A New Approach to Sample Deconvolution

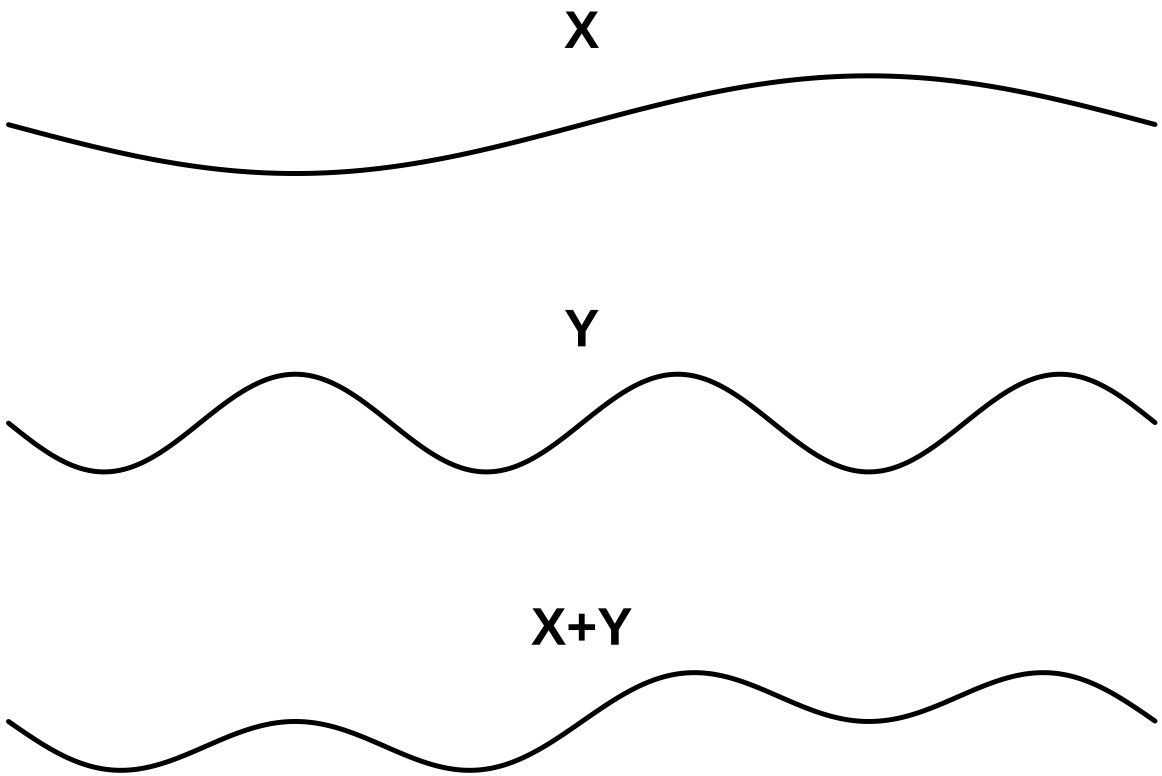
Greg Hunt

University of Michigan

December 9, 2016

Deconvolution = Decomposing Mixtures

Convolution: $X, Y \rightarrow X + Y$, Deconvolution $X + Y \rightarrow X, Y$



Sample Deconvolution = Decomposing Expression Data

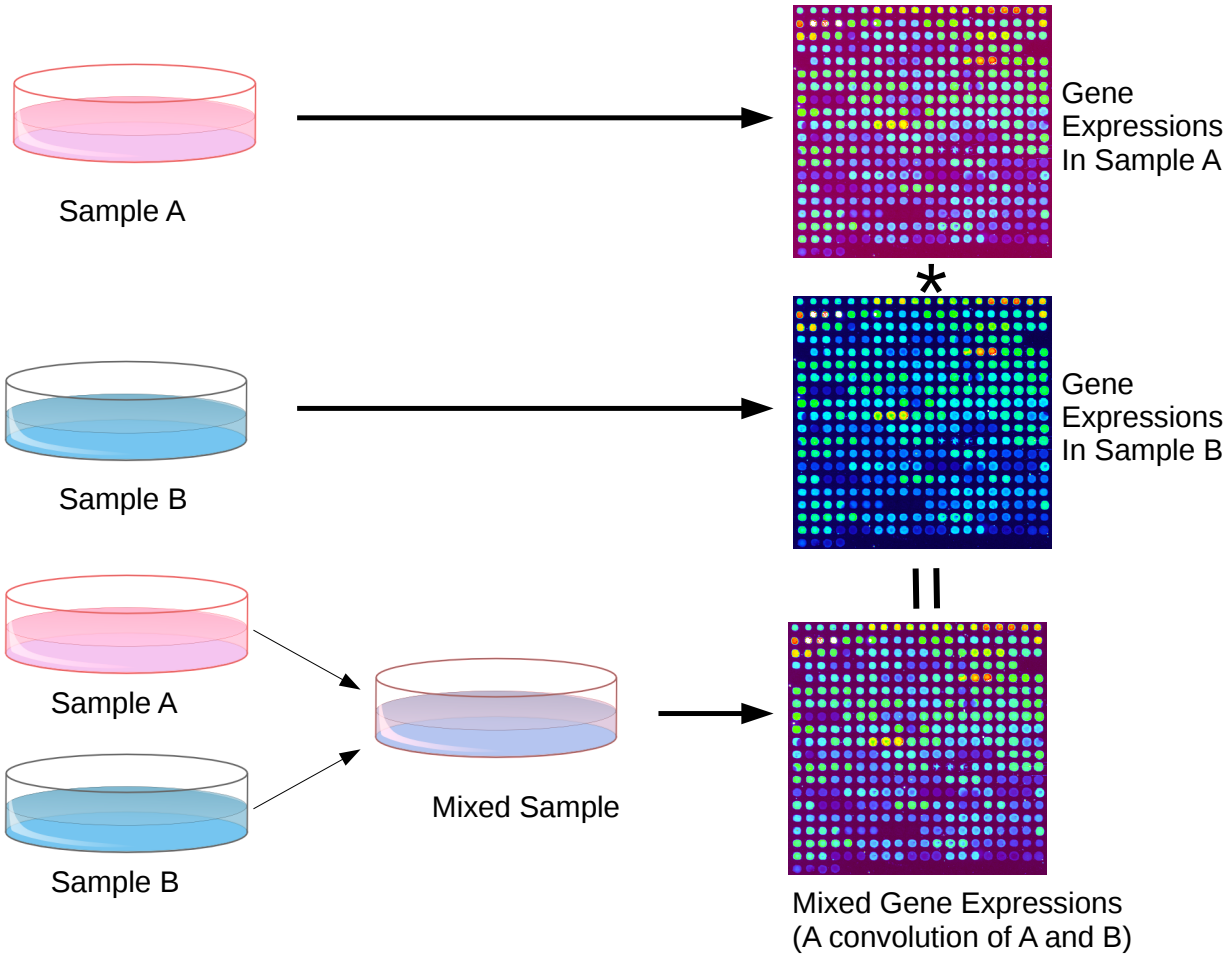


Image Credits: [1,2]

Scientific Background

Measure Gene Expression by Counting mRNA

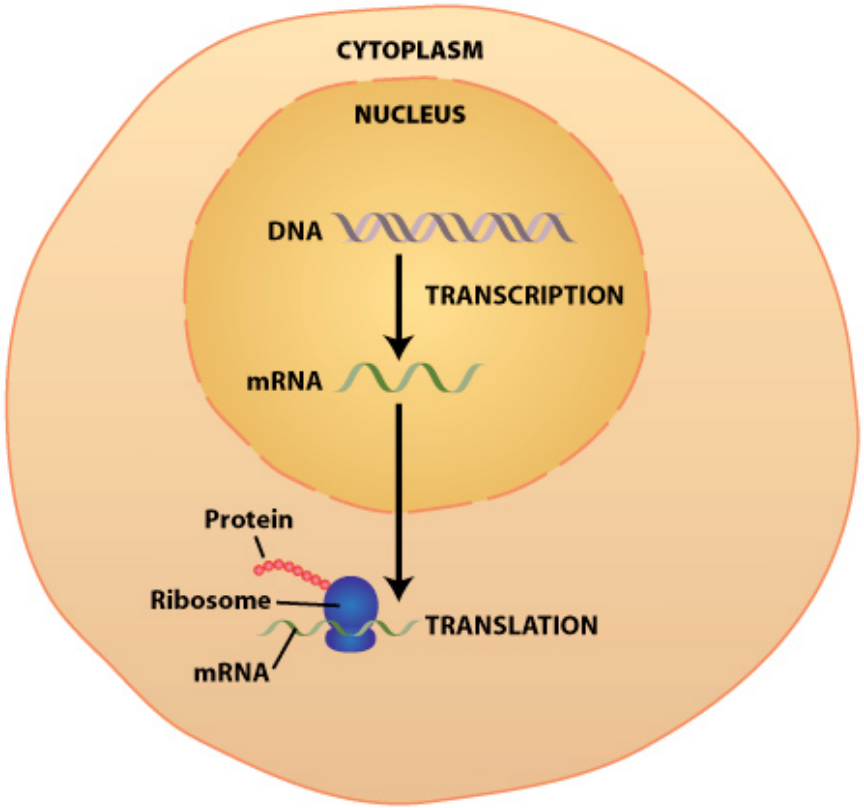


Image Credits: [4]

Every time a gene is expressed an mRNA specific to that gene is created.

Microarrays Determine which Oligos are Present

(1) Extract mRNA chunks (oligonucleotides or oligos).

(2) Determine which oligos present with microarray.

Image Credits: [5,6]

Microarrays are a Collection of Probes

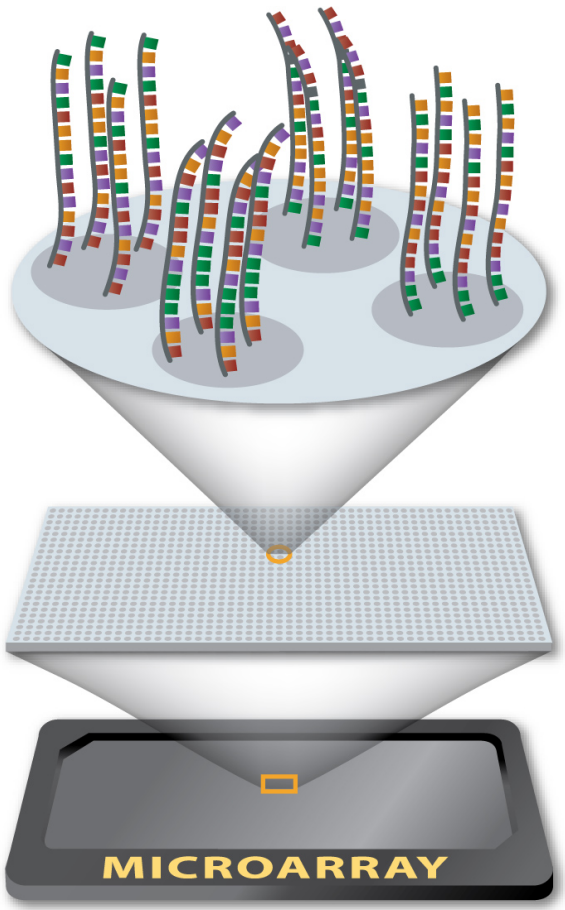
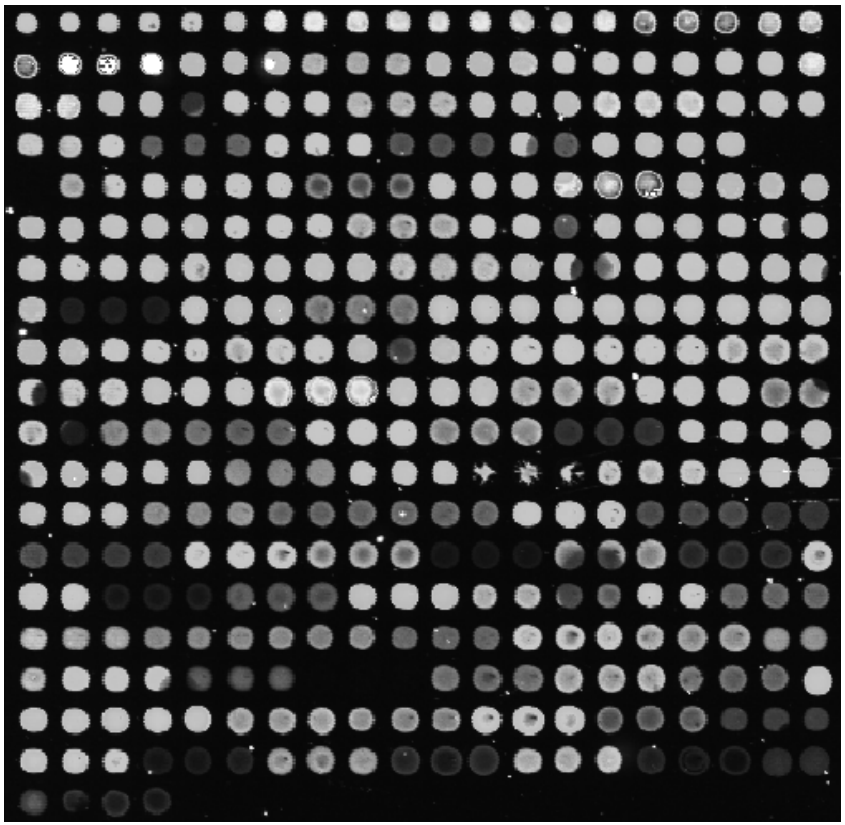


Image Credits: [7]

Probe spots on microarray each tell about abundance of a different oligo.

Fluorescence of a Probe = Abundance of an Oligo



Fluorescence of each probe tells us how much of that type of oligo is present.

Microarray Data is Probe Intensities

Microarray data is intensity measurements $I_1, \ldots, I_N$

Typically $N \approx 500,000$

Example:

Probe Name	1367452_at	1367453_at	1367454_at	1367455_at	$\cdots$
Intensity	84	5063	140	5065	$\cdots$

A common data pre-processing step is to logarithmically tranform.

Example:

Probe Name	1367452_at	1367453_at	1367454_at	1367455_at	$\cdots$
log(Intensity)	6.39	12.305	7.129	12.306	$\cdots$

Literature Review

The Common Model is Linear

Consider S samples

1. Each sample mixture of K cell types.
2. Measurements of N oligos in each sample.

The Model:

$$\underbrace{\left[\; X_{S\times N} \;\right]}_{\text{data matrix}} = \underbrace{\left[\; M_{S\times K} \;\right]}_{\text{mixing matrix}} \underbrace{\left[\; U_{K\times N} \;\right]}_{\text{characteristic expressions}} + E$$

for error matrix E.

Our Goal: Predict the Mixing Proportions

We always know X and model it as

$$X = MU + E$$

We may not know U or M.

Two types of deconvolution:

Partial Deconvolution

1. Know U and predict M
2. Know M and predict U

Full Deconvolution

3. Know neither M nor U and predict jointly

We are interested in (1): predicting M given U.

Literature: Similar Model, Different Fitting

Problem

Assume known X, U,

$$X = MU + E$$

and solve for M.

Constraint: M must be a row-wise probability matrix.

Solutions:

1. Regression: regress X on U so that elements of M are regression coefficients. Somehow enforce constraints.
2. Bayesian: Similar to LDA. Estimate as MAP.

Idea: marker oligos.

Marker Oligos are Oligos Expressed in Only One Cell Type

Empirically models have better fit if restricted to marker oligos.

Idea: Find marker oligos for each cell type. Restrict analyis to marker oligos only.

1. Can be as simple as fitting using submatrices.
2. Many different ways to select markers. Usually chosen by looking at pure samples.

Pure Samples are Training Data

Most methods require a pure sample of each of the K cell types.

1. Used to find markers.
2. Can give us U.

Other Minor Variations in the Literature

Model:

$$X = MU + E$$

Ways to solve this partial deconvolution problem differ by

1. Fitting methods.
2. Marker oligo choices.
3. Construction of U
4. **Transformations:** e.g. do a logarithmic transformation.
5. **Summarizations:** summarize probes into genes using RMA or MAS5.

New Methodology

Toy Example: Two Cell Types

Assume we have two cell types and three samples.

1. A = pure sample of type one.
2. B = pure sample of type two.
3. C = mixture sample.

Define:

$$\eta_{An} = \text{concentration of oligo n in sample } A$$

Similarly for η_{Bn} and η_{Cn}.

Our Concentration/Expression Model is Linear

$Y_{An} = \log_2(\text{expression of oligo } n \text{ in sample } A.)$

Similarly define Y_{Bn} and Y_{Cn}.

Assume the linear relationship between concentration and expression

$$Y_{An} = \theta_n + \gamma \log_2(\eta_{An}) + \epsilon_{An}$$
$$Y_{Bn} = \theta_n + \gamma \log_2(\eta_{Bn}) + \epsilon_{Bn}$$
$$Y_{Cn} = \theta_n + \gamma \log_2(\eta_{Cn}) + \epsilon_{Cn}$$

for all $n = 1, \ldots, N$. Where the ϵ are i.i.d with mean zero and constant variance.

Truth is Non-Linear but Linear Model is Reasonable

Remember the linear model: $Y_n = \theta_n + \gamma \log_2(\eta_n) + \epsilon$

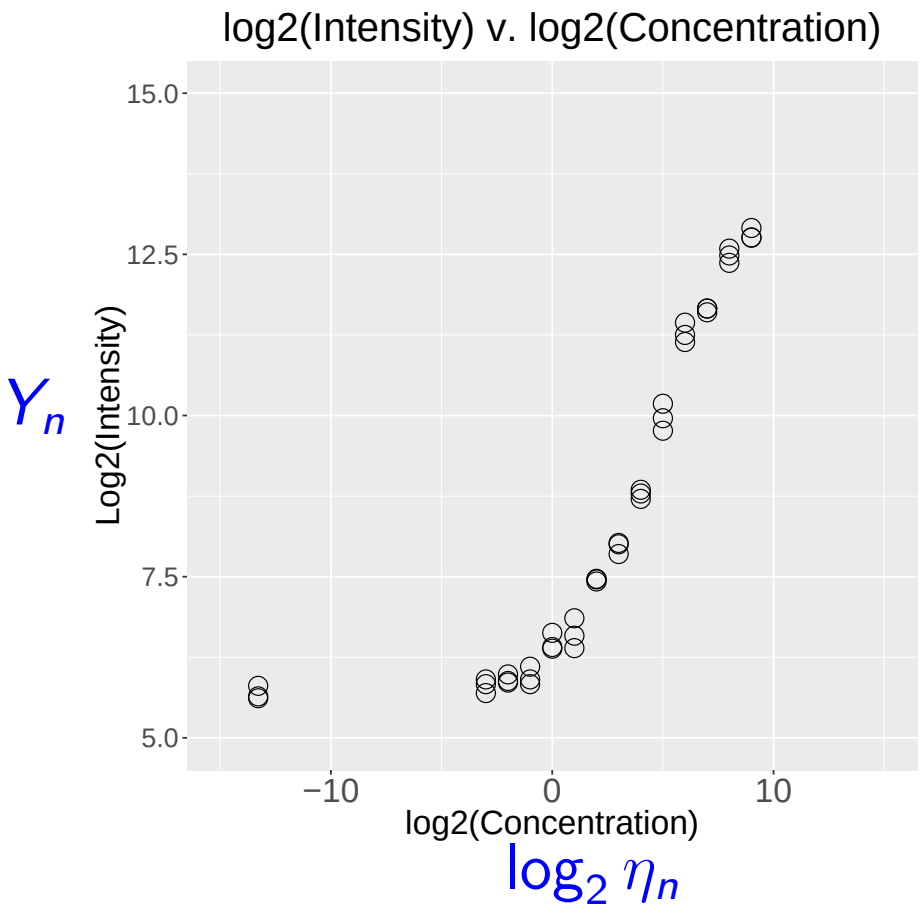


Figure: Relationship between concentration and expression.

Truth is Non-Linear but Linear Model is Reasonable

Remember the linear model: $Y_n = \theta_n + \gamma \log_2(\eta_n) + \epsilon$

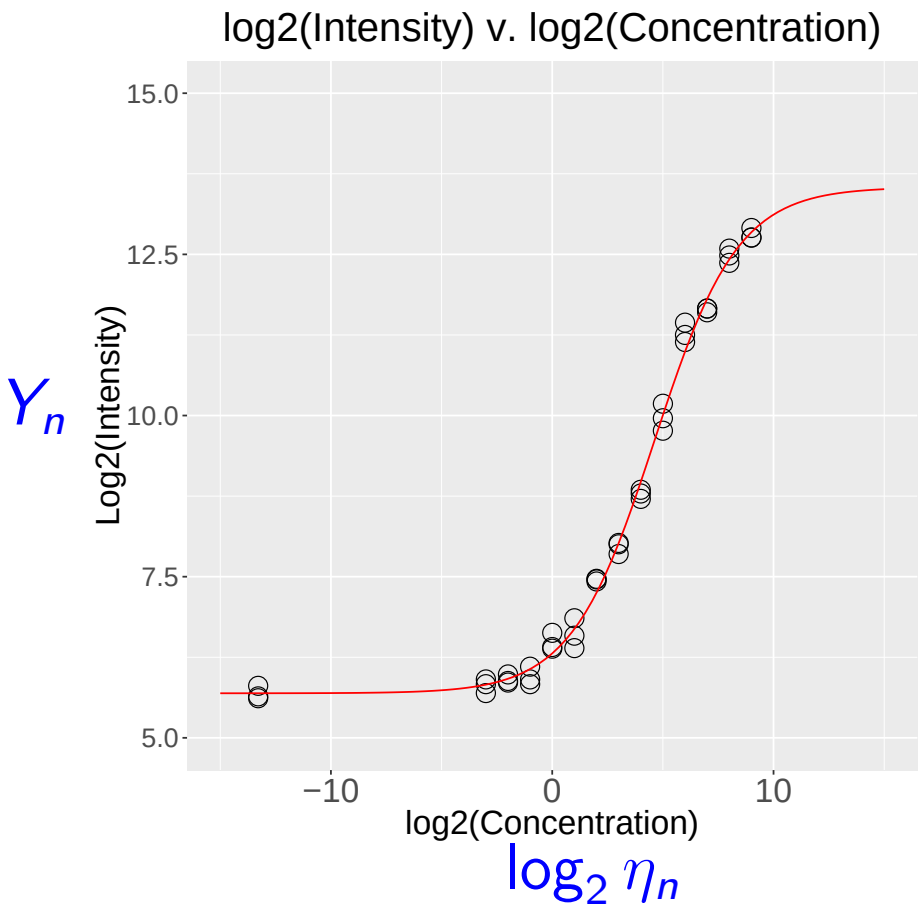


Figure: Relationship between concentration and expression.

Truth is Non-Linear but Linear Model is Reasonable

Remember the linear model: $Y_n = \theta_n + \gamma \log_2(\eta_n) + \epsilon$

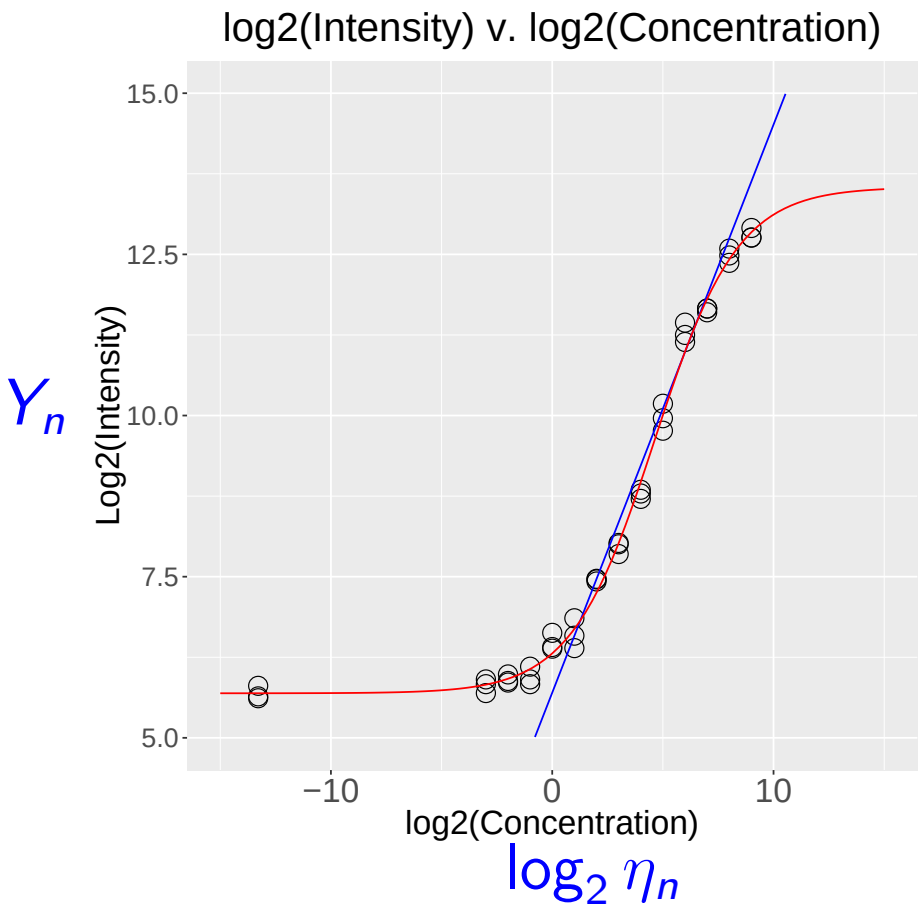


Figure: Relationship between concentration and expression.

Oligo Concentrations: Mixture is Combination of Pure

Sample C is mixture in proportions $p_1, p_2 \geq 0$ so that $p_1 + p_2 = 1$.

Thus

$$\eta_{Cn} = p_1 \eta_{An} + p_2 \eta_{Bn}$$

plugged into model gives

$$\begin{aligned} Y_{Cn} &= \theta_n + \gamma \log_2 (\eta_{Cn}) + \epsilon_{Cn} \\ &= \theta_n + \gamma \log_2 (p_1 \eta_{An} + p_2 \eta_{Bn}) + \epsilon_{Cn}. \end{aligned}$$

Marker Oligos Separate Cell Types

Generally

$$Y_{Cn} = \theta_n + \gamma \log_2 (p_1 \eta_{An} + p_2 \eta_{Bn}) + \epsilon_{Cn}.$$

Define:

$$n_1 = \text{marker oligo for type one}$$
$$n_2 = \text{marker oligo for type two}$$

then

$$\eta_{An_2} = 0 \text{ and } \eta_{Bn_1} = 0.$$

Hence

$$Y_{Cn_1} = \theta_{n_1} + \gamma \log_2 (p_1 \eta_{An_1}) + \epsilon_{Cn_1}$$

and

$$Y_{Cn_2} = \theta_{n_2} + \gamma \log_2 (p_2 \eta_{An_2}) + \epsilon_{Cn_2}$$

Marker Oligos Allow Isolation of Proportions

Subtract off expression of marker oligos in pure samples:

$$\begin{aligned} Y_{Cn_1} - Y_{An_1} &= \theta_{n_1} + \gamma \log_2 (p_1 \eta_{An_1}) + \epsilon_{Cn_1} \\ &\quad - (\theta_{n_1} + \gamma \log_2 (\eta_{An_1}) + \epsilon_{An_1}) \\ &= \gamma \log_2 (p_1) + \epsilon_{Cn_1} - \epsilon_{An_1} \end{aligned}$$

hence

$$\exp_2 \left(\frac{Y_{Cn_1} - Y_{An_1}}{\gamma} \right) = \lambda_1 p_1$$

where $\lambda_1 = \exp_2 \left(\frac{\epsilon_{Cn_1} - \epsilon_{An_1}}{\gamma} \right)$ is some error term.

We can do something similar for p_2.

Replace γ with $\widehat{\gamma}$ to Get Estimator

Replace γ with $\widehat{\gamma}$ in

$$\exp_2\left(\frac{Y_{Cn_1} - Y_{An_1}}{\gamma}\right) = \lambda_1 p_1$$

to get estimator

$$\widehat{q_1} = \exp_2\left(\frac{Y_{Cn_1} - Y_{An_1}}{\widehat{\gamma}}\right) \approx \lambda_1 p_1$$

since $\widehat{\gamma} \approx \gamma$.

Similarly define an estimator $\widehat{q_2}$.

Re-normalize to get Final Estimators

Define

$$\widehat{p_1} = \frac{\widehat{q_1}}{\widehat{q_1} + \widehat{q_2}}$$
$$= \mathit{logistic}^{-1}\left(\frac{(Y_{Cn_1} - Y_{Cn_2}) - (Y_{An_1} - Y_{Bn_2})}{\widehat{\gamma}}\right)$$

Reasons

1. Guarantees that $0 \leq \widehat{p_1}, \widehat{p_2} \leq 1$ and $\widehat{p_1} + \widehat{p_2} = 1$
2. Nice interpretation as $\mathit{logistic}^{-1}$(baseline corrected difference)

Similarly for p_2.

A General Model: K cell types, multiple markers

Assumptions:

1. K cell types
2. ν_k pure samples of type k
3. log-level microarray data for each pure sample

$$Z_{kr} \in \mathbb{R}^{1\times N}, \quad r = 1, \ldots, \nu_k.$$

4. Set of marker oligos G_k for each cell type, $|G_k| = \Gamma_k$
5. Heterogeneous mixture of the K cell types in proportions $p_1, \ldots, p_K$
6. log-level expression measurements of heterogeneous sample: Y_n

General Estimators Look Similar

For the simple case we had

$$\widehat{q_1} = \exp_2\left(\frac{Y_{Cn_1} - Y_{An_1}}{\widehat{\gamma}}\right)$$

Generally define

$$\widehat{q_k} = \exp_2\left(\frac{\frac{1}{\Gamma_k}\sum_{n\in G_k}\left(Y_n - \overline{Z_{kn}}\right)}{\widehat{\gamma}}\right) \approx \lambda_k p_k$$

for $k = 1, \ldots, K$ and

$$\widehat{p_k} = \frac{\widehat{q_k}}{\sum_{t=1}^{K}\widehat{q_t}}$$

as our estimator of p_k.

Estimate γ from Benchmark Data Sets

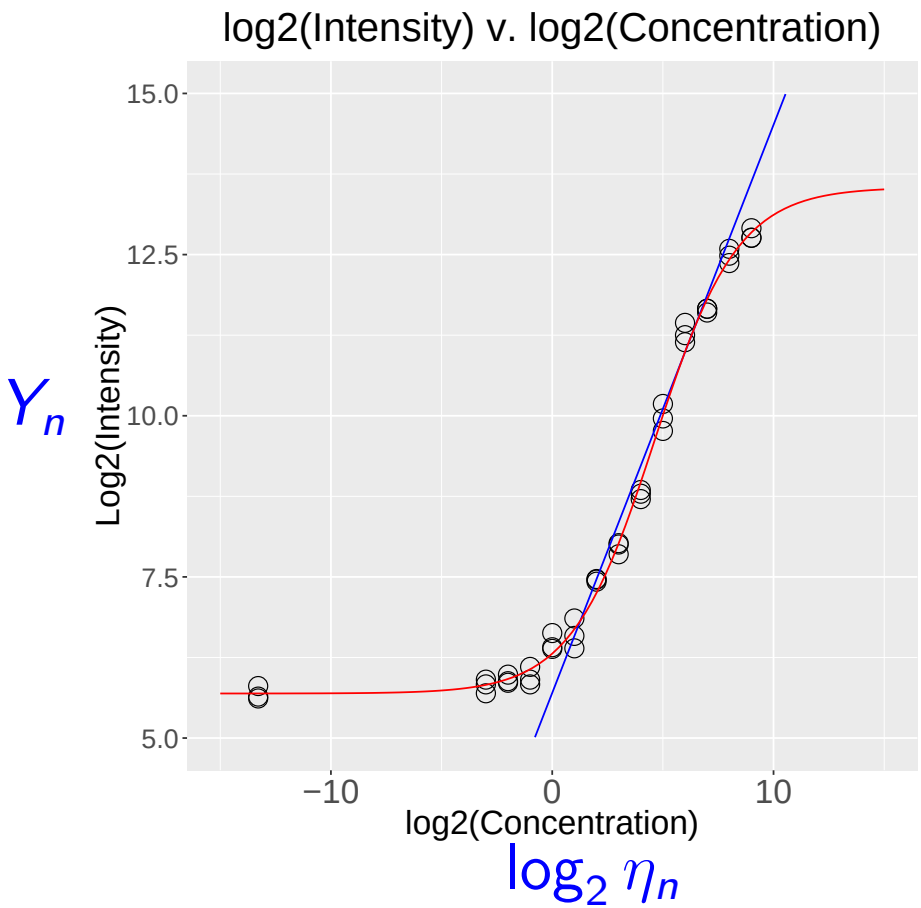


Figure: Relationship between concentration and expression.

Choose Markers From Pure Samples

Goal: Determine which oligos are highly expressed in some cell types but not others.
Method: Some type of differential expression analysis among pure sample expressions.

Contrast with the Literature

1. Our model:

$$Y_n = \theta_n + \gamma \log_2(\eta_n)$$

effective model in the literature:

$$Y_n = \log_2(\eta_n)$$

2. **Our model fit**: background correct and solve directly.
Fitting in the literature: (1) regression or (2) bayesian.

Data Analysis

GSE19830

Data set from Shen-Orr et al. (2010).

1. Affymetrix Rat Genome 230 2.0 DNA microarray
2. Mixtures of brain, liver and lung cells
3. 14 different mixing proportions

GSE11058

Data set from Abbas et al. (2009)

1. HG-U133 Plus 2 Affymetrix Human Genome DNA microarray
2. Mixtures of white blood cell types (Jurkat, IM-9, Raji and THP-1)

Future Work

1. Determine pre-processing normalization. Use negative-controls and apply RUV.
2. Refining how we estimate γ
3. Estimate how many marker genes are appropriate.
4. Refine how we chose marker genes and use negative controls to account for unwanted variation.
5. Estimate proportion of sample that isn't one of K cell types.

Thanks!

1. https://www.wpclipart.com/science/tools/petri_dish.png.html
2. http://www.ditabis.com/resources/images/OEM/Biochip%20rechts.PNG
3. https://en.wikipedia.org/wiki/White_blood_cell#/media/File:SEM_blood_cells.jpg
4. https://sites.duke.edu/apep/module-2-the-abcs-of-intoxication/biology-and-chemistry-connections/dna-transcription-translation-synthesis-of-proteins/
5. http://www.bbc.co.uk/news/special/sci_environment/11/forensics/slideshows/html/img/4_cells.jpg
6. http://ipo.lbl.gov/wp-content/uploads/sites/8/2014/08/22291.png
7. http://learn.genetics.utah.edu/content/labs/