Briefings in Functional Genomics Advance Access published December 30, 2011

doi:10.1093/bfgp/elr041

Differential expression—the next generation and beyond

Paul L. Auer, Sanvesh Srivastava and R.W. Doerge

Abstract

RNA-sequencing (RNA-seq) technologies have not only pushed the boundaries of science, but also pushed the computational and analytic capacities of many laboratories. With respect to mapping and quantifying transcriptomes, RNA-seq has certainly established itself as the approach of choice. However, as the complexities of experiments continue to grow, there is still no standard practice that allows for design, processing, normalization, efficient dimension reduction and/or statistical analysis. With this in mind, we provide a brief review of some of the key challenges that are general to all RNA-seq experiments, namely experimental design, statistical analysis and dimensionality reduction.

Keywords: *experimental design; statistical bioinformatics; curse of dimensionality; dimension reduction; RNA-seq; differential expression*

INTRODUCTION

Next-generation sequencing, or high-throughput deep sequencing (HTDS), is a versatile technology that is being applied in a variety of ways. Variant discovery [1], profiling of histone modifications [2], identification of transcription factor binding sites [3] and resequencing [4] are among the most popular applications. It also had a significant impact on molecular biology [5], the estimation of allele frequencies [6], as well as the identification of induced mutations [7]. By comparison to these applications, RNA-sequencing (RNA-seq) may be leading the pack in popularity because of its ability to characterize transcriptomes [8], to assess differential gene expression [9] and to essentially challenge the continued use of microarray technology for studying transcription. No one would argue that microarray technology certainly served more than a few very important purposes. It enabled whole genome studies in a variety of applications and organisms, it supplied vast amounts of data that continue to challenge both the biological and bioinformatics communities, and it brought to bear the importance that good experimental design and proper statistical analysis have on good science. Considering the time and resources dedicated to studying the statistical design, processing and analysis of microarrays [10–12], a similar investment needs to be made for RNA-seq data [13, 14].

Even though HTDS, and all of its applications, is a major leap forward in the biological sciences the statistics community has been slow to embrace it with the same gusto that microarray technology received [15–17]. This has left the biological communities wondering how to design an RNA-seq experiment, how to deal with the huge RNA-seq data files, and finally how to statistically test and analyze the data. We are concerned that next-generation technologies will, most likely, remain expensive for a while, and that there may be an inclination to revert to single sample science that is void of any ability to estimate biological and/or technical variation, or to test scientific hypotheses. With this in mind, we provide a brief review of some of the key

Corresponding author. R.W. Doerge, Department of Statistics, Purdue University, West Lafayette, IN 47907, USA. Tel: +1 765 494 6030; Fax: +1 765 494 0558; E-mail: doerge@purdue.edu

Paul L. Auer completed his PhD under the direction of R.W. Doerge. He is a Staff Scientist at the Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

Sanvesh Srivastava is currently a PhD student under the direction of R.W. Doerge in the Department of Statistics, Purdue University. His research is focused on statistical inference for next-generation sequencing data using hierarchical Bayesian models.

R.W. Doerge is Professor Statistics and Professor of Agronomy, Purdue University, West Lafayette, IN USA. She is Director of the Statistical Bioinformatics Center, and Head of the Department of Statistics.

challenges that are general to all RNA-seq experiments, namely experimental design, statistical analysis and dimensionality reduction.

UNDERSTANDING THE DATA

In practice, any HTDS technology can be used to provide RNA-seq data. The three most popular, commercially available platforms are those produced by Applied Biosystems (SOLiD), 454 Life Sciences (454-Sequencing) and Illumina (Solexa) (others include Helicos Biosciences, Pacific Biosciences, Complete Genomics and Oxford Nanopore). The Solexa and SOLiD platforms offer enriched depth of coverage over 454 (18 GB per run, 30 GB per run and 450 MB per run, respectively) at the cost of read length (75, 50 and 330 bp, respectively) [18]. The trade-offs of this balance are entirely application dependent and all three have been used to study the transcriptomes of various organisms. For simplicity, we base our discussion generally on a flow cell composed of multiples lanes, each of which is capable of sequencing independent genomic samples, and thus producing tens of millions of short sequence reads per lane.

Raw data from a single lane of a flow cell contain sequencing reads of fixed length (e.g. 50 bases), accompanied by quality scores for each base. These reads must be assembled *de novo* or aligned to a reference database (e.g. the known transcriptome or the reference genome). Aligning or assembling many millions of short sequence reads presents new and difficult challenges, namely computational efficiency, resolving mapping ambiguities, map bias and inaccuracies in both the sequence reads, as well as the reference database [19, 20]. With respect to the aligned reads, they occupy a distinct data file for each lane, typically requiring at least 10 GB of computer disk space per file. Depending on the application and/or question asked (e.g. alternative splicing, coding region discovery, etc.), the size of the data files can be greatly reduced to provide only data required for the statistical analysis. For example, a typical Solexa flow cell with eight lanes can provide more than 80 GB of raw data that can be bioinformatically reduced to a much more manageable file of approximately 10–30 MB.

RNA-seq data may be used to investigate many different phenomena, such as alternative splicing [21], isoform expression [22] and allele specific expression [23]. However, for this discussion and for clarity, we will rely on the context of a gene expression study exploring differential expression between two treatment groups. If we consider two samples, A and B, biological replicates of each sample are needed to both acknowledge the biological variation and to assess the statistically validity of the comparison. For example, each sample is processed independently such that each gene's expression is measured by sequencing transcripts and counting the number of reads that map back to each gene. Obviously, these are count data that should be considered as discrete random variables; again very different from the continuous data that microarray technology supplies.

Once the data have been tabulated, normalization is required. Because lane-specific coverage, substitution errors and the resulting alignment are all known issues [9, 19, 24], any of the previous analogies that we have made to microarrays end here. The standard normalizing techniques for microarray data do not apply. For example, quantile normalization, which effectively forces the quantiles of the empirical distribution to be identical across samples, renders the upper and lower ends of the distributions indistinguishable across samples, thus robbing RNA-seq data of one of its most advantageous features, dynamic range. To date, everyone acknowledges the issue and need for normalization, but there are few standard normalizing techniques [25, 26]. The most popular approach (RPKM, [8]) effectively normalizes the gene counts in each sample by gene length and the total number of mapped reads in that sample, but this unfortunately ignores underlying data patterns. Therefore, it seems that estimating depth of coverage at the exon or gene level provides the most effective means to normalizing across samples.

Since the normalized RNA-seq data are based on counts, one immediately considers the Poisson distribution when describing the population of counts or when formulating any sort of model. It turns out that when the data have no variability (i.e. no replication) or are distributed similarly (i.e. technical replicates), then indeed the data are Poisson distributed [9]. Unfortunately, these two scenarios (i.e. no replication and/or technical replicates) provide no information about the biological question at hand (i.e. differential expression). In fact, biological variability in count data is not accurately described by a Poisson distribution [27–29]. Further, if one ignores that variation between biological samples exists (i.e. the sampling error), typically it results in an increased

false positive rate. Even though the amount of technical variation from RNA-seq experiments is low [9, 30, 31], it is important to remember that biological variation provides the information for testing scientific questions.

EXPERIMENTAL DESIGN

Designing an experiment that minimizes error, maximizes information and answers the question at hand requires three important principles that can be credited to R.A. Fisher, randomization, replication and blocking [13, 32]. When dealing with RNA-seq data, randomizing samples across the lanes on the flow cell in a completely randomized fashion effectively averages out any effect that lanes might have on the gene counts. In a similar fashion, randomizing treatments across lanes assures that potential lane effects are not confounded with the treatment effects. Meaning, one can separate lane effects from the effect of interest, namely the treatment. If more than one machine is used, or more than one flow cell on the same machine, blocking will account for the variation associated to these differences. Table 1 illustrates a single flow cell with eight lanes used to compare two treatments (A and B) each with four biological replicates. An analysis of these data proceeds in a similar manner as a microarray experiment with the same design.

Auer and Doerge [13] detail the implications of proper experimental design for both unreplicated and replicated experiments. For replicated experiment, they rely on well-known statistical designs that partition the sources of variation and suggest statistical models for testing hypotheses for differential expression. Through randomization, replication and blocking it is possible to design an experiment that allows more refined questions to be asked of the data. Given technical and biological replicates, appropriately allocated to lanes and sequencing runs, one can simultaneously test for differential expression and lane effects. A suitable experimental design for this purpose is called a D-optimal split plot design [33]. D-optimal refers to the design that minimizes the generalized variance [34]. Blocking by sequencing run removes the possibility of confounding lane effects with any effects introduced by differences in sequencing runs. In keeping with the first example (Table 1), if we assume four biological and four technical replicates, the D-optimal design for which lane and sequencing run are not confounded is illustrated in Table 2.

Of course, this is a hypothetical example meant to demonstrate how a researcher can optimize the information extracted from an experiment. In reality, the choice of biological and technical replicates is governed by time, resources, research goals and variability in the population of interest. The microarray literature is rich with suggestions on how to choose between biological and technical replicates [15, 35] and for the most part these suggestions hold true for RNA-seq experiments. Keep in mind, however, that although technical variation has been studied extensively in microarrays, by comparison there are relatively few investigations of technical variation for next-generation sequencing technologies [9, 30, 31].

Table 1: Biological samples have been randomly assigned to lanes

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
a_1	b_4	b_2	a_2	b_3	a_4	b_1	a_3

There are two treatment groups A and B, with four biological replicates each $a_1, \ldots, b_4$.

Table 2: D-optimal split plot design [33] that minimizes the generalized variance

	Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
Run 1	a_1	b_1	b_1	a_1	b_1	a_1	b_1	a_1
Run 2	a_2	b_2	a_2	a_2	b_2	b_2	b_2	a_2
Run 3	b_3	a_3	b_3	a_3	a_3	b_3	a_3	b_3
Run 4	a_4	b_4	b_4	b_4	a_4	a_4	a_4	b_4

Within each run there are two treatments (A and B) each with four technical replicates (a_i and b_i; $i = 1, 2, 3, 4$). Four biological replicates correspond to the four runs.

THE CURSE OF DIMENSIONALITY

One very certain outcome of HDTS technologies, independent of application, is that the dimensionality of the data is unwieldy. Since more and more of the genome is being explored at the base pair level, asking questions using a relatively small number of (biological) samples is becoming a real challenge. This challenge is also known as the curse of dimensionality. Small samples provide only so much information from which to ask questions. In the statistical community, the issue of having more independent or predictor variables (e.g. genetic markers, SNPs, gene counts, etc.) measured on relatively few samples is known as the 'large p, small n' problem; the data are called high-dimensional data. Basically, there are too many parameters to estimate, with not enough information to do it. Since the biological samples provide the information that describes the behavior of the total population, the more samples one observes, the better one can describe a biological phenomenon such as transcript abundance or differences in transcript abundance. Theoretically, with a large enough number of samples (i.e. upwards of tens of thousands) one could employ genome-wide predictors to accurately estimate differential expression. In the absence of many biological samples, one option is to rely on proper experimental design, proper modeling of the data to partition variation and test biological phenomena, and to incorporate the available biological information in the models. Oshlack *et al.* [36] and Salzman *et al.* [14] provide an excellent review of the current statistical methods for analyzing RNA-seq data. Efron [37] provides a thorough overview of the underlying theoretical motivation and practical guidelines for analyzing high-dimensional data, like RNA-seq data. All these approaches reduce the dimensionality of the data by identifying a small fraction of genes suitable for further exploration.

MACHINE LEARNING

High-dimensional data are not limited to genomics. They also occur in every avenue of the world (finance, communication, retail, etc.), and as a result this area of study has seen an influx of statistical research. Specific to genomics, microarray technologies were the first to introduce high-dimensional data, which resulted in significant advances in the theory of multiple hypotheses testing [38–40], variable selection [41] and false discovery rates (FDRs) for multiple testing [42–44]. Issues for analysis of data from HDTS technologies are similar to those found in microarray analysis, but they are much greater because of the complexity of the data and because the applications are more sophisticated.

Machine learning is a relatively new interdisciplinary area that rests at the intersection of computer science and statistics. It has grown from computer engineering applications that use algorithms to learn trends and to look for patterns in data, with an emphasis on the computational efficiency. The ideal application to genomics combines the computational efficiency of machine learning methods with the widely used statistical procedures for discovering patterns in large amounts of data; specifically for situations in which classical statistical approaches would be intractable.

Very recently, Srivastava and Doerge [45] proposed a flexible approach, similar to clustering, for identifying latent patterns in high-dimensional count data. When applied to count data from HDTS technologies, the approach identifies subsets of genes with similar expression patterns and that explain a large portion of variability. The fundamental concept behind their modeling strategy is that a small fraction of genes, organized into groups, are responsible for a significant amount of biological variation. Specifically, their model is a three-level hierarchical Bayesian model that discovers the underlying biological functions, or latent variables, by modeling the Poisson data and sample-specific functional probabilities as a mixture distribution, respectively. This modeling approach is referred to as latent process decomposition (LPD) and has been implemented in an R/Bioconductor package called *themes*. Due to the biological motivation behind the model, LPD provides interpretable parameter estimates that other clustering-based approaches cannot. LPD gains its power and flexibility from the underlying Bayesian modeling strategy that makes it modular and extensible, and very different from existing classification and prediction methods. The approach developed by Srivastava and Doerge [45] is easily extendable to many other applications.

DISCUSSION

It is very clear that RNA-seq technologies have pushed the boundaries of science, as well as the computational limits of many laboratories. To date,

there is no standard practice that allows for design, processing, normalization, efficient dimension reduction and statistical analysis. Specific to mapping and quantifying transcriptomes, RNA-seq has certainly established itself as the state of the art tool. Early indications suggest that results from RNA-seq experiments are highly reproducible and the range of scientific questions that can be addressed by harnessing the full potential of this technology is just now being explored.

Where we go from here is highly dependent on the steps that the scientific community takes. Unfortunately, statisticians have not been as engaged with RNA-seq data as compared to the explosion of research generated by those involved in microarray experiments. RNA-seq raw data files are disorganized and too large to handle on a personal computer. Alignment and assembly require server-sized processors, and there are no standard bioinformatics pipelines for summarizing the raw data into organized matrices. Combined with the fact that RNA-seq is still relatively new, it is easy to see why the statistics community has been slow to catch on. However, this situation will not persist. As noted by Wang *et al.* [46], 'As the cost of sequencing continues to fall, RNA-seq is expected to replace microarrays in many applications.' With this rise in popularity, our hope is that statisticians will start to take a closer look at the quantitative aspects of this new tool.

Key Points

- Next-generation sequencing data experiments are complex and require proper experimental design.
- There are unresolved statistical issues that deal with normalization, biological replication and dimensionality reduction.
- Machine learning techniques have great potential in genomic applications for the purpose of reducing the dimensionality.
- The amount of next-generation sequencing data, and its dimensionality, will continue to increase, so the statistical issues associated with these data require qualified statisticians' attention.

References

1. Hillier LW, Marth GT, Quinlan AR, *et al*. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* 2008;**5**:183–8.
2. Mikkelsen TS, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007; **448**:553–60.
3. Valouev A, Johnson DS, Sundquis A, *et al*. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 2008;**5**:829–34.
4. Bentley DR. Whole-genome re-sequencing. *Curr Opin Genet Dev* 2006;**16**:545–52.
5. Futschik A, Schlötterer C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 2010;**186**:207–18.
6. Lynch M. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* 2009;**182**: 295–301.
7. Blumenstiel JP, Noll AC, Griffiths JA, *et al*. Identification of EMS-induced mutations in drosophila melanogaster by whole-genome sequencing. *Genetics* 2009;**182**:25–32.
8. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
9. 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.
10. Larkin JE, Frank BC, Gavras H, *et al*. Independence and reproducibility across microarray platforms. *Nat Methods* 2005;**2**:337–44.
11. Nettleton D. A discussion of statistical methods for design and analysis of microarray experiments for plant scientists. *Plant Cell* 2006;**18**:2112–21.
12. Irizarry RA, Hobbs B, Collin F, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;**4**:249.
13. Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. *Genetics* 2010;**185**:405–16.
14. Salzman J, Jiang H, Wong WH. Statistical modeling of RNA-seq data. *Statistical Science* 2011;**26**(1):62–83.
15. Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 2002;**32**:490–5.
16. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. *Genetics Res* 2001;**77**: 123–8.
17. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostatistics* 2001;**2**:183–201.
18. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet* 2010;**11**:31–46.
19. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;**18**:1851–8.
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**: 1754–60.
21. Trapnell C, Williams BA, Pertea A, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–5.
22. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 2009;**25**:1026–32.
23. 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.
24. Hutchison III CA. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* 2007;**35**:6227–37.
25. Bullard JH, Purdom E, Hansen K, *et al*. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010; **11**:94.

26. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biol* 2010;**11**:R25.
27. Robinson MD, Smyth GK. Moderated statistical test for assessing differences in tag abundance. *Bioinformatics* 2007; **23**:2881–87.
28. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA sequencing differential expression analysis with Myrna. *Genome Biol* 2010;**11**:R83.
29. Auer PL, Doerge RW. A two-stage Poisson model for testing RNA-seq data. *Stat Appl Genet Mol Biol* 2011;**10**: 1–26.
30. 't Hoen PAC, Ariyurek Y, Thygesen HH, *et al*. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 2008;**36**: e141.
31. McIntyre LM, Lopiano KK, Morse AM, *et al*. RNA-seq: technical variability and sampling. *BMC Genomics* 2011;**12**: 293.
32. Fisher RA. *The Design of Experiments, 1935*. 3rd edn. London: Oliver & Boyd Ltd.
33. Jones B, Goos P. A candidate-set-free algorithm for generating D-optimal split-plot designs. *J R Stat Soc Ser C* 2007; **56**:347–64.
34. Wit E, Nobile A, Khanin R. Simulated annealing for near-optimal dual-channel microarray designs. University of Glasgow, 2004.
35. Wit E, McClure J. *Statistics for Microarrays; Design, Analysis and Inference*. Chichester: John Wiley & Sons, 2004.
36. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol* 2010; **11**(12):220.
37. Efron B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, 2010.
38. Efron B, Tibshirani R, Storey JD, *et al*. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 2001; **96**(456):1151–60.
39. Efron B. Large-scale simultaneous hypothesis testing. *J Am Stat Assoc* 2004;**99**(465):96–104.
40. Efron B. Size, power and false discovery rates. *Ann Stat* 2007;**35**(4):1351–77.
41. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softwr* 2010;**33**(1):1.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;**57**(1):289–300.
43. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001;**29**:1165–88.
44. Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat Sci* 2008;**23**(1):1–22.
45. Srivastava S, Doerge RW. Latent process decomposition of high-dimensional count data. Technical Report 11–03, Purdue University, 2011.
46. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**: 57–63.