

Pooling in Microarray Experiment

Kenny Ye, Anil Dhundale
SUNY at Stony Brook

Pooling Technique

- Combine RNA samples from more than one subject and apply the combined sample for hybridization
- Good for Comparing gene-expression difference among several groups, (e.g. Cases vs. Controls) to identify genes for further investigation

When Pooling should not be considered

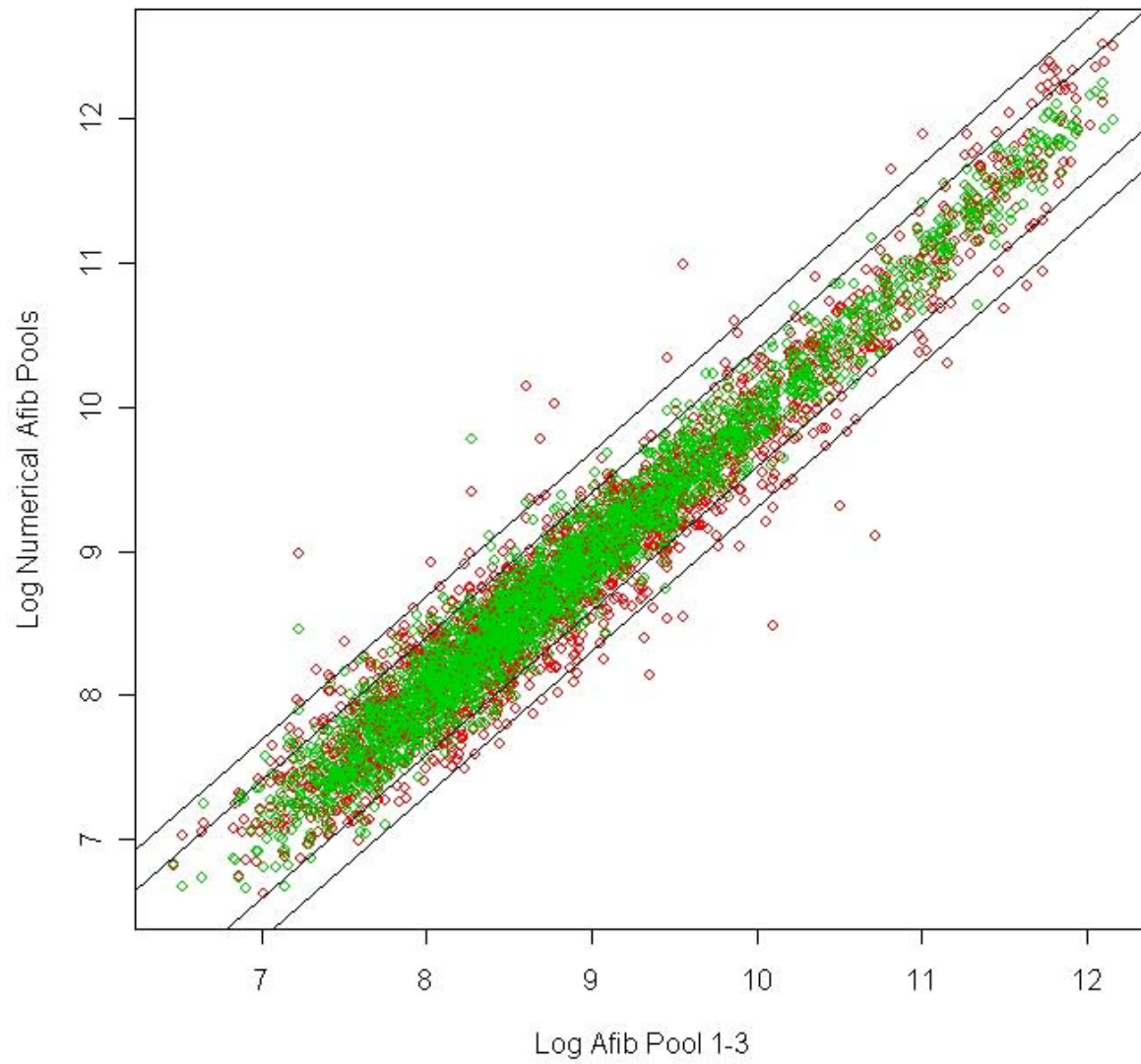
- Goal is to explore relations among the genes, e.g. clustering
- Goal is to establish predictive models based on gene-expression profiles
- Essential individual information permanently lost

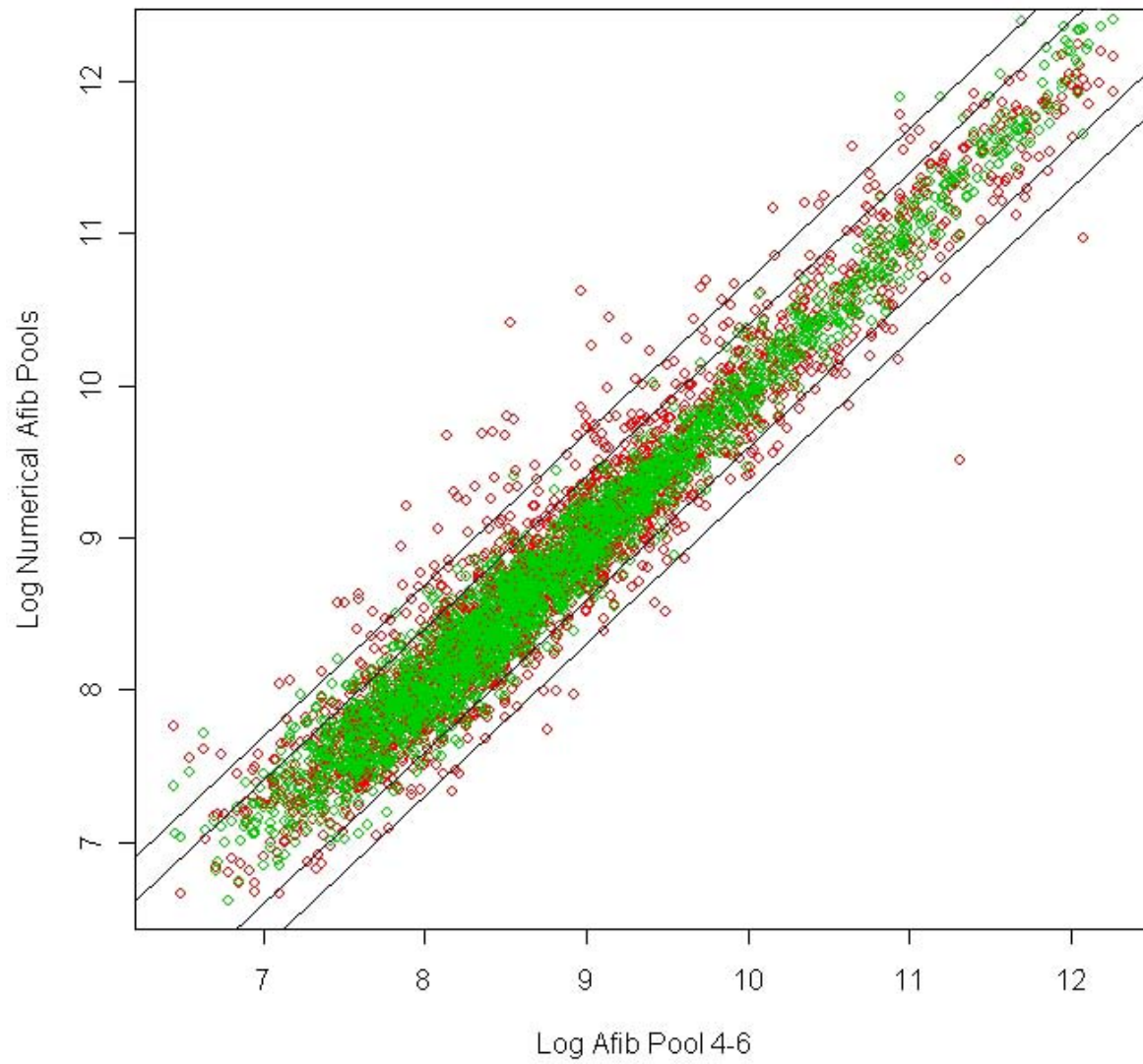
Why Pooling

- RNA not enough from a single sample
 - Tissue from mice and other small animal models
 - Tissue from rare cell types
- Save money
 - Significant cost of microarrays (especially commercially available ones such as Affymetrix)
- Lose Statistical power
 - Replicates of the same pooled sample only gives estimates of measurement error
 - More than two pools of different samples allow estimates of biological variation

Expression level of pooled samples

- Equivalent to average of expression level of individual samples?
- A small experiment with Affymetrix
 - Six subjects into two pools
 - Expression levels obtained for individuals
 - Expression levels obtained for pooled samples





Variance Components

- Biological Variation
 - Between individuals
 - Within individuals
- Measurement Error
 - DNA sample preparation
 - Affymetrix performance
- Normalization adds complication

An Error Model

$$\text{Log}(y_{ijk}) = \mu_i + \epsilon_j + \delta_k$$

- μ : treatment effect
- ϵ : variations that are averaged in the pooled sample (Including biological variations)
- δ : errors that are not averaged in the pooled sample (Including microarray performance)

Power without pooling

- Sample size of two groups: n_1, n_2

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_\epsilon^2 + \sigma_\delta^2}{n_1}\right), \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_\epsilon^2 + \sigma_\delta^2}{n_2}\right)$$

- Non-central t-distribution $T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
 - Non-central parameter

$$\log(\gamma) / \sqrt{(\sigma_\epsilon^2 + \sigma_\delta^2) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- $\gamma = e^{\mu_1 - \mu_2}$: Fold change

Power with pooling

- Pool size : m

$$\tilde{X}_1 \sim N\left(\mu_1, \frac{\sigma_\epsilon^2/m + \sigma_\delta^2}{n_1/m}\right), \tilde{X}_2 \sim N\left(\mu_2, \frac{\sigma_\epsilon^2/m + \sigma_\delta^2}{n_2/m}\right)$$

- Non-central t-distribution $T = \frac{\tilde{X}_1 - \tilde{X}_2}{S_p \sqrt{\frac{m}{n_1} + \frac{m}{n_2}}}$
 - Non-central parameter

$$\log(\gamma) / \sqrt{(\sigma_\epsilon^2 + m\sigma_\delta^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Degrees of freedom: $(n_1+n_2)/m-2$

ε and δ

- Need their variances to find power
- Normal biological variation is highly variable
- Microarray R&R have not been thoroughly studied
 - Good Repeatability
 - Less Reproducibility
 - Microarrays not re-usable, batch-to-batch variation might exist
- Variances can be estimated empirically

A small study

- Normal vs. Atrial fibrillation (most common cardiac arrhythmia)
- Six subjects each, discarded muscle tissues collected during surgeries
- DNA samples from each Afib individuals are tested
- For each group, two pooled samples are tested (n=3)

Moment estimator of σ_ϵ^2 and σ_δ^2

$$\hat{\sigma}_\epsilon^2 = \frac{m_1 - m_2}{2/3}, \quad \hat{\sigma}_\delta^2 = \frac{3m_1 - m_2}{2}$$

$$m_1 = \sum_{i=1}^N \sum_{j=1}^6 \frac{(z_{ij} - \bar{z}_i)^2}{(6-1)N}$$

$$m_2 = \sum_{i=1}^N \sum_{j=1}^2 \frac{(\tilde{z}_{ij} - \bar{\tilde{z}}_i)^2}{N}$$

Assume the variances are constant for all genes or treat the estimators as the average over the genes.

Heart muscle tissues

- MAS 5.0

$$\hat{\sigma}_\epsilon^2 = 0.0621, e^{\hat{\sigma}_\epsilon} = 1.283$$

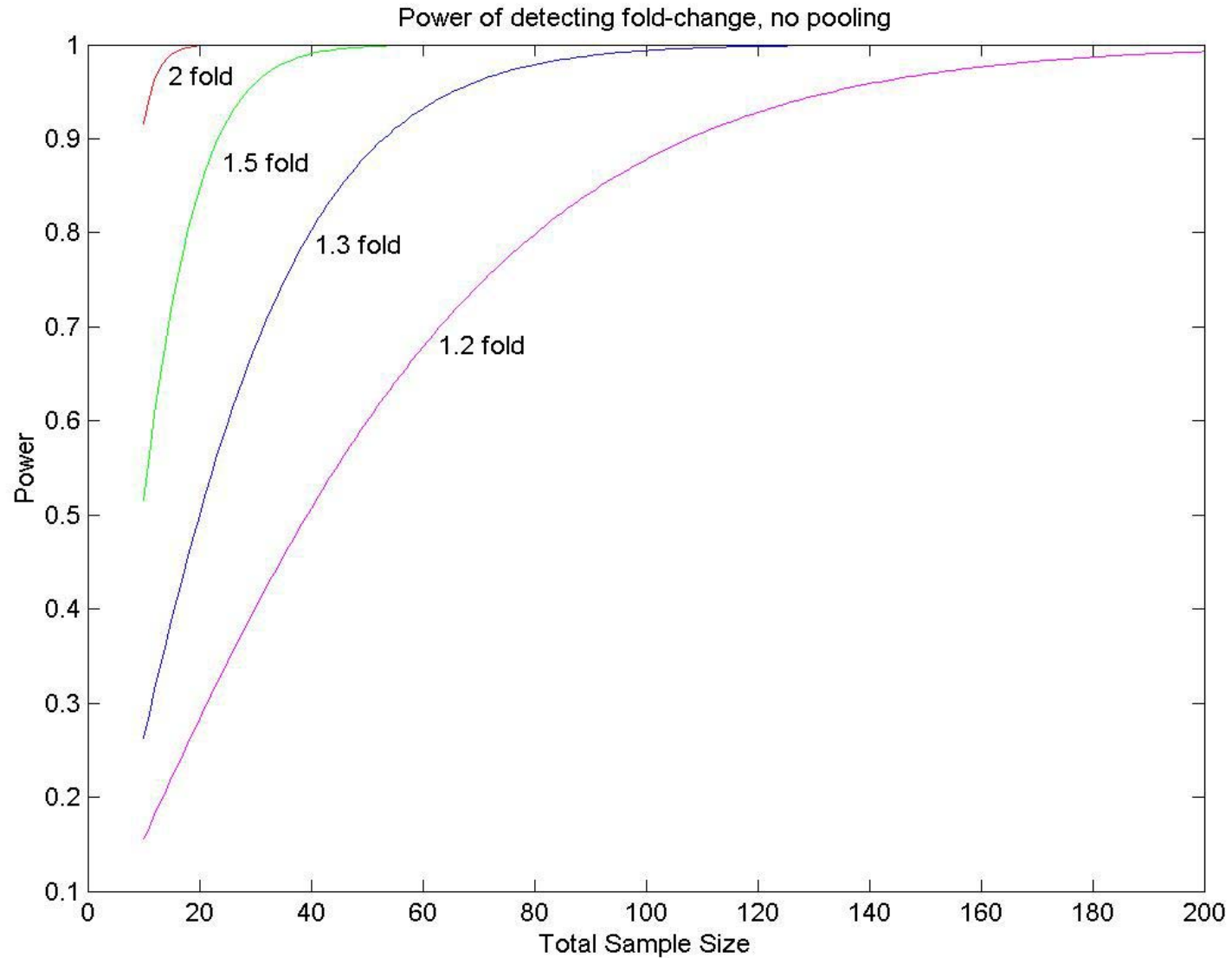
$$\hat{\sigma}_\delta^2 = 0.0161, e^{\hat{\sigma}_\delta} = 1.135$$

- Bioconductor (RMA)

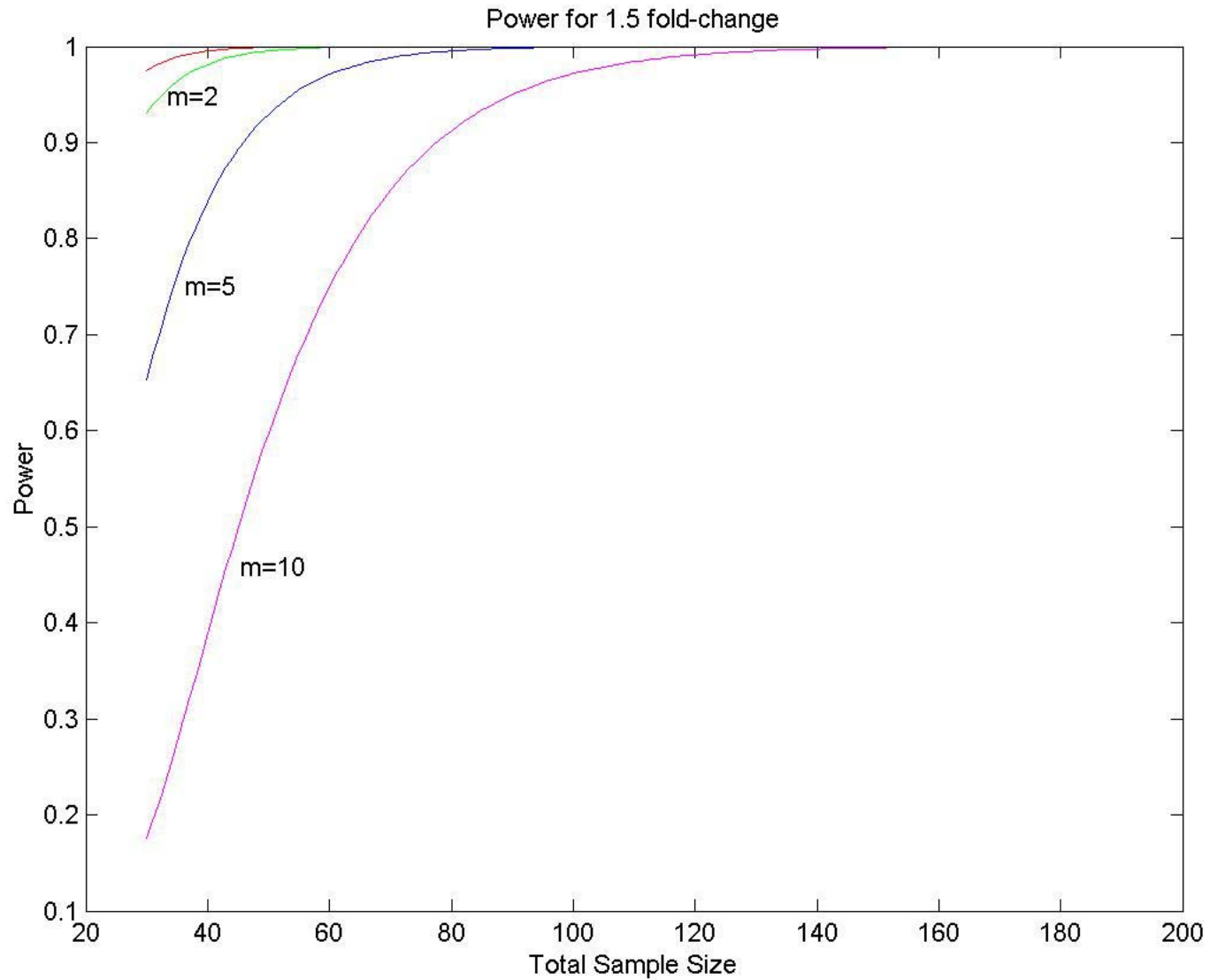
$$\hat{\sigma}_\epsilon^2 = 0.0257, e^{\hat{\sigma}_\epsilon} = 1.174$$

$$\hat{\sigma}_\delta^2 = 0.0161, e^{\hat{\sigma}_\delta} = 1.135$$

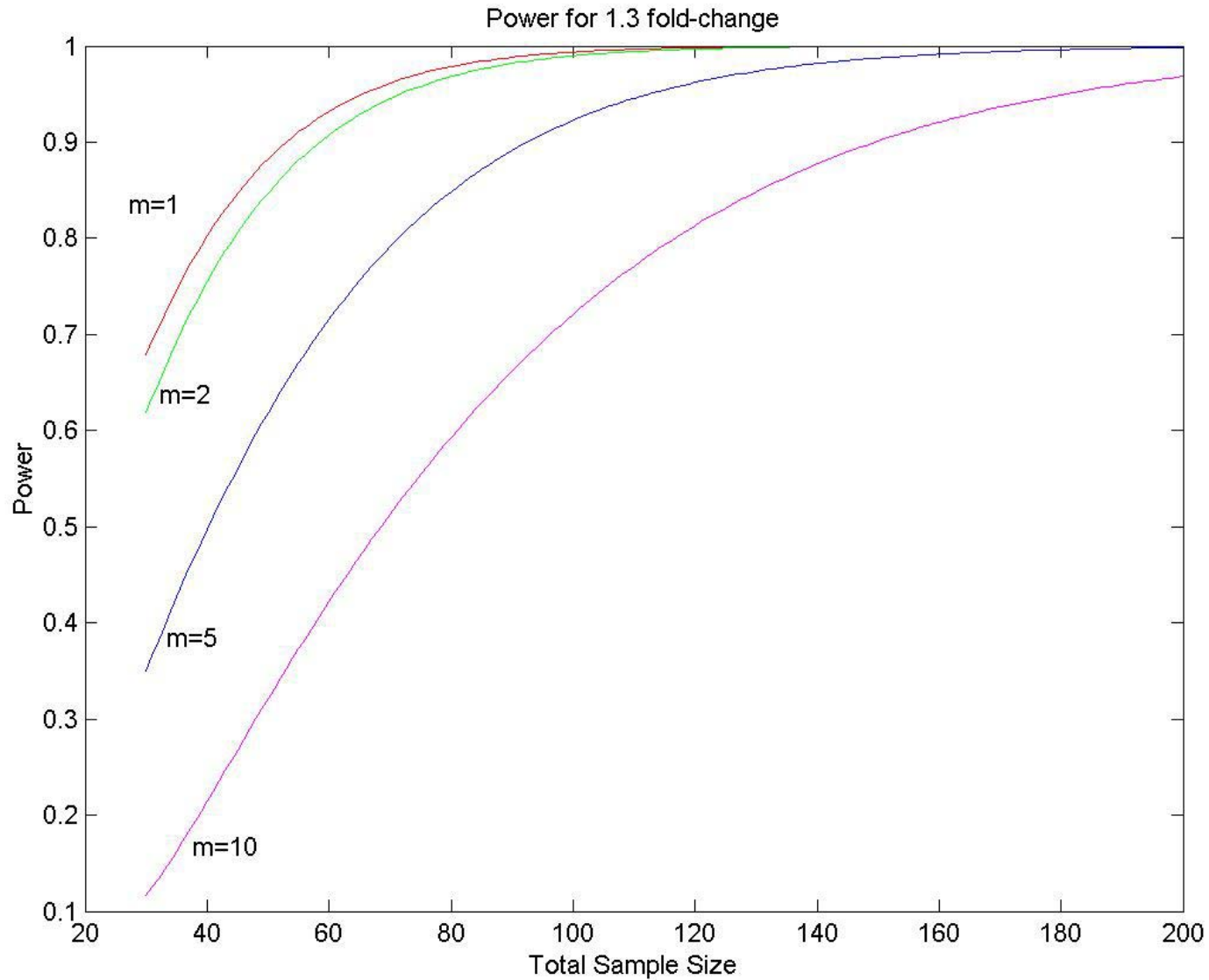
Power Curves without pooling



Pooling Power Curves



Power Curves with pooling



More on the Afib Study

- Do not expect large fold-changes
- Some 1.3-1.5 fold-changes are expected and seen
- Find several genes with statistical significant (without adjustment of multiple test) fold-change and biological interest
- Further confirmation are underway

Pooling with cDNA arrays

- **Balanced Blocked Design**

Treatment	Dye	Block (array)
A	R	1
B	G	1
A	G	2
B	R	2
A	R	3
B	G	3
A	G	4
B	R	4

Conclusion

- When measurement error is small, pooling is very effective with large sample size ($n > 100$). More detailed math can be done in the future.
- More attention are needed for R&R evaluation of the measurement system
- Normal biological variation need to be better understood