The EM-algorithm for modeling Serial analysis of Gene Expression (SAGE) data

Michèle Ampe

promotor :
Prof. dr. Tomasz BURZYKOWSKI,
De heer Dirk VALKENBORG

Eindverhandeling voorgedragen tot het bekomen van de graad
Master of Science in Biostatistics

universiteit hasselt

Universiteit Hasselt
Center for Statistics

The EM Algorithm for Modeling Serial Analysis of Gene Expression (SAGE) Data

by

Michèle Ampe

Supervisors: Prof. dr. Tomasz Burzykowski and Dirk Valkenborg

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Biostatistics.

Academic year 2006-2007

Contents

1 Introduction

Serial Analysis of Gene Expression (SAGE), a technique that has been developed at Johns Hopkins University in the USA, allows the analysis of overall gene expression patterns. It is an open platform because SAGE does not require a preexisting clone, unlike microarrays. So SAGE can be used for the identification and quantification of known genes as well as new genes.
A SAGE experiment, from a statistical points of view, consists of the following 7 steps:

1. Extract a sample of mRNA fragments from a biological sample.
2. Convert the mRNA fragment into cDNA clones.
3. Generate tags by cutting 10 or 17 base long segments from a certain site of cDNA. These tags are what we call the *true tags*.
4. Apply the PCR (Polymerase Chain Reaction) procedure to boost the counts of the tags.
5. Link the tags to form long sequences.
6. Take a sample of those sequences.
7. Read off tag counts by sequencing these chosen sequences. The resulting tags are called *sequenced tags* and the resulting counts are the *observed counts*.

Note that no true tags are lost before, during or after sequencing, hence the number of sequenced tags is equal to the number of true tags. In the following sections we will assume that the true tags uniquely identify mRNA fragments that are present in the biological sample. The result of a SAGE experiment, called a SAGE library, contains the observed counts. Hence a SAGE experiment can only measure the expression levels of the tags. We can get the gene expression levels from a SAGE library by mapping the tags onto the genes.

The aspects of SAGE experiments that bias the outcomes have been studied by simulating libraries by Stollberg *et al.* (2000). The following four sources of errors are considered:

(1) sampling errors in tag selection;

(2) sequencing errors;

(3) non uniqueness of tag sequences; and

(4) non randomness of DNA sequences.

The authors have provided a maximum likelihood approach to estimate the number of unique transcripts and their frequency distribution.
In what follows, we will focus on sequencing errors. Sequencing errors have a large impact on the outcome of a SAGE experiment: non-existing tags may be introduced at low abundance

and the real abundance of the other tags may decrease.
Colinge and Feger (2001) introduced an approach to identify tags whose abundance is biased by sequencing errors. Their approach is based on a concept of neighbourhood, i.e. abundant tags can contaminate tags whose sequence is very close. They assume constant error probabilities and use matrix inversion to correct for sequencing errors.
There are also more biological approaches to the problem of sequencing errors as in Blades *et al.* (2004a,b). In Blades *et al.* (2004a), the fact that frequency distributions of tags display a regularity across cell types and species is used to

- automatically discount low counts that are not reliable for the comparison of expression levels across conditions for a specific gene;
- to transform the tag counts to a scale that provides a more reliable correlation and clustering of genome-wide expression profiles.

They state that the transformation enhances the ability to distinguish between signal and noise in SAGE data. Blades *et al.* (2004b) observed a linear relationship between the copy number of a given tag and the number of observed tags which differ from the given tag by a single base. By transforming the slope of this relationship, an estimate of the sequencing error rate can be found.
Akmaev and Wang (2004) estimated error rates based on a mathematical model that includes the PCR and sequencing error contributions. About 3.5% of Long SAGE tags (10-17 base pair tags) will inherit errors from the PCR amplification and 17.3% of the Long SAGE tags will have sequencing errors.
Beissbarth *et al.* (2004) introduced a statistical model for the propagation of sequencing errors and proposed an Expectation-Maximization (EM) algorithm to correct for the sequencing errors given a library of observed sequences and base-calling error estimates. The suggested correction method adjusts the tag counts to be closer to the true counts and the bias introduced by the sequencing errors can be partly corrected. In the article, they make use of the sequence neighbourhood of SAGE tags. This means that they assume that sequencing errors can only come from the first order neighbours tags.
First order neighbours tags are tags that differ from each other by only 1 nucleotide, e.g. AAAA and AAAC are first order neighbour tags.
The authors simulate the true tag counts by sampling from a Poisson distribution with mean $p\lambda$, with p the proportion of a tag in the library and λ a parameter for setting the size of the library. An observed tag sequence is generated from a true tag sequence using the simulated quality values (given by a base-calling program and in function of the probability of a base-calling error) of the true tag sequence as the multinomial probabilities, i.e. replacing each base with either one of the three bases with the probability specified by the sequencing quality value of that base. The counts of the observed tags are then summed to represent the observed tags. The implementation of the algorithm is done in R.

We also propose a statistical model for the propagation of sequencing errors in the case that we have multiple SAGE libraries and correct for the sequencing error through an EM algorithm by using a similar strategy as Beissbarth *et al.* (2004). We use MATLAB for the

implementation.
There are, however, some differences between our method and the one developed by Beissbrath *et al.* (2004). We assume that the true tag counts follow a multinomial distribution with parameters $\underline{\pi}$ and N, where $\underline{\pi}$ is the vector of probabilities that represent the relative expression levels of the DNA fragment and N is the number of true tags. The error estimates which we propose are partly based on the estimate given in Akmaev and Wang (2004). Another difference is that we assume that the sequencing errors are such that a tag can be misread as one of all possible tags, instead of only restricting this to the first order neighbours. Finally, in paper of Beissbarth *et al.* (2004), they work with Long SAGE sequences, while we work with sequences of four base pairs because we do not use the restriction of the first order neighbours.
In section 2, we explain the notation and the settings that we will use throughout this thesis. In section 3, we give a detailed mathematical description of the EM algorithm with the expressions for the estimates of the expression probabilities $\underline{\pi}$ and the corresponding Variance-Covariance matrix. In section 4, we simulate SAGE libraries to study the following:

- the potential gain in terms of bias when we use estimates obtained by the EM algorithm instead of the observed expression probabilities;
- the potential gain in terms of bias when we use multiple libraries instead of a single library;
- the effect of the probabilities of sequencing errors;
- the comparison of the bias using our method and using the method of Beissbarth *et al.* (2004).

The results of the simulations are given in section 5.

2 Description of the settings and notation

A tag can be represented by a l-long vector $T_j = (j_1, \ldots, j_l)$ with $j = 1, \ldots, K$ and j_m ($m = 1, \ldots, l$) can take the values 1,2,3 or 4 if the mth nucleotide in the tag is **A**, **C**, **G** or **T**, respectively. Considering tags of length of l base pairs (bp), there are $K = 4^l$ such tags theoretically possible.

We will perform experiments with single SAGE libraries. Such a library consists of observed tag counts, denoted as the observed vector $(n_1^*, \ldots, n_K^*)$, which represent the set of N sequenced tags. The corresponding set of N true tags is represented by the vector $(n_1, \ldots, n_K)$. Note that since no true tags are lost before, during or after sequencing, we have that the number of sequenced tags is equal to the number of true tags.
In the experiments, the distribution of the true tags in the samples is the object of interest. The distribution is represented by the vector of probabilities $\underline{\pi} = (\pi_1, \ldots, \pi_K)$, with the property that $\sum_{j=1}^{K} \pi_j = 1$. The probability π_j can be interpreted as a representation of the relative expression level of the DNA fragment corresponding to the tag T_j. We will refer to the vector $\underline{\pi}$ as the true expression probabilities.

As we will also work with multiple SAGE libraries, we need to extend the notation above. We will consider g independent libraries $(L_1, \ldots, L_g)$.
Let N_i be the number of true and sequenced tags in the ith library L_i. The set of N_i true tags will be represented by the vector of counts $(n_{i1}, \ldots, n_{iK})$, where the index i ($i = 1, \ldots, g$) corresponds to the ith library L_i. The set of N_i sequenced tags will be represented by the vector of counts $(n_{i1}^*, \ldots, n_{iK}^*)$. These counts are observed.
So a sample of observed counts of g libraries is denoted as $(\underline{n}_1^*, \ldots, \underline{n}_K^*)$, with $\underline{n}_j^* = (n_{1j}, \ldots, n_{gj})$, which represents $\underline{N}$ sequenced tags with $\underline{N} = (N_1, \ldots, N_g)$.
In the experiments with multiple libraries, the distribution of the true tags is again the object of interest. We will assume that the distribution parameter $\underline{\pi} = (\pi_1, \ldots, \pi_K)$ is the same across the g libraries, since we will consider them to be generated under similar conditions.

During the sequencing process, the DNA sequence of each of the sampled true tags can be misread. The probability of a sequencing error will be assumed to be constant for each nucleotide and independent of the position in the tag. Let ω_{jk} be the probability of the change of nucleotide j into nucleotide k during the sequence process, where j and k equal to 1,2,3 or 4 for nucleotides **A**, **C**, **G** or **T**, respectively. We will consider the following 4×4 matrix Ω that contains the probabilities of sequencing errors:

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & \omega_{23} & \omega_{24} \\ \omega_{31} & \omega_{32} & \omega_{33} & \omega_{34} \\ \omega_{41} & \omega_{42} & \omega_{43} & \omega_{44} \end{pmatrix},$$

with $\sum_{k=1}^{4} \omega_{jk} = 1$. We assume that the matrix Ω is the same across the g libraries. In the following, we will denote the matrix Ω as the transition matrix.
One can consider the most general form of the matrix Ω where the probability of change of

a true tag T_r into a sequenced tag T_s depends on the nucleotides in the sequences. Then the probability is equal to

$$P(T_r \rightarrow T_s|\Omega) \equiv \varphi_{rs} = \prod_{m=1}^{l} \omega_{r_m s_m}.$$

The probabilities φ_{rs} for all pairs of true and sequenced tags r and s $(r, s = 1, \ldots, K)$ are stored in the $K \times K$ matrix Φ. The matrix Φ is constructed using the Kronecker tensor product $\bigotimes$ and the transition matrix Ω :

$$\Phi = \bigotimes_{i=1}^{l} \Omega. \tag{1}$$

In Beissbarth *et al.* (2004), the matrix Φ has non-zero probabilities on the diagonal and for the first order neighbours. We will call this matrix Φ_B. Hence they assume that the sequencing errors can only come from the first order neighbours. We will assume that a tag can be misread as one of all possible tags and so the matrix Φ will only contain non-zero probabilities.

3 Expectation-Maximization Algorithm for multiple libraries

The EM algorithm is an iterative procedure for the computation of maximum likelihood estimates in situations where, beside the fact that some additional data is missing, maximum likelihood estimation would be straightforward. In our context, the observed tag counts are the incomplete data and the true tag counts are the complete data. The incomplete data is an observable 'function' of the complete data. The problem of solving the likelihood equation of the incomplete data is tackled by proceeding iteratively in terms of the log-likelihood function of the complete data. All that is needed is the specification of the complete data and its conditional expectation given the observed incomplete data, which will be used in the E-step. The details of the mathematical formulation of the EM algorithm is given in this section.

The probability of getting a sample $(\underline{n}_1, \ldots, \underline{n}_K)$ of $\underline{N}$ true tags is a multinomial with parameters $\underline{N}$ and $\underline{\pi}$:

$$\begin{aligned} P(\underline{n}_1, \ldots, \underline{n}_K | \underline{\pi}, \underline{N}) &= \prod_{i=1}^{g} P(n_{i1}, \ldots, n_{iK} | \underline{\pi}, N_i) \\ &= \prod_{i=1}^{g} \left(\frac{N_i!}{\prod_{j=1}^{K} (n_{ij}!)} \prod_{j=1}^{K} \pi_j^{n_{ij}} \right) \\ &= \frac{\prod_{i=1}^{g} N_i!}{\prod_{i=1}^{g} \prod_{j=1}^{K} (n_{ij}!)} \prod_{i=1}^{g} \prod_{j=1}^{K} \pi_j^{n_{ij}} \\ &= \frac{\prod_{i=1}^{g} N_i!}{\prod_{i=1}^{g} \prod_{j=1}^{K} (n_{ij}!)} \prod_{j=1}^{K} \pi_j^{\sum_{i=1}^{g} n_{ij}}, \end{aligned} \tag{2}$$

where $\underline{N} = (N_1, \ldots, N_g)$ with N_i the number of true tags in the ith library L_i.

The sequencing procedure for n_{ij} true tags T_{ij} has as outcome a vector of counts $(n'_{ij1}, \ldots, n'_{ijK})$ for all possible sequenced tags with $0 \leq n'_{ijk}$ and $\sum_{k=1}^{K} n'_{ijk} = n_{ij}$ $(i = 1, \ldots, g$ and $j = 1, \ldots, K)$.
The observed sequenced tag counts $(\underline{n}^*_1, \ldots, \underline{n}^*_K)$ are functions of the counts $\underline{n}'_{jk}$ because

$n^*_{ik} = \sum_{j=1}^{K} n'_{ijk}$. Moreover, we have that:

$$\begin{aligned}
P(\underline{n}'_{11}, \ldots, \underline{n}'_{KK} | \underline{\pi}, \Omega, \underline{N}) &= \prod_{i=1}^{g} P(n'_{i11}, \ldots, n'_{iKK} | \underline{\pi}, \Omega, N_i) \\
&= \prod_{i=1}^{g} P(n_{i1}, \ldots, n_{iK}) P(n'_{i11}, \ldots, n'_{iKK} | n_{i1}, \ldots, n_{iK}) \\
&= \prod_{i=1}^{g} \left[\left\{ \frac{N_i!}{\prod_{j=1}^{K}(n_{ij}!)} \prod_{j=1}^{K} \pi_j^{n_{ij}} \right\} \prod_{j=1}^{K} \left\{ \frac{(n_{ij}!)}{\prod_{k=1}^{K}(n'_{ijk}!)} \prod_{k=1}^{K} \varphi_{jk}^{n'_{ijk}} \right\} \right] \\
&= \prod_{i=1}^{g} \left\{ \frac{N_i!}{\prod_{j=1}^{K}\prod_{k=1}^{K}(n'_{ijk}!)} \prod_{j=1}^{K}\prod_{k=1}^{K} (\pi_j \varphi_{jk})^{n'_{ijk}} \right\} \\
&= \frac{\prod_{i=1}^{g}(N_i!)}{\prod_{i=1}^{g}\prod_{j=1}^{K}\prod_{k=1}^{K}(n'_{ijk}!)} \prod_{j=1}^{K}\prod_{k=1}^{K} (\pi_j \varphi_{jk})^{\sum_{i=1}^{g} n'_{ijk}}. \qquad (3)
\end{aligned}$$

Expression (3) specifies a multinomial distribution since

$$\sum_{i=1}^{g}\sum_{j=1}^{K}\sum_{k=1}^{K} n'_{ijk} = \sum_{i=1}^{g}\sum_{j=1}^{K} n_{ij} = \sum_{i=1}^{g} N_i =: N'$$

and

$$\sum_{j=1}^{K}\sum_{k=1}^{K} \pi_j \varphi_{jk} = 1.$$

Hence, it follows that

$$E[\underline{n}'_{jk}] = N' \pi_j \varphi_{jk} \qquad (4)$$

and

$$E[\underline{n}'_{jk} | \underline{n}^*_k] = \sum_{i=1}^{g} \frac{\pi_j \varphi_{jk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} n^*_{ik}. \qquad (5)$$

The probabilities of the observed counts $(\underline{n}^*_1, \ldots, \underline{n}^*_K)$ for the sets of $\underline{N}$ sequenced tags $(\sum_{i=1}^{g}\sum_{j=1}^{K} n^*_{ij} = \sum_{i=1}^{g} N_i)$ can be written as follows by using (2) and (3):

$$\begin{aligned}
P(\underline{n}^*_1, \ldots, \underline{n}^*_K | \underline{\pi}, \Omega, \underline{N}) &= \sum_{S^*} \{ P(\underline{n}_1, \ldots, \underline{n}_K) P(\underline{n}^*_1, \ldots, \underline{n}^*_K | \underline{n}_1, \ldots, \underline{n}_K) \} \qquad (6) \\
&= \sum_{S^*} \left[\frac{\prod_{i=1}^{g} N_i!}{\prod_{i=1}^{g}\prod_{j=1}^{K}(n_{ij}!)} \prod_{j=1}^{K} \pi_j^{\sum_{i=1}^{g} n_{ij}} \sum_{S} \left\{ \prod_{i=1}^{g}\prod_{j=1}^{K} \frac{(n_{ij}!)}{\prod_{k=1}^{K}(n'_{ijk}!)} \prod_{k=1}^{K} \varphi_{jk}^{n'_{ijk}} \right\} \right].
\end{aligned}$$

The first summation in (6) is over the set $S = \left\{ (\underline{n}_1, \ldots, \underline{n}_K) : 0 \le n_{ij}, \sum_{j=1}^{K} n_{ij} = N_i \right\}$ of all possible combinations of counts of $\underline{N}$ true tags. The second summation is over the set

$$S = \{ (\underline{n}'_{11}, \ldots, \underline{n}'_{1K}, \ldots, \underline{n}'_{K1}, \ldots, \underline{n}'_{KK}) : 0 \le n^*_{ij}, \sum_{j=1}^{K} n'_{ijk} = n^*_{ik}, \sum_{k=1}^{K} n'_{ijk} = n_{ij} \}$$

of all possible series of counts $(\underline{n}'_{11}, \ldots, \underline{n}'_{KK})$ of sequenced tags arising from the sets counts $(\underline{n}^*_1, \ldots, \underline{n}^*_K)$ of $\underline{N}$ true tags.
Now looking at the likelihood function for the g libraries, we have the following expression

$$L(\underline{n}^*_1, \ldots, \underline{n}^*_g | \underline{\pi}, \Omega, \underline{N}) = \prod_{i=1}^{g} P(\underline{n}^*_i | \underline{\pi}, \Omega, N_i) = \prod_{i=1}^{g} P(n^*_{i1}, \ldots, n^*_{iK} | \underline{\pi}, \Omega, N_i) \tag{7}$$

The relevant part of the logarithm of (7) using (3) is

$$l(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{g} \sum_{j=1}^{K} \sum_{k=1}^{K} n'_{ijk} \log(\pi_j) = \sum_{i=1}^{g} \sum_{j=1}^{K} n_{ij} \log(\pi_j). \tag{8}$$

The loglikelihood expression in (8) depends on the probabilities $\pi_1, \ldots, \pi_K$, which we will estimate using the EM-algorithm.

The E-step:

In the Expectation step, the conditional expected value of (8) is computed, conditional on the observed counts $(n^*_{i1}, \ldots, n^*_{iK})$ and the current values of the parameters $(\pi_1, \ldots, \pi_K)$:

$$Q(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{g} \sum_{j=1}^{K} E[n_{ij} | n^*_{i1}, \ldots, n^*_{iK}, \pi^{(0)}_1, \ldots, \pi^{(0)}_K] \log(\pi_j). \tag{9}$$

Using (5) we have that

$$\begin{aligned} E[n_{ij} | n^*_{i1}, \ldots, n^*_{iK}, \pi^{(0)}_1, \ldots, \pi^{(0)}_K] &= E[\sum_{k=1}^{K} n'_{ijk} | n^*_{i1}, \ldots, n^*_{iK}, \pi^{(0)}_1, \ldots, \pi^{(0)}_K] \\ &= \sum_{k=1}^{K} E[n'_{ijk} | n^*_{i1}, \ldots, n^*_{iK}, \pi^{(0)}_1, \ldots, \pi^{(0)}_K] \\ &= \sum_{k=1}^{K} \sum_{i=1}^{g} \frac{\pi^{(0)}_j \varphi_{jk}}{\sum_{l=1}^{K} \pi^{(0)}_l \varphi_{lk}} n^*_{ik}. \end{aligned} \tag{10}$$

So we have that

$$Q(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{g} \sum_{j=1}^{K} \left\{ \log(\pi_j) \sum_{k=1}^{K} \frac{\pi^{(0)}_j \varphi_{jk}}{\sum_{l=1}^{K} \pi^{(0)}_l \varphi_{lk}} n^*_{ik} \right\} \tag{11}$$

The M-step:

In the Maximization step, (11) is maximized over $(\pi_1, \ldots, \pi_K)$, while taking into account that $\sum_{j=1}^{K} \pi_j = 1$. If we take $\pi_K = 1 - (\pi_1 + \ldots + \pi_{K-1})$, then we need to solve the following

set of equations simultaneously for $j = 1, \ldots, K-1$:

$$\left\{\sum_{k=1}^{K} \frac{\pi_j^{(0)}\varphi_{jk}}{\sum_{l=1}^{K}\pi_l^{(0)}\varphi_{lk}} \sum_{i=1}^{g} n_{ik}^*\right\} \log \pi_j = \left\{\sum_{k=1}^{K} \frac{\pi_K^{(0)}\varphi_{Kk}}{\sum_{l=1}^{K}\pi_l^{(0)}\varphi_{lk}} \sum_{i=1}^{g} n_{ik}^*\right\} \log(1 - (\pi_1 + \ldots + \pi_{K-1})). \tag{12}$$

To give the solution of (12), we will first simplify some of the expressions given above. First, one can rewrite (11) as

$$Q(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{g} \left[\left(\sum_{j=1}^{K-1} C_{ij} \log(\pi_j) \right) + C_{iK} \log(1 - (\pi_1 + \ldots + \pi_{K-1})) \right], \tag{13}$$

where

$$C_{ij} = \sum_{k=1}^{K} \frac{\pi_j^{(0)}\varphi_{jk}}{\sum_{l=1}^{K}\pi_l^{(0)}\varphi_{lk}} n_{ik}^*, \qquad \text{with } i = 1, \ldots, g \text{ and } j = 1, \ldots, K. \tag{14}$$

To maximize (13) over $(\pi_1, \ldots, \pi_K)$, we first set the derivative of Q over π_j for all $j \neq K$ equal to zero:

$$\frac{\partial Q}{\partial \pi_j} = \sum_{i=1}^{g} \left[\frac{C_{ij}}{\pi_j} - \frac{C_{iK}}{1 - (\pi_1 + \ldots + \pi_{K-1})} \right] = 0, \ \forall\, j \neq K,$$

and solve the obtained set of equations to derive the estimates of $(\pi_1, \ldots, \pi_K)$.
Hence we have that

$$\begin{aligned}
&\begin{cases} \sum_{i=1}^{g} [C_{i1} - (C_{i1} + C_{iK})\pi_1 - \ldots - C_{i1}\pi_{K-1}] = 0 \\ \vdots \\ \sum_{i=1}^{g} [C_{iK-1} - C_{iK-1}\pi_1 - \ldots - (C_{iK-1} + C_{iK})\pi_{K-1}] = 0 \end{cases} \\
\Leftrightarrow &\begin{cases} \sum_{i=1}^{g} [(C_{i1} + C_{iK})\pi_1 + \ldots + C_{i1}\pi_{K-1}] = \sum_{i=1}^{g} C_{i1} \\ \vdots \\ \sum_{i=1}^{g} [C_{iK-1}\pi_1 + \ldots + (C_{iK-1} + C_{iK})\pi_{K-1}] = \sum_{i=1}^{g} C_{iK-1} \end{cases} \\
\Leftrightarrow &\begin{bmatrix} \sum_{i=1}^{g}(C_{i1} + C_{iK}) & \ldots & \sum_{i=1}^{g} C_{i1} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{g} C_{iK-1} & \ldots & \sum_{i=1}^{g}(C_{iK-1} + C_{iK}) \end{bmatrix} \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_{K-1} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{g} C_{i1} \\ \vdots \\ \sum_{i=1}^{g} C_{iK-1} \end{bmatrix},
\end{aligned}$$

from which we obtain the estimate for each π_j

$$\pi_j = \frac{\sum_{i=1}^{g} C_{ij}}{\sum_{i=1}^{g} C_{i1} + \ldots + \sum_{i=1}^{g} C_{iK}}, \ \forall\, j = 1, \ldots, K. \tag{15}$$

Next, we will estimate the Variance-Covariance matrix corresponding to the estimates given in equation (15) for $j = 1, \ldots, K$. The estimation of the Variance-Covariance matrix is based on the information matrix.
To estimate the information matrix $I(\underline{\pi}, \underline{n}^*)$, we will use the following formula:

$$\begin{aligned} I(\underline{\pi}, \underline{n}^*) &= \mathcal{I}_c(\underline{\pi}, \underline{n}^*) - \mathcal{I}_m(\underline{\pi}, \underline{n}^*) && (16) \\ &= E_{\underline{\pi}}\left[I_c(\underline{\pi}, \underline{n})|\underline{n}^*\right] - \mathcal{I}_m(\underline{\pi}, \underline{n}^*), && (17) \end{aligned}$$

where $I_c(\underline{\pi}, \underline{n}) = -\partial^2 \log L_c(\underline{\pi})/\partial\underline{\pi}\partial\underline{\pi}^T$ with $L_c(\underline{\pi})$ the complete-data likelihood function.
$\mathcal{I}_m(\underline{\pi}, \underline{n}^*)$ can be viewed as the "missing information" and hence equation (17) can be interpreted as: the observed informations equals the conditional expected complete information minus the missing information.
The missing information matrix $\mathcal{I}_m(\underline{\pi}, \underline{n}^*)$ can be expressed in the form (Louis (1982))

$$\begin{aligned} \mathcal{I}_m(\underline{\pi}, \underline{n}^*) &= Cov_{\underline{\pi}}\left[S_c(\underline{n}, \underline{\pi})|\underline{n}^*\right] && (18) \\ &= E_{\underline{\pi}}\left[S_c(\underline{n}, \underline{\pi})S_c(\underline{n}, \underline{\pi})^T|\underline{n}^*\right] - S(\underline{n}^*, \underline{\pi})S(\underline{n}^*, \underline{\pi})^T && (19) \end{aligned}$$

$$\begin{aligned} \Rightarrow I(\hat{\underline{\pi}}, \underline{n}^*) &= \mathcal{I}_c(\hat{\underline{\pi}}, \underline{n}^*) - \mathcal{I}_m(\hat{\underline{\pi}}, \underline{n}^*) && (20) \\ &= E_{\underline{\pi}}\left[I_c(\hat{\underline{\pi}}, \underline{n})|\underline{n}^*\right] - Cov_{\underline{\pi}}\left[S_c(\underline{n}, \underline{\pi})|\underline{n}^*\right]_{\underline{\pi}=\hat{\underline{\pi}}}. && (21) \end{aligned}$$

To obtain the Variance-Covariance matrix of $\hat{\underline{\pi}}$, we need to invert the information matrix, i.e.

$$Cov(\hat{\underline{\pi}}) = (I(\hat{\underline{\pi}}, \underline{n}^*))^{-1} = (\mathcal{I}_c(\hat{\underline{\pi}}, \underline{n}^*) - \mathcal{I}_m(\hat{\underline{\pi}}, \underline{n}^*))^{-1}.$$

Calculation of $\mathcal{I}_c(\underline{\pi}, \underline{n}^*)$:

$$\begin{aligned} l(\pi_1, \ldots, \pi_K) &= \sum_{i=1}^{g}\sum_{j=1}^{K} n_{ij}\log(\pi_j) && (22) \\ &= \sum_{i=1}^{g}\left[\sum_{j=1}^{K-1} n_{ij}\log(\pi_j) + n_{iK}\log(1 - \pi_1 - \ldots - \pi_{K-1})\right], && (23) \end{aligned}$$

taking into account that $\sum_{j=1}^{K}\pi_j$ and thus $\pi_K = 1 - \pi_1 - \ldots - \pi_{K-1}$.
The elements of the score matrix $S_c(\underline{n}, \underline{\pi}) = \frac{\partial l(\pi_1, \ldots, \pi_K)}{\partial \underline{\pi}}$ are of the form

$$S_{j,c}(\underline{n}, \underline{\pi}) = \sum_{i=1}^{g}\left\{\frac{n_{ij}}{\pi_j} - \frac{n_{iK}}{1 - \pi_1 - \ldots - \pi_{K-1}}\right\} \qquad \forall j \neq K.$$

We calculate the matrix $I_c(\underline{\pi}, \underline{n})$ by using the score matrix, i.e.

$$I_c(\underline{\pi}, \underline{n}) = -\frac{\partial S_c(\underline{n}, \underline{\pi})}{\partial \underline{\pi}}.$$

So $I_c(\underline{\pi}, \underline{n})$ will be a $(K-1) \times (K-1)$-matrix.

diagonal elements of $I_c(\underline{\pi}, \underline{n})$:

$$-\frac{\partial S_{j,c}(\underline{n}, \underline{\pi})}{\partial \pi_j} = \sum_{i=1}^{g} \left\{ \frac{n_{ij}}{\pi_j^2} + \frac{n_{iK}}{(1-\pi_1-\ldots-\pi_{K-1})^2} \right\} \tag{24}$$

off-diagonal elements of $I_c(\underline{\pi}, \underline{n})$:

$$-\frac{\partial S_{j,c}(\underline{n}, \underline{\pi})}{\partial \pi_k} = \sum_{i=1}^{g} \frac{n_{iK}}{(1-\pi_1-\ldots-\pi_{K-1})^2} \tag{25}$$

Now we can calculate the information matrix $\mathcal{I}_c(\underline{\pi}, \underline{n}^*) = E_{\underline{\pi}}\left[I_c(\underline{\pi}, \underline{n}) | \underline{n}^*\right]$

diagonal elements of $\mathcal{I}_c(\underline{\pi}, \underline{n}^*)$:

$$\begin{aligned}
\mathcal{I}_{jj,c}(\underline{\pi}, \underline{n}^*) &= \frac{\sum_{i=1}^{g} E_{\underline{\pi}}\left[n_{ij} | \underline{n}_i^*\right]}{\pi_j^2} + \frac{\sum_{i=1}^{g} E_{\underline{\pi}}\left[n_{iK} | \underline{n}_i^*\right]}{(1-\pi_1-\ldots-\pi_{K-1})^2} \\
&= \frac{\sum_{i=1}^{g} \sum_{k=1}^{K} \frac{\pi_j \varphi_{jk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} n_{ik}^*}{\pi_j^2} + \frac{\sum_{i=1}^{g} \sum_{k=1}^{K} \frac{\pi_K \varphi_{Kk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} n_{ik}^*}{(1-\pi_1-\ldots-\pi_{K-1})^2} \\
&= \frac{\sum_{i=1}^{g} C_{ij}}{\pi_j^2} + \frac{\sum_{i=1}^{g} C_{iK}}{(1-\pi_1-\ldots-\pi_{K-1})^2}
\end{aligned} \tag{26}$$

off-diagonal elements of $\mathcal{I}_c(\underline{\pi}, \underline{n}^*)$:

$$\begin{aligned}
\mathcal{I}_{jj',c}(\underline{\pi}, \underline{n}^*) &= \frac{\sum_{i=1}^{g} E_{\underline{\pi}}\left[n_{iK} | \underline{n}_i^*\right]}{(1-\pi_1-\ldots-\pi_{K-1})^2} \\
&= \frac{\sum_{i=1}^{g} C_{iK}}{(1-\pi_1-\ldots-\pi_{K-1})^2}
\end{aligned} \tag{27}$$

Calculation of $\mathcal{I}_m(\underline{\pi}, \underline{n}^*)$:

$$\mathcal{I}_m(\underline{\pi}, \underline{n}^*) = Cov_{\underline{\pi}}\left[S_c(\underline{n}, \underline{\pi}) | \underline{n}^*\right]$$

diagonal elements:

$$\begin{aligned}
\mathcal{I}_{jj,m}(\underline{\pi}, \underline{n}^*) &= Var_{\underline{\pi}}\left[S_{j,c}(\underline{n}, \underline{\pi}) | \underline{n}^*\right] \\
&= Var_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{ij}}{\pi_j} - \frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}} | \underline{n}^*\right] \qquad (28) \\
&= \frac{\sum_{i=1}^{g} Var\left[n_{ij} | \underline{n}_i^*\right]}{\pi_j^2} + \frac{\sum_{i=1}^{g} Var\left[n_{iK} | \underline{n}_i^*\right]}{(1-\pi_1-\ldots-\pi_{K-1})^2} \\
&= F_1 + F_2 \qquad (29)
\end{aligned}$$

Next we calculate the expressions F_1 and F_2.

$$\begin{aligned} Var[n_{ij}|\underline{n}_i^*] &= Var\left[\sum_{k=1}^{K} n'_{ijk}|\underline{n}_i^*\right] \\ &= \sum_{k=1}^{K} Var\left[n'_{ijk}|\underline{n}_i^*\right] \\ &= \sum_{k=1}^{K}\left\{n_{ik}^*\left(\frac{\pi_j\varphi_{jk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\left(1-\frac{\pi_j\varphi_{jk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\right\} \end{aligned} \tag{30}$$

So, we have that

$$F_1 = \sum_{i=1}^{g}\sum_{k=1}^{K}\left\{n_{ik}^*\left(\frac{\pi_j\varphi_{jk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\left(1-\frac{\pi_j\varphi_{jk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\right\}/\pi_j^2 \tag{31}$$

and

$$F_2 = \frac{\sum_{i=1}^{g}\sum_{k=1}^{K}\left\{n_{ik}^*\left(\frac{\pi_K\varphi_{Kk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\left(1-\frac{\pi_K\varphi_{Kk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\right\}}{(1-\pi_1-\ldots-\pi_{k-1})^2}. \tag{32}$$

Hence the diagonal elements of the matrix $\mathcal{I}_m(\underline{\pi},\underline{n}^*)$ are of the following form:

$$\begin{aligned} \mathcal{I}_{jj,m}(\underline{\pi},\underline{n}^*) &= \frac{\sum_{i=1}^{g}\sum_{k=1}^{K}\left\{n_{ik}^*\left(\frac{\pi_j\varphi_{jk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\left(1-\frac{\pi_j\varphi_{jk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\right\}}{\pi_j^2} \\ &\quad + \frac{\sum_{i=1}^{g}\sum_{k=1}^{K}\left\{n_{ik}^*\left(\frac{\pi_K\varphi_{Kk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\left(1-\frac{\pi_K\varphi_{Kk}}{\sum_{l=1}^{K}\pi_l\varphi_{lk}}\right)\right\}}{(1-\pi_1-\ldots-\pi_{k-1})^2}. \end{aligned} \tag{33}$$

off-diagonal elements:

$$\begin{aligned} \mathcal{I}_{j,j+1,m}(\underline{\pi},\underline{n}^*) &= Cov_{\underline{\pi}}\left[S_{j,c}(\underline{n},\underline{\pi}), S_{j+1,c}(\underline{n},\underline{\pi})|\underline{n}^*\right] \\ &= Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{ij}}{\pi_j} - \frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}, \frac{\sum_{i=1}^{g} n_{i,j+1}}{\pi_{j+1}} - \frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}|\underline{n}^*\right] \\ &= Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{ij}}{\pi_j}, \frac{\sum_{i=1}^{g} n_{i,j+1}}{\pi_{j+1}} - \frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}|\underline{n}^*\right] \\ &\quad -Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}, \frac{\sum_{i=1}^{g} n_{i,j+1}}{\pi_{j+1}} - \frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}|\underline{n}^*\right] \\ &= Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{ij}}{\pi_j}, \frac{\sum_{i=1}^{g} n_{i,j+1}}{\pi_{j+1}}|\underline{n}^*\right] - Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}, \frac{\sum_{i=1}^{g} n_{i,j+1}}{\pi_{j+1}}|\underline{n}^*\right] \\ &\quad -Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{ij}}{\pi_j}, \frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}|\underline{n}^*\right] + Var_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{iK}}{1-\pi_1-\ldots-\pi_{K-1}}|\underline{n}^*\right] \\ &= F_3 - F_4 - F_5 + F_2 \end{aligned}$$

Using the fact that the g libraries are independent,

$$\begin{aligned}
F_3 &= Cov_{\underline{\pi}}\left[\frac{\sum_{i=1}^{g} n_{ij}}{\pi_j}, \frac{\sum_{i=1}^{g} n_{i,j+1}}{\pi_{j+1}} | \underline{n}^*\right] = \sum_{i_1=1}^{g}\sum_{i_2=1}^{g} Cov_{\underline{\pi}}\left[\frac{n_{i_1 j}}{\pi_j}, \frac{n_{i_2,j+1}}{\pi_{j+1}} | \underline{n}^*\right] \\
&= \sum_{i=1}^{g} Cov_{\underline{\pi}}\left[\frac{n_{ij}}{\pi_j}, \frac{n_{i,j+1}}{\pi_{j+1}} | \underline{n}_i^*\right] = \sum_{i=1}^{g} Cov_{\underline{\pi}}\left[\frac{\sum_{k=1}^{K} n'_{ijk}}{\pi_j}, \frac{\sum_{k=1}^{K} n'_{i,j+1,k}}{\pi_{j+1}} | \underline{n}_i^*\right] \\
&= \frac{1}{\pi_j \pi_{j+1}} \sum_{i=1}^{g}\sum_{k_1}\sum_{k_2} Cov_{\underline{\pi}}\left[n'_{ijk_1}, n'_{i,j+1,k_2} | \underline{n}_i^*\right] \\
&= \frac{1}{\pi_j \pi_{j+1}} \sum_{i=1}^{g}\sum_{k=1}^{K} Cov_{\underline{\pi}}\left[n'_{ijk}, n'_{i,j+1,k} | \underline{n}_i^*\right] \\
&= -\frac{1}{\pi_j \pi_{j+1}} \sum_{i=1}^{g}\sum_{k=1}^{K} \left(n_{ik}^* \frac{\pi_j \varphi_{jk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \frac{\pi_{j+1}\varphi_{j+1,k}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right). \qquad (34)
\end{aligned}$$

The expressions for F_4 and F_5 are similar to the expression (34):

$$F_4 = -\frac{1}{\pi_j(1-\pi_1-\ldots-\pi_{K-1})} \sum_{i=1}^{g}\sum_{k=1}^{K} \left\{ n_{ik}^* \frac{\pi_j \varphi_{jk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \frac{\pi_K \varphi_{K,k}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right\} \qquad (35)$$

$$F_5 = -\frac{1}{(1-\pi_1-\ldots-\pi_{K-1})\pi_{j+1}} \sum_{i=1}^{g}\sum_{k=1}^{K} \left\{ n_{ik}^* \frac{\pi_K \varphi_{Kk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \frac{\pi_{j+1}\varphi_{j+1,k}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right\} \qquad (36)$$

Hence, using equations (34), (35),(36) and (32), we have that

$$\begin{aligned}
\mathcal{I}_{j,j+1,m}(\underline{\pi}, \underline{n}^*) &= -\frac{1}{\pi_j \pi_{j+1}} \sum_{i=1}^{g}\sum_{k=1}^{K} \left(n_{ik}^* \frac{\pi_j \varphi_{jk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \frac{\pi_{j+1}\varphi_{j+1,k}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right) \\
&\quad + \frac{1}{\pi_j(1-\pi_1-\ldots-\pi_{K-1})} \sum_{i=1}^{g}\sum_{k=1}^{K} \left\{ n_{ik}^* \frac{\pi_j \varphi_{jk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \frac{\pi_K \varphi_{K,k}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right\} \\
&\quad + \frac{1}{(1-\pi_1-\ldots-\pi_{K-1})\pi_{j+1}} \sum_{i=1}^{g}\sum_{k=1}^{K} \left\{ n_{ik}^* \frac{\pi_K \varphi_{Kk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \frac{\pi_{j+1}\varphi_{j+1,k}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right\} \\
&\quad + \frac{\sum_{i=1}^{g}\sum_{k=1}^{K} \left\{ n_{ik}^* \left(\frac{\pi_K \varphi_{Kk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right) \left(1 - \frac{\pi_K \varphi_{Kk}}{\sum_{l=1}^{K} \pi_l \varphi_{lk}} \right) \right\}}{(1-\pi_1-\ldots-\pi_{k-1})^2}. \qquad (37)
\end{aligned}$$

4 Simulations

In this section, we give the description of the settings for the simulations. We will perform 1000 simulations in the case that there is only a single library available and 1000 simulations in the case that there are multiple (20) libraries available. We will generate 21000 libraries containing 2000 tags under six different conditions: three different true probabilities of expression $\underline{\pi}$ and two different transition matrices Ω. The parameter $\underline{\pi} = (\pi_1, \ldots, \pi_K)$ will be estimated using the EM Algorithm as explained in Section 3. The number of iterations that we use for the EM algorithm is the same as in Beissbarth *et al.* (2004), namely 50 iterations. We will also use the observed tag counts as starting values for the EM algorithm.
The goals of our simulations are:

1. the potential gain in terms of bias comparing the estimates obtained by using the EM algorithm with the estimates when there is no correction for the sequencing errors, i.e., the observed expression probabilities ;

2. the potential gain in terms of bias resulting from the use of 20 libraries as compared with the estimates obtained using only one library;

3. the gain in terms of bias when the true transition matrix is known;

4. the gain in terms of bias comparing smaller sequencing errors with larger sequencing errors;

5. the comparison of the bias of the estimates resulting from the use of Φ with the estimates obtained by using Φ_B.

4.1 The true probabilities of expression of a tag

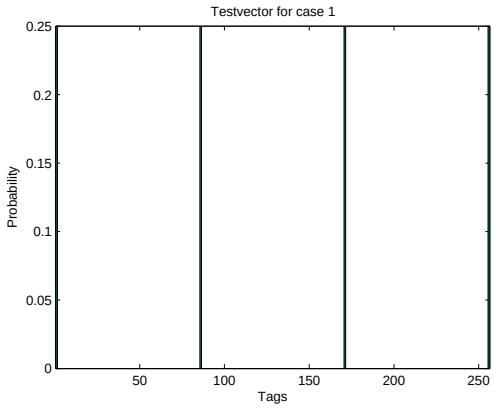
The three different true probabilities of expression that we will use, are displayed in Table 1. The plots of the true probabilities of expression are shown in Figure 1.

$\underline{\pi}_1$	Tag	'AAAA'	'CCCC'	'GGGG'	'TTTT'
	Percentage	25%	25%	25%	25%
$\underline{\pi}_2$	Tag	'AAAA'	'AAAC'	'CCCC'	'CCCG'
	Percentage	12.5%	12.5%	12.5%	12.5%
	Tags	'GGGG'	'GGGT'	'TTTA'	'TTTT'
	Percentage	12.5%	12.5%	12.5%	12.5%
$\underline{\pi}_3$	Tag	'AAAA'	'AAAC'	'CCCC'	'CCCG'
	Percentage	20%	5%	20%	5%
	Tag	'GGGG'	'GGGT'	'TTTA'	'TTTT'
	Percentage	20%	5%	5%	20%

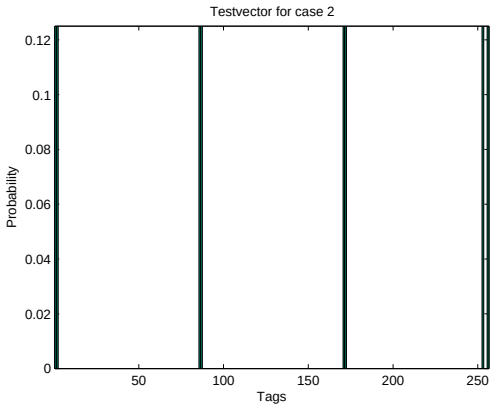
Table 1: The three different true probabilities of expression

The ordering of the tags is such a that 'AAAA' has tag number 1, 'AAAC' tag number

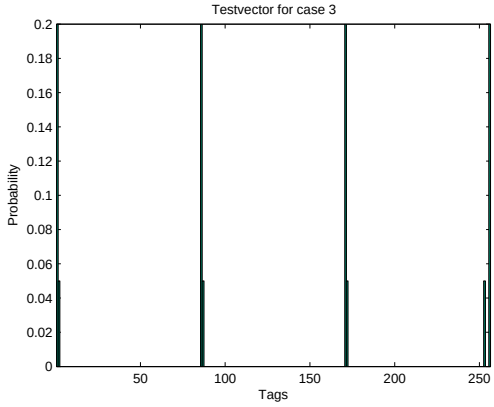
2, 'CCCC' tag number 86, 'CCCG' tag number 87, 'GGGG' tag number 171, 'GGGT' tag number 172, 'TTTA' tag number 253 and 'TTTT' tag number 256.



(a) True probabilities of expression 1



(b) True probabilities of expression 2



(c) True probabilities of expression

Figure 1: True probabilities of expression

These true parameter configurations have been chosen because they can give us a good inside on how the algorithm works, on how the algorithm correct for sequencing errors. The first vector $\underline{\pi}_1$ represents the simple situation of four tags for which the probabilities of expression are all equal to 25%. With only 4 true tags, we can study how the amount of their expression probability that we expect to be underestimated, is scattered across tags that have a zero true tag count. This situation is our primary interest and thus we will only show the plots concerning the parameter $\underline{\pi}_1$ is the result section. The second $\underline{\pi}_2$ and third parameter $\underline{\pi}_3$ are chosen to study the effect of the presence of true first order neighbour tags.

4.2 The transition matrix

To assess the effect of the chosen sequencing errors on the estimation of the parameter $\underline{\pi}$, we will use two different error rates for one tag. This will result into two different transition matrices Ω_1 and Ω_2.
The first error rate for a nucleotide we consider is 10%. This results in an error rate of $1 - 0.9^4 = 34.4\%$ for the 4-bp tags and gives us the following matrix:

$$\Omega_1 = \begin{pmatrix} 0.9 & 0.0333 & 0.0333 & 0.0333 \\ 0.0333 & 0.9 & 0.0333 & 0.0333 \\ 0.0333 & 0.0333 & 0.9 & 0.0333 \\ 0.0333 & 0.0333 & 0.0333 & 0.9 \end{pmatrix}.$$

The second error rate for a nucleotide we consider is 5%. This results in an error rate of $1 - 0.95^4 = 18.6\%$ for the 4-bp tags and gives us the following matrix:

$$\Omega_2 = \begin{pmatrix} 0.95 & 0.0167 & 0.0167 & 0.0167 \\ 0.0167 & 0.95 & 0.0167 & 0.0167 \\ 0.0167 & 0.0167 & 0.95 & 0.0167 \\ 0.0167 & 0.0167 & 0.0167 & 0.95 \end{pmatrix}.$$

The true transition matrix is unknown and need to be estimated before we estimate the parameter $\underline{\pi}$. We mimic the estimation of $\hat{\Omega}_1$ and $\hat{\Omega}_2$ by disturbing the matrices Ω_1 and Ω_2, respectively, with a small error. This is done by sampling from a Dirichlet distribution with mean Ω and a certain variance factor, so that

$$\hat{\Omega} = \begin{pmatrix} \omega_{11} + \epsilon_{11} & \omega_{12} + \epsilon_{12} & \omega_{13} + \epsilon_{13} & \omega_{14} + \epsilon_{14} \\ \omega_{21} + \epsilon_{21} & \omega_{22} + \epsilon_{22} & \omega_{23} + \epsilon_{23} & \omega_{24} + \epsilon_{24} \\ \omega_{31} + \epsilon_{31} & \omega_{32} + \epsilon_{32} & \omega_{33} + \epsilon_{33} & \omega_{34} + \epsilon_{34} \\ \omega_{41} + \epsilon_{41} & \omega_{42} + \epsilon_{42} & \omega_{43} + \epsilon_{43} & \omega_{44} + \epsilon_{44} \end{pmatrix},$$

with $E[\epsilon_{ij}] = 0$.
The matrices $\hat{\Omega}_1$ and $\hat{\Omega}_2$ are then used to construct the matrix Φ. Note that the matrices $\hat{\Omega}_1$ and $\hat{\Omega}_2$ still have the property that for each row the elements sum up to one. The estimation of $\hat{\Omega}$ is done for each simulation. This means that we estimate $\hat{\Omega}$ 1000 times.
However we also run simulations assuming that Ω_2 is the true transition matrix, i.e. $\hat{\Omega}_2 = \Omega_2$. So, the construction of Φ in this case is based on the matrix Ω_2. The goal of this set of simulations is to compare the bias introduced by $\hat{\Omega}_2$ with the bias resulting from working with the known true transition matrix Ω_2.

4.3 Generation of the libraries

The SAGE libraries we use for the estimation of the parameter $\underline{\pi}$ consist of the observed counts.
To generate the observed counts, we need to generate the true counts first. The true counts are generated by sampling from a multinomial distribution with parameters $\underline{\pi}_1, \underline{\pi}_2$ or $\underline{\pi}_3$, the

number of tags per library N_i $(i = 1, \ldots, g)$ and the number of libraries g. The number of tags per library is set equal to 2000 for all libraries.
To obtain the observed counts we need to mimic the sequencing error process. This is a process in which the true counts are scattered around the most probable error tags. This will lead to non-zero counts for tags which have a true zero count. So we need to do the following:

1. Build the matrix Φ as given in equation (1).
2. For tag j we use the jth column of the Φ-matrix. We call this vector τ. Then sampling the multinomial with parameters τ and sample size n_{ij} gives us the subset of the observed tag counts for tag j.
3. Repeating the previous step for all true tags and summing over all subsets gives us a library with the observed tag counts.

We generate 21000 libraries with sample size $N = 2000$ so that we can perform 1000 simulations for a single library and 1000 simulations for 20 libraries. We do this for the six different situations (i.e., the three vectors $\underline{\pi}_1$, $\underline{\pi}_2$ and $\underline{\pi}_3$ each with the two transition matrices Ω_1 and Ω_2). Note that the construction of the Φ-matrix in the second step of the generation process is based on the matrices Ω_1 and Ω_2 and not on their estimates.
Figure 2 shows the mean of the observed counts $\underline{n}^*$ over the libraries for the situation of the first parameter $\underline{\pi}_1$ with the two transition matrices. Figures 2a and 2c represent the mean of the observed counts $\underline{n}^*$ for the single library case, where the mean is over 1000 libraries. Figures 2b and 2d represent the mean of the observed counts $\underline{n}^*$ for the 20 libraries case, where the mean is over 20000 libraries. In Figure 2 the effect of the different sequencing error rates can be observed. In the case of the matrix Ω_2 with a small sequencing error rate, the scatter of the true counts is smaller and thus the observed counts of the true tags are larger than in the case of Ω_1 with a larger sequence error rate. The same conclusion can be made based the plots for the two other parameters (see Figures 21 and 22 in the appendix A).

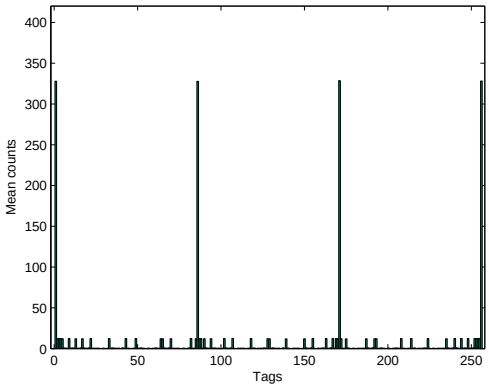
4.4 Convergence monitoring

The convergence of the estimation of $\hat{\underline{\pi}}$ is monitored through the relative distance between the current estimate for the parameter and the previous estimate for the parameter in each iteration step. We use the following formula for the relative distance:

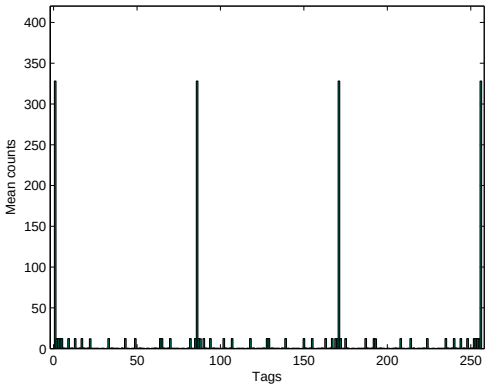
$$d(k) = \max\left(\frac{|\hat{\underline{\pi}}_{k-1} - \hat{\underline{\pi}}_k|}{\hat{\underline{\pi}}_{k-1}}\right), \tag{38}$$

where $k = 1, \ldots, 50$ is the index of the iteration steps. The relative distance is calculated for each simulation.
In addition, if $\hat{\underline{\pi}}_k$ converges to $\hat{\underline{\pi}}$, then the largest eigenvalue of the matrix

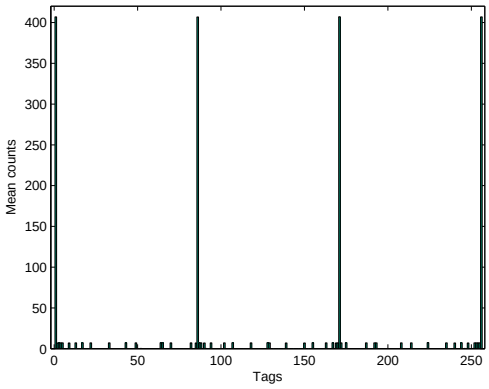
$$J(\hat{\underline{\pi}}) = \mathcal{I}_c^{-1}(\hat{\underline{\pi}}, \underline{n}^*)\mathcal{I}_m(\hat{\underline{\pi}}, \underline{n}^*) \tag{39}$$



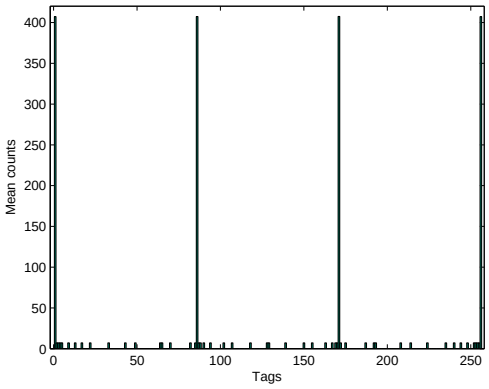
(a) 1 library: $\underline{\pi}_1$, Ω_1



(b) 20 libraries:$\underline{\pi}_1$, Ω_1



(c) 1 library: $\underline{\pi}_1$, Ω_2



(d) 20 libraries:$\underline{\pi}_1$, Ω_2

Figure 2: Mean of observed counts for $\underline{\pi}_1$

gives the rate of convergence of the EM algorithm. The expression given in equation (39) measures the fraction of information about $\underline{\pi}$ that is missing. So if $J(\underline{\hat{\pi}})$ increases, the convergence speed decreases. The fraction of missing information may vary across the different components of $\underline{\hat{\pi}}$.

5 Results

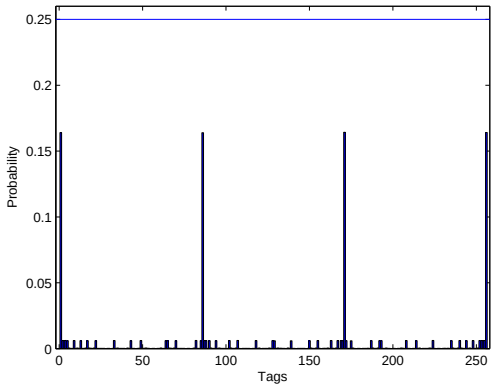
In this section we give the results of the simulations explained in section 4. Only the plots concerning the first true expression probabilities $\underline{\pi}_1$ are displayed. The plots concerning the results for the parameters $\underline{\pi}_2$ and $\underline{\pi}_3$ can be found in the appendix sections B-F.

5.1 The observed expression probabilities

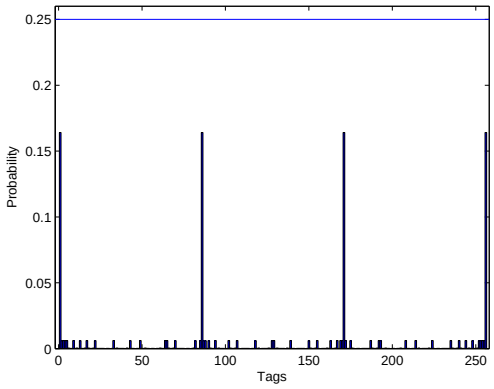
We can estimate the parameter $\underline{\pi}$ based on the observed counts without correcting for the sequencing errors:

$$\hat{\pi}_j^* = \frac{\sum_{i=1}^{g} n_{ij}^*}{gN}, \qquad \text{for } j = 1, \ldots, K \tag{40}$$

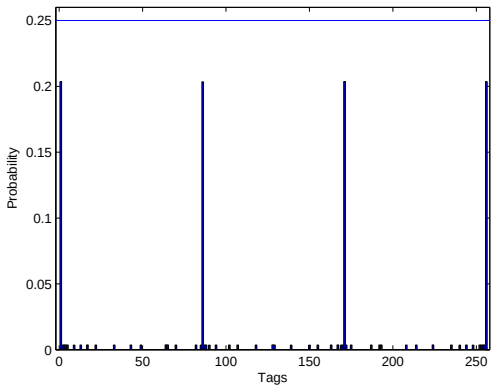
with g the number of libraries per simulation. The estimator $\hat{\underline{\pi}}^* = (\hat{\pi}_1^*, \ldots, \hat{\pi}_K^*)$ is the vector of the observed expression probabilities and is a maximum likelihood estimator.
In Figure 3 we see the comparison of the vector of the observed expression probabilities with the parameter $\underline{\pi}_1$. Figures 3a and 3c show the results for the single library case and Figures 3b and 3d for the multiple libraries case. We can observe that the expression probabilities of the true tags (see Table 1) are underestimated. Tags, which should have count zero, are introduced with very small expression probabilities. The same conclusions can be made from Figures 23 and 24 (see appendix B) of the results for the parameters $\underline{\pi}_2$ and $\underline{\pi}_3$.



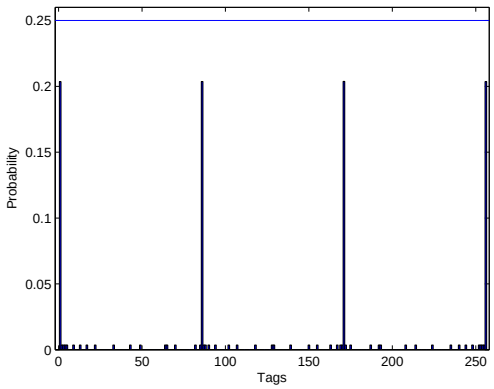
(a) 1 library:$\underline{\pi}_1$, Ω_1



(b) 20 libraries:$\underline{\pi}_1$, Ω_1



(c) 1 library: $\underline{\pi}_1$, Ω_2

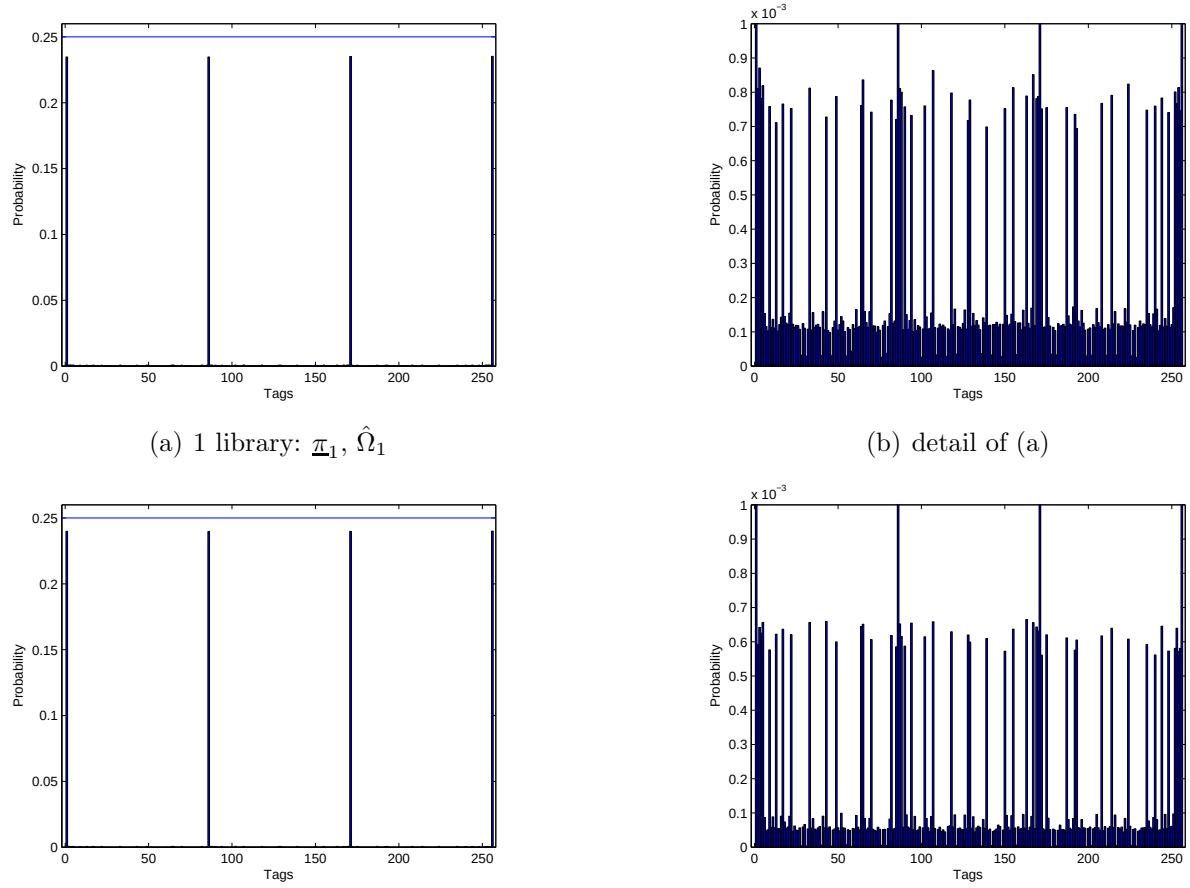


(d) 20 libraries:$\underline{\pi}_1$, Ω_2

Figure 3: The observed expression probabilities $\hat{\underline{\pi}}_1^*$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 25%. The other true expression probabilities are equal to zero.

5.2 Single library

Figure 4 shows the mean of the estimator $\hat{\underline{\pi}}_1$ over the 1000 simulations. The mean of the estimates obtain with the matrix $\hat{\Omega}_1$ for a single library is shown in Figure 4a and for a single library with matrix $\hat{\Omega}_2$ in Figure 4c. The expression probabilities of the true tags (see Table 1) are underestimated. In Figure 4a, the amount of the expression probabilities that is underestimated for the true tags is approximately the same for the four tags. The same observation can be made from Figure 4c. However, in the case of the matrix $\hat{\Omega}_2$, the underestimation is smaller than when we use the matrix $\hat{\Omega}_1$. Figures 4b and 4c are the enlarged



(a) 1 library: $\underline{\pi}_1$, $\hat{\Omega}_1$

(b) detail of (a)

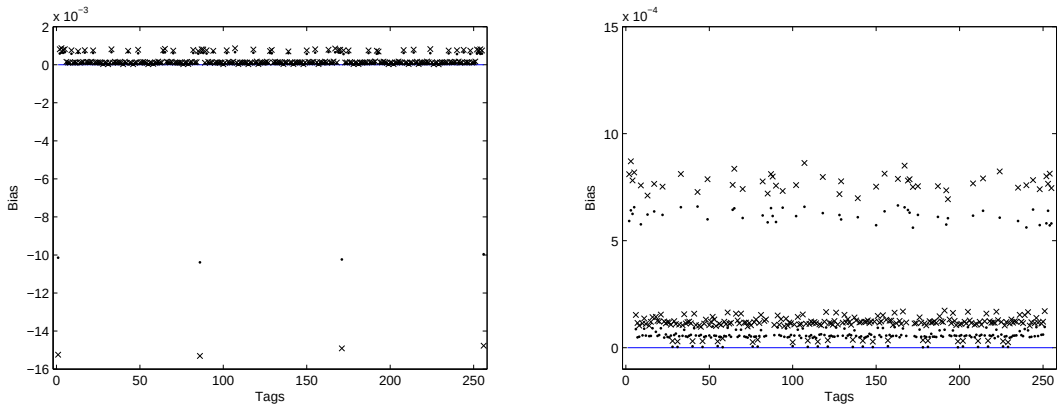
(c) 1 library: $\underline{\pi}_1$, $\hat{\Omega}_2$

(d) detail of (c)

Figure 4: The estimated expression probabilities $\hat{\underline{\pi}}_1$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 25% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.

plots of the very small expression probabilities in the Figures 4a and 4c, respectively. From these enlarged plots can be observed that all the tags with a true zero count have a small probability of being expressed and that the expression probabilities of the tags, with true count zero, is smaller in the case of the matrix $\hat{\Omega}_2$ than in the case of the matrix $\hat{\Omega}_1$.
The smaller underestimations of $\underline{\pi}_1$ and the smaller expression probabilities of the tags with

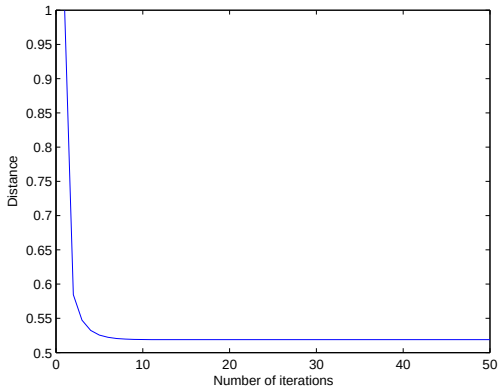
true count zero when we use the matrix $\hat{\Omega}_2$, is what we expected because of the smaller sequencing errors.
Figure 5 shows the bias of the mean of $\hat{\underline{\pi}}_1$ over the 1000 simulations for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. We can make several conclusions from Figure 5. First of all, the bias in case of $\hat{\Omega}_1$ is larger than in case of $\hat{\Omega}_2$. Secondly, the underestimation of the expression probabilities is approximately of the same magnitude for the four true tags. In Figure 5(b), we zoom in on the smallest bias



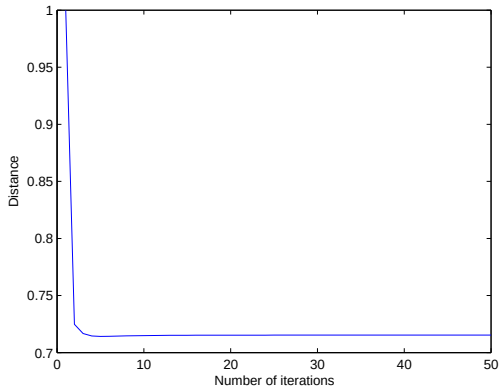
(a) (b) detail of (a)

Figure 5: Single library: Bias of $\hat{\underline{\pi}}_1$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. The crosses ($\times$) represent the bias of the estimates resulting from the use of $\hat{\Omega}_1$ and the dots ($\bullet$) the bias of the estimates resulting from the use of $\hat{\Omega}_2$.

values. These values correspond to the tags with a true zero count. Figure 5(b) clearly shows that there no estimates with a zero bias, although the values are close to zero. Even for the smallest bias values are the ones resulting from the use of $\hat{\Omega}_1$ larger than those resulting from $\hat{\Omega}_2$.
The convergence of the estimation of $\underline{\pi}_1$ is monitored through the relative distance (see equation (38)). In Figure 6 the mean relative distance over the 1000 simulations is given. The relative distance in both case stabilizes after approximately 10 iterations. So there is a plateau after 10 iterations, indicating that we cannot get a better convergence after more than 50 iterations than we already have.
We are not able to calculate the Variance-Covariance matrix because the information matrix corresponding to $\hat{\underline{\pi}}_1$ is singular. Also the fraction of missing information can not be calculated due to the singularity of the complete-data information matrix $\mathcal{I}_c$. This means that we have 100% missing information. This may be an indication of identifiability problems.
Next, we assume that the matrix Ω_2 is the true transition matrix and hence we use Ω_2 for the estimation of $\underline{\pi}_1$ instead of $\hat{\Omega}_2$. Figure 7 shows us the mean of the estimator $\hat{\underline{\pi}}_1$ over the 1000 simulations. Again the expression probabilities of the true tags are underestimated and from the enlarged plot (Figure 7b) we observe that all the tags with a true zero count have a small probability of being expressed. Now we can compare the bias of the estimates for $\underline{\pi}_1$ obtained with Ω_2 with the results obtained with $\hat{\Omega}_2$. Figure 8 shows that the bias is larger when we estimate the transition matrix. Hence, if we know the true transition matrix, we

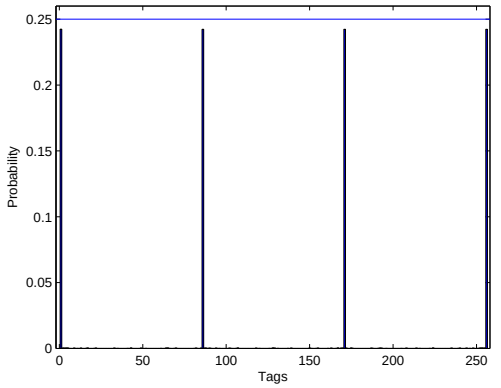


(a) 1 library: $\underline{\pi}_1$, Ω_1

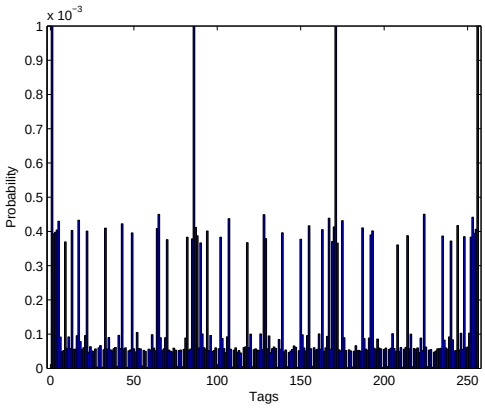


(b) 1 library: $\underline{\pi}_1$, Ω_2

Figure 6: Convergence monitoring through the relative distance



(a)



(b) detail of (a)

Figure 7: The estimated expression probabilities $\hat{\underline{\pi}}_1$ for Ω_2. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 25% in subfigure (a). The other true expression probabilities are equal to zero.

can reduce the bias.
Finally, we can also compare the bias of $\hat{\underline{\pi}}_1^*$, the observed expression probabilities, with the estimates from the EM algorithm. In Figure 9, it is shown that the bias of the observed expression probabilities is larger, as expected because there is no correction for the sequencing errors.
Let us now discuss the results of the other two parameters $\underline{\pi}_2$ and $\underline{\pi}_3$. The plots to illustrate the results can be found in the appendix section C. In these scenarios we have 8 tags (see Table 1) that are uniformly expressed in case of $\hat{\underline{\pi}}_2$ or where the first order neighbours tags are less expressed than the four main tags in case of $\hat{\underline{\pi}}_3$.
Figures 25 and 26 show the mean of the estimators $\hat{\underline{\pi}}_2$ and $\hat{\underline{\pi}}_3$ over the 1000 simulations. The expression probabilities of the true tags (see Table 1) are underestimated and all the

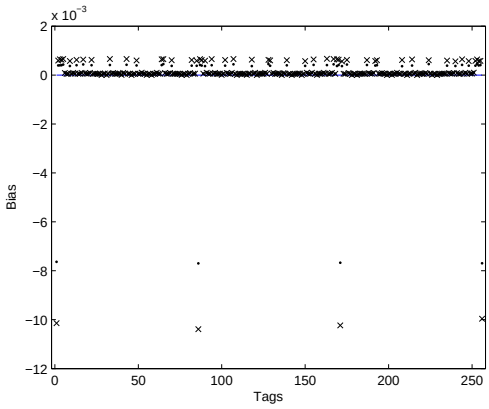
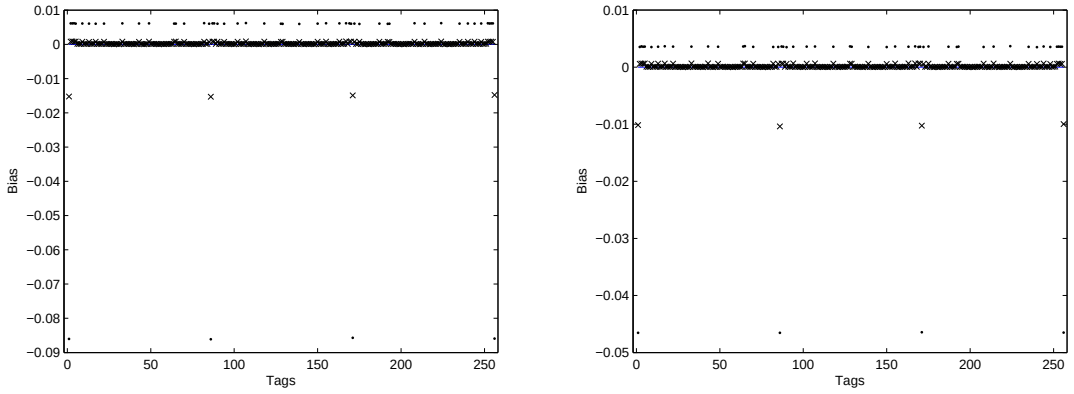


Figure 8: Bias comparison between $\hat{\Omega}_2(\times)$ and $\Omega_2(\bullet)$



(a) Ω_1 (b) Ω_2

Figure 9: Bias of the observed expression probabilities $(\bullet)$ versus the estimates from the EM algorithm $(\times)$

tags with a true zero count have a small probability of being expressed. From the enlarged plots (Figures 25b and 25d, Figures 26b and 26d) it can be observed that the probabilities given to the tags with a true zero count are smaller in the case of the transition matrix $\hat{\Omega}_2$ than in the case of the the matrix $\hat{\Omega}_1$ for $\underline{\hat{\pi}}_2$, but for $\underline{\hat{\pi}}_3$ there does not seem to be a obvious difference.

The bias plots of $\underline{\hat{\pi}}_2$ and $\underline{\hat{\pi}}_3$ (Figures 27 and 28) show that the bias is smaller in the case of the matrix $\hat{\Omega}_2$, as expected.

Figures 29 and 30 show the monitoring of the convergence through the relative distance. For $\hat{\Omega}_1$ the relative distance stabilizes after approximately 15 iterations and for $\hat{\Omega}_2$ after approximately 5 iterations. Again there is no indication for a better convergence after more than 50 iterations.

We are not able to calculate the Variance-Covariance matrices because the information matrices corresponding to $\underline{\hat{\pi}}_2$ and $\underline{\hat{\pi}}_3$ are singular. Also the fraction of missing information can not be calculated due to the singularity of the complete-data information matrices $\mathcal{I}_c$.

This means that we have 100% missing information. Again, we may have an indication of identifiability problems.
Assuming that the matrix Ω_2 is the true transition matrix, Figures 31 and 32 show the mean of the estimator $\underline{\hat{\pi}}_1$ over the 1000 simulations. Again the expression probabilities of the true tags are underestimated and from the enlarged plots (Figures 31b and 32b) we observe that all the tags with a true zero count have a small probability of being expressed. So, again, knowing the true transition matrix reduces the bias of the estimates. We also compare the bias of the observed expression probabilities $\underline{\hat{\pi}}_2^*$ and $\underline{\hat{\pi}}_3^*$ with the estimates resulting from the EM algorithm. As expected, Figures 36 and 36 show that the bias of the observed expression probabilities is larger.

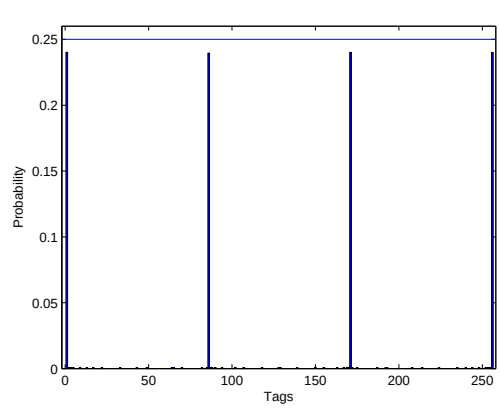
5.3 Multiple libraries

Figure 10 shows the mean of the estimator $\underline{\hat{\pi}}_1$ over the 1000 simulations. The enlarged plots (Figures 10b and 10d) show that tags with a true zero count have a small probability of being expressed. We can make the same observations as in the single library case, namely that

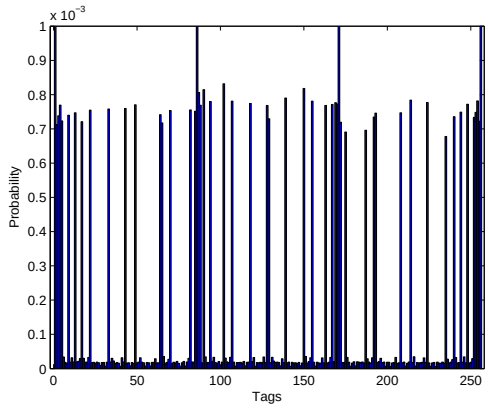
- the amount of the expression probabilities that is underestimated for the true tags is almost the same for the 4 tags;
- the probabilities given to the tags with true count zero, are smaller in the case of the matrix $\hat{\Omega}_2$;
- the underestimation is smaller in the case of the matrix $\hat{\Omega}_2$.

The bias of $\underline{\hat{\pi}}_1$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$ is shown in Figure 11. As is the case for a single library, we can see that the bias of the estimates resulting from the use of $\hat{\Omega}_2$ is smaller than the bias of the estimates when $\hat{\Omega}_1$ was used. Again we have that the underestimation of the expression probabilities is approximately of the same magnitude for the four true tags.
Focusing on the smallest bias values (these correspond to the tags with a true zero count) (see Figure 11(b)), the observation that there are no estimates with a zero bias can be made, although the values are close to zero. Within the smallest values is the bias resulting from the use of $\hat{\Omega}_1$ are larger than the bias resulting from working with $\hat{\Omega}_2$.
The monitoring of the convergence of the estimation is given in Figure 12, where we show the mean relative distance over the 1000 simulations. The relative distance is both case seems to stabilizes after approximately 40 iterations and again there is no indication for a better convergence after more than 50 iterations.
As in the single library case, we are not able to calculate the Variance-Covariance matrix because the information matrix corresponding to $\underline{\hat{\pi}}_1$ is singular. Also the fraction of missing information can not be calculated due to the singularity of the complete-data information matrix $\mathcal{I}_c$. This means that we have 100% missing information and probably identifiability problems.

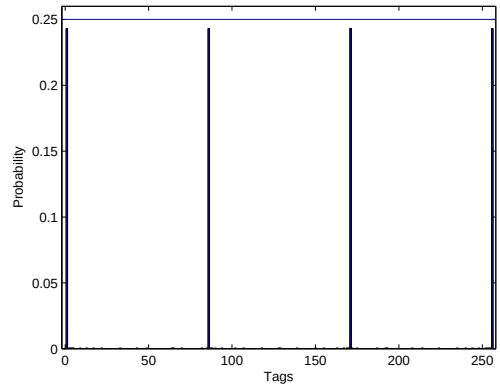
Now we assume that the matrix Ω_2 is the true transition matrix. Figure 13 shows us the



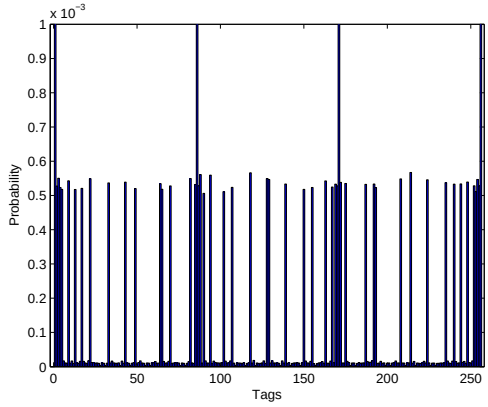
(a) 20 libs: $\underline{\pi}_1$, $\hat{\Omega}_1$



(b) detail of (a)



(c) 20 libs: $\underline{\pi}_1$, $\hat{\Omega}_2$



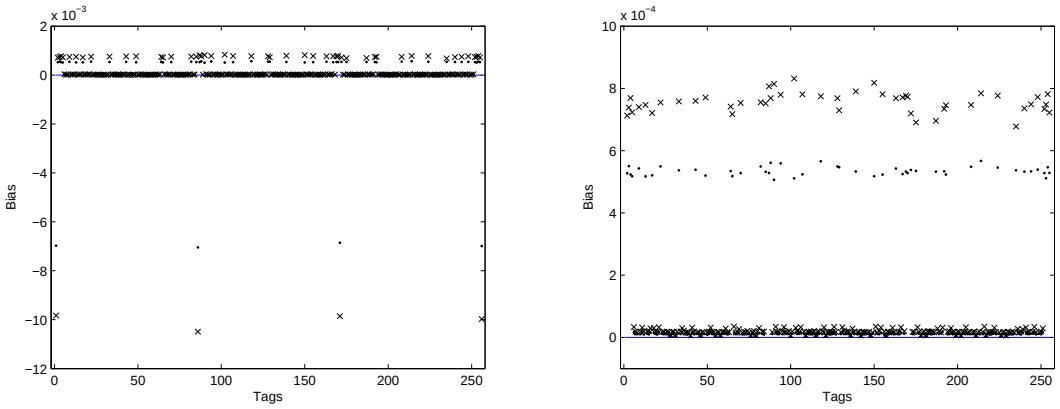
(d) detail of (c)

Figure 10: The estimated expression probabilities $\hat{\underline{\pi}}_1$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 25% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.

mean of the estimator $\hat{\underline{\pi}}_1$ over the 1000 simulations. The expression probabilities of the true tags are underestimated. From the enlarged plot (Figure 13b) we observe that all the tags with a true zero count have a small probability of being expressed. Figure 14 shows comparison of the bias of the estimates for $\hat{\underline{\pi}}_1$ obtained with Ω_2 with the results obtained with $\hat{\Omega}_2$. As is the case for a single library, we have that the bias is larger for the case where we estimate the transition matrix.

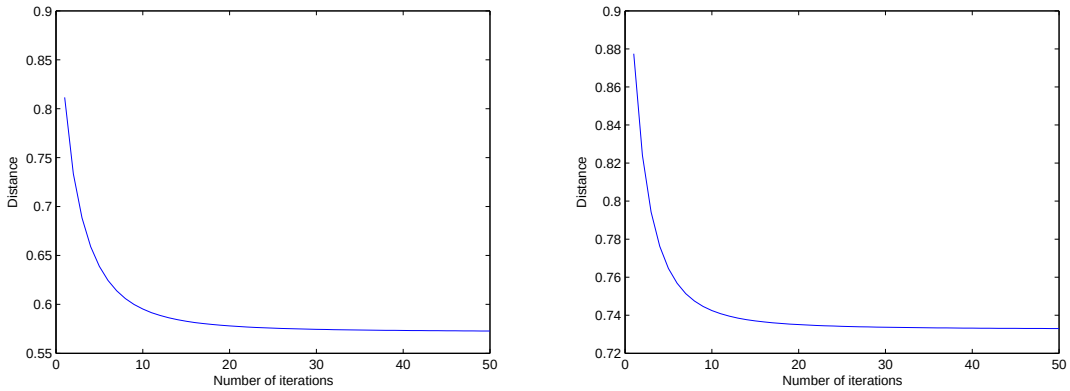
Finally, we can compare the bias of the observed probabilities $\hat{\underline{\pi}}_1^*$ with the estimates from the EM algorithm. From Figure 15, the larger bias of the observed probabilities is observed, as expected.

Now we discuss the results of the other two parameters $\underline{\pi}_2$ and $\underline{\pi}_3$. The plots to illustrate the results can be found in the appendix section D.
Figures 37 and 38 show the mean of the estimators $\hat{\underline{\pi}}_2$ and $\hat{\underline{\pi}}_3$ over the 1000 simulations.



(a) (b) detail of (a)

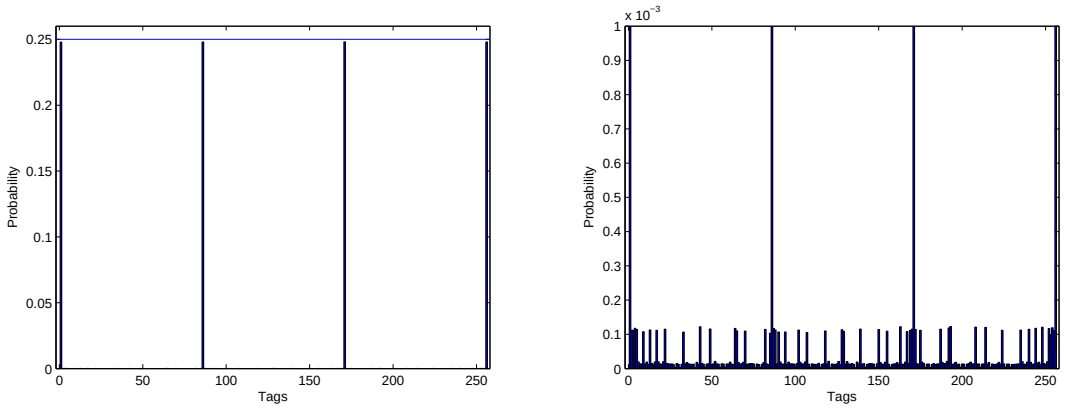
Figure 11: 20 libs: Bias of $\underline{\hat{\pi}}_1$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. The crosses ($\times$) represent the bias of the estimates resulting from the use of $\hat{\Omega}_1$ and the dots ($\bullet$) the bias of the estimates resulting from the use of $\hat{\Omega}_2$.



(a) 20 libs: $\underline{\pi}_1$, Ω_1 (b) 20 libs: $\underline{\pi}_1$, Ω_2

Figure 12: Convergence monitoring through the relative distance

The expression probabilities of the true tags (see Table 1) are underestimated and all the tags with a true zero count have a small probability of being expressed. From the enlarged plots (Figures 25b and 25d, Figures 38b and 38d) it can be observed that the probabilities given to the tags with a true zero count are smaller in the case of the transition matrix $\hat{\Omega}_2$ than in the case of the matrix $\hat{\Omega}_1$.
The bias plots of $\underline{\hat{\pi}}_2$ and $\underline{\hat{\pi}}_3$ (Figures 39 and 40) show that the bias is smaller in the case of the transition matrix $\hat{\Omega}_2$, because of the smaller sequencing errors. Figures 41 and 42 show the monitoring of the convergence through the relative distance. For both $\hat{\Omega}_1$ and for $\hat{\Omega}_2$ the relative distance stabilizes after approximately 40 iterations. The obtained plateau does not indicates towards a better convergence after more than 50 iterations.
As above, we are not able to calculate the Variance-Covariance matrix because the informa-



(a) (b) detail of (a)

Figure 13: The estimated expression probabilities $\underline{\hat{\pi}}_1$ for Ω_2. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 25% in subfigure (a). The other true expression probabilities are equal to zero.

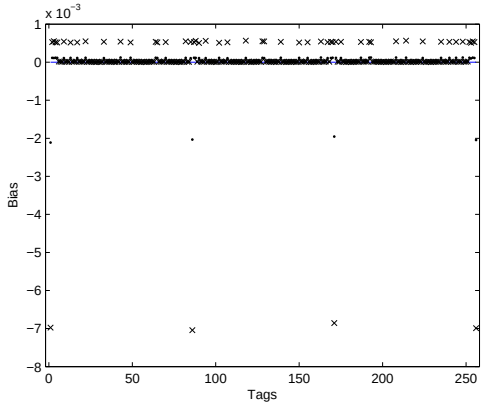
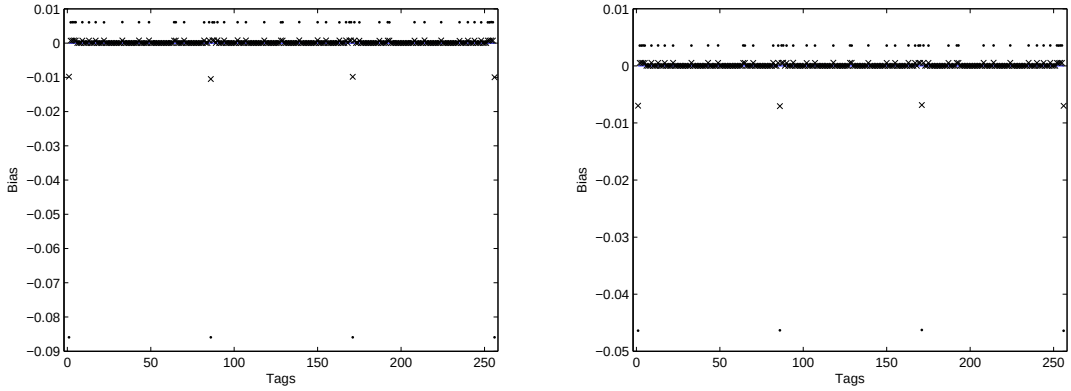


Figure 14: Bias comparison between $\Omega_2(\bullet)$ and $\hat{\Omega}_2(\times)$

tion matrix corresponding to $\underline{\hat{\pi}}_1$ is singular. The fraction of missing information can not be calculated due to the singularity of the complete-data information matrix $\mathcal{I}_c$. This means that we have 100% missing information and probably identifiability problems.
We assume now that Ω_2 is the true transition matrix. Figures 43 and 44 show that the expression probabilities of the true tags are underestimated and from the enlarged plots (Figures 43b and 44b) we observe that all the tags with a true zero count have a small probability of being expressed. The comparison of the bias of the estimates resulting from using $\hat{\Omega}_2$ with the estimates for which we used Ω_2, in Figures 45 and 46, shows that the bias is larger for the case where we estimate the transition matrix.
Finally, we compare the bias of the observed expression probabilities $\underline{\hat{\pi}}_2^*$ and $\underline{\hat{\pi}}_3^*$ with the estimates resulting from the EM algorithm. As expected, Figures 47 and 48 show that the bias of the observed expression probabilities is larger.



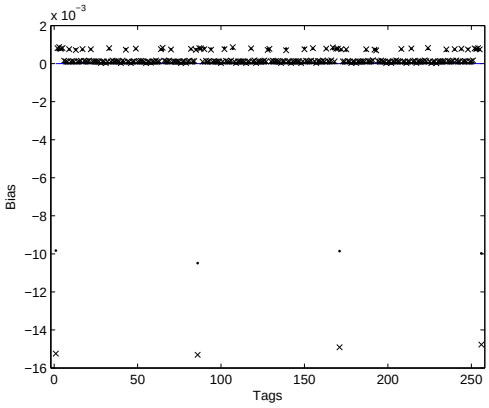
(a) T_1 (b) T_2

Figure 15: Bias of the observed expression probabilities (•) versus the estimates from the EM algorithm (×)

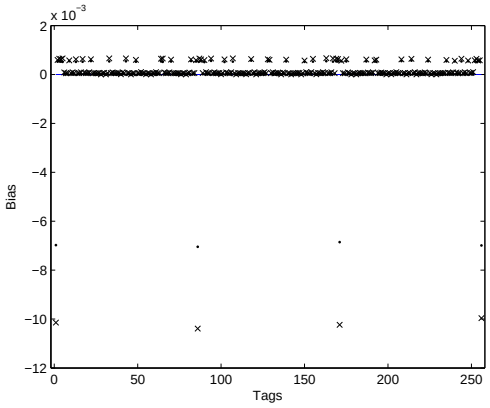
5.4 Single library versus 20 libraries

In this subsection, we want to compare the bias of $\underline{\hat{\pi}}_1$ obtained from the simulations for a single and 20 libraries. In this following Figures 16a, 16b and 16c,we show the bias of the mean of $\underline{\hat{\pi}}_1$ over the 1000 simulations for the three possible transition matrices $\hat{\Omega}_1$, $\hat{\Omega}_2$ and Ω_2, respectively.

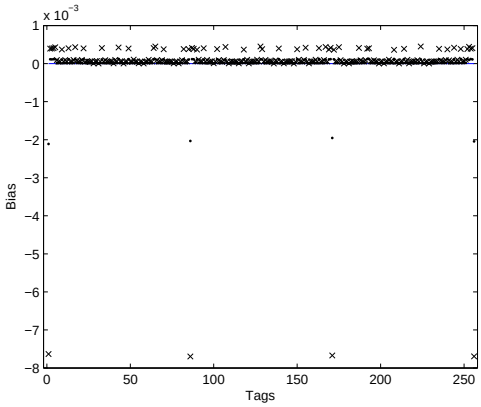
In all three Figures we can see the benefit in terms of bias if we can use 20 libraries rather



(a) Bias of 1 library vs 20 libraries for $\hat{\Omega}_1$



(b) Bias of 1 library vs 20 libraries for $\hat{\Omega}_2$



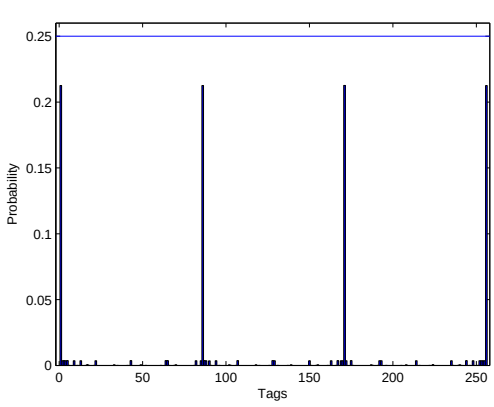
(c) Bias of 1 library vs 20 libraries for Ω_2

Figure 16: Bias of 1 library ($\times$) vs 20 libraries ($\bullet$) for $\hat{\Omega}_1$, $\hat{\Omega}_2$ and Ω_2

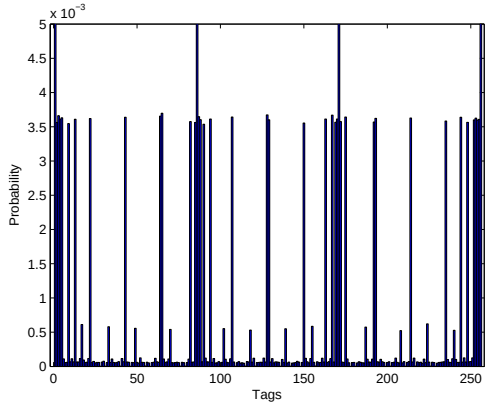
than only a single library. Focusing on the true tags (see Table 1), the bias for these tags seems to be the largest in the case of $\hat{\Omega}_1$ and the smallest in the case of Ω_2, as expected. Next we look at the results of the other two parameters $\underline{\pi}_2$ and $\underline{\pi}_3$. The Figures 49 and 50 both show the reduces of the bias if we can use multiple libraries rather than only a single library.

5.5 Restriction of first order sequencing errors

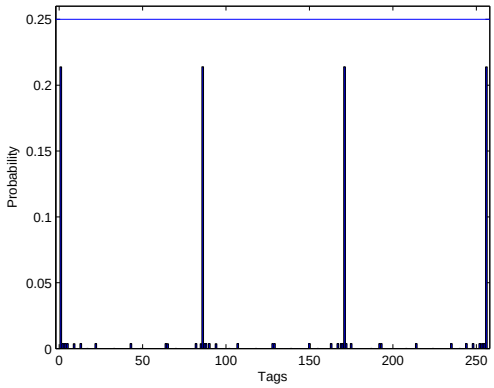
In this subsection, we present the results from the simulations for a single library and multiple libraries where the method of Beissbarth *et al.* was adopted, i.e. the sequencing errors come from the first order neighbours. Hence the matrix Φ_B, as defined in section 2 is used for the estimation of the parameter $\underline{\pi}$. We assume that Ω_2 is the true transition matrix and so Ω_2 is used to build Φ. The matrix Ω_2 is also used for the construction of Φ_B, but with the restriction that sequencing errors can only coming first order neighbours.
Figure 17 gives the mean over the 1000 simulations of a single library and 20 libraries for $\underline{\pi}_1$ with the method of Beissbarth *et al.* (2004).



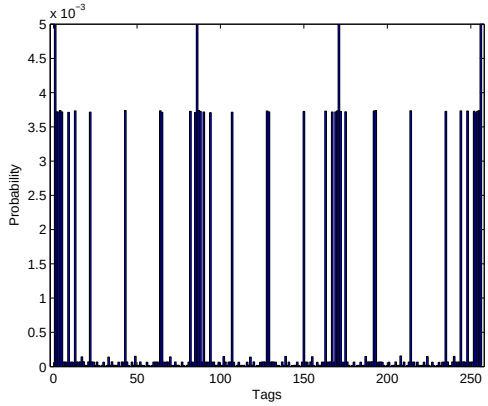
(a) 1 library: $\underline{\pi}_1$, Φ_B



(b) detail of (a)



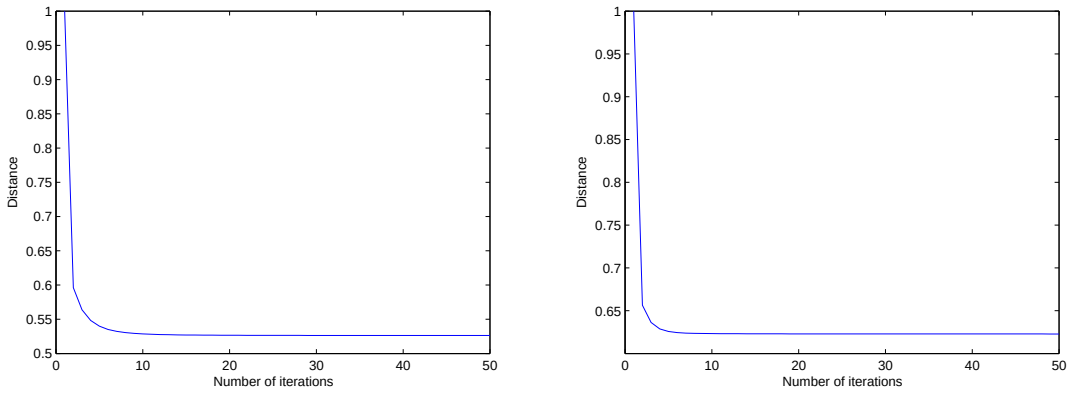
(c) 20 libraries: $\underline{\pi}_1$, Φ_B



(d) detail of (c)

Figure 17: The estimated expression probabilities $\hat{\underline{\pi}}_1$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 25% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.

We can see that if we use the Beissbarth method then there is also an underestimation of the four true tags (see Table 1). Again, all the tags with a true zero count have a small probability of being expressed. The relative distance (see Figure 18), that monitors the convergence, appears to stabilize after approximately 10 iterations, which is faster than most of the cases discussed in the previous subsections. Here, the plateau also does not give an indication towards a better convergence after more than 50 iterations.
As is the case in the previous subsections, we are not able to calculate the Variance-Covariance matrix because the information matrix corresponding to $\underline{\hat{\pi}}_1$ is singular. So the fraction of missing information can not be calculated due to the singularity of the complete-data information matrix $\mathcal{I}_c$, meaning that we have 100% missing information. This may be again an indication of identifiability problems.

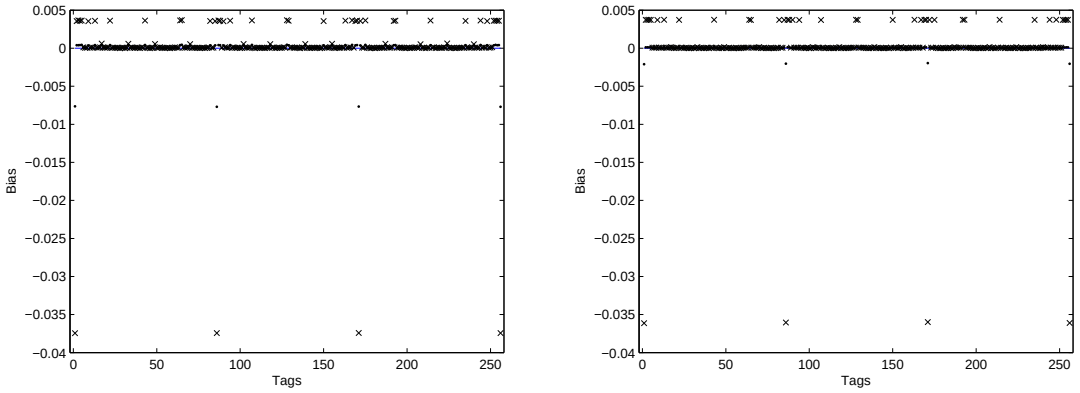


(a) 1 library: $\underline{\pi}_1$, Φ_B (b) 20 libraries: $\underline{\pi}_1$, Φ_B

Figure 18: Convergence monitoring through the relative distance for $\underline{\hat{\pi}}_1$

Let us now compare the results of the case where we use the matrix Φ (i.e. our method) with the case where we use Φ_B. From Figure 19, it is clear that in both cases the bias of $\underline{\hat{\pi}}_1$ for Φ_B is larger than for Φ. So, although the relative distance seems to stabilize quite fast, the estimates using the method developed by Beissbarth *et al.* (2004) are worse than when our method. We show a comparison of the bias of $\underline{\hat{\pi}}_1$ resulting from using Φ_B for a single library with multiple libraries in Figure 20. The bias is smaller for the multiple libraries case. However the difference between the bias resulting from the multiple library case and the bias resulting from a single library is not as large as in the situation shown in Figure 16c.

Finally, we look at the results for the parameters $\underline{\pi}_2$ and $\underline{\pi}_3$. Figure 51 and Figure 52 give the mean over the 1000 simulations of a single library and 20 libraries for $\underline{\pi}_2$ and $\underline{\pi}_3$, respectively, with the method of Beissbarth *et al.* (2004). If the Beissbarth method is used then there is an underestimation of the eight true tags (see Table 1). Again, all the tags with a true zero count have a small probability of being expressed.
The relative distance (see Figures 55 and 56), that monitors the convergence, appears to sta-



(a) 1 library: $\underline{\pi}_1$ (b) 20 libraries: $\underline{\pi}_1$

Figure 19: Bias of $\hat{\underline{\pi}}_1$ for $\Phi_B(\times)$ versus $\Phi(\bullet)$

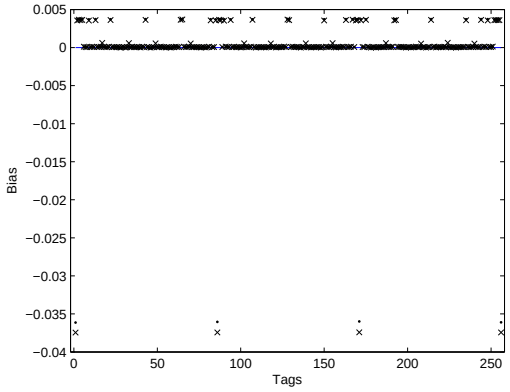


Figure 20: Bias of 1 library ($\times$) vs 20 libraries ($\bullet$) for Φ_B

bilize after approximately 10 iterations, which is faster than most of the cases discussed in the previous subsections but again there is no indication for a better convergence after more than 50 iterations. We are not able to calculate the Variance-Covariance matrix because the information matrix corresponding to $\hat{\underline{\pi}}_1$ is singular. So the fraction of missing information can not be calculated due to the singularity of the complete-data information matrix $\mathcal{I}_c$, meaning that we have 100% missing information.

Let us now compare the results of the case where we use the matrix Φ with the case where we use Φ_B. From Figures 53 and 54, it is clear that the bias of $\hat{\underline{\pi}}_2$ and $\hat{\underline{\pi}}_3$ for Φ_B is larger than for Φ. So, although the relative distance seems to stabilize quite fast, the estimates using the method developed by Beissbarth *et al.* (2004) are again worse than when our method. We show the comparison of the bias of $\hat{\underline{\pi}}_2$ and $\hat{\underline{\pi}}_3$(using Φ_B) for a single library with multiple libraries in Figure 57. The bias is smaller for the multiple libraries case.

6 Conclusions

In this project we proposed a statistical model for the propagation of sequencing errors in the case that we have multiple SAGE libraries and correct for the sequencing error through an EM algorithm.
A general conclusion that can be made is that the model does not 100% correct the sequencing errors. Thus the estimates are not entirely free of sequencing errors. This can be seen through the underestimation of the expression probabilities for the true tags (see Table 1) and the introduction of small expression probabilities for tags with a true zero tag count.
In the simulation study, we studied the potential gain in terms of bias comparing the estimates obtained by using the EM algorithm with the estimates when there is no correction for the sequencing errors, i.e., the observed expression probabilities. The conclusion that can be made from the Figures 9, 35 and 36 is that in the single library case there is a reduction in the bias of the estimates obtained by the EM algorithm with both transition matrices Ω_1 and Ω_2. The same conclusion can be made for the multiple library case. The conclusions are what we expected because the observed expression probabilities are not corrected for sequencing errors.
In subsection 5.4, we studied the potential gain in terms of bias resulting from the use of 20 libraries as compared with the estimates obtained using only one library. From Figure 16, we can clearly see the reduction of the bias when the estimates are obtained in the multiple library case. The same canclusion is true for the other parameters $\underline{\pi}_2$ and $\underline{\pi}_3$. This conclusion is in the line of our expectations because there is more information available for the estimation in the multiple case.
For the single and multiple library case, we studied the potential gain in terms of bias when the true transition matrix is known. From the Figures 8 and 14, the conclusion can be made that there is a gain in terms of bias when the true transition matrix is known, as expected. Again, we can make the same observation for the other parameters $\underline{\pi}_2$ and $\underline{\pi}_3$.
We also studied the potential gain in terms of bias comparing smaller sequencing errors with larger sequencing errors through the comparison of the transition matrix Ω_2 with the transition matrix Ω_1, respectively. For the single and multiple library case we can make the same conclusion, namely that the bias of the estimates is reduced when the transition matrix Ω_2 (corresponding to smaller sequencing errors) is used.
Finally, in subsection 5.5, we studied the comparison of the bias of the estimates resulting from the use of Φ with the estimates obtained by using Φ_B. Several conclusion can be made. Firstly, when using the method of Beissbarth *et al.* (2004), the expression probabilities of the true tags are underestimated and tags with a true zero count are introduced with small expression probabilities. Secondly, there is a gain in terms of bias for the estimates resulting from the multiple libraries case compared to the estimates resulting from the use of only a single library. However, the difference between the bias of these two cases is not as large as when we apply our method (described in section 3). We also compared the estimates resulting from the method developed in section 3 with the method of Beissbarth *et al.* (2004). As well in the multiple libraries as in the single library case, the bias of the estimates resulting from our method is smaller than the bias of the estimates resulting from the method of Beissbarth *et al.* (2004).

We encounter several problems with respect to the convergence of the EM algorithm. We use 50 iterations in the EM algorithm as suggested in Beissbarth *et al.* (2004). Within the 50 cycles, the relative distance stabilizes around a value between 0.5 and 0.8. Beissbarth *et al.* (2004) faced a similar problem. From the monitoring of the convergence, we see that the tags that made it difficult to converge, are in fact the tags with a very small probability of being expressed. The expression probabilities for these tags were smaller than 1×10^{-3}.
Another problem is the estimation of the Variance-Covariance matrix of the estimates $\hat{\underline{\pi}}_1$, $\hat{\underline{\pi}}_2$ and $\hat{\underline{\pi}}_3$. First we construct the information matrix and by inverting this matrix we obtain the Variance-Covariance matrix. However, the problem is the inversion of the information matrix, because of its singularity. Due to the singularity, we are also not able to calculate the fraction of missing information as given in the equation (39). Hence, we most likely have 100% missing information. The problematic estimation of the Variance-Coviance matrix may be an indication of identifiability problems.

References

[1] Akmaev,V.R. and Wang,C.J. (2004) Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics*, **20**, 1254-1263.

[2] Blades,N.J., Jones,J.B., Kern,S.E. and Parmigiani,G. (2004a) Denoising of data from serial analysis of gene expression., *Bioinformatics* (in press).

[3] Blades,N.J., Velculescu,V. and Parmigiani,G. (2004a) Estimation of sequencing error rates in SAGE libraries., *Genome Biol.* (in press).

[4] Burzykowski,T. (2005) Likelihood approach to sequencing errors in SAGE., Technical Report, CenStat, Hasselt University.

[5] Beissbarth,T., Hyde,L., Smyth,G.K., Job,C., Boon,W-M, Tan,S-S, Scott,H.S. and Speed,T.P. (2004) Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics*, **20**, Suppl. 1, i31-i39.

[6] Chu,T. (2002) A statistical analysis of the SAGE data.
http://www.phil.cmu.edu/projects/genegroup/papers/chu2002b.pdf

[7] Colinge,J. and Feger,G. (2001) Detecting the impact of sequencing errors on SAGE data. *Bioinformatics*, **17**, 840-842.

[8] Louis,T.A. (1982) Finding the Observed Information Matrix when Using the EM Algorithm, *J. R. Statist. Soc.***B**, **44**, No. 2, 226-233.

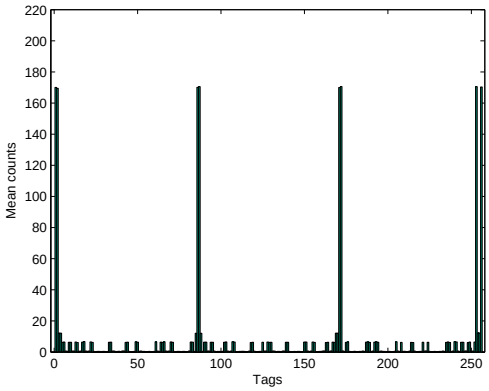
[9] McLachlan,G.J. and Krishnan,T. (1997) *The EM Algorithm and Extensions*, John Wiley & Sons Inc., New York.

[10] Stollberg,J., Urschitz,J., Urban,Z. and Boyd,C.D. (2000) A quantitative evaluation of SAGE. *Genome Research*, **10**, 1241-1248.

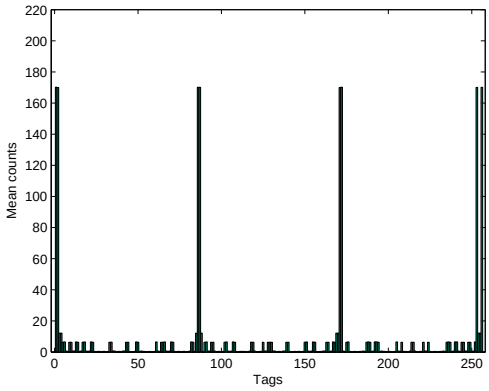
[11] Valkenborg,D. and Burzykowski,T. (2005) Report SAGE data analysis: Correcting sequence errors using methods of the moments., Technical Report, CenStat, Hasselt University.

[12] For an introduction on SAGE and a more detailed description on the working of SAGE:
http://www.sagenet.org
http://www.embl-heidelberg.de/info/sage/

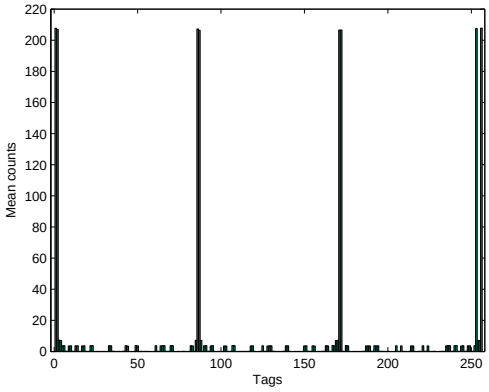
A Plots of section 4.3



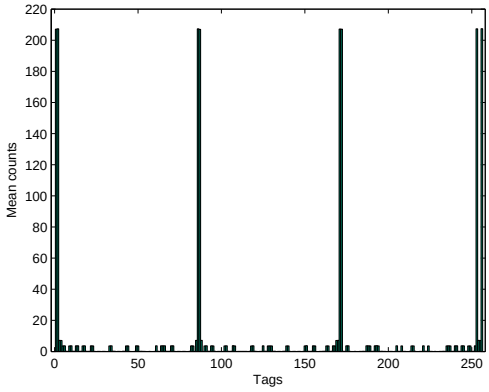
(a) 1 library: $\underline{\pi}_2$, Ω_1



(b) 20 libraries:$\underline{\pi}_2$, Ω_1

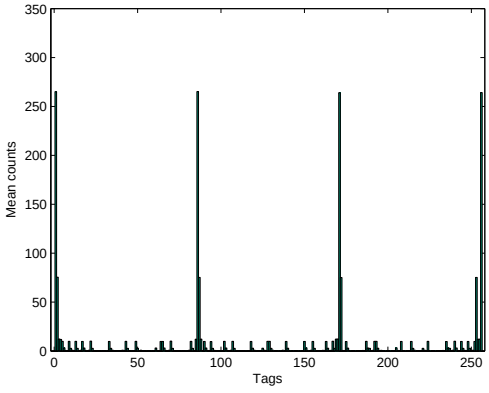


(c) 1 library: $\underline{\pi}_2$, Ω_2

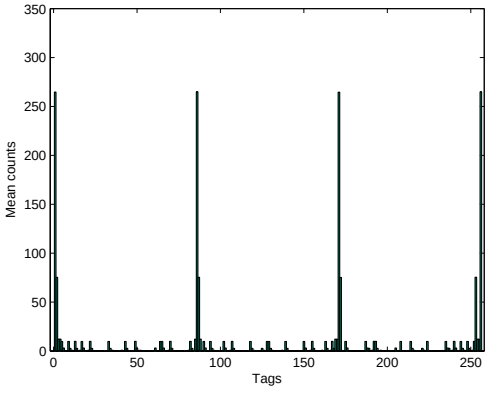


(d) 20 libraries:$\underline{\pi}_2$, Ω_2

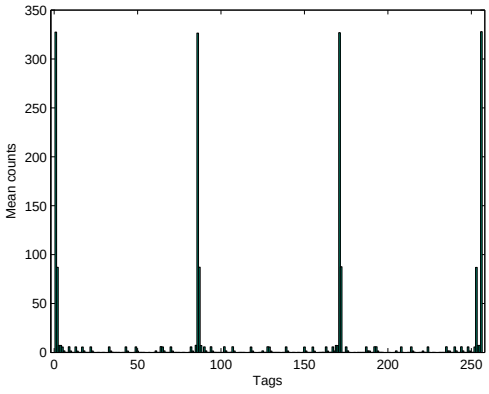
Figure 21: Mean of observed counts for $\underline{\pi}_2$



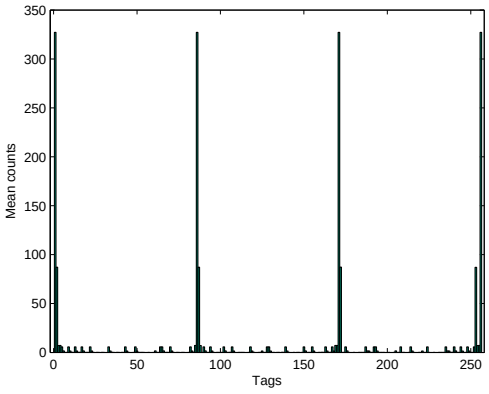
(a) 1 library: $\underline{\pi}_3$, Ω_1



(b) 20 libraries:$\underline{\pi}_3$, Ω_1



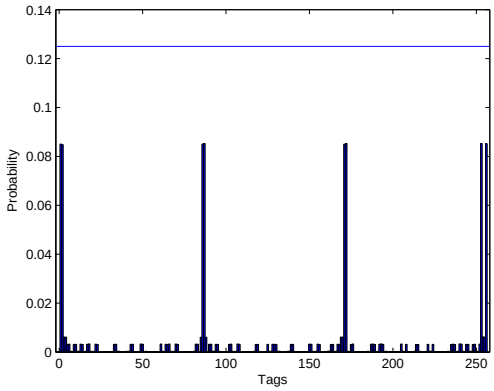
(c) 1 library: $\underline{\pi}_3$, Ω_2



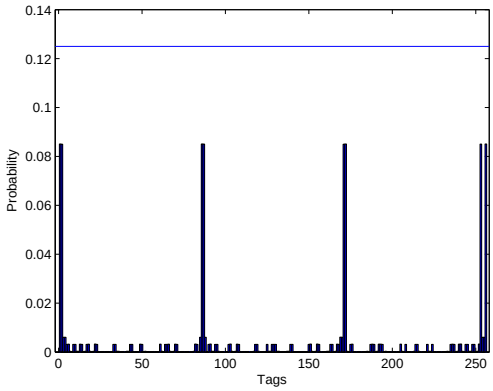
(d) 20 libraries:$\underline{\pi}_3$, Ω_2

Figure 22: Mean of observed counts for $\underline{\pi}_3$

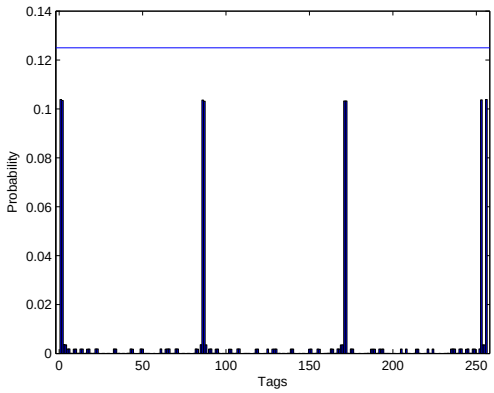
B Plots of section 5.1



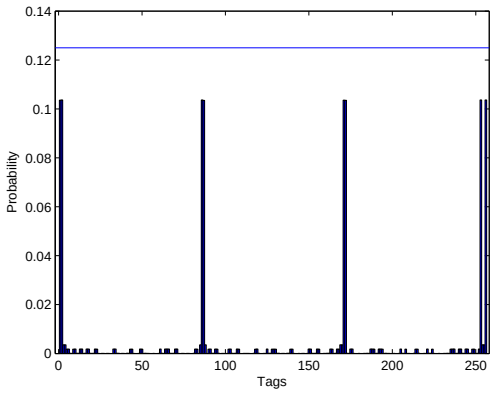
(a) 1 library: $\underline{\pi}_2$, Ω_1



(b) 20 libraries:$\underline{\pi}_2$, Ω_1

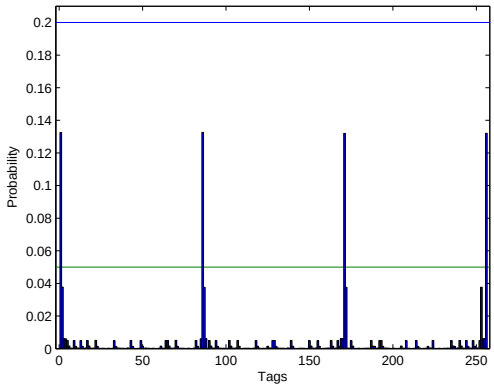


(c) 1 library: $\underline{\pi}_2$, Ω_2

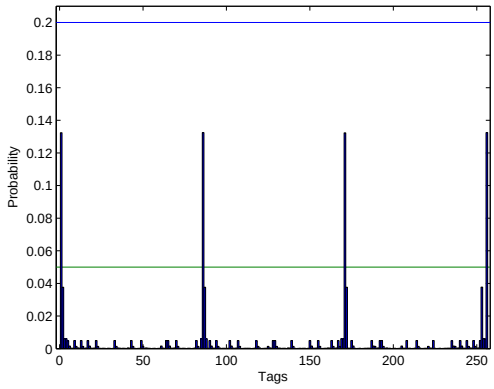


(d) 20 libraries:$\underline{\pi}_2$, Ω_2

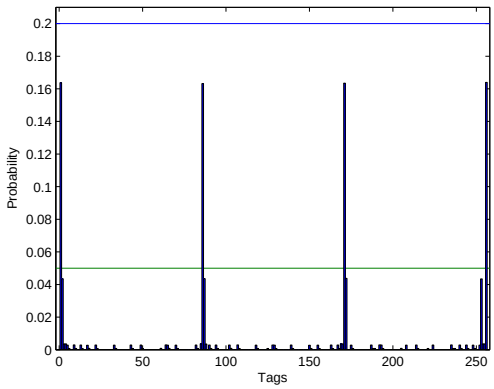
Figure 23: The observed expression probabilities $\hat{\underline{\pi}}_2^*$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 12.5%. The other true expression probabilities are equal to zero.



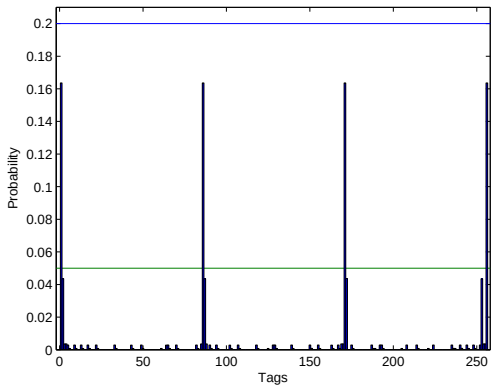
(a) 1 library: $\underline{\pi}_3$, Ω_1



(b) 20 libraries:$\underline{\pi}_3$, Ω_1



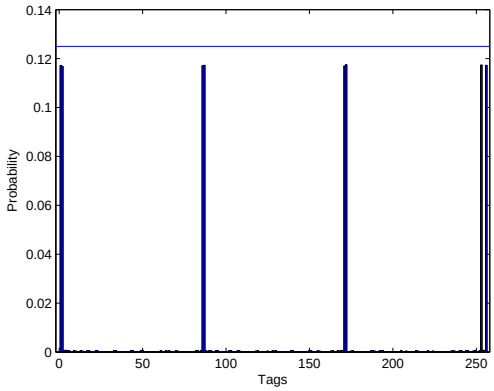
(c) 1 library: $\underline{\pi}_3$, Ω_2



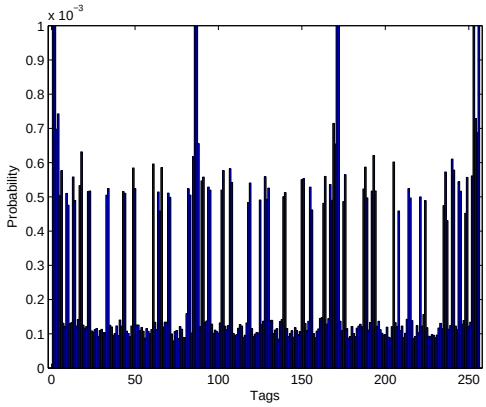
(d) 20 libraries:$\underline{\pi}_3$, Ω_2

Figure 24: The observed expression probabilities $\hat{\underline{\pi}}_3^*$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal lines at 20% and 5%. The other true expression probabilities are equal to zero.

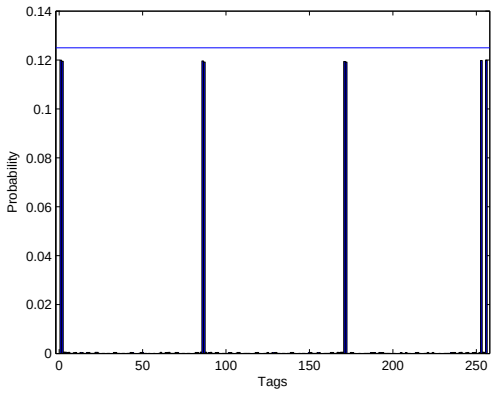
C Plots of section 5.2



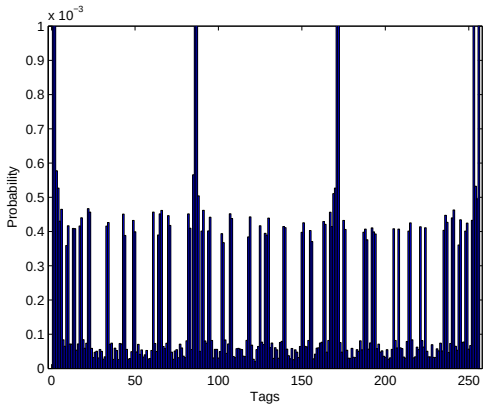
(a) 1 library: $\underline{\pi}_2$, $\hat{\Omega}_1$



(b) detail of (a)

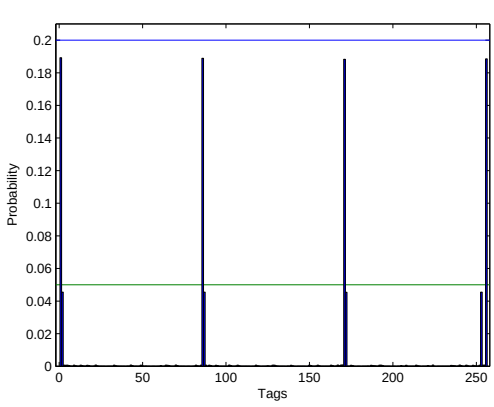


(c) 1 library: $\underline{\pi}_2$, $\hat{\Omega}_2$

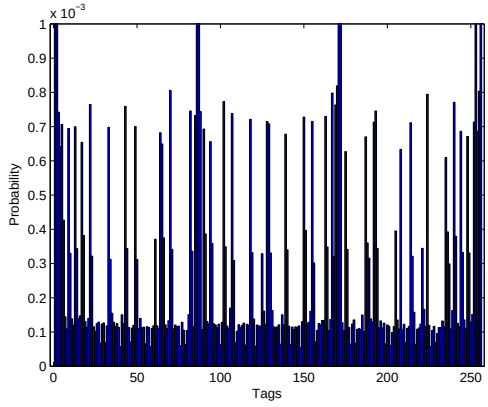


(d) detail of (c)

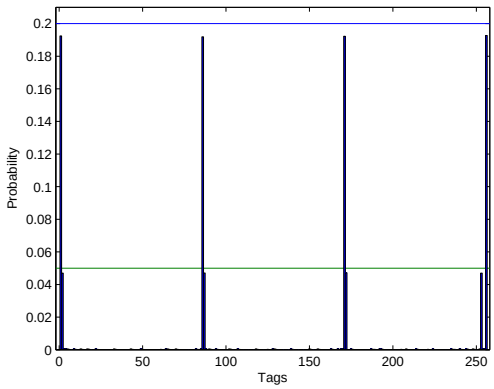
Figure 25: The estimated expression probabilities $\hat{\underline{\pi}}_2$. The true expression probabilities of the eight tags (see Table 1) are represented by the horizontal line at 12.5% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.



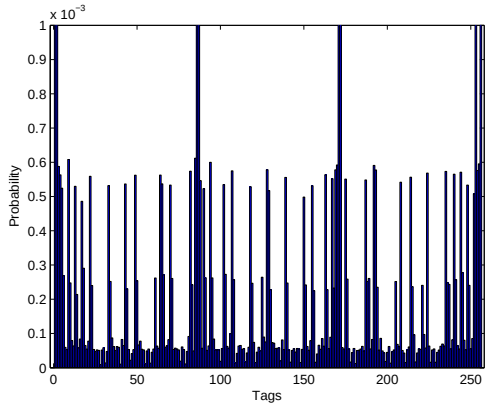
(a) 1 library: $\underline{\pi}_3$, $\hat{\Omega}_1$



(b) detail of (a)



(c) 1 library: $\underline{\pi}_3$, $\hat{\Omega}_2$



(d) detail of (c)

Figure 26: The estimated expression probabilities $\hat{\underline{\pi}}_3$. The true expression probabilities of the eight tags (see Table 1) are represented by the horizontal lines at 20% and 5% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.

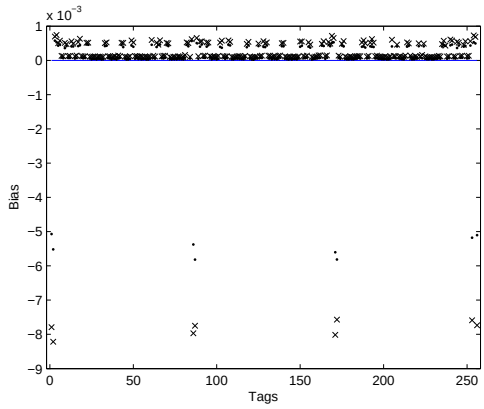


Figure 27: 1 library: Bias of $\underline{\hat{\pi}}_2$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. The crosses ($\times$) represent the bias of the estimates resulting from the use of $\hat{\Omega}_1$ and the dots ($\bullet$) the bias of the estimates resulting from the use of $\hat{\Omega}_2$.

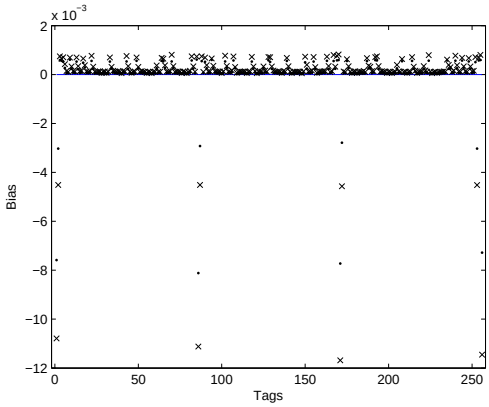
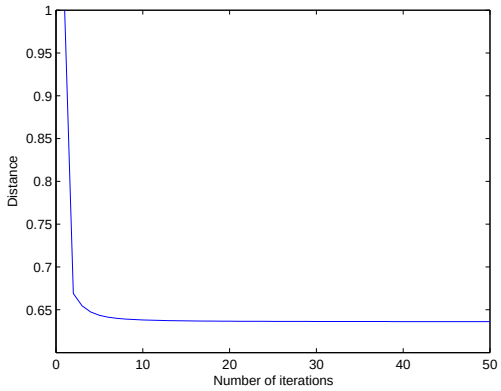
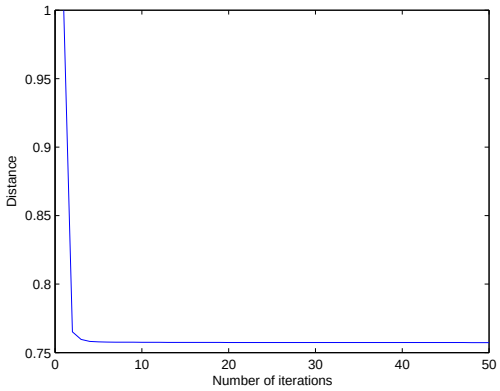


Figure 28: 1 library: Bias of $\underline{\hat{\pi}}_3$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. The crosses ($\times$) represent the bias of the estimates resulting from the use of $\hat{\Omega}_1$ and the dots ($\bullet$) the bias of the estimates resulting from the use of $\hat{\Omega}_2$.

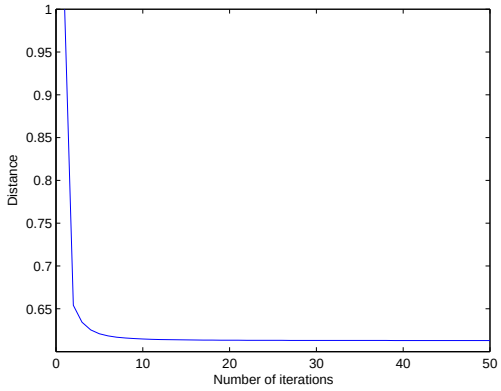


(a) 1 library: $\underline{\pi}_2$, $\hat{\Omega}_1$

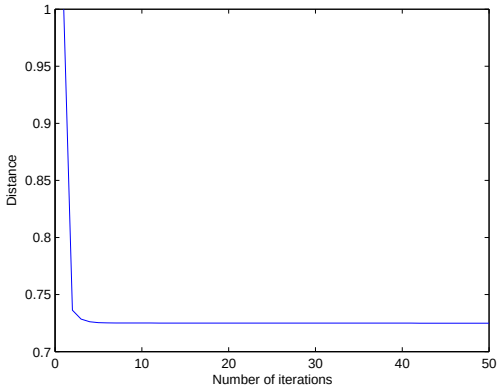


(b) 1 library: $\underline{\pi}_2$, $\hat{\Omega}_2$

Figure 29: Convergence monitoring through the relative distance for $\hat{\underline{\pi}}_2$

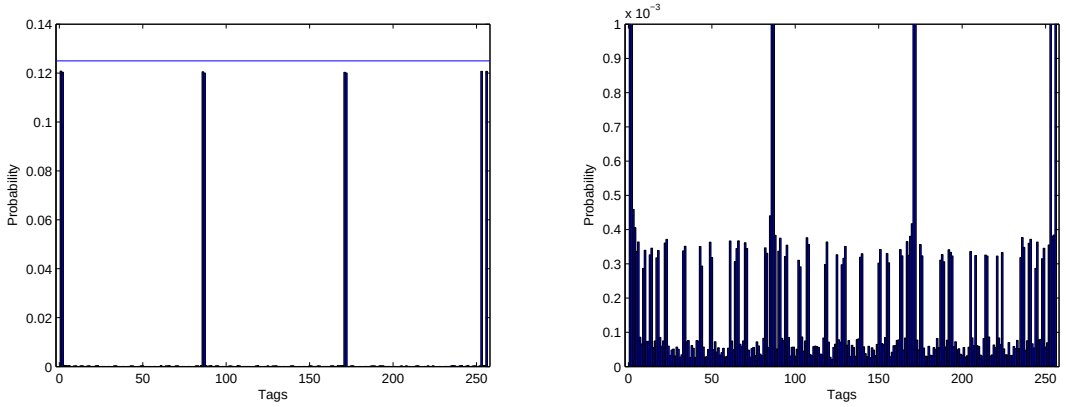


(a) 1 library: $\underline{\pi}_3$, $\hat{\Omega}_1$



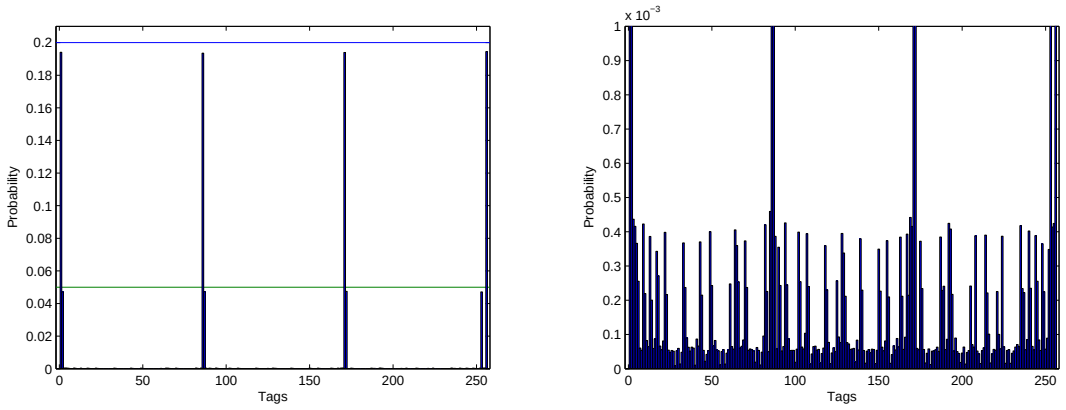
(b) 1 library: $\underline{\pi}_3$, $\hat{\Omega}_2$

Figure 30: Convergence monitoring through the relative distance for $\hat{\underline{\pi}}_3$



(a) (b) detail of (a)

Figure 31: The estimated expression probabilities $\underline{\hat{\pi}}_2$ for Ω_2. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 12.5% in subfigure (a). The other true expression probabilities are equal to zero.



(a) (b) detail of (a)

Figure 32: The estimated expression probabilities $\underline{\hat{\pi}}_3$ for Ω_2. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal lines at 20% and 5% in subfigure (a). The other true expression probabilities are equal to zero.

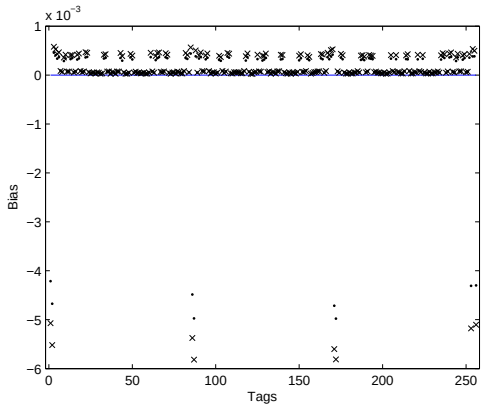


Figure 33: Bias comparison between $\Omega_2(\bullet)$ and $\hat{\Omega}_2(\times)$ for $\underline{\hat{\pi}}_2$

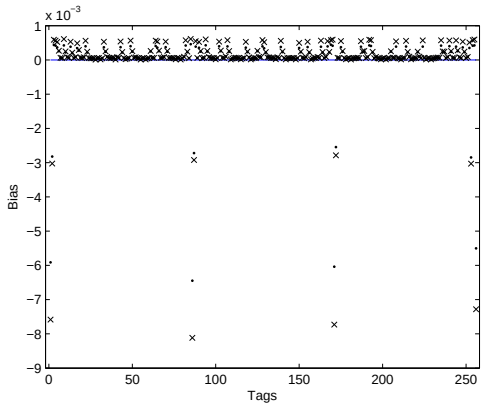
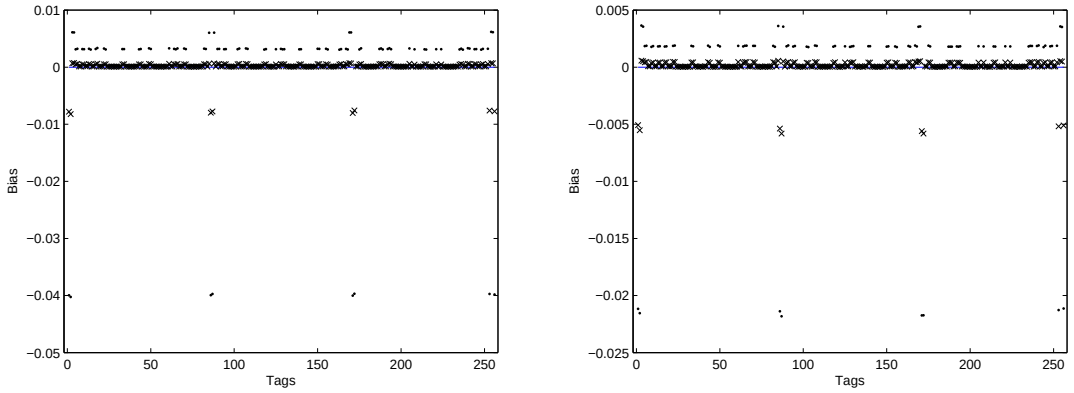
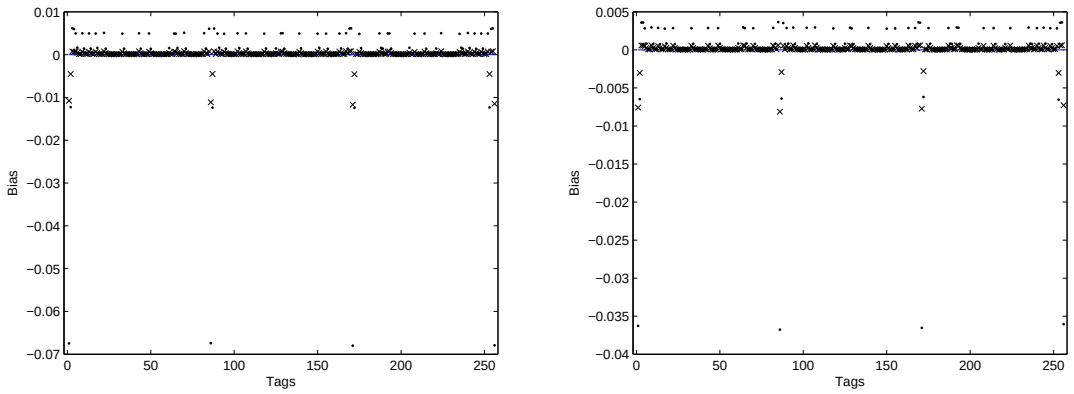


Figure 34: Bias comparison between $\Omega_2(\bullet)$ and $\hat{\Omega}_2(\times)$ for $\underline{\hat{\pi}}_3$



(a) Ω_1 (b) Ω_2

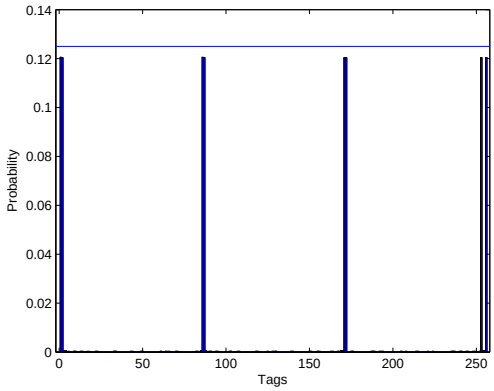
Figure 35: Bias of the observed expression probabilities (•) versus the estimate $\underline{\hat{\pi}}_2$ from the EM algorithm (×)



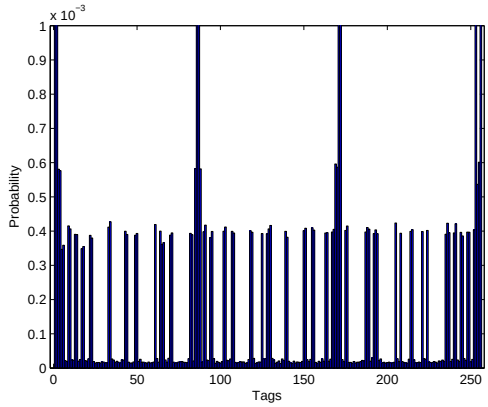
(a) Ω_1 (b) Ω_2

Figure 36: Bias of the observed expression probabilities (•) versus the estimate $\underline{\hat{\pi}}_3$ from the EM algorithm (×)

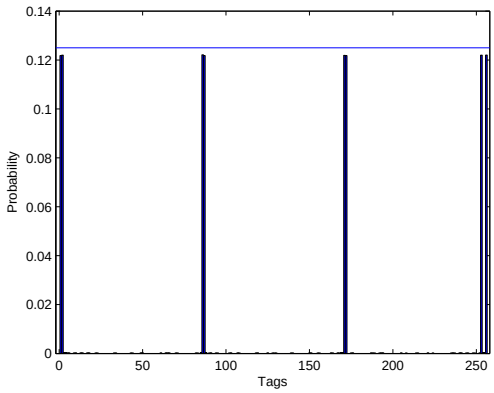
D Plots of section 5.3



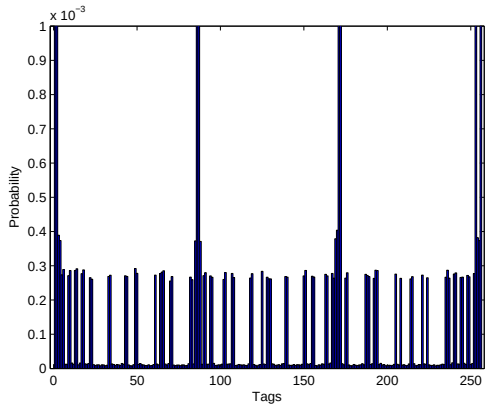
(a) 20 libs: $\underline{\pi}_2$, $\hat{\Omega}_1$



(b) detail of (a)

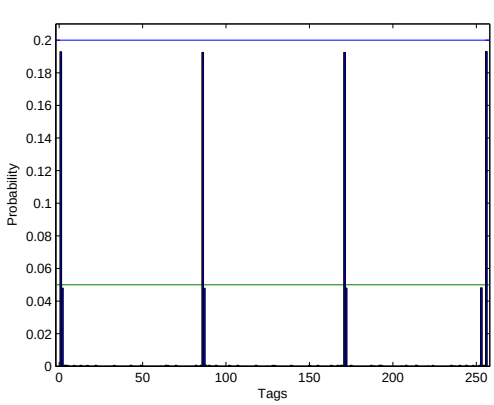


(c) 20 libs: $\underline{\pi}_2$, $\hat{\Omega}_2$

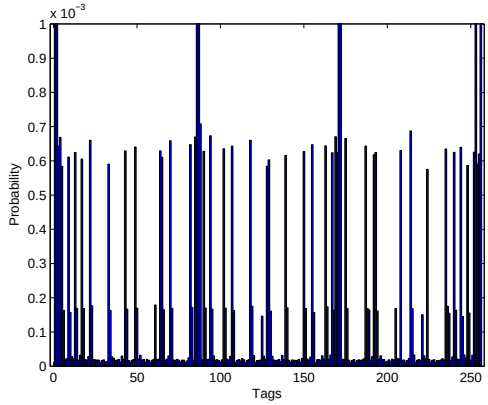


(d) detail of (c)

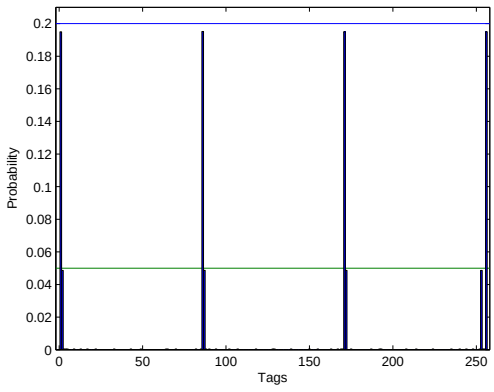
Figure 37: The estimated expression probabilities $\hat{\underline{\pi}}_2$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 12.5% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.



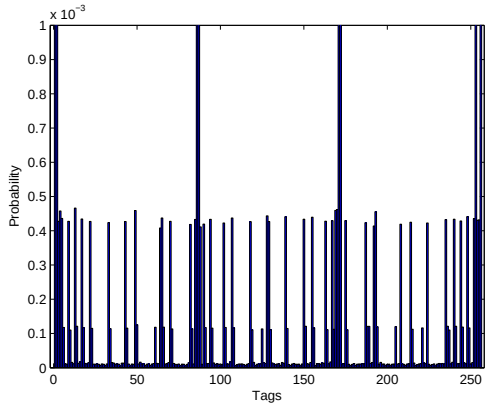
(a) 20 libs: $\underline{\pi}_3$, $\hat{\Omega}_1$



(b) detail of (a)



(c) 20 libs: $\underline{\pi}_3$, $\hat{\Omega}_2$



(d) detail of (c)

Figure 38: The estimated expression probabilities $\hat{\underline{\pi}}_3$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal lines at 20% and 5% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.

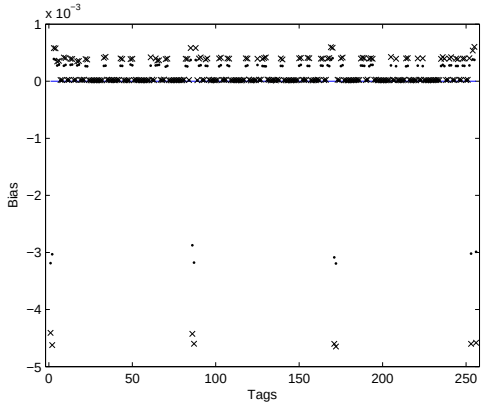


Figure 39: 20 libs: Bias of $\underline{\hat{\pi}}_2$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. The crosses ($\times$) represent the bias of the estimates resulting from the use of $\hat{\Omega}_1$ and the dots ($\bullet$) the bias of the estimates resulting from the use of $\hat{\Omega}_2$.

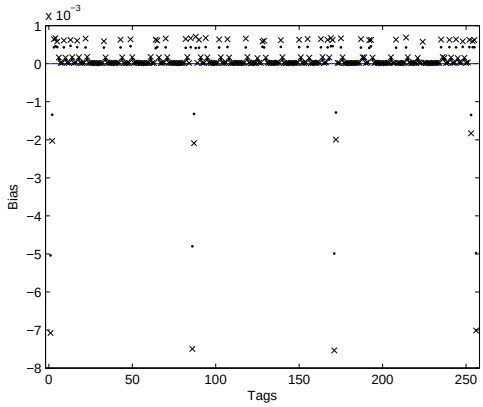
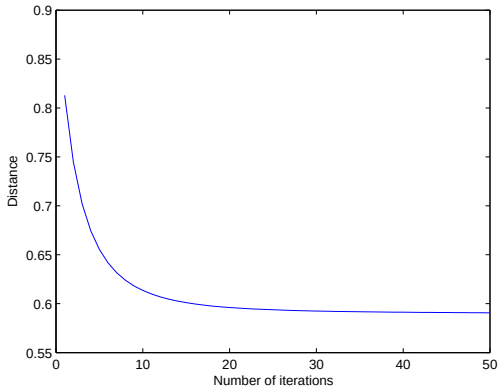
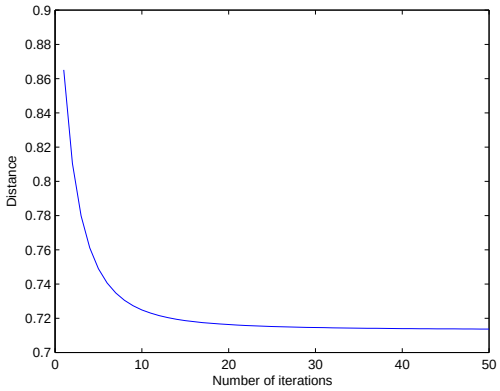


Figure 40: 20 libs: Bias of $\underline{\hat{\pi}}_3$ for $\hat{\Omega}_1$ versus $\hat{\Omega}_2$. The crosses ($\times$) represent the bias of the estimates resulting from the use of $\hat{\Omega}_1$ and the dots ($\bullet$) the bias of the estimates resulting from the use of $\hat{\Omega}_2$.

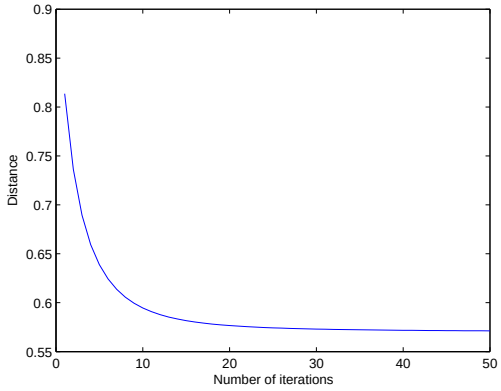


(a) 20 libs: $\underline{\pi}_2$, $\hat{\Omega}_1$

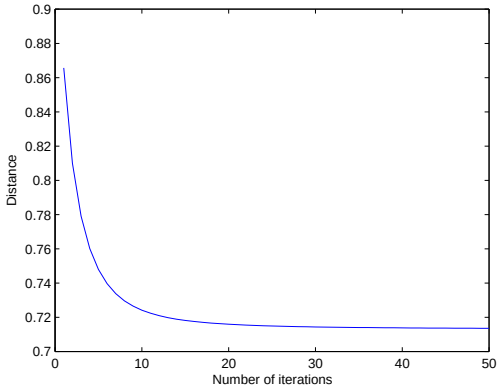


(b) 20 libs: $\underline{\pi}_2$, $\hat{\Omega}_2$

Figure 41: Convergence monitoring through the relative distance for $\hat{\underline{\pi}}_2$

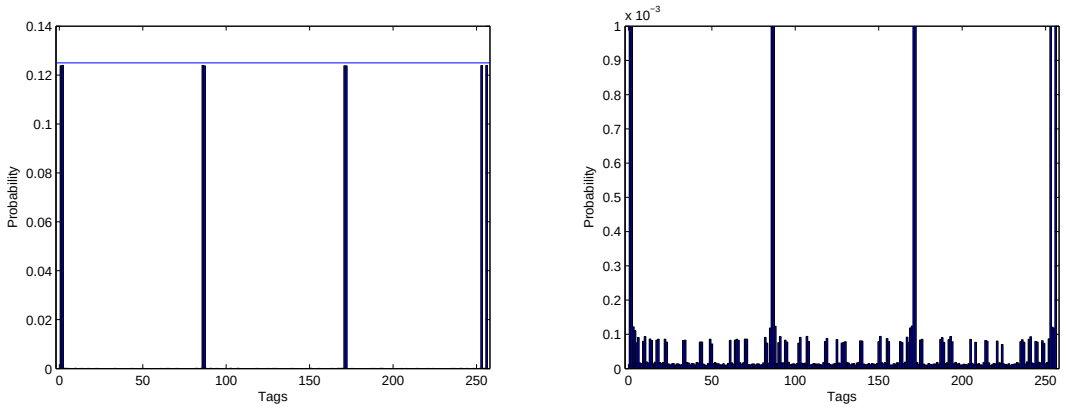


(a) 20 libs: $\underline{\pi}_3$, $\hat{\Omega}_1$



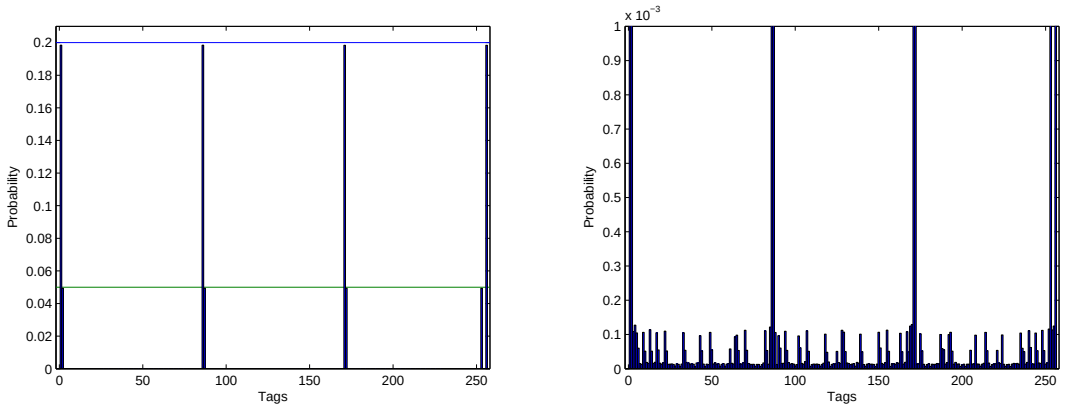
(b) 20 libs: $\underline{\pi}_13$, $\hat{\Omega}_2$

Figure 42: Convergence monitoring through the relative distance for $\hat{\underline{\pi}}_3$



(a) (b) detail of (a)

Figure 43: The estimated expression probabilities $\underline{\hat{\pi}}_2$ for Ω_2. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 12.5% in subfigure (a). The other true expression probabilities are equal to zero.



(a) (b) detail of (a)

Figure 44: The estimated expression probabilities $\underline{\hat{\pi}}_3$ for Ω_2. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal lines at 20% and 5% in subfigure (a). The other true expression probabilities are equal to zero.

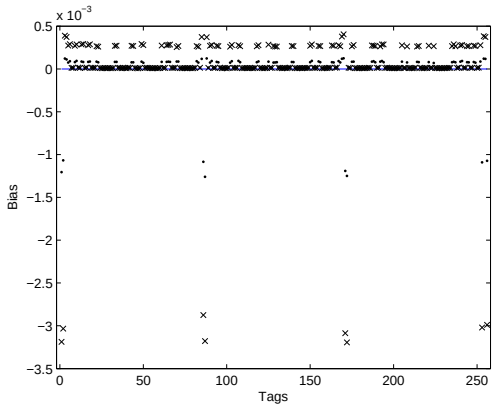


Figure 45: Bias comparison between $\Omega_2(\bullet)$ and $\hat{\Omega}_2(\times)$ for $\underline{\hat{\pi}}_2$

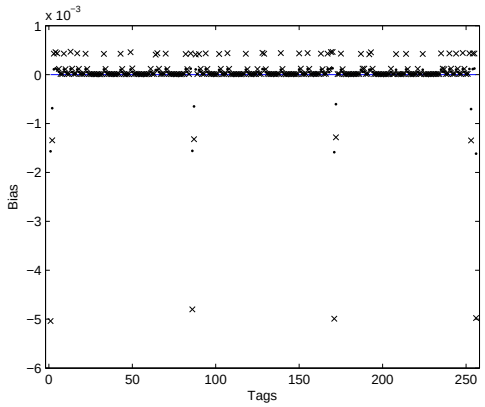
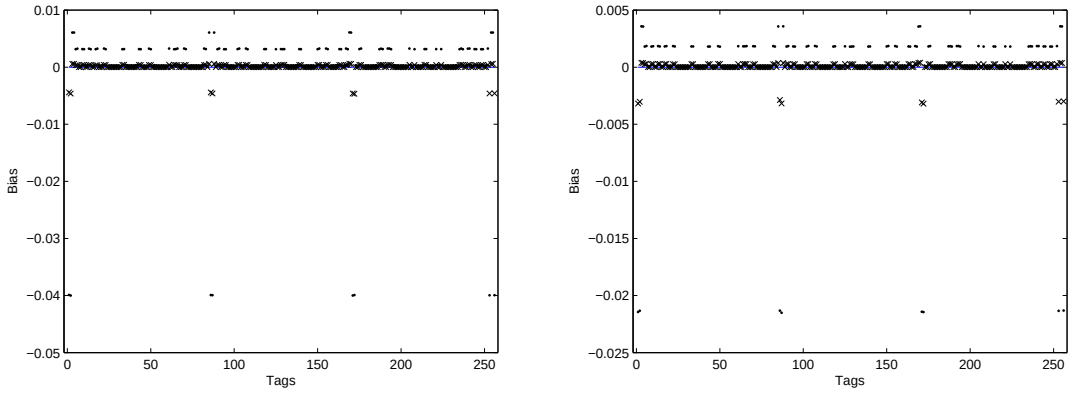
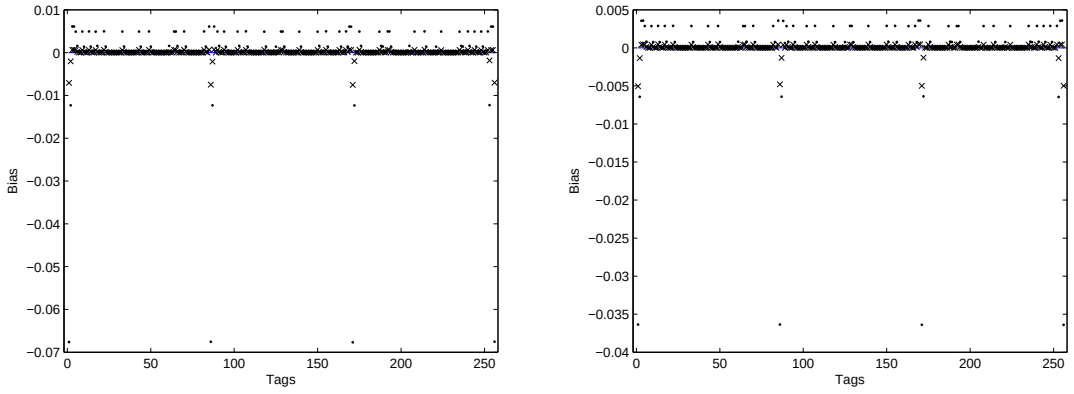


Figure 46: Bias comparison between $\Omega_2(\bullet)$ and $\hat{\Omega}_2(\times)$ for $\underline{\hat{\pi}}_3$



(a) $\hat{\Omega}_1$ (b) $\hat{\Omega}_2$

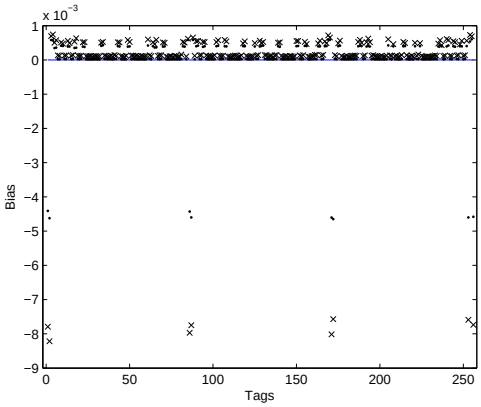
Figure 47: Bias of the observed expression probabilities (•) versus the estimate $\underline{\hat{\pi}}_2$ from the EM algorithm (×)



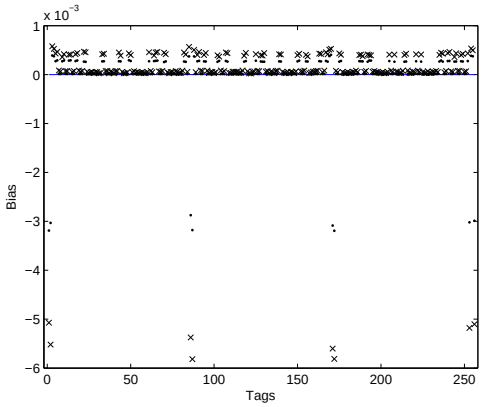
(a) $\hat{\Omega}_1$ (b) $\hat{\Omega}_2$

Figure 48: Bias of the observed expression probabilities (•) versus the estimate $\underline{\hat{\pi}}_3$ from the EM algorithm (×)

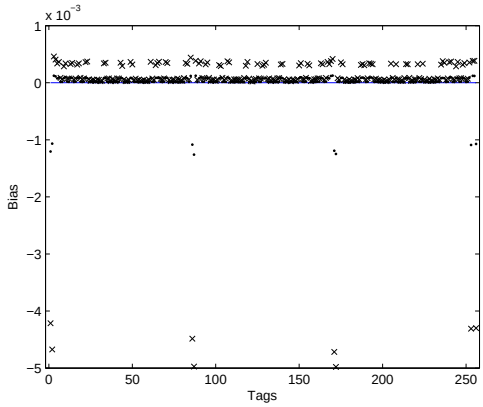
E Plots of subsection 5.4



(a) Bias of 1 library vs 20 libraries for $\hat{\Omega}_1$

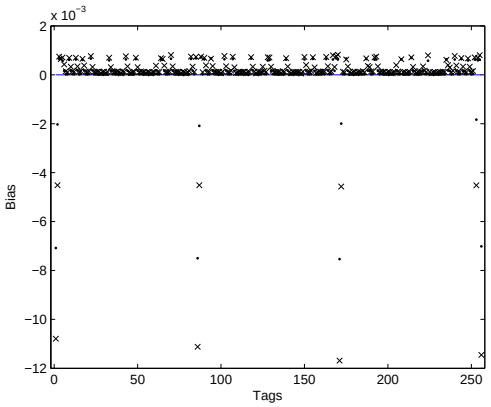


(b) Bias of 1 library vs 20 libraries for $\hat{\Omega}_2$

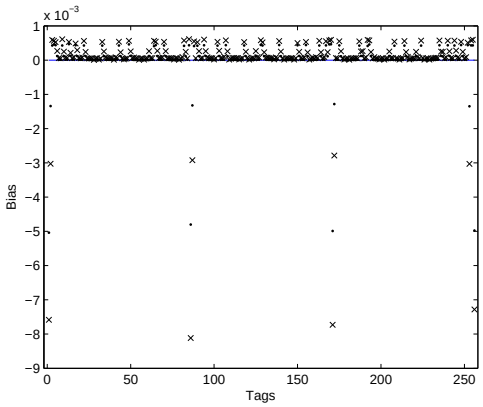


(c) Bias of 1 library vs 20 libraries for Ω_2

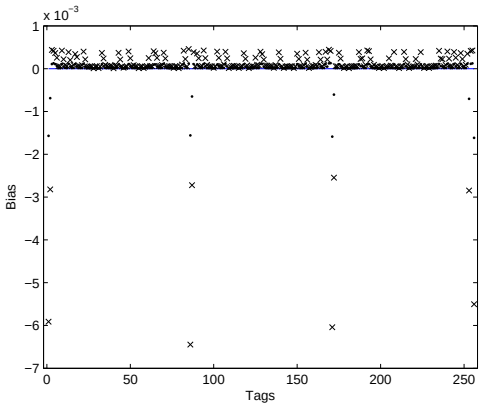
Figure 49: Bias of 1 library ($\times$) vs 20 libraries ($\bullet$) for $\hat{\Omega}_1$, $\hat{\Omega}_2$ and Ω_2 ($\underline{\hat{\pi}}_2$)



(a) Bias of 1 library vs 20 libraries for $\hat{\Omega}_1$



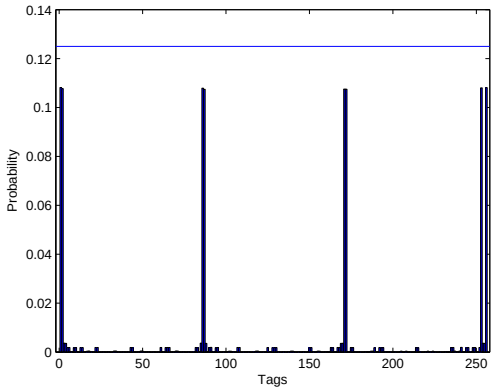
(b) Bias of 1 library vs 20 libraries for $\hat{\Omega}_2$



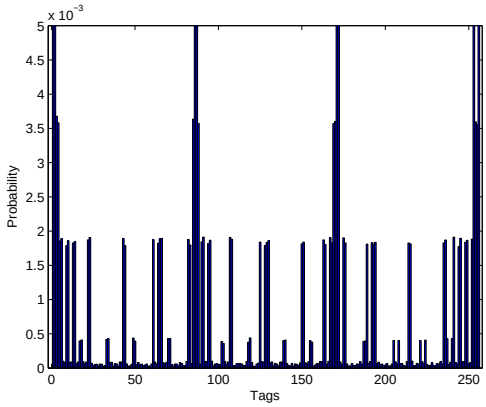
(c) Bias of 1 library vs 20 libraries for Ω_2

Figure 50: Bias of 1 library ($\times$) vs 20 libraries ($\bullet$) for $\hat{\Omega}_1$, $\hat{\Omega}_2$ and Ω_2 ($\underline{\hat{\pi}}_3$)

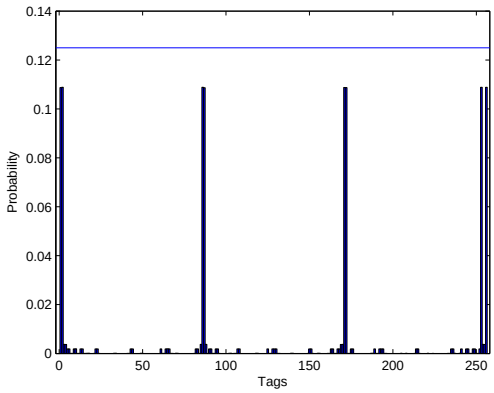
F Plots of section 5.5



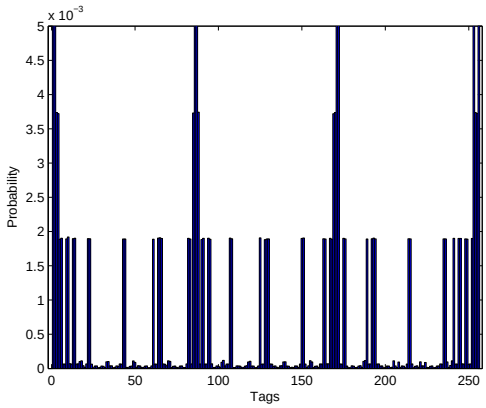
(a) 1 library: $\underline{\pi}_2$, Φ_B



(b) detail of (a)

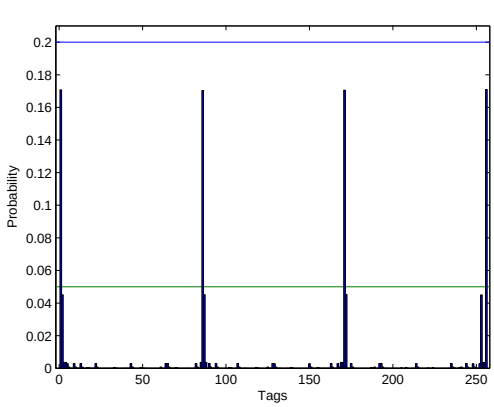


(c) 20 libraries: $\underline{\pi}_2$, Φ_B

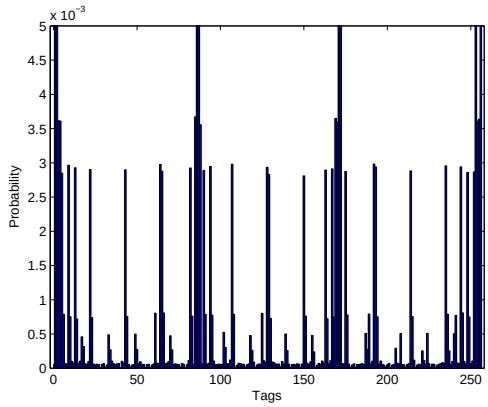


(d) detail of (c)

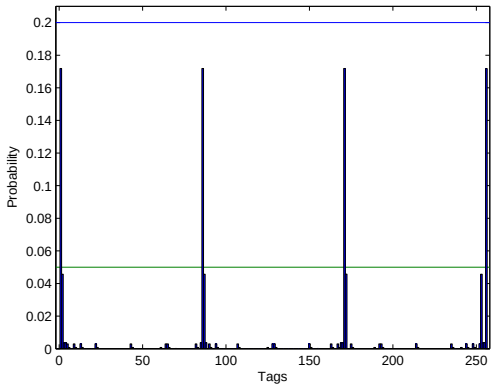
Figure 51: The estimated expression probabilities $\hat{\underline{\pi}}_2$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal line at 12.5% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.



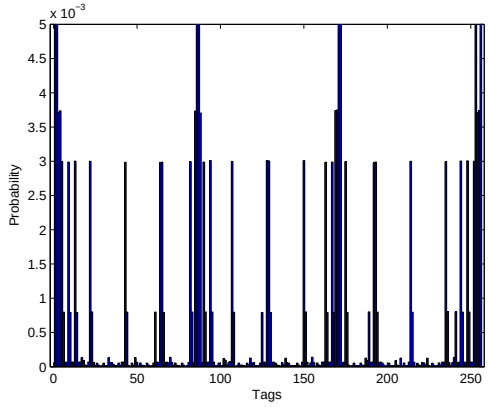
(a) 1 library: $\underline{\pi}_3$, Φ_B



(b) detail of (a)

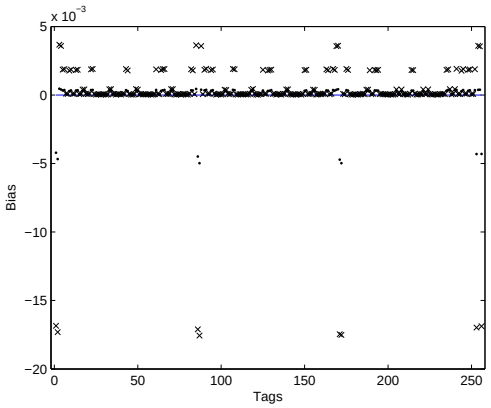


(c) 20 libraries: $\underline{\pi}_3$, Φ_B

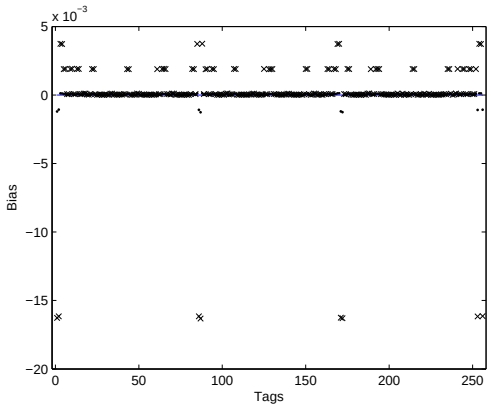


(d) detail of (c)

Figure 52: The estimated expression probabilities $\hat{\underline{\pi}}_3$. The true expression probabilities of the four tags (see Table 1) are represented by the horizontal lines at 20% and 5% in the subfigures (a) and (c). The other true expression probabilities are equal to zero.

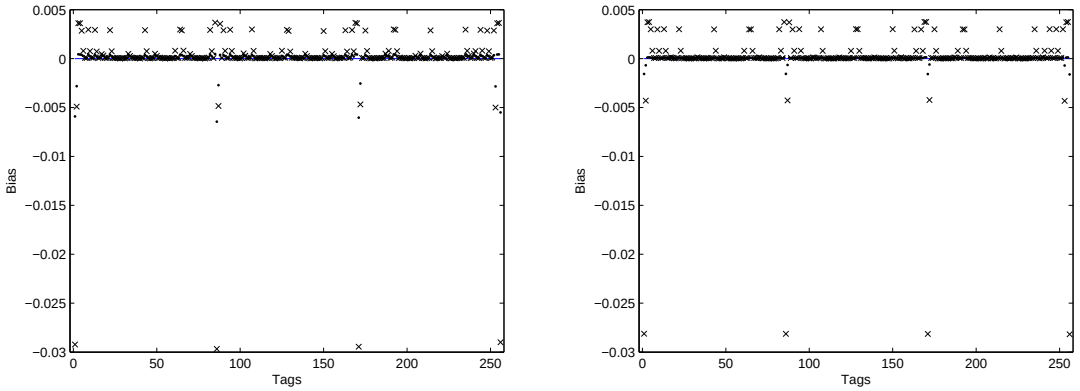


(a) 1 library: $\underline{\pi}_2$



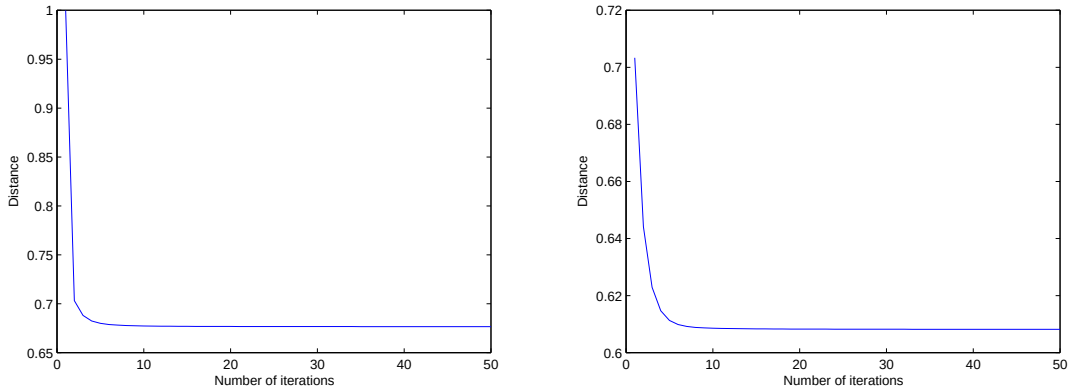
(b) 20 libraries: $\underline{\pi}_2$

Figure 53: Bias of $\hat{\underline{\pi}}_2$ for $\Phi_B(\times)$ versus $\Phi(\bullet)$



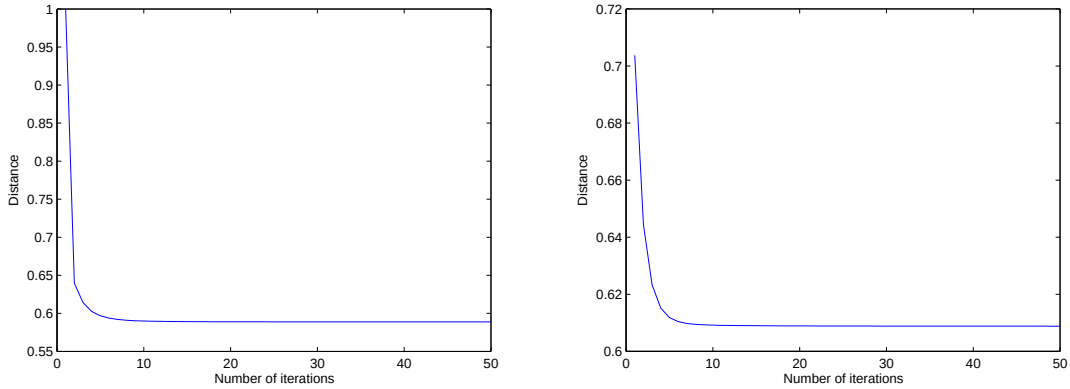
(a) 1 library: $\underline{\pi}_3$ (b) 20 libraries: $\underline{\pi}_3$

Figure 54: Bias of $\hat{\underline{\pi}}_3$ for $\Phi_B(\times)$ versus $\Phi(\bullet)$



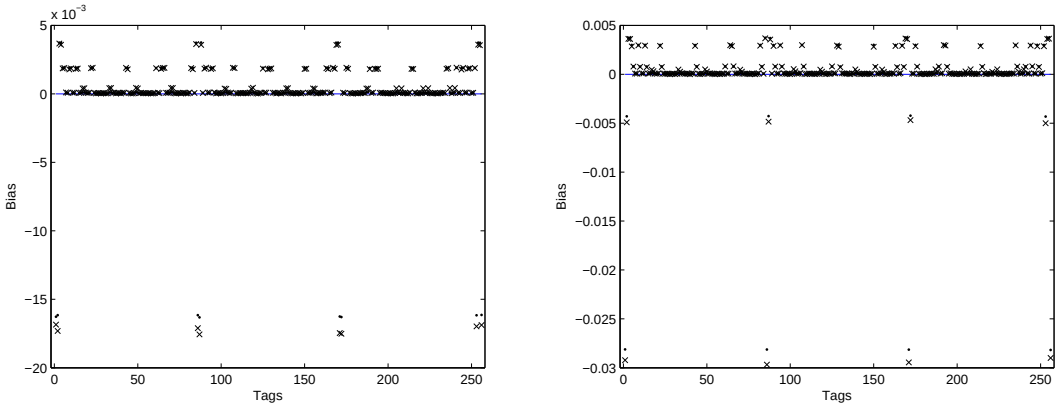
(a) 1 library: $\underline{\pi}_2$ (b) 20 libraries: $\underline{\pi}_2$

Figure 55: Convergence monitoring through the relative distance for $\hat{\underline{\pi}}_2$



(a) 1 library: $\underline{\pi}_3$ (b) 20 libraries: $\underline{\pi}_3$

Figure 56: Convergence monitoring through the relative distance for $\hat{\underline{\pi}}_3$



(a) $\underline{\hat{\pi}}_2$ (b) $\underline{\hat{\pi}}_3$

Figure 57: Bias of 1 library ($\times$) vs 20 libraries for $\Phi_B(\bullet)$

Auteursrechterlijke overeenkomst

Opdat de Universiteit Hasselt uw eindverhandeling wereldwijd kan reproduceren, vertalen en distribueren is uw akkoord voor deze overeenkomst noodzakelijk. Gelieve de tijd te nemen om deze overeenkomst door te nemen, de gevraagde informatie in te vullen (en de overeenkomst te ondertekenen en af te geven).

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
The EM-algorithm for modeling Serial analysis of Gene Expression (SAGE) data
Richting: **Master of Science in Biostatistics** Jaar: **2007**
in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Ik ga akkoord,

Michèle Ampe

Datum: **03.01.2007**