mRNA abundance estimation with RNAseq

Hector Corrada Bravo

CMSC858B Spring 2012

Many slides courtesy of Ben Langmead @ JHSPH

RNA-seq differential expression

Sample A

Align

```
GTCGCAGTANCTGTCT
|||||||||| ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||| ||||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||||||||||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
||||||||| |||||
TCTCTCCCAGGAGAGC
```

Aggregate

```
                GTCGCAGTATCTGTCT
                GTCGCAGTATCTGTCT
                GTCGCAGTATCTGTCT
                GTCGCAGTATCTGTCT
                GTCGCAGTATCTGTCT
               TGTCGCAGTATCTGTC
            TATGTCGCAGTATCTG
            TATATCGCAGTATCTG
            TATATCGCAGTATCTG
            TATATCGCAGTATCTG
         CCCTATATCGCAGTAT
AGCACCCTATGTCGCA
AGCACCCTATATCGCA
AGCACCCTATGTCGCA
   GAGCACCCTATGTCGC
CCGGAGCACCCTATAT
CCGGAGCACCCTATAT
GCCGGAGCACCCTATG
```

Statistics

```
Gene 1
differentially
expressed?: YES

p-value: 0.0012
```

Gene 1

```
GCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCT
```

```
                TGTCGCAGTATCTGTC
            AGCACCCTATGTCGCA
         GCCGGAGCACCCTATG
```

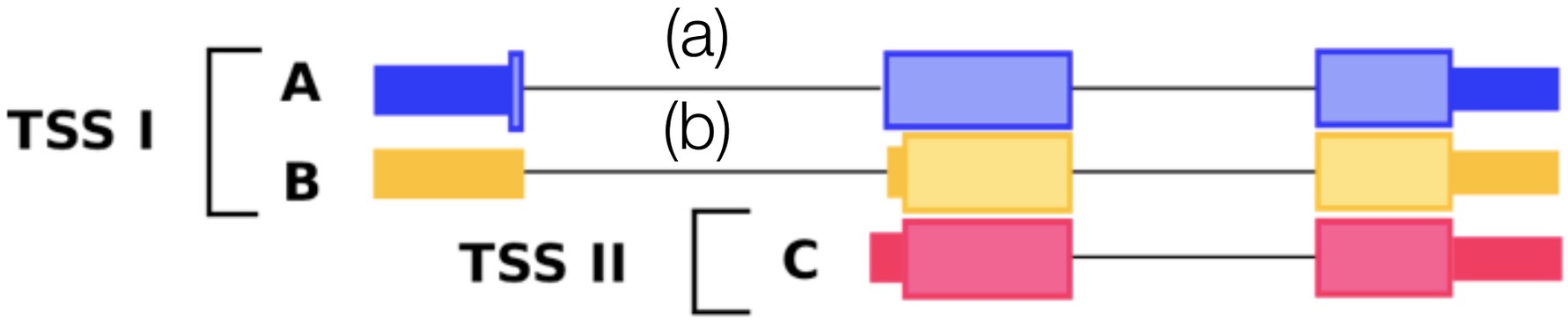

Sample B

Align

```
GTCGCAGTANCTGTCT
|||||||||| ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||| ||||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||||||||||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
```

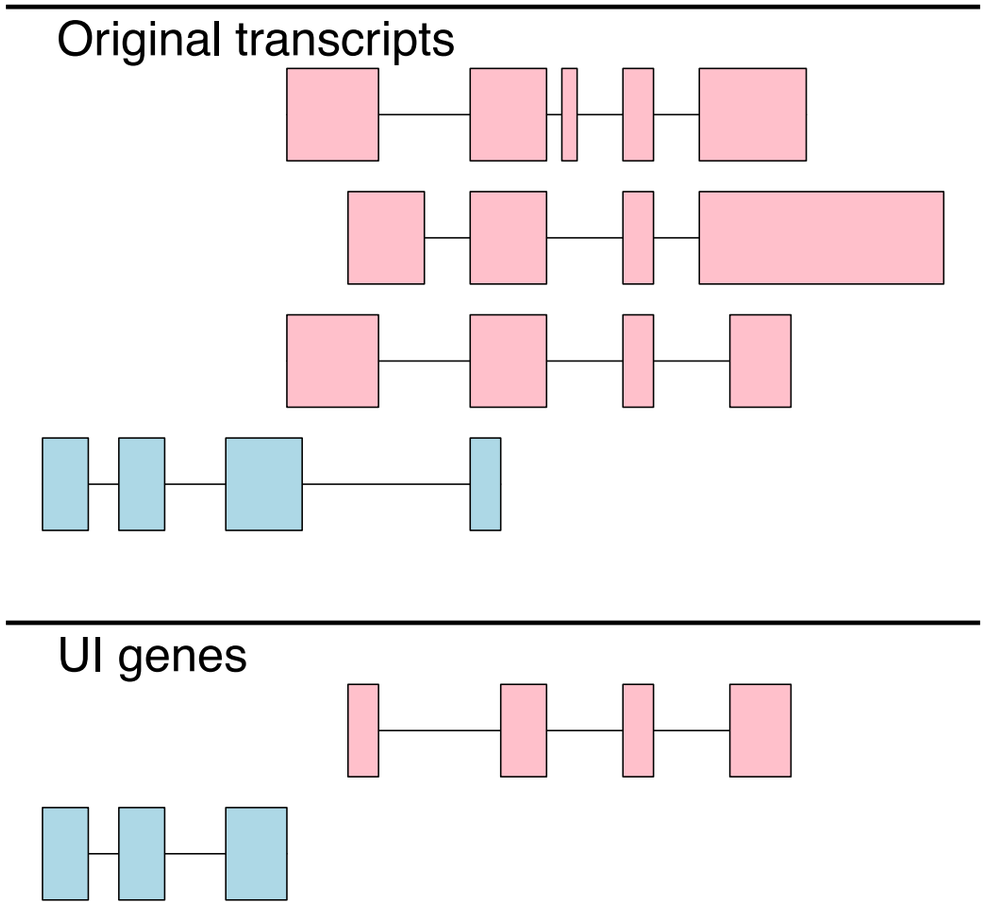
Aggregate

How should expression levels be estimated?

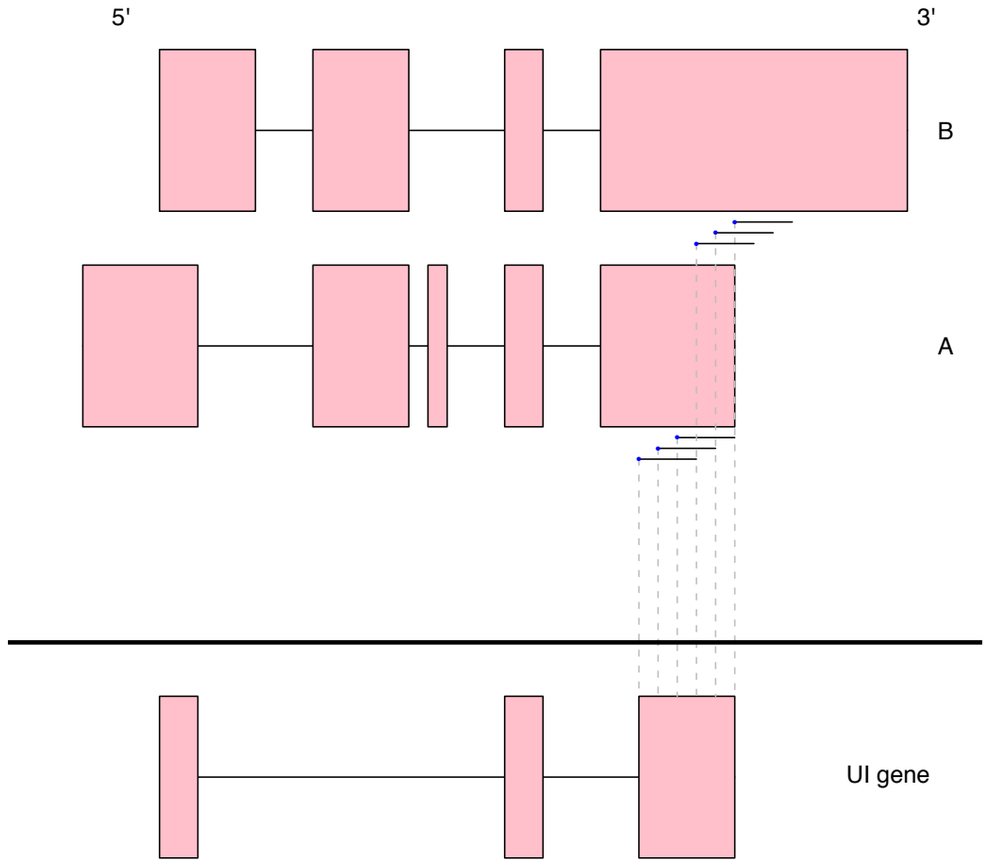


- A-B are distinguished by the presence of splice junction (a) or (b).
- A-C are distinguished by the presence of splice junction (a) and change in UTR
- B-C are distinguished by the presence of splice junction (b) and change in UTR

Union-Intersection Genes

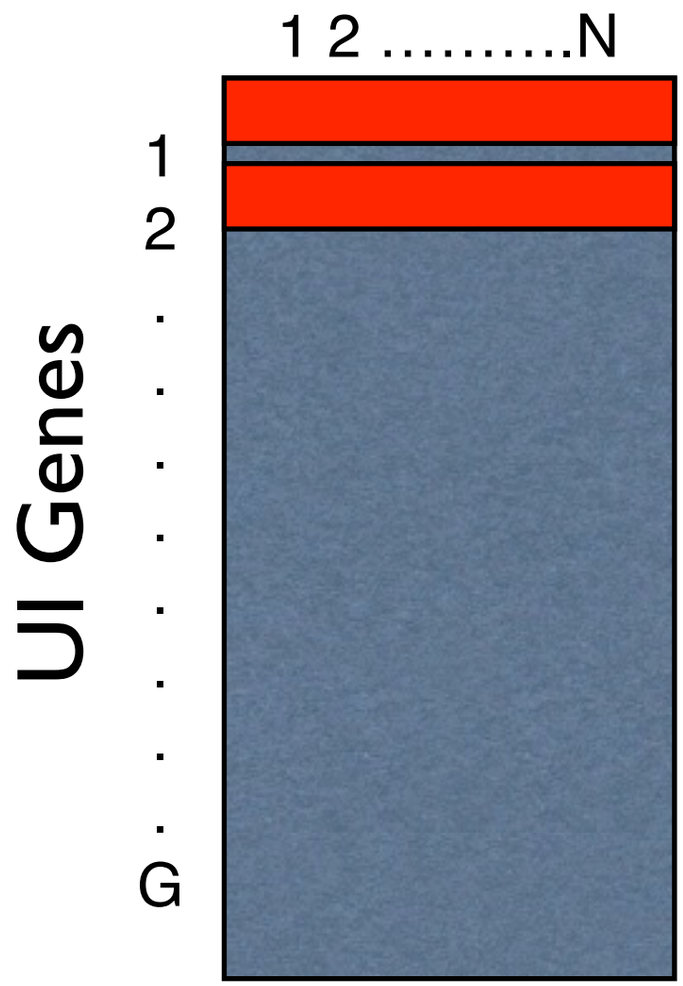


(a) UI vs. Ensembl genes



(b) UI genes and read-counting

Measurements

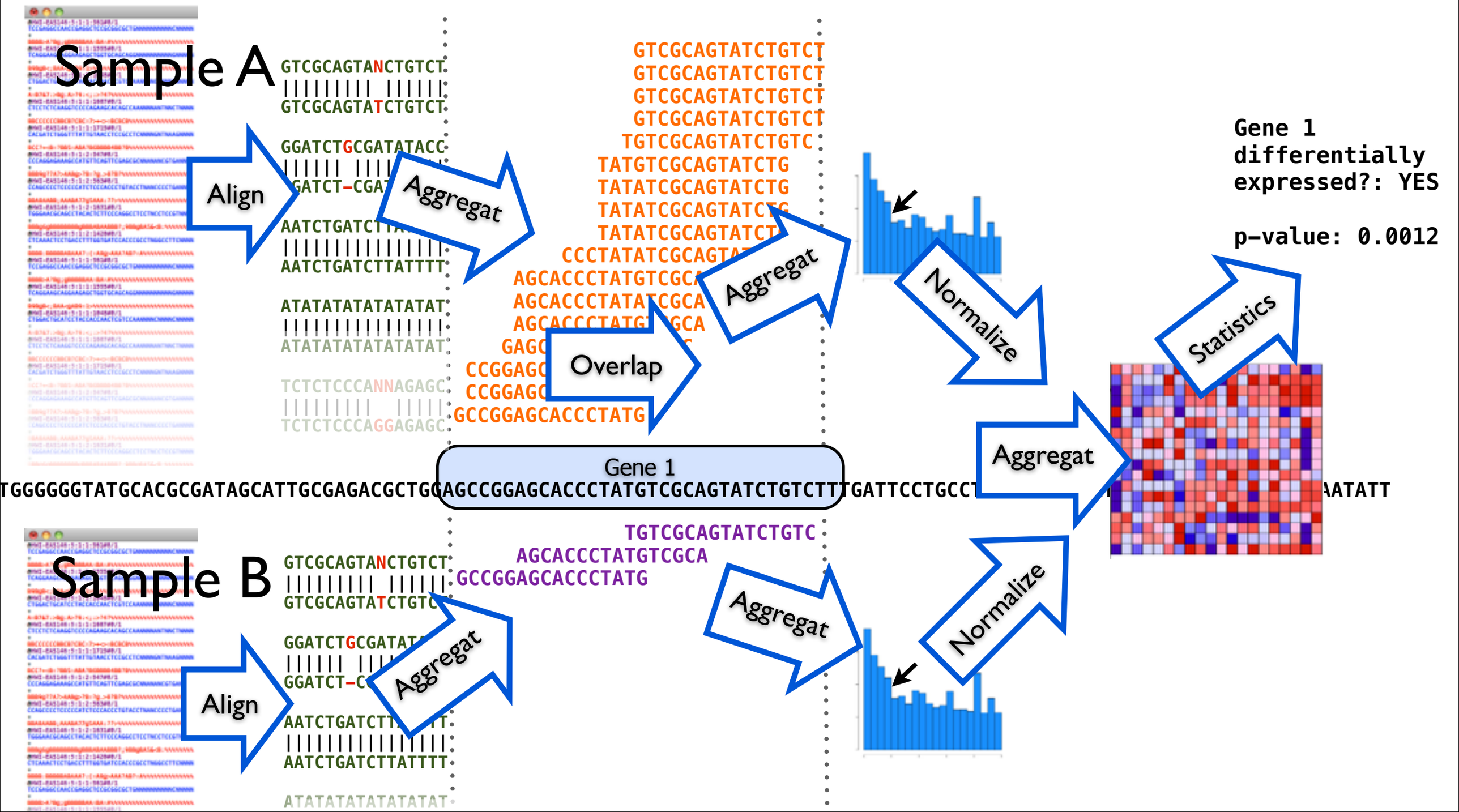


Measurements are now read counts (# of fragments) per UI-gene per sample

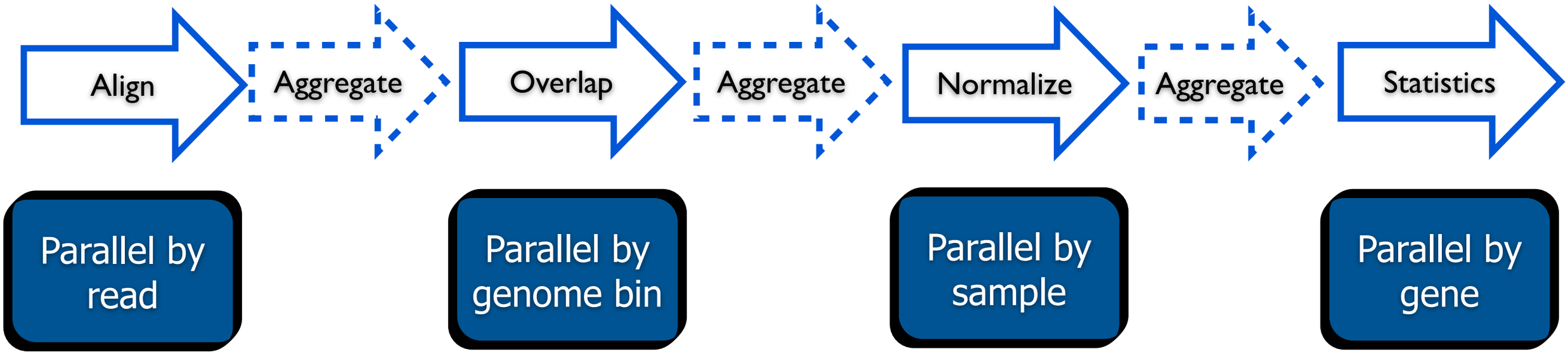
RNAseq

- Relatively new assay, some kinks are still being worked out
- The statistics for moderate sample sizes are somewhat under-developed
- Computational requirements are going to get worse
    - Cloud architectures might be the way to go for many analysts

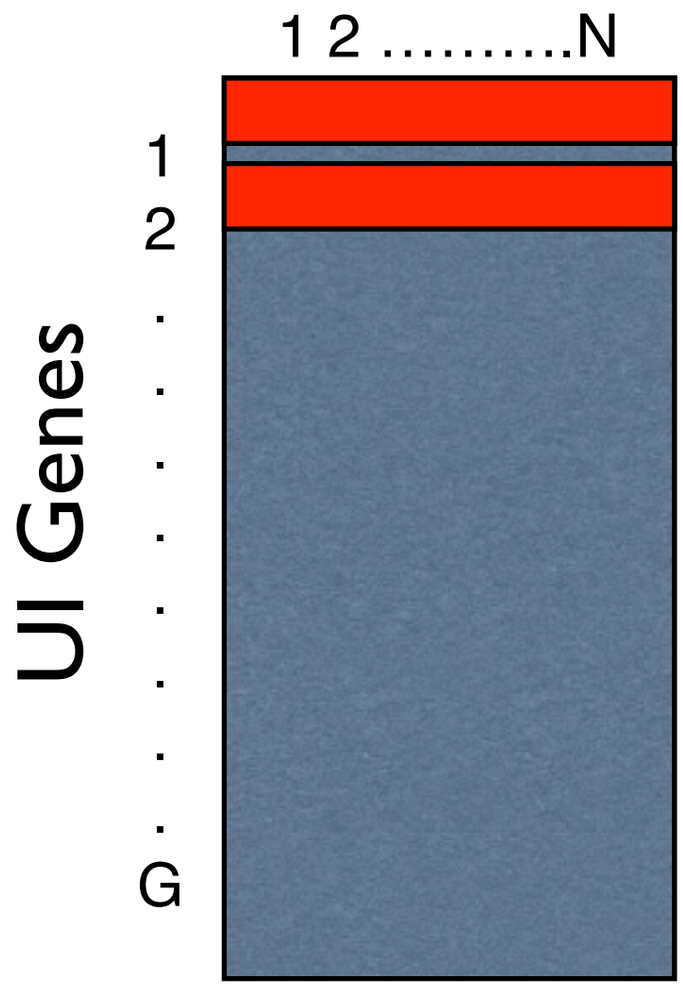
Myrna


Sample A
GTCGCAGTANCTGTCT
GTCGCAGTATCTGTCT
GGATCTGCGATATACC
GATCT-CGA
AATCTGATCTTATTTT
ATATATATATATATAT
ATATATATATATATAT
TCTCTCCCANNAGAGC
TCTCTCCCAGGAGAGC
Align
Aggregat
Overlap
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
TGTCGCAGTATCTGTC
TATGTCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCT
CCCTATATCGCAGTAT
AGCACCCTATGTCGCA
AGCACCCTATATCGCA
AGCACCCTATGTCGCA
GAGC
CCGGAGC
CCGGAGC
GCCGGAGCACCCTATG
Aggregat
Normalize
Gene 1
differentially
expressed?: YES
p-value: 0.0012
Statistics
Aggregat
Gene 1
TGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCT
AATATT
Sample B
GTCGCAGTANCTGTCT
GTCGCAGTATCTGTC
GGATCTGCGATAT
GGATCT-C
AATCTGATCTT
AATCTGATCTTATTTT
ATATATATATATATAT
Align
Aggregat
TGTCGCAGTATCTGTC
AGCACCCTATGTCGCA
GCCGGAGCACCCTATG
Aggregat
Normalize

Myrna


Align
Aggregate
Overlap
Aggregate
Normalize
Aggregate
Statistics
Parallel by read
Parallel by genome bin
Parallel by sample
Parallel by gene

Measurements

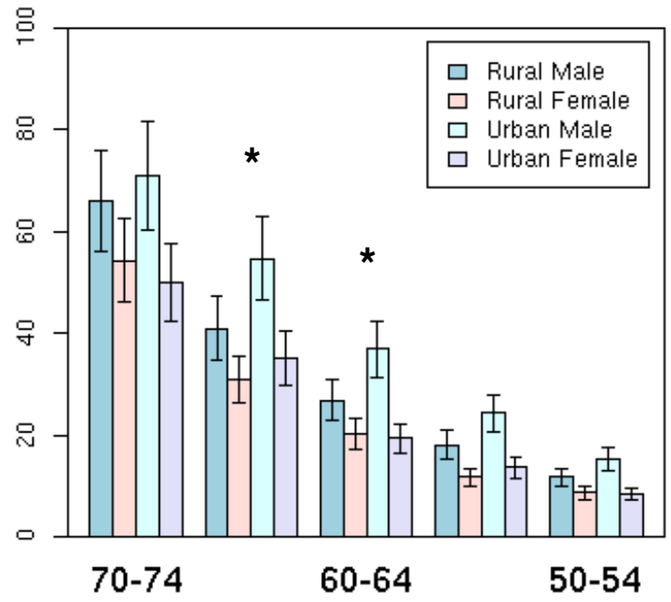


Measurements are now read counts (# of fragments) per UI-gene per sample

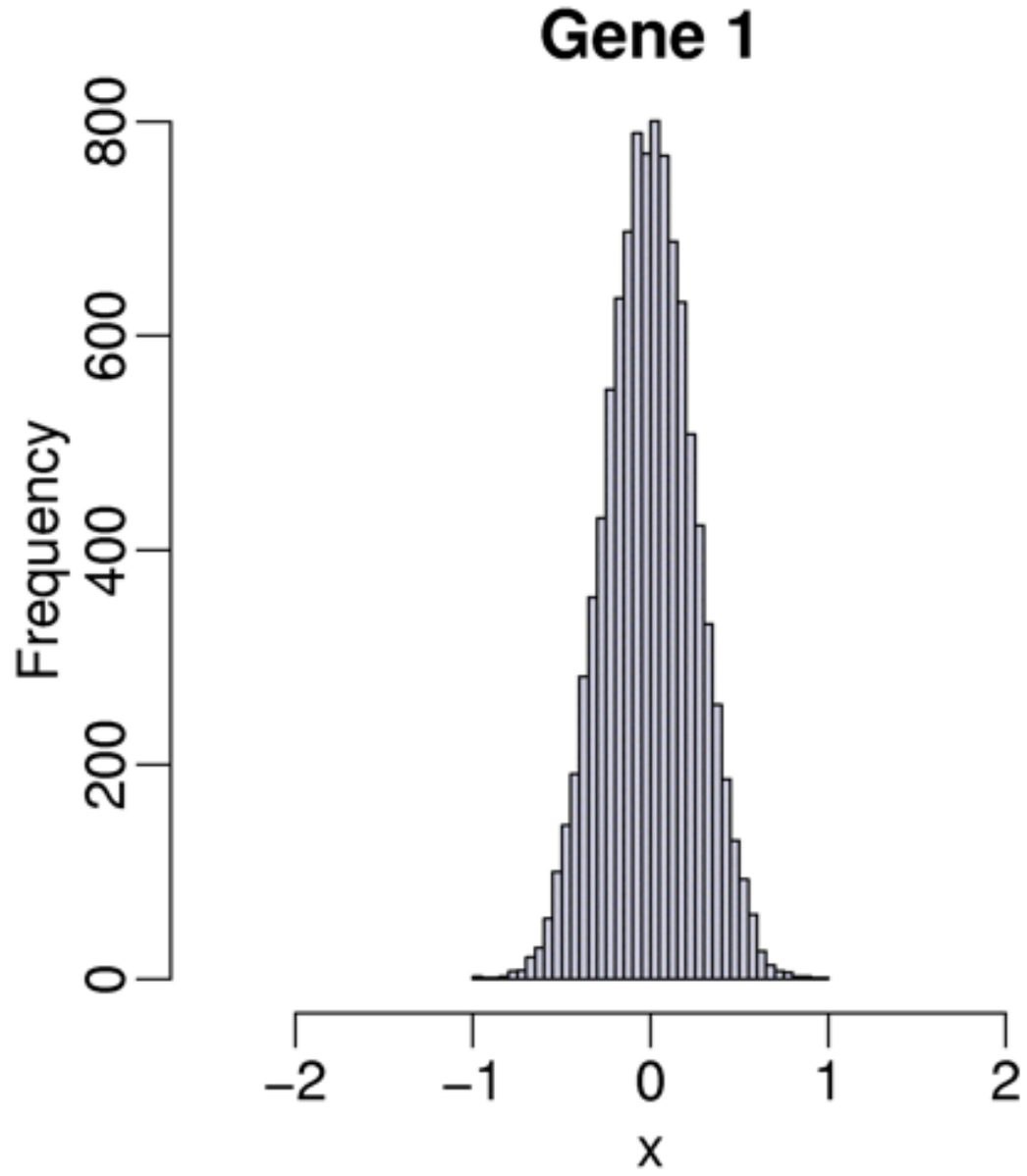
Differential Expression Analysis (DESeq)

Example

- **Fold changes: the ratio of average expression for two conditions**
- **Consider a case were we have observed two genes with fold changes of 2**
- **Is this worth reporting? Some journals require *statistical significance*. What does this mean?**

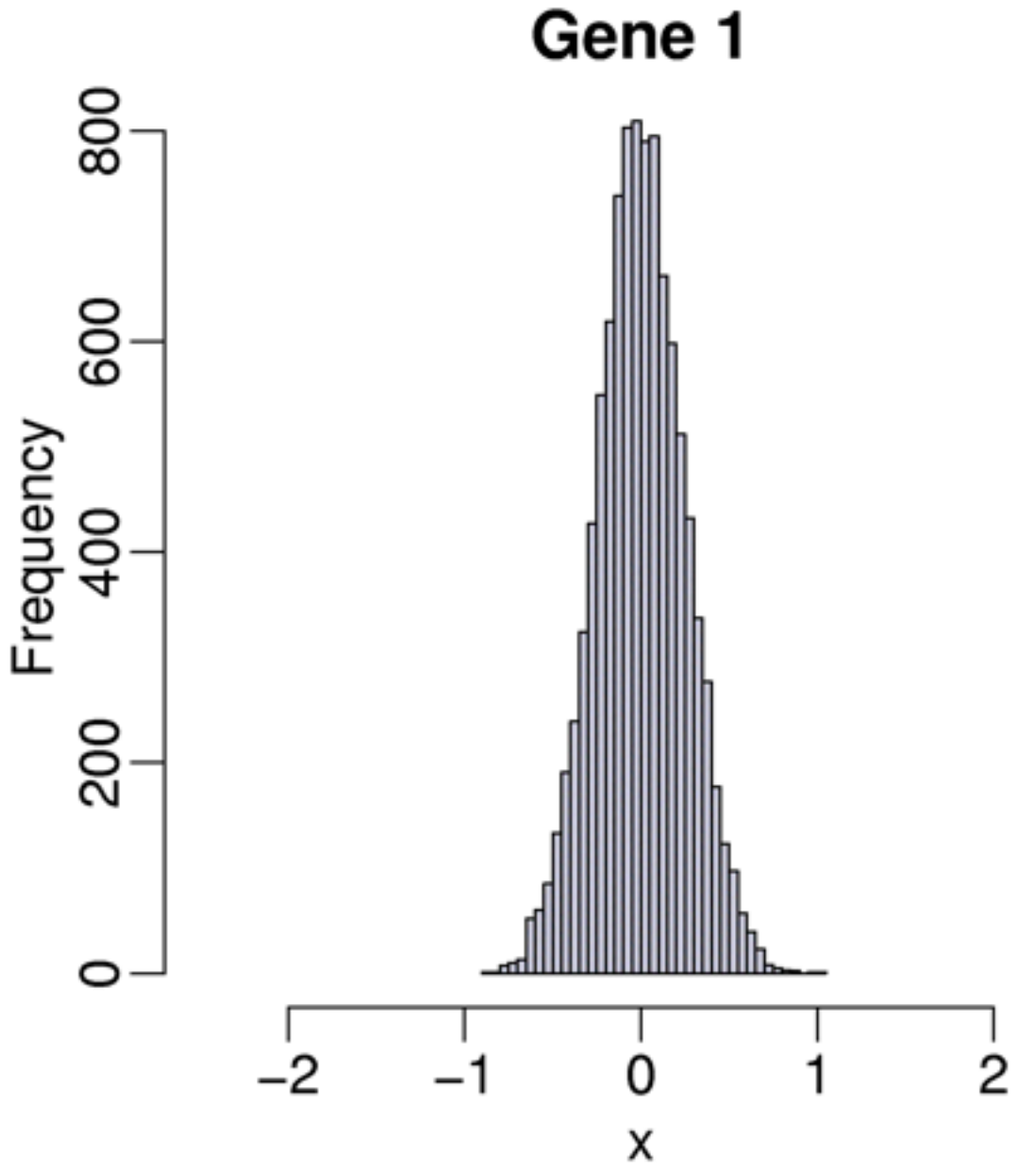


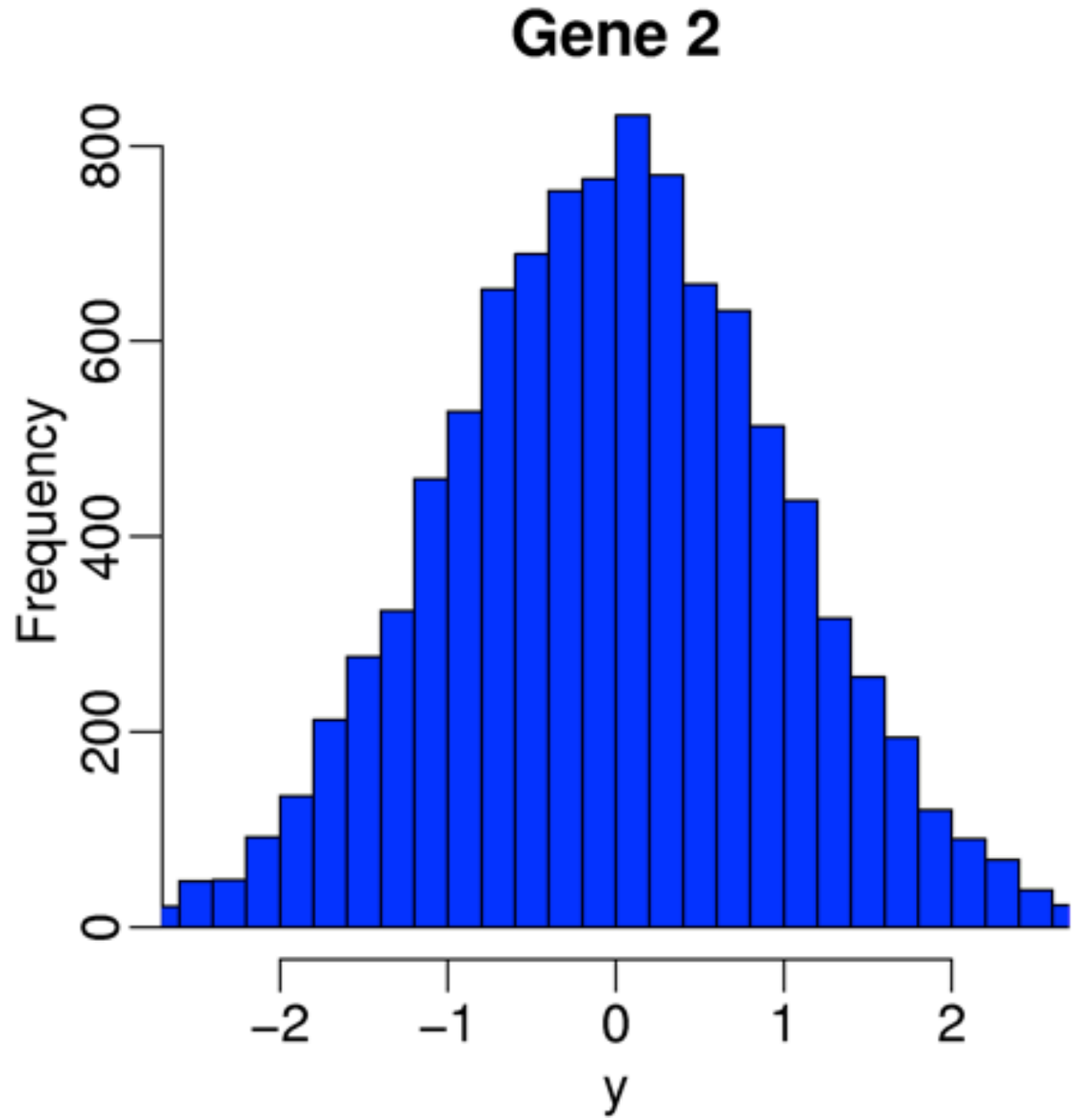
Repeated Experiment



log fold-change

Repeated Experiment





Review of Statistical Inference

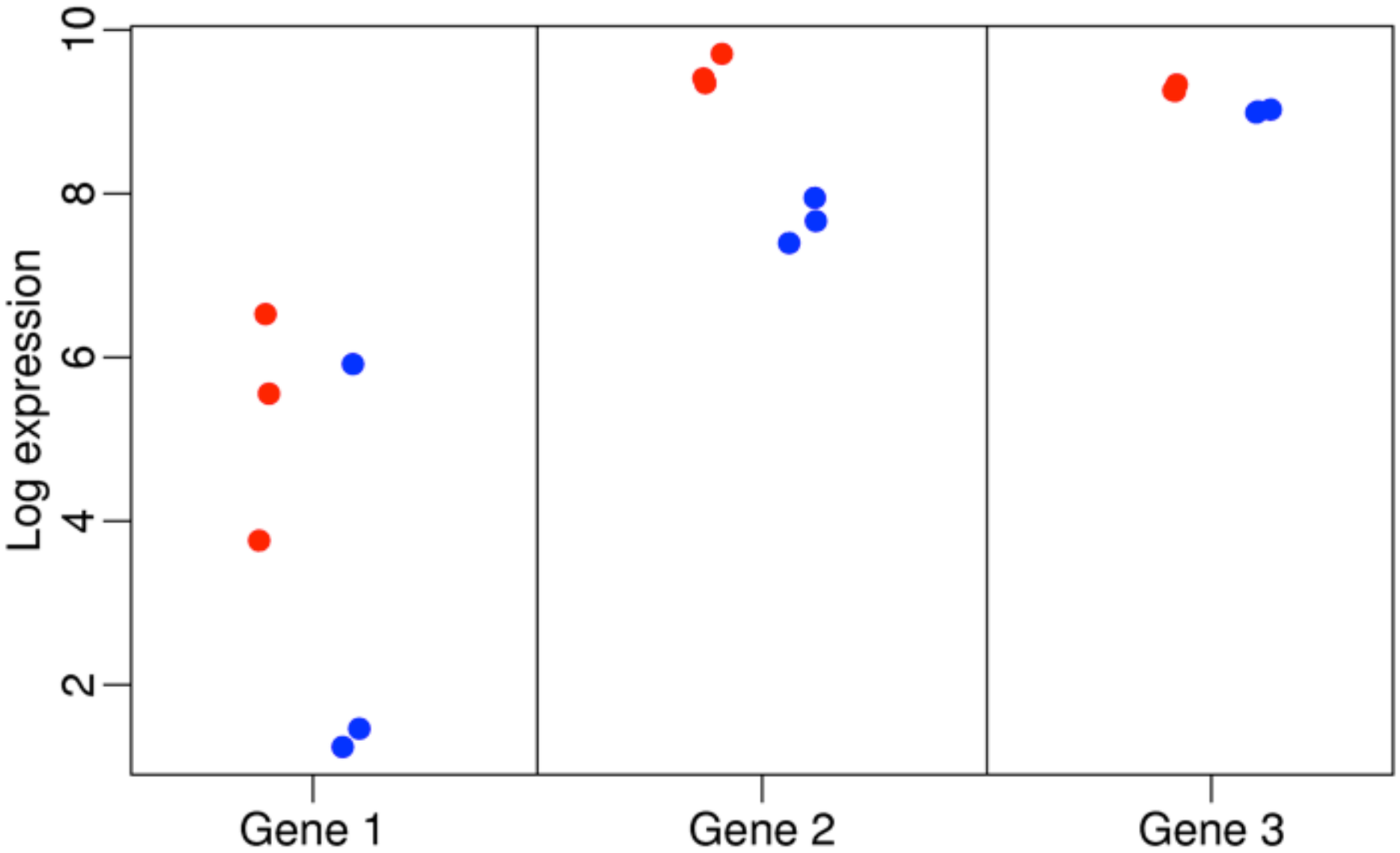
- Given statistics of differential expression (e.g., sum of counts per condition)
- What is the typical null hypothesis?

- DE stat may have a different distribution under the null hypothesis for different genes
  - Why do we want a *distribution* here?
- More specifically the standard deviation σ of *DE* stat may be different for different genes.

Inference of Ranking

- **Are we really interested in inference?**
- **Sometimes all we are after is a list of candidate genes**
- **If we are just ranking should we still consider variance?**

Should we consider gene-specific variance?



How do we summarize?

- Seems that we should consider variance even if not interested in inference

–More on this later

- Today we concentrate on ***probabilistic methods***

Questions

- **Normalization:**

–**why is normalization between samples necessary?**

–**how does DESeq normalize between samples?**

- **Distribution: What probability distribution does DESeq assume for count data?**

–**What are other possibilities?**

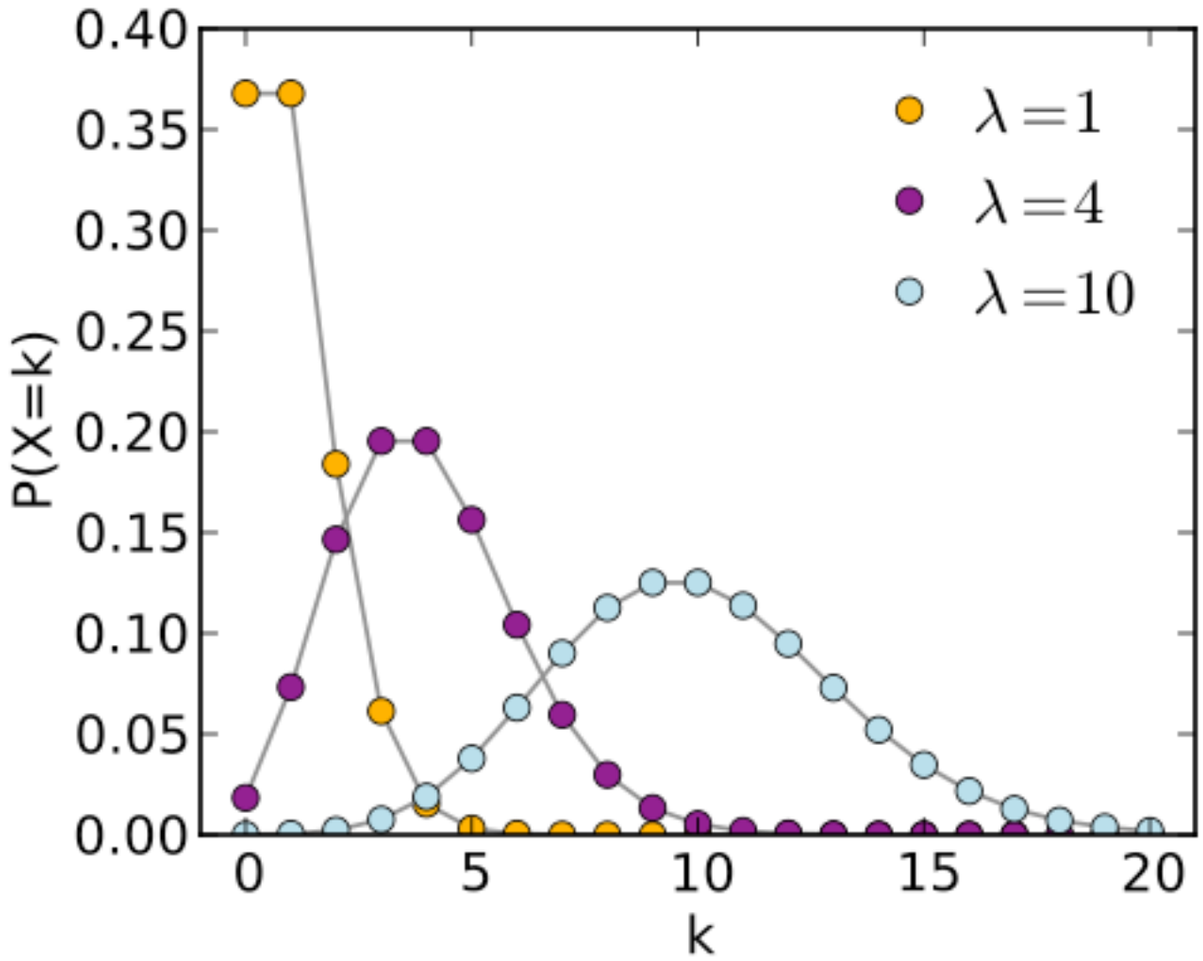
–**Why are these not appropriate**

Questions

- **Testing: What is the null hypothesis?**
  - **What do we need to compute under the null hypothesis?**

- **Suppose I wanted to build a classification model based on gene expression data, is using DESeq a sensible way of doing *feature selection*?**

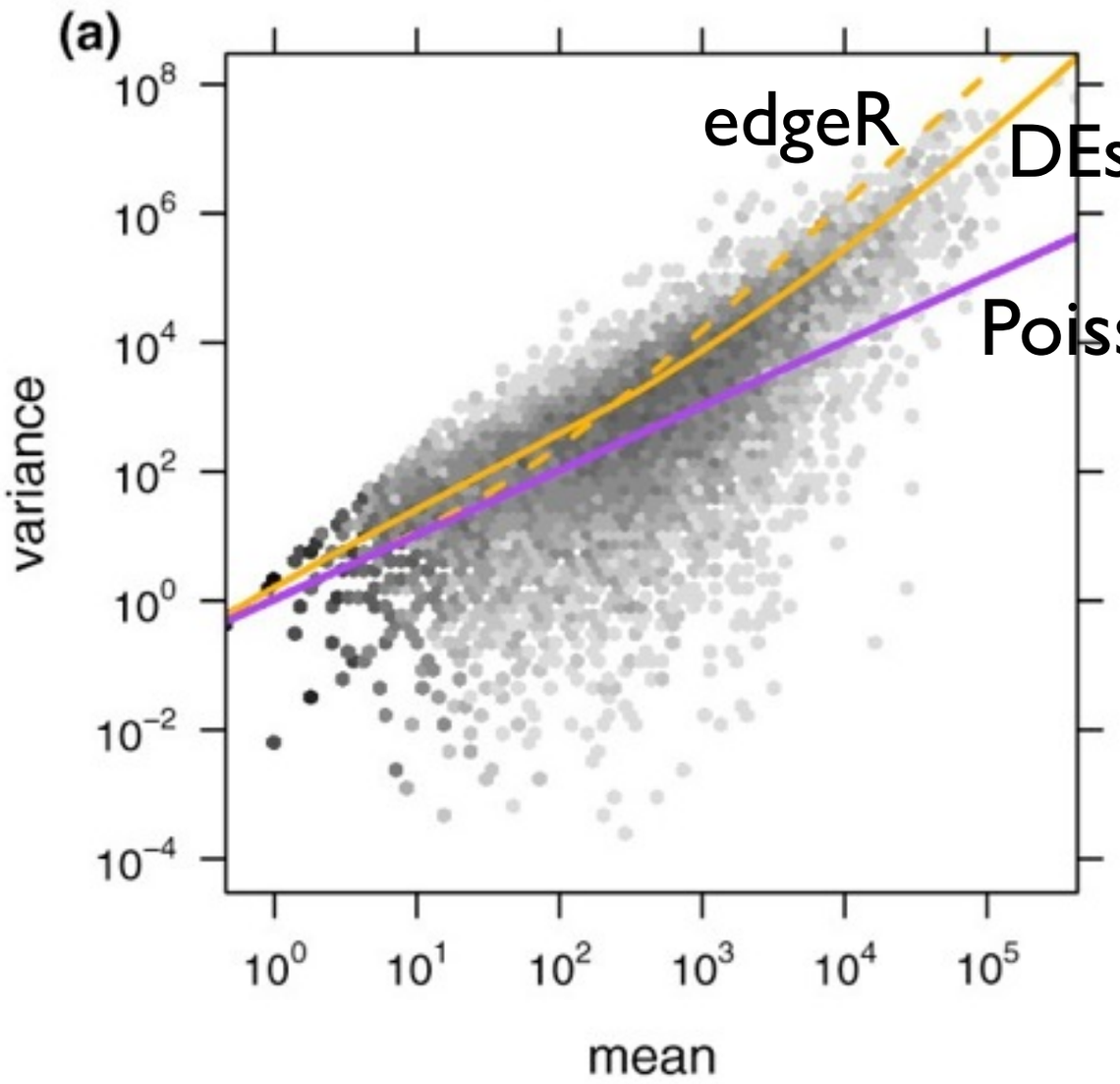
Poisson Distribution



$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

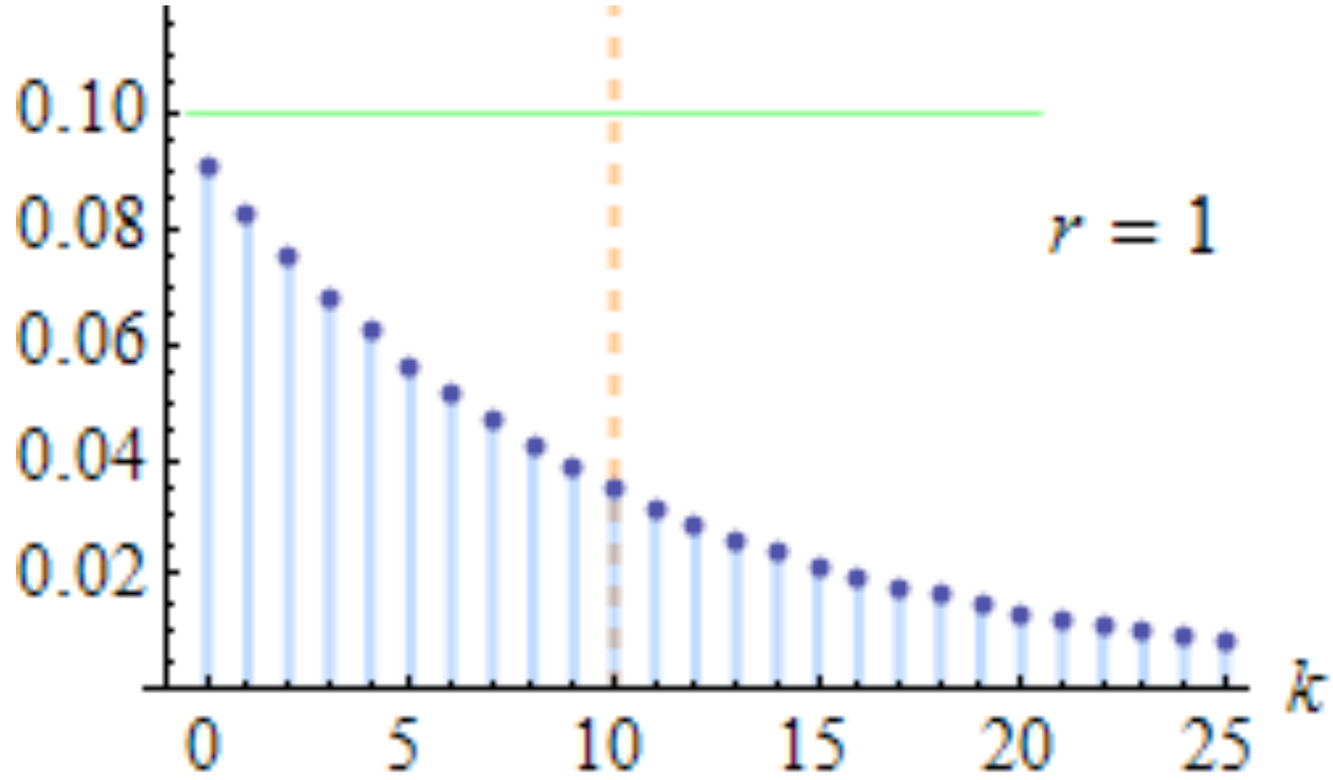
Mean & Variance: λ

Overdispersion



From: Anders et al. Genome Biology 2010

Negative Binomial



$$f(k; r, p) = \binom{k+r-1}{r-1} p^r (1-p)^k$$

Also parametrized by mean and variance

$$p = \frac{\mu}{\sigma^2} \qquad r = \frac{\mu^2}{\sigma^2 - \mu}$$

Negative Binomial

- Assume count data has a *negative binomial* distribution
  - This is a common technique to model count data where over-dispersion is apparent
  - This is what DESeq uses
- One interpretation: number of successful trials before *r* failures occur, where probability of failure is *p*