

Published in final edited form as:

Cell. 2012 October 26; 151(3): 476–482. doi:10.1016/j.cell.2012.10.012.

## Revisiting Global Gene Expression Analysis

Jakob Lovén<sup>1,\*</sup>, David A. Orlando<sup>1,\*</sup>, Alla A. Sigova<sup>1</sup>, Charles Y. Lin<sup>1,2</sup>, Peter B. Rahl<sup>1</sup>, Christopher B. Burge<sup>3</sup>, David L. Levens<sup>4</sup>, Tong Ihn Lee<sup>1,†</sup>, and Richard A. Young<sup>1,3,†</sup>

<sup>1</sup>Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142

<sup>2</sup>Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139

<sup>3</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139

<sup>4</sup>Gene Regulation Section, Laboratory of Pathology, National Cancer Institute, Center for Cancer Research, Bethesda, MD, 20892

### Summary

Gene expression analysis is a widely used and powerful method for investigating the transcriptional behavior of biological systems, for classifying cell states in disease and for many other purposes. Recent studies indicate that common assumptions currently embedded in experimental and analytical practices can lead to misinterpretation of global gene expression data. We discuss these assumptions and describe solutions that should minimize erroneous interpretation of gene expression data from multiple analysis platforms.

### Global Gene Expression Analysis

Global gene expression analysis provides quantitative information about the population of RNA species in cells and tissues. It is an exceptionally powerful tool of molecular biology that is used to explore basic biology, diagnose disease, facilitate drug discovery and development, tailor therapeutics to specific pathologies and generate databases with information about living processes. Consequently, expression analysis is among the most commonly used methods in modern biology; there are over 750,000 expression datasets in the NCBI Gene Expression Omnibus (GEO) public database (Edgar et al., 2002).

Global gene expression analysis uses DNA microarrays, RNA-Seq, and other methods to measure the levels of RNA species in biological systems (Geiss et al., 2008; Heller, 2002; Lockhart and Winzler, 2000; Ozsolak and Milos, 2011; Schena et al., 1998; Wang et al., 2009). DNA microarrays, which have been most frequently used for expression analysis, consist of millions of individual oligonucleotide probes fixed to a solid surface. The oligonucleotide probes typically have sequences representative of known RNA species and are generally used to quantitate the relative levels of RNA species that hybridize to the

© 2012 Elsevier Inc. All rights reserved.

Corresponding Authors: Tong Ihn Lee, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, Tel: (617) 258-7181, Fax: (617) 258-0376, tlee@wi.mit.edu. Richard A. Young, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, Tel: (617) 258-5218, Fax: (617) 258-0376, young@wi.mit.edu.

\*These authors contributed equally

†These corresponding authors contributed equally

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

probes. Massively parallel sequencing technologies, developed more recently, provide a measure of the frequency of RNA species through sequencing of RNA-derived cDNA populations. Other approaches, such as digital molecular barcoding, represent a fusion of the hybridization and counting approaches. For instance, the nCounter digital quantification platform relies on hybridization of labeled probes to RNA molecules and single molecule imaging to provide a measurement of the frequency of particular RNA species.

## Assumptions and Interpretation

Almost all global expression analysis involves isolation of RNA from two or more cellular sources, introducing similar amounts of RNA from the sources into the experimental platform and analyzing the data using algorithms that normalize the signal from the samples (Kulkarni, 2011; Mortazavi et al., 2008; Quackenbush, 2002; Schulze and Downward, 2001). If the cellular sources produce equivalent amounts of RNA/cell, and the yields of RNA and its derivatives are equivalent throughout experimental manipulation, then normalized expression data should produce an accurate representation of the relative levels of each gene product.

We recently found that cells with high levels of c-Myc can amplify their gene expression program, producing 2–3 times more total RNA and generating cells that are larger than their low-Myc counterparts (Lin et al., 2012; Nie et al., 2012). This discovery has led us to question the common assumption that cells produce similar levels of RNA/cell and the general practice of introducing similar amounts of total RNA into analysis platforms without including standardized controls that would reveal transcriptional amplification or repression. As described below, it is likely that this assumption and practice has led to erroneous interpretations. We describe here an experimental approach to genome-wide analysis of RNA expression that is more likely to produce accurate assessments of changes in steady-state levels of RNA.

Consider two different models for changes in gene expression (Figure 1). In the first, RNA levels for a minority of genes are elevated, but the levels of total RNA in the two cells is similar (Figure 1A). The absolute levels of most RNA species are therefore similar in the two cells, and when the total signal for the RNA population is normalized by standard algorithms, the resulting expression data appropriately indicates an increase in the relative RNA levels for a set of genes (Figure 1B). In the second model, the two cells express a similar set of genes, but one cell produces and accumulates 2–3 times more RNA/gene for many of the same genes expressed in the other cell (Figure 1C), an effect that has been termed transcriptional amplification (Lin et al., 2012; Nie et al., 2012). In the conventional approach to expression analysis, similar amounts of RNA from the two cells are introduced into the assay, thus masking the fact that one of the cells has 2–3 times more RNA than the other (Figure 1D). This potential source of error is typically overlooked because of the commonly believed, though rarely stated, assumption that the absolute amount of total mRNA in each cell is similar across different cell types or experimental perturbations. Furthermore, the most commonly used analysis methods are primarily intended to account for technical variations in signal to noise and assume that the signals for different samples from different experiments should be scaled to have the same median or average value, or that the distributions of signal intensities for each experiment within a set should all be the same (Bolstad et al., 2003; Huber et al., 2002; Irizarry et al., 2003; Kalocsai and Shams, 2001; Li and Wong, 2001; Reimers, 2010; Wu et al., 2004). Normalization of signal from cells that experience transcriptional amplification can thus have the net result of equalizing values that are actually different and producing the erroneous perception that some genes have elevated expression while a similar number of genes have reduced expression.

## Experimental Approach

To produce a reliable gene expression analysis protocol that addresses this experimental and data normalization issue, we investigated the use of spiked-in standards (Benes and Muckenthaler, 2003; Hartemink et al., 2001; Hill et al., 2001; Jiang et al., 2011; Mortazavi et al., 2008). We implemented an approach that uses spiked-in RNA standards to allow normalization to cell number and permit correction for differences in yields during experimental manipulation (Figure 2A). We performed genome-wide analysis on P493-6 cells expressing low or high levels of c-Myc (Pajic et al., 2000; Schuhmacher et al., 1999) where cells with high levels of the transcription factor were found to produce 2-3 -fold higher levels of the same RNA species found in cells with low levels (Lin et al., 2012). Cell number was determined by counting cells using C-Chip disposable hemocytometers (Digital Bio, Hopkinton, MA) and equivalent numbers of high- and low-Myc cells were harvested. The DNA content of the two samples was measured and found to be equivalent. Following total RNA extraction, spiked-in RNA standards were added in proportion to the number of cells present in the sample. Samples were then split and prepared for microarray, RNA-seq and digital analysis using NanoString.

DNA-microarrays were first used to compare the high-Myc versus low-Myc cell RNA populations (Figure 2B; Table S1). Similar amounts of RNA from the low- and high-Myc cells were introduced into the Affymetrix DNA microarray assay following the manufacturer's protocol, which is the most common approach used in expression analysis. The resulting data were processed by using standard normalization methods and by using the spike-in standards for normalization. The results obtained using standard approaches can be interpreted to mean that the expression levels of some genes is unchanged, while others increase or decrease (Figure 2B). The interpretation is quite different when the same data is normalized using spike-in standards that reflect cell number: over 90% of the genes show increases in expression in high-Myc cells relative to low-Myc cells (Figure 2B).

RNA-Seq analysis was then used to compare the high-Myc versus low-Myc cell RNA populations (Figure 2C; Table S2). Similar amounts of RNA from the low- and high-Myc cells were subjected to sequencing. The resulting data were processed by using standard normalization methods and by using the spike-in standards for normalization. Again, the results obtained using standard approaches suggest that the expression levels of some genes is unchanged, while others increase or decrease (Figure 2C), yet when the same data is normalized using spike-in standards that reflect cell number, there is an increase in transcript levels for the vast majority of genes (Figure 2C).

We then used whole-sample, digital gene expression quantification (NanoString, Seattle, WA) to compare transcript levels in the high-Myc and low-Myc cells. In one experiment, equal amounts of RNA from the high- and low-Myc cells were compared using this method. The results of this analysis suggest that the expression levels of some genes is unchanged, while others increase or decrease. In a second experiment, equal numbers of high- and low-Myc cells were used to prepare RNA and these total RNA populations were subjected to digital gene expression quantification. Here the data indicate there is an increase in transcript levels for the vast majority of genes in high- vs. low-Myc cells (Figure 2D; Table S3).

In summary, three of the major technologies typically used for global gene expression analysis - microarray, RNA-sequencing and digital quantification - detect a widespread increase in transcripts/cell in cells that experience transcriptional amplification by c-Myc. Significantly, all three of these major technologies used for gene expression fail to detect the widespread increase of transcription when inappropriate normalization methods are used.

Instead, they erroneously suggest the interpretation that a similar number of genes show increases and decreases in expression.

## Implications

Our results indicate that spike-in controls of the type described here are a robust, cross-platform method to allow normalization to cell number and thus enable more accurate detection of differential gene expression and changes in gene expression programs. The clear implication is that the use of spike-in controls normalized to cell number should become the default standard for all expression experiments, as opposed to their more limited use in experiments where gross changes in RNA levels are already anticipated, as exemplified by transcription shutdown experiments (Bar-Joseph et al., 2012). When cell counting may be problematic, as for expression experiments from solid tumors or tissues, DNA content may be used as a surrogate if ploidy and DNA replication profiles are also characterized to prevent the introduction of a DNA content-based artifact.

The discovery of transcriptional amplification and the realization that common experimental methods may lead to erroneous interpretation of gene expression experiments has implications for much current biological research. How prevalent is misinterpretation of genome-wide expression data due to the assumption that cells produce similar levels of total RNA? The answer is likely related to the prevalence of regulatory mechanisms that globally amplify or suppress transcription. What are the implications for classifying cell states in disease? Significant effort is being devoted to expression profiling cancer cells and these studies use standard normalization methods (Alizadeh et al., 2000; Beer et al., 2002; Berger et al., 2010; Bhattacharjee et al., 2001; Bittner et al., 2000; Golub et al., 1999; Lapointe et al., 2004; Northcott et al., 2012; Ramaswamy et al., 2001; Ross et al., 2000; Schmitz et al., 2012; Shipp et al., 2002; Su et al., 2001; The Cancer Genome Atlas TCGA, 2012; van't Veer et al., 2002; van de Vijver et al., 2002; Yeoh et al., 2002). Because c-Myc expression occurs at widely varying levels in various tumor cells, transcriptional amplification is likely having a profound impact on cancer cell signatures. Where expression data is being used to gain insights into cancer cell behavior and regulation, it should be interpreted with added caution.

## Experimental Procedures

### Cell Culture

P493-6 cells were kindly provided by Chi Van Dang, University of Pennsylvania. Cells were propagated in RPMI-1640 supplemented with 10% fetal bovine serum and 1% GlutaMAX (Invitrogen, 35050-061). The conditional *myc*-tet construct in P493-6 cells was repressed with 0.1 µg/mL tetracycline (Sigma, T7660) for 72 hours. Cells were then washed three times with RPMI-1640 medium containing 10% tetracycline system approved FBS (Clontech, 631105) and 1% GlutaMAX and re-cultured in tetracycline-free culture conditions. All experiments were performed in the absence of EBNA2 activation. Cell numbers were determined by manually counting cells using C-Chip disposable hemocytometers (Digital Bio, DHC-N01) prior to lysis and RNA extraction.

### RNA Extraction and Synthetic RNA Spike-In

Ten million P493-6 cells were homogenized in 1 mL of TRIzol Reagent (Life Technologies, 15596-026), purified using the mirVANA miRNA isolation kit (Ambion, AM1560) following the manufacturer's instructions and re-suspended in 100 µL nuclease-free water (Ambion, AM9938). Total RNA was spiked-in with ERCC ExFold RNA spike-in controls, treated with DNA-free™ DNase I (Ambion, AM1906) and analyzed on Agilent 2100 Bioanalyzer for integrity. The external control spike-ins used in the microarray and RNA-

Seq analysis were obtained from the ERCC ExFold RNA Spike-In kit (Ambion, 4456739). The ERCC RNA Spike-In Control Mixes used here comprise a set of 92 polyadenylated transcripts that mimic natural eukaryotic mRNAs. The RNAs range in size from 250 – 2000 nucleotides in length and span an approximately  $10^6$ -fold concentration range.

After extracting total RNA from equal numbers of cells, a fixed amount of diluted ERCC Spike-In Mix #1 was added. The amount of spike-in added was calibrated to the RNA yield of the high-Myc cells to ensure the spike-in signal was in the appropriate dynamic range (ERCC User Guide, Table 4). For these experiments, 1  $\mu$ l of a 1:10 dilution of Mix #1 was added to total RNA extracted from  $1 \times 10^6$  cells. RNA with the RNA Integrity Number (RIN) above 9.8 was used for library generation for RNA-Seq or hybridized to GeneChip® PrimeView Human Gene Expression Arrays (Agilent), using 10  $\mu$ g or 100ng of total RNA, respectively.

Spike-in controls can also be added prior to RNA purification, if desired, but ideally, the controls used in that case should have additional features of mRNAs, such as 5' caps, that would limit the potential for degradation during lysis.

### Microarray Sample Preparation and Analysis

For microarray analysis, 100ng of total RNA containing ERCC ExFold Mix #1 RNA spike-in controls (see above) was used to prepare biotinylated aRNA (cRNA) according to the manufacturer's protocol (3' IVT Express Kit, Affymetrix 901228). GeneChip arrays (Primeview, Affymetrix 901837) were hybridized and scanned according to standard Affymetrix protocols. All samples were processed in technical duplicate. Images were extracted with Affymetrix GeneChip Command Console (AGCC), and analyzed using GeneChip Expression Console. A Primeview CDF that included probe information for the ERCC controls, provided by Affymetrix, was used to generate CEL files. We processed the CEL files using standard tools available within the *affy* package in R. The CEL files were processed with the *expresso* command to convert the raw probe intensities to probeset expression values. The parameters of the *expresso* command were set to generate Affymetrix MAS5-normalized probeset values. We used a loess regression to re-normalize these MAS5 normalized probeset values, using only the spike-in probesets to fit the loess. The *affy* package provides a function, *loess.normalize*, which will perform loess regression on a matrix of values (defined using the parameter *mat*) and allows for the user to specify which subset of data to use when fitting the loess (defined using the parameter *subset*, see the *affy* package documentation for further details). For this application the parameters *mat* and *subset* were set as the MAS5-normalized values and the row-indices of the ERCC control probesets, respectively. The default settings for all other parameters were used. The result of this was a matrix of expression values normalized to the control ERCC probes. The probeset values from the duplicates were averaged together and the log2 fold change from the low-Myc to the high-Myc samples are shown.

### RNA-Seq Sample Preparation and Analysis

Using 10  $\mu$ g of total RNA containing ERCC ExFold Mix #1 RNA spike-in controls (see above), sequencing libraries were prepared according to the following protocol. Polyadenylated RNA was purified by two rounds of selection using Dynabeads® mRNA Purification Kit for mRNA Purification from total RNA (Life Technologies, 610-06) following the manufacturer instructions. This resulting RNA was then further processed for RNA-Seq assays. Briefly, polyadenylated RNA was fragmented with divalent cations under elevated temperature. First strand cDNA synthesis was performed with random hexamers and Superscript III reverse transcriptase (Life Technologies, 18080-051). Second strand cDNA synthesis was performed using RNase H and DNA Polymerase I. In the second-



strand synthesis reaction, dTTP was replaced with dUTP. Following cDNA synthesis, the double stranded products were end repaired, a single “A” base was added, and Illumina PE adaptors were ligated onto the cDNA products. The ligation products with an average size of 300 bp were purified using agarose gel electrophoresis. Following gel purification, the strand of cDNA containing dUTP was selectively destroyed during incubation of purified double-stranded DNA with HK-UNG (Epicentre, HU59100). The adapter ligated single-stranded cDNA was then amplified with 15 cycles of PCR and PCR products were purified using gel electrophoresis. These RNA-Seq libraries were subsequently sequenced on Illumina HiSeq 2000. Sequences were aligned using Bowtie (version 0.12.2) to build version NCBI36/HG18 of the human genome where the sequences of the ERCC synthetic spike-in RNAs (<http://tools.invitrogen.com/downloads/ERCC92.fa>) had been added. The RPKM (Reads Per Kilobase of exon per Million) was then computed for each gene and synthetic spike-in RNA. We used a loess regression to re-normalize the RPKM values, using only the spike-in values to fit the loess. The *affy* package in R provides a function, *loess.normalize*, which will perform loess regression on a matrix of values (defined using the parameter *mat*) and allows for the user to specify which subset of data to use when fitting the loess (defined using the parameter *subset*, see the *affy* package documentation for further details). For this application the parameters *mat* and *subset* were set as a matrix of all RPKM values and the row-indices of the ERCC spike-ins, respectively. The default settings for all other parameters were used. The result of this was a matrix of RPKM values normalized to the control ERCC spike-ins. The log<sub>2</sub> fold change from the low-Myc to the high-Myc samples were then calculated and shown as a heatmap.

### NanoString nCounter Gene Expression Assay Sample Preparation and Analysis

For digital gene expression using NanoString nCounter Gene Expression CodeSets,  $1 \times 10^6$  cells were collected and lysed directly either in 100µL RLT buffer (Qiagen, 74104) to yield a concentration of 10,000 cells per µL or in 500µL Lysis Buffer using the mirVANA miRNA isolation kit (Ambion, AM1560). Samples were processed according to the Cell Lysate Protocol (nCounter Gene Expression Protocol, NanoString) or the Total RNA Extraction Protocol (Ambion). Four µL of cell lysate (for Cell-Count normalization) or 100ng of total RNA (for Total RNA normalization) was subsequently incubated overnight at 65°C in nCounter Reporter CodeSet, Capture ProbeSet and hybridization buffer. Following hybridization, samples were immediately processed with the nCounter PrepStation and subsequently analyzed on an nCounter Digital Analyzer. All samples were processed in biological duplicate.

We used two custom nCounter Reporter CodeSets encompassing 429 genes. These codsets encompassed sets of known cancer related genes (CodeSet CS-1, CS-2) (Delmore et al., 2011). For each NanoString dataset, we used a piecewise linear interpolation of control RNAs (added after hybridization as part of the nCounter PrepStation protocol) to their known concentrations to normalize each dataset. The normalized signals for the biological duplicates were averaged together. 266 genes showing active expression (a normalized value of 1.0 or greater) in both the average low-Myc Total-RNA, and average low-Myc Cell-Count samples were selected, and the log<sub>2</sub> fold ratio between the low-Myc and high-Myc samples were calculated and shown as a heatmap.

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Tom Volkert, Jeong-Ah Kwen, Jennifer Love, Sumeet Gupta, at the Whitehead Genome Technologies Core for Solexa sequencing and microarray processing and Ziv Bar-Joseph for critical comments. This work was supported by National Institutes of Health grants HG002668 (RAY) and CA146445 (RAY, TL), an American Cancer Society Postdoctoral Fellowship PF-11-042-01-DMC (PBR) and a Swedish Research Council Postdoctoral Fellowship VR-B0086301 (JL).

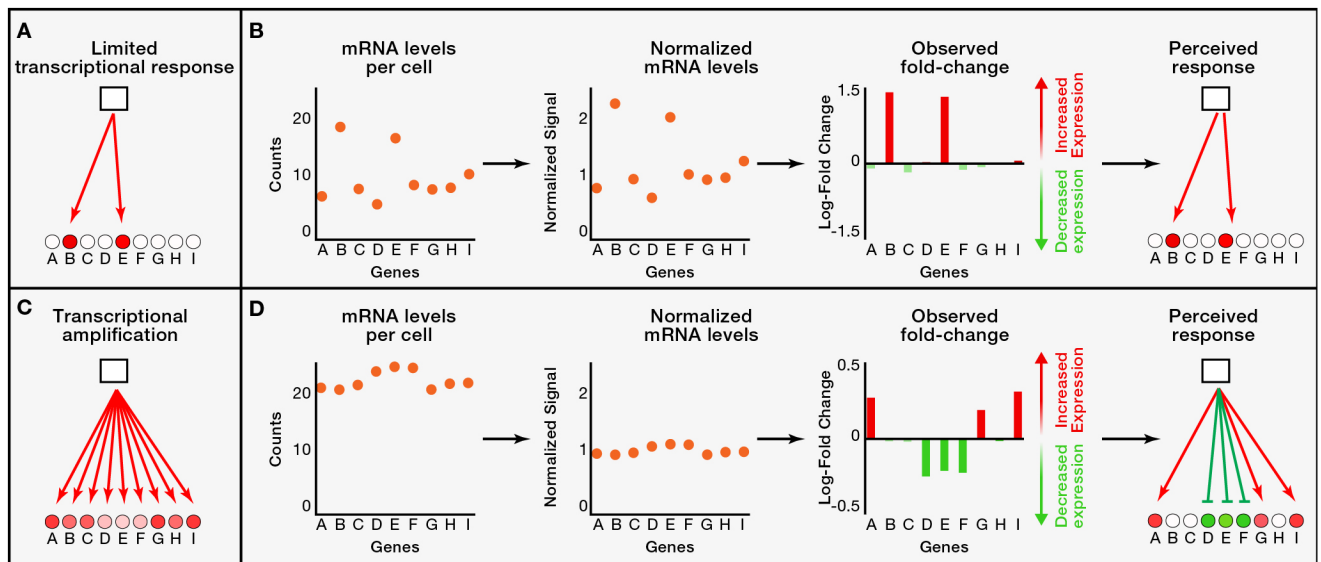
## References

- Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403:503–511. [PubMed: 10676951]
- Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet*. 2012; 13:552–564. [PubMed: 22805708]
- Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*. 2002; 8:816–824. [PubMed: 12118244]
- Benes V, Muckenthaler M. Standardization of protocols in cDNA microarray analysis. *Trends Biochem Sci*. 2003; 5:244–249.
- Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, et al. Integrative analysis of the melanoma transcriptome. *Genome Res*. 2010; 20:413–427. [PubMed: 20179022]
- Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*. 2001; 98:13790–13795. [PubMed: 11707567]
- Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*. 2000; 406:536–540. [PubMed: 10952317]
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19:185–193. [PubMed: 12538238]
- Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi JW, Jacobs HM, Kastiris E, Gilpatrick T, Paranal RM, Qi J, et al. BET Bromodomain Inhibition as a Therapeutic Strategy to Target c-Myc. *Cell*. 2011; 146:903–916.
- Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30:207–210. [PubMed: 11752295]
- Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol*. 2008; 26:317–325. [PubMed: 18278033]
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286:531–537. [PubMed: 10521349]
- Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Maximum likelihood estimation of optimal scaling factors for expression array normalization. *P Soc Photo-Opt Ins*. 2001; 2:132–140.
- Heller MJ. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*. 2002; 4:129–153. [PubMed: 12117754]
- Hill AA, Brown EL, Whitley MZ, Tucker-Kellogg G, Hunter CP, Slonim DK. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol*. 2001; 2.
- Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*. 2002; 18(Suppl 1):S96–104. [PubMed: 12169536]

- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4:249–264. [PubMed: 12925520]
- Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. *Genome Res*. 2011; 9:1543–1551. [PubMed: 21816910]
- Kalocsai P, Shams S. Use of bioinformatics in arrays. *Methods Mol Biol*. 2001; 170:223–236. [PubMed: 11357685]
- Kulkarni MM. Digital multiplexed gene expression analysis using the NanoString nCounter system. *Curr Protoc Mol Biol*. 2011; Chapter 25(Unit 25B):10. [PubMed: 21472696]
- Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*. 2004; 101:811–816. [PubMed: 14711987]
- Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *P Natl Acad Sci USA*. 2001; 98:31–36.
- Lin, Charles Y.; Lovén, J.; Rahl, Peter B.; Paranal, Ronald M.; Burge, Christopher B.; Bradner, James E.; Lee, Tong I.; Young, Richard A. Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell*. 2012; 151:56–67. [PubMed: 23021215]
- Lockhart DJ, Winzler EA. Genomics, gene expression and DNA arrays. *Nature*. 2000; 405:827–836. [PubMed: 10866209]
- 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]
- Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green Douglas R, Tessarollo L, Casellas R, et al. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. *Cell*. 2012; 151:68–79. [PubMed: 23021216]
- Northcott PA, Shih DJ, Peacock J, Garzia L, Morrissy AS, Zichner T, Stutz AM, Korshunov A, Reimand J, Schumacher SE, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*. 2012; 488:49–56. [PubMed: 22832581]
- Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*. 2011; 12:87–98. [PubMed: 21191423]
- Pajic A, Spitkovsky D, Christoph B, Kempkes B, Schuhmacher M, Staeger MS, Brielmeier M, Ellwart J, Kohlhuber F, Bornkamm GW, et al. Cell cycle activation by c-myc in a Burkitt lymphoma model cell line. *Int J Cancer*. 2000; 87:787–793. [PubMed: 10956386]
- Quackenbush J. Microarray data normalization and transformation. *Nat Genet*. 2002; 32:496–501. [PubMed: 12454644]
- Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*. 2001; 98:15149–15154. [PubMed: 11742071]
- Reimers M. Making informed choices about microarray data analysis. *PLoS Comput Biol*. 2010; 6:e1000786. [PubMed: 20523743]
- Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*. 2000; 24:227–235. [PubMed: 10700174]
- Schena M, Heller RA, Thériault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*. 1998; 16:301–306. [PubMed: 9675914]
- Schmitz R, Young RM, Ceribelli M, Jhavar S, Xiao W, Zhang M, Wright G, Shaffer AL, Hodson DJ, Buras E, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*. 2012
- Schuhmacher M, Staeger MS, Pajic A, Polack A, Weidle UH, Bornkamm GW, Eick D, Kohlhuber F. Control of cell growth by c-Myc in the absence of cell division. *Curr Biol*. 1999; 9:1255–1258. [PubMed: 10556095]
- Schulze A, Downward J. Navigating gene expression using microarrays--a technology review. *Nat Cell Biol*. 2001; 3:E190–195. [PubMed: 11483980]



- Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med.* 2002; 8:68–74. [PubMed: 11786909]
- Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF Jr, et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* 2001; 61:7388–7393. [PubMed: 11606367]
- TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* 2012; 487:330–337. [PubMed: 22810696]
- van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002; 415:530–536. [PubMed: 11823860]
- van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002; 347:1999–2009. [PubMed: 12490681]
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009; 10:57–63. [PubMed: 19015660]
- Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc.* 2004; 99:909–917.
- Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell.* 2002; 1:133–143. [PubMed: 12086872]

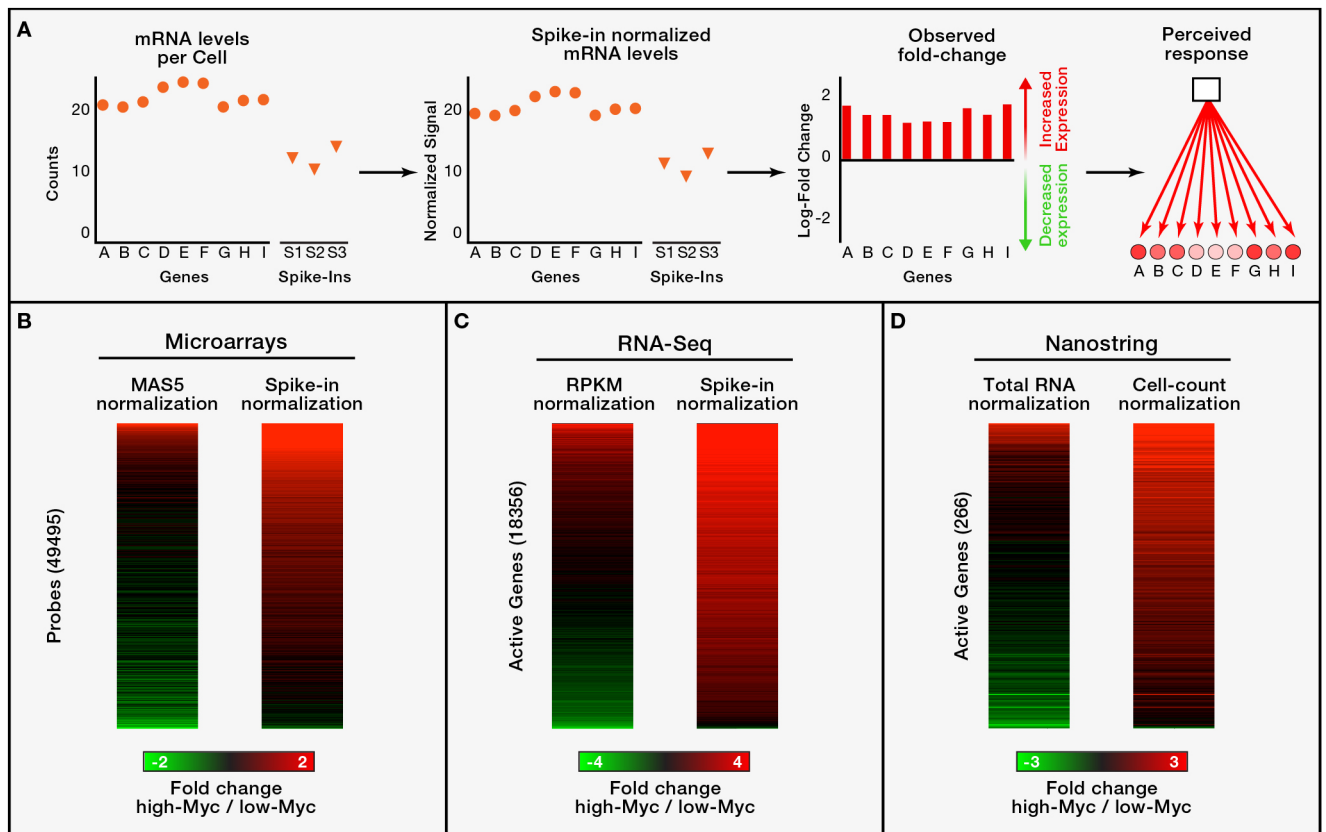
**Figure 1.**

A) Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells is similar. The square box represents a perturbation such as increased expression of a gene regulator or a change in environment or cell state. Red arrows point to target genes affected by the perturbation, which are represented as circles. Red shading of circles indicates relative transcriptional increase.

B) Schematic representation of microarray normalization when the overall levels of mRNA per cell are not changing in two conditions. Relative mRNA levels for 9 different genes (A–I) are indicated along the y-axis for condition 1 (black) and condition 2 (orange). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of median normalization; the calculated fold-changes based on median normalization, with increased expression represented by red bars above the midline and decreased expression represented by green bars below the midline; and, the perceived transcriptional response of a limited transcriptional increase in gene expression.

C) Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells is different such as in transcriptional amplification, where most genes are expressed at higher levels. The square box represents a perturbation such as increased expression of a gene regulator or a change in environment or cell state. Red arrows point to target genes affected by the perturbation, which are represented as circles. Red shading of circles indicates relative transcriptional increase.

D) Schematic representation of microarray normalization when the overall levels of mRNA per cell are increased in one condition compared to another. Relative mRNA levels for 9 different genes (A–I) are indicated along the y-axis for condition 1 (black) and condition 2 (orange). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of median normalization; the calculated fold-changes based on median normalization, with increased expression represented by red bars above the midline and decreased expression represented by green bars below the midline; and, the perceived transcriptional response following transcriptional amplification of gene expression.



**Figure 2.**

A) Schematic representation of microarray normalization when the total level of mRNA per cell is different, as in transcriptional amplification, but spike-in RNAs are used as standards for normalization. mRNA levels are indicated along the y-axis for condition 1 (black) and condition 2 (orange); individual genes are represented along the x-axis. Spike-in standards in the mRNA for condition 1 are represented by black triangles and spike-in standards in the mRNA for condition 2 are represented by orange triangles (S1–S3). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of normalization using the spike-in standards; the resulting fold-changes from condition 1 and condition 2, where increased expression is represented by red bars above the midline; and the perceived transcriptional response following transcriptional amplification of gene expression normalized with spike-in RNAs.

B) Heatmap showing the results of different normalization methods on the interpretation of microarray data. The data represent fold-change of expression in high-Myc vs. low-Myc cells. Each line represents data for individual probes on the microarray. Red indicates increased expression in high-Myc vs. low-Myc cells. Green indicates decreased expression in high-Myc vs. low-Myc cells. Black indicates no change in expression. The left panel displays data using a standard microarray normalization method (MAS5). The right panel shows the same data, now re-normalized using spike-in standards.

C) Heatmap showing the results of different normalization methods on the interpretation of RNA-sequencing data. The data represent fold-change of expression in high-Myc vs. low-Myc cells. Each line represents data for an individual gene. Red indicates increased expression in high-Myc vs. low-Myc cells. Green indicates decreased expression in high-Myc vs. low-Myc cells. Black indicates no change in expression. The left panel displays data using a standard sequencing normalization (reads per kilobase of exon model per

million mapped reads). The right panel shows the same data, now re-normalized using spike-in standards.

D) Heatmap showing the results of different sample preparation methods on the interpretation of digital quantification data. The data represent counts of mRNA molecules in high-Myc vs. low-Myc cells. Each line represents data for an individual gene. Red indicates increased expression in high-Myc vs. low-Myc cells. Green indicates decreased expression in high-Myc vs. low-Myc cells. Black indicates no change in expression. The left panel displays the results if the quantification is performed with equal amounts of total RNA for the high-Myc vs. low-Myc cells. The right panel displays the results if the quantification is performed with RNA from equal numbers of high-Myc and low-Myc cells.