UCI Genomics
High Throughput Facility

Statistical Analysis of Differential Expression

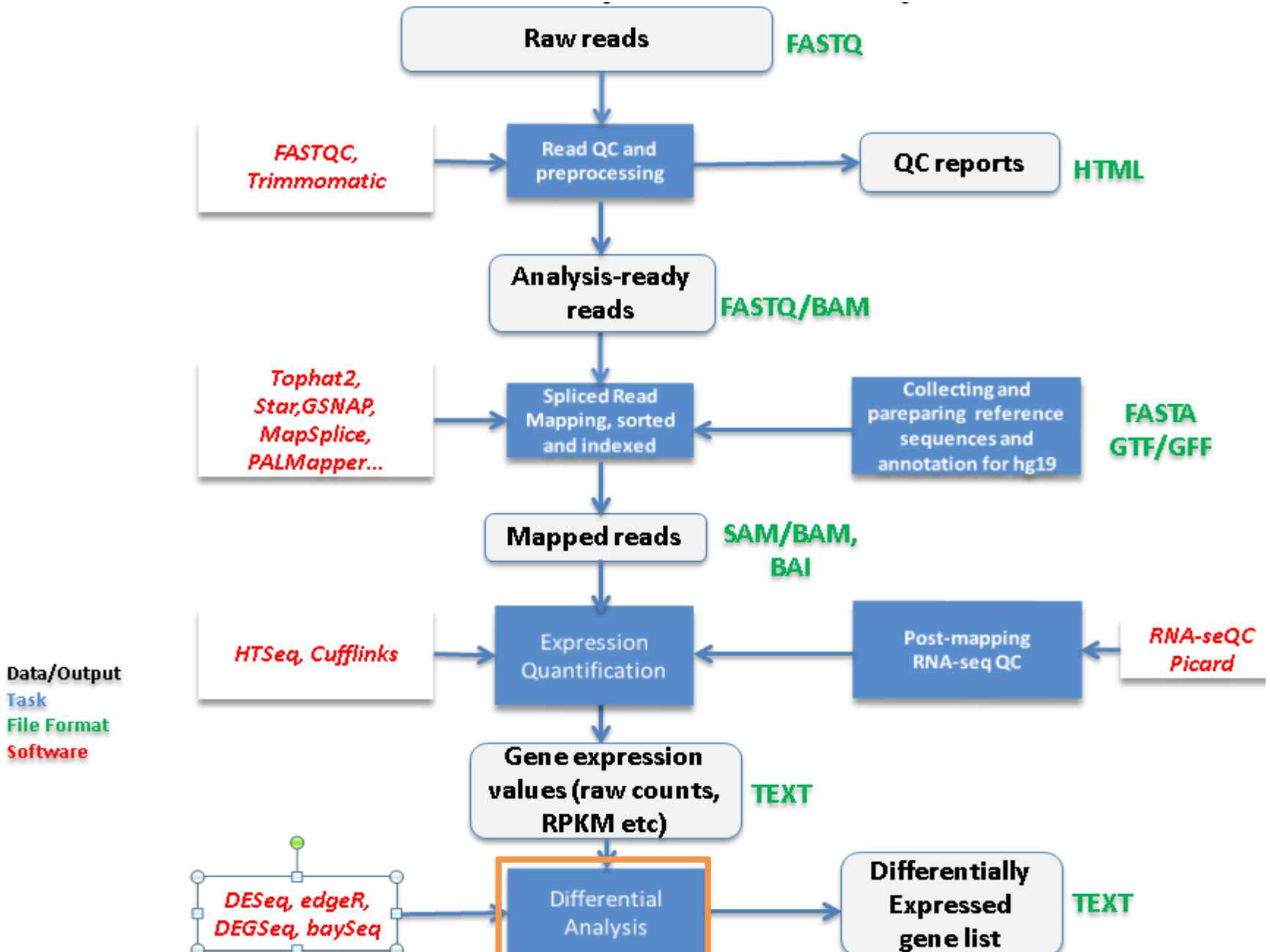
Jenny Wu

Director of Bioinformatics

Genomics Research and Technology Hub

Chao Family Comprehensive Cancer Center

RNA-Seq Pipeline for **Gene Level** DE



What does differential expression (DE) mean?

- A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant, i.e. whether or not the difference is greater than what would be expected just due to random variation.

Differential Expression

Statistical Framework:

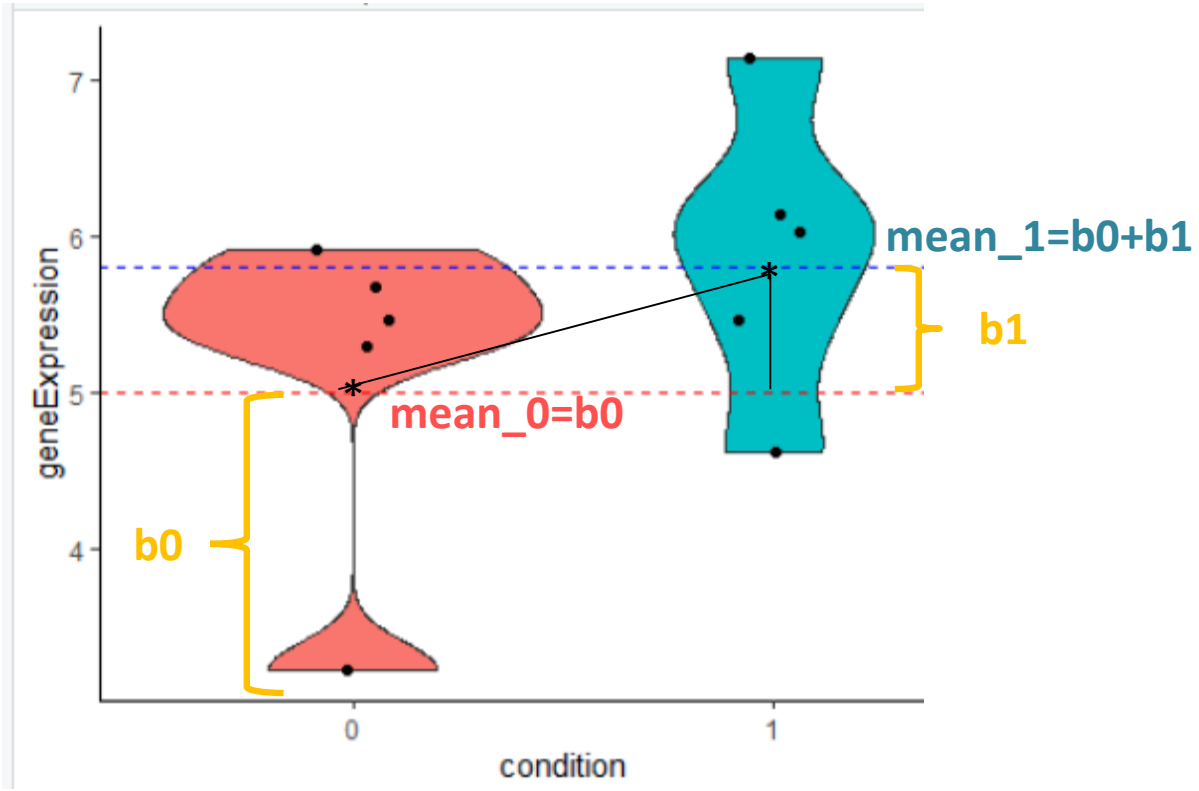
	Sample 1	Sample 2	…	Sample N
Feature 1	K_{11}	K_{12}	…	K_{1N}
Feature 2	K_{21}	K_{22}	…	K_{2N}
…			…	
Feature p	K_{p1}	K_{p2}	…	K_{pN}

- Goal: test for differential expression across different sample groups.
- K_{ij} is discrete positive, skewed, large dynamic range.
- p$\gg N$ small number of replicates.
- Using generalized linear model (GLMs) to detect differential expression:

$$\boldsymbol{log2(q_{ij}) = \beta_0 + \beta_1 x_{1i}}\text{, where } \boldsymbol{x_{1i} = 0\ (control)\ or\ 1\ (diseased)}$$

- For EACH expressed gene (feature), do hypothesis testing: H_0: $\beta_1 = 0$

Classic Framework for Differential Analysis: Regression

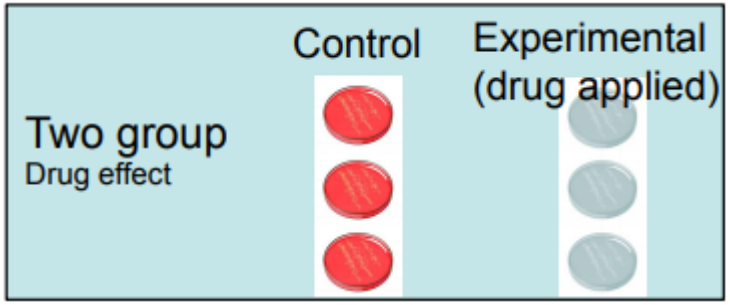


Condition: x = 0 $(control)$ *or* 1 $(diseased)$

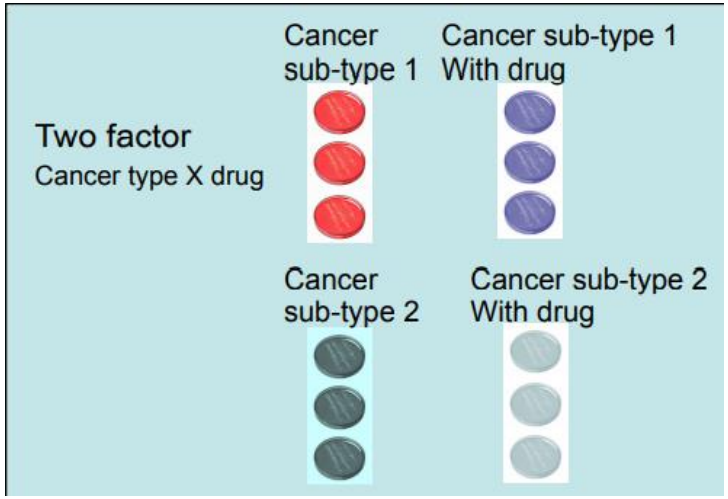
Gene Expression: y = $log2(q_i) = b_0 + b_1 x$

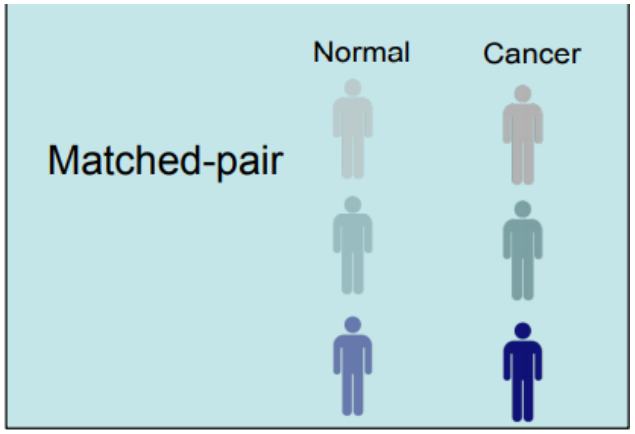
Experimental Design

✓Simple Design



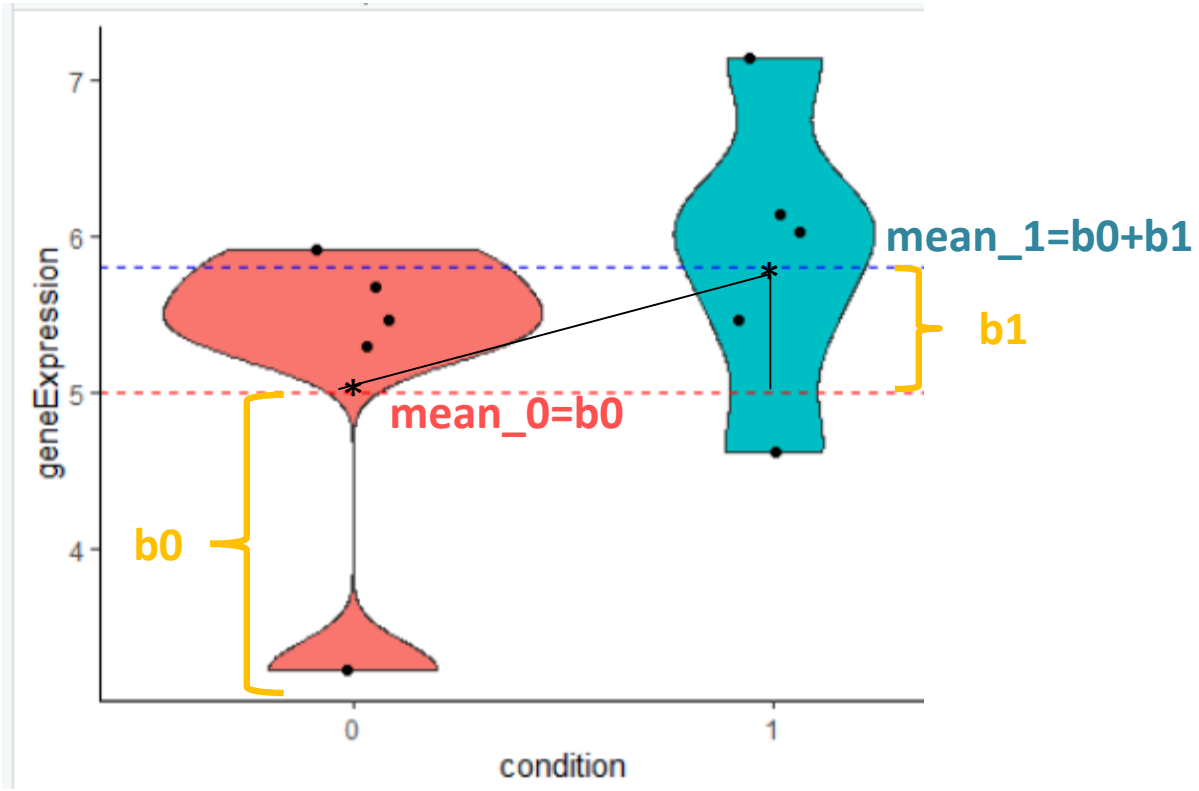
✓Complex Design





Statistical analysis of data with complex design can be done using R based tools such as DESeq2 and limma

Classic Framework for Differential Analysis: Generative Approach: Regression



Condition: x = 0 $(control)$ *or* 1 $(diseased)$

Batch, gender

Gene Expression: y = $log2(q_i) = b_0 + b_1 x$

$$y = log2(q_i) = b_0 + b_1 x_1 + b_2 x_2$$

GLMs in RNA-seq: DESeq2 Implementation

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

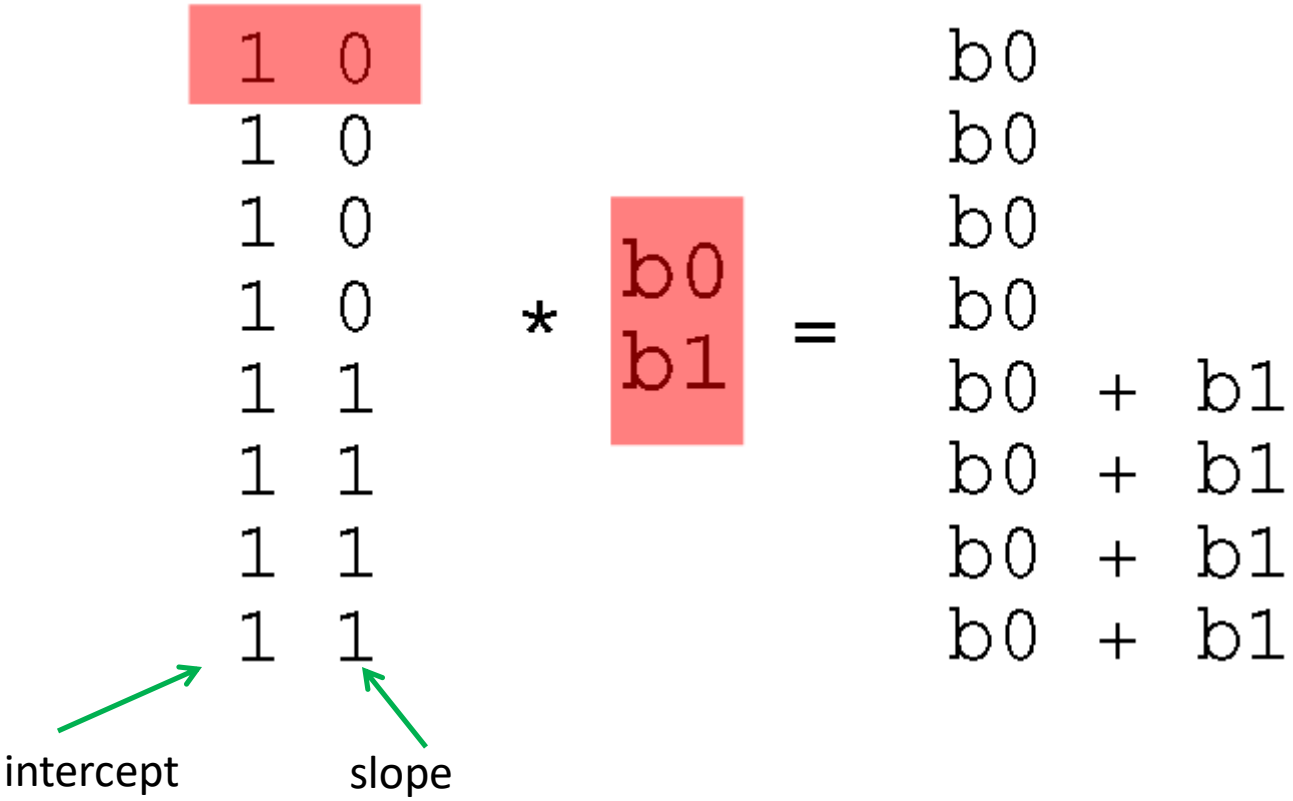
$$\log_2(q_{ij}) = x_{j*}\vec{\beta}_i$$

K_{ij}	counts of reads for gene i, sample j
μ_{ij}	fitted mean
α_i	gene-specific dispersion
s_j	sample-specific size factor
q_{ij}	parameter proportional to the expected true concentration of fragments
x_{j*}	the j-th row of the design matrix X
$\vec{\beta}_i$	the log fold changes for gene i for each column of X

https://bioconductor.org/packages/release/bioc/manuals/DESeq2/man/DESeq2.pdf

Design Matrix *X*

$$X \quad * \quad \beta = \quad log2(\mu)$$



Hypothesis Testing in RNA-Seq

Null hypothesis (H_0)

- The experimental condition r has no influence on the expression of the gene under consideration:

$$\mu_{\rho_1} = \mu_{\rho_2}$$

Alternative hypothesis (H_0)

$$\mu_{\rho_1} \neq \mu_{\rho_2}$$

Hypothosis Testing with GLM

- H_0: $\beta_1 = 0$
- Likelihood Ratio Test

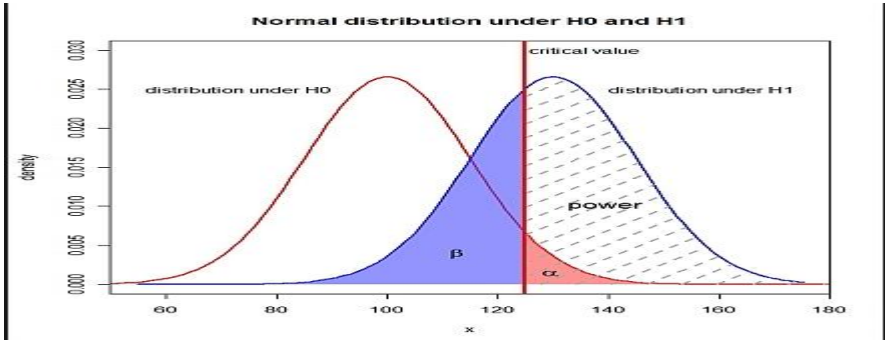
$$D = -2 * \log\left(\frac{L_0}{L_{alt}}\right)$$

Under H_0, $D \sim \chi^2$ and p value can be calculated using χ^2 distribution.

Error Types in Hypothesis Testing

Reality

Decision	Truth (H_1)	Truth (H_0)	Total
Test result $reject\ H_0$	True Positive	False Positive (Type I error α)	R
Test result $do\ not\ reject\ H_0$	False Negative (Type II error β)	True Negative	m-R
Total	m_0	m_1	m



Differential Analysis: Two Group

```
> library(DESeq2)
> sampleFiles <- c("C1_R1.counts.txt", "C1_R2.counts.txt", "C1_R3.counts.txt",
"C2_R1.counts.txt", "C2_R2.counts.txt", "C2_R3.counts.txt")
> sampleCondition <- factor(substr(sampleFiles, 1, 2))
> sampleTable <- data.frame(sampleName=sampleFiles, fileName=sampleFiles,
condition=sampleCondition)
> sampleTable
         sampleName          fileName condition
1 C1_R1.counts.txt C1_R1.counts.txt        C1
2 C1_R2.counts.txt C1_R2.counts.txt        C1
3 C1_R3.counts.txt C1_R3.counts.txt        C1
4 C2_R1.counts.txt C2_R1.counts.txt        C2
5 C2_R2.counts.txt C2_R2.counts.txt        C2
6 C2_R3.counts.txt C2_R3.counts.txt        C2
dds <- DESeqDataSetFromHTSeqCount(sampleTable=sampleTable, directory=".",
design=~condition)
dds <- DESeq(dds)
results <- results(dds)
results <- results[order(results$FDR), ]
```

http://www.bioconductor.org/packages/2.12/bioc/html/DESeq2.html

Output from DESeq2

```
log2 fold change (MLE): treatmentTR.genotypeMT
Wald test p-value: treatmentTR.genotypeMT
DataFrame with 6 rows and 6 columns
         baseMean log2FoldChange      lfcSE      stat       pvalue         padj
        <numeric>      <numeric>  <numeric> <numeric>    <numeric>    <numeric>
Crabp1  3885.6382      0.5944463 0.03712528  16.01190 1.055350e-57 1.387785e-53
C1qtnf3 1533.2677      0.9165233 0.06142923  14.91999 2.443265e-50 1.606447e-46
Zcchc5   912.0010      0.8111819 0.06129395  13.23429 5.562018e-40 2.438018e-36
Fabp4    988.1449     -1.9359619 0.14925925 -12.97046 1.799357e-38 5.915385e-35
Aqp1     679.7191      1.2630634 0.09789543  12.90217 4.375994e-38 1.150886e-34
Tgfbi   1187.7486      0.5448491 0.04634381  11.75668 6.525354e-32 1.430140e-28
```

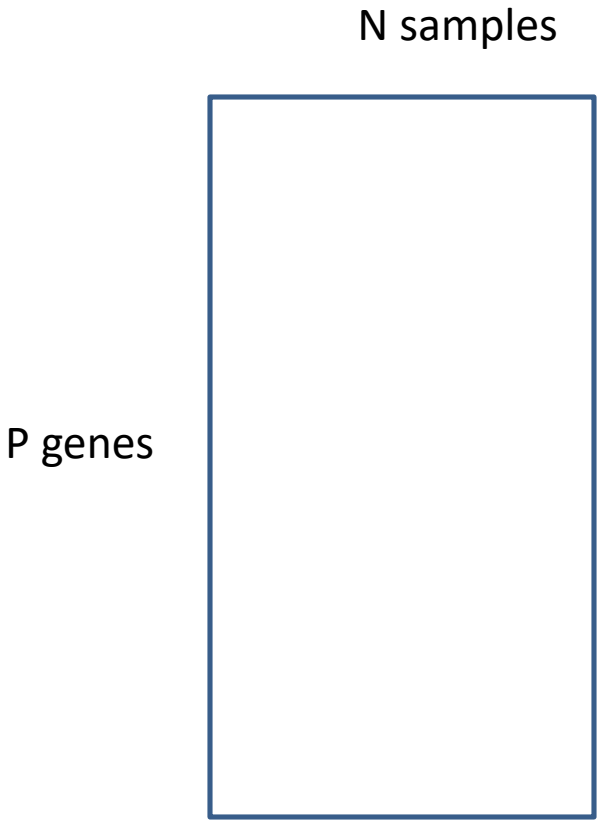
Differential Analysis: Complex Design

```
> library(DESeq2)
> sampleFiles <- c("C1_R1.counts.txt", "C1_R2.counts.txt", "C1_R3.counts.txt",
"C2_R1.counts.txt", "C2_R2.counts.txt", "C2_R3.counts.txt")
> sampleCondition <- factor(substr(sampleFiles, 1, 2))
> sampleTable <- data.frame(sampleName=sampleFiles, fileName=sampleFiles,
condition=sampleCondition)
> sampleTable
        sampleName          fileName condition
1 C1_R1.counts.txt C1_R1.counts.txt        C1
2 C1_R2.counts.txt C1_R2.counts.txt        C1
3 C1_R3.counts.txt C1_R3.counts.txt        C1
4 C2_R1.counts.txt C2_R1.counts.txt        C2
5 C2_R2.counts.txt C2_R2.counts.txt        C2
6 C2_R3.counts.txt C2_R3.counts.txt        C2
dds <- DESeqDataSetFromHTSeqCount(sampleTable=sampleTable, directory=".",
design=~condition)
dds <- DESeq(dds)
results <- results(dds)
results <- results[order(results$FDR), ]
```

~batch + condition

http://www.bioconductor.org/packages/2.12/bioc/html/DESeq2.html

Why adjusted p value in Genomics



- Lots of data in genomics that have lots of hypothesis tests.
- In RNA-seq we are doing *p* simultaneous tests! H1, H2, H3, ..., Hp
- For a 10k gene experiment, a standard p value cutoff 0.05 will give ***500*** DEGs by chance.

Integrative Analysis: Annotation

- Gene identifiers (e.g., *org.Hs.eg.db*) and models (e.g., *TxDb.Hsapiens.UCSC.hg19.knownGene)*
- Web-based resources (e.g., *biomaRt, KEGGREST, UniProt.ws*)
- Whole-genome annotations via AnnotationHub, e.g., *Ensembl, UCSC*

Hands on session

Differential Expression Analysis with *DESeq2*

- Workshop data:

  /dfs6/pub/ucightf/workshop/
- URL to access HPC3 Jupyter environment:

  biojhub-3.oit.uci.edu