Fast-track: Calibrating a DNA microarray and testing the calibration with real data

Original citation for this work *Pozhitkov, A., P. A. Noble, J. Bryk and D. Tautz. (2012) A New Design for Microarray Experiments to Account for Experimental Noise and the Uncertainty of Probe Response. PlosOne (submitted).*

Rationale Although microarrays are extensively used for examining the expression of genes, and for detecting single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) in genomic DNA, uncertainty remains on the interpretation of microarray output. The reason for this uncertainty is that different signal quantification algorithms often yield different interpretations of the same microarray dataset and different pre-processing and post-processing steps (normalization and scaling methods) can yield different results. For example, several studies have shown that about 20 to 30% of the genes are identified as up- or down- regulated solely depending on the algorithm used (2,7,13). We argue that the source of these uncertainty lies in the poor knowledge of the physicochemistry of the probe-target hybridizations. Many bioinformatic approaches to data analysis deal with assumed "noise" in the experimental data, without properly defining it. Alternatively, bioinformatic approaches are based on parameters known from the physicochemistry of solution-based hybridizations -- although these have not proven to be applicable for arrays where surface interactions play a role (5,11,14,15).

Hence, even two decades after the introduction of microarrays, the interpretation of the output remains an elusive problem (3,6,12). In retrospect, a signal quantification approach that *does* capture the behavior of probes on a microarray surface without extensive parameterization, normalization, or scaling, would be highly desired because it would substantially improve the interpretation of microarray data, eliminate the need for confirmatory analyses, such as qPCR (1,4,8), and enhance our ability to understand biological systems. Here we provide such a solution.

Factoid: Probe behavior on a microarray surface cannot be determined *in silico* (9,10) because every probe on a microarray surface has its own behavior governed by its binding affinities to targets, the concentration of the targets in solution, and the extent of noise in the fluorescent signal, all of which are unknown.

Solution: Characterize the behaviors of all probes on a microarray empirically. This can be accomplished by pooling labeled samples, making a dilution series, hybridizing each diluted sample to different arrays, and recording signal intensities. We built three programs that users can use to average signal intensities, calculate the hybridization isotherm (i.e., the relationship between signal intensity and target concentration), and test the isotherm with experimental data. Below is a synoposis of each program.

First software program: Averages the signal intensity of the probes (each probe is replicated ten times). Input files #1 to #6 contain information on the probe identity and corresponding signal intensities, from 0.0625 to 2 times the recommended target concentration. The output file contains information on the target concentration, the probe identity, the signal intensity average, and the standard deviation divided by average divided by the number of probes. This information is needed for input to calibrate probe signal intensities (see next).

Second software program: Takes the output from the first program and fits the data to three different models (linear, Ln; Freundlich, F; Langmuir, (L)). The fit of the best model is selected based of the R^2. The output of the second program is the probe identity, the coefficients of the selected model, the R^2 of the model, and the model selected. For example, in the case of Probe A_53_P100398, the probe signal intensities (*SI*) best fit a Langmuir model with the coefficients of 0.00069116 and 0.00325183 with a R^2=0.997242. The final formula for this probe is thus:

$$SI = y_{\max}\frac{Kx}{1+Kx}, where$$
$$y_{\max} = \frac{1}{0.00325183}; K = \frac{0.00325183}{0.00069116}$$

Third software program: Allows a user to test the models with experimental data. Here, the output file from the second program (containing model parameters) are used as input. Also, the file containing signal intensities from an experiment are used as input. The program determines the relative target concentrations based on the calibrated probes as well as provides an error term. For example, in the case of Probe A_67_P18309499, the relative target concentration is 0.861463 dilution units, meaning it was less than 1 time of the recommended target concentration of the pooled sample. The

error term is 0.0310935, which is the standard deviation divided by square root of the number of replicates. In other words, the error is 3.1%.

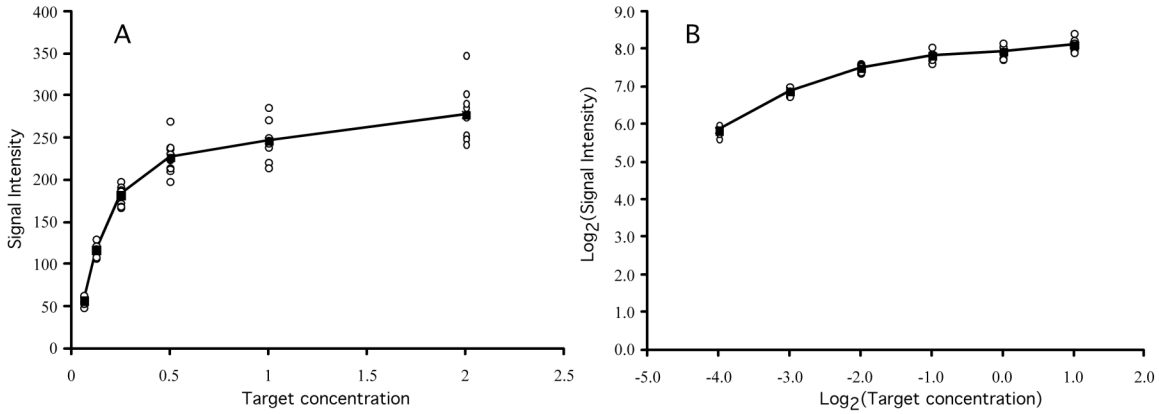

Figure 1. Signal intensity of raw signal intensities (open circles) and averaged signal intensities (closed circles) by the pooled target dilution series. In this example, the probe's behavior was best fitted to the Langmuir model. Panel A, raw values; Panel B, $\log_2$ transformed values.

Summary Calibration and testing of microarray data requires three programs. The first program averages signal intensities of the probes. The second program calibrates the probe dilution series so that signal intensities of each probe is fitted to a model. The third program allows users to test experimental data to determine the relative target concentration of a biological sample.

References

1. Ach RA, Wang H, Curry B. (2008). Measuring microRNAs: comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. BMC Biotechnol. 8:69. PMID: 18783629
2. Barash Y, Dehan E, Krupsky M, Franklin W, Geraci M, Friedman N, Kaminski N. (2004). Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. Bioinformatics. 20:839-46. PMID: 14751998
3. Chagovetz A, Blair S. (2009) Real-time DNA microarrays: reality check. *Biochem Soc Trans.* 37:471-5. PMID: 19290884
4. Gaj S, Eijssen L, Mensink RP, Evelo CT. (2008) Validating nutrient-related gene expression changes from microarrays using RT(2) PCR-arrays. Genes Nutr. 153-7. PMID: 19034552

5. Gong P, Levicky R. (2008). DNA surface hybridization regimes. Proc Natl Acad Sci U S A. 105:5301-5306.
6. Marshall E. (2004). Getting the noise out of gene arrays. Science. 306:630-1. PMID: 15499004
7. Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LA, Peeters AJ. (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. BMC Bioinformatics. 7:137. PMID: 16539732
8. Morey JS, Ryan JC, Van Dolah FM. (2006). Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biol Proced Online. 2006 8:175-93. PMID: 17242735
9. Mueckstein, U, Leparc, GG; Posekany, A; Hofacker, I; Kreil, DP. (2010). Hybridization thermodynamics of NimbleGen Microarrays. BMC Bioinformatics 11, art.no.-35.
10. Pozhitkov A, Noble PA, Domazet-Loso T, Nolte AW, Sonnenberg R, Staehler P, Beier M, Tautz D. (2006). Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucl. Acids Res.* 2006: 34: e66. PMID: 16707658
11. Pozhitkov AE, Boube I, Brouwer MH, Noble PA. (2010). Beyond Affymetrix arrays: expanding the set of known hybridization isotherms and observing pre-wash signal intensities. *Nucl Acids Res* 8: e2. PMID: 19969547
12. Pozhitkov AE, Tautz D, Noble PA. (2007). Oligonucleotide microarrays: widely applied--poorly understood. Brief Funct Genomic Proteomic. 6:141-8. PMID: 17644526
13. Seo J, Bakay M, Chen YW, Hilmer S, Shneiderman B, Hoffman EP. (2004). Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection *p*-value weighting in Affymetrix microarrays. Bioinformatics. 20:2534-44. PMID: 15117752
14. Vainrub A, Pettitt BM. (2003). Surface electrostatic effects in oligonucleotide microarrays: Control and optimization of binding thermodynamics. Biopolymers. 68:265-270.
15. Vainrub A, Pettitt BM. (2000). Thermodynamics of association to a molecule immobilized in an electric double layer. JChemical Physics Letters. 323:160-166.