A MARKOV RANDOM FIELD-BASED APPROACH TO CHARACTERIZING HUMAN BRAIN DEVELOPMENT USING SPATIAL–TEMPORAL TRANSCRIPTOME DATA[1]

By Zhixiang Lin[*], Stephan J. Sanders[†], Mingfeng Li[*],
Nenad Sestan[*], Matthew W. State[†] and Hongyu Zhao[*]

Yale University[*] and University of California, San Francisco[†]

Human neurodevelopment is a highly regulated biological process. In this article, we study the dynamic changes of neurodevelopment through the analysis of human brain microarray data, sampled from 16 brain regions in 15 time periods of neurodevelopment. We develop a two-step inferential procedure to identify expressed and unexpressed genes and to detect differentially expressed genes between adjacent time periods. Markov Random Field (MRF) models are used to efficiently utilize the information embedded in brain region similarity and temporal dependency in our approach. We develop and implement a Monte Carlo expectation–maximization (MCEM) algorithm to estimate the model parameters. Simulation studies suggest that our approach achieves lower misclassification error and potential gain in power compared with models not incorporating spatial similarity and temporal dependency.

1. Introduction. Human neurodevelopment is a dynamic and highly regulated biological process. Abnormalities in neurodevelopment may lead to psychiatric and neurological disorders, such as Autism Spectrum Disorders (ASD) [Geschwind and Levitt (2007), Sestan et al. (2012), Walsh, Morrow and Rubenstein (2008)]. The statistical methodology developed in this paper was motivated by our interest in studying human brain development using a microarray gene expression data set, which was collected from 1340 tissue samples of 57 developing and adult post-mortem brains (including 39 with both hemispheres) [Johnson et al. (2009), Kang et al. (2011)]. These 57 post-mortem brains spanned from embryonic development to late adulthood. A 15-period system, demonstrated in Table 1, was defined to represent distinct stages of brain development [Johnson et al. (2009), Kang et al. (2011)]. Except for periods 1 and 2, tissue samples from 16 brain regions were collected from both hemispheres in each brain, including the cerebellar cortex (CBC), mediodorsal nucleus of the thalamus (MD), striatum (STR),

TABLE 1
The 15-*period system in* Kang et al. *(*2011*). M, postnatal months*;
PCW, post-conceptional weeks; Y, postnatal years

Period	Description	Age
1	Embryonic	4 PCW ≤ Age < 8 PCW
2	Early fetal	8 PCW ≤ Age < 10 PCW
3	Early fetal	10 PCW ≤ Age < 13 PCW
4	Early mid-fetal	13 PCW ≤ Age < 16 PCW
5	Early mid-fetal	16 PCW ≤ Age < 19 PCW
6	Late mid-fetal	19 PCW ≤ Age < 24 PCW
7	Late fetal	24 PCW ≤ Age < 38 PCW
8	Neonatal and early infancy	0 M (birth) ≤ Age < 6 M
9	Late infancy	6 M ≤ Age < 12 M
10	Early childhood	1 Y ≤ Age < 6 Y
11	Middle and late childhood	6 Y ≤ Age < 12 Y
12	Adolescence	12 Y ≤ Age < 20 Y
13	Young adulthood	20 Y ≤ Age < 40 Y
14	Middle adulthood	40 Y ≤ Age < 60 Y
15	Late adulthood	Age ≥ 60 Y

amygdala (AMY), hippocampus (HIP) and 11 areas of the neocortex, including the orbital prefrontal cortex (OFC), dorsolateral prefrontal cortex (DFC), ventrolateral prefrontal cortex (VFC), medial prefrontal cortex (MFC), primary motor cortex (M1C), primary somatosensory cortex (S1C), posterior inferior parietal cortex (IPC), primary auditory cortex (A1C), posterior superior temporal cortex (STC), inferior temporal cortex (ITC) and the primary visual cortex (V1C) [Johnson et al. (2009), Kang et al. (2011)]. Details on the brain regions are described in the supplementary material Section 1 [Lin et al. (2015)].

The goal of our analysis is to characterize human neurodevelopment through the dynamics of gene expression, such as the identification of expressed and unexpressed genes, and differentially expressed (DE) genes over time in each brain region. The unique challenge presented for statistical analysis of this data set is the appropriate modeling and analysis of the spatial–temporal structure. For gene expression data with only temporal structure (e.g., time course gene expression data), various methods have been proposed to model the temporal dependency to better identify DE genes. However, as far as we know, none of the existing methods utilizes the information embedded in the spatial similarity between brain regions, as indicated by the high correlation in gene expression levels between brain regions in the same period [supplementary material Section 2, Lin et al. (2015) and Kang et al. (2011)]. For time course gene expression data, the existing methods can be classified into two broad categories: (1) methods that identify DE genes between multiple biological conditions [Hong and Li (2006), Storey et al. (2005), Tai and

Speed (2006), Yuan and Kendziorski (2006)]; and (2) methods that identify DE genes over time in one biological condition [Liu and Yang (2009), Storey et al. (2005), Tai and Speed (2006), Wu et al. (2007)]. Statistical models that have been proposed to incorporate the temporal structure include Hidden Markov Models [Wu et al. (2007), Yuan and Kendziorski (2006)], functional models using basis function expansions [Hong and Li (2006), Storey et al. (2005), Wu et al. (2007)], function principal component analysis [Liu and Yang (2009)] and multivariate empirical Bayes models [Tai and Speed (2006)].

To efficiently capitalize on brain region similarity and temporal dependency, we propose a two-step Markov Random Field (MRF)-based approach to answer the following two biological questions: 1. Which genes are expressed/unexpressed in each period and in each brain region? 2. Which genes are differentially expressed over time in each brain region? We note that MRF models have been used to model dependency in genomics data, such as neighboring genes defined by biological pathways [Chen, Cho and Zhao (2011), Li, Wei and Li (2010), Wei and Li (2007, 2008)] and marker dependencies defined by linkage disequilibrium [Li, Wei and Maris (2010)]. Across all the brain regions and time periods, the histogram of the observed gene expression levels has a bimodal distribution, where the two components likely represent expressed and unexpressed genes [supplementary material Section 4, Lin et al. (2015) and Kang et al. (2011)]. In this paper, we first use a Gaussian mixture model-based approach to identify the unexpressed and expressed genes. The model fit and the robustness of the Gaussian mixture model are discussed in the supplementary material Section 4 [Lin et al. (2015)]. We note that an "unexpressed" gene does not necessarily suggest that there is no mRNA molecules of that gene in the cell, but rather the gene's expression level is very low and the observed variation in the expression values may be mostly due to noise in the microarray experiment. In the second step, our methodology utilizes the local false discovery rate (f.d.r.) framework [Efron (2004)] to identify DE genes between adjacent time periods. We propose an efficient Monte Carlo expectation–maximization (MCEM) algorithm [Wei and Tanner (1990)] to estimate the model parameters and a Gibbs sampler to estimate the posterior probabilities.

The key feature of our approach is to simultaneously consider spatial similarity and temporal dependency of gene expression levels to better extract biologically meaningful results from the data. We introduce the MRF model in Section 2 and present the Monte Carlo expectation–maximization (MCEM) algorithm for statistical inference in Section 3. We also present the posterior probability estimation and the FDR controlling procedure in Section 3. In Section 4 we apply our method to analyze the human brain microarray data reported in Kang et al. (2011). Results from simulation studies are summarized in Section 5. We conclude the paper with a brief discussion in Section 6.

2. Statistical models and methods.

2.1. *Biological question* 1: *Identify expressed and unexpressed genes*.

2.1.1. *Gaussian mixture model for microarray data.*   In our human brain microarray data, expression levels were measured for $G = 17,568$ genes on the Affymetrix GeneChip Human Exon 1.0 ST Array platform. For quality control, RMA background correction, quantile normalization, mean probe set summarization and $\log_2$-transformation were performed [Kang et al. (2011)]. Details for the quality control procedures are described in the supplementary material Section 3 [Lin et al. (2015)]. The number of brains that were collected varies across time periods and for some brains, tissue samples are missing for certain brain regions. So the number of samples varies among brain regions and time periods. We treated samples from the same brain region and time period as biological replicates. Periods 1 and 2 correspond to embryonic and early fetal development, when most of the 16 brain regions sampled in future periods have not differentiated (i.e., most of the 16 brain regions are missing data in periods 1 and 2). Therefore, samples in periods 1 and 2 are excluded in our analysis. In total, we consider $B = 16$ brain regions sampled in $T = 13$ periods of brain development. Let n_{bt} denote the number of replicates for brain region b in period t, $\mathbf{N}_b = (n_{b1}, \ldots, n_{bt}, \ldots, n_{bT})'$ is the column vector for the number of replicates for brain region b, and $\mathbf{N} = (\mathbf{N}_1, \ldots, \mathbf{N}_b, \ldots, \mathbf{N}_B)$ is the matrix summarizing the number of replicates across brain regions and periods. The entries in $\mathbf{N}$ range from 1 to 16 and the median is 5. Let y_{bgtk} denote the observed gene expression value for gene g in the kth replicate of samples in brain region b and period t, and let $\mathbf{y}_{bgt} = (y_{bgt1}, \ldots, y_{bgtn_{bt}})$ denote the expression values for all the replicates. We assume that y_{bgtk}, for $k = 1, \ldots, n_{bt}$, follows the same normal distribution with mean μ_{bgt} and standard deviation σ_0^2:

$$y_{bgtk} \sim \mathcal{N}(\mu_{bgt}, \sigma_0^2).$$

Let x_{bgt} be the binary latent state representing whether gene g is expressed in brain region b and period t, that is, $x_{bgt} = 1$ if the gene is expressed and 0 otherwise. Conditioning on x_{bgt}, we assume that μ_{bgt} follows a Gaussian distribution:

$$\mu_{bgt}|x_{bgt} = 0 \sim \mathcal{N}(\mu_{1b}, \sigma_{1b}^2),$$
$$\mu_{bgt}|x_{bgt} = 1 \sim \mathcal{N}(\mu_{2b}, \sigma_{2b}^2).$$

Marginally, μ_{bgt} follows a Gaussian mixture distribution. We assume that the mean and the variance for the mixture components are brain region specific. Denote by $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ the vectors of parameters for all brain regions. It is easy to see that the distribution of y_{bgtk} conditioning on x_{bgt} has the following form:

$$y_{bgtk}|x_{bgt} = 0 \sim \mathcal{N}(\mu_{1b}, \sigma_{1b}^2 + \sigma_0^2),$$
$$y_{bgtk}|x_{bgt} = 1 \sim \mathcal{N}(\mu_{2b}, \sigma_{2b}^2 + \sigma_0^2).$$

Given the latent state array $\mathbf{X}$, conditional independence is assumed:

$$f(\mathbf{Y}|\mathbf{X}) = \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{t=1}^{T} f(\mathbf{y}_{bgt}|x_{bgt}),$$

where

$$f(\mathbf{y}_{bgt}|x_{bgt}) = \prod_{k=1}^{n_{bt}} f(y_{bgtk}|x_{bgt}).$$

2.1.2. *A MRF model for $p(\mathbf{X})$.* One key component in the above model and the inferential objective is the latent state array $\mathbf{X}$, which is unknown to us. Now we discuss how to specify the prior on $\mathbf{X}$, denoted by $p(\mathbf{X})$, through a MRF model that takes into account both temporal dependency and spatial similarity. For each gene g, we construct an undirected graph $G_g = \{V_g, E_g\}$, where $V_g = \{x_{bgt} : b = 1, \ldots, B, t = 1, \ldots, T\}$ is the set of nodes and E_g is the set of edges. E_g can be divided into two subsets, E_{g1} and E_{g2}, where $E_{g1} = \{(x_{bgt}, x_{b'gt'}) : b \neq b'$ and $t = t'\}$ and $E_{g2} = \{(x_{bgt}, x_{b'gt'}) : b = b'$ and $|t - t'| = 1\}$. E_{g1} contains the edges capturing spatial similarity between brain regions and E_{g2} contains the edges capturing temporal dependency between adjacent periods. For the joint distribution of $p(\mathbf{X})$, we construct a pairwise interaction MRF model [Besag (1986)] with the following form:

(1)
$$\begin{aligned}
p(\mathbf{X}|\mathbf{\Phi}) \propto \prod_{g=1}^{G} \exp\Big\{ &\gamma_0 \sum_{V_g} I_0(x_{bgt}) + \gamma_1 \sum_{V_g} I_1(x_{bgt}) \\
&+ \beta_1 \sum_{E_{g1}} [I_0(x_{bgt})I_0(x_{b'gt'}) + I_1(x_{bgt})I_1(x_{b'gt'})] \\
&+ \beta_2 \sum_{E_{g2}} [I_0(x_{bgt})I_0(x_{b'gt'}) + I_1(x_{bgt})I_1(x_{b'gt'})] \Big\},
\end{aligned}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the indicator functions. Letting $\gamma = \gamma_1 - \gamma_0$, the conditional probability can be derived (see Appendix for the details of derivation):

(2)
$$p(x_{bgt}|\mathbf{X}/x_{bgt}; \mathbf{\Phi}) = \frac{\exp\{x_{bgt}F(x_{bgt}, \mathbf{\Phi})\}}{1 + \exp\{F(x_{bgt}, \mathbf{\Phi})\}},$$

where

$$\begin{aligned}
F(x_{bgt}, \mathbf{\Phi}) = \gamma &+ \beta_1 \sum_{b' \neq b} (2x_{b'gt} - 1) \\
&+ \beta_2 \{I_{t \neq 1}[2x_{bg(t-1)} - 1] + I_{t \neq T}[2x_{bg(t+1)} - 1]\},
\end{aligned}$$

where "/" means other than; $\mathbf{\Phi} = (\gamma, \beta_1, \beta_2)$ and $\gamma, \beta_1, \beta_2 \in \mathbf{R}$; β_1 is the parameter capturing the spatial similarity and β_2 is the parameter capturing the temporal dependency.

2.2. *Biological question* 2: *Identify DE genes over time.*

2.2.1. *A latent state model for DE.*   For DE analysis, we first transform the observed data into an array where the entries are then used in the follow-up analysis. This is accomplished by performing t-tests between adjacent periods and transforming the t-statistics into z-scores. Let $\mathbf{y}_{bg(t-1)}$ and $\mathbf{y}_{bgt}$ denote the vectors of expression values for gene g in region b and in periods $t-1$ and t, respectively. The two-sample t-statistic is obtained by

$$t_{bg(t-1)} = \frac{\bar{\mathbf{y}}_{bgt} - \bar{\mathbf{y}}_{bg(t-1)}}{s},$$

where s is an estimate of the standard error for $\bar{\mathbf{y}}_{bgt} - \bar{\mathbf{y}}_{bg(t-1)}$. The test statistic $t_{bg(t-1)}$ is then transformed into $z_{bg(t-1)}$:

$$z_{bg(t-1)} = \Phi^{-1}\big(F_{n_{bt}+n_{b(t-1)}-2}(t_{bg(t-1)})\big),$$

where $n_{b(t-1)}$ and n_{bt} are the numbers of replicates in $\mathbf{y}_{bg(t-1)}$ and $\mathbf{y}_{bgt}$; Φ and $F_{n_{bt}+n_{b(t-1)}-2}$ are the c.d.f.s for standard normal and t distribution with $n_{bt} + n_{b(t-1)} - 2$ degrees of freedom. As a result, the gene expression data are represented by a $B \times G \times (T-1)$ z-score array $\mathbf{Z}$. The entry z_{bgt} represents the evidence of DE between periods t and $t+1$ for gene g in brain region b. Some entries in the array are not assigned values because of the presence of unexpressed genes. The variations in the expression values of unexpressed genes may be mostly caused by noise in the microarray experiments and we do not want to include that noise in identifying DE genes; the transitions from unexpressed to expressed and vice versa are already captured in biological question 1. Therefore, no t-test is performed if the gene is unexpressed in at least one of the adjacent periods. Let s_{bgt} denote the binary latent state representing whether gene g is differentially expressed in brain region b between periods t and $t+1$, which is the objective of our inference. Let $\mathbf{S}$ be the latent state array of dimensions $B \times G \times (T-1)$. Conditioning on s_{bgt}, we assume that z_{bgt} follows a mixture distribution:

$$f(z_{bgt}|s_{bgt}) = (1 - s_{bgt})f_0(z_{bgt}) + s_{bgt}f_1(z_{bgt}),$$

where $f_0(z)$ is the null density and $f_1(z)$ is the nonnull density. We assume that the null density follows a standard normal $\mathcal{N}(0, 1)$ distribution. We adopt the nonparametric empirical Bayesian framework for DE [Efron (2004)] by fitting the nonnull density with a natural spline using the R package *locfdr*. Given $\mathbf{S}$, conditional independence is assumed:

$$f(\mathbf{Z}|\mathbf{S}) = \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{t=1}^{T-1} f(z_{bgt}|s_{bgt}).$$

2.2.2. *A MRF model for $p(\mathbf{S})$.* Next, we present a MRF model for the prior distribution $p(\mathbf{S})$, taking into account both temporal dependency and spatial similarity. We separate the 16 brain regions into two groups: 11 neocortex regions, represented by $\mathbf{B}_c$, and 5 nonneocortex regions, represented by $\mathbf{B}_n$. The joint probability is similar to (1), except that different spatial parameters are assumed for the two groups. The conditional probability can be calculated and has the following form:

$$(3) \qquad p(s_{bgt}|\mathbf{S}/s_{bgt}; \boldsymbol{\Phi}_{\mathrm{DE}}) = \frac{\exp\{s_{bgt}F_{\mathrm{DE}}(s_{bgt}, \boldsymbol{\Phi}_{\mathrm{DE}})\}}{1 + \exp\{F_{\mathrm{DE}}(s_{bgt}, \boldsymbol{\Phi}_{\mathrm{DE}})\}},$$

if $b \in \mathbf{B}_c$,

$$F_{\mathrm{DE}}(s_{bgt}, \boldsymbol{\Phi}_{\mathrm{DE}}) = \gamma_{\mathrm{DE}} + \beta_{\mathrm{cc}} \sum_{b' \in \mathbf{B_c}/b} (2s_{b'gt} - 1) + \beta_{\mathrm{cn}} \sum_{b' \in \mathbf{B}_n} (2s_{b'gt} - 1)$$
$$+ \beta_t \{I_{t\neq 1}[2s_{bg(t-1)} - 1] + I_{t\neq T}[2s_{bg(t+1)} - 1]\},$$

else if $b \in \mathbf{B}_n$,

$$F_{\mathrm{DE}}(s_{bgt}, \boldsymbol{\Phi}_{\mathrm{DE}}) = \gamma_{\mathrm{DE}} + \beta_{\mathrm{nn}} \sum_{b' \in \mathbf{B}_{n/b}} (2s_{b'gt} - 1) + \beta_{\mathrm{nc}} \sum_{b' \in \mathbf{B}_c} (2s_{b'gt} - 1)$$
$$+ \beta_t \{I_{t\neq 1}[2s_{bg(t-1)} - 1] + I_{t\neq T}[2s_{bg(t+1)} - 1]\},$$

where $\boldsymbol{\Phi}_{\mathrm{DE}} = (\beta_{\mathrm{cc}}, \beta_{\mathrm{nn}}, \beta_{\mathrm{cn}}, \beta_{\mathrm{nc}})$, β_{cc} is the between neocortex coefficient, β_{nn} is the between nonneocortex coefficient, β_{cn} is the neocortex to nonneocortex coefficient, and β_{nc} is the nonneocortex to neocortex coefficient. For symmetry, we assume that $\beta_{\mathrm{cn}} = \beta_{\mathrm{nc}}$. In the MRF model in Section 2.1.2, we did not separate the brain regions into two groups because the latent states for all brain regions were quite similar, which will be shown in Section 4.

3. Parameter and posterior probability estimation.

3.1. *Parameter estimation for biological question* 1: *Identify expressed and unexpressed genes.* In the model, the MRF parameters $\boldsymbol{\Phi} = (\gamma, \beta_1, \beta_2)$ and the Gaussian mixture model parameters $\boldsymbol{\Theta} = (\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)$ need to be estimated. Given the latent state $\mathbf{X}$, both $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ can be estimated by the maximum likelihood estimates (MLE). However, the latent state is unobserved and needs to be estimated as well. Although the expectation–maximization (EM) algorithm is generally implemented for missing data estimation, it is not applicable to our model as the expectation term is not tractable. Therefore, we propose the following Monte Carlo EM Algorithm [Wei and Tanner (1990)] to estimate $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$:

1. Estimate σ_0 by the unbiased estimator:

$$\hat{\sigma}_0^2 = \frac{1}{G \times \sum_{b=1}^{B} \sum_{t=1}^{T}(n_{bt} - 1)} \sum_{g=1}^{G} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{k=1}^{n_{bt}} (y_{bgtk} - \bar{y}_{bgt})^2.$$

2. Obtain the initial estimates $\hat{\mathbf{X}}$ and $\hat{\boldsymbol{\Theta}}$ by the simple Gaussian mixture model, without considering spatial and temporal dependency.

3. Because there is no explicit MLE for $\boldsymbol{\Phi}$, an initial estimate $\hat{\boldsymbol{\Phi}}$ is chosen which maximizes the following pseudolikelihood function $l(\hat{\mathbf{X}}; \boldsymbol{\Phi})$ [Besag (1974)]:

$$l(\hat{\mathbf{X}}; \boldsymbol{\Phi}) = \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{t=1}^{T} p(\hat{x}_{bgt} | \hat{\mathbf{X}}/\hat{x}_{bgt}; \boldsymbol{\Phi}),$$

where $p(\hat{x}_{bgt} | \hat{\mathbf{X}}/\hat{x}_{bgt}; \boldsymbol{\Phi})$ is as defined in (2).

4. Let $\boldsymbol{\Psi} = (\boldsymbol{\Phi}, \boldsymbol{\Theta})$. The expected complete data log-likelihood in the EM algorithm is approximated by the Monte Carlo sum [Wei and Tanner (1990)]:

$$(4) \qquad Q_m(\boldsymbol{\Psi} | \hat{\boldsymbol{\Psi}}^{(r)}) = \frac{1}{m} \sum_{l=1}^{m} \ln f(\mathbf{Y}, \mathbf{X}_l^{(r)} | \boldsymbol{\Psi}),$$

where $\mathbf{X}_1^{(r)}, \ldots, \mathbf{X}_m^{(r)}$ are obtained by Gibbs sampling. From $\mathbf{X}_l^{(r)}$ to $\mathbf{X}_{(l+1)}^{(r)}$, all entries in $\mathbf{X}_l^{(r)}$ are updated, and they are updated sequentially by

$$(5) \qquad p(x_{bgt} | \mathbf{Y}, \mathbf{X}/x_{bgt}; \hat{\boldsymbol{\Psi}}^{(r)}) \propto p(x_{bgt} | \mathbf{X}/x_{bgt}; \hat{\boldsymbol{\Phi}}^{(r)}) f(\mathbf{y}_{bgt} | x_{bgt}; \hat{\boldsymbol{\Theta}}^{(r)}).$$

5. Update $\boldsymbol{\Psi}$ by $\hat{\boldsymbol{\Psi}}^{(r+1)}$, which maximizes (4):

$$\hat{\boldsymbol{\Psi}}^{(r+1)} = \arg\max_{\boldsymbol{\Psi}} Q_m(\boldsymbol{\Psi} | \hat{\boldsymbol{\Psi}}^{(r)}).$$

Same as in step 3, we replace the likelihood by the pseudolikelihood function in $Q_m(\boldsymbol{\Psi} | \hat{\boldsymbol{\Psi}}^{(r)})$. The terms that contain $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are separable, therefore, they can be optimized separately.

6. Repeat steps 4 and 5 until convergence.

3.2. *Parameter estimation for biological question* 2: *Identify DE genes over time.* In the model, only the parameters $\boldsymbol{\Phi}$ in the MRF prior need to be updated iteratively. The algorithm shares some similarity with that in the previous section:

1. Pool the z-scores in $\mathbf{Z}$ and estimate f_1 by the *locfdr* procedure.

2. Obtain an initial estimate $\hat{\mathbf{S}}$ by the simple mixture model, without considering spatial and temporal dependency.

3. Obtain an initial estimate $\hat{\boldsymbol{\Phi}}_{\text{DE}}$, which maximizes the pseudolikelihood function:

$$l(\hat{\mathbf{S}}; \boldsymbol{\Phi}_{\text{DE}}) = \prod_{b=1}^{B} \prod_{g=1}^{G} \prod_{t=1}^{T-1} p(\hat{s}_{bgt} | \hat{\mathbf{S}}/\hat{s}_{bgt}; \boldsymbol{\Phi}_{\text{DE}}),$$

where $p(\hat{s}_{bgt} | \hat{\mathbf{S}}/\hat{s}_{bgt}; \boldsymbol{\Phi}_{\text{DE}})$ is as defined in (3).

4. Approximate the expected complete data log-likelihood by the Monte Carlo sum:

$$(6) \qquad Q_m\big(\boldsymbol{\Phi}_{\mathrm{DE}}|\hat{\boldsymbol{\Phi}}_{\mathrm{DE}}^{(r)}\big) = \frac{1}{m} \sum_{l=1}^{m} \ln f\big(\mathbf{Z}, \mathbf{S}_l^{(r)}|\boldsymbol{\Phi}_{\mathrm{DE}}\big),$$

where $\mathbf{S}_1^{(r)}, \ldots, \mathbf{S}_m^{(r)}$ are obtained by Gibbs sampling. From $\mathbf{S}_l^{(r)}$ to $\mathbf{S}_{(l+1)}^{(r)}$, all entries in $\mathbf{S}_l^{(r)}$ are updated, and they are updated sequentially by

$$(7) \qquad p\big(s_{bgt}|\mathbf{Z}, \mathbf{S}/s_{bgt}; \hat{\boldsymbol{\Phi}}_{\mathrm{DE}}^{(r)}\big) \propto p\big(s_{bgt}|\mathbf{S}/s_{bgt}; \hat{\boldsymbol{\Phi}}_{\mathrm{DE}}^{(r)}\big) f(z_{bgt}|s_{bgt}).$$

5. Update $\boldsymbol{\Phi}_{\mathrm{DE}}$ by $\hat{\boldsymbol{\Phi}}_{\mathrm{DE}}^{(r+1)}$, which maximizes (6).
6. Repeat steps 4 and 5 until convergence.

3.3. *Posterior probability estimation and FDR controlling procedure.* To acquire an estimate of the posterior probability, we implement a separate Gibbs sampler and keep the model parameters fixed at the estimated values by the MCEM algorithm. The latent states in biological questions 1 and 2 are updated sequentially according to (5) and (7).

For the inference of expressed/unexpressed genes, we use 0.5 as the cutoff for the posterior probability. For the inference of DE genes, we adapt the posterior probability-based definition of FDR [Li, Wei and Maris (2010), Newton et al. (2001)]. The posterior local f.d.r. $q_{bgt} = p(s_{bgt} = 0|\mathbf{Z})$ is estimated by the Gibbs sampler. Let $q_{(s)}$ be the sorted values of q_{bgt} in ascending order. Find $k = \max\{t : \frac{1}{t} \sum_{s=1}^{t} q_{(s)} \leq \alpha\}$ and reject all the null hypotheses $H_{(s)}$, for $s = 1, \ldots, k$. In the analysis of human brain gene expression data, we chose $\alpha = 0.05$.

4. Application to the human brain microarray data.

4.1. *Identify expressed and unexpressed genes.* We first applied the MRF model to infer whether a gene is expressed or not in a certain brain region and time period. In the parameter estimation, we first ran 20 iterations of MCEM by a Gibbs sampler with 500/1500 (1500 iterations in total and 500 as burn-in), then 20 iterations with 1000/6000 and, finally, 20 iterations with 1000/10,000. We gradually increased the number of iterations in the Gibbs sampler to make the estimate of the parameters more stable. The posterior probability was then estimated by a Gibbs sampler with 10,000 iterations and 1000 as burn-in. A diagnosis for the number of iterations is presented in the supplementary material Section 5 [Lin et al. (2015)].

The estimated parameters for the Gaussian mixture model are shown in Table 2. The estimated parameters for the MRF prior were $\gamma = 0.30$, $\beta_1 = 0.22$ and $\beta_2 = 6.44$. The large coefficient in β_2 indicates strong temporal dependency. Compared with the total number of genes (17,568), only a small number of genes changed

TABLE 2
The estimated parameters for the Gaussian mixture model

Region	μ_1	μ_2	σ_1	σ_2
MFC	4.58	7.82	0.59	1.57
OFC	4.57	7.83	0.59	1.58
VFC	4.56	7.84	0.58	1.59
DFC	4.58	7.83	0.58	1.58
STC	4.62	7.8	0.58	1.56
ITC	4.61	7.81	0.58	1.57
A1C	4.6	7.82	0.58	1.57
IPC	4.61	7.81	0.58	1.57
S1C	4.61	7.82	0.58	1.58
M1C	4.60	7.82	0.58	1.58
V1C	4.63	7.78	0.59	1.55
AMY	4.65	7.76	0.6	1.52
HIP	4.64	7.77	0.61	1.54
STR	4.65	7.78	0.62	1.55
MD	4.62	7.81	0.63	1.59
CBC	4.61	7.76	0.65	1.58

their latent states between adjacent periods (Figure 1). The table for the numbers are presented in supplementary material Section 8 [Lin et al. (2015)]. For all brain regions, a general trend can be observed: the number of genes that changed their latent states first increased, peaked in periods 6 to 7, the number in periods 7 to 8
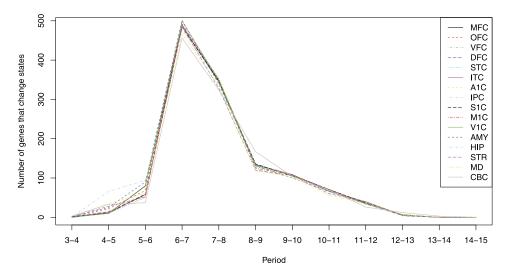

FIG. 1. *The number of genes that changed from expressed to unexpressed and vice versa in adjacent periods. Each line represents a brain region.*

TABLE 3
Summary of the latent states by pooling brain regions. "0" represents the total count of genes that were unexpressed in all brain regions and "16" represents the total count of genes that were expressed in all brain regions

0	89,347
1	2560
2	541
3	218
4	95
5	62
6	31
7	52
8	31
9	26
10	19
11	46
12	42
13	94
14	99
15	297
16	134,824

was also large, then gradually decreased, starting from periods 12 to 13, fewer than 15 genes changed their latent states. Period 8 corresponds to birth to 6 postnatal months. The observation that the changes in gene expression peaked from periods 6 to 8 suggests that robust changes in gene expression occurred close to birth.

Moreover, we observed that the latent states for the same gene in all brain regions tended to agree with each other. These are summarized in Table 3, where we considered all genes by time combinations, that is, $G \times T = 17{,}568 \times 13 = 228{,}384$, and counted the number of genes that were expressed in a given number of brain regions. Although the MRF prior encourages the agreement of latent states, the observation is unlikely driven by the model, as we observed a similar trend when the spatial coefficient β_1 was fixed to be 0 (supplementary material Section 8 [Lin et al. (2015)]).

Genes that changed states over time may be of biological interest for the study of brain development. We conducted Gene Ontology (GO) enrichment analysis using DAVID, which takes a list of genes as input and outputs the enriched Gene Ontology (GO) terms [Huang et al. (2008), Sherman et al. (2009)]. A GO term represents the functional annotation of a list of genes and may belong to any of the following three categories: (a) genes that participate in the same biological process, (b) genes that have the same molecular function, and (c) genes that are located in the same

cellular component. Only GO terms in categories (a) and (b) were included in our analysis, as genes located in the same cellular component do not necessarily share similar functions. We observed enrichment of GO terms only from periods 6 to 7 (0.05 threshold for Bonferroni-adjusted p-value). From periods 6 to 7, genes that switched from expressed to unexpressed in all brain regions were enriched for "DNA binding" (Bonferroni adjusted p-value $= 1.6 \times 10^{-9}$), "regulation of transcription, DNA-dependent" (Bonferroni adjusted p-value $= 2.5 \times 10^{-4}$) and "zinc ion binding" (Bonferroni adjusted p-value $= 9.5 \times 10^{-5}$); there were no enriched GO terms for genes that switched from unexpressed to expressed. The enrichment of transcription regulation and DNA binding proteins (including zinc-finger proteins coordinated by the binding of zinc ions) is consistent with our previous observation that robust changes in transcription occurred close to birth. Changes in transcriptional regulation may also lead to the peak of differentially expressed genes (see Section 4.2). Details for the GO enrichment analysis are presented in the supplementary material Section 6 [Lin et al. (2015)].

4.2. *Identify DE genes over time.* After excluding genes that were unexpressed in all brain regions and all periods, 11,370 genes remained. We then applied the MRF model to identify DE genes between adjacent periods. The settings for the MCEM algorithm and the Gibbs sampler were the same as that in the previous section.

The estimated MRF parameters were $\gamma_{\text{DE}} = -0.10$, $\beta_{\text{cc}} = 0.32$, $\beta_{\text{nn}} = 0.53$, $\beta_{\text{cn}} = 0.06$, and $\beta_t = 0.15$. The temporal coefficient β_t was much smaller compared with that in the previous section (where $\beta_2 = 6.44$), which suggests lower temporal dependency. The neocortex to nonneocortex coefficient β_{cn} was much smaller than the neocortex to neocortex coefficient β_{cc} and the nonneocortex to nonneocortex coefficient β_{nn}, which indicates the group difference between neocortex and nonneocortex regions.

When no spatial and temporal dependency is assumed, the model reduces to a simple empirical Bayesian (EB) model. Based on the posterior FDR control procedure described in Section 3, the thresholds in the MRF and EB models were 0.26 and 0.12, respectively. The numbers of genes identified as DE in the two models were 356,207 (MRF) and 77,330 (EB), with 74,228 (96%) overlap. The higher threshold led to more genes identified as DE in the MRF model. The numbers of DE genes identified are presented in Figure 2, where each line represents a brain region. The table of the exact numbers is presented in the supplementary material Section 9 [Lin et al. (2015)]. For the number of DE genes, the trend over time was slightly different from that in the previous section. In addition to the peak close to birth, there was another peak that spanned from early childhood (period 10) to adolescence (period 12). The peak was less obvious in the 5 nonneocortex regions (AMY, HIP, STR, MD and CBC). During these periods, motor skills, social skills, emotional skills and cognitive skills are rapidly developed. The second peak may correspond to the development of these essential skills. Genes that were DE in the
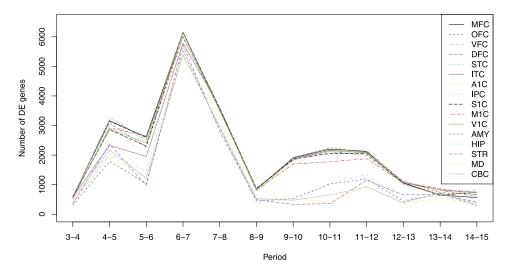
FIG. 2. *The number of DE genes identified in each time window of adjacent periods. Each line represents a brain region.*

second peak may be of interest to researchers studying these behaviors. Note that there was a slight decrease in DE genes in periods 5–6 compared with that in periods 4–5. The decrease was most obvious in brain region STR. Further biological studies are needed to understand the trend. We randomly split the data into two subsets and implemented the algorithm separately for each subset. Compared with the EB model, the genes identified as DE by the MRF model were more likely to overlap: 56.2% vs. 12.4% (supplementary material Section 9 [Lin et al. (2015)]). The information for the direction of changes in gene expression was not utilized in the model. However, we observed that DE genes in all neocortex regions tended to have the same direction of changes (Table 4). Therefore, the MRF model is able to detect consistent changes in gene expression among the brain regions, which may be missed by other approaches not considering temporal and spatial similarity.

Autism Spectrum Disorders (ASD) are a group of syndromes characterized by fundamental impairments in social reciprocity and language development accompanied by highly restrictive interests and/or repetitive behaviors [American Psychiatric Association (2000)]. By exome sequencing, loss of function (LoF) mutations with large biological effects have been shown to affect ASD risk [Iossifov et al. (2012), Kong et al. (2012), Neale et al. (2012), O'Roak et al. (2011, 2012), Sanders et al. (2012)]. A set of nine high-confidence ASD risk genes have been identified recently: ANK2, CHD8, CUL3, DYRK1A, GRIN2B, KATNAL2, POGZ, SCN2A, TBR1 [Willsey et al. (2013)]. These nine genes carry LoF mutations in ASD patients. Details for the genes are described in the supplementary material Section 7

TABLE 4

Summary for the direction of changes in gene expression by pooling neocortex regions. Each row represents a time window. The "0" column represents the counts of genes that were down-regulated in all neocortex regions and the "11" column represents the counts of genes that were up-regulated in all neocortex regions

	0	1	2	3	4	5	6	7	8	9	10	11
Periods 3–4	163	5	1	0	0	0	0	0	0	0	1	47
Periods 4–5	1039	31	3	3	0	0	1	3	0	4	18	436
Periods 5–6	539	30	3	1	1	0	1	0	0	2	20	417
Periods 6–7	3475	28	3	2	1	1	2	2	0	2	29	1238
Periods 7–8	1014	14	1	0	0	0	0	0	0	1	3	1640
Periods 8–9	387	5	0	0	0	0	0	0	0	0	1	146
Periods 9–10	1034	1	0	0	0	0	0	0	0	1	1	351
Periods 10–11	342	2	0	0	0	0	0	0	0	0	3	1124
Periods 11–12	915	9	0	0	0	0	0	0	0	0	1	485
Periods 12–13	450	0	0	0	0	0	0	0	0	0	1	204
Periods 13–14	263	5	0	0	0	0	0	0	0	0	2	39
Periods 14–15	107	22	0	0	0	0	0	0	0	0	5	149

[Lin et al. (2015)]. Next we analyzed the nine ASD risk genes in the human brain gene expression data set. Among the nine genes, KATNAL2 and CHD8 were unexpressed. The other seven genes were expressed in all brain regions and all periods. Gene expression study on postmortem autistic brains and structural magnetic resonance imaging studies have highlighted the frontal cortex as pathological in ASD patients [Amaral, Schumann and Nordahl (2008), Voineagu et al. (2011)]. In the brain gene expression data, five regions were sampled in the frontal cortex: OFC, DFC, VFC, MFC and M1C. The gene expression curves for TBR1 and CHD8 are shown in Figure 3. The five frontal cortex regions shared similar dynamics for the two genes. TBR1 was differentially expressed in periods 4–5 and 6–7, while CHD8 remained unexpressed. We performed a binomial test to see whether the ASD gene set was enriched for DE genes, compared with the overall distribution (Table 5). In the binomial test, a gene was counted as DE only if it was DE in all five frontal cortex regions. We observed an increased fold change of DE genes in the ASD gene set in periods 4–5, 5–6, 6–7, 9–10 and 10–11. It is interesting to note the gap that spanned periods 7 to 9, when the ASD genes tended to be equally expressed. For periods 4–5 and 9–10, the enrichment was significant (<0.05). Period 10 corresponds to early childhood ($1 \leq \text{Age} \leq 6$), when social, emotional and cognitive skills are observed [Kang et al. (2011)]. The most obvious signs of autism tend to emerge between 2 and 3 years of age. In periods 9–10, there were four DE genes: SCN2A, CUL3, ANK2, GRIN2B. These four genes are of potential interest, as a malfunction of these genes in ASD patients may directly affect the development of social and cognitive skills in early childhood.
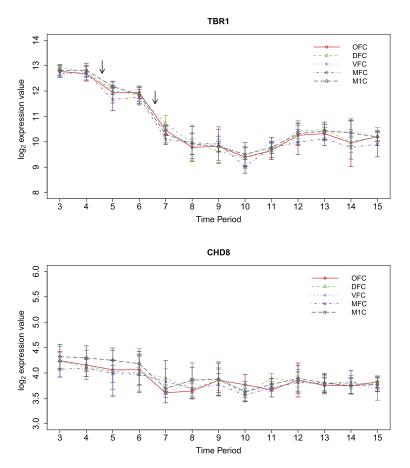
FIG. 3. *The dynamics of gene expression for TBR1 and CHD8 in frontal cortex regions. In periods 4–5 and 6–7, TBR1 was differentially expressed in all frontal cortex regions, as indicated by the arrows in the figure.*

5. Simulation studies.

5.1. *Identify expressed and unexpressed genes.* We conducted simulation studies to evaluate the performance of our proposed MRF model. The expression values for 100 genes in 16 brain regions and 13 periods were simulated. The number of replicates was set to be 3. The latent state array was first simulated and we considered two simulation settings:

Simulation setting 1. The latent state array was simulated by Gibbs sampling. The sampler started from a random array with equal probability of being expressed or unexpressed. The latent states were updated sequentially by (2) and the MRF parameters were set to $\gamma = 0.08$, $\beta_1 = 0.20$ and $\beta_2 = 1.5$. We conducted three rounds of Gibbs sampling to obtain the latent state array $\mathbf{X}$.

TABLE 5
Enrichment analysis of DE genes in the ASD gene set

	# of DE (expected)	# of DE (ASD)	Fold change	*p*-value
Periods 3–4	0.3	0	0	0.62
Periods 4–5	1.6	4	2.5	0.03
Periods 5–6	1.2	3	2.5	0.06
Periods 6–7	3.7	6	1.6	0.05
Periods 7–8	2.1	0	0	0.96
Periods 8–9	0.4	0	0	0.67
Periods 9–10	1.0	4	3.9	0.006
Periods 10–11	1.1	2	1.8	0.19
Periods 11–12	1.1	1	0.9	0.50
Periods 12–13	0.6	0	0	0.72
Periods 13–14	0.3	0	0	0.64
Periods 14–15	0.2	0	0	0.60

Simulation setting 2. In period 1, all genes had equal probability of being unexpressed/expressed. The latent states evolved over time by a Hidden Markov Model with 0.1 transition probability. The latent states for the 16 brain regions were initially set to be the same. Then we let different proportions $(0.1, 0.2, 0.5)$ of the latent states flip randomly.

The gene expression levels were simulated based on the latent states. The mean gene expression array μ was generated from $\mathbf{X}$ by a Gaussian mixture model, where $\mu_1 = 4.5$, $\sigma_1 = 0.75$, $\mu_2 = (5, 5.5, 6, 6.5, 7, 7.5, 8)$ and $\sigma_2 = 1.5$. We varied μ_2 and kept the other parameters unchanged to test the model in different scenarios. Parameters were set to be the same for all brain regions. The gene expression levels $\mathbf{Y}$ were then simulated from a normal distribution, with mean μ and variance $\sigma_0^2 = 0.25$. The MCEM algorithm and the Gibbs sampler were implemented the same as in the previous sections. A comparison of misclassification rates was made between the MRF model and the simple Gaussian mixture model with no temporal and spatial dependency assumed (Table 6). For all simulation settings, the MRF model achieved significant improvement in misclassification rates compared with the simple Gaussian mixture model.

5.2. *Identify DE genes over time.* In the simulation study, data were generated for 100 genes, 16 brain regions and 12 periods. We considered three simulation settings:

Simulation setting 1. The latent state array $\mathbf{S}$ was updated sequentially by (7) and the MRF parameters were set to $\gamma_{DE} = -0.10$, $\beta_{cc} = 0.31$, $\beta_{nn} = 0.52$, $\beta_{cn} = 0.06$ and $\beta_t = 0.14$. To keep the ratio of DE genes roughly the same as

TABLE 6
Comparison of misclassification rates between the simple Gaussian mixture model (GMM) and the MRF model. The standard deviations in 100 independent runs are shown in the brackets. The results for simulation settings 1 and 2 are presented, the numbers after the model names represent the proportions (0.1, 0.2, 0.5) of purturbation in simulation setting 2

μ_2	GMM	MRF	GMM (0.1)	MRF (0.1)
5	0.421 (0.025)	0.093 (0.008)	0.426 (0.012)	0.131 (0.004)
5.5	0.346 (0.017)	0.084 (0.006)	0.375 (0.013)	0.11 (0.004)
6	0.275 (0.011)	0.071 (0.005)	0.31 (0.014)	0.093 (0.002)
6.5	0.203 (0.006)	0.055 (0.004)	0.242 (0.009)	0.083 (0.002)
7	0.144 (0.004)	0.041 (0.003)	0.185 (0.006)	0.072 (0.003)
7.5	0.101 (0.003)	0.029 (0.002)	0.137 (0.004)	0.053 (0.002)
8	0.067 (0.002)	0.020 (0.001)	0.096 (0.004)	0.037 (0.002)
μ_2	GMM (0.2)	MRF (0.2)	GMM (0.5)	MRF (0.5)
5	0.423 (0.008)	0.233 (0.005)	0.421 (0.004)	0.344 (0.008)
5.5	0.378 (0.011)	0.208 (0.005)	0.377 (0.005)	0.312 (0.011)
6	0.31 (0.012)	0.18 (0.004)	0.309 (0.004)	0.261 (0.007)
6.5	0.242 (0.009)	0.144 (0.004)	0.243 (0.004)	0.187 (0.004)
7	0.185 (0.004)	0.106 (0.003)	0.185 (0.004)	0.133 (0.003)
7.5	0.137 (0.004)	0.075 (0.002)	0.138 (0.003)	0.093 (0.002)
8	0.096 (0.003)	0.051 (0.002)	0.096 (0.003)	0.060 (0.002)

that in the real data, the sampler started from a random array with 0.4 probability of being DE. 10% of the genes were then randomly selected to be unexpressed in all brain regions from periods 1 to t or t to $T = 12$, where t was randomly picked from $1, \ldots, T$. The presence of unexpressed genes reflects the fact that a small portion of genes switched their states of unexpressed/expressed in the real data. We conducted three rounds of Gibbs sampling to obtain the latent state array $\mathbf{S}$. The z-score array $\mathbf{Z}$ was then generated from $\mathbf{S}$ by a mixture model. For EE, the z-score was generated from $\mathcal{N}(0, 1)$; for DE, it was generated from $\mathcal{N}(-2, 1)$ or $\mathcal{N}(2, 1)$, with equal probability.

Simulation setting 2. The latent state array $\mathbf{S}$ was simulated by Gibbs sampling with the same setting as in *simulation setting* 1. The mean gene expression array $\boldsymbol{\mu}$ was then generated from $\mathbf{S}$. In period 1, all the genes had mean expression values at 0. From period t to $t + 1$, $\mu_{bg(t+1)} = \mu_{bgt} + s_{bgt}\delta$, where $\delta \sim \mathcal{N}(0, 1)$. Finally, the gene expression array $\mathbf{Y}$ was generated from $\boldsymbol{\mu}$ by Gaussian distribution with variance $\sigma_0^2 = 0.25$ and the number of replicates was set to be 3.

Simulation setting 3. In period 1, all the genes had 0.15 probability of being DE. From periods t to $t + 1$, 70% of the DE genes in period t randomly switched to EE, and the same number of EE genes randomly switched to DE, to keep the

number of DE genes constant over time. To represent the neocortex and nonneo-cortex regions, the first 11 brain regions were set to have the same latent states and the other 5 brain regions were set to be the same. Compared with the first 11 brain regions, 40% of the DE genes randomly switched to EE in the other 5 brain regions. Then we randomly selected different proportions (0.1, 0.2, 0.5) of the DE states to switch to EE; the same number of EE states were randomly selected to switch to DE. 10% of the genes were randomly selected to be unexpressed in all brain regions as in *simulation setting* 1. Finally, the *z*-score array **Z** was generated in the same way as in *simulation setting* 1.

The settings for the MCEM algorithm and the Gibbs sampler were the same as those in the previous section. We calculated the sensitivity and specificity by varying the threshold for the posterior local-f.d.r. We compared the proposed MRF model with the empirical Bayesian (EB) model, which assumes no temporal and spatial dependency (Figure 4). As the neocortex group and the nonneocortex group have different numbers of brain regions (11 vs. 5), the ROC curves were plotted separately for the two groups. Compared with the EB model, the MRF model per-formed better in both the neocortex and nonneocortex regions. The improvement was more significant in the neocortex regions, as there were more brain regions and the MRF model benefits more from the spatial similarity.

6. Conclusions and discussion. The statistical methods developed in this pa-per were motivated from the analysis of human brain development microarray data. These data represent expression profiles in different brain regions at different de-velopmental stages and they allow us to infer (1) whether a gene is expressed or not in a specific brain region in a specific period, and (2) whether a gene is dif-ferentially expressed between two adjacent periods in a specific brain region. To efficiently utilize the spatial similarity between brain regions and temporal depen-dency, we have developed a two-step modeling framework that is based on the Markov Random Field model and local FDR methodology to facilitate statistical inference. Our simulation studies suggest that this model has a lower misclassifi-cation rate compared with commonly used Gaussian mixture models without con-sidering spatial similarity and temporal dependency. Simulation results and real data analysis also suggest that the proposed model improves the power to identify DE genes.

The analysis of the human brain microarray data by our proposed model pro-duces biologically meaningful results. The inferred latent states of "expressed" or "unexpressed" were similar in all brain regions. The number of genes that switched their latent states first increased and peaked at birth, then gradually decreased in adulthood. In periods 6–7, the list of genes that switched from expressed to un-expressed was enriched for transcriptional regulatory genes. For the purpose of identifying DE genes between adjacent periods, we observed a similar trend in the number of DE genes. However, there was an additional peak in periods that cor-respond to childhood and adolescence. These observations reflect the dynamics of
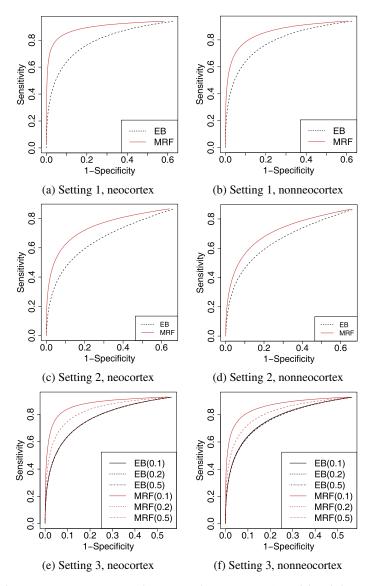
FIG. 4. *The ROC curves comparing the empirical Bayesian (EB) model and the proposed MRF model. The curves were averaged over 100 simulations.*

the neurodevelopment process. We also observed that genes carrying a high risk for neurodevelopment disorders, such as ASD, tended to be differentially expressed, especially during periods when cognitive and social skills were developed.

We have also proposed and implemented an MCEM algorithm to estimate the model parameters and a separate Gibbs sampler to estimate the posterior probability. In previous studies, the iterated conditional mode (ICM) algorithm was imple-

mented to estimate the MRF parameters [Besag (1986), Li, Wei and Maris (2010), Wei and Li (2008)]; however, our simulation study suggested that the ICM algorithm may lead to biased parameter estimates (supplementary material Section 10 [Lin et al. (2015)]). One limitation of the MCEM algorithm is the high computing cost. Under the current setting for the MCEM algorithm, the computing time for the whole data set took ten days (five days for biological question 1 and five days for biological question 2) on the Yale Louise high performance cluster (Dell m620 system, 8 core processor, 48 GB of memory). To accelerate convergence, we started the model from the estimation which does not consider the spatial and temporal dependency. Another limitation of the MCEM algorithm is that the Monte Carlo sum is an approximation to the expectation and may lead to instability in parameter estimation. In the diagnosis of the MCEM algorithm (supplementary material Section 5 [Lin et al. (2015)]), we demonstrated that our model is robust to unstable parameter estimation. Levine and Casella (2001) provided a detailed discussion on the setting of the MCEM algorithm.

APPENDIX

We provide details on the derivation of the conditional probability (2) from the joint probability (1).

For $t \neq 1$ and $t \neq T$,

$$
\frac{p(x_{bgt} = 1 | \mathbf{X}/x_{bgt}; \mathbf{\Phi})}{p(x_{bgt} = 0 | \mathbf{X}/x_{bgt}; \mathbf{\Phi})}
$$

$$
= \frac{p(x_{bgt} = 1, \mathbf{X}/x_{bgt}; \mathbf{\Phi})}{p(x_{bgt} = 0, \mathbf{X}/x_{bgt}; \mathbf{\Phi})}
$$

$$
= \exp\left\{ \gamma_1 - \gamma_0 + \beta_1 \sum_{b' \neq b} \left[I_1(x_{b'gt}) - I_0(x_{b'gt}) \right] \right.
$$

$$
\left. + \beta_2 \left[I_1(x_{bg(t-1)}) - I_0(x_{bg(t-1)}) + I_1(x_{bg(t+1)}) - I_0(x_{bg(t+1)}) \right] \right\}
$$

$$
= \exp\left\{ \gamma + \beta_1 \sum_{b' \neq b} (2x_{b'gt} - 1) + \beta_2 [2x_{bg(t-1)} - 1 + 2x_{bg(t+1)} - 1] \right\},
$$

$p(x_{bgt} = 1 | \mathbf{X}/x_{bgt}; \mathbf{\Phi}) + p(x_{bgt} = 0 | \mathbf{X}/x_{bgt}; \mathbf{\Phi}) = 1$, so we have

$$
p(x_{bgt} = 1 | \mathbf{X}/x_{bgt}; \mathbf{\Phi}) = \frac{\exp\{F(x_{bgt}, \mathbf{\Phi})\}}{1 + \exp\{F(x_{bgt}, \mathbf{\Phi})\}},
$$

where

$$
F(x_{bgt}, \mathbf{\Phi}) = \gamma + \beta_1 \sum_{b' \neq b} (2x_{b'gt} - 1) + \beta_2 \{2x_{bg(t-1)} - 1 + 2x_{bg(t+1)} - 1\}.
$$

For $t = 1$ and $t = T$, the conditional probability can be derived similarly.

SUPPLEMENTARY MATERIAL

Supplement to "A Markov random field-based approach to characterizing human brain development using spatial–temporal transcriptome data" (DOI: 10.1214/14-AOAS802SUPP; .pdf). Section 1: More information on the brain regions. Section 2: Spatial and temporal similarity. Section 3: Microarray quality control procedures. Section 4: Model fit and the robustness of the Gaussian mixture model. Section 5: Diagnosis for the MCEM algorithm. Section 6: Gene Ontology (GO) enrichment analysis. Section 7: High confidence ASD genes. Section 8: Supplementary data for Section 4.1. Section 9: Supplementary data for Section 4.2. Section 10: Comparison between the ICM algorithm and the MCEM algorithm.

REFERENCES

AMARAL, D. G., SCHUMANN, C. M. and NORDAHL, C. W. (2008). Neuroanatomy of autism. *Trends Neurosci.* **31** 137–145.

AMERICAN PSYCHIATRIC ASSOCIATION (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR®*. American Psychiatric Publishing, Arlington, VA.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

CHEN, M., CHO, J. and ZHAO, H. (2011). Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.* **7** e1001353.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing. *J. Amer. Statist. Assoc.* **99** 96–104.

GESCHWIND, D. H. and LEVITT, P. (2007). Autism spectrum disorders: Developmental disconnection syndromes. *Curr. Opin. Neurobiol.* **17** 103–111.

HONG, F. and LI, H. (2006). Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics* **62** 534–544. MR2236847

HUANG, D., SHERMAN, B. T., LEMPICKI, R. A. et al. (2008). Systematic and integrative analysis of large gene lists using David bioinformatics resources. *Nat. Protoc.* **4** 44–57.

IOSSIFOV, I., RONEMUS, M., LEVY, D., WANG, Z., HAKKER, I., ROSENBAUM, J., YAMROM, B., LEE, Y.-H., NARZISI, G., LEOTTA, A. et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron* **74** 285–299.

JOHNSON, M. B., KAWASAWA, Y. I., MASON, C. E., KRSNIK, Ž., COPPOLA, G., BOGDANOVIĆ, D., GESCHWIND, D. H., MANE, S. M., SESTAN, N. et al. (2009). Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62** 494–509.

KANG, H. J., KAWASAWA, Y. I., CHENG, F., ZHU, Y., XU, X., LI, M., SOUSA, A. M., PLETIKOS, M., MEYER, K. A., SEDMAK, G. et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature* **478** 483–489.

KONG, A., FRIGGE, M. L., MASSON, G., BESENBACHER, S., SULEM, P., MAGNUSSON, G., GUDJONSSON, S. A., SIGURDSSON, A., JONASDOTTIR, A., JONASDOTTIR, A. et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488** 471–475.

LEVINE, R. A. and CASELLA, G. (2001). Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Statist.* **10** 422–439. MR1939033

LI, C., WEI, Z. and LI, H. (2010). Network-based empirical Bayes methods for linear models with applications to genomic data. *J. Biopharm. Statist.* **20** 209–222. MR2752203

LI, H., WEI, Z. and MARIS, J. (2010). A hidden Markov random field model for genome-wide association studies. *Biostatistics* **11** 139–150.

LIN, Z., SANDERS, S. J., LI, M., SESTAN, N., STATE, M. W. and ZHAO, H. (2015). Supplement to "A Markov random field-based approach to characterizing human brain development using spatial–temporal transcriptome data." DOI:10.1214/14-AOAS802SUPP.

LIU, X. and YANG, M. C. (2009). Identifying temporally differentially expressed genes through functional principal components analysis. *Biostatistics* **10** 667–679.

NEALE, B. M., KOU, Y., LIU, L., MA'AYAN, A., SAMOCHA, K. E., SABO, A., LIN, C.-F., STEVENS, C., WANG, L.-S., MAKAROV, V. et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485** 242–245.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. and TSUI, K.-W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* **8** 37–52.

O'ROAK, B. J., DERIZIOTIS, P., LEE, C., VIVES, L., SCHWARTZ, J. J., GIRIRAJAN, S., KARAKOC, E., MACKENZIE, A. P., NG, S. B., BAKER, C. et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43** 585–589.

O'ROAK, B. J., VIVES, L., GIRIRAJAN, S., KARAKOC, E., KRUMM, N., COE, B. P., LEVY, R., KO, A., LEE, C., SMITH, J. D. et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485** 246–250.

SANDERS, S. J., MURTHA, M. T., GUPTA, A. R., MURDOCH, J. D., RAUBESON, M. J., WILLSEY, A. J., ERCAN-SENCICEK, A. G., DILULLO, N. M., PARIKSHAK, N. N., STEIN, J. L. et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485** 237–241.

SESTAN, N. et al. (2012). The emerging biology of autism spectrum disorders. *Science* (*New York, NY*) **337** 1301.

SHERMAN, B. T., LEMPICKI, R. A. et al. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37** 1–13.

STOREY, J. D., XIAO, W., LEEK, J. T., TOMPKINS, R. G. and DAVIS, R. W. (2005). Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA* **102** 12837–12842.

TAI, Y. C. and SPEED, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* **34** 2387–2412. MR2291504

VOINEAGU, I., WANG, X., JOHNSTON, P., LOWE, J. K., TIAN, Y., HORVATH, S., MILL, J., CANTOR, R. M., BLENCOWE, B. J. and GESCHWIND, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474** 380–384.

WALSH, C. A., MORROW, E. M. and RUBENSTEIN, J. L. (2008). Autism and brain development. *Cell* **135** 396–400.

WEI, Z. and LI, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23** 1537–1544.

WEI, Z. and LI, H. (2008). A hidden spatial–temporal Markov random field model for network-based analysis of time course gene expression data. *Ann. Appl. Stat.* **2** 408–429. MR2415609

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

WILLSEY, A. J., SANDERS, S. J., LI, M., DONG, S., TEBBENKAMP, A. T., MUHLE, R. A.,
    REILLY, S. K., LIN, L., FERTUZINHOS, S., MILLER, J. A. et al. (2013). Coexpression net-
    works implicate human midfetal deep cortical projection neurons in the pathogenesis of autism.
    Cell **155** 997–1007.

WU, H., YUAN, M., KAECH, S. M. and HALLORAN, M. E. (2007). A statistical analysis of memory
    CD8 T cell differentiation: An application of a hierarchical state space model to a short time
    course microarray experiment. *Ann. Appl. Stat.* **1** 442–458. MR2415750

YUAN, M. and KENDZIORSKI, C. (2006). Hidden Markov models for microarray time course data
    in multiple biological conditions. *J. Amer. Statist. Assoc.* **101** 1323–1332. MR2307565

Z. LIN
INTERDEPARTMENTAL PROGRAM IN COMPUTATIONAL
    BIOLOGY AND BIOINFORMATICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: zhixiang.lin@yale.edu

M. LI
N. SESTAN
DEPARTMENT OF NEUROBIOLOGY
KAVLI INSTITUTE FOR NEUROSCIENCE
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
USA
E-MAIL: mingfeng.li@yale.edu
        nenad.sestan@yale.edu

S. J. SANDERS
M. W. STATE
DEPARTMENT OF PSYCHIATRY
UNIVERSITY OF CALIFORNIA
SAN FRANCISCO, CALIFORNIA 94143
USA
E-MAIL: stephan.sanders@ucsf.edu
        matthew.state@ucsf.edu

H. ZHAO
DEPARTMENT OF BIOSTATISTICS
YALE SCHOOL OF PUBLIC HEALTH
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
USA
E-MAIL: hongyu.zhao@yale.edu