John Storey

John Storey provides his take on the importance of new statistical methods for high-throughput sequencing.

The challenges of deriving statistically robust conclusions from high-throughput technologies such as microarrays are well known. Yet the exploration of similar challenges for data from sequencers, which can generate orders of magnitude more data than an array, is largely at the beginning stages. In this edited interview, John Storey, associate professor in the Lewis-Sigler Institute and the Department of Molecular Biology at Princeton University in New Jersey, and creator of the influential *q* value method for estimating false discovery rates[1], discusses the challenges of using next-generation sequencing as a quantitative measurement tool in such applications as RNA-Seq.

***Nature Biotechnology*: What are the statistical challenges of analyzing next-generation sequence data?**

John Storey: Statistics becomes particularly important when sequencing technology is being used quantitatively. In terms of traditional applications of just sequencing a genome, I think that there is an abundance of methods already available that addresses many of the major questions. But as soon as any sort of quantification comes into play, such as using read depth to quantify copy number variation or RNA transcript abundances (RNA-Seq), then statistical modeling—and particularly normalization—becomes much more important.

What is meant by statistical modeling?

JS: One of the main roles that statistics play in science is explaining variation—variation of observed data. That variation can actually be true signal that you're interested in, but there can also be variations due to noise or confounding signal. So I think of statistical modeling as the process of explaining variation in the data according to concrete variables that have been meausured.

Statistical modeling is largely enumerating the sources of variation, and then coming up with a mathematical way of explaining variation of one variable in terms of that in other variables. To test a model, you take some measurements and see how well the mathematics can predict the actual observations. Just like in science in a general sense, statisticians try to explain a complex phenomenon with simpler parts that we can understand and control better.

What are examples of sources of variation in a sequencing experiment?

JS: One is that short reads cannot always be precisely mapped back to the genome, and this presently causes many of the reads to be discarded. What sorts of biases are introduced by this imprecise mapping process? This question is not well understood.

A second example is that there is a nontrivial upper bound on the number of reads that can be produced per 'lane' [reaction chamber] of sequencing, depending on the technology and the lab's set-up. This implies that there is now a competition among the genomic features being measured. For example, in many of the early RNA-Seq experiments, a small percentage of genes made up almost half of the reads. This 'competition' among genes is a feature not really present in gene expression microarrays. Differential expression therefore becomes differential relative abundance among genes for RNA-Seq, which is not the same as differential expression in the microarray context.

Explain this competition effect more.

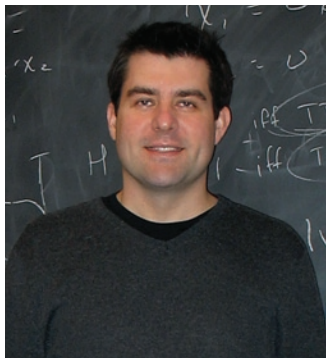
JS: Let's say my sequencer can measure about 50 million reads per lane. Then there's competition to be represented among those 50 million reads. In other words, if the expression of a subset of genes is very high in one sample, that's going to push down the number of reads you see for other genes.

Suppose that you had a particular gene that had a statistically equivalent number of reads across all of your lanes—about 100 reads—and also suppose that any fluctuations were no different from what you would expect by chance. Although you might think the gene is not differentially expressed, that isn't necessarily the case because the absolute mRNA level could be changing from sample to sample. All it means is that the relative abundance of the gene, with respect to all of the RNA that was extracted, is not changed.

So sequencing depth, or coverage, is a critical parameter?

JS: It really is. Continuing with the previous example, you have a fixed measurement capacity of 50 million reads. If one gene is really overexpressed, it could essentially 'use up' most of those reads and leave fewer reads left over for other transcripts. Microarrays are less susceptible to this because the probes presumably don't cross-hybridize too badly, and so even if one probe matches a ton of stuff in the sample, the excess all just gets washed off.

"The sources of variation in these experiments are not well understood."



How can unwanted variation be 'removed' using statistics?

JS: This process of accounting for, and possibly removing, sources of variation that are not of biological interest is called normalization. There are two distinct approaches to normalization.

One of them I would call 'unsupervised' in that it does not take into account the study design. These are the most popular methods because they require the least amount of statistical modeling and knowledge of statistics.

The other approach, which is the one I strongly favor, is what I would call 'supervised normalization'. This approach directly takes into account the study design. I find this appealing because if one is trying to parse sources of variation, then it seems that all sources of variation should be considered. If I perform an experiment with 20 microarrays, say 10 treated and 10 control, then I want to utilize this information when separating and removing technical sources of variation.

Another component of normalization, which is gaining popularity, is normalizing by principal components. Again, I think this should be done in the context of the study design, which was the goal of a recent method I worked on called 'surrogate variable analysis'[2].

Walk me through a simple example of when normalization is needed.

JS: Let's say you are doing an RNA-Seq experiment. The sequencer may produce a different number of total reads from lane to lane, and that is more than likely driven by technology, not biology. And so, normalization would be about accounting for those differences.

Unsupervised normalization might involve simply just dividing by the number of lanes and taking each gene as a percentage of the reads in the lane. Why is that less than ideal? Suppose that you have two batches of data, one flow cell that was done in November and another flow cell that was done in December. If you're actually accounting for this variation in the total number of reads per lane, my inclination would be to take into account the fact that these two flow cells were processed in different months. And it can be more complicated than that, too. Maybe you've taken clinical samples and there were some differences in the clinical conditions under which they were taken. In supervised normalization, you would actually take that information into account. For example, the adjustment made to the raw reads may be based on a model that includes the total number of reads per lane as well as the information about the study design, such as batch and biological variables.

Why can't we just use the same normalization methods that have been developed for microarrays?

JS: Microarrays start with probes that have been designed or constructed before any profiling or genotyping is performed. This means that microarray measurements are based on currently available genome annotations. Next-gen sequencing runs in the reverse direction: the measurements, in the form of short reads, are products of the biological sample. How the reads are processed determines how the genome is probed, and this can be influenced by many factors not present in microarrays.

Why is it hard to normalize sequencing data?

JS: The most immediate challenge is that the sources of variation in these experiments are not well understood. Until we have accurate knowledge and models of these sources of variation, normalization will be somewhat of a guessing game. The best way to build better models is to produce several comprehensive, well-designed pilot experiments. I believe that collaboration between industry and academia could rapidly produce such data sets. The pilot experiments I have seen so far (some of them made freely available from companies) have too many confounding sources of variation to be that useful for understanding normalization.

At a more conceptual level, I believe that normalization is challenging for sequencing experiments because it should heavily depend on the sequencing platform, the genomic background, the method by which RNA or DNA is isolated, the library preparation technique, the level of genetic variation that is present, the study design, the read length, the total number of reads per lane, whether paired-end sequencing is performed and other variables. That is a long list to consider.

Box 1 Suggested reading

John Storey identifies papers that readers can consult to learn more about the basic concepts in normalization of DNA microarrays, supervised normalization of DNA microarrays and batch effects (and other confounding sources of variation due to the study design, which are a concern for all high-throughput data utilized for quantification, even next-generation sequencing). In terms of next-generation sequencing, he recommends that readers build an understanding in the following areas: new statistical methods specifically for next-generation platforms (currently, most papers fail to provide a sufficient treatment of normalization, reflecting the fact that the tools just aren't yet there); sources of bias in next-generation sequencing data and robust and comprehensive normalization methods for dealing with bias (which are being developed earlier than they were for DNA microarrays); and statistical issues in RNA-Seq. Papers in each of these areas are provided in the lists below.

Unsupervised normalization of DNA microarrays

Li, C. & Wong, W. *Genome Biol.* **2**, 0032.1–0032.11 (2001).

Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. *Bioinformatics* **19**, 185–193 (2003).

Yang, Y. *et al. Nucleic Acids Res.* **30**, e15 (2002).

Supervised normalization of DNA microarrays

Kerr, M.K., Martin, M. & Churchill, G.A. *J. Comput. Biol.* **7**, 819–837 (2000).

Mecham, B.H., Nelson, P.S. & Storey, J.D. *Bioinformatics* **26**, 1308–1315 (2010).

Batch effects

Leek, J.T. *et al. Nat. Rev. Genet.* **11**, 733–739 (2010).

New statistical models in next-generation sequencing (annotations by John Storey)

Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
This paper addresses the fact that the variance of a gene's expression measurements heavily depends on the total number of reads for that gene. A consequence of this is that genes with more reads will have a higher statistical power and therefore be preferentially selected for differential expression.

Katz, Y., Wang, E., Airoldi, E.M. & Burge, C.B. *Nat. Methods* **7**, 1009–1015 (2010).

Gresham, D. *et al. Genetics* **187**, 299–317 (2011).
This paper highlights the fact that sequencing measures relative abundance (barcodes here), and additional information is needed for absolute abundance. We were able to quantify both in this paper.

Pickrell, J.K. *et al. Nature* **464**, 768–772 (2010).
The authors sequenced RNA from 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals from the HapMap Project. They identified new putative protein-coding exons, established extensive use of unannotated untranslated regions, detected allele-specific expression and allele-specific splicing. A number of innovative statistical ideas are utilized, but they are far from being fleshed out for general use. The large sample size here proved to be very powerful for overcoming some of the drawbacks of having a low number of reads for most of the genes.

Sources of bias in next-generation sequencing data

Bravo, H.C. & Irizarry, R.A. *Biometrics* **66**, 665–674 (2010).

Degner, J.F. *et al. Bioinformatics* **25**, 320712 (2009).

Hansen, K.D., Brenner, S.E. & Dudoit, S. *Nucleic Acids Res.* **38**, e131 (2010).

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. *Bioinformatics* **26**, 493–500 (2010).

Li, J., Jiang, H. & Wong, W.H. *Genome Biol.* **11**, R50 (2010).

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. *Genome Res.* **18**, 1509–1517 (2008).

Robinson, M.D. & Oshlack, A. *Genome Biol.* **11**, R25 (2010).

Young, M.D., Wakefield, M.J., Smyth, G.K. & Oshlack, A. *Genome Biol.* **11**, R14 (2010).

Statistical issues in RNA-Seq

Auer, P.L. & Doerge, R.W. *Genetics* **185**, 405–416 (2010).

Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. *BMC Bioinform.* **11**, 94 (2010).

Oshlack, A., Robinson, M.D. & Young, M.D. *Genome Biol.* **11**, 220 (2010).

Wang, Z., Gerstein, M. & Snyder, M. *Nat. Rev. Genet.* **10**, 57–63 (2009).

What else?

JS: The fact that the technological developments are moving so fast is another challenge. Microarray technology more or less hasn't changed since the early days. Whereas, it's clear with next-generation sequencing that really fundamental aspects of the data could change every three or four months. Methods development goes much more slowly than the technology developments.

It took about seven years from the publication of the early papers describing microarrays in 1995 and 1996 until a lot of the fundamental ideas behind modeling microarrays were established in 2002 and later. However, there's a possibility that what we have learned from microarrays can be leveraged to speed up that process with next-generation sequencing.

For a biologist doing RNA-Seq, what are telltale signs that better normalization of the data is needed?

JS: Not seeing any signal when they expected to see signal is one sign. In other words, they're missing some positive controls. Another sign might be seeing much more signal than they expected to see, say 80% of the genome is differentially expressed. The biggest, the easiest way—the way that I discovered the importance of normalization in the microarray context—is the lack of reproducibility across different studies. You can have three studies that are all designed to study the same thing, and you just see basically no reproducibility, in terms of differentially expressed genes. And every time I encountered that, it could always be traced back to the normalization. So, I'd say that the biggest sign and the biggest reason why you want to use normalization is to have a clean signal that's reproducible.

That said, it's very hard to define what one means by 'correct normalization'. It's very difficult, in a single study, to know the normalization has been performed correctly. You have to look at a whole body of studies before you can really start getting your head wrapped around it, unless that study was carried out specifically to develop normalization methods.

What about statistical methods for data from 'third-generation' sequencers?

JS: From my understanding, third-generation sequencing will give us longer reads and more accurate sequence information. This will mitigate sequence biases, read mapping biases and other technical biases of this nature. However, it does not appear that third-generation sequencing will necessarily increase the total number of reads. The fact that most genes are currently measured by just tens of reads in RNA-seq experiments is a major drawback of the next-gen sequencing technology. I would like to have more than 20 or 40 counts per gene when capturing RNA abundance variation. Microarray measurements are based on hybridization events several orders of magnitude higher. I hope that sooner rather than later, the total number of reads available in a study will increase by a couple orders of magnitude. This will of course put an even greater strain on computational costs, but from a statistics perspective it will be an improvement.

What papers on statistics would you recommend?

JS: A good place to start is in reading the key papers from the microarray literature that have been most influential over the past decade, particularly those on experimental design. Next-gen sequencing is not immune to technical biases, poor study design, batch effects, etc. For normalization, I recommend that one achieves a firm understanding of the difference between unsupervised and supervised normalization. The fundamental statistical difference in next-gen sequencing experiments is that they produce count data, as opposed to the more continuous data coming from arrays. I recommend any well-written paper that introduces basic methods for count data and covers Poisson models, Poisson regression and overdispersion (**Box 1**).

Final thoughts?

JS: Overall, I believe that the problem of normalization for sequencing data is wide open right now. Also, there needs to be more dialog about transparency in being able to fully reproduce statistical analyses all the way from the raw data to the final results. This will be extremely crucial for next-gen sequencing as there are many preprocessing and inference steps occurring along the way from raw data to tables and figures. But biology and statistics have a long history of working well together, and there are many of us working at the interface, helping to bridge the two fields even more.

H. Craig Mak is Associate Editor, Nature Biotechnology

1. Storey, J.D. & Tibshirani, R. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
2. Leek, J.T. & Storey, J.D. *PLoS Genet.* **3**, 1724–1735 (2007).