Front. Math. China 2011, 6(6): 1131–1145
DOI 10.1007/s11464-011-0123-z

RESEARCH ARTICLE

Generate gene expression profile from high-throughput sequencing data

Hui LIU[1], **Zhichao JIANG**[1], **Xiangzhong FANG**[1], **Hanjiang FU**[2], **Xiaofei ZHENG**[2], **Lei CHA**[3], **Wuju LI**[3]

1 School of Mathematical Sciences, Statistical Center, LMAM, Peking University, Beijing 100871, China
2 Beijing Institute of Radiation Medicine, Beijing 100850, China
3 Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing 100850, China

Abstract This work presents two methods, the Least-square and Bayesian method, to solve the multiple mapping problem in extracting gene expression profiles through the next-generation sequencing. We parallel the tag sequences to genome, and partition them to improving the methods' efficiency. The essential feature of these methods is that they can solve the multiple mapping problem between genes and short-reads, while generating almost the same estimation in single-mapping situation as the traditional approaches. These two methods are compared by simulation and a real example, which was generated from radiation-induced lung cancer cells (A549), through mapping short-reads to human ncRNA database. The results show that the Bayesian method, as realized by Gibbs sampler, is more efficient and robust than the Least-square method.

Keywords Next-generation sequencing, multiple mapping, Gibbs sampler, least-square, Bayesian
MSC 62F15, 62J05, 62P10

1 Introduction

Next-generation sequencing (NGS) is a high-throughout sequencing method with high sensitivity and specificity. It can produce an enormous volume of data precisely and cheaply without prior knowledge of a particular gene, and provide information regarding alternative splicing and sequence variation in identified

Received October 15, 2010; accepted March 9, 2011
Corresponding authors: Xiangzhong FANG, Wuju LI
E-mail: xzfang@math.pku.edu.cn, wujuli@yahoo.com

genes. Many important applications can be carried out by NGS, including transcription start site mapping, strand-specific measurements, gene fusion detection, small RNA characterization, and detection of alternative splicing events [7,10]. With the development of this new technology, the major commercial NGS platforms are Illumina, Roche 454, Helicos BioSciences, and Life Technologies, which constitute various strategies of template preparation, sequencing and imaging, genome alignment, and assembly approaches [4].

In the alignment approaches of NGS, the short-reads are either aligned to a reference genome or reference transcripts, or assembled *de novo* without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene. However, a significant portion of the short reads cannot be uniquely mapped to the reference genome or reference transcripts [3]. Most of studies are based on single-map reads, therefore, the multiple mapping problem is a challenge for NGS data analysis. One solution of this problem is applying some approaches to improve the single mapping rate, such as assigning these multi-matched reads proportionally based on the number of reads mapped to their neighboring unique sequences [2,3,6]. Another solution of the multiple mapping is adjusting the reference dataset, which is not proper to large transcriptomes, such as regarding miRNAs that have the same sequence are considered as a single sequence in the detection of the expression levels [9]. However, the expression profiles from above approaches are still generated through the improved single-mapping results, which is still not using the multiple-mapping data. For fully utilizing the information of sequenced short-reads, we will present two methods, the Least-square and Bayesian method, to extract gene expression profiles through the NGS with all the single-mapping and multiple-mapping data together. Additionally, in view of the fact that tens of thousands of short-reads can be obtained from one NGS experiment, mismatch or insertion or deletion is often disallowed in mapping short-reads to the reference genome or RNA sequence [5]. Here, we take the same strategy to process the short-reads data. The whole report is arranged as follows. In Section 2, we will introduce our NGS data and how serious the multiple mapping is. Section 3 is the data preparation. Section 4 provides two schemes to get the estimation of gene expression levels through the NGS. Section 5 shows the differences among the above methods through simulation and real examples. Section 6 concludes this paper with a brief discussion. The proofs are gathered in Appendix.

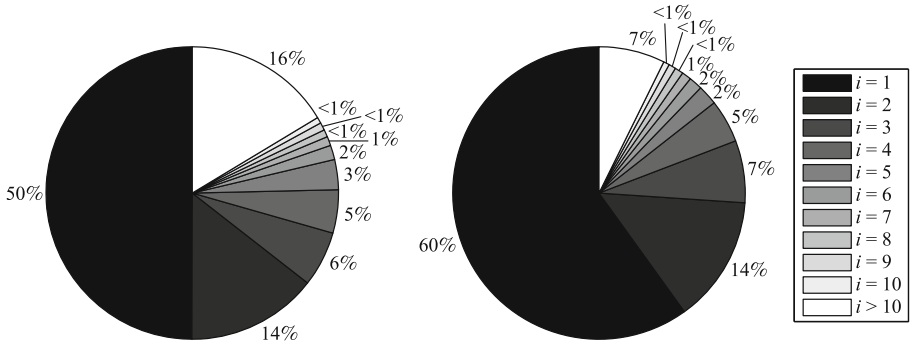
2 Problem description

Our NGS data was generated using Illumina sequencing technology from radiation-induced lung cancer cells (A549) (data are available in Additional file 1), and aligned with human ncRNA database by the Blast program.

The principle and procedure of the experiment are using beads of OligodT to enrich mRNA in the total RNA, and they are transferred into double-

stranded cDNA through reverse transcription. The enzyme NlaIII with 4-base recognition (CATG) is used to digest this cDNA, and Illumina adapter 1 is linked. Mmel is used to digest at 17bp downstream of CATG site; Illumina adapter 2 is linked at 3′ end. Primer GX1 and Primer GX2 are added for PCR. Then, regain 85bp strips through 6% TBE PAGE. The DNA is purified and followed by Solexa sequencing. Sequencing-received raw image data is transformed by base calling into sequence data, which is called raw data or raw short-reads, then these raw short sequences are transformed into clean tags after certain steps of data-processing. With the tens of thousands of clean short sequences (often called short-reads in this paper), we convert them into FastA format. Then, Blast program is used to compare these short-reads with human ncRNA database, and a perl program is designed to parsing the blast output file. If there is a hit, short read copy number and ncRNA name will be written into a txt file, based on which, we can find out the relationship between short-reads and human ncRNA sequences.

The Illumina sequencing uses NlaIII to recognize and cut off the CATG sites on cDNA, then Mmel is used to digest at 17bd downstream of CATG site. Hence, the length of human ncRNA sequences is much longer than the short-reads, and ncRNAs can have multiple CATG sites from the 5′ end while one short-read can be aligned to multiple ncRNAs. Therefore, the multiple mapping problem is serious in our data, such as one ncRNA can be aligned with various short-read sequences with count number varying from 1 to 523, while the most popular short-read is matched with 150 different ncRNAs. Fig. 1 provides the distributions of ncRNAs having multiple short-reads and short-reads mapping to multiple ncRNAs. There are only 50% of ncRNAs can be identified by one short-read, while 16% of ncRNAs have more than 10 short-reads. At the same time, there are 60% of short-reads are single mapping, while 7% of short-reads are multiple matched with 10 ncRNAs (data are available in Additional file 3).



(a) (b)

Fig. 1 Distribution of multiple mapping

(a) One ncRNA having i short-reads; (b) one short-read matched to i ncRNAs

If we take the sum of copy numbers of short-reads, which are single mapped to ncRNA, as the ncRNA expression levels, then only 1743 ncRNAs expressions

can be abstracted, which account only 42% of the whole expression profile. Since the multiple mapping problem is serious here, such as the short-read sequence `GATCACAACCAGTTACAGAT` is multiple matched with 150 ncRNAs in our dataset, we cannot compute the sum of short-reads' counts for one ncRNA as Wang et al. [9] did, who combine these miRNAs having the same short-reads as one sequence. Therefore, the existing expression profile abstracting approaches, which based on single-mapping, is not proper to our NGS dataset, and new methods needed to solve the multiple mapping problem.

3 Data preparation

Before presenting the expression profile abstracting approaches based on multiple mapping, we need some data preparation. Because one short-read can be multiple mapped to multiple genes and plenty of short-reads can be aligned to one gene, the form of our data is a matrix $A = \{a_{ij}\}_{n\times m}$, i.e.,

$$\begin{array}{c} \\ S_1 \\ S_2 \\ \vdots \\ S_n \end{array} \begin{array}{c} \begin{array}{cccc} G_1 & G_2 & \cdots & G_m \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \end{array},$$

where a_{ij} is the i-th short-read count generated from the j-th gene, $\{S_i\}_{i=1,\ldots,n}$ are the given counts of the short-read sequences, $\{G_j\}_{j=1,\ldots,m}$ are the gene expression levels we want to estimate, furthermore, $\{S_i\}_{i=1,\ldots,n}$ are the row sums of A, and $\{G_j\}_{j=1,\ldots,m}$ are the column sums of A.

Furthermore, when aligning the short-read sequences with the genes, another matrix $T = \{T_{ij}\}_{n\times m}$ is got to identify whether one short-read is aligned to one gene or not. If the i-th short-read sequence is aligned with the j-th gene, $T_{ij} = 1$; otherwise, $T_{ij} = 0$ (the results for our dataset are available in Additional file 2).

However, our reference transcripts is human ncRNA database, therefore, matrices A and T denote the relationship between short-reads and human ncRNA sequences. Since the new methods can be used to different kind of NGS data, we use gene expression in methods introduction and ncRNA sequences expression only in real data analysis.

The main objective is to estimate the gene expression levels $\{G_j\}_{j=1,\ldots,m}$ by the given $\{S_i\}_{i=1,\ldots,n}$ and $T = \{T_{ij}\}_{n\times m}$. There are thousands of genes and short-reads in one sample, and hence, T is a huge sparse matrix and it is difficult to solve directly. To improving the question, we partition the genes into subgroups. The data is partitioned by separating the genes sharing short-reads into the same subgroup and corresponding short-reads forming the short-reads' subgroup. The genes and short-reads are divided into subgroups according the following criterions. (1) If two genes are matched to the same short-reads, they

belong to the same genes' subgroup, and those short-reads matched to either of them belong to the same short-reads' subgroup. (2) Different subgroups have no common genes or sequences. Then T can be rewritten as a diagonal matrix:

$$T = \mathrm{diag}(T_1, T_2, T_3, \ldots),$$

where T_r is an $n_r \times m_r$ matrix, which presenting the relationship between the r-th gene subgroup G^r and the corresponding r-th short-read sequences group S^r. Therefore, our original dataset is separated into 2014 subgroups, and the largest subgroup has 1268 ncRNAs and 6375 short-reads.

4 Schemes to get estimation of gene expression levels

4.1 Nonnegative constrained least-square estimation

In data preparation, we partitioned genes and short-read sequences into non-intersect subgroups. Therefore, the following methods are focused on the gene expression levels' estimation of each subgroup. For the r_{th} genes' subgroup G^r and its corresponding short-reads' subgroup S^r, we naturally try to estimate the gene expression levels by solving the equation

$$S^r = T^r G^r.$$

However, we cannot use this equation directly, the estimation of G^r does not satisfy the precondition that the sum of gene expression levels' estimations must equal the sum of short-reads' counts. Therefore, some translations are made by dividing each element in T^r by the sum of the column which this element in. With the transformed matrix $\widetilde{T}^r$, the gene expression levels are generated by the nonnegative constrained least squares estimation (NCLSE), while the answers of the equation $S^r = \widetilde{T}^r G^r$ satisfy the precondition.

Since the relations between the numbers of genes and short-reads in subgroups are different, in our dataset, about 7% subgroups are indeterminate equations, i.e., $n_r < m_r$, which NCLSE cannot be used. And about 93% subgroups can be solved by NCLSE, since they have $n_r \geqslant m_r$. However, about 3% of the NCLSE cannot work since $(\widetilde{T}^r \widetilde{T}^{r\mathrm{T}})^{-1}$ does not exist, and the largest subgroup of real dataset is always in this situation. Therefore, NCLSE can be used in a part of subgroups, and only about 50% ncRNAs expression levels are estimated. However, NCLSE is better than the traditional single-mapping approaches, which estimates the expression profile by the sum of single-mapping short-reads copy numbers and only gets 42% of the expression profile. Furthermore, we can prove that these 42% of the expression profile, got from traditional approaches, can be solved by NCLSE, and the results of these two methods are equal by single-mapping data (the proof is given in Appendix).

4.2 Bayesian method

Based on the above analysis, the NCLSE can solve half of the expression profile, and hence, we try to work by Bayesian method which uses a data augmentation algorithm and works in all subgroups.

4.2.1 Distribution assumptions

Let $A^r = \{a^r_{ij}\}_{i=1,\ldots,n_r, j=1,\ldots,m_r}$ be our augmented data for the r_{th} subgroup, where a^r_{ij} means the i-th short-read's count generated from the j-th gene. The row sums of A^r are the given counts of the short-read sequences and the column sums are the gene expression levels, i.e.,

$$\sum_{i=1}^{n_r} a^r_{ij} = G^r_j, \quad \sum_{j=1}^{m_r} a^r_{ij} = S^r_i.$$

Suppose that each row of A^r is sampled from a Multinomial distribution, i.e.,

$$\begin{aligned}(a^r_{i1}, a^r_{i2}, \ldots, a^r_{im_r}) &\sim Pr(a^r_{ij} \mid q^r_{ij},\, S^r_i,\, j = 1, 2 \ldots, m_r)\\ &= \text{Multinomial}(q^r_{i1}, q^r_{i2}, \ldots, q^r_{i,m_r-1}, S^r_i),\end{aligned}$$

and the parameters' prior distribution is Dirichlet distribution, i.e.,

$$\begin{aligned}(q^r_{i1}, q^r_{i2}, \ldots, q^r_{im_r}) &\sim Pr(q^r_{ij} \mid \beta^r_{ij},\, j = 1, 2, \ldots, m_r)\\ &= \text{Dirchlet}(\beta^r_{i1}, \beta^r_{i2}, \ldots, \beta^r_{im_r}),\end{aligned}$$

where the initial value

$${}^0\beta^r_{ij} = \frac{T_{ij}}{\sum_{j=1}^{m_r} T_{ij}},$$

which means that the probability of i-th short-read generated from the j-th gene is inversely proportional to the number of short-reads aligned to the j-th gene. Then the posterior distributions of the parameters and A^r are

$$\begin{aligned}&q^r_{i1}, q^r_{i2}, \ldots, q^r_{im_r} \mid a^r_{ij}, \beta^r_{ij},\, j = 1, \ldots, m_r\\ &\quad\sim \text{Dirichlet}(\beta^r_{i1} + a^r_{i1}, \beta^r_{i2} + a^r_{i2}, \ldots, \beta^r_{im_r} + a^r_{im_r}),\end{aligned}$$

and

$$\begin{aligned}&a^r_{i1}, a^r_{i2}, \ldots, a^r_{im_r} \mid q^r_{ij}, \beta^r_{ij},\, j = 1, \ldots, m_r\\ &\quad\sim \text{Multinomial}(q^r_{i1}, q^r_{i2}, \ldots, q^r_{i,m_r-1}, S^r_i),\end{aligned}$$

while the proof is given in Appendix.

Following is an example of Bayesian method. As showed in Fig. 2, the genes' subgroup has two genes $G^r = \{G^r_1, G^r_2\}$ and the correspondingly short-reads' group has three short-reads $S^r = \{S^r_1, S^r_2, S^r_3\}$. Then the matrices

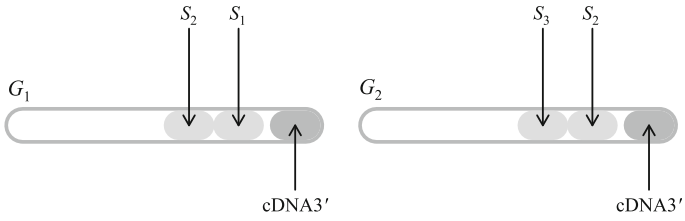


Fig. 2 Example of genes sharing same short-reads

$$T^r = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad A^r = \begin{pmatrix} a_{11}^r & 0 \\ a_{21}^r & a_{22}^r \\ 0 & a_{32}^r \end{pmatrix}.$$

We have the following results that

$$a_{11}^r = S_1^r, \quad a_{21}^r + a_{22}^r = S_2^r, \quad a_{32}^r = S_3^r$$
$$a_{11}^r + a_{21}^r = G_1^r, \quad a_{22}^r + a_{32}^r = G_2^r,$$

the prior distribution of the parameters is

$$(q_{21}^r, q_{22}^r) \sim \text{Dirchlet}(0.5, 0.5),$$

and the posterior distributions of the parameters and A^r are

$$(q_{21}^r, q_{22}^r) \sim \text{Dirchlet}(0.5 + a_{21}^r, 0.5 + a_{22}^r),$$
$$(a_{21}^r, a_{22}^r) \sim \text{Multinomial}(q_{21}^r, q_{22}^r).$$

4.2.2 Gibbs sampling

Next, we use the Gibbs sampling method to get the Bayesian estimation. With the prior distribution

$$(q_{i1}^r, q_{i2}^r, \ldots, q_{im_r}^r) \sim \text{Dirchlet}\left({}^0\beta_{i1}^r = \frac{T_{i1}}{\sum_{j=1}^{m_r} T_{ij}}, {}^0\beta_{i2}^r = \frac{T_{i2}}{\sum_{j=1}^{m_r} T_{ij}}, \ldots, {}^0\beta_{im_r}^r = \frac{T_{im_r}}{\sum_{j=1}^{m_r} T_{ij}}\right),$$

the l-th iteration of Gibbs sampling is

Step 1 draw ${}^la_{i\cdot}^r$ from

$$Pr({}^la_{i\cdot}^r \mid {}^{l-1}q_{i\cdot}^r, S_i^r) = \text{Multinomial}({}^{l-1}q_{i\cdot}^r, S_i^r), \quad i = 1, 2, \ldots, n_r;$$

Step 2 draw ${}^lq_{i\cdot}^r = ({}^lq_{i1}^r, {}^lq_{i2}^r, \ldots, {}^lq_{im_r}^r)$ from

$$Pr({}^lq_{i\cdot}^r \mid {}^{l-1}a_{i\cdot}^r, S_i^r) = \text{Dirchlet}({}^{l-1}\beta_{i\cdot}^r + {}^la_{i\cdot}^r), \quad i = 1, 2, \ldots, n_r;$$

Step 3

$${}^l\beta_{ij}^r = {}^{l-1}\beta_{ij}^r + {}^{l-1}a_{ij}^r, \quad i = 1, 2, \ldots, n_r, \ j = 1, 2, \ldots, m_r.$$

Under the regularity conditions [8], the Gibbs sampler will converge after enough iteration, and we parallel independent run the Gibbs sampler to realize an i.i.d. sample ${}^1A^r, {}^2A^r, \ldots, {}^KA^r$ from the posterior distribution of A^r. Furthermore, Benjamini-Hochberg step-up procedure [1] is used for the multiple testing problem in the iteration stop criteria, while the iteration always less than 100 times by our real data.

4.2.3 Estimation

Since the gene expression levels $\{G_j^r\}_{j=1,\ldots,m_r}$ are the row sums of A^r, we

generate an i.i.d. sample of $\{G_j^r\}_{j=1,\ldots,m_r}$ by the row sums of ${}^1A^r, {}^2A^r, \ldots, {}^KA^r$, respectively, i.e.,

$$ {}^kG_j^r = \sum_{i=1}^{n_r} {}^ka_{ij}^r, \quad j = 1, \ldots, m_r, \ k = 1, \ldots, K. $$

Then the posterior moments and marginal distributions of the gene expression levels $\{G_j^r\}_{j=1,\ldots,m_r}$ can be estimated by its i.i.d. sample.

We estimate the gene expression levels by its posterior moments, i.e.,

$$ \hat{G}_j^r = \frac{1}{K}\sum_{k=1}^{K} {}^kG_j^r = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_r} {}^ka_{ij}^r, \quad j = 1, \ldots, m_r, $$

and get the following convergence theory:

$$ \hat{G}_j^r \overset{\text{a.s.}}{\to} E\left(\sum_{i=1}^{n_r} a_{ij}^r\right) = \sum_{i=1}^{n_r} S_i^r q_{ij}^r, \quad K \to \infty, \ j = 1, \ldots, m_r, $$

and the equation

$$ \sum_{j=1}^{m_r} \hat{G}_j^r = \sum_{i=1}^{n_r} S_i^r. $$

These results indicate that the sum of our gene expression levels' estimation is equal to the sum of short-reads' counts, which satisfies the preconditions of the question.

Since $A^r = (A_1^r, A_2^r, \ldots, A_{N_r}^r)^{\mathrm{T}}$ and the marginal distribution of $A_i^r = (a_{i1}^r, a_{i2}^r, \ldots, a_{im_r}^r)$ is Multinomial$(q_{i1}^r, q_{i2}^r, \ldots, q_{i,m_r-1}^r, S_i^r)$, the marginal means of A_i^r are $E_i^r = S_i^r(q_{i1}^r, q_{i2}^r, \ldots, q_{im_r}^r)$ and the marginal variance-covariance matrices of A_i^r are

$$ V_i^r = S_i^r \begin{pmatrix} q_{i1}^r(1-q_{i1}^r) & -q_{i1}^r q_{i2}^r & \cdots & -q_{i1}^r q_{im_r}^r \\ -q_{i2}^r q_{i1} & q_{i2}^r(1-q_{i2}^r)^r & \cdots & -q_{i2}^r q_{im_r}^r \\ \vdots & \vdots & & \vdots \\ -q_{im_r}^r q_{i1}^r & -q_{im_r}^r q_{i2}^r & \cdots & q_{im_r}^r(1-q_{im_r}^r) \end{pmatrix}, \quad i = 1, \ldots, n_r. $$

Since*

$$ G^r = (G_1^r, G_2^r, \ldots, G_{n_r}^r)^{\mathrm{T}} = [I_{n_r\times n_r} \otimes 1_{n_r\times 1}^{\mathrm{T}}](A_1^r, A_2^r, \ldots, A_{n_r}^r)^{\mathrm{T}}, $$

the mean of $\hat{G}^r = (\hat{G}_1^r, \hat{G}_2^r, \ldots, \hat{G}_{n_r}^r)^{\mathrm{T}}$ is

$$ E^r = [I_{n_r\times n_r} \otimes 1_{n_r\times 1}^{\mathrm{T}}](E_1^r, E_2^r, \ldots, E_{n_r}^r)^{\mathrm{T}}, $$

*The Kronecker product $A \otimes B$ is defined as the partitioned matrix $(a_{ij}B)_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n}$.

the variance-covariance matrix of $\hat{G}^r$ is

$$V^r = [I_{n_r \times n_r} \otimes 1^{\mathrm{T}}_{n_r \times 1}] \operatorname{diag}(V^r_1, V^r_2, \ldots, V^r_{n_r}) [I_{n_r \times n_r} \otimes 1^{\mathrm{T}}_{n_r \times 1}]^{\mathrm{T}},$$

and the normal approximation is

$$\hat{G}^r \xrightarrow{\mathrm{d}} N(E^r, V^r), \quad K \to \infty.$$

5 Comparison and examples

5.1 Comparison of presumptions

Actually, the above two methods have their presumptions respectively. When transforming the matrix T^r for the NCLSE, we assumed that if one gene have multiple short-reads, then each short-read has the same possibility to be generated from this gene, which is true in single-mapping situation. On the other hand, when supposing the initial value ${}^0\beta^r_{ij}$ for the parameters' prior, we assumed that the probability of i-th short-read generated from the j-th gene is inversely proportional to the number of short-reads aligned to the j-th gene. Here, we will give two examples based on these two assumptions to discuss their differences.

First, we assume that the real value of matrix A^r is

$$\begin{array}{c} \\ 285 \\ 135 \\ 420 \\ 445 \\ 335 \end{array} \begin{array}{c} \begin{array}{cccc} 350 & 330 & 540 & 400 \end{array} \\ \begin{pmatrix} 175 & 110 & 0 & 0 \\ 0 & 0 & 135 & 0 \\ 175 & 110 & 135 & 0 \\ 0 & 110 & 135 & 200 \\ 0 & 0 & 135 & 200 \end{pmatrix} \end{array},$$

which exactly satisfies the presumption of NCLSE. With the given short-read counts $S^r = (285, 135, 420, 445, 335)$, the estimation results are in Table 1. These results show that, in the ideal condition, the NCLSE provides the true value of the gene expression, while the average deviation of estimators, got by the Bayesian method, is 39.27.

Table 1 Estimation results under NCLSE's presumption

	G^r_1	G^r_2	G^r_3	G^r_4	average deviation
assumed gene expression levels	350	330	540	400	
estimation by NCLSE	350	330	540	400	0
estimation by Bayesian method	369.41	372.82	461.46	416.30	39.27

Second, we assume that the real value of matrix A^r is

$$\begin{array}{c} \\ 240 \\ 300 \\ 400 \\ 400 \\ 280 \end{array} \begin{array}{c} \begin{array}{cccc} 350 & 330 & 540 & 400 \end{array} \\ \begin{pmatrix} 150 & 90 & 0 & 0 \\ 0 & 0 & 300 & 0 \\ 200 & 120 & 80 & 0 \\ 0 & 120 & 80 & 200 \\ 0 & 0 & 80 & 200 \end{pmatrix} \end{array},$$

which exactly satisfies the presumption of Bayesian estimation. With the given short-read counts $S^r = (240, 300, 400, 400, 280)$, the estimation results by the two models are in Table 2. These results show that the average deviation of estimators got by the Bayesian method is extremely smaller than the one got by the NCLSE.

Table 2 Estimation results under Bayesian method's presumption

	G_1^r	G_2^r	G_3^r	G_4^r	average deviation
assumed gene expression levels	350	330	540	400	
estimation by NCLSE	146.67	360.00	1013.33	53.33	263.35
estimation by Bayesian method	329.16	328.65	576.525	385.675	18.26

The results in Tables 1 and 2 reflect that Bayesian method is less dependent on the presumption and more robust than the Least-square method. Especially, the Beysian method works on all subgroups, while the Least-square method only works on about 90% of the subgroups and 50% of ncRNA for our dataset.

5.2 Comparison of ncRNA expression profiles produced by different methods

For the 2014 subgroups of the real dataset, 200 parallel independent runs of the Gibbs sampler generate i.i.d. sample of ncRNA expression levels G^r. The posterior moments of $\{G_j^r\}_{j=1,\ldots,m_r}$ are the Bayesian estimators of the ncRNA expression levels. At the same time, NCSLE is used with the same dataset to get the ncRNA expression profile. Then the estimation results of these two methods are available in Additional files 4 and 5, respectively.

Table 3 shows the top 20 ncRNA expression levels by NCSLE and Bayesian methods. The first group of data is the top 20 ncRNA expression levels from the nonnegative constrained least squares estimation, and their corresponding expression levels from Bayesian method and traditional single-mapping method are following. The expression levels' differences between NCLSE and Bayesian method are small, and they get the same order of the expression levels. At the same time, the traditional single-mapping method gets the same results as the NCLSE in the single-mapping case. The second group is the top 20 ncRNA expression levels generated from the Bayesian method, but only 13 of them have NCLSE and 9 of them have traditional single-mapping estimation. Furthermore, these 13 ncRNAs are the top 14 ncRNAs in the Bayesian group, while the other 7 ncRNAs do not have the least squares estimators since the equations, which their are in, do not have the inverse matrices $(\widetilde{T}^r\widetilde{T}^{r\mathrm{T}})^{-1}$. Thus, it shows that the Bayesian method can solve the sparse matrix problem in the NCLSE and woke on all subgroups.

In our real example, the ncRNA FR185008 has the highest expression level in Bayesian method. The subgroup this ncRNA in has 12 ncRNAs and 130 short-reads, and this subgroup cannot be solved by the NCLSE and the traditional single-mapping approaches. Furthermore, we can estimate the distribution of ncRNA expression levels by the i.i.d. samples which we have got by Bayesian method. Therefore, for the above subgroup, which can solved only by Bayesian method, the histograms of ncRNA expression levels are displayed in Figure 3.

Table 3 Top 20 ncRNA expression levels (ppb) of Least-square and Bayesian method

top 20 of NCLSE				top 20 of Bayesian estimation			
ncRNA ID	NCLSE	Bayesian	single-mapping	ncRNA ID	Bayesian	NCLSE	single-mapping
FR185665	6044.35	5888.53	—	FR185008	7955.55	—	—
FR181176	4252.66	4143.66	—	FR185665	5888.53	6044.35	—
FR233520	3277.55	3193.96	3277.55	FR181176	4143.66	4252.66	—
FR233568	3072.50	2994.11	3072.50	FR233520	3193.96	3277.55	3277.55
FR211846	2216.16	1969.82	—	FR233568	2994.11	3072.50	3072.50
FR068264	2072.44	2019.57	2072.44	FR308205	2892.03	—	—
FR380873	1925.57	1876.49	1925.57	FR068264	2019.57	2072.44	2072.44
FR353964	1374.21	1339.17	1374.21	FR328519	2002.05	—	—
FR035930	864.56	842.52	864.56	FR211846	1969.82	2216.16	—
FR028853	837.71	804.04	—	FR380873	1876.49	1925.57	1925.57
FR231770	793.18	772.94	793.18	FR062134	1843.53	—	—
FR181979	762.49	743.05	762.49	FR353964	1339.17	1374.21	1374.21
FR332029	743.22	429.89	—	FR035930	842.52	864.56	864.56
FR343785	688.15	670.60	688.15	FR025554	818.28	—	—
FR273148	658.12	641.34	658.12	FR028853	804.04	837.71	—
FR308195	654.24	637.56	654.24	FR338471	788.10	—	—
FR395528	583.52	566.91	—	FR231770	772.94	793.18	793.18
FR173131	517.49	504.28	517.49	FR020984	746.86	—	—
FR186572	451.90	440.37	451.90	FR181979	743.05	762.49	762.49
FR208556	449.31	437.85	449.31	FR343785	670.60	688.15	688.15

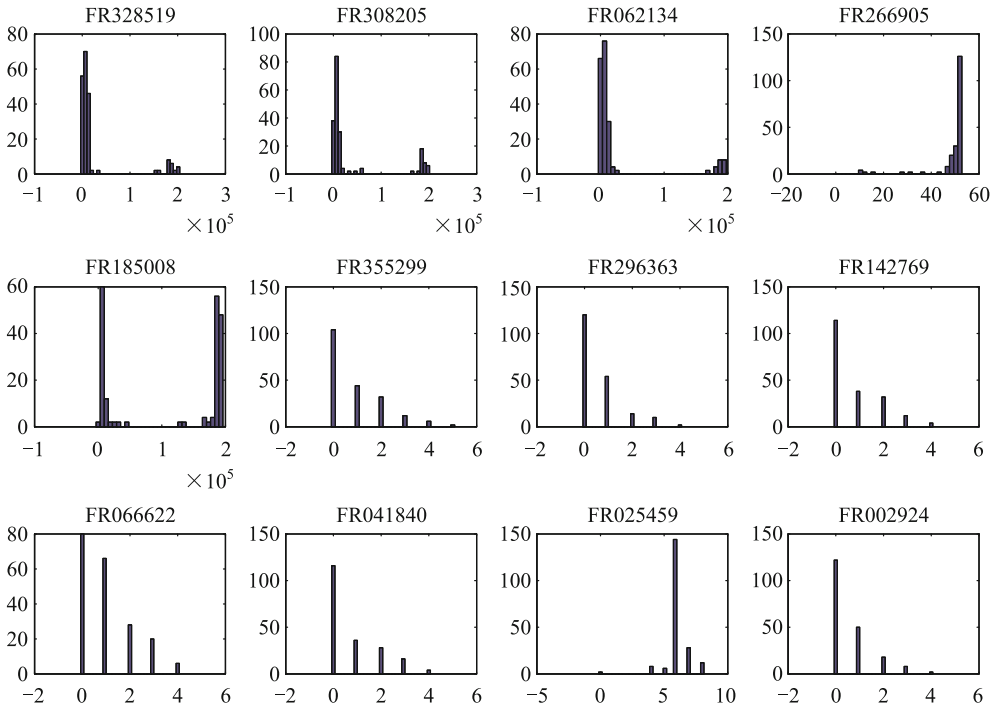


Fig. 3 Histograms of ncRNA expression levels

The largest subgroup in our data has 1268 ncRNAs and 6375 short-reads. Though $n_r > m_r$ in this subgroup, the Least-square method still cannot work since $(\widetilde{T}^r \widetilde{T}^{r\mathrm{T}})^{-1}$ does not exist. This is also the reason why the Least-square method only estimates about 50% of ncRNA expression levels in our dataset.

According to the above analysis, we found that NCSLE can cover the traditional single-mapping approache and get the same estimation for single-mapping data, while NCSLE and Bayesian method have similar results for the ncRNAs which can be estimated by the NCLSE. Furthermore, the Bayesian method can estimate the ncRNA expression levels for all subgroups, and it depends less on the model assumption. At the same time, the Bayesian method can estimate the ncRNA expression levels distributions and has more statistical inference.

6 Conclusion

In this paper, we constructed two methods, the Least-Square and Bayesian Method, to solve the multiple mapping problem and extract gene expression profiles through the next-generation sequencing. The short-read sequences, coming from the NGS, were paralleled to genome through Blast. To improve the efficiency of these methods, we described a partition scheme, which was followed by the implementation of these methods in each subgroup. The NCSLE can cover the traditional approach in single-mapping case, and get more gene expression levels with multiple mapping data. We compared the robustness of NCSLE and Bayesian methods by simulations, which revealed that the Bayesian method is less dependent on the assumptions. Then the above methods were used to solve a real example, which showed that the Bayesian method, using the Gibbs sampler, worked in all subgroups, while the nonnegative constrained least squares estimation only solved 90% of the subgroups and 50% of the genes expression levels.

The essential feature of these methods is that they can solve the multiple mapping problem between genes and short-read sequences, which cannot be solved by traditional single-mapping approaches. To satisfy the precondition that the sum of gene expression levels' estimations must equal the sum of short-reads' counts, some translation is made for the design matrix, then nonnegative constrained least squares is used to get the positive estimation of gene expression levels. Furthermore, the data augmentation algorithm under Bayesian framework is more proper and powerful than the NCLSE method. Since the posterior distribution of the augmentation data A^r is multinomial, the samples of gene expression levels from the Gibbs sampler are integrate positive, which is consistent with the facts. At the same time, this Bayesian method can estimate the gene expression levels' distributions and has more statistical inference.

Acknowledgements This work was supported by the National Key Basic Research and Development Program (973) (Grant No. 2010CB912801), the Ph. D. Programs Foundation of Ministry of Education of China (No. 20090001110005), and the National Natural Science Foundation of China (Grant No. 10731010).

Appendix

A1 Proof of relationship between NCLSE and traditional approaches

The traditional approaches based on the single-mapping data requires each short-read is single mapped to one gene. Suppose that this gene matches to k short-read $S_1, S_2, \ldots, S_k$. Then the NCLSE equation is

$$\widetilde{T}G = \Big[\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\Big]G = [S_1, S_2, \ldots, S_n] = S,$$

where G is the expressive level of the gene, and S_i is the copy number of i-th short-read. Due to the formulation of NCLSE method,

$$G = \mathrm{inv}(\widetilde{T}'\widetilde{T})\widetilde{T}'S = S_1 + S_2 + \cdots + S_n,$$

i.e., the NCLSE of the expression level G is the sum of single-mapping short-read copy numbers, which is also the solution of the traditional approaches.

A2 Proof of posterior

Let

$$\begin{aligned}\{x_k,\, k = 1, \ldots, n\} &\sim Pr(x_k \mid p_k, N,\, k = 1, 2, \ldots, n)\\ &= \mathrm{Multinomial}(p_k, N,\, k = 1, 2, \ldots, n)\\ &= N! \prod_{k=1}^{n} \frac{p_k^{x_k}}{x_k!},\end{aligned}$$

where $\sum_{k=1}^{n} p_k = 1$ and $\sum_{k=1}^{n} x_k = N$.

The prior distributions of the parameters are

$$\begin{aligned}\{p_k,\, k = 1, \ldots, n\} &\sim f(p_k \mid \alpha_k,\, k = 1, 2, \ldots, n)\\ &= \mathrm{Dirichlet}(\alpha_k,\, k = 1, 2, \ldots, n)\\ &= \Gamma\Big(\sum_{k=1}^{n} \alpha_k\Big) \prod_{k=1}^{n} \frac{p_k^{\alpha_k - 1}}{\Gamma(\alpha_k)},\end{aligned}$$

where $\alpha_k,\ k = 1, 2, \ldots, n$, is know.

Then the join density function is

$$\begin{aligned}f(x_k, p_k, N,\, k = 1, 2, \ldots, n) &= Pr(x_k \mid p_k, N,\, k = 1, 2, \ldots, n)\\ &\quad \cdot f(p_k \mid \alpha_k, k = 1, 2, \ldots, n)\\ &= N!\,\Gamma\Big(\sum_{k=1}^{n} \alpha_k\Big) \prod_{k=1}^{n} \frac{p_k^{x_k + \alpha_k - 1}}{x_k!\,\Gamma(\alpha_k)},\end{aligned}$$

and the conditional distributions are

$$\begin{aligned} f(p_k \mid x_k, N,\, k=1,2,\ldots,n) &= \frac{f(x_k, p_k, N,\, k=1,2,\ldots,n)}{\int_{\sum_{k=1}^{n} p_k=1} f(x_k, p_k, N,\, k=1,2,\ldots,n)\mathrm{d}p_k} \\ &\propto \prod_{k=1}^{n} p_k^{x_k+\alpha_k-1} \\ &= \text{Dirichlet}(\alpha_k^* = \alpha_k + x_k,\, k=1,2,\ldots,n), \end{aligned}$$

$$\begin{aligned} Pr(x_k \mid p_k, N,\, k=1,2,\ldots,n) &= \frac{f(x_k, p_k, N,\, k=1,2,\ldots,n)}{\sum_{\sum_{k=1}^{n} x_k=N} f(x_k, p_k, N,\, k=1,2,\ldots,n)} \\ &\propto \prod_{k=1}^{n} p_k^{\alpha_k} \\ &= \text{Multinomial}(p_k, N,\, k=1,2,\ldots,n). \end{aligned}$$

Therefore, using this proof, we get the posterior distributions of parameters are

$$\begin{aligned} &q_{i1}^r, q_{i2}^r, \ldots, q_{im_r}^r \mid a_{ij}^r, \beta_{ij}^r,\, j=1,\ldots,m_r \\ &\quad\sim \text{Dirichlet}(\beta_{i1}^r + a_{i1}^r, \beta_{i2}^r + a_{i2}^r, \ldots, \beta_{im_r}^r + a_{im_r}^r), \quad i=1,\ldots,n_r, \end{aligned}$$

and

$$\begin{aligned} &a_{i1}^r, a_{i2}^r, \ldots, a_{im_r}^r \mid q_{ij}^r, \beta_{ij}^r,\, j=1,\ldots,m_r \\ &\quad\sim \text{Multinomial}(q_{i1}^r, q_{i2}^r, \ldots, q_{i,m_r-1}^r, S_i^r). \end{aligned}$$

A3 Proof of convergence theory

Since $\{{}^kA^r\}_{k=1,\ldots,K}$ is an i.i.d. sample from the A^r's posterior distribution and rows of A^r are also independent on each other, $\{{}^kG_j^r\}_{k=1,\ldots,K}$ is an i.i.d. sample with finite mean and variance. By the Kolmogorov strong law of large number, we have

$$\hat{G}_j^r \overset{\text{a.s.}}{\to} E\left(\sum_{i=1}^{n_r} a_{ij}^r\right) = \sum_{i=1}^{n_r} S_i^r q_{ij}^r, \quad K\to\infty,\; j=1,\ldots,m_r,$$

and

$$\begin{aligned} \sum_{j=1}^{m_r} \hat{G}_j^r &= \sum_{j=1}^{m_r} \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_r} {}^k a_{ij}^r \\ &= \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_r} \sum_{j=1}^{m_r} {}^k a_{ij}^r \\ &= \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_r} S_i^r \\ &= \sum_{i=1}^{n_r} S_i^r. \end{aligned}$$

Additional files

Additional file 1:
Our original data from second generation sequencing ($a05_04_blast.txt$);

Additional file 2:
The aligned result of genes with short-reads ($a5_04h_copy.txt$);

Additional file 3:
Partition Results (g1group1.txtg1group2.txtg1group3.txt);

Additional file 4:
The gene expression levels getting from least squares estimation (lsereresult.txt);

Additional file 5:
The gene expression levels getting from Bayesian method (gibbsresult.txt);

Additional file 6:
Program direction Bayesian method (ProgramDirection.txt).

These additional files can be found at
ftp://162.105.69.120/teachers/fangxz/public/MathFrontData/

References

1. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist, 2001, 29(4): 1165–1188
2. Cloonan N, Forrest A R R, Kolle G, Gardiner B B A, Faulkner G J, Brown M K, Taylor D F, Steptoe A L, Wani S, Bethel G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods, 2008, 5(7): 613–619
3. Faulkner G J, Forrest A R R, Chalk A M, Schroder K, Hayashizaki Y, Carninci P, Hume D A, Grimmond S M. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics, 2008, 91(3): 281–288
4. Metzker M L. Sequencing technologies—the next generation. Nat Rev Genet, 2009, 11(1): 31–46
5. Morin R D, O'Connor M D, Griffith M, Kuchenbauer F, Delaney A, Prabhu A L, Zhao Y, McDonald H, Zeng T, Hirst M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research, 2008, 18(4): 610
6. Mortazavi A, Williams B A, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 2008, 5(7): 621–628
7. Ozsolak F, Milos P M. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet, 2011, 12(2): 87–98
8. Tanner M A. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Berlin: Springer-Verlag, 1996
9. Wang W C, Lin F M, Chang W C, Lin K Y, Huang H D, Lin N S. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinformatics, 2009, 10(1): 328
10. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10(1): 57–63