Published in final edited form as:

Stat Appl Genet Mol Biol. 2016 April ; 15(2): 139–150. doi:10.1515/sagmb-2015-0070.

A Markov random field-based approach for joint estimation of differentially expressed genes in mouse transcriptome data

Zhixiang Lin,

Department of Statistics, Stanford University, Stanford, CA 94305, USA

Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA

Mingfeng Li,

Department of Neurobiology, Kavli Institute for Neuroscience, Yale School of Medicine, 06510 New Haven, CT, USA

Nenad Sestan, and

Department of Neurobiology, Kavli Institute for Neuroscience, Yale School of Medicine, 06510 New Haven, CT, USA

Hongyu Zhao

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520, USA

Department of Genetics, Yale School of Medicine, New Haven, Connecticut 06520, USA

Hongyu Zhao: hongyu.zhao@yale.edu

Abstract

The statistical methodology developed in this study was motivated by our interest in studying neurodevelopment using the mouse brain RNA-Seq data set, where gene expression levels were measured in multiple layers in the somatosensory cortex across time in both female and male samples. We aim to identify differentially expressed genes between adjacent time points, which may provide insights on the dynamics of brain development. Because of the extremely small sample size (one male and female at each time point), simple marginal analysis may be underpowered. We propose a Markov random field (MRF)-based approach to capitalizing on the between layers similarity, temporal dependency and the similarity between sex. The model parameters are estimated by an efficient EM algorithm with mean field-like approximation. Simulation results and real data analysis suggest that the proposed model improves the power to detect differentially expressed genes than simple marginal analysis. Our method also reveals biologically interesting results in the mouse brain RNA-Seq data set.

Keywords

Markov random field model; neurodevelopment; RNA-Seq and differential expression

Correspondence to: Hongyu Zhao, hongyu.zhao@yale.edu.

Conflict of interest statement: None declared.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2015-0070) offers supplementary material, available to authorized users.

1 Introduction

Neurodevelopment is a dynamic and highly regulated biological process. Abnormalities in neurodevelopment in humans may lead to psychiatric and neurological disorders, such as Autism Spectrum Disorders (ASD) (Geschwind and Levitt, 2007; Walsh et al., 2008; Sestan and State, 2012). The cerebral neocortex plays central roles in many higher-order functions such as cognition, language, consciousness, and the control of voluntary behavior. It is divided into functionally distinct areas and layers. Each layer contains a distinct set of neurons and synaptic connections (Kwan et al., 2012). The primary somatosensory cortex (S1C) is a part of the neocortex that plays a key role in processing of sensory modalities. The statistical methodology developed in this study was motivated by our interest in studying neurodevelopment using the mouse brain RNA-Seq data set reported in Fertuzinhos et al. (2014). At each time point, including postnatal day 4 (P4), P6, P8, P10, P14 and P180 (adult), samples from one male and one female mice were collected. For each mouse, three layers of samples in S1C were collected: IgL (layers 2/3), L4 (layer 4), and SgL (layers 5 and 6).

RNA-sequencing (RNA-Seq) technology produces transcriptomic data in the form of short reads (Kwan et al., 2012). Compared with other transcriptomic methods, such as microarray, RNA-Seq offers larger dynamic range for more accurate expression level estimation and also allows the evaluation of allele and isoformspecific expression (Wang et al., 2009). Methods developed for the quantification of RNA-Seq data can be broadly divided into two categories: gene-level evaluation and transcript-level evaluation. The major focus of this paper is differential expression (DE) estimation in gene-level read count data. For the estimation of gene-level DE, DESeq (Anders and Huber, 2010) and edgeR (Robinson and Smyth, 2007, 2008; Robinson et al., 2010; McCarthy et al., 2012; Zhou et al., 2014) are among the most commonly used methods. Both methods modeled the gene-level read count data by negative binomial distribution and used exact tests. One major difference between edgeR and DESeq is the estimation of the over-dispersed variance: edgeR used empirical Bayes estimation and DESeq implemented local regression. For transcript-level evaluation, numerous methods are available, including Cufflinks (Trapnell et al., 2010; Trapnell et al., 2013), IsoEM (Nicolae et al., 2011), RSEM (Li and Dewey, 2011), BitSeq (Glaus et al., 2012), TIGAR (Nariai et al., 2013), EBSeq (Leng et al., 2013) and Casper (Rossell et al., 2014).

The goal of our analysis is to characterize neurodevelopment through the dynamics of gene expression, such as the identification of DE genes between adjacent time points. The unique challenge in our analysis is to appropriately model the complex data structure, including the spatial similarity between brain layers, the temporal dependency and the similarity between males and females. We propose a Markov Random Field (MRF)-based approach to identify DE genes between adjacent time points. We note that MRF models have been used to model dependency in genomics data, such as neighboring genes defined by biological pathways (Wei and Li, 2007, 2008; Li et al., 2010a; Chen et al., 2011), and marker dependencies defined by linkage disequilibrium (Li et al., 2010b). Our methodology utilizes the general methodology of edgeR and local false discovery rate (FDR) (Efron, 2004). We implemented an efficient expectation-maximization (EM) algorithm to estimate the model parameters and

a separate Gibbs sampler to estimate the posterior probabilities. In our previous work, we proposed a MRF-based approach to identify DE genes in human brain microarray gene expression data and proposed an Monte-Carlo EM (MCEM) algorithm to estimate the model parameters (Lin et al., 2015). In this study, we extend the methodology to RNA-Seq read count data and implemented a much faster EM algorithm with mean field-like approximation.

The key feature of our approach is to simultaneously consider the similarities between layers, time points and sex, and extract biologically meaningful results from the data. Our method can be naturally extended to other data set with complex structures. We introduce the statistical models and methods in Section 2, including the MRF model, the EM algorithm, the posterior probability estimation and the FDR control procedure. In Section 3, we apply our method to analyze the mouse brain RNA-Seq data reported in Fertuzinhos et al. (2014). Results from simulation studies are summarized in Section 4. In Section 5, the computational cost is briefly covered. We conclude the manuscript with a brief discussion in Section 6.

2 Statistical models and methods

2.1 The mouse RNA-Seq data set

In the mouse RNA-Seq data set, mRNA read counts for 12,729 protein-coding genes were measured. Details for the experimental procedures and data preprocessing steps were discussed in Fertuzinhos et al. (2014). Brain samples in six time points were collected: postnatal day 4 (P4), P6, P8, P10, P14, and P180 (adult). The time points span from early postnatal periods to adulthood. At each time point, two mice were sacrificed, one male and one female. For each brain, samples from three layers in S1C were collected: IgL, L4, and SgL. The whole data set can be represented by an array of dimension $B \times T \times S \times G$, where $B=3$ denotes the number of layers, $T=6$ denotes the number of time points, $S=2$ denotes the two sexes, and $G=12,729$ denotes the number of genes.

2.2 A latent state model for differential expression

To acquire evidence of differential expression between adjacent time points, we first adapted the general framework of edgeR (Robinson and Smyth, 2007, 2008; Robinson et al., 2010; McCarthy et al., 2012; Zhou et al., 2014). The RNA-Seq read count data often demonstrate over-dispersed variance. Estimation of dispersion generally requires replicates. However, there are no replicates in our data set. In the hierarchical clustering, samples from male and female mice (same layer and time point) tend to form the smallest cluster. In our analysis, we collapsed the male and female samples to estimate the gene-specific dispersion. Then we adapted the exact binomial test in edgeR with the following slight modification. In edgeR, the original read count data was transformed into pseudo-counts. The two-sided p -value was defined as the sum of the probabilities that were no more likely than the probability of the observed pseudo-counts, conditioning on the total sum of the pseudo-counts (Robinson and Smyth, 2008). In our analysis, we calculated the quantile (or equivalently, the one-sided p -value) by summing over the probability of one side. Let q_{bts_g} denote the quantile for the test

between time point t and $t+1$ for gene g in layer b and sex s . Note that both small and large q indicate evidence of differential expression. We then transformed the quantile into z -score:

$$z_{btsg} = \Phi^{-1}(q_{btsg}),$$

As a result, the RNA-Seq data were represented by a $B \times (T-1) \times S \times G$ z -score array $\mathbf{Z}$. Let x_{btsg} denote the binary latent state indicating whether gene g is differentially expressed between t and $t+1$ in layer b and sex s , which is the objective of our inference. Let $\mathbf{X}$ be the $B \times (T-1) \times S \times G$ dimension array with entries x_{btsg} . Conditioning on x_{btsg} , we assume that z_{btsg} follows a mixture distribution:

$$p(z_{btsg} | x_{btsg}) = (1 - x_{btsg})f_0(z_{btsg}) + x_{btsg}f_1(z_{btsg}),$$

where $f_0(\cdot)$ is the null density and $f_1(\cdot)$ is the non-null density. We adapted the nonparametric empirical Bayesian framework (Efron, 2004) by fitting the densities by a natural spline using the R package *locfdr*. We used empirical null for the null density f_0 in both simulation studies and real data analysis. Given $\mathbf{X}$, we assume conditional independence among the z_{btsg} :

$$p(\mathbf{Z} | \mathbf{X}) = \prod_{b=1}^B \prod_{t=1}^{T-1} \prod_{s=1}^S \prod_{g=1}^G p(z_{btsg} | x_{btsg})$$

2.3 A MRF model for the prior on $\mathbf{X}$

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 distribution on $\mathbf{X}$, through a MRF model that takes into account the similarity between layers, time points and sex. For each gene g , we construct an undirected graph $G_g = \{V_g, E_g\}$, where $V_g = \{x_{btsg} : b=1, \dots, B; t=1, \dots, T-1; s=1, \dots, S\}$ is the set of nodes and E_g is the set of edges. E_g is divided into three groups, E_{g1} , E_{g2} and E_{g3} , where $E_{g1} = \{(x_{btsg}, x_{b's't'g}) : b=b', t=t' \text{ and } s=s'\}$, $E_{g2} = \{(x_{btsg}, x_{b's't'g}) : b=b', |t-t'|=1 \text{ and } s=s'\}$, $E_{g3} = \{(x_{btsg}, x_{b's't'g}) : b=b', t=t' \text{ and } s \neq s'\}$. E_{g1} capture the spatial similarity between layers, E_{g2} capture the temporal dependency and E_{g3} capture the similarity between sex. We do not consider the direction of the edges: $(x_{btsg}, x_{b's't'g})$ and $(x_{b's't'g}, x_{btsg})$ represent the same edge. We construct a pairwise interaction MRF prior on $\mathbf{X}$ (Besag, 1986) with the following form,

$$p(\mathbf{X} | \Phi) \propto \prod_{g=1}^G \exp \left\{ \gamma_0 \sum_{V_g} (1 - x_{btsg}) + \gamma_1 \sum_{V_g} x_{btsg} + \sum_{E_g} \beta(E_g) [x_{btsg} x_{b's't'g} + (1 - x_{btsg})(1 - x_{b's't'g})] \right\}, \quad (1)$$

with slight abuse of notation, is $\beta((b, t, s, g), (b', t', s', g))$ has been written as $\beta(E_g)$ for brevity. $\beta(E_g)$ represents the parameter for the edge: for edges in E_{g2} , $\beta(E_g) = \beta_b$, where we assume that the parameter is constant over time; for edges in E_{g3} , $\beta(E_g) = \beta_s$; for edges in E_{g1} , we assume the parameters to be layer specific, it can be β_{st} , $\beta_{st'}$ or β_{lt} , which captures the similarity between SgL-L4, SgL-IgL, and L4-IgL correspondingly. Let $\Phi = (\gamma_0, \gamma_1, \beta_{ls}, \beta_{st}, \beta_{st'}, \beta_b, \beta_s)$. In the MRF model, the normalizing constant is generally computationally infeasible to calculate as it involves the summation over all possible configurations of $\mathbf{X}$. Let $\gamma = \gamma_1 - \gamma_0$, the conditional probability can be derived:

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

where,

$$F(x_{btsg}, \Phi) = \gamma + \sum_{\mathcal{N}(x_{btsg})} \beta(E_g) (2x_{bt's't'g} - 1),$$

where “/” means other than; $\mathcal{N}(x_{btsg})$ denote all the neighbors of x_{btsg} , which is specified by the edge set E_g . Note that in (1), the edge set E_g can be adjusted to incorporate other complex structures.

2.4 Parameter and posterior probability estimation

2.4.1 Parameter estimation—In our model, both the latent state array $\mathbf{X}$ and the model parameters Φ have to be estimated. The standard expectation-maximization (EM) algorithm is not applicable in our model because the expectation term is not computationally tractable. We propose to implement an EM algorithm with mean field-like approximation (Celeux et al., 2003). The mean field approximation principle originates from statistical mechanics (Chandler, 1987), and the key idea is that the fluctuation of the neighbors of a node is ignored by fixing the neighbors to some configuration (Zhang, 1992). In standard mean field approximation, the configuration is taken to be the mean of the probability distribution. This principle makes the nodes in the graph independent with each other and the computation becomes tractable. In our analysis, instead using the mean, we used random samples from the Gibbs sampler as the configuration. This approach is termed simulated field in Celeux et al. (2003). In our simulation studies, the simulated field approach was faster and led to less biased parameter estimation, compared with the approach using the mean (data not shown). Let $\tilde{\mathbf{X}}$ denote a configuration, $p(\mathbf{X} | \Phi)$ was approximated by $p_{\tilde{\mathbf{X}}}(\mathbf{X} | \Phi)$ which has the following form:

$$p_{\tilde{\mathbf{X}}}(\mathbf{X} | \Phi) = \prod_{b,t,s,g} p(x_{btsg} | \tilde{\mathcal{N}}(x_{btsg}); \Phi), \quad (3)$$

where $\tilde{\mathcal{N}}(x_{btsg})$ denotes the neighbors of x_{btsg} that were fixed to the corresponding entries in $\tilde{\mathbf{X}}$. $p(x_{btsg} | \tilde{\mathcal{N}}(x_{btsg}); \Phi)$ has the same form as that in (2). The log complete likelihood is:

$$\log p_{\tilde{X}}(X, Z|\Phi) = \log p(Z|X) + \log p_{\tilde{X}}(X|\Phi) = \sum_{b,t,s,g} \log p(z_{btsg}|x_{btsg}) + \sum_{b,t,s,g} \log p(x_{btsg}|\tilde{\mathcal{N}}(x_{btsg});\Phi), \quad (4)$$

where the first sum in (4) does not depend on Φ and is ignored when updating Φ in the EM algorithm. The Q function then takes the form:

$$Q(\Phi|\hat{\Phi}) = \sum_{b,t,s,g} \{p(x_{btsg}=0|\tilde{\mathcal{N}}(x_{btsg}), Z;\hat{\Phi}) \log p(x_{btsg}=0|\tilde{\mathcal{N}}(x_{btsg});\Phi) + p(x_{btsg}=1|\tilde{\mathcal{N}}(x_{btsg}), Z;\hat{\Phi}) \log p(x_{btsg}=1|\tilde{\mathcal{N}}(x_{btsg});\Phi)\}, \quad (5)$$

where $p(x_{btsg}|\tilde{\mathcal{N}}(x_{btsg}), Z;\hat{\Phi}) \propto p(z_{btsg}|x_{btsg}) p(x_{btsg}|\tilde{\mathcal{N}}(x_{btsg});\hat{\Phi})$. With mean field-like approximation, the normalizing constant in the MRF prior disappears and the Q function becomes analytically tractable. The EM algorithm consists of the following steps:

1. Initialization: pool the z -scores in Z , and estimate f_0 and f_1 by the *locdr* procedure. Obtain an initial $\tilde{X}$ by the simple mixture model.
2. E-step: Denote $\hat{\Phi}^{(r)}$ the estimated parameters in the r th iteration. With $\tilde{X}$, $Q(\Phi|\hat{\Phi}^{(r)})$ becomes tractable as in (5).
3. M-Step: Update Φ which maximizes $Q(\Phi|\hat{\Phi}^{(r)})$:

$$\hat{\Phi}^{(r+1)} = \underset{\Phi}{\operatorname{argmax}} Q(\Phi|\hat{\Phi}^{(r)}).$$

The Q function takes the same form as the likelihood function for weighted logistics regression and the maximization can be efficiently solved by the Newton-Raphson method.

4. Update $\tilde{X}$: update $\tilde{x}_{bst}$ sequentially by the Gibbs sampler with probability $\propto p(z_{btsg}|\tilde{x}_{btsg}) p(\tilde{x}_{btsg}|\tilde{\mathcal{N}}(\tilde{x}_{btsg});\hat{\Phi}^{(r+1)})$. All entries in $\tilde{X}$ are updated once.
5. Repeat steps 2–4.

Diagnosis for the number of iterations is discussed in the Supplementary Materials Section 1. In practice, we used the estimated parameters at the 200th iteration. A comparison between EM algorithm with mean field-like approximation and MCEM algorithm is included in the Supplementary Materials Section 3. The estimated parameters are comparable, though they tend to fluctuate more during the iterations in the EM algorithm with mean field-like approximation.

2.4.2 Posterior probability estimation and FDR control—To acquire an estimate of the posterior probability, we implemented a separate Gibbs sampler and fixed the model

parameters at the estimated values by the EM algorithm. The latent states were updated sequentially by

$$p(x_{\text{btsg}}|\cdot) \propto p(z_{\text{btsg}}|x_{\text{btsg}})p(x_{\text{btsg}}|\mathcal{N}(x_{\text{btsg}};\hat{\Phi})).$$

We then adapted the posterior probability based definition of FDR (Newton et al., 2001; Li et al., 2010b). The posterior local FDR $p_{\text{btsg}}=p(x_{\text{btsg}}=0|\mathbf{Z};\hat{\Phi})$ can be estimated from the Gibbs sampler. Let $p_{(s)}$ be the sorted values of p_{btsg} in ascending order. Let

$k=\max\left\{t:\frac{1}{t}\sum_{s=1}^t p_{(s)}\leq\alpha\right\}$, then we reject all the null hypotheses $H_{(s)}$, $s=1, \dots, k$. In the simulation studies, we found that the true FDR was slightly inflated in both the empirical Bayesian model (EB) not considering the complex data structure and the MRF model. This is likely caused by the fact that the null distribution of the quantiles does not strictly follow uniform distribution as it is calculated from discrete random variables. We selected $\alpha=0.01$ in both simulation studies and real data analysis. $\alpha=0.01$ was able to control the FDR under 0.05 in all our simulations. The achieved FDR for various α vs. power curves are shown in the Supplementary Materials Figure 7S.

3 Application to the mouse RNA-Seq data

We applied the MRF model to the mouse RNA-Seq data. We ran 200 iterations of the EM algorithm. Because of the randomness in the mean field-like approximation, the estimated parameters do not converge to a single point. In practice, we set the parameters equal to the estimates in the last iteration. Diagnosis for the parameter estimation is provided in the Supplementary Materials Section 1. With the parameters fixed, a separate Gibbs sampler with 20,000 total iterations and 10,000 as burn-in was implemented to estimate the posterior probabilities. The estimated MRF parameters were $\gamma=-0.21$, $\beta_s=1.18$, $\beta_{Is}=2.33$, $\beta_{sI}=0.23$, $\beta_{If}=2.51$, and $\beta_{fI}=0.25$. It is not surprising that β_s was large, as the expression levels in male and female samples are quite similar. This can be visualized in the expression profiles for some representative genes in Figure 1 and is also shown by hierarchical clustering (Supplementary Materials Figure 4S), where male and female samples tend to form the smallest cluster. In the mouse brain, layer L4 lies in between layers SgI and IgI. Based on the estimated MRF parameters, there is less similarity between SgI and IgI, as β_{sI} was much smaller compared with β_{Is} and β_{If} . The observation that β_{Is} and β_{If} were large (even larger than β_s) suggest that although the absolute gene expression levels may be different across layers, the differential expression patterns tend to be consistent. The layers represent distinct cell types and our observation suggests that there may be some common regulatory machineries across layers during neurodevelopment.

In the MRF model, following the FDR control procedure in 2.4.2, the threshold for the posterior marginal probability $p(x_{\text{btsg}}=1|\cdot)$ was set at 0.935. 7.5% of genes were identified as DE, as $28,599/(12,729\times 5\times 3\times 2)=7.5\%$. We compared the results from using the MRF model with those from edgeR. In edgeR, Benjamini-Hochberg procedure was implemented and $\alpha=0.05$. Our MRF model identified more DE genes compared with edgeR: 28,599 vs. 16,747,

with 85.5% overlap. The number of genes identified as DE was plotted against time (Figure 2).

The trend was similar among all layer-sex combinations. The number of DE genes first decreased sharply in P6–P8 then gradually increased. There are very few genes that are DE between P6 and P8. This may reflect the temporal dynamics of the gene expression in S1C, similar to an “hourglass” temporal gene expression pattern observed in the human neonatal neocortex (Pletikos et al., 2014). Female mice become sexually mature at 6 weeks after birth (P42) and males mature at 8 weeks (P56). When the mice become mature, we would expect fewer DE genes between adjacent time points as the brain development stabilizes. Except from the valley, we did not observe a decrease in DE genes as there is still robust changes in gene expression at P14. The time gap between P180 and P14 is large and there is a large number of genes that are DE between the two time points. For genes that are DE in all layer-sex combinations, we conducted Gene Ontology (GO) enrichment analysis using DAVID, which takes a list of genes as input and outputs the enriched 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. The results for the enrichment analysis is available on https://github.com/linzx06/Mouse_brain_RNAseq.

The information for the direction of changes in gene expression was not utilized in the MRF model. For genes that are DE in both male and female in all three layers, we summarized the direction of changes (Table 1). We observed that those genes tend to have the same direction of changes. The direction of changes in Table 1 were based on normalized read counts adjusted for library size. Compared with edgeR, the MRF model was able to detect more genes that have consistent changes in gene expression.

Finally, as three examples, we plotted the expression levels of *Gna12*, *Dok4* and *Cntnap1* against time (Figure 1). *Gna12* was equally expressed (EE) and the expression levels remained flat over time. Among all genes, only three genes were DE between every adjacent time points in all layer-sex combinations: *Dok4*, *Cntnap1* and *Rasgrf1*. The expression levels of *Cntnap1* and *Rasgrf1* (data not shown) both increased over time, while that of *Dok4* decreased over time. *Cntnap1* is transcribed predominantly in brain and the protein product may be the signaling subunit of contactin, enabling recruitment and activation of intracellular signaling pathways in neurons [provided by RefSeq, Jan 2009]. The protein encoded by *Rasgrf1* is a guanine nucleotide exchange factor (GEF) implicated in learning and memory in neonatal mice (Fernandez-Medarde et al., 2007). *Dok4* is involved in neuronal differentiation (Grimm et al., 2001).

4 Simulation studies

We conducted simulations to evaluate the performance of our proposed MRF model. Read counts for 500 genes, three layers, 6 time points, male and female with three replicates were

simulated. We first simulated the latent states and then the read counts. We considered two simulation settings:

Simulation setting 1: The latent state array was simulated by a Gibbs sampler. The sampler started from a random array with equal probability of being EE or DE. In each round of Gibbs sampling, every latent state was updated once sequentially according to (2). We conducted three rounds of Gibbs sampling to obtain the latent state array X . The parameter were: $\gamma = -1.1$, $\beta_{Is} = 2$ and $\beta_{sf} = 0.1$, $\beta_{If} = 2$, $\beta_f = 0.2$ and $\beta_s = 1$.

Simulation setting 2: In time point 1, all genes have 0.1 probability of being DE. The latent states evolved over time by a Hidden Markov Model: if a gene was DE at $t-1$, with probability 0.5 it will become EE at t ; to keep the number of DE genes constant over time, we randomly selected the same number of EE genes at $t-1$ to switch to DE at t . The latent states for the three layers, male and female were initially set to be the same. 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 also switched to DE, to keep the total number of DE genes constant.

The read counts were simulated from a negative binomial distribution, similarly as in Robinson and Smyth (2007). The dispersion was gene specific and was randomly sampled from gamma distribution (shape = 0.85, scale = 0.5). Across biological conditions (layers, time points and sex), the dispersion for a gene was assumed to be the same. At $t=1$, the mean read count for layer b , sex s and gene g , μ_{btsg} was simulated from uniform distribution $U[100, 300]$; if $x_{btsg}=1$, $\mu_{b(t+1)sg} = \lambda \mu_{btsg}$, where $\lambda = 0.25$ or 4 with equal probability; if $x_{btsg}=0$, $\mu_{b(t+1)sg} = \mu_{btsg}$. To avoid low read counts, if $\mu_{b(t+1)sg} < 10$, let $\mu_{b(t+1)sg} = 4 \mu_{btsg}$.

The plots for false discovery rate (FDR) vs. true positive rates are shown in Figure 3. For the MRF model, the curve was generated by thresholding the posterior marginal probabilities of DE; for the EB model, it was generated by directly implementing the local-FDR procedure on the z -score as defined in Section 2.2; and for edgeR, it was generated by thresholding the p -values. As expected, the performance of EB and edgeR was similar. In all simulation settings, our model (MRF) performed equally well or better, compared with EB and edgeR, where both methods do not consider the complex data structure. Next we compared the models in terms of sensitivity, specificity and FDR (Table 2). For the MRF and EB model, we selected $\alpha = 0.01$. For edgeR, we chose 0.05 as the cutoff for p -values. For edgeR-BH, we implemented the Benjamini-Hochberg procedure and chose $\alpha = 0.05$. MRF, EB and edgeR-BH were able to control the FDR under 0.05. Our method (MRF) achieved higher sensitivity compared with EB and edgeR-BH. Note that when the similarity decreases, i.e. from HMM (0.1) to HMM (0.5), the FDR slightly increases in our model, but not quite in the EB model. The plots for other classification measures, including the precision-recall curve, and the accuracy curve are shown in the Supplementary Materials Figures 5S and 6S, respectively. Results for other simulation settings are shown in Supplementary Materials Figures 8S, 9S, and 10S, including simulations with larger scale (10,000 genes), more DE genes (20%), and no replicates, respectively. The trends are similar to that in Figure 3.

5 Computational cost

In our previous work analyzing the human brain microarray data, we implemented an MCEM algorithm to estimate the MRF parameters. In each iteration of the MCEM algorithm, the Q function, which is the expectation over X , was approximated by Monte-Carlo sum over random samples of X . The computational cost for the MCEM algorithm is high because X has to be sampled a large number of times within each E-step. With mean field-like approximation, the computational cost is substantially reduced as the configuration $\hat{X}$ is only updated once in each EM iteration. The estimated parameters were comparable in both algorithms (data not shown). The computational time for simulation setting 1 with increasing number of genes is shown in Figure 4. The computations in Figure 4 were implemented on a dual-core CPU 2.4 GHz laptop running OS X 10.9.5 using R 3.1.0.

6 Conclusions and discussions

The statistical methods developed in this study were motivated from the analysis of mouse brain development RNA-Seq data. To efficiently utilize the information embedded in the complex data structure, we have developed a Markov Random Field-based model to capitalize on the between layers similarity, temporal dependency and the similarity between sexes. We also adapted the exact test in edgeR and the local FDR methodology to facilitate statistical inference. To estimate the model parameters, we implemented a fast and efficient EM algorithm with mean field-like approximation. Simulation results and real data analysis suggest that the proposed model improves the power to identify DE genes.

The analysis of the mouse brain RNA-Seq data by our proposed method yields biologically meaningful results. Although the absolute gene expression levels may be different across layers, they tend to change together over time. The number of DE genes first decreased sharply in P6–P8, then gradually increased, which reflects the dynamics of the development of SIC. In our approach, we did not directly model the direction of changes in gene expression. However, we found that DE genes tend to have the same direction of changes in gene expression across layers and sex. As an extension, the performance of our model may be improved by taking into account the direction of changes in gene expression.

Our proposed method is designed for gene-level read count data and the overall modeling framework is flexible to incorporate statistical tests for transcript-level evaluation as well. The coverage for the mouse brain data is not deep, ~10 million reads per sample, making the evaluation of transcript-level expression challenging. For gene-level analysis, we lose the resolution for transcripts, but gain robustness and power. In the simulation, we generated read counts for the genes, rather than raw sequencing reads. The simulation settings suffice the problem and model setups, but do not account for the full complexity of RNA-Seq experiments, including base-level errors and sequencing biases. The modeling framework can be extended to identify DE genes between sexes and DE genes between layers. The R package is available on https://github.com/linzx06/Mouse_brain_RNAseq.

7 Supplementary materials

Supplementary materials are available. Section 1: diagnosis for the EM algorithm. Section 2: hierarchical clustering of the samples. Section 3: comparison of the parameters estimated by MCEM and EM algorithm with mean field-like approximation. Section 4: more simulation results, including other classification measures (precision-recall, ROC), the achieved false discovery rate-power curves, and other simulation settings (larger scale, more DE genes and no replicates).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Matthew W. State for the financial support of the first author. All computations except that in Figure 4 were performed on the Yale University Biomedical High Performance Computing Center. This study was supported by National Institutes of Health [GM59507, CA154295, MH106934 and NS051869 (N.S.)], and National Science Foundation [DMS 1106738].

References

- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010; 11:R106. [PubMed: 20979621]
- Besag J. On the statistical analysis of dirty pictures. *J. R. Stat. Soc. B Methodol.* 1986;259–302.
- Celeux G, Forbes F, Peyrard N. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recogn.* 2003; 36:131–144.
- Chandler, D. Introduction to modern statistical mechanics. Vol. 5. Oxford University Press; Oxford, UK: 1987. p. 119-158.
- Chen M, Cho J, Zhao H. Incorporating biological pathways via a markov random field model in genome-wide association studies. *PLoS Genet.* 2011; 7:e1001353. [PubMed: 21490723]
- Efron B. Large-scale simultaneous hypothesis testing. *J. Am. Stat. Assoc.* 2004; 99:96–104.
- Fernández-Medarde A, Porteros A, De Las Rivas J, Núñez A, Fuster J, Santos E. Laser microdissection and microarray analysis of the hippocampus of ras-grf1 knockout mice reveals gene expression changes affecting signal transduction pathways related to memory and learning. *Neuroscience.* 2007; 146:272–285. [PubMed: 17321057]
- Fertuzinhos S, Li M, Kawasawa YI, Ivic V, Franjic D, Singh D, Crair M, Šestan N. Laminar and temporal expression dynamics of coding and noncoding rnas in the mouse neocortex. *Cell Rep.* 2014; 6:938–950. [PubMed: 24561256]
- Geschwind DH, Levitt P. Autism spectrum disorders: developmental disconnection syndromes. *Curr. Opin. Neurobiol.* 2007; 17:103–111. [PubMed: 17275283]
- Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from rna-seq data with biological variation. *Bioinformatics.* 2012; 28:1721–1728. [PubMed: 22563066]
- Grimm J, Sachs M, Britsch S, Di Cesare S, Schwarz-Romond T, Alitalo K, Birchmeier W. Novel p62dok family members, dok-4 and dok-5, are substrates of the c-ret receptor tyrosine kinase and mediate neuronal differentiation. *J. Cell Biol.* 2001; 154:345–354. [PubMed: 11470823]
- Huang D, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.* 2008; 4:44–57.
- Kwan KY, Šestan N, Anton E. Transcriptional co-regulation of neuronal migration and laminar identity in the neocortex. *Development.* 2012; 139:1535–1546. [PubMed: 22492350]

- Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C. Eseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*. 2013; 29:1035–1043. [PubMed: 23428641]
- Li B, Dewey CN. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*. 2011; 12:323. [PubMed: 21816040]
- Li C, Wei Z, Li H. Network-based empirical bayes methods for linear models with applications to genomic data. *J. Biopharm. Stat.* 2010a; 20:209–222. [PubMed: 20309755]
- Li H, Wei Z, Maris J. A hidden markov random field model for genome-wide association studies. *Biostatistics*. 2010b; 11:139–150. [PubMed: 19822692]
- Lin Z, Sanders SJ, Li M, Sestan N, State MW, Zhao H. A markov random field-based approach to characterizing human brain development using spatial-temporal transcriptome data. *Ann. Appl. Stat.* 2015; 9:429–451.
- McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nuc. Acids Res.* 2012; 40:4288–4297.
- Nariai N, Hirose O, Kojima K, Nagasaki M. Tigar: transcript isoform abundance estimation method with gapped alignment of rna-seq data by variational bayesian inference. *Bioinformatics*. 2013; 29:2292–2299. [PubMed: 23821651]
- Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui K-W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* 2001; 8:37–52. [PubMed: 11339905]
- Nicolae M, Mangul S, Mandoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms Mol. Biol.* 2011; 6:9. [PubMed: 21504602]
- Pletikos M, Sousa AM, Sedmak G, Meyer KA, Zhu Y, Cheng F, Li M, Kawasaki YI, Sestan N. Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron*. 2014; 81:321–332. [PubMed: 24373884]
- Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007; 23:2881–2887. [PubMed: 17881408]
- Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*. 2008; 9:321–332. [PubMed: 17728317]
- Robinson MD, McCarthy DJ, Smyth GK. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26:139–140. [PubMed: 19910308]
- Rossell D, Attolini CS-O, Kroiss M, Stöcker A. Quantifying alternative splicing from paired-end rna-sequencing data. *Ann. Appl. Stat.* 2014; 8:309. [PubMed: 24795787]
- Sestan N, State MW. The emerging biology of autism spectrum disorders. *Science*. 2012; 337:1301. [PubMed: 22984058]
- Sherman BT, Lempicki RA, Huang da W. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nuc. Acids Res.* 2009; 37:1–13.
- Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nat. Biotechnol.* 2013; 31:46–53. [PubMed: 23222703]
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010; 28:511–515. [PubMed: 20436464]
- Walsh CA, Morrow EM, Rubenstein JL. Autism and brain development. *Cell*. 2008; 135:396–400. [PubMed: 18984148]
- Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 2009; 10:57–63. [PubMed: 19015660]
- Wei Z, Li H. A markov random field model for network-based analysis of genomic data. *Bioinformatics*. 2007; 23:1537–1544. [PubMed: 17483504]
- Wei Z, Li H. A hidden spatial-temporal markov random field model for network-based analysis of time course gene expression data. *Ann. Appl. Stat.* 2008; 2:408–429.

- Zhang J. The mean field theory in em procedures for markov random fields. IEEE T. Signal Proces. 1992; 40:2570–2583.
- Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in rna sequencing data using observation weights. Nuc. Acids Res. 2014; 42:e91.

Figure 1.
The dynamics of gene expression for *Gna12*, *Dok4* and *Cntnap1*.

Figure 2.
The number of DE genes identified in every adjacent time points.

Figure 3.

The false discovery rate vs. true positive rate curves comparing the proposed MRF model, empirical bayesian (EB) and edgeR. Each plot represents the average of 100 simulations. (A) Setting 1. (B) Setting 2: HMM (0.1). (C) Setting 2: HMM (0.2). (D) Setting 2: HMM (0.5).

Figure 4.
Computational time for 10 iterations of the EM algorithm with mean field-like approximation.

Table 1

Summary for the direction of changes in gene expression.

	0	1	2	3	4	5	6
Overall	9108	7971	9163	11382	9087	7517	9417
edgeR	255	0	0	0	0	0	216
MRF	1064	1	7	0	4	4	2041

“0” column represents the counts of genes that were down-regulated in all three layers in both male and female, and “6” column represents the counts of genes that were up-regulated in all three layers in both male and female. “Overall” represents the result by pooling all the genes.

Table 2

Comparison of the three methods: MRF, EB and edgeR.

Simulation	Model	Sensitivity	Specificity	FDR
MRF	MRF	0.654(0.031)	0.997(0.001)	0.035(0.012)
	EB	0.500(0.035)	0.998(0.001)	0.037(0.011)
	edgeR	0.758(0.020)	0.960(0.002)	0.319(0.019)
	edgeR-BH	0.513(0.031)	0.998(0.000)	0.041(0.008)
HMM(0.1)	MRF	0.698(0.042)	0.998(0.001)	0.030(0.011)
	EB	0.498(0.051)	0.998(0.001)	0.035(0.010)
	edgeR	0.762(0.026)	0.960(0.002)	0.319(0.013)
	edgeR-BH	0.514(0.043)	0.998(0.000)	0.041(0.007)
HMM(0.2)	MRF	0.628(0.040)	0.997(0.001)	0.037(0.011)
	EB	0.496(0.045)	0.998(0.001)	0.035(0.010)
	edgeR	0.757(0.026)	0.960(0.002)	0.321(0.013)
	edgeR-BH	0.510(0.040)	0.998(0.001)	0.041(0.008)
HMM(0.5)	MRF	0.535(0.036)	0.997(0.001)	0.049(0.013)
	EB	0.511(0.037)	0.998(0.001)	0.038(0.010)
	edgeR	0.759(0.017)	0.960(0.002)	0.320(0.012)
	edgeR-BH	0.518(0.030)	0.998(0.000)	0.041(0.008)