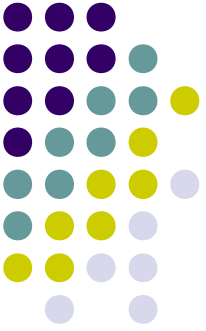
Lecture #4

Power and sample size



Outline

- Importance of sample size
- Statistical terms
- Confidence and power calculations
- Sample size calculations
- Example
- Replicate concordance
  - Individual array results
  - Combined array results

Replicates

- Reliability of statistical inference requires replicate data
  - Hypothesis testing
  - Feature selection
  - Classification
- Variance estimates are highly dependent on an adequate sampling
- Microarrays can be fairly costly, so the minimum number of arrays is optimal for experiment and analysis

Statistical Terms

- Hypothesis tests
  - H_0: the means of two samples are the same (null)
  - H_1: the means of two samples are not the same (alternative)

Rejecting or disproving the null hypothesis – and thus concluding that there are grounds for believing that there is a relationship between two phenomena or that a potential treatment has a measurable effect – is a central task in the modern practice of science

	H_0 is true **Truly not guilty**	**H_1 is true** **Truly guilty**
Accept Null Hypothesis Acquittal	Right decision	Wrong decision Type II Error
Reject Null Hypothesis Conviction	Wrong decision Type I Error	Right decision

- Type I error (false positive – alpha value)
  - Probability of accepting the alternative hypothesis, when the means are the same

- Type II error (false negative – beta value)
  - Probability of accepting the null hypothesis, when the means are different

Statistical Terms (cont.)

- Confidence level
  - Probability of accepting the null hypothesis, when the means are the same
  - 1-α (where α is the size of the test)
  - is used to indicate the reliability of an estimate
- Power
  - Probability of accepting the alternative hypothesis, when the means are different (1-β)
- Sample size determination is made, such that confidence and power can reach predefined values
  - e.g. 95% confidence; 80% power

Calculations (two sample case)[1]

- Power can help estimate the minimum sample size necessary to test for the effect size
- The t-statistic for the hypothesis test:

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2$$

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

- The H_0 distribution for all classes having the same mean is defined as:

$$H_0 : t = \frac{(\bar{y}_1 - \bar{y}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(N-2)$$

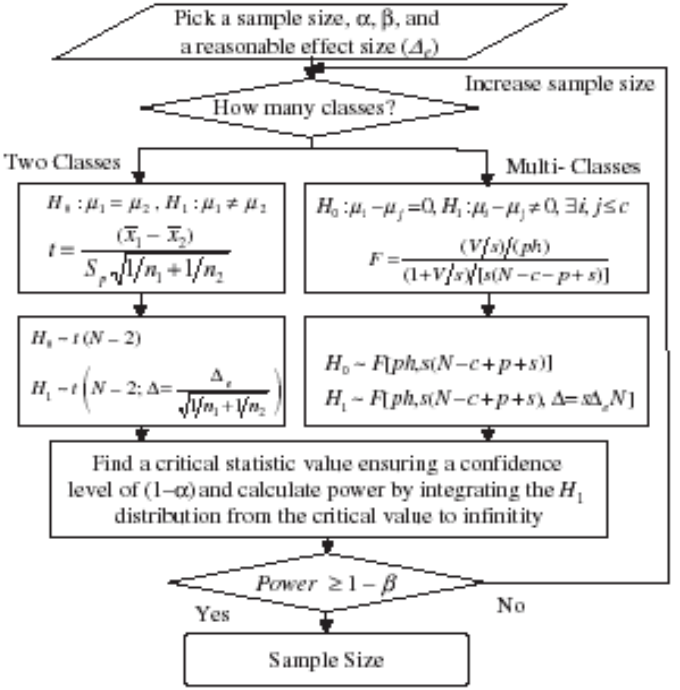
- The H_1 distribution for all classes having different means is defined as:

$$H_1 : t = \frac{(\bar{y}_1 - \bar{y}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t\left(N-2; \Delta = \frac{\Delta_e}{\sqrt{1/n_1 + 1/n_2}}\right)$$

- The effect size is the critical difference between populations that is set in advance:

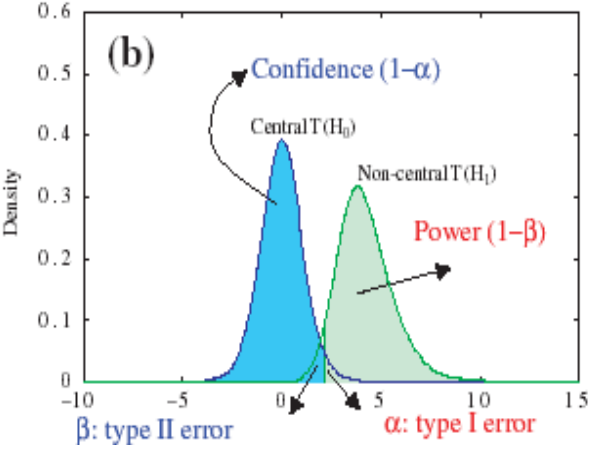
$$\Delta_e = \frac{(\bar{y}_1 - \bar{y}_2)_{crit}}{S_p}.$$

Flow diagram[1]



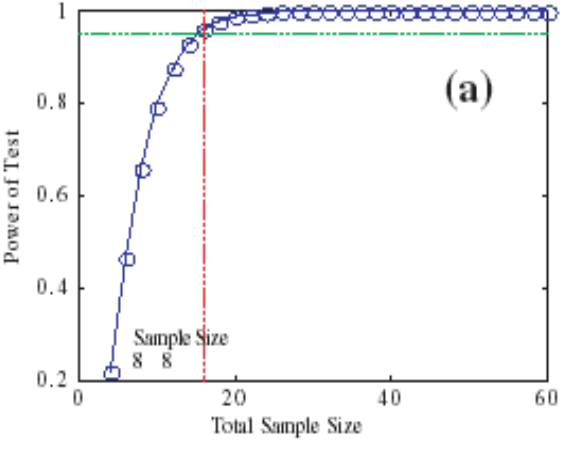
Calculations (cont.)[1]

- Confidence and power are calculated using the distributions of the null and alternative hypotheses[1]
- An initial sample size is assumed, along with a given effect size[1]
- A critical value is identified to ensure a pre-selected confidence level (95% in this example) from the null distribution (blue) [1]
- The power is then calculated by integrating the alternative distribution (green) from the critical value to positive infinity[1]
- If the power falls below the predefined value (1-β), the sample size is increased until the power reaches this threshold[1]



Calculations (cont.)[1]

- Power curve
- Sample size in the plot represent the total number of samples (both classes)
- Assumes that the standard deviation matrix is the same for each class[1]



Sample size calculation

- Sample size is a function of multiple factors
  - Effect size
  - Desired power (1-type II error probability)
  - Confidence level (type I error probability)
  - Variability (CVs)

- There is difficulty in representing variability in microarray data because it tends to vary across genes
  - Effect size is expected difference between classes (e.g. fold change)
  - Power is a pre-determined threshold (e.g. 95%)
  - Confidence level is 1-α (e.g. 99% for a size=.01 test)

- To get a single statistic for n genes, we must assume a single estimator (constant variance) across a microarray
  - This is unrealistic for each gene to have similar variability

- For calculating the power in at least h genes that are thought be regulated between classes, the binomial probabilities must be summed
  - $(1-\beta) = \sum x!/(h!(x-h)!)(1-B)hB(x-h)$
    where h=# regulated genes detected
    x=# of actually regulated genes
    B= type II error

Sample size calculation – two sample, two-sided test

given: z_α= critical value at specific size of test

$k = n_2/n_1$ projected ratio of 2 sample sizes

σ_1^2 & σ_2^2= sample variances

μ_1 & μ_2= sample means

Power $(1-\beta) = \Phi \left[-z_{1-\alpha/2} + (\sqrt{n_1}|\mu_1 - \mu_2|) / (\sqrt{\sigma_1^2 + \sigma_2^2/k}) \right]$

$n_1 = \{ (\sigma_1^2 + \sigma_2^2/k)(z_{1-\alpha/2} + z_{1-\beta})^2 \} / |\mu_1 - \mu_2|^2$

$n_2 = \{ (k\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2 \} / |\mu_1 - \mu_2|^2$

assuming near equal sample sizes

$n = \{ (\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2 \} / |\mu_1 - \mu_2|^2$

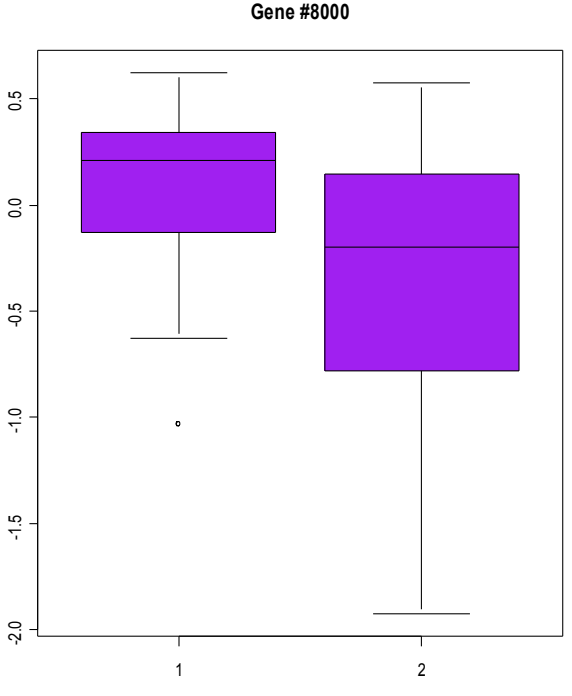
Example

- Multiple gene power calculations are beyond the scope of this course
  - We can calculate sample sizes and power based on single gene statistics

- Utilizing only a few selected genes, we can get an idea of how many replicates would be required to detect a specified mean difference between classes

Example with colon data

gene #8,000 boxplots




```
 Welch Two Sample t-test

data:  x and y
t = 2.226, df = 32.726, p-value =
0.03302
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 0.03484702 0.77814245
sample estimates:
  mean of x   mean of y
 0.06089474 -0.34560000
```

Sample size

- The t-test finds a significant result (p=0.03) at a difference of ~0.41 between the means
- To detect a 3 fold difference (log scale) for gene #8,000 with 80% power and confidence=95%
  - Data is z-score normalized, so detectable fold change is difficult to infer

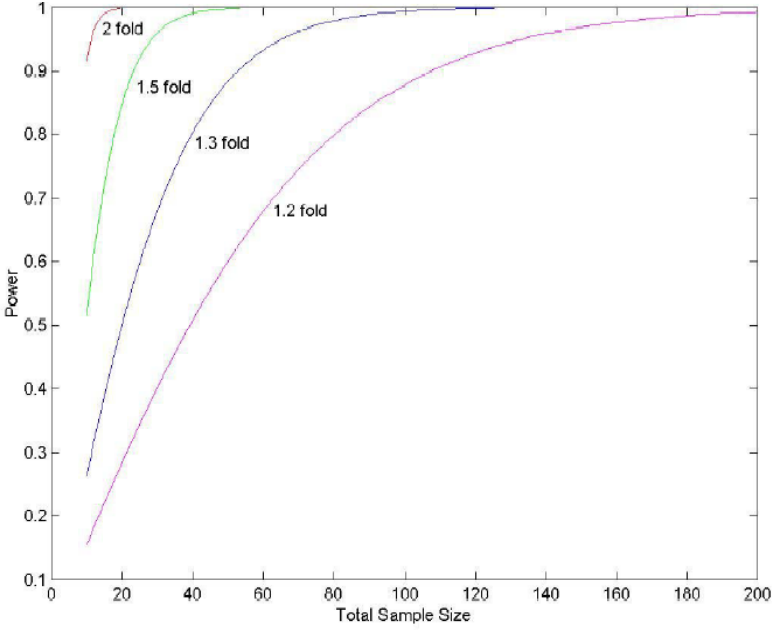
```
Two-sample t test power calculation

              n = 7
          delta = 1.099
             sd = 0.680
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Power curves[3]

- Assuming constant variance across all genes (false assumption) at ~ 1.44, the replicate numbers can be represented by the calculated power at the specified fold change detections



False Discovery Rate as opposed to power and confidence for sample size determination

- The p-value is associated with specificity of a test
  - p-value<0.05 means that specificity = 0.95
- Multiple testing procedures can be too conservative
  - Will discuss this concept in later lectures
- False discovery rate (FDR) is proposed as an alternative to simple p-values
  - FDR is expected proportion of FPs among declared significant results
  - e.g: if 100 genes are declared differentially expressed, and set the FDR to 0.10, 10 of these genes will be FPs

Properties of the FDR

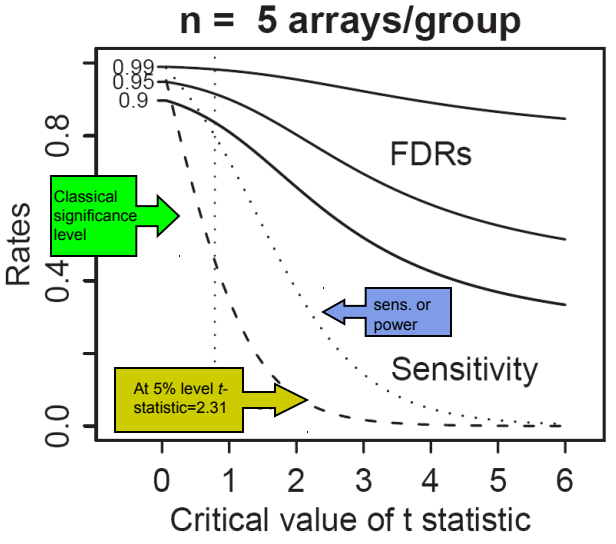
- The FDR relies on:
  - The proportion of truly differentially expressed genes
  - Distribution of the true differences
  - Variability
  - Sample size (only factor under the experimentalists control)

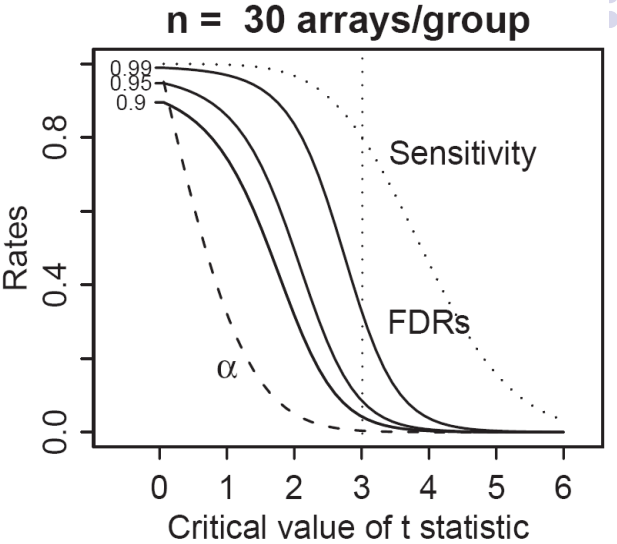
10,000 gene example[5]

	Test result: non differentially regulated	Test result: differentially expressed	Total
True: non differentially expressed	A = 9025	B = 475	9500
True: differentially expressed	C = 100	D = 400	500
Total	9125	875	10,000

- FP rate (1-specificity) = B/(A+B) = 5%
- Sensitivity = D/(C+D) = 80%
- FDR = B/(B+D) = 54%
- FNR = C/(C+D) = 20%
  - Over half of genes that hypothesis test says are differentially expressed, are not
- Using significance test, 80% power and 95% confidence gives a high FDR
  - Can reduce FPs by reducing p-value threshold

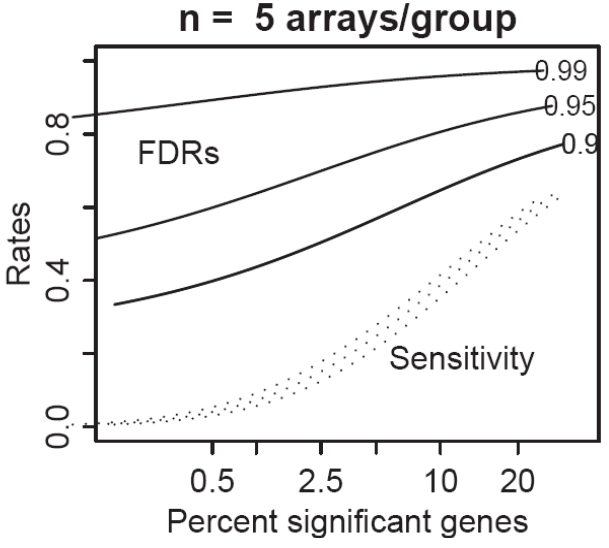
FDR curves for non-differentially expressed genes[5]

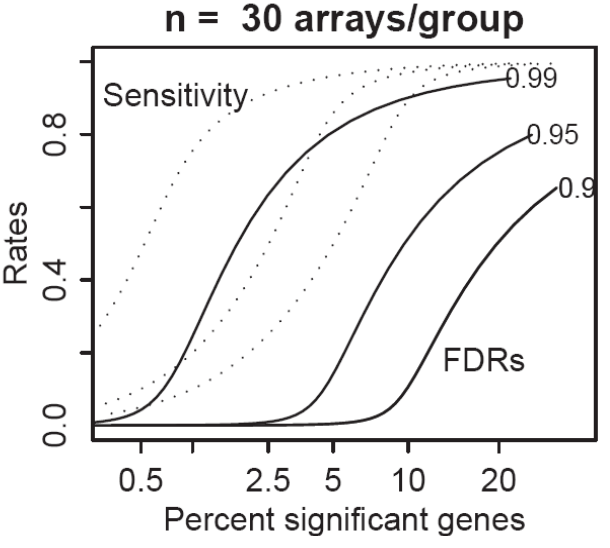




- Each curve is labeled by the percentage of truly non-differentially expressed genes
- In experiments with small *n*, where the percentage of non-differentially expressed genes is expected to be high, FDR can be high, even when using large *t*-statistic critical values
  - *e.g.* if the proportion of non-differentially expressed genes = 0.90, this provides a 60% FDR, with a sample size of 5
- When *n* is increased to 30 (per group), FDR improves
  - e.g. at a *t*-statistic critical value of 3 (p-value=0.004), there is <10% FDR, if 0.90 of genes are non-differentially expressed; sensitivity ~ 0.80

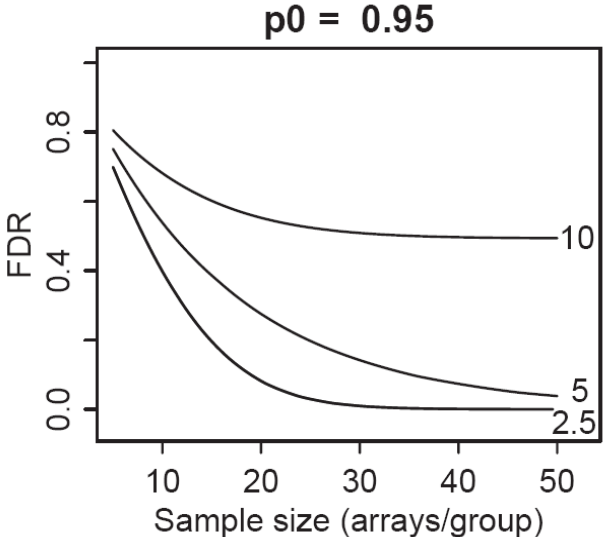
FDR curves for differentially expressed genes[5]

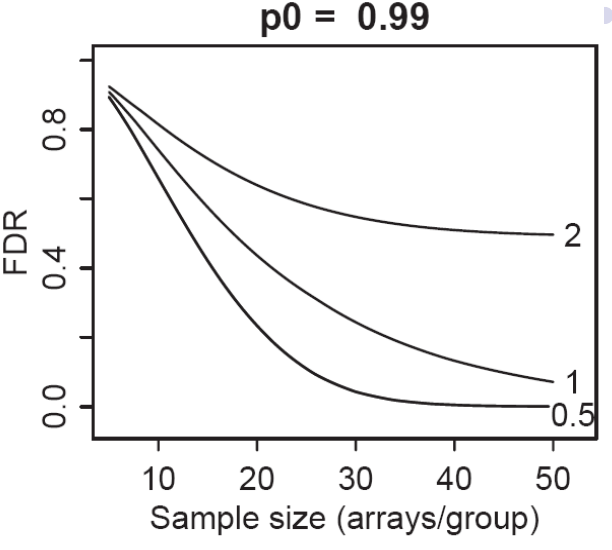




- Assume
  - Genes with top 1% highest absolute *t*-statistics are truly differentially expressed
  - Proportion of non-differentially expressed genes = 0.99
- FDR > 80% for *n*=5 (per group)
- As *n* is increased, FDR increases

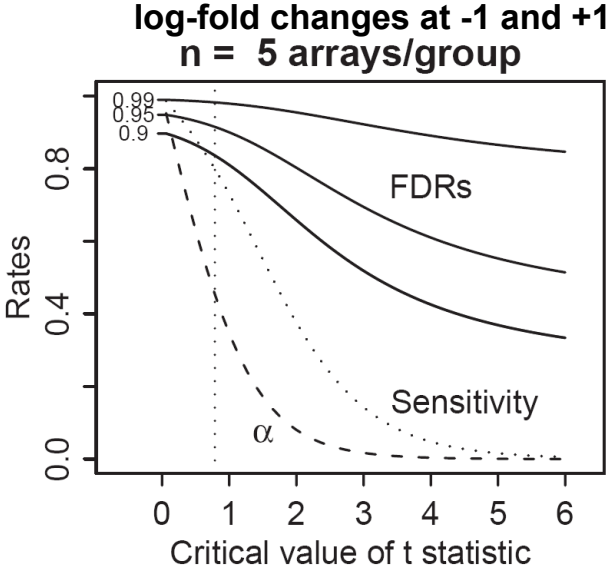
FDR curves vs. *n* for non-differentially expressed genes (p0)[5]

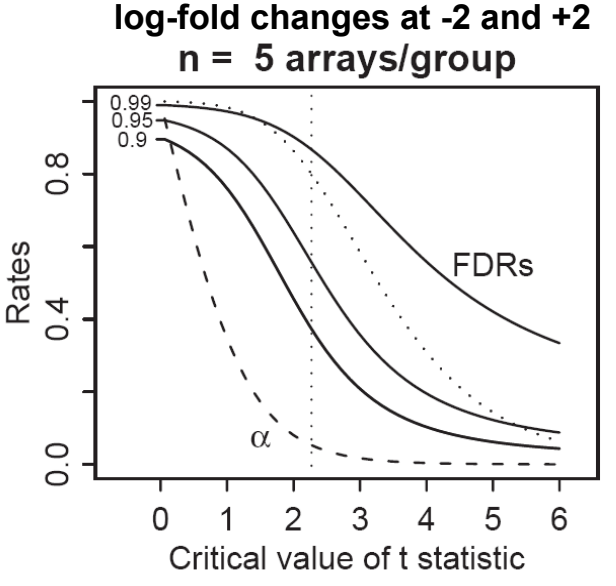




- Each curve is labeled with a fixed percentage of truly differentially expressed genes
- If the number of differentially expressed genes is known to be around a certain amount for an array, increasing the probes will only increase the proportion of non-differentially expressed genes
  - This will result in larger FDRs

Increase log-fold changes for truly differentially expressed genes[5]





- With increased fold changes FDR is reduced

References

- [1]Hwang D, Schmitt W, Stephanopoulos G, and Stephanopoulos G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*. **18**, 1184-1193.
- [2]Lee M, Kuo F, Whitmore G, and Sklar J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *PNAS*. **97**, 9834-9839.
- [3]SUNY
  - http://www.ams.sunysb.edu/~kye/talks/PoolingPoster.pdf
- [4]Leeds University
  - http://www.amsta.leeds.ac.uk/~edwin/m1830/lectures/m1830_7.htm
- [5]Pawitan Y, Michiels S, Koscielny S, Gusnanto A, and Ploner A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 21(13): 3017-3024.

R Code

```
# import eisen data
dat <- read.table("eisen.txt",header=T)
dimnames(dat)[[1]] <- as.character(dat[,1])
dat <- dat[,-1]
dat <- as.data.frame(dat)

# import annotation file
ann <- read.table("eisenClasses.txt",header=T)

# subset dat by samples of interest
cl <- as.character(ann[,2])
dat <- dat[,cl]

# two classes of DLBCL
gc <- cl[1:19]
act <- cl[20:39]

# split up classes and look at both samples for gene #8000
x <- as.numeric(dat[8000,gc])
y <- as.numeric(dat[8000,act])

# remove "NAs"
x <- x[!is.na(x)];    y <- y[!is.na(y)]

# plot both samples
xy.list <- list(x,y)
boxplot(xy.list,col='purple',main='Gene #8000')
```

R Code

```
# calculate two-sample Welch's t-test (unequal variances) between normal and tumor for gene #8000
xy.ttest <- t.test(x, y, alternative ="two.sided",paired = FALSE, var.equal = FALSE,conf.level = 0.95)

# determine sd of each group and choose max
x.sd <- sd(x)
y.sd <- sd(y)

# calculate number of replicates to detect 3 fold change (1.1 on log scale) at 80% power
power.t.test(delta=log(3),sd=y.sd,power=.8)
```

Backup slides

Replicate concordance

- An alternative method of viewing the effect of replicate experiments is to estimate the concordance of various parameters in each replicate individually
  - Probability of detecting a gene
- This method can give you insight into the similarity between each replicate independently
- Then, observe how this changes when the replicates are pooled

Statistical Model[2]

- We assume a model for the detection of a particular gene *g (g=1…,G)* in replicate *j (j=1…,J)*, subject to the following considerations[2]:
  - Expression of a gene is taken as the log ratio Y_{gj}
  - Y_{gj} has two distinct distributions:
    - Gene *g* is not in the sample tissue, distributed as $N(\mu_{Uj}\sigma^2_{Uj})$, where U refers to being unexpressed

      Probability density function is $Y_{gj}|\boldsymbol{E.bar}_g$ given by $f_{Uj}(y)$
    - Gene *g* is in the sample tissue, distributed as $N(\mu_{Ej}\sigma^2_{Ej})$, where E refers to being expressed

      Probability density function is $Y_{gj}|E_g$ given by $f_{Ej}(y)$
  - Prior probability of observing a gene is $Pr\{E_g\}=p$

Statistical Model (cont.)[2]

- The log-ratio, Y_{gj} for replicate *j* will be distributed according to the following mixture model

  $f_j(y) = pf_{Ej}(y) + (1-p)f_{Uj}(y)$

- Manipulating the equation above gives the posterior probabilities for whether gene *g* is expressed, based on the expression value $Y_{gj} = y$

  $Pr\{E_g \mid Y_{gj} = y\} = pf_{Ej}(y) / f_j(y)$

Model Parameters[2]

- Using the following parameters, we can estimate the posterior probabilities from the two previous equations
  - p = prior prob. of observing a gene (controlled experiment showed 32/288 (0.111) as expressed)
  - μ_{Uj} & σ^2_{Uj} = mean and variance for gene g being unexpressed
  - μ_{Ej} & σ^2_{Ej} = mean and variance for gene g being expressed
  - We would expect a large difference between the 2 mean parameters ($\mu_{Ej} > \mu_{Uj}$)

- First solve the MLE (maximum likelihood estimates) of the parameters above in each of the 3 replicates alone and see how similar they are
  - MLE is a method of determining the values of n unknown variables, such that the function is maximized
  - We solve for these parameters in the first equation and compare how they differ between replicates

Equation #1 Model Parameters[2]

Table 1. Separate analysis for each experimental replicate

	Replicate		
Parameter	$j = 1$	$j = 2$	$j = 3$
p	0.285	0.124	0.274
μ_{U_j}	0.384	0.410	0.442
μ_{E_j}	0.968	2.203	1.233
$\sigma^2_{U_j}$	0.070	0.076	0.062
$\sigma^2_{E_j}$	1.186	0.114	1.079

Parameter estimates of the mixed normal model (Eq. **1**).

- Replicate #2 is fairly different for 3 parameters
- The approximations of *p* in j=1 and 3 are too large, as compared to the controlled study (0.111)
- These 3 replicates show the differences in replicate mean and variance between identical samples

Equation #2 Posterior Probabilities[2]

Table 2. Posterior probability of expression in sample tissue

	Replicate 1		Replicate 2		Replicate 3	
Gene g	$Y_{g1} = y$	$Pr\{\mathscr{E}_g \mid Y_{g1} = y\}$	$Y_{g2} = y$	$Pr\{\mathscr{E}_g \mid Y_{g2} = y\}$	$Y_{g3} = y$	$Pr\{\mathscr{E}_g \mid Y_{g3} = y$
1	2.043	1.0000	1.6804	0.9993	2.6251	1.0000
2	0.6549	0.1356	0.5551	0.0000	0.6874	0.1134
3	0.4940	0.0877	0.3791	0.0000	0.5065	0.0682
17	0.6646	0.1404	0.2662	0.0000	1.7204	1.0000
18	2.4397	1.0000	2.3081	1.0000	2.2481	1.0000
19	2.2331	1.0000	2.0549	1.0000	2.5257	1.0000

Log ratios $Y_{gj} = y$ and estimates of posterior probabilities $Pr\{\mathscr{E}_g \mid Y_{gj} = y\}$ for a few illustrative genes g, for replicates $j = 1, 2, 3$.

- The posterior probability that a gene is expressed is at a threshold of 0.5
  - $E_g \mid Y_g > 0.5$ (gene is expressed);
  - $E_g \mid Y_g < 0.5$ (gene is not expressed);
- Gene #17 has very different estimate of posterior probability (prob. of being expressed) in replicate 3, as compared to 1 and 2.

Single Replicate vs. Combined

- The differences in both model parameters and posterior probabilities (prob. that the gene is expressed) are significant when looking at individual replicates

- How can these estimates be improved when utilizing combined replicate data?
  - Model parameters
  - Misclassification percentages (stratified by replicate combinations)

Combined Data Model Parameters[2]

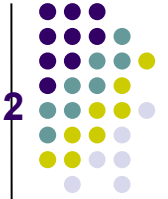


Table 4. Analysis of the combined data from all three replicates

Parameter	Estimate	Est. Std. Err.
p	0.118	0.013
μ_U	−0.204	0.009
μ_E	1.524	0.058
σ_U^2	0.044	0.003
σ_E^2	0.126	0.036

Parameter estimates of the mixed normal model (Eq. **5**) derived from the estimated main effects for genes $\hat{\alpha}_g$.

- Prior probability is more consistent with controlled study results
  - 0.118 vs. 0.111
- Difference between means is large
  - 1.524 >> -0.204

Combined Data Misclassification Rates[2]

Table 5. Misclassification percentages for different combinations of replicates

Classification Outcome	Combination of Replicates						
	(1)	(2)	(3)	(1, 2)	(1, 3)	(2, 3)	(1, 2, 3)
False positive, %	8.3	1.4	9.0	1.0	2.1	0.7	0.7
False negative, %	0.3	0.0	0.0	0.3	0.3	0.0	0.0
Misclassified, %	8.7	1.4	9.0	1.4	2.4	0.7	0.7

- Misclassification rates are highest in individual replicates 1 and 3
- All three replicates provide the lowest misclassification rate