursors [Krause (1995)]	16 muscle and 1 ganglion	Muscle
		178–190 Dxx	180 Dxx		8 muscle cells	Muscle
	12–24 MS	51–87 Msxxx	24 MSxx	Transiently in MS [Krause (1995)]	22 muscle, 3 ganglion, 2 coelomocyte 1 mesoderm and 21 pharynx	Muscle and pharynx
isw-1	Every stage	87–350	350	Most [Andersen, Lu and Horvitz (2006)]	Most	Most but not all
tbx-38	24 ABaxxx	45–51 ABaxxxx		Descendants of ABa [Good et al. (2004)]	9 connective tissue, 27 hypodermis, 97 nerve tissue, 48 pharynx	ABa descendents

time points and all expression segments. Two expression segments are merged if they are separated by no more than two data points.

DEGEO reports no false positive from the 6 negative control data files, indicating the good specificity of our method. For other data files, we try to compare our results with available results in current literatures. Table 4 shows our results of several genes together with supporting evidences in current literatures and results obtained by Murray et al. (2008) (denoted by ROM), which used an ad hoc threshold to report the expression onset among all the data sets. It shows that the onset reported in current literatures and Murray et al. (2008) are also detected by DEGEO, but DEGEO detects more exact onset times and more onset locations. Note that Krause (1995) detected disparate onset times in various expression branches of gene *hlh*-1, which suggests that iterative runs of step 2 and estimating exact onset for different paths in step 4 are necessary.

The real data results also show that DEGEO has the capability to handle the case where the tree contains no expression branch or more than one expression branch,

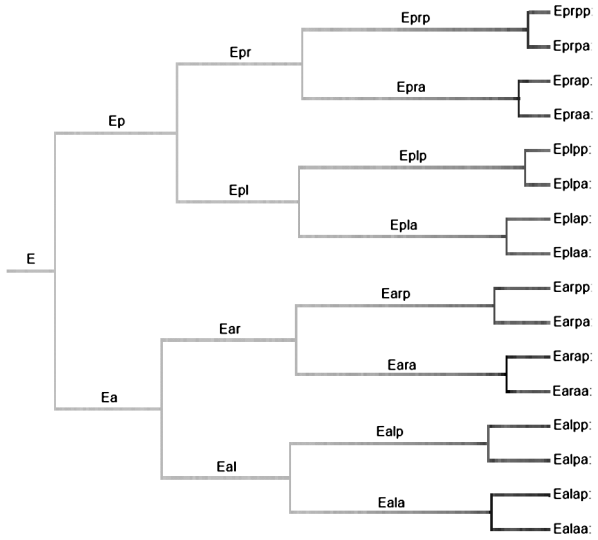


FIG. 5. *The E branch detected from CD*20070319*_pha*4*_I*1*LBBB.csv. The expression of the target gene is found in every path of this expression branch, and the expression increases faster to high values.*

although the change-point-in-tree model assumes that the tree contains exactly one expression branch. DEGEO finds the expression branches one by one, and tends to first detect the more outstanding expression branch, which contains more cells and whose expression grows faster to high values, with a bigger SVR output value. Using the data file *CD*20070319*_pha*4*_I*1*LBBB.csv* as an example, the *E* branch shown in Figure 5 is detected with a SVR output value of 0.87 before the *ABarapa* branch is detected with a SVR output value of 0.33. This tendency is also shown in Supplement D, where branches with bigger β values are detected earlier.

After the expression branches are detected, DEGEO moves to estimate the exact expression onset time. For example, DEGEO reports the cells *Exxx* and *ABarapaxx* as expression onset in Figures 5 and 6, respectively. Here the *x* in cell names works as a wildcard character.

DEGEO also seems resistant to the false expression phenomenon which may result from noise fluctuation or fluorescent absorption. For example, as shown in Figure 7, the *MSap* branch from *CD*20060627*_cnd*1_4-2.*csv* and the *ABpraapp* branch from *CD*20080128*_elt*-1_3.*csv* are correctly classified as false expression branches by DEGEO.

DEGEO may perform poorly if almost all paths in a tree show increasing trends, such as in *CD*20080504*_C*01*B*7_1_6.*csv*. This is because DEGEO assumes all cells outside the selected expression branch follow a normal distribution, which is invalid if most of these cells are actually from expression branches. As a result, the MCMC algorithm will report many expression branches, but may report relatively weaker expression branches earlier before stronger expression branches. In

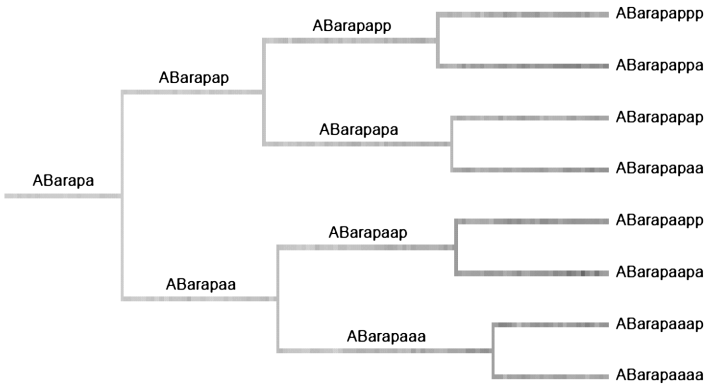
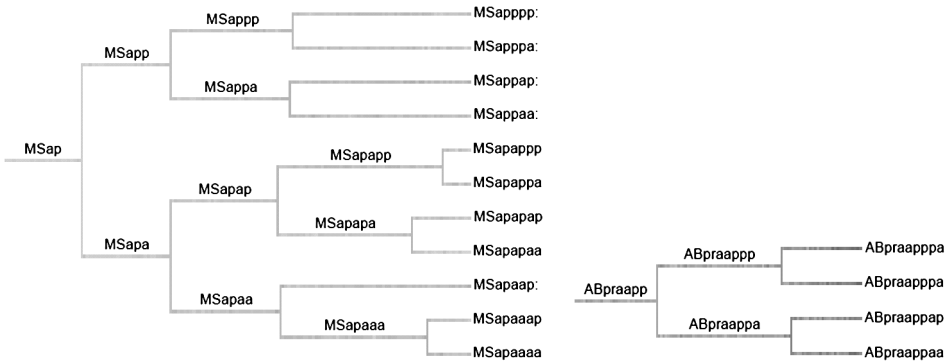


FIG. 6. *The ABarapa branch detected from CD*20070319*_pha*4*_I*1*LBBB.csv. The expression of the target gene is found in every path of this expression branch, but the expression only increases slowly to relatively low values.*

this case, we can actually use step 4 of the DEGEO procedure to directly detect expression onset on each path without the need to sort out expression branches in steps 2 and 3.

4. Conclusion. We provide a principled automatic procedure to detect expression onset from 4D confocal data of *C. elegans* embryos. Both simulation studies and real data examples show that our method can detect both fast and slow expression lineages. On the other hand, it efficiently excludes false posi-



(A) Msap branch (B) ABpraapp branch

FIG. 7. *Examples of false expression phenomenon.* (A) *The MSap branch from CD*20060627*_cnd*1_4-2*.csv. Its values show a weak uptrend which probably results from fluorescent absorption. This branch is classified as a false expression branch by the trained SVR*; (B) *The ABpraapp branch from CD*20080128*_elt*-1_3*.csv. Its values show a weak uptrend which probably results from noise fluctuation. This branch is classified as a false expression branch by the trained SVR.*

tive ones. Along the paths of detected expression lineages, we detect exact onset times of the target gene's expression. Meanwhile, we are able to estimate the parameters of data files, such as expression rate and distribution of background noise.

In general, our algorithm can handle most cases well except for the case where a gene is expressed in almost all cells, because this case does not fit our model assumption. Extending our model for multiple change points is a natural choice, but the unknown number of change points may make the problem computationally very hard. In this paper, we stick to the assumption of one change point and test its detection power on the data with multiple change points. For cases when the gene is not widely expressed, DEGEO can accurately detect all change points one by one, while for broadly expressed genes, we come up with a solution by constructing the background noise distribution from early expression values instead of all values outside the selected expression branch.

Except for the embryonic gene onset problem on the cell lineage tree, our algorithm can also be applied on other change point problems as long as the data points form a known tree structure. For example, the information flow on social networks may form such a lineage tree, thus our algorithm can be used to detect information change points, such as sentiment formation and propagation [Liben-Nowell and Kleinberg (2008)]. The propagation of contagious disease may also form a lineage tree, and we can detect the virus mutation on the lineage.

APPENDIX: DETAILED DESCRIPTION OF THE REAL DATA

One fundamental question in biology is how a zygote develops into an embryo with different tissues. To approach this question, large-scale 4D confocal movies of *C. elegans* embryos have been produced by experimental biologists. The first objective is to detect when and where a gene is expressed in an embryo. Our real data files are obtained by automated analysis of reporter gene expression in *C. elegans* with cellular resolution during embryogenesis [Murray et al. (2012)]. Basically, an embryo is measured once per minute to report simultaneously the fluorescence intensity of each cell which is living in the embryo at that time. Each real data file can be viewed as a binary tree, where each node is a cell represented by a time series and each branch indicates a parent–child relationship during cell division. Since the cell lineage is invariant for all *C. elegans* embryos, the binary trees from different data files have the same topology. But a cell's lifetime may vary across different embryos. Overall, the real data set contains 201 real data files. 5 of them are negative control files, and each of the remaining files measures an individual gene's expression during embryogenesis. In total, 111 genes are measured, and 51 genes are measured by replicated embryos. The 25% quantile, mean, median, 75% quantile and standard deviation of the distribution of all cell lifetimes are 20, 28.55, 27, 35 and 12.85 minutes, respectively. Some characteristics of the real data

TABLE 5
Distributions of some statistics across the 201 *real data files*

Distribution summary	25% quantile	Mean	Median	75% quantile	Standard deviation
No. of data points	14,369	17,210.4	16,425	20,032	4424.4
Observation time (min)	144	173	159	199	39
No. of observed cells	708	697.7	713	726	101.5
Mean fluorescent intensity	516	7260.2	1834.3	6444.4	13,328.2
SD of fluorescent intensity	1714.4	9207	4780.3	12,201.2	10,813.5

files are summarized in Table 5. For more details about the experiment and the data, please refer to Bao et al. (2006) and Murray et al. (2008, 2012).

Acknowledgment. We thank two anonymous reviewers, the Associate Editor and the area Editor for their very helpful comments. Supplemental materials are available online and the R code for the DEGEO algorithm is available upon request.

SUPPLEMENTARY MATERIAL

Supplement A: Model checking (DOI: 10.1214/15-AOAS820SUPPA; .pdf). We provide the justification of our 3 model assumptions in Section 2.2.

Supplement B: Hyperparameters of prior distributions (DOI: 10.1214/15-AOAS820SUPPB; .pdf). The settings and the sensitivity analysis of hyperparameters are shown in detail.

Supplement C: Classification and stopping criterion based on SVR (DOI: 10.1214/15-AOAS820SUPPC; .pdf). We provide plots and tables to demonstrate the good performance of the SVR method in classifying expression and nonexpression branches.

Supplement D: Convergence diagnosis and parameter estimation (DOI: 10.1214/15-AOAS820SUPPD; .pdf). Proofs of successful convergence and good parameter estimation are provided in additional figures and table.

Supplement E: Detection of size-biased sampling (DOI: 10.1214/15-AOAS820SUPPE; .pdf). We supply some details in detection of the size-biased sampling problem.

Supplement F: Detection results of real data files (DOI: 10.1214/15-AOAS820SUPPF; .zip). All SVR reported expression branches, all exact expression onset time points and all expression segments in real data files are listed in a folder.

REFERENCES

ANDERSEN, E. C., LU, X. and HORVITZ, H. R. (2006). *C. elegans* ISWI and NURF301 antagonize an Rb-like pathway in the determination of multiple cell fates. *Development* **133** 2695–2704.

BAO, Z., MURRAY, J. I., BOYLE, T., OOI, S. L., SANDEL, M. J. and WATERSTON, R. H. (2006). Automated cell lineage tracing in *caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **103** 2707–2712.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GOOD, K., CIOSK, R., NANCE, J., NEVES, A., HILL, R. J. and PRIESS, J. R. (2004). The t-box transcription factors tbx-37 and tbx-38 link glp-1/notch signaling to mesoderm induction in *C. elegans* embryos. *Development* **131** 1967–1968.

GURALNIK, V. and SRIVASTAVA, J. (1999). Event detection from time series data. In *KDD'99 Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **17** 33–42. ACM, San Diego, CA.

HARRIS, D., BURGES, J. C. C., KAUFMAN, L., SMOLA, J. A. and VLADIMIR, N. V. (1997). Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **9** 155–161.

HU, J., ZHAO, Z., YALAMANCHILI, H., WANG J., YE, K. and FAN, X. (2015). Supplement to "Bayesian detection of embryonic gene expression onset in *C. elegans*." DOI:10.1214/15-AOAS820SUPPA, DOI:10.1214/15-AOAS820SUPPB, DOI:10.1214/15-AOAS820SUPPC, DOI:10.1214/15-AOAS820SUPPD, DOI:10.1214/15-AOAS820SUPPE, DOI:10.1214/15-AOAS820SUPPF.

KRAUSE, M. (1995). Myod and myogenesis in *C. elegans*. *BioEssays* **17** 228.

LIBEN-NOWELL, D. and KLEINBERG, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA* **105** 4633–4638.

LIU, X., LONG, F., PENG, H., AERNI, S. J., JIANG, M., BLANCO, A. S., MURRAY, J. I., PRESTON, E., MERICLE, B., BATZOGLOU, S., MYERS, E. W. and KIM, S. K. (2009). Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. *Cell* **139** 623–633.

LONG, F., PENG, H., LIU, X., KIM, S. K. and MYERS, E. (2009). A 3D digital atlas of *C. elegans* and its application to single-cell analyses. *Nat. Methods* **6** 667–672.

MADUROA, M. F., HILLB, R. J., HEIDC, P. J., SMITHA, E. D. N., ZHU, J., PRIESS, J. R. and ROTHMAN, J. H. (2005). Genetic redundancy in endoderm specification within the genus caenorhabditis. *Dev. Biol.* **284** 522.

MURRAY, J. I., BAO, Z., BOYLE, T. J., BOECK, M. E., MERICLE, B. L., NICHOLAS, T. J., ZHAO, Z., SANDEL, M. J. and WATERSTON, R. H. (2008). Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nature Methods* **5** 703–709.

MURRAY, J. I., BOYLE, T. J., PRESTON, E., VAFEADOS, D., MERICLE, B., WEISDEPP, P., ZHAO, Z., BAO, Z., BOECK, M. and WATERSTON, R. H. (2012). Multidimensional regulation of gene expression in the *C. elegans* embryo. *Genome Research* **22** 1282–1294.

PERREAULT, L., BERNIER, J., BOBEE, B. and PARENT, E. (2000). Bayesian change-point analysis in hydrometeorological time series. *Journal of Hydrology* **235** 221–241.

PICARD, D. (1985). Testing and estimating change-points in time series. *Adv. in Appl. Probab.* **17** 841–867. MR0809433

SPENCER, W. C., ZELLER, G., WATSON, J. D., HENZ, S. R., WATKINS, K. L., MCWHIRTER, R. D., PETERSEN, S., SREEDHARAN, V. T., WIDMER, C., JO, J., REINKE, V., PETRELLA, L., STROME, S., STETINA, S. E. V., KATZ, M., SHAHAM, S., RATSCH, G. and MILLER, D. M. (2011). A spatial and temporal map of *C. elegans* gene expression. *Genome Research* **21** 325–341.

SULSTON, J. E., SCHIERENBERG, E., WHITE, J. G. and THOMSON, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100** 64–119.

YALAMANCHILI, H. K., YAN, B., LI, M. J., QIN, J., ZHAO, Z., CHIN, F. Y. and WANG, J. (2013). Dynamic delay gene network inference from high temporal data using gapped local alignment. *Bioinformatics* **30** 377–383.

J. HU
X. FAN
DEPARTMENT OF STATISTICS
CHINESE UNIVERSITY OF HONG KONG
SHATIN
HONG KONG
E-MAIL: hujiechelsea@gmail.com
xfan@sta.cuhk.edu.hk
URL: http://www.sta.cuhk.edu.hk/xfan/

H. K. YALAMANCHILI
J. WANG
DEPARTMENT OF BIOCHEMISTRY
UNIVERSITY OF HONG KONG
POKFULAM
HONG KONG
E-MAIL: hari@hku.hk
junwen@hku.hk

Z. ZHAO
DEPARTMENT OF BIOLOGY
HONG KONG BAPTIST UNIVERSITY
KOWLOON TONG
HONG KONG
E-MAIL: zyzhao@hkbu.edu.hk

K. YE
DEPARTMENT OF EPIDEMIOLOGY
AND POPULATION HEALTH
ALBERT EINSTEIN COLLEGE OF MEDICINE
1300 MORRIS PARK AVENUE
BRONX, NEW YORK 10461
USA
E-MAIL: kenny.ye@einstein.yu.edu