

1. Abstract

The biggest challenge in applying RNA-Seq was obtaining the required amount of input total RNA of at least 1 μ g. To overcome this obstacle, BGI has optimized the procedures to enable RNA-Seq (Quantification) with low input RNA amount (100 ng), which would not compromise high-quality results. To test the performance of this new methodology, we evaluated the quality and accuracy of the data, by conducting a comparative RNA-Seq analysis using Illumina HiSeq™ 2000 platform for the MAQC UHRR* and HBRR* samples with 1 μ g and 100 ng input RNA amounts. Results demonstrated high technical reproducibility, high correlation between gene expression analysis from 100 ng and 1 μ g sample input, and high concordance between the RNA-Seq data and data generated by quantitative PCR (qPCR). This breakthrough enables the use of low input samples in many relevant research fields, such as stem cell studies, paleoarcheology, evolutionary biology, forensics and cancer.

Note: *UHRR: Universal Human Reference RNA
*HBRR: Human Brain Reference RNA

2. Introduction

RNA-Seq (Quantification) is a powerful Next Generation Sequencing (NGS) based strategy for transcript quantification that brings several advantages over conventional methods. Compared with microarray technology, RNA sequencing offers comprehensive assessment of RNA expression profiles, high-throughput data, low background noise, high sensitivity, high repeatability and larger dynamic range [1-3]. The optimized method of RNA-Seq (Quantification) in BGI extends the power and flexibility of this technology to very small amounts (100 ng) of input total RNA.

3. Methods

3.1 Samples Preparation, Library Construction and Sequencing

Standard RNA samples in MAQC test [4] UHRR and HBRR were used for RNA-Seq by Illumina HiSeq™ 2000 sequencer (50-base single-end runs).

Samples were prepared following Illumina “mRNA Sequencing Sample Preparation Guide” with 100 ng and 1 μ g of total RNA. RNA libraries were constructed by two operators (batch 1 and batch 2). Libraries were pooled in different lanes for sequencing. The detailed experimental design is illustrated in Table 1.

Table 1 Detailed experimental design of 100 ng input RNA library construction

| Batch of library construction | Lane of sequencing | Sample | | |
|-------------------------------|--------------------|-----------|-----------|-----------|
| Batch 1 | Lane 1 | UHRR (A1) | UHRR (A2) | HBRR (B1) |
| | Lane 2 | UHRR (A3) | HBRR (B2) | HBRR (B3) |
| Batch 2 | Lane 3 | UHRR (A4) | UHRR (A5) | HBRR (B4) |
| | Lane 4 | UHRR (A6) | HBRR (B5) | HBRR (B6) |

3.2 Bioinformatics Analysis

10 M sequencing reads were generated for each duplicate and mapped to RefSeq (version updated June 16, 2011). RPKM method was employed to normalize the gene expression, and then the gene expression values were used to evaluate the technical reproducibility of the RNA-Seq (Quantification) and the correlation of 100 ng vs. 1 μ g input total RNA. Furthermore, the result of gene expression was validated by qPCR (Gene Expression Omnibus).

4. Results

4.1 Statistics of Detected Genes

After mapping of all reads to the reference RefSeq, for duplicates A1 to A6 and B1 to B6, an average of 80.22% and 78.86% of the total reference could be detected by RNA-Seq (Quantification) with 100 ng total RNA of UHRR and HBRR samples, respectively (Table 2). These results are highly similar to that of RNA-Seq with 1 μ g total RNA input (A and B).

Table 2 Statistics of detected genes (Reads > 0, mismatches \leq 2) for UHRR and HBRR

| Sample ID | UHRR | | | Sample ID | HBRR | | |
|-----------|-------|----------|------------|-----------|-------|----------|------------|
| | Total | Detected | % Detected | | Total | Detected | % Detected |
| A1 | 22748 | 18134 | 79.72% | B1 | 22748 | 17948 | 78.90% |
| A2 | 22748 | 18228 | 80.13% | B2 | 22748 | 17953 | 78.92% |
| A3 | 22748 | 18297 | 80.43% | B3 | 22748 | 17901 | 78.69% |
| A4 | 22748 | 18280 | 80.36% | B4 | 22748 | 17956 | 78.93% |
| A5 | 22748 | 18270 | 80.31% | B5 | 22748 | 17960 | 78.95% |
| A6 | 22748 | 18276 | 80.34% | B6 | 22748 | 17918 | 78.77% |
| A | 22748 | 19331 | 80.58% | B | 22748 | 17976 | 78.90% |

4.2 Technical Reproducibility Analysis

As illustrated in MA plot [5] (Fig.1) below, the values of median and IQR indicate high reproducibility between technical replicates, different samples in different lanes and batches.

We further used spearman’s correlation R to evaluate the correlation of technical replicates. Log transformed values are used to calculate differential expression ratios. Fig.2 below shows RNA-Seq (Quantification) technical replicates of 100 ng total RNA with high correlation R values ranging from 0.9294 to 0.9656.

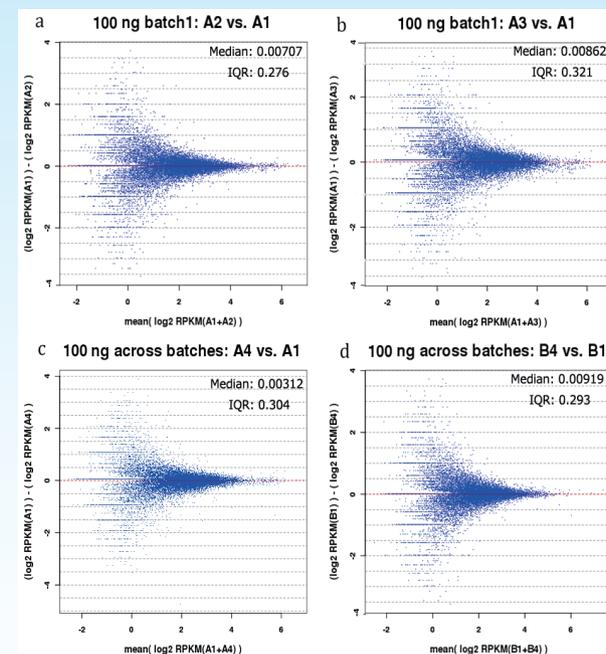


Figure 1 Correlation of gene expression between different lanes and batches
a: correlation of UHRR gene expression in the same lane and same batch;
b: correlation of UHRR gene expression in different lanes but the same batch;
c: correlation of UHRR gene expression in different lanes and different batches;
d: correlation of HBRR gene expression in different lanes and different batches.

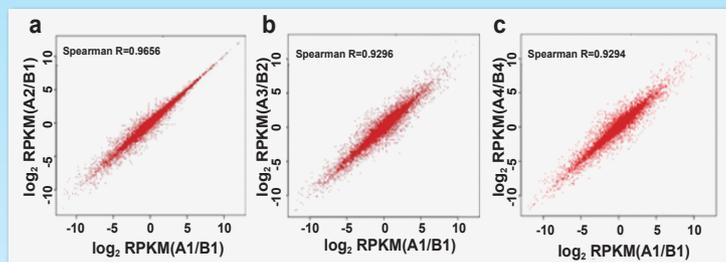


Figure 2 Comparison analysis of 100 ng RNA-Seq between different lanes and batches
a: correlation of gene expression for 100 ng RNA-Seq in the same lane and same batch;
b: correlation of gene expression for 100 ng RNA-Seq in different lanes but the same batch;
c: correlation of gene expression for 100 ng RNA-Seq in different lanes and batches.

4.3 Correlation between RNA-Seq (Quantification) with 100 ng and 1 μ g Input

Correlation between RNA-Seq (Quantification) with 100 ng and 1 μ g input was evaluated by spearman correlation (R). As shown in Fig.3, correlation values, ranging from 0.9114 to 0.9188, indicates good correlation between 100 ng and 1 μ g inputs of total RNA.

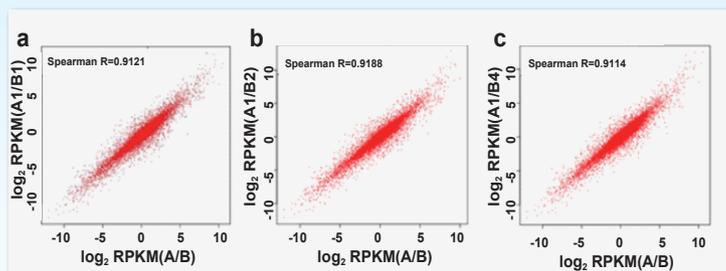


Figure 3 Correlation of gene expression between 100 ng and 1 μ g input
a: correlation of gene expression of 100 ng RNA-Seq in the same lane and same batch vs. 1 μ g;
b: correlation of gene expression of 100 ng RNA-Seq in different lanes but the same batch vs. 1 μ g;
c: correlation of gene expression of 100 ng RNA-Seq in different lanes and batches vs. 1 μ g.

4.4 Correlation of Gene Expression between RNA-Seq and qPCR

The gold standard qPCR data was compared to calculate the accuracy of gene expression. R values, ranged from 0.9184 to 0.9210 for 100 ng and 0.9170 for 1 μ g indicate that RNA-Seq with 100 ng input has the same high accuracy as RNA-Seq with 1 μ g.

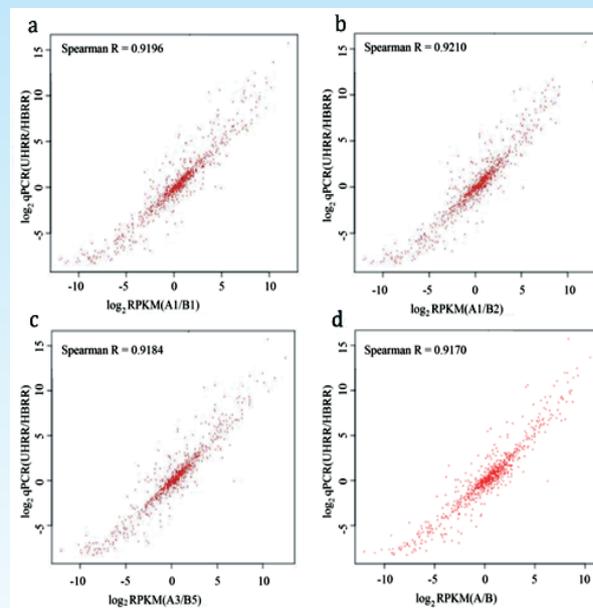


Figure 4 Correlation of gene expression between RNA-Seq and qPCR
a: correlation of gene expression of 100 ng RNA-Seq in the same lane and same batch vs. qPCR;
b: correlation of gene expression of 100 ng RNA-Seq in different lanes but the same batch vs. qPCR;
c: correlation of gene expression of 100 ng RNA-Seq in different lanes and batches vs. qPCR;
d: correlation of gene expression of 1 μ g RNA-Seq vs. qPCR.

5. Conclusions

- ▶ The input amount of 100 ng total RNA using BGI's developed method shows similar capacity for detecting genes as previously used input of 1 μ g total RNA.
- ▶ Gene expression results on technical replicates originated from 100 ng total RNA shows high reproducibility, with median value around zero in the MA plots and spearman correlation R values over 0.9.
- ▶ Good correlation between 100 ng and 1 μ g of total RNA indicates small amount of input total RNA can be adopted for RNA-Seq (Quantification) without affecting performance.
- ▶ Strong concordance with gene expression measured by gold standard qPCR demonstrates high accuracy of results.

6. References

- [1] Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2010, 11(1): 31-46.
- [2] Marioni JC, Mason CE, Mane SM, *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008, 18(9): 1509-1517.
- [3] Carninci P. Is sequencing enlightenment ending the dark age of the transcriptome? *Nature Methods.* 2009, 6(6): 711-713.
- [4] Shi L, Reid LH, Jones WD, *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol.* 2006, 24(9): 1151-1161.
- [5] Yang YH, Dudoit S, Luu P, *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002, 30(4): e15.

Contact Us

China (Mainland)
BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen, 518083, China
Tel: 400-706-6615 +86-755-25281960
Email: bgi-PGx@genomics.cn
www.genomics.cn

Asia Pacific (Hong Kong)
BGI Hong Kong Co. Limited, 16th Dai Fu Street, Tai Po Industrial Estate, Tai Po, Hong Kong
Tel: +852-3610-3510
Email: bgihk.enquiry@genomics.cn
www.bgisequence.com

Europe (Copenhagen)
Ole Maaløes Vej 3, DK-2200 Copenhagen N, Denmark
Tel: +45-7026 0806
Email: bgi-europe@genomics.cn
www.bgisequence.com

North America (Boston)
BGI Americas Corporation, One Broadway, 14th Floor, Cambridge, MA 02142, USA
Tel: +1-617-500-2741
Email: info@bgiamericas.com
www.bgiamericas.com www.bgisequence.com