

to a wider range of journals. Many biotech companies, at some point in their growth cycle, will strike a deal with a big pharma partner to help develop one or more molecules in their pipeline. These deals can be structured in a variety of ways and usually include milestone payments for the biotech partner. The smaller biotechs should ask for access to their pharmaceutical company partner's online journal subscriptions as a component of these deals. The cost to add access for 20 scientists, for example, to in-house research staffs numbering in the thousands should be minimal. An obvious added benefit of this arrangement is that the online access should facilitate the ability of the smaller biotech partner to move the joint projects as well as their own projects forward.

What happens, though, to scientists working for unpartnered companies, or with partners that can't or won't provide such access? Another solution would be for small biotechs to form some type of cooperative or consortium that would provide biomedical journal access. The consortium could be modeled on the cable TV business, which sells a range of tiered services to different customers. The consortium would purchase online subscriptions to a core group of journals and members of the group would pay a set fee per number of employees in their organization to gain access. As with the cable business, those wanting to have access to additional tiers of journals (e.g., those with a focus on clinical studies, oncology or veterinary science) would pay additional fees. The charge for the basic access tier needs to be widely affordable, and the arrangement would benefit both the journal publishers and the researchers. The journals that participated in providing access to consortium members would gain additional (albeit reduced) revenues from organizations that otherwise could not afford their subscriptions. On the other side of the equation, scientists working at the consortium companies would gain access to journals that they otherwise could not afford. This should help facilitate innovation and stimulate both basic as well as translational research. It would be preferable if access to the older scientific literature (say, greater than two years old) were provided for free, and for tiered subscriptions to newly published articles to be the focus of what consortium members are paying for.

The biotech industry's advocacy group, the Biotechnology Industry Organization (BIO; Washington, DC), would be a logical candidate to create and manage such a consortium within the auspices of their own organization. Providing this service would

have benefits in three specific areas. It would help motivate small companies to join BIO, thereby strengthening the organization and its ability to assist its members. Companies who joined the consortium would bolster the ability of their scientists to do cutting edge R&D work in creating new drugs. Journal publishers would increase their revenues by drawing in a much larger pool of subscribers, which is why they could afford to give them a price break. As an alternative to BIO, a pioneering information services company might garner a strong financial reward for creating and operating an online journal access consortium.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

Stewart Lyman

Lyman BioPharma Consulting LLC, Seattle, Washington, USA.

e-mail: lymanbiopharma@comcast.net

1. Lyman, S. *Xconomy/Seattle* (18 January 2011) <http://www.xconomy.com/seattle/2011/01/18/what-does-biotech-really-suffer-from-information-overload-or-underload/?single_page=true>
2. Kling, J. *Nat. Biotechnol.* **29**, 197–200 (2011).
3. Samaha, L. *Investing by Suite 101* (9 July 2010) <<http://www.suite101.com/content/estimating-drug-probabilities-of-success-in-fda-clinical-trials-a259518>>

Sequencing technology does not eliminate biological variability

To the Editor:

RNA sequencing technology provides various advantages over DNA microarrays. For example, it is possible to measure alternative transcription¹ or measure transcription for noncoding regions² *de novo*. Another potential advantage is low technical variation^{2–4}. This has led to rapid adoption of the technology and a recent surge of publications⁵. We would like to caution, however, that the euphoria surrounding the technology has led many of these publications to discount the influence of biological variability, forgetting perhaps that unwanted variability in gene expression measurements is not due only to measurement error. Gene expression is a stochastic process⁶ and is known to vary between units considered to be of the same population, for example, in samples from a specific healthy tissue across individuals⁷. In a typical experiment, variation in gene expression measurements [Var(Expr)] can be decomposed⁸ as the following:

$$\text{Var(Expr)} = \text{Across Group Variability} + \text{Measurement Error} + \text{Biological Variability}$$

'Across Group Variability' is the variation in gene expression due to the groups under consideration in an experiment. For example, it is well known that gene expression profiles for tumor samples differ from expression profiles for matched healthy controls⁹. This type of variability can be measured by comparing samples from different biological groups and is typically the outcome of interest. The second component of gene expression variation, 'Measurement Error', can be estimated with technical replicates—different aliquots of the same sample measured with a technology multiple times. This is the

type of variation that may be reduced with technology improvements⁴. Well-known sources of technical variability in both sequencing and microarray studies are laboratory^{10,11} and batch¹² effects. The third component of expression variation is true 'Biological Variability', which can be measured only by considering expression measurements taken from multiple biological samples within the same group. Regardless of the technology used to measure expression levels, the true gene expression levels will vary among individuals because expression is inherently a stochastic process⁶. In an experiment where the group comparison is of primary interest, both measurement error and biological variation may be confused with the outcome of interest: the estimated difference in expression between groups.

To illustrate how biological variability among individuals within the same group is not eliminated by sequencing technology, we collected public data from two of the only RNA-sequencing experiments with a large number of biological replicates, $n = 60$ and $n = 69$, respectively^{13,14}. We compared a subset of these sequencing data ($n = 43$ and 51 samples, respectively) with microarray data from two different platforms^{15,16}. In each comparison, the exact same cell lines were analyzed on both technologies. In study one, $m = 14,797$ genes had expression measurements from both sequencing and microarrays on all samples. In study two, $m = 7,157$ genes had expression measurements from both technologies on all samples (**Supplementary Methods**).

For each expressed gene in each of the two studies, we calculated an estimate of the variability in expression levels across

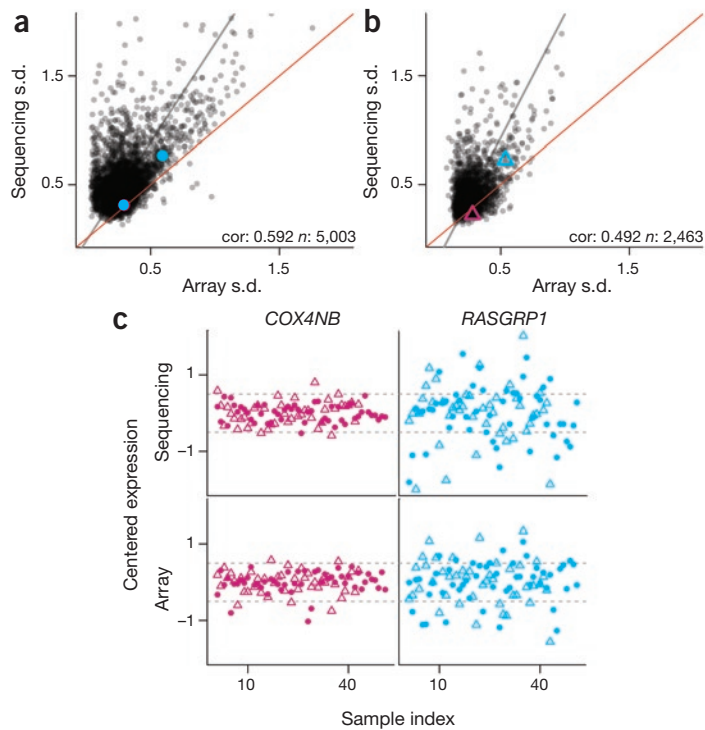


Figure 1 Biological variability measured with sequencing and microarrays. (a) A plot of the s.d. of expression values as measured with microarrays in the Stranger *et al.* study¹⁵ (x axis) and sequencing in the Montgomery *et al.* study¹³ (y axis). The estimates of expression variability from sequencing are similar to the estimates from microarrays. (b) A plot of the s.d. of expression values as measured with microarrays in the Choy *et al.* study¹⁶ (x axis) and the Pickrell *et al.* study¹⁴ (y axis). The estimates of expression variability from sequencing are again almost the same as estimates from microarrays. In each plot, the black line is the best linear fit and the red line is the line $y = x$. (c) A plot of the expression for two genes *COX4NB* (left column, pink) and *RASGRP1* (right column, blue) as measured with sequencing (top row) and microarrays (bottom row) versus biological sample. Mean-centered measurements from the two studies are plotted as circles and triangles, respectively. The s.d. for the two genes are highlighted in a,b. The plot shows that regardless of the measurement technology or study, *COX4NB* expression is much less variable than *RASGRP1* expression.

individuals as measured with microarrays and sequencing (Supplementary Methods). We found that variability in expression for each gene was similar in microarray and sequencing technologies (Fig. 1a,b). The same trend existed for different choices of variability measures (Supplementary Fig. 1a,b) and for different methods of calculating expression from sequencing (Supplementary Fig. 1c,d). We also found that transcripts showed substantial differences in biological variability. For example, *COX4NB* was not strongly variable in either population, whereas *RASGRP1* was highly variable for both populations, again regardless of technology (Fig. 1c). The technical variability for both genes was substantially smaller than the total variability (Supplementary Fig. 2a). These results are consistent with biological variability being a property of gene expression itself, rather than the technology used to measure expression. To confirm

this result, we estimated the proportion of the total variability for each gene that is attributable to biology by applying a mixed effects model to data from the sequencing (11 samples) and DNA microarray (14 samples) experiments for which we had two technical replicates. In general, most of the observed variation was biological, rather than technical (Supplementary Fig. 2b).

Biological variability has important implications for the design, analysis and interpretation of RNA-sequencing experiments. For example, a large observed difference in expression of *COX4NB* between two groups is likely important because the expression of this gene varies little across individuals. Meanwhile, that same difference in expression for *RASGRP1* may be meaningless because the expression for that gene is highly variable. If only a few biological replicates are available, it will be impossible to estimate the level of biological variability in expression for

each gene in a study. Supplementary Table 1 summarizes a large number of published RNA-sequencing studies over the past three years. In every case, except for the two studies we analyzed here, conclusions were based on a small number ($n \leq 2$) of biological replicates. One goal of RNA-sequencing studies may be simply to identify and catalog expression of new or alternative transcripts. However, all of these studies make broader biological statements on the basis of a very small set of biological replicates.

Our analysis has two important implications for studies performed with a small number of biological replicates. First, significant results in these studies may be due to biological variation and may not be reproducible; and second, it is impossible to know whether expression patterns are specific to the individuals in the study or are a characteristic of the study populations. These ideas are now widely accepted for DNA microarray experiments, where a large number of biological replicates are now required to justify scientific conclusions. Our analysis suggests that as biological variability is a fundamental characteristic of gene expression, sequencing experiments should be subject to similar requirements.

Note: Supplementary information is available on the Nature Biotechnology website.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Kasper D Hansen¹, Zhijin Wu²,
Rafael A Irizarry¹ & Jeffrey T Leek¹

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. ²Department of Community Health, Section of Biostatistics, Brown University, Providence, Rhode Island, USA.
e-mail: rafa@jhu.edu or jleek@jhsph.edu

1. Wang, E.T. *et al.* *Nature* **456**, 470–476 (2008).
2. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
3. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. *BMC Bioinformatics* **11**, 94 (2010).
4. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. *Genome Res.* **18**, 1509–1517 (2008).
5. Wang, Z., Gerstein, M. & Snyder, M. *Nat. Rev. Genet.* **10**, 57–63 (2009).
6. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. *Science* **297**, 1183–1186 (2002).
7. Whitney, A.R. *et al.* *Proc. Natl. Acad. Sci. USA* **100**, 1896–1901 (2003).
8. Churchill, G.A. *Nat. Genet.* **32** Suppl, 490–495 (2002).
9. Golub, T.R. *et al.* *Science* **286**, 531–537 (1999).
10. Irizarry, R.A. *et al.* *Nat. Methods* **2**, 345–350 (2005).
11. Shi, L. *et al.* *Nat. Biotechnol.* **24**, 1151–1161 (2006).
12. Leek, J.T. *et al.* *Nat. Rev. Genet.* **11**, 733–739 (2010).
13. Montgomery, S.B. *et al.* *Nature* **464**, 773–777 (2010).
14. Pickrell, J.K. *et al.* *Nature* **464**, 768–772 (2010).
15. Stranger, B.E. *et al.* *Science* **315**, 848–853 (2007).
16. Choy, E. *et al.* *PLoS Genet.* **4**, e1000287 (2008).