Bioinformatics Advance Access published December 6, 2013

Gene expression

RNA-seq differential expression studies: more sequence, or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology, [3]Department of Human Genetics, University of Chicago

Associate Editor: Dr. Janet Kelso

ABSTRACT

Motivation: RNA-seq is replacing microarrays as the primary tool for gene expression studies. Many RNA-seq studies have used insufficient biological replicates, resulting in low statistical power and inefficient use of sequencing resources.

Results: We show the explicit trade-off between more biological replicates and deeper sequencing in increasing power to detect differentially expressed (DE) genes. In the human cell line MCF-7, adding more sequencing depth after 10M reads gives diminishing returns on power to detect DE genes, while adding biological replicates improves power significantly regardless of sequencing depth. We also propose a cost-effectiveness metric for guiding the design of large scale RNA-seq DE studies. Our analysis showed that sequencing less reads and perform more biological replication is an effective strategy to increase power and accuracy in large scale differential expression RNA-seq studies, and provided new insights into efficient experiment design of RNA-seq studies.

Contact: kpwhite@uchicago.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

RNA-seq has been widely used for differential expression studies (Oshlack, *et al*., 2010; Ozsolak and Milos, 2011). Despite the large number of studies performed for transcriptome comparisons, little empirical optimization has been made for RNA-seq based experimental designs. Critical issues include biological replication and sequencing depth (Auer and Doerge, 2010), and inefficient designs of RNA-seq studies can lead to sub-optimal power and waste of resources, especially in large scale treatment-control studies.

Although for most RNA-seq studies high technical reproducibility means that technical replicates are not necessary (Marioni, *et al*., 2008), this fact does not ameliorate the need for biological replicates in making statistical inferences (Hansen, *et al*., 2011). Yet frequently large scale RNA-seq studies with extensive differential expression analyses have employed limited biological replication, instead favoring a strategy of low level biological replication with very deep sequencing (e.g. (Brawand, *et al*., 2011; Graveley, *et al*., 2011; Hah, *et al*., 2011).

In addition to replication number, the choice for sequencing depth is often unguided. It is clear that higher sequencing depth generates more informational reads, which increases the statistical power to detect differentially expressed (DE) genes (Tarazona, *et al*., 2011). However, high sequencing depth comes with cost, and resources will be wasted in scenarios where more sequencing brings diminishing returns as a saturation level is approached.

To achieve maximum power to detect DE genes within a budget, a compromise must be made between sequencing depth and biological replication. There are a few previous studies on experimental design issues for RNA-seq studies (Auer and Doerge, 2010; Fang and Cui, 2011; Tarazona, *et al*., 2011), but they do not directly address the specific question raised here of the trade-offs between replication, sequencing depth and cost: should we sequence more samples with low depth, or should we sequence fewer samples with high depth?

2 METHODS

MCF-7 cells (from ATCC) were seeded in complete medium in 6cm2 plates until reaching 40% confluence, followed by incubation in medium with 10% charcoal-stripped serum for 3 days. The cells were then treated with either 10nM E2 or control for 24hrs. Qiagen RNeasy columns were used to extract mRNAs from these cells. Bioanalyzer was used to measure the integrity of all mRNAs samples to make sure all the samples have RIN number greater than 9.

RNA-seq libraries were made with Illumina TruSeq RNA sample preparation protocol by Institute for Genomics and Systems Biology Sequencing Center. The libraries were multiplexed with Illumina barcodes and 6 samples were sequenced per lane by Illumina HiSeq 2000. 50bp single end reads were generated for the datasets. 7 biological replicates of both control and E2-treated MCF7 cells were sequenced. All libraries have > 30 million reads sequenced.

All libraries were aligned to the hg18 human genome using Tophat (Trapnell, *et al*., 2012). We then randomly down-sampled the RNA-seq reads of each sample to generate datasets of 2.5M, 5M, 10M, 15M, 20M, 25M and 30M reads using Picard Version 1.61(Wysoker, *et al*., 2012). In all subsequent analysis, the total number of reads refers to total number of

*To whom correspondence should be addressed.

aligned reads. Using these down-sampled sequence reads, we generated raw counts of number of tags on each gene by using coverageBED program in the BEDTools package Version 2.16.2 (Quinlan and Hall, 2010).

edgeR (Robinson, *et al*., 2010) package (Version 2.6.9) was used to detect significantly differential expressed genes between control and E2-treated samples. Upper-quantile normalization was performed to normalize tag counts among different samples. Tag-wise dispersion of negative binomial distribution for each gene was estimated and used in the exactTest function in edger package to identify DE genes. Genes with fewer than 5 reads are removed from calculation. In the simulation, under each sequencing depth, treatment samples are randomly picked (without replacement) to compare with same number of control samples, and number DE genes were calculated using edgeR, with FDR < 0.05 (BH adjusted) as the cutoff. Each sequencing depth and biological replication is simulated 100 times.

For the power calculation and generation of ROC curves, a list of 3,292 genes is used as "true positives" for E2 regulated genes, which are the DE genes detected by edgeR, using 7 biological replicates, 30M sequencing depth, with a FDR cutoff of 0.001. Using this gene list we computed true positive rates and true negative rates for each replication level and each sequencing depth on varying FDR rates, then computed the power, and constructed the ROC curves based on these rates.

The coefficient of variation for the logFC was computed using the top 100 differentially expressed genes (defined as having the lowest FDR in edgeR using 7 replicates, 30M reads per replicate). Estimated logFC computed at each level of replication and sequencing depth was simulated 100 times the same way as above and CV was computed. logCPM (logarithm of counts per million reads) was used here as a proxy for the estimation accuracy for expression level instead of FPKM, because genes with similar tag counts will have similar level of randomness in expression estimation which made across genes comparison possible. CV of logCPM was calculated similar to CV of logFC. The high expression level genes were defined as genes with logCPM rank 1-100, medium expression level genes were defined as genes with logCPM rank 2001 – 2100, low expression level genes were defined as genes with logCPM rank 12001 - 12100.

When calculating cost per DE gene, we made the following assumptions: Illumina sequencing cost per lane is $1200 (including reagents, personnel, equipment depreciation and contracts), for each lane 150M reads can be produced, and maximum multiplexing for each lane is 24x. The fixed cost for each sample is the library preparation cost, which is assumed to be $250 (reagents and personnel).

3 RESULTS

3.1 Trade-off between sequencing depth and biological replication

We calculated the number of significantly differentially expressed genes between E2-treated MCF7 cells and control-treated MCF7 cells under various levels of biological replication and sequencing depth (Figure 1a; See Methods). The number of DE genes increases with both increased number of biological replicates and increased number of reads in each sample. However, the increase in number of DE genes with sequencing depth has diminishing returns after 10 million (10M) reads. For example, at a sequencing depth of 10M reads, using 2 biological replicates for a total of 20M combined reads, the average number of DE genes identified is 2,011. If we use 15M reads and 2 biological replicates for a total of 30M combined reads the number is 2,139, a 6 % increase for a 50% increase in reads. If instead we apply an additional 10M reads to another biological replicate (3 biological replicates for a total of 30M combined reads) we obtain an average of 2,709 DE genes, a 35% increase. Even if we triple the reads for the two biological replicates to 30M each (60M combined total), we find an average of 2,522 DE genes, an increase of only 27%. Similar results were observed when we used different significance cutoffs or using different software package DESeq (Anders and Huber, 2010) (Supplementary figure S1).

Moreover, as one might expect based on most other biological measurements (Sokal and Rohlf, 1995), substantial increase in power through replication occurs regardless of sequencing depth. At 30M depth, 2 replicates gives 2,553 DE genes, and 3 replicates gives 3,447 DE genes, a 35% increase. If samples are available, adding more biological replicates almost always increases power significantly. Adding biological replicates has diminishing returns only when number of replications is very high. Increase from 2 biological replicates to 3 biological replicates at 10M depth yielded a 34.7% increase in number of DE genes, but increase from 6 replicates to 7 replicates still added 26.3% more DE genes at this sequencing depth (Figure 1a). When we split genes into high, medium, and low expressers and plot the relationship between DE genes, sequencing depth and replication level separately, we see that biological replicates increase DE genes for genes of all expression levels, and are more effective than adding sequencing depth for all expression levels (Figure S3).

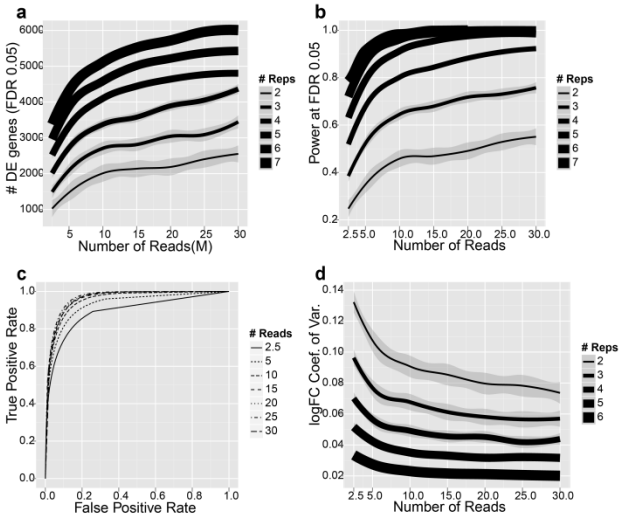

Fig. 1. (a) Increase in number of biological replication significantly increases the number of DE genes identified, while number of sequencing reads have diminishing return after 10M reads. Different color indicates different number of replication, with 2 replicate the darkest and 7 replicate lightest. The lines are smoothed average line of each replication level, with the shade corresponding to 95% confidence interval of the mean number of DE genes. (b) Power of detecting DE genes increases with both sequencing depth and biological replication level. Similar to the trends in (a), the power increases after 10M become smaller. (c) ROC curve for 3 biological replicates. Sequencing deeper than 10M reads does not significantly improve statistical power and precision for detecting DE genes. (d) The coefficient of variation (CV) of logFC for the top 100 differentially expressed genes. The CV of the logFC estimates decreases significantly as we add more biological replicates, while adding sequencing depth after 10M reads has much less effect.

Concordant with the total number of DE genes, statistical power also increases as more sequence or biological replicates are added (Figure 1b). Similar to the trends in total numbers of DE genes, we observed diminishing returns on power after 10M reads per sample. For example, with 2 replicates, 10M reads per sample (20M reads combined), we calculated a power of 0.46. When we tripled the number of reads to 30M reads per sample (60M reads combined), we observed a power of 0.55, only a 19.6% increase. In contrast, if we add another biological replicate at 10M reads (30M reads combined), we reach a power of 0.65, a 41.3% increase. When we split the genes into high, medium, and low expressers and plotted the relationship between power, sequencing depth and replication (Figure S4), similar trends were observed: replication adds significant power to detect DE genes regardless of expression, and is more effective than adding sequencing depth. If this strategy is adopted, one possible concern is that with lower sequencing depth, more genes will be dropped from the DE calculation, as most software packages remove genes with fewer than 5 reads. However, in our dataset as long as number of reads exceeds 10M, reducing sequencing depth has very small effects on the number of genes being removed. (Figure S5).

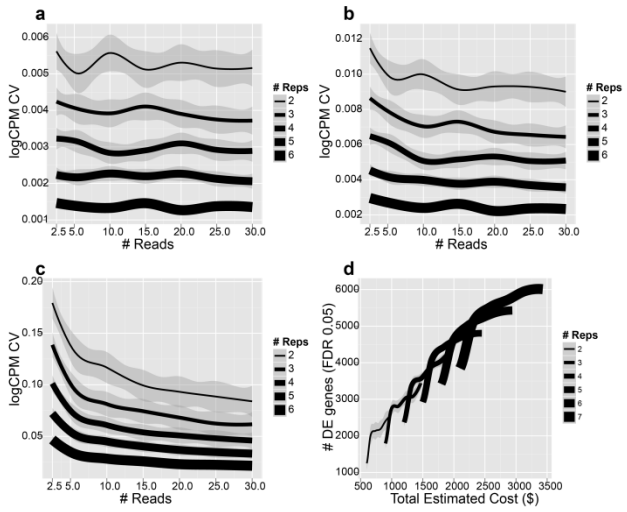

Fig. 2. (a-c) The coefficient of variation (CV) of logCPM (count per million reads) for high expression level genes (a), medium expression level genes (b), and low expression level genes (c) (See Methods for definition). High/medium expression level genes have very low CV for expression level estimates, adding sequencing depth do not have significant effect on accuracy of estimation, while adding biological replicates still improves accuracy significantly. For low expression level genes, both adding sequencing depth and adding biological replication level improves expression level estimation accuracy. (d) Number of DE genes plotted against the total estimated sequencing cost. If higher number of #DE is needed, increased number of biological replicates has to be used.

To look further into the false positive rates and false negative rates under these conditions, we constructed ROC curves for all sequencing depth and replication level (Figure 1c; see methods for details). At 3 biological replicates, 10M reads is nearly as good as 30M reads in terms of statistical power and precision (percentage of true positives among all positives). Curves for other replication levels showed very similar trends (Supplementary Figure S2). For ROC curves at 10M reads, similar to the trends in the power curves, 4 replicates is very close to 6 replicates, while power and precision gains from 2 replicates to 3 replicates, and 3 to 4 replicates, are more substantial.

We also examined individual gene log fold changes (logFC) and expression level estimation accuracy under different levels of replication and sequencing depth, to gain a quantitative idea of how accurate these estimates are under different conditions. For logFC estimates, we calculated the logFC coefficient of variation (CV) for the top 100 most differentially expressed genes (Figure 1d). For these 100 genes, adding sequencing reads after 10M reads barely has any effect on CV when replication is high, while biological replication continues to improve accuracy of logFC estimation significantly, High replication level gives accuracies that are probably not practically achievable by adding sequencing depth at low replication levels.

For expression level estimation, we examined three groups of genes: high, medium, and low expression level (See Methods). For these three groups of genes, the CV of logarithm counts per million reads (logCPM) was calculated and plotted against sequencing depth and replication level (Figure 2a-c). For highly expressed genes, expression level estimate accuracy is already very high (Figure 2a), and adding more reads has little effect on accuracy, while biological replicates still improves accuracy. For low expression genes (Figure 2c), CV for expression estimates are much larger, and accuracy is improved when either more reads or more replicates were added. For genes with medium expression level (Figure 2b), the situation is somewhat in between, as expected: adding more sequencing reads reduced CVs slightly, while biological replicates still reduced CV significantly. These results indicate that biological replicates improve the accuracy in estimating expression level for all genes, regardless of expression level, while adding sequencing depth will improve estimation accuracy mostly for low expression genes.

3.2 A metric for cost effectiveness

When choosing an experimental design for an RNA-seq differential expression study, the trade-off between number of biological replicates and sequencing depth is an important consideration, especially for large projects where many perturbation experiments are performed. Our results indicate that biological replicates are very important for increasing the power for DE gene detection regardless of the sequencing depth used.

In order to guide experimental designs of RNA-seq studies for differential expression, we propose the following simple metric:

> *Cost per 1% power given a particular design = (fixed costs per sample * number of samples + sequencing costs) / power*

The cost per 1% power metric measures the cost effectiveness of a given study design. Fixed costs per sample include library construction costs, sample costs and labor costs. Sequencing costs are variable costs for each sample depending on the sequencing depth and multiplexing scheme used. In study designs for RNA-seq DE studies, we can compare different designs using cost per 1% power after defining our total budget, and desired power.

Using this formula and some cost assumptions (see methods for details), we calculated the cost per 1% power for different designs

of our experiment (Table 1). For our samples, the lowest cost per 1% power was achieved at the 10M sequencing depth for 2-6 replicates. The cost per 1% power did increase slightly when we added more biological replicates, but having more biological replicates also means higher power (Figure 2b). If a larger number of DE genes is desired in the study, the number of samples has to be employed in the study can be decided based on such "standard curves". However, note that our cost calculation here does not reflect the sample collection cost, because it varies hugely from project to project. For human cell line studies we presented here, sample collection cost is relatively low, but for other projects, the sample collection cost can dominate the cost calculation. The investigator should definitely take sample collection cost into consideration when designing the project.

Table 1. Cost efficiency for power to detect DE genes (cost per 1% power given each experimental design where the variables are). Assumptions made during calculations are described in Methods. * indicates lowest cost per 1% power in each replication level. Units are in dollars.

Relative Cost	2.5M	5M	10M	15M	20M	25M	30M
2 replicates	24.2	17.2	14.4*	15.8	16.7	17.0	17.8
3 replicates	23.4	17.2	15.3*	16.3	17.1	18.5	19.4
4 replicates	23.1	17.7	16.5*	17.5	18.6	19.8	21.2
5 replicates	23.8	19.0	18.1*	19.4	21.0	22.8	24.9
6 replicates	25.0	20.7	20.6*	22.4	24.6	27.0	29.4
7 replicates	26.8	23.0*	23.5	26.0	28.7	31.5	34.3

4 CONCLUSION

We conclude that in a typical DE study using RNA-seq, sequencing deeper for each sample generates diminishing returns for power of detecting DE genes once beyond a certain sequencing depth. Instead, increasing the number of biological replications consistently increases the power significantly, regardless of sequencing depth. Additionally estimation accuracy for log fold changes and absolute expression levels greatly improve across the board when more biological replicates are added, while sequencing depth improves the accuracy of these estimations only in some situations, So, when possible, using more biological replication with lower sequencing depth, instead of sequencing few samples in great depth, is a more efficient strategy for RNA-seq DE studies. In the specific case of MCF7 breast cancer cell samples, our cost metric suggests that sequencing more than 10M reads per sample gives diminishing returns compared to adding replication. Obviously, for other species and perhaps other samples such as heterogeneous tumor samples, the exact sequencing depth will be different, but the overall guideline of replication rather than deeper sequencing should still remain the same. A similar set of standard curves could be constructed for each type of sample to guide experimental designs. We argue that such a metric is useful in designing large-scale genomic studies to optimize cost effectiveness. Almost all individual laboratories are mindful of budgets, but the stakes are particularly high in studies such as ENCODE or TCGA where millions of dollars are being spent on sequencing. Careful consideration needs to be given to cost effectiveness.

We have focused on differential expression studies using RNA-seq with the aim to improve a single target: power to detect differentially expressed genes between samples. Of course, there are cases where sequencing very deeply is advantageous (such as differential expression of exons, and transcript specific expression,). In these applications, much higher sequencing depths are required, because the informative genomic regions are much shorter. However, if gene differential expression is the primary goal, it would be a sensible choice to optimize sequencing depth and number of biological replicates according to the simple guidelines we propose here.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Institute of General Medical Sciences (P50GM081892), and by the Searle Funds at The Chicago Community Trust from the Chicago Biomedical Consortium.

REFERENCES

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data, *Genome Biol*, **11**.

Auer, P.L. and Doerge, R.W. (2010) Statistical design and analysis of RNA sequencing data, *Genetics*, **185**, 405-416.

Brawand, D., *et al.* (2011) The evolution of gene expression levels in mammalian organs, *Nature*, **478**, 343-348.

Fang, Z. and Cui, X. (2011) Design and validation issues in RNA-seq experiments, *Brief Bioinform*, **12**, 280-287.

Graveley, B.R., *et al.* (2011) The developmental transcriptome of Drosophila melanogaster, *Nature*, **471**, 473-479.

Hah, N., *et al.* (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells, *Cell*, **145**, 622-634.

Hansen, K.D., *et al.* (2011) Sequencing technology does not eliminate biological variability, *Nat Biotech*, **29**, 572-573.

Marioni, J.C., *et al.* (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays, *Genome Research*, **18**, 1509-1517.

Oshlack, A., Robinson, M.D. and Young, M.D. (2010) From RNA-seq reads to differential expression results, *Genome Biol*, **11**, 220.

Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities, *Nat Rev Genet*, **12**, 87-98.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841-842.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139-140.

Sokal, R.R. and Rohlf, F.J. (1995) The principles and practice of statistics in biological research, *New York.: Edition*, **3**.

Tarazona, S., *et al.* (2011) Differential expression in RNA-seq: a matter of depth, *Genome Res*, **21**, 2213-2223.

Trapnell, C., *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat Protoc*, **7**, 562-578.

Wysoker, A., Tibbetts, K. and Fennell, T. (2012) Picard: http://picard.sourceforge.net.