Published in final edited form as:

Nat Immunol. ; 13(3): 199–203. doi:10.1038/ni.2244.

Gene-expression measurement: variance-modeling considerations for robust data analysis

Shankar Subramaniam and Gene Hsiao

Department of Bioengineering, the Department of Cellular and Molecular Medicine and the Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA

Shankar Subramaniam: shankar@ucsd.edu

Abstract

System-wide measurements of gene expression by DNA microarray and, more recently, RNA-sequencing strategies have become *de facto* tools of modern biology and have led to deep understanding of biological mechanisms and pathways. However, analyses of the measurements have often ignored statistically robust methods that account for variance, resulting in misleading biological interpretations.

The main objectives of gene-expression profiling in a biological system are to delineate the mechanisms associated with the triggering of transcriptional changes and to postulate putative downstream events that are a consequence of the function of protein products that result from altered expression of genes^{1,2}. Commonly, comparisons are made of biological systems subject to different stimuli or of normal and diseased states to elucidate the differences in the expression of genes that lead to altered endpoint phenotypes. As several hundreds to thousands of genes show differences in expression, it becomes a challenge first to identify the changes that are statistically significant and then to identify changes in genes that lead to altered biological mechanisms. Instead of using statistical methods, researchers have adopted an approach assessing change as ‘-fold’ differences (‘fold-change’) or some variant thereof for most gene-expression studies; this leads to inaccurate results, especially for the transcripts of low abundance that dominate gene-expression measurements. Furthermore, in studies with fewer replicates, mandated by either availability of samples or cost, fold-change analysis leads to a large number of false-positive results.

One of the most fundamental questions in gene-expression analysis is whether a particular stimulus produces a difference in the expression of a given gene. For example, an experimenter may want to determine which genes are induced in macrophages after treatment with an inflammatory stimulus. Several RNA samples may be collected from cells that have been treated with the stimulus or with vehicle alone. The biological question (whether this produces a difference in gene expression) can then be addressed through the

use of a statistical test to identify those genes with the greatest change in expression relative to 'measurement noise'. Unfortunately, the massively parallel nature of microarray studies limits the effectiveness of traditional statistical tests in addressing this question. As typically on the order of 1×10^5 genes are probed by each array, multiple-testing correction must be used to limit the number of false-positive results detected by the test. For example, for the application of a conventional *t*-test, scientists often choose a type I error (α) value of 0.05 as a significance threshold. Then, if only a single gene is tested, there is only a 1-in-20 chance that a significant change in gene expression has occurred by random chance alone. However, when 1×10^5 genes are tested simultaneously at this threshold, 5×10^3 false-positive results are expected. Two multiple-testing corrections are commonly used for microarray analysis: the Bonferroni correction³ and empirical false-discovery rates⁴. Application of these thresholds with conventional statistical tests frequently yields very few, and often no, statistically significant changes, given the limited number of experimental replicates commonly used. It is partially because of this result that statistical methods have not been rapidly adopted by the biological-research community.

In conventional microarray expression measurements, as well as those obtained by RNA-sequencing (RNA-seq) methods, most genes with differences in regulation have low intensity or tag count, and this makes the analysis problem difficult to resolve. Traditional statistical methods such as the *t*-test, which attempt to identify genes with differences in expression, fail when the average intensity is low and provide misleading insights into biological mechanisms. In essence, prohibitively large sample sizes are necessary for sufficient statistical power and reliable results through the use of traditional methods. Here we will provide an analysis of the error structure of gene-expression measurements, common pitfalls associated with measurement mean-based approaches and the power of variance modeling; we hope this will help investigators in immunology to use more robust analysis methods for microarray or RNA-seq data sets.

The error structure of microarray data

Measurement of variance is an indication of the reproducibility of several measurement obtained with the same experimental conditions. Many potential sources of variability exist, including noise introduced by the experimental equipment and fluorescence quality, which we categorize as gene expression-independent variability, as well as hybridization strength and RNA-input amount, which we categorize as gene expression-dependent variability. Biological variability (the changes in gene expression that are variable between controlled measurements) may contribute to both classes of variance. Both gene expression-independent and gene expression-dependent variability lie beyond the control of the experimenter, and many of these variables can influence individual genes differently. It has been widely observed that the degree of measurement uncertainty is dependent on the amount of gene expression and, in many cases, the variance decreases as a fraction of the total signal as expression increases. At greater signal intensities, both classes of variability are small relative to the signal. This error structure of high-throughput data can itself be a reliable measure of noise and may be used to devise tests with better accuracy. For determining whether a particular gene measured on an array has different expression in two physiological conditions, it is necessary to account for uncontrollable sources of variation.

To offset the typically poor estimation of variance and to increase statistical power, investigators have developed methods that borrow information from genes across the array. Distribution parameters or pseudo-replicates can be shared for genes, with the assumption that different genes may have similar patterns of variability. This strategy is largely agreed on by many groups and has been implemented in many software packages, including but not limited to SAM (significance analysis of microarrays)⁵, Cyber-T⁶, Limma (linear models for microarray data)⁵ and VAMPIRE (variance-modeled posterior inference with regional exponentials)^{8,9}. Detailed descriptions of the statistical modeling that underlies these analysis programs have been provided elsewhere^{5–9}. SAM augments estimates of the standard deviation with an estimated positive constant, so that variances closer to zero are proportionately penalized. Cyber-T uses a posterior variance estimate based on a Bayesian probability conditional on observed data, which combines sample variance and a background pooling of variances from genes with similar expression intensities contained in a window size defined by the user. Limma instead uses a posterior variance that represents a weighted combination of sample variance and a more stable grand variance estimate that is approximated from all data on the array. VAMPIRE is distinct in that it uses variance estimates derived from a global model of the intrinsic variance structure of microarray data. The global model that describes the relationship between expression intensity and variance (σ^2) can be written as follows:

$$\rho^2 = \mu^2 A + B$$

where A is the expression-dependent variance, B is the expression-independent variance and μ is the true expression. The Markov chain Monte Carlo (MCMC) algorithm is used for the computation of estimates for the hyper-parameters A and B . As with least-squares regression, μ is not known before the observation of array measurements; thus, the observed sample mean is often used as a surrogate. This choice, although necessary, can be problematic at low intensities at which poor signal-to-noise ratios can cause underestimation of true variability. Therefore, estimation techniques that identify parameters that are stable relative to an expression cutoff are often used. From this global variance model, estimates of variance in gene expression are extrapolated and integrated into a statistical test.

In a two-channel microarray, measurements are obtained from two RNA samples, each labeled with a different dye. These samples are subsequently hybridized to microarray probes on a single array. Typically, one RNA sample represents a control and the other represents some kind of experimental treatment. In this system, we refer to each array as a single replicate ($n = 1$). Parallel RNA samples from the control and experimental group can be subsequently labeled in reverse and hybridized to a second array to create a 'dye-swap' replicate pair. These may either be aliquots of the previously used RNA samples (for array replicates) or samples from parallel experiments (for biological replicates). The simultaneous quantification of gene expression by two RNA samples leads to covariation of the measured signals. This covariation can be modeled by the introduction of correlation coefficients for expression-dependent and expression-independent variance. A variance model of this type contains a total of six parameters. These parameters are related to the

treatment condition variances (σ_1 and σ_2) and correlation (ρ) through the following equations:

$$\begin{aligned}\sigma_1^2 &= \mu_1^2 A_1 + B_1 \\ \sigma_2^2 &= \mu_2^2 A_2 + B_2 \\ \rho \sigma_1 \sigma_2 &= \rho_A \mu_1 \mu_2 \sqrt{A_1 A_2} + \rho_B \sqrt{B_1 B_2}\end{aligned}$$

where σ_1 and σ_2 are variance in the control and experimental sample, respectively, for a single probe set; A_1 and A_2 are the expression-dependent variances; B_1 and B_2 are the expression-independent variances; ρ_A is the expression-dependent correlation; ρ_B is the expression-independent correlation; and μ_1 and μ_2 are the true expression in each treatment condition. Parameters for the variance model can be computed by MCMC simulation to determine maximum-likelihood estimates for each of the six variance parameters, as achieved with the VAMPIRE microarray analysis web application. Standard statistical testing and significance thresholds that correct for multiple-parallel tests are deployed in variance-modeling strategies⁶⁻⁹.

Sample size and variance

Small sample size remains a characteristic feature of most high-throughput biology experiments. This leads to unstable estimates of measurement variances after data analysis and thus leads to inaccurate detection of measured concentration differences of biomolecules. This is evident in the increasingly divergent results obtained with different statistical methods as sample size decreases. For the Student's *t*-test and Welch's *t*-test, small sample size is especially debilitating, as it leads to a low-powered test statistic¹⁰.

To illustrate unstable variance estimation at small sample size, we present a two-replicate subset from the Affymetrix Latin Square data set¹¹ processed along with a complete data set composed of 42 replicates (Fig. 1a). The sample variance for two replicates has wide vertical scatter for features with low expression. Other two-replicate combinations show similar poor fit. As expected, the sample variance estimates are lower with 42 replicates. The superimposed VAMPIRE plots show that these two-replicate variance estimates are more stable and accurate than standard sample variance for small sample size. In addition to modeling the expression-to-variance relationship, VAMPIRE's model penalizes features with low expression with an additive variance component.

The importance of variance modeling of microarray data sets with fewer replicate measurements is further illustrated by biological analysis of RAW264.7 mouse macrophages treated with an inflammatory stimulus (Fig. 1b). Analysis of data from three pairs of dye-swapped samples across six time points⁹ shows that VAMPIRE identifies a nearly equal number of genes with statistically significant regulation with as little as a single pair of dye-swapped measurements as it would if six replicates were used. Cyber-T requires more replicates for identifying the same number of genes with a significant result (Fig. 1b). Methods that involve mean-based approaches, and even SAM, miss hundreds of highly significant gene-expression changes.

Sensitivity in variance modeling

The most incisive view of the importance of variance in microarray measurements has been provided by the analysis of highly precise microarrays in which a known abundance of each RNA is hybridized to the microarray ('spike-in' samples). A Platinum Spike data set from an Affymetrix array allows investigators to characterize in detail the repeatability of their analysis pipeline¹². In one such analysis, over 2,000 spike-in RNAs, representing expression changes from 1.2- to fourfold in two experimental conditions (A and B), are mixed with a background of over 3,500 equally abundant RNAs. To simulate an experiment with a smaller sample size, we present a three-replicate subset of the Platinum Spike as an MA plot (dye-pair intensity ratio (M) plotted by average intensity (A)); Fig. 2. Here, the spikes are increasingly difficult to detect as average expression decreases. In the region of low expression, not only are the spikes embedded with background and equally spiked features, but also the upregulated and down-regulated features lie on both sides of the axis, thereby leading to potentially false-positive results (Fig. 2a). With the use of VAMPIRE at a multiple-comparison corrections Bonferroni cutoff of 5% ($\alpha_{\text{Bonf}} = 0.05$), 77% of the spikes are detected correctly, with a false-positive rate of 2%. Cyber-T detects 71% of the spikes at the same cutoff, and the *t*-test detects 4% of these (data not shown).

We also present the results of an RNA spiked in equally in conditions A and B and therefore with no differences in expression (Fig. 2b). This equally spiked feature is found to be significant by the *t*-test but not by VAMPIRE, as VAMPIRE's variance model systematically modulates low-expression measurements. Thus, this mRNA expression is a false-positive result incorrectly identified as being significant by *t*-test analysis of the microarray data set. For comparison, we present the results of a different RNA spiked in differently in conditions A and B and therefore with truly different expression (Fig. 2c). This true-positive result is considered to represent different expression by the *t*-test, which is correct. VAMPIRE, which results in less modulation of measurements of higher expression (and therefore measurements with more confidence), also considers this probe set to represent different expression. Given these results of spike-in experiments with well-defined constant background cRNA, variance-modeling approaches such as VAMPIRE and Cyber-T perform well in unpaired and paired experimental design settings.

Variance modeling and fold change

It is widely recognized for gene-expression data that the signal-to-noise ratio is far worse for measurements of lower intensity than for those of higher intensity. Because of this, microarray users often define an arbitrary threshold, statistical or otherwise, to remove data of low reliability before analysis. Then, an observed change of twofold or greater is deemed significant.

Although it is not a statistically robust approach, the fold-change method can give fairly reliable results, detecting many genes known to be regulated by their corresponding stimuli. Still, this approach is limited because of the need to specify arbitrary thresholds that are presumed to indicate significance. Furthermore, these thresholds are rarely, if ever, determined a priori or correlated with statistical concepts of significance. These thresholds

are therefore subject to experimenter bias in identifying genes with differences in expression. To illustrate the consequences of using a fold-change approach, we present an analysis of genes that are regulated differently in adipose tissue of human patients who are insulin resistant and of a control population who are insulin sensitive^{13,14} (Fig. 3).

The region that falls short of the twofold range contains a rich repertoire of genes related to the immune system established as being important in insulin resistance in human adipose tissue (Fig. 3a). Conversely, genes identified as having a difference in expression of over twofold relate to neuroactive ligand-receptor interactions or regulation of actin cytoskeleton and are not biologically important (Fig. 3a). We also present results for two pathways related to the immune system showing statistically significant enrichment: antigen processing and presentation and leukocyte transendothelial migration (Fig. 3b,c); these are composed of genes identified above (Fig. 3a). Methods such as the *t*-test do not have the sensitivity to capture these biologically relevant mechanisms through analysis of changes in gene expression.

Variance and RNA-seq methods

Analog microarray methods for studying gene expression are being replaced by RNA-seq methods¹⁵. However, such digital sequencing methods have a variance structure similar to that of analog microarray data. This concern is illustrated by an example in which the relationship between variance and average expression is analyzed for human kidney samples (two replicates) subjected to RNA-seq by Illumina's next-generation sequencing method¹⁶ (Fig. 4a). The strong relationship between variance and mean expression demonstrates the failure of the normal assumptions that residual variances are normally distributed and hence are uncorrelated. Change presented as a function of expression in human liver versus kidney demonstrates the importance of analyzing the variance structure for the digital gene-expression data (Fig. 4b), similar to the results presented above (Fig. 3). The DESeq approach¹⁷, like VAMPIRE, exploits the expression-variance relationship of assay-wide data to better approximate gene-expression variances for experiments with few replicates. Similar to the microarray data sets presented above (Figs. 2a and 3a), transcripts with differences in expression are systematically distributed so that more confidence is given to measurements with a high sequence-tag count, whereas a greater change (higher 'fold' value) is required for confident determination of the significance of features with a low sequencetag count.

Summary

Gene-expression measurements have become integral tools in modern immunology¹⁸. Given the paucity of samples, especially in human studies, it is essential that analysis methods help experimental designs with minimal sample size and number. The use of mean and fold-change ratios alone can be misleading and can lead to erroneous assignment of mechanistic changes or can miss important mechanisms that can help diagnostic and therapeutic processes. There is rich information in both analog and digital expression data, and use of the inherent error structure of data, as well as analysis of the variance and its use in modeling expression changes, can provide valuable information about markers and

mechanisms of disease. Methods such as VAMPIRE, in addition to providing accurate analysis of samples of a limited size, can also be sufficiently sensitive to identify genes with different expression in regions of low tag count with low average expression. Most importantly, such methods avoid the need for arbitrary choices of fold-change cut-offs or assessment of significance with the *t*-test, which suffers from low statistical power.

Acknowledgments

We thank A. Hsiao for discussions. Supported by the US National Institute of Health (U19 AI090023, P01 DK074868, U54 GM069338 and R33 HL087375).

References

1. Jenner RG, Young RA. *Nat Rev Microbiol.* 2005; 3:281–294. [PubMed: 15806094]
2. Papin J, et al. *Nat Rev Mol Cell Biol.* 2005; 6:99–111. [PubMed: 15654321]
3. Allison DB, et al. *Nat Rev Genet.* 2006; 7:55–65. [PubMed: 16369572]
4. Benjamini Y, Hochberg Y. *J R Stat Soc B.* 1995; 57:289–300.
5. Tusher VG, et al. *Proc Natl Acad Sci USA.* 2001; 98:5116–5121. [PubMed: 11309499]
6. Baldi P, Long AD. *Bioinformatics.* 2001; 17:509–519. [PubMed: 11395427]
7. Smyth GK. *Stat Appl Genet Mol Biol.* 2004; 3:1–25.
8. Hsiao A, Subramaniam S. *Syst Synth Biol.* 2008; 2:95–104. [PubMed: 19680790]
9. Hsiao A, et al. *Bioinformatics.* 2004; 20:3108–3127. [PubMed: 15217816]
10. Jeanmougin M, et al. *PLoS ONE.* 2010; 5:1–9.
11. Irizarry RA, et al. *Biostatistics.* 2003; 4:249–264. [PubMed: 12925520]
12. Zhu Q, et al. *BMC Bioinformatics.* 2010; 11:285–302. [PubMed: 20507584]
13. Sears DD, et al. *Proc Natl Acad Sci USA.* 2009; 106:18745–18750. [PubMed: 19841271]
14. Hsiao G, et al. *Am J Physiol.* 2011; 300:164–174.
15. Mortazavi A, et al. *Nat Methods.* 2008; 5:621–628. [PubMed: 18516045]
16. Marioni JC, et al. *Genome Res.* 2008; 18:1509–1517. [PubMed: 18550803]
17. Anders S, Huber W. *Genome Biol.* 2010; 11:R106–R118. [PubMed: 20979621]
18. Pulendran B. *Nat Rev Immunol.* 2009; 9:741–747. [PubMed: 19763148]

Figure 1.

Expression-variance plot of an Affymetrix Latin Square 42-replicate data set. **(a)** Sample variance from two replicates (light blue) shows poorer approximation than the variance estimates from all 42 replicates (dark blue). In contrast, VAMPIRE two-replicate variance estimates (red) are much more stable and accurate. Global variance modeling stabilizes variance estimates, which leads to detection of more genes with statistically significant differences in expression while maintaining false-positive error control. **(b)** Comparison of variance methods and replicates of lipopolysaccharide-treated macrophage RAW264.7 cells. VAMPIRE can use as little as a single pair of dye-swapped measurements to identify similar numbers of statistically significant gene-expression changes ($\alpha_{\text{Bonf}} = 0.05$) at all six time points regardless of how many replicates are used in the analysis. The results obtained with Cyber-T approach those of VAMPIRE as the number of replicates increases.

Figure 2.

MA plot of dye-pair intensity ratio and average intensity. **(a)** MA plot characterization of features detected and missed by VAMPIRE analysis of a three-replicate subset of a Platinum Spike data set¹², alongside two probe set samples showing how VAMPIRE systematically modulated the variance estimate for the statistical tests. Here, the spikes (red and green) are increasingly difficult to detect as average expression decreases. In the region of low expression, not only the spikes embedded with background and equally spiked transcripts (gray and yellow), but also the upregulated and downregulated transcripts lay on both sides of the axis. Correct spike detection, bold green and red circles; false-positive rate, bold yellow squares. **(b,c)** Gray triangles indicate mean expression of a probe set for one condition; black bars indicate estimates of the standard deviation used as a basis for the *t*-test; red dashed bars indicate estimates of the standard deviation used as a basis for VAMPIRE.

Figure 3.

Gene-expression changes in insulin resistance. **(a)** MA plot, or change in expression (fold) as a function of expression for an Affymetrix microarray of insulin-resistant and insulin-sensitive (control) human adipose tissue¹³. Brown symbols represent transcripts that would have been exclusively considered differently expressed by the fold-change method. Blue symbols represent transcripts considered differently expressed by Student's *t*-test at a Bonferroni multiple-comparisons correction threshold of 0.05 ($\alpha_{\text{Bonf}} = 0.05$). Red symbols represent transcripts considered differently expressed by VAMPIRE at an α_{Bonf} value of 0.05. orange shading indicates transcripts found to be overexpressed in adipose tissue of insulin-resistant subjects by VAMPIRE but not captured by the fold-change method (because the change in expression is less than twofold) or *t*-test; transcripts in this region include subtle but overexpressed immune-related genes. **(b,c)** Statistical enrichment of functional terms by Goby, a gene and pathway annotation tool integrated into the VAMPIRE web framework, showing significant enrichment for genes identified as statistically significant by VAMPIRE in **a** that encode molecules of immune pathways such as antigen processing and presentation **(b)** and leukocyte transendothelial migration **(c)** Red rectangles indicate molecules encoded by genes with significant overexpression in adipose tissue of insulin-resistant subjects relative to their expression in insulin sensitive subjects; white rectangles indicate genes without differences in expression. Inflammation is a common feature of the insulin-resistant state. TCR, T cell antigen receptor; nK, natural killer; RoS, reactive oxygen species. Adapted from the Kyoto Encyclopedia of Genes and Genomes.

Figure 4.

RNA-Seq data and microarray data shares similar variance structure. **(a)** Sample variance as a function of expression in human kidney samples¹⁶ for Illumina sequencing of two technical replicates ($n = 2$). The assay-wide variance structure is similar to microarrays, with a strong mean-variance relationship. **(b)** MA plot (change in expression (fold) as a function of expression) for human liver versus human kidney. Bold blue points indicate transcripts with statistically significant differences in expression, as determined by DESeq software at a Bonferroni multiple-comparisons corrections threshold of 0.05.