Revisiting Global Gene Expression Analysis

Jakob Lovén,[1,5] David A. Orlando,[1,5] Alla A. Sigova,[1] Charles Y. Lin,[1,2] Peter B. Rahl,[1] Christopher B. Burge,[3] David L. Levens,[4] Tong Ihn Lee,[1,6,*] and Richard A. Young[1,3,6,*]
[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA
[2]Computational and Systems Biology Program
[3]Department of Biology
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Gene Regulation Section, Laboratory of Pathology, National Cancer Institute, Center for Cancer Research, Bethesda, MD, 20892, USA
[5]These authors contributed equally to this work and are co-first authors
[6]These authors contributed equally to this work and are co-senior authors
*Correspondence: tlee@wi.mit.edu (T.I.L.), young@wi.mit.edu (R.A.Y.)
http://dx.doi.org/10.1016/j.cell.2012.10.012

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 data sets in the National Center for Biotechnology Information 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 Winzeler, 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 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 by 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 two to three 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 are 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
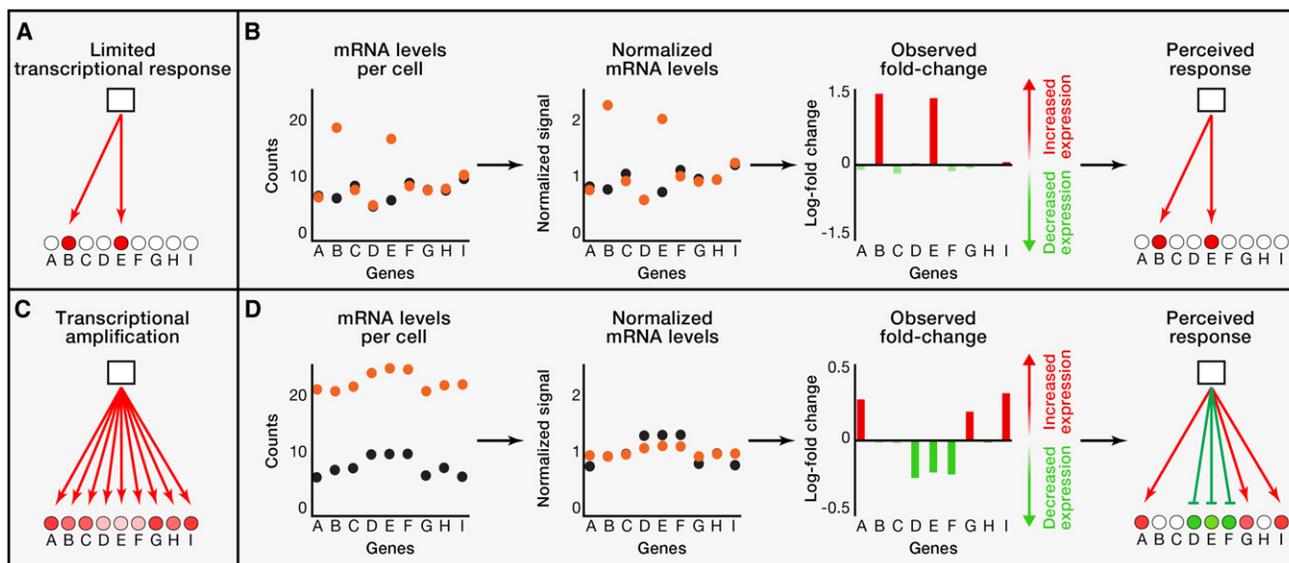
Figure 1. Normalization and Interpretation of Expression Data

(A) Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells are 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 nine 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.
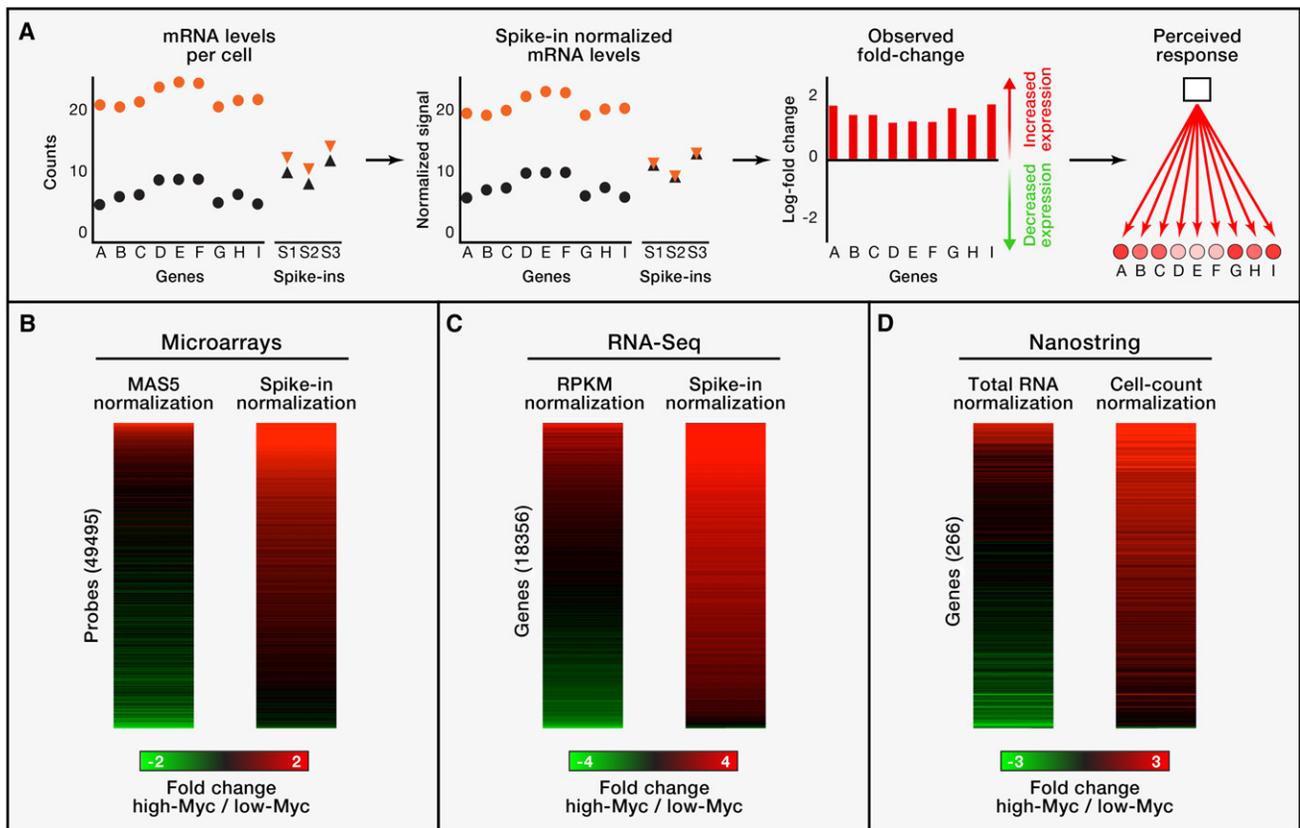
(D) Schematic representation of microarray normalization when the overall levels of mRNA per cell are increased in one condition compared to another. Relative mRNA levels for nine different genes (A–I) are indicated along the y axis for condition 1 (black) and condition 2 (orange). The panels, from left 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.

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 two to three 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 two to three 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, whereas 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) in which cells with high levels of the transcription factor were found to produce 2- to 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 with C-Chip disposable hemocytometers

Figure 2. Spike-In Controls, Normalized to Cell Number, Enable Accurate Interpretation of Transcriptional Changes

(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 versus low-Myc cells. Each line represents data for individual probes on the microarray. Red indicates increased expression in high-Myc versus low-Myc cells. Green indicates decreased expression in high-Myc versus low-Myc cells. Black indicates no change in expression. Left: data using a standard microarray normalization method (MAS5). Right: the same data, now renormalized by 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 versus low-Myc cells. Each line represents data for an individual gene. Red indicates increased expression in high-Myc versus low-Myc cells. Green indicates decreased expression in high-Myc versus low-Myc cells. Black indicates no change in expression. Left: data using a standard sequencing normalization (reads per kilobase of exon model per million mapped reads). Right: the same data, now renormalized by using spike-in standards.

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

(Digital Bio) 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 by using NanoString.

DNA-microarrays were first used to compare the high-Myc versus low-Myc cell RNA populations (Figure 2B; Table S1 available online). 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 by using standard approaches can be interpreted to mean that the expression levels of some genes are unchanged, whereas others increase or decrease (Figure 2B). The interpretation is quite different when the same data is normalized by using spike-in standards that reflect cell number: 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 by using standard approaches suggest that the expression levels of some genes are unchanged, whereas others increase or decrease (Figure 2C), yet when the same data are normalized by 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 by using this method. The results of this analysis suggest that the expression levels of some genes is unchanged, whereas 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- versus 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/cells 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; Cancer Genome Atlas, 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 are 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 p*myc*-tet construct in P493-6 cells was repressed with 0.1 μg/ml tetracycline (Sigma, T7660) for 72 hr. Cells were then washed three times with RPMI-1640 medium containing 10% tetracycline system approved FBS (Clontech, 631105) and 1% GlutaMAX and recultured in tetracycline-free culture conditions. All experiments were performed in the absence of EBNA2 activation. Cell numbers were determined by manually counting cells with 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 with the mirVANA miRNA isolation kit (Ambion, AM1560) following the manufacturer's instructions and resuspended in 100 μl nuclease-free water (Ambion, AM9938). Total RNA was spiked-in with the External RNA Controls Consortium (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 250–2,000 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, 1μl of a 1:10 dilution of Mix #1 was added to total RNA extracted from 1×10^6 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 (Affymetrix) by using 10 μg or 100 ng of total RNA, respectively.

For these experiments, we followed the manufacturer's recommendation and added the spike-in controls to total RNA

following RNA extraction. However, we have found that spike-in controls can also be added directly to the sample-Trizol homogenate prior to RNA purification if desired.

Microarray Sample Preparation and Analysis

For microarray analysis, 100 ng 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 by using GeneChip Expression Console. A Primeview Chip Definition File that included probe information for the ERCC controls, provided by Affymetrix, was used to generate CEL files. We processed the CEL files by 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 probe set expression values. The parameters of the expresso command were set to generate Affymetrix MAS5-normalized probe set values. We used a loess regression to renormalize these MAS5 normalized probe set values by using only the spike-in probe sets to fit the loess. The *affy* package provides a function, *loess.normalize*, which will perform loess regression on a matrix of values (defined by using the parameter *mat)* and allows for the user to specify which subset of data to use when fitting the loess (defined by 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 probe sets, 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 probe set 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 μg of total RNA containing ERCC ExFold Mix #1 RNA spike-in controls (see above), we prepared sequencing libraries according to the following protocol. Polyadenylated RNA was purified by two rounds of selection with 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 by using RNase H and DNA Polymerase I. In the second-strand synthesis reaction, dTTP was replaced with dUTP. After 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 by 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 (Epicenter, HU59100). The adaptor ligated single-stranded cDNA was then amplified with 15 cycles

of PCR and PCR products were purified by using gel electrophoresis. These RNA-Seq libraries were subsequently sequenced on Illumina HiSeq 2000. Sequences were aligned by 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 renormalize the RPKM values by 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 by using the parameter *mat)* and allows for the user to specify which subset of data to use when fitting the loess (defined by 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. Eighteen thousand five hundred and thirty-six genes with a RPKM value of 1.0 or greater in the low-Myc sample were selected, and the log2 fold ratio between the low-Myc and high-Myc samples were 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, 1×10^6 cells were collected and lysed directly either in 100 μl RLT buffer (QIAGEN, 74104) to yield a concentration of 10,000 cells per μl or in 500 μl lysis buffer with 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 100 ng 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 codesets encompassed sets of known cancer related genes (CodeSets CS-1 and CS-2) (Delmore et al., 2011). For each NanoString data set, 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 data set. Two hundred and sixty-six genes showing expression with a normalized value of 1.0 or greater in both the low-Myc Total-RNA and low-Myc Cell-Count samples were selected, and the log2 fold ratio between the low-Myc and high-Myc samples were calculated and shown as a heatmap.

ACCESSION NUMBERS

Raw and normalized microarray and RNA-Seq data can be found online associated with the GEO Accession ID GSE40784.

SUPPLEMENTAL INFORMATION

ACKNOWLEDGMENTS

REFERENCES

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. Nat. Rev. Genet. *13*, 552–564.

Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. *8*, 816–824.

Benes, V., and Muckenthaler, M. (2003). Standardization of protocols in cDNA microarray analysis. Trends Biochem. Sci. *28*, 244–249.

Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Ma-guire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C., et al. (2010). Integrative analysis of the melanoma transcriptome. Genome Res. *20*, 413–427.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adeno-carcinoma subclasses. Proc. Natl. Acad. Sci. USA *98*, 13790–13795.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406*, 536–540.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A compar-ison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Cancer Genome Atlas Network. (2012). Comprehensive molecular character-ization of human colon and rectal cancer. Nature *487*, 330–337.

Delmore, J.E., Issa, G.C., Lemieux, M.E., Rahl, P.B., Shi, J.W., Jacobs, H.M., Kastritis, E., Gilpatrick, T., Paranal, R.M., Qi, J., et al. (2011). BET bromodo-main inhibition as a therapeutic strategy to target c-Myc. Cell *146*, 904–917.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Geiss, G.K., Bumgarner, R.E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D.L., Fell, H.P., Ferree, S., George, R.D., Grogan, T., et al. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat. Bio-technol. *26*, 317–325.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molec-ular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. P. Soc. Photo-Opt. Ins. *2*, 132–140.

Heller, M.J. (2002). DNA microarray technology: devices, systems, and appli-cations. Annu. Rev. Biomed. Eng. *4*, 129–153.

Hill, A.A., Brown, E.L., Whitley, M.Z., Tucker-Kellogg, G., Hunter, C.P., and Slonim, D.K. (2001). Evaluation of normalization procedures for oligonucleo-tide array data based on spiked cRNA controls. Genome Biol. *2*, 1-0055.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18*(Suppl 1), S96–S104.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experi-ments. Genome Res. *21*, 1543–1551.

Kalocsai, P., and Shams, S. (2001). Use of bioinformatics in arrays. Methods Mol. Biol. *170*, 223–236.

Kulkarni, M.M. (2011). Digital multiplexed gene expression analysis using the NanoString nCounter system. Curr. Protoc. Mol. Biol., *Chapter 25*.

Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Fer-rari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. (2004). Gene expres-sion profiling identifies clinically relevant subtypes of prostate cancer. Proc. Natl. Acad. Sci. USA *101*, 811–816.

Li, C., and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA *98*, 31–36.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional Amplification in Tumor Cells with Elevated c-Myc. Cell *151*, 56–67.

Lockhart, D.J., and Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. Nature *405*, 827–836.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D.R., Tessarollo, L., Casellas, R., et al. (2012). c-Myc Is a Universal Amplifier of Ex-pressed Genes in Lymphocytes and Embryonic Stem Cells. Cell *151*, 68–79.

Northcott, P.A., Shih, D.J., Peacock, J., Garzia, L., Morrissy, A.S., Zichner, T., Stütz, A.M., Korshunov, A., Reimand, J., Schumacher, S.E., et al. (2012). Subgroup-specific structural variation across 1,000 medulloblas-toma genomes. Nature *488*, 49–56.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Pajic, A., Spitkovsky, D., Christoph, B., Kempkes, B., Schuhmacher, M., Staege, M.S., Brielmeier, M., Ellwart, J., Kohlhuber, F., Bornkamm, G.W., et al. (2000). Cell cycle activation by c-myc in a burkitt lymphoma model cell line. Int. J. Cancer *87*, 787–793.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nat. Genet. *32*(Suppl), 496–501.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA *98*, 15149–15154.

Reimers, M. (2010). Making informed choices about microarray data analysis. PLoS Comput. Biol. *6*, e1000786.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic varia-tion in gene expression patterns in human cancer cell lines. Nat. Genet. *24*, 227–235.

Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., and Da-vis, R.W. (1998). Microarrays: biotechnology's discovery platform for func-tional genomics. Trends Biotechnol. *16*, 301–306.

Schmitz, R., Young, R.M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Shaffer, A.L., Hodson, D.J., Buras, E., et al. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature 490, 116–120.

Schuhmacher, M., Staege, M.S., Pajic, A., Polack, A., Weidle, U.H., Bornkamm, G.W., Eick, D., and Kohlhuber, F. (1999). Control of cell growth by c-Myc in the absence of cell division. Curr. Biol. 9, 1255–1258.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays—a technology review. Nat. Cell Biol. 3, E190–E195.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8, 68–74.

Su, A.I., Welsh, J.B., Sapinoso, L.M., Kern, S.G., Dimitrov, P., Lapp, H., Schultz, P.G., Powell, S.M., Moskaluk, C.A., Frierson, H.F., Jr., and Hampton, G.M. (2001). Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. 61, 7388–7393.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. 347, 1999–2009.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Wu, Z.J., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. J. Am. Stat. Assoc. 99, 909–917.

Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1, 133–143.