BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 3. 280–287
Advance Access published on 15 April 2011

doi:10.1093/bib/bbr004

Design and validation issues in RNA-seq experiments

Zhide Fang and Xiangqin Cui

Submitted: 30th August 2010; Received (in revised form): 2nd February 2011

Abstract

The next-generation sequencing technologies are being rapidly applied in biological research. Tens of millions of short sequences generated in a single experiment provide us enormous information on genome composition, genetic variants, gene expression levels and protein binding sites depending on the applications. Various methods are being developed for analyzing the data generated by these technologies. However, the relevant experimental design issues have rarely been discussed. In this review, we use RNA-seq as an example to bring this topic into focus and to discuss experimental design and validation issues pertaining to next-generation sequencing in the quantification of transcripts.

Keywords: *experimental design; next-generation sequencing; RNA-seq; replicates; sample size; blocking*

INTRODUCTION

The next-generation sequencing is a group of new sequencing technologies that are based on randomly amplifying and shotgun sequencing techniques. Short sequences (often around 30 bases) are obtained in extremely high throughput. The total sequences generated by each run can be hundreds of millions of bases due to the extremely high parallel nature of these technologies [1–3]. Since the marketing in 2004, next-generation sequencing technologies have dramatically improved, and their application is growing exponentially [4]. The next-generation sequencing technologies greatly reduce the cost for sequencing new genomes [5] and for resequencing genomes with reference genome sequences for genetic variant discovery [6–8]. They have also been widely used to characterize DNA–protein interaction (ChIP-seq) [9, 10], to survey the transcriptome for expression level and splicing variants [11–13], and to study epigenomics [14–16] with more precise digital count outcomes.

The applications of next-generation sequencing technologies will grow more rapidly as the technologies continue to improve and the cost continues to drop. However, in the rapid growth of the applications, the experimental design issues related to the next-generation sequencing have rarely been discussed until very recently [17]. Here, we will review some statistical experimental design principles and provide some thoughts on the design and validation. We will use RNA-seq as an example to discuss the principles of experimental design and technology-specific issues. We will briefly mention a few other issues specific to other applications of next-generation sequencing at the end.

The next-generation sequencing technologies have been widely applied to measure gene expression levels and composition (termed as RNA-seq) [16, 18–21]. The power of RNA-seq in quantifying and annotating transcriptomes is striking. By obtaining tens of millions of short sequence reads from the transcript population of interest and by mapping these reads to the reference genome, RNA-seq produces digital signals (counts), and thus leads to highly reproducible results with relatively little technical variation [12, 22, 23]. When enough reads are

Corresponding author. Xiangqin Cui, Assistant Professor, Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, 327 Ryals Public Health Building, 1665 University BLVD, Birmingham, AL 35294, USA. Tel: +1-205-996-4154; Fax: +1-205-975-2540; E-mail: xcui@uab.edu

Zhide Fang obtained his PhD in Statistics in 1999. He is currently an Associate Professor in Biostatistics Program, School of Public Health, at Louisiana State University Health Sciences Center at New Orleans.

Xiangqin Cui obtained her PhD in Genetics in 2001 and did her postdoctoral training in Statistical Genetics from 2001 to 2004. She is currently an assistant professor in Department of Biostatistics, Section on Statistical Genetics, at the University of Alabama at Birmingham.

collected from a sample, it has the potential to detect and quantify RNAs from all biologically relevant classes, including those with low and moderate abundance [12, 24].

EXPERIMENTAL DESIGN PRINCIPLES

For any experiment that has variation, there are well-established experimental design principles for achieving validity and efficiency [25, 26]. These principles were originally established from low throughput experiments, but have been widely accepted for microarray experiments. For detailed discussion, please refer to book chapters dedicated to this topic [27, 28]. Although next-generation sequencing technologies have many characteristics different from microarray technologies, most of the experimental design principles still apply. Here, we briefly review the principles and remind people about their application in experiments employing the next-generation sequencing technologies.

Randomization

Randomization dictates that the experimental subjects should be randomly assigned to the treatments or conditions to be studied in order to eliminate unknown factors that potentially affect results [29]. For RNA-seq experiments, besides the randomization in preparing the research subjects, there are many other steps to consider for randomization due to the complexity of the technologies. For example, we can randomize the sample order for various steps in the library construction and the order/ location of the samples in the sequencer.

Replication

Replication is essential for estimating and decreasing the experimental error, and thus to detect the biological (treatment) effect more precisely. A true replication is an independent repetition of the same experimental process and independent acquisition of the observations [26]. As in the expression microarray experiments, there are different levels of replications in RNA-seq experiments. The most desirable replicates are the biological replicates, which are true replicates and provide us the variation among biological samples [30, 31]. In the current RNA-seq publications, some studies include biological replicates [13, 18, 32–35], while many others only have technical replicates that are repeated measurements from the same biological sample [12, 20, 22, 23]. If the goal is to evaluate the technology, technical replicates alone are sufficient. Otherwise, if the goal is to investigate the biological differences between conditions/tissues/treatments, biological replicates are essential. In addition to the intrinsic biological variation in gene expression, the sequence library construction often includes PCR amplification. The PCR amplification artifacts can result in a large number of identical sequences, which can be confounded with gene expression level. If consistent results are obtained from biological replicates, the count of reads from a gene is more likely to reflect the expression level [36]. However, it is worth pointing out that there may be sequence polymorphisms between biological replicates, which can result in a gain or a loss in reads from different biological replicates depending on their sequence consistency with the reference genome sequence [37]. Some sequence reads containing sequence polymorphisms compared with the reference sequences are more likely to be discarded during mapping. Therefore, there could appear to be an extra biological variation in addition to the gene expression variation. It might be less of an issue for inbred model organisms, such as inbred mice, when comparisons are within an inbred strain. However, for human studies, this may not be a negligible issue.

Different levels of replicates often require different statistical analysis methods. For example, when there are only strict technical replicates of different runs from the same library, the variation comes from the random sampling variance of sequencing. This variation can be modeled fairly well using a Poisson distribution [20]. However, when biological replicates are used, a Bayesian hierarchical error model or a negative binomial (NB) model is more appropriate for modeling the variation resulted from both the random sampling of sequencing and the natural variation among biological replicates [38, 39].

RNA-SEQ SPECIFIC EFFECTS AND BLOCKING

As in microarray studies, RNA-seq experiments can be affected by the variability coming from nuisance factors, often called technical effects in the RNA-seq literature. Besides processing date, technician and reagent batch, which are commonly known to investigators, there are some recognized technical effects specific to the RNA-seq procedures. One of these technical effects comes from the generation of libraries of cDNA fragments, which involves various ligations of adaptors and PCR amplifications. Besides the library preparation effect, there are also other technology-specific effects. For example, the commonly used Illumina-sequencing technology can sequence eight samples simultaneously in the eight lanes in one flow cell, of which one lane is often used for the control sample. Thus, there is variation from one flow cell to another resulting in flow cell effect. In addition, there exits variation between the individual lanes within a flow cell due to systematic variation in sequencing cycling and/or base-calling. Among these sources of variation, the library preparation effect is the largest [40]. The flow cell and lane effects are relatively small [20, 41].

From the experimental design point of view, there are some steps that can be taken to properly handle these effects besides the technology improvement. For the library preparation effect, introducing replicates before this step (often biological replicates) provides a way to estimate this effect and to properly handle it in the statistical inference. Blocking design can be used to eliminate the flow cell and lane effects. Blocking is also an experimental design principle. It dictates comparisons within a block, a known uninteresting factor that causes variation, such as flow cell effect. Either the randomized complete block design (RCBD) or the balanced incomplete block design (BIBD) can be used to achieve this goal, depending on the number of treatments/groups to be compared. Sequencing lanes can also serve as blocks when bar-coding during library preparation (for the protocol for Illumina platform, see http://www.illumina.com/Documents/products/datasheets/data sheet_sequencing_multiplex.pdf) is used for multiplexing [17]. However, it has been shown that multiplexing reduces sensitivity and reproducibility in miRNA detection [42]. Therefore, caution needs to be taken when multiplexing is considered for the purpose of reducing flow cell and lane effects.

SEQUENCING DEPTH

One unique characteristic for RNA-seq and some other applications of next-generation sequencing technologies is sequencing depth or coverage, which is often estimated as the number of total mapped sequences. Genes are expressed at different levels in each transcriptome. Due to the random sampling nature of RNA-seq, it will take a large number of sequences to measure the transcripts that are expressed at low level. For a given budget, it is critical to decide whether to increase the sequencing depth to have more accurate measurements on the genes expressed at low level or increase the sample size with limited sequencing depth for each sample. Since gene expressions roughly follow power law in terms of expression level and number of genes [43, 44], Bashir *et al.* [24] has modeled the sequencing depth for RNA-seq to examine the number of reads needed to reach the low expression level. Based on Marioni's data, they showed that more than 90% of the transcripts were sampled with one million sequence reads. For experimental design, they recommend a small pilot sequencing (about 1-million sequences) to estimate the distribution of all transcripts in the population before deciding the actual sequencing depth for the whole experiment. However, as the authors acknowledged, the sequencing process is not an unbiased random sampling process. There are several recognized biases as discussed later in this review and the potential lack of fit of the power law or other distributions to the gene expression in the genome. Therefore, their model might be too optimistic and the required total number of reads could be much higher than estimated to achieve the target coverage. In addition, it is well known that the low count reads show low consistency between technical replicates and the low count transcripts (less than a few reads) are often excluded from analysis [40] for comparisons across conditions. Therefore, more than one runs of the same sample are sometimes used to increase the depth of the sequencing. However, it is still not clear what the lower limit of gene expression level is for being functionally relevant. If the extremely low expression is stochastic instead of biologically meaningful, sequencing may not need to reach an incredible depth. Thus, given the total cost of the experiment, one needs to carefully consider whether to increase the sequencing depth per sample or to increase the biological replicates.

RNA-seq is sometimes used to detect the differential expression between the two alleles of the same gene at the heterozygous locations in the genome when the expressed sequences contain sequence polymorphisms [45–47]. Heap *et al.* [45] ran a simple power analysis based on a Chi-square test at each SNP and found that it will take about 50 reads covering a SNP to detect a 2-fold difference between the two alleles with 19% power at significance level of 0.001. Therefore, they only focused on the SNPs with at least 50 reads for discovering allelic differential expression. It would take extremely deep coverage in order to detect allelic differential expression for genes expressed at a fairly low level.

PAIRED-END SEQUENCING

The development of the paired-end sequencing technique brought more improvement in the next-generation sequencing [48–50]. In RNA-seq experiments, the sequencing of both ends of RNA fragments adds more information especially in the detection of alternative splicing and chimeric transcripts [51, 52]. At the same sequencing depth, the pair-end sequences increase the sensitivity and specificity of the detection of the alternative splicing and chimeras in comparison with the single end sequencing. Therefore, paired-end sequencing is a more efficient strategy for characterizing and quantifying transcriptome.

BIASES OF NEXT-GENERATION SEQUENCING

The shotgun short sequences in RNA-seq experiments are expected to be randomly obtained from the transcripts. Therefore, the number of reads from a transcript depends on the transcript length [12]. This nature makes the comparison across samples more efficient for longer transcripts than shorter ones [40, 53]. In addition, sequence reads are not exactly randomly obtained from transcripts in reality. Biases have been found to be related to GC content of the sequence [54], the use of the random hexamer primers [55], $3'$ and $5'$ depletion or bias towards $3'$-end [36], and bias toward specific RNA species [56]. Most of these biases are related to library preparation methods. They directly affect the transcription level comparisons across genes, gene end characterization, and alternative splicing characterization. Some attempts for removing these sequence/spatial biases have been explored both from the analysis level [55] and the protocol improvement [57]. From the experimental design point of view, these biases increase the required samples size and sequence depth, which emphasize the importance of choosing better protocols and selecting the right analysis methods.

SAMPLE SIZE CALCULATION FOR RNA-SEQ

The complexity of RNA-seq experiments makes it difficult to determine sample sizes. To our knowledge, there has not yet been any literature published on this topic. The sample size may be determined at two levels—the number of lanes for technical replicates in one treatment or the number of biological replicates for each treatment.

In the cases when there are only technical replicates and the library preparation effects and lane effects are negligible or mitigated by proper designs, sample sizes can be calculated gene-by-gene based on Poisson models. Here we provide some details on how this can be done. For the RNA-seq data (counts) from two tissues/treatments, X_{ijk}, where i (=1,2) is the index for the tissues/treatments and $j(=1, 2, \ldots, n_i)$ is the index for the replicates (lanes) within each treatment for the kth gene, we assume Poisson models as $X_{ijk} \sim$ Poisson($L_{ij}\gamma_{ik}$), with L_{ij} being the total number of mapped reads in this replicate and γ_{ik} being the transcript frequency for this gene. Let $L_i (i = 1, 2)$ be the sum, over j, of L_{ij}. Then, to obtain a power $(1 - \beta)$ under the alternative hypothesis $H_a : \gamma_{1k}/\gamma_{2k} = \rho > 1$ at the level of significance α, by using a Wald-type Z-statistic, we can have the relation

$$\lambda_1 = L_1\gamma_{1k} = \frac{(\rho/d + 1)\ (Z_{1-\alpha} + Z_{1-\beta})^2}{(\rho - 1)^2}, \tag{1}$$

where $d = L_1/L_2$, and $Z_{1-\alpha}, Z_{1-\beta}$ are normal quantiles [58]. The values for α, β and ρ are pre-selected. The values of d, $\bar{L}_i$ (the average of mapped reads from each replicate i), and γ_{1k} can be determined by preliminary data. We will have approximate sample sizes [59]

$$n_1 = \frac{\lambda_1}{\bar{L}_1\gamma_{1k}}, \quad n_2 = \frac{d\bar{L}_1 n_1}{\bar{L}_2}. \tag{2}$$

If the likelihood ratio test is employed for testing the hypothesis, the sample sizes can also be determined similarly, but the calculation is more complicated. See Gu *et al.* [60] for the formula. However, it needs to be pointed out that the sample size calculated based on the Poisson model above is the most optimistic scenario given that only the Poisson random variation is considered in the calculation.

When there are biological replicates and the over-dispersion problem exists, NB distributions are more appropriate than Poisson distributions to model the RNA-seq data [38, 39]. If we denote $NB(\mu, \tau)$ as a negative binomial distribution with mean μ and variance $(\mu + \tau\mu^2)$, then we can assume that $X_{1jk} \sim \text{NB}(\mu, \tau)$, and $X_{2jk} \sim \text{NB}(\delta\mu, \tau)$, where δ is the tissue/condition ratio difference. Since there is no close form for the maximum likelihood estimate (MLE) of τ [61], there is no analytical relation between power and sample size. As these authors suggest, we can perform Monte Carlo simulations to obtain the simulated power for a fixed sample size n (assuming $n_1 = n_2 = n$). Specifically, for an effect size ($\delta \neq 1$), a numerical relation between power and the sample size (n) can be built with the following steps:

(1) Obtain two independent random samples, $\{x_{1jk}\}_{j=1}^{n}$ and $\{x_{2jk}\}_{j=1}^{n}$, from $\text{NB}(\hat{\mu}, \hat{\tau})$ and $\text{NB}(\delta\hat{\mu}, \hat{\tau})$ separately, where $\hat{\mu}, \hat{\tau}$ are the corresponding MLEs from the preliminary data;
(2) Repeat the process in (1) w times;
(3) Calculate the simulated power as the percentage of times that the null hypothesis is rejected by likelihood ratio test or Wald-type Z-test.

The sample size formulas discussed above are for a single gene. They cannot be directly applied to the RNA-seq experiments since there are thousands of genes in one RNA-seq experiment. However, we can handle this as in microarray experiments—first obtain the sample sizes for one gene and then determine the overall sample size based on the overall average power. Another way to compute the power and sample size could be based on the setting of effect size, number of non-differential genes, and the expected number of false positives as that for the microarray data [62] although error models would need to be adjusted. A third potential way is based on modeling the P-values as a mixture of distribution from genes that are not differentially expressed and genes that are differentially expressed as the method behind the *PowerAtlas* (http://www.poweratlas.org/) software [63].

VALIDATION

Validation has been an important part in expression microarray literature. The differentially expressed genes (at least some) identified using microarray are often validated using quantitative RT-PCR (qRT-PCR). Although validation is required by a lot of journals for publication, it is still debatable whether qRT-PCR validation of differentially expressed genes is still necessary after the huge number of publications with qRT-PCR validation of the microarray technologies and how to do it exactly if the answer is yes [64, 65]. In RNA-seq studies, RT-PCR has also been used for validation in some studies. For example, Camarena *et al.* [66] validated a handful of differentially expressed genes identified using RNA-seq in pathogen *Acinetobacter baumannii*. They showed that the fold changes estimated from RNA-seq had high correlation with that from the qRT-PCR. High consistency between RNA-seq and qRT-PCR results have also been observed in other studies [13, 32, 67]. A detailed RT-PCR analysis by Ramskold *et al.* [68] showed that UTRs especially the 3′-UTR are actually quite variable. Excluding the UTRs from the RNA-seq data improves the consistency of RNA-seq and RT-PCR results significantly. This finding is consistent with the observations that high consistency exists between microarray and qRT-PCR results for genes with microarray probes and PCR primers interrogating exactly the same transcript while the questionable genes show lower consistency [69]. In addition, studies on allelic expression and alternative splicing also validated their RNA-seq findings using RT-PCR [46, 70]. It is worth pointing out that validation using qRT-PCR on the same RNA samples assayed in the RNA-seq analysis only validates the technology. It does not validate the conclusion about the treatments/conditions. It is the validation using different biological replicates from the same populations that can further validate the biological conclusions from RNA-seq experiments [65].

SOME DESIGN ISSUES RELATED TO OTHER APPLICATIONS OF THE NEXT-GENERATION SEQUENCING

Next-generation sequencing has been used to identify genetic variants in a population. For example, the Thousand Genomes project (http://www.1000genomes.org/page.phpis) is sequencing at least 1000 individuals to identify all the genetics variants in humans. To most efficiently identify the genetic structure variants (such as CNV, Indels) using the next-generation sequencing technologies, one factor to consider is the insert length of the library. Bashir *et al.* [24] pointed out that having two libraries with different insert lengths is beneficial (for example, 200 bp and 20 kb libraries) for mapping the break points of genetic-structure variants.

ChIP-seq is another common application of the next-generation sequencing technologies. The DNA fragments enriched in chromatin immunoprecipitation (ChIP) are compared with the total DNA to identify binding sites in the genome for a given protein [10, 12, 24, 71, 72]. For some DNA binding factors, the total ChIP-seq sequence needs to be more than what is obtained from one sequencing lane. In this case, whether increasing the sequencing depth by sequencing the same sample in multiple lanes (technical replicates) or sequencing different biological samples in different lanes (biological replicates) needs to be considered. Tuteja *et al.* [73] found that using biological replicates can increase the number of peaks identified and increase the enrichment for consensus sequences for the binding factor. One explanation is that results from technical replicates tend to be affected more by the PCR artifacts than the independent biological replicates.

Key Points

- Biological replicates are important in most RNA-seq experiments.
- Sequencing depth and sample size are interrelated. They are often limited by experiment budget.
- Sequencing biases potentially increase the required sample size and sequencing depth.
- Validation using biological replicates is more meaningful.

References

1. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol* 2009;**25**:195–203.
2. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods* 2008;**5**:16–18.
3. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**11**:31–46.
4. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet* 2010;**11**:476–86.
5. Huber JA, Mark Welch DB, Morrison HG, *et al.* Microbial population structures in the deep marine biosphere. *Science* 2007;**318**:97–100.
6. Wheeler DA, Srinivasan M, Egholm M, *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;**452**:872–6.
7. Korbel JO, Urban AE, Affourtit JP, *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**:420–6.
8. Alkan C, Kidd JM, Marques-Bonet T, *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 2009;**41**:1061–7.
9. Johnson DS, Mortazavi A, Myers RM, *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007; **316**:1497–502.
10. Visel A, Blow MJ, Li Z, *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;**457**: 854–8.
11. Pan Q, Shai O, Lee LJ, *et al.* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**:1413–5.
12. Mortazavi A, Williams BA, McCue K, *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
13. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 2008;**322**:1845–8.
14. Park PJ. Epigenetics meets next-generation sequencing. *Epigenetics* 2008;**3**:318–21.
15. Brunner AL, Johnson DS, Kim SW, *et al.* Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res* 2009;**19**:1044–56.
16. Lister R, O'Malley RC, Tonti-Filippini J, *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;**133**:523–36.
17. Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. *Genetics* 2010;**185**:405–16.
18. Cloonan N, Forrest AR, Kolle G, *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 2008;**5**:613–9.
19. Pickrell JK, Marioni JC, Pai AA, *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**:768–72.
20. 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**:1509–17.
21. Sultan M, Schulz MH, Richard H, *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008;**321**:956–60.
22. Bainbridge MN, Warren RL, Hirst M, *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 2006;**7**: 246.
23. Hashimoto S, Qu W, Ahsan B, *et al.* High-resolution analysis of the 5′-end transcriptome using a next generation DNA sequencer. *PLoS One* 2009;**4**:e4108.
24. Bashir A, Bansal V, Bafna V. Designing deep sequencing experiments: structural variation, haplotype assembly, and transcript abundance. *BMC Genomics* 2010;**11**:385.
25. Fisher RA. The arrangement of field experiments. *J Minist Agric Great Britain* 1926;**33**:503–13.
26. Kuehl RO. *Design of Experiments: Statistical Principles of Research Design and Analysis*. Pacific Grove, California, USA: Duxbury Press, 2000.
27. Simon R, Korn EL, McShane LM, *et al. Design and Analysis of DNA Microarray Investigations*. New York, USA: Springer, 2003;11–35.
28. Cui X. Experimental designs on high-throughput biological experiments. In: Lee JK, (ed). *Statistical Bioinformatics*. Hoboken, New Jersey: Wiley-Blackwell, 2010;201–17.
29. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.

30. Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 2002;**32**(Suppl 2):490–5.

31. Yang YH, Speed T. Design issues for cDNA microarray experiments. *Nat Rev Genet* 2002;**3**:579–88.

32. Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**:1344–9.

33. Oliver HF, Orsi RH, Ponnala L, *et al*. Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics* 2009;**10**:641.

34. Lu T, Lu G, Fan D, *et al*. Function annotation of rice transcriptome at single nucleotide resolution by RNA-seq. *Genome Res* 2010;**20**:1238–49.

35. Martin J, Zhu W, Passalacqua KD, *et al*. Bacillus anthracis genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics* 2010;**11**(Suppl 3):S10.

36. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.

37. Degner JF, Marioni JC, Pai AA, *et al*. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 2009;**25**:3207–12.

38. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 2008;**9**:321–32.

39. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.

40. Bullard JH, Purdom E, Hansen KD, *et al*. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010; **11**:94.

41. Balwierz PJ, Carninci P, Daub CO, *et al*. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol* 2009;**10**:R79.

42. Willenbrock H, Salomon J, Sokilde R, *et al*. Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. *RNA* 2009;**15**:2028–34.

43. Furusawa C, Kaneko K. Zipf's law in gene expression. *Phys Rev Lett* 2003;**90**:088102.

44. Ueda HR, Hayashi S, Matsuyama S, *et al*. Universality and flexibility in gene expression from bacteria to human. *Proc Natl Acad Sci USA* 2004;**101**:3765–9.

45. Heap GA, Yang JH, Downes K, *et al*. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet* 2010;**19**:122–34.

46. Serre D, Gurd S, Ge B, *et al*. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet* 2008;**4**:e1000006.

47. Wang X, Sun Q, McGrath SD, *et al*. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* 2008;**3**:e3839.

48. Ni T, Corcoran DL, Rach EA, *et al*. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* 2010;**7**:521–7.

49. Bashir A, Volik S, Collins C, *et al*. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* 2008;**4**:e1000051.

50. Dempsey MP, Nietfeldt J, Ravel J, *et al*. Paired-end sequence mapping detects extensive genomic rearrangement and translocation during divergence of Francisella tularensis subsp. tularensis and Francisella tularensis subsp. holarctica populations. *J Bacteriol* 2006;**188**:5904–14.

51. Au KF, Jiang H, Lin L, *et al*. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 2010;**38**:4570–8.

52. Maher CA, Palanisamy N, Brenner JC, *et al*. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci USA* 2009;**106**:12353–8.

53. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009;**4**:14.

54. Dohm JC, Lottaz C, Borodina T, *et al*. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008;**36**:e105.

55. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 2010;**38**:e131.

56. Linsen SE, de Wit E, Janssens G, *et al*. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* 2009;**6**:474–6.

57. Mamanova L, Andrews RM, James KD, *et al*. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 2010;**7**:130–2.

58. Thode HC Jr. Power and sample size requirements for tests of differences between two poisson rates. *J Royal Stat Soc Series D* 1997;**46**:227–30.

59. Ng HK, Tang ML. Testing the equality of two Poisson means using the rate ratio. *Stat Med* 2005;**24**:955–65.

60. Gu K, Ng HK, Tang ML, *et al*. Testing the ratio of two poisson rates. *Biom J* 2008;**50**:283–98.

61. Aban IB, Cutter GR, Mavinga N. Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. *Comput Stat Data Anal* 2008;**53**:820–33.

62. Lee ML, Whitmore GA. Power and sample size for DNA microarray studies. *Stat Med* 2002;**21**:3543–70.

63. Page GP, Edwards JW, Gadbury GL, *et al*. The PowerAtlas: a power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics* 2006;**7**:84.

64. Rockett JC, Hellmann GM. Confirming microarray data - is it really necessary? *Genomics* 2004;**83**:541–9.

65. Allison DB, Cui X, Page GP, *et al*. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;**7**:55–65.

66. Camarena L, Bruno V, Euskirchen G, *et al*. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. *PLoS Pathog* 2010;**6**:e1000834.

67. Feng L, Liu H, Liu Y, *et al*. Power of deep sequencing and agilent microarray for gene expression profiling study. *Mol Biotechnol* 2010;**45**:101–10.

68. Ramskold D, Wang ET, Burge CB, *et al*. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 2009;**5**: e1000598.

69. Dallas P, Gottardo N, Firth M, *et al*. Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR - how well do they correlate? *BMC Genomics* 2005;**6**:59.

70. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res* 2009;**37**:e75.

71. Robertson AG, Bilenky M, Tam A, *et al*. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* 2008;**18**:1906–17.

72. Johnson DS, Li W, Gordon DB, *et al*. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res* 2008;**18**:393–403.

73. Tuteja G, White P, Schug J, *et al*. Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res* 2009;**37**: e113.