HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2012 January 11.

Published in final edited form as:

Nat Biotechnol. ; 29(7): 572–573. doi:10.1038/nbt.1910.

Sequencing technology does not eliminate biological variability

Kasper D. Hansen[1], **Zhijin Wu**[2], **Rafael A. Irizarry**[1,*], and **Jeffrey T. Leek**[1,*]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[2] Department of Community Helath, Section of Biostatistics, Brown University, Providence, RI 02912, USAof

Abstract

RNA sequencing has generated much excitement for the advantages offered over microarrays. This excitement has led to a barrage of publications discounting the importance of biological variability; as microarray publications did in the 1990s. By comparing microarray and sequencing data, we demonstrate that expression measurements exhibit biological variability across individuals irrespective of measurement technology. Our analysis suggests RNA-sequencing experiments designed to estimate biological variability are more likely to produce reproducible results.

RNA sequencing (RNAseq) technology provides various advantages over microarrays. For example, it is possible to measure alternative transcription[1] or measure transcription for non-coding regions[2]*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 publications[5]. However, the euphoria 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 process[6] and is known to vary between units considered to be of the same population - for example in samples from a specific healthy tissue across individuals[7]. In a typical experiment, variation in gene expression measurements can be decomposed[8] as:

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

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 controls[9]. 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

* To whom correspondence should be addressed rafa@jhu.edu, jleek@jhsph.edu.

may be reduced with technology improvements[4]. Well-known sources of technical variability in both sequencing and microarray studies are laboratory[10, 11] and batch[12] effects. The third component of expression variation is true *biological variability*, which can only be measured 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 process[6]. 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 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 while *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 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, since the expression of this gene varies little across individuals. Meanwhile, that same difference in expression for *RASGRP1* may be meaningless, since 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 last three years. In every case, except for the

two studies we analyzed here, conclusions were based on a small number ($n \leq 2$) 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: (1) significant results in these studies may be due to biological variation and may not be reproducible and (2) 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 microarray experiments, where a large number of biological replicates are now required to justify scientific conclusions. Our analysis suggests that since biological variability is a fundamental characteristic of gene expression, sequencing experiments should be subject to similar requirements.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. [PubMed: 18978772]
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
3. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. [PubMed: 20167110]
4. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18:1509–1517. [PubMed: 18550803]
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]
6. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002; 297:1183–1186. [PubMed: 12183631]
7. Whitney AR, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A. 2003; 100:1896–1901. [PubMed: 12578971]
8. Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet. 2002; 32(Suppl):490–495. [PubMed: 12454643]
9. Golub TR, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–537. [PubMed: 10521349]
10. Irizarry RA, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods. 2005; 2:345–350. [PubMed: 15846361]
11. Shi L, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006; 24:1151–1161. [PubMed: 16964229]
12. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11:733–739. [PubMed: 20838408]
13. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
14. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]

15. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–853. [PubMed: 17289997]

16. Choy E, et al. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet. 2008; 4:e1000287. [PubMed: 19043577]

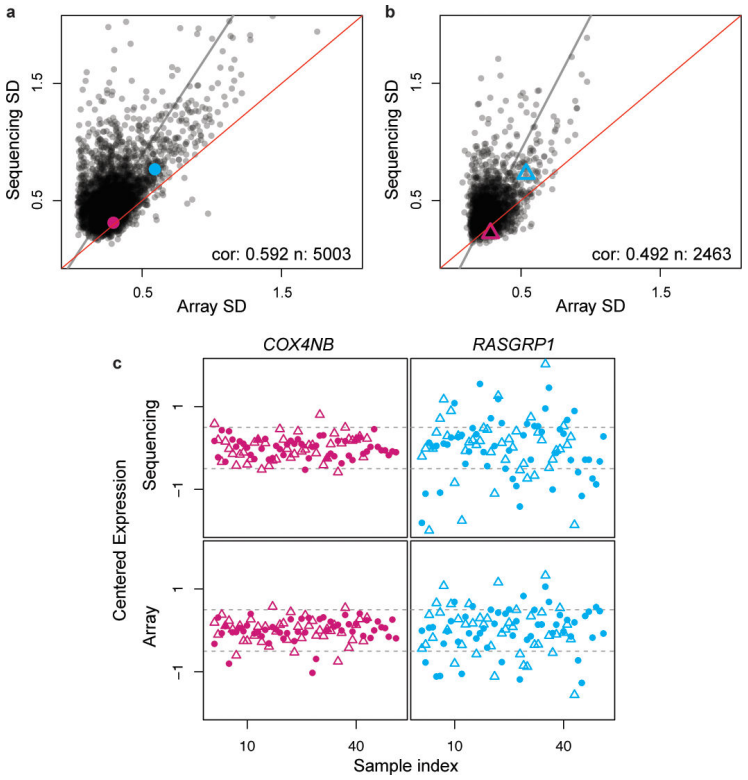

Figure 1.

Biological variability measured with sequencing and microarrays. **(a)** A plot of the standard deviation of expression values as measured with microarrays in the Stranger *et al*. study[15] (x-axis) and sequencing in the Montgomery *et al*. study [13] (y-axis). The estimates of expression variability from sequencing are similar to the estimates from microarrays. **(b)** A plot of the standard deviation of expression values as measured with microarrays in the Choy *et al*. study[16] (x-axis) and the Pickrell *et al*. study [14] (y-axis). The estimates of expression variability from sequencing are again almost the same as estimates from microarrays. **(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 standard deviations 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.