



University of  
Zurich<sup>UZH</sup>



URPP Evolution

# URPP tutorial

## RNAseq

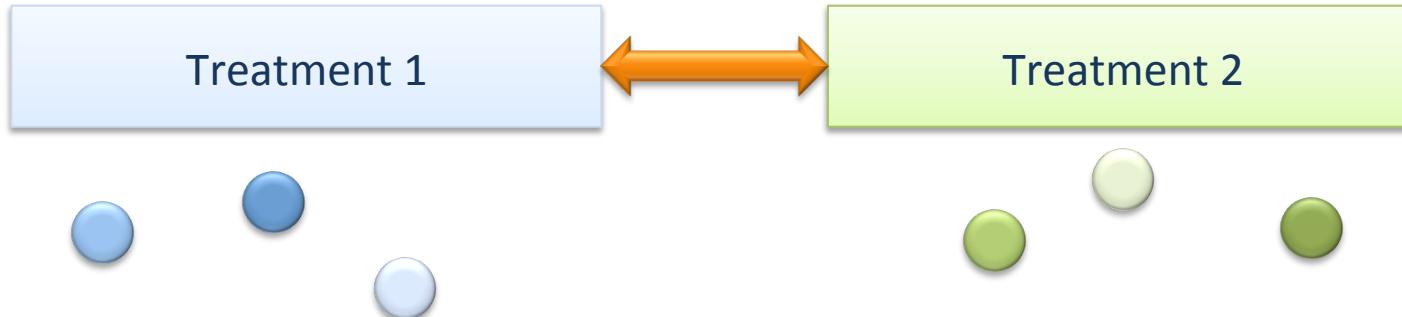
Dr. Heidi E.L. Lischer  
University of Zurich  
Switzerland

26 Mai 2014

# Introduction

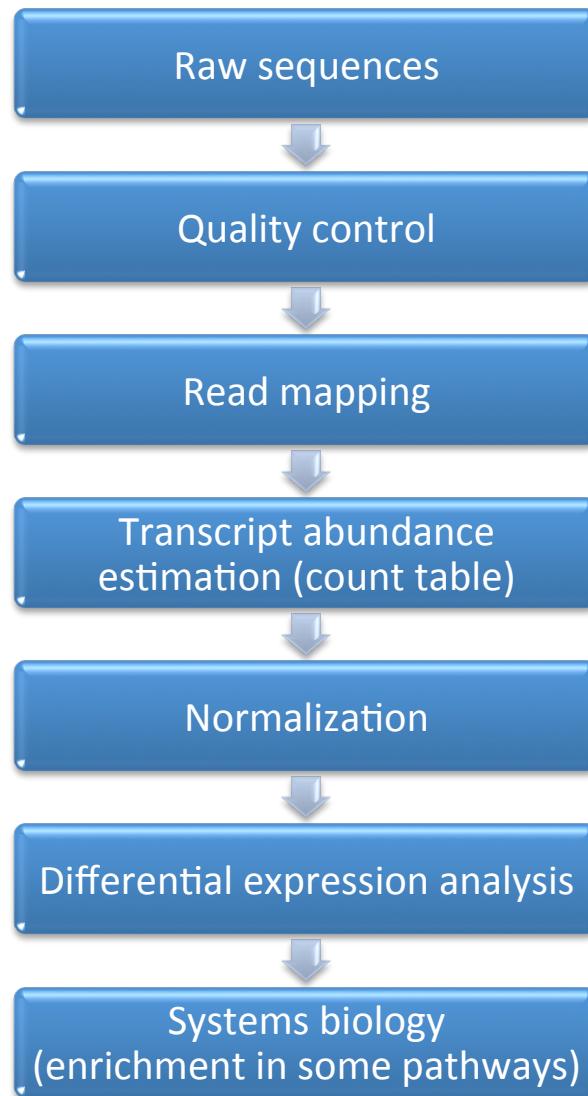
- **Gene expression**
    - essential for understanding the evolution of organisms
    - cause phenotypic variations (e.g.: between sexes, along development)
    - allows to respond to spatial and temporal changes in environment
  - **RNA-Seq:** next-generation sequencing of cDNA libraries
    - Measure gene expression in all transcripts (Microarrays: limited to array design)
    - Find new transcribed regions/genes
    - Detect low abundance transcripts
    - Study alternative splicing and allele specific expression
- Possible for non-model organisms

# Design

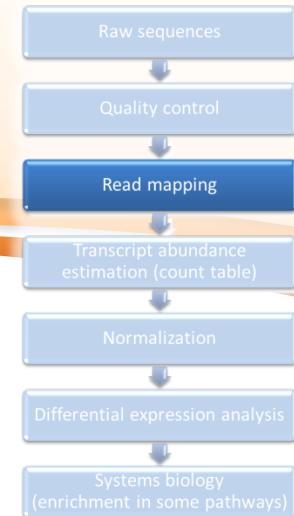


- Multiple biological replicate per treatment group
  - Increase confidence that differentially expressed genes are due to treatment and not biological variance
  - does not account for technical variance
  - Biological variance > technical variance  
→ biological replicates are more useful than technical replicates

# RNAseq pipeline



# Read mapping



- Map against transcriptome (cDNA)



- Use standard reference mapping tools (e.g.: BWA, Bowtie2)
- Transcript level expression
- Problem: shared exons → reads map to several positions
  - Gene level expression: map them randomly
  - Transcript level expression: map them proportional

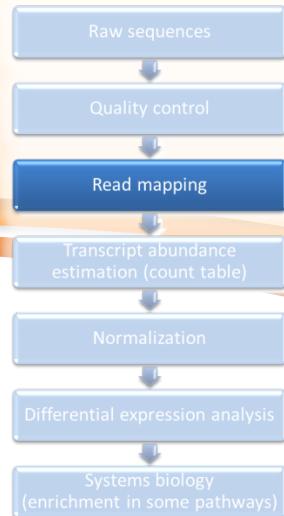
- Map against genome



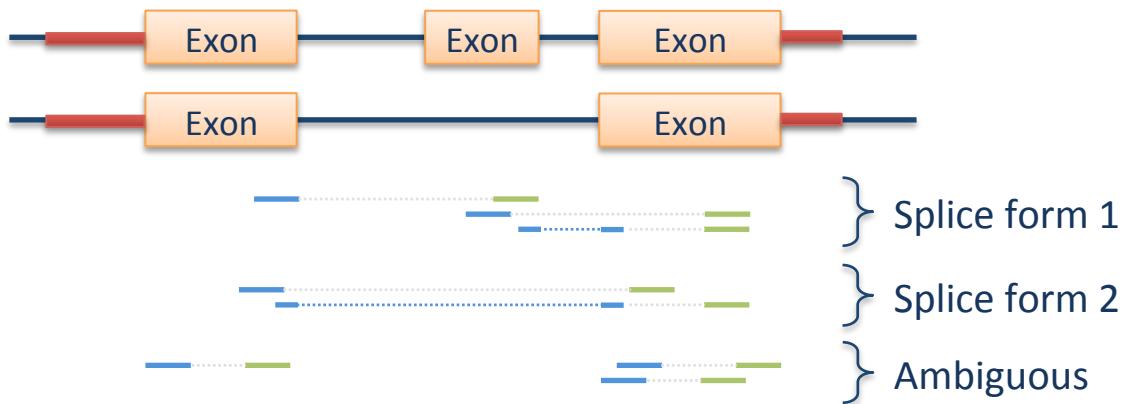
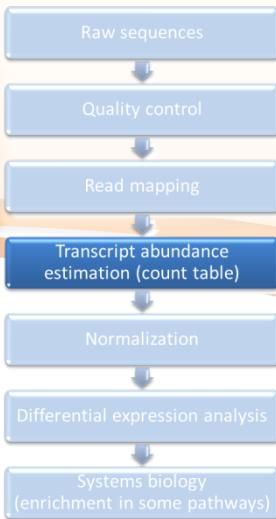
- Need splice junction reference mapping tools (e.g.: TopHat, rna-star)
- Problem: transcript level expression more difficult to estimate

# TopHat

- Automatically detects splice junctions
- Can provide annotation file (GFF/GTF)
  - Map reads first against transcriptome
  - Unassembled reads are then mapped against whole reference genome
- Requires Bowtie2
- Outputs BAM files

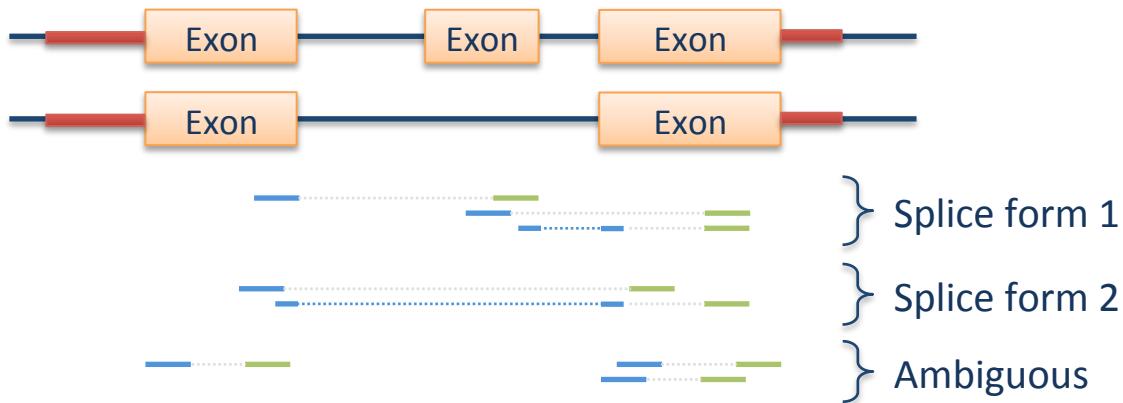
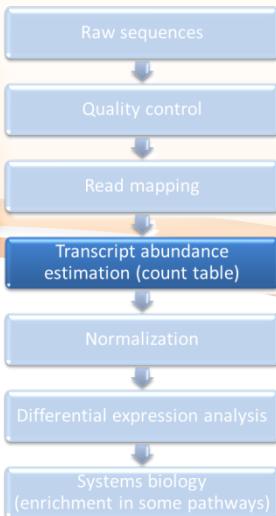


# Get count data



- Gene level (ignore splice variants):
    - Number of reads per gene (sum up reads from different splice forms)
    - Simple
    - Powerful
    - Inaccurate in some cases
- BEDtools (multicov)      → HTSeq (htseq-count)      } SAM/BAM, GFF

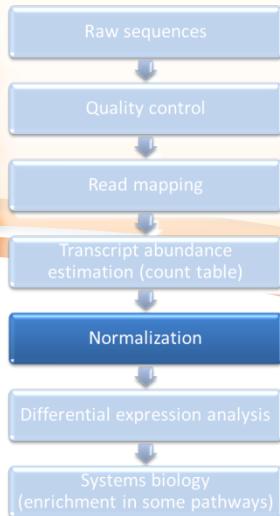
# Get count data



- Transcript level:
  - Get number of reads per splice form
  - Cleaner
  - More powerful signal
  - Some degree of uncertainty:  
Ambiguous reads are assigned proportional to unique ones

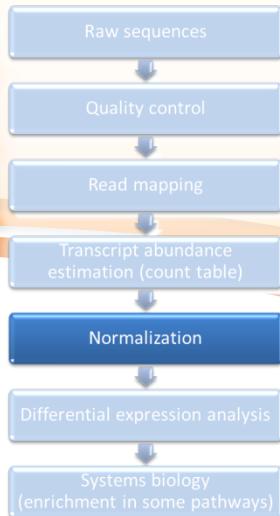
→ Cufflinks      } BAM, GFF/GTF

# Normalization



- Why normalization?
  - Suppose cDNA from treatment 1 was sequenced deeper as cDNA from treatment 2
    - sequenced on different lanes
    - differences in DNA concentration
  - Everything in treatment 1 will appear as up regulated
- RPM (reads per million reads)
  - Correct for differences in coverage
  - Allows comparisons between treatments/samples
- RPKM (reads per kilobase per million reads) / FPKM (paired-end)
  - Correct for differences in coverage
  - Correct for gene length
  - Allows comparisons between treatments/samples and genes

# Normalization



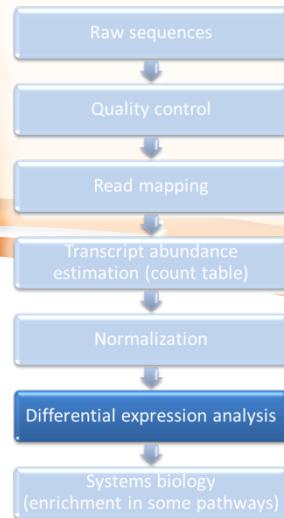
- Problems of RPM/RPKM/FPKM
    - Small changes in highly expressed genes  
→ cause global shifts in all values  
→ as highly expressed genes consume substantial proportion of total number of reads
  - EdgeR:
    - Estimates a scaling factor
    - Uses a trimmed mean of M-values (TMM)    Robinson and Oshlack (2010)
  - DESeq:
    - Calculates a size factors for each sample
    - For each gene: counts of the samples are divided by the geometric means over all samples
    - Size factor: median of all gene ratios
- Do not correct for gene length

# Differential expression analysis

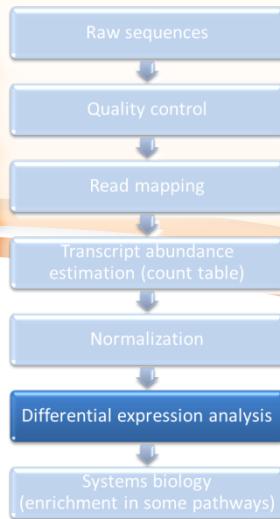
- Test if the expression strength of a gene between two treatments is larger as compared to the variation within each treatment

Bioconductor packages (R) estimate means and variances of read counts under a

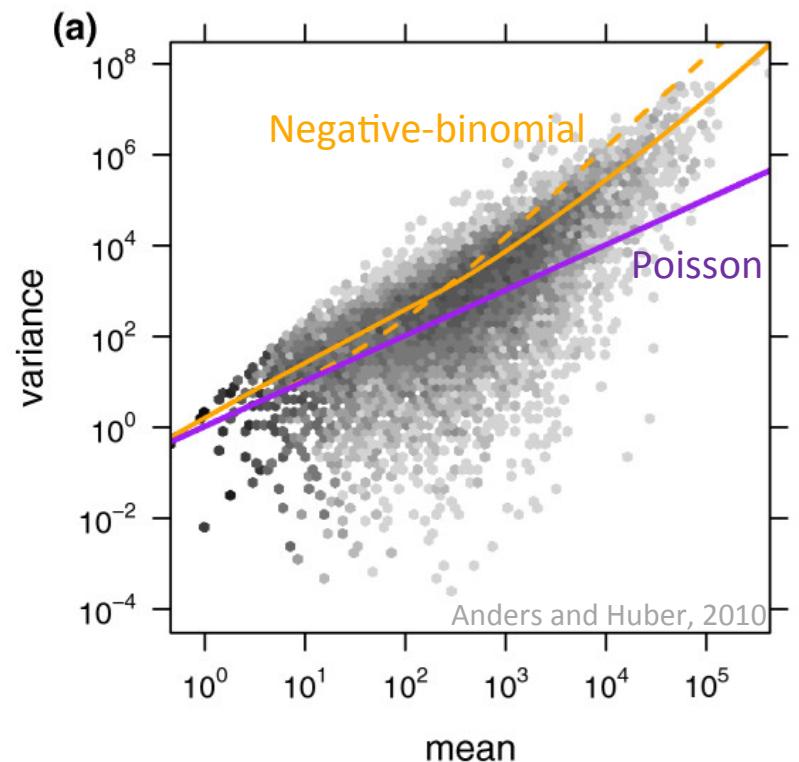
- Poisson distribution:
  - DEGSeq (Wang et al.)
- negative-binomial distribution:
  - DESeq (Anders, Huber)
  - edgeR (Robinson, McCarthy, Smyth)
  - BaySeq (Hardcastle, Kelly)



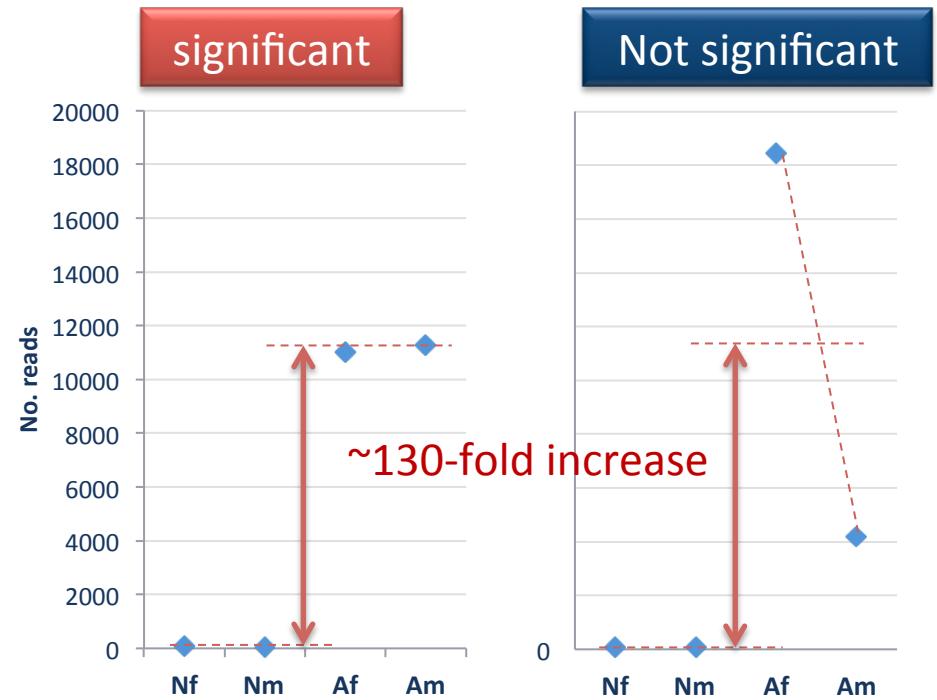
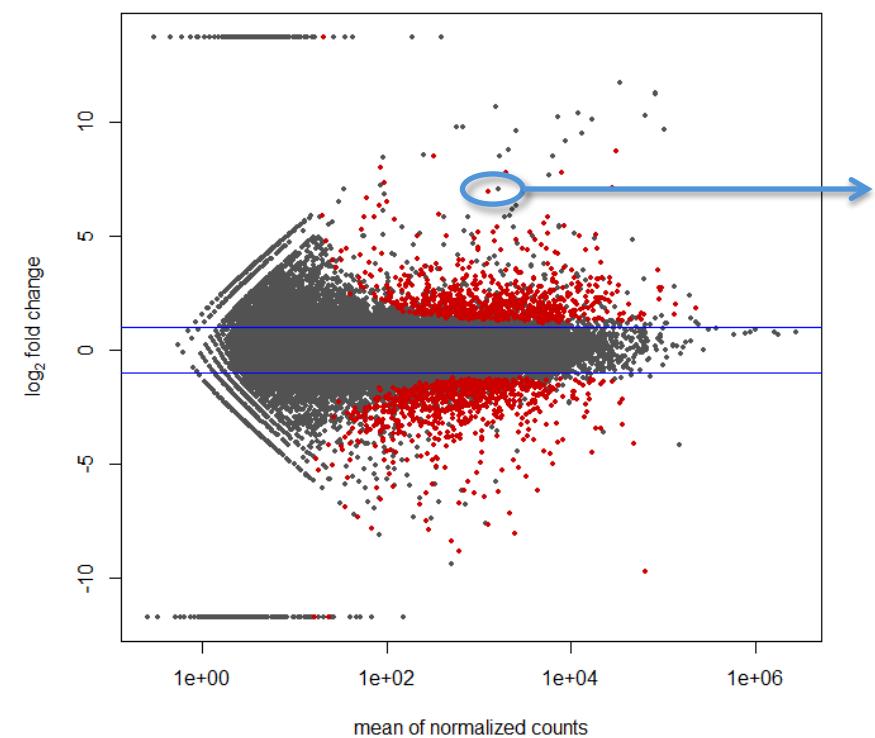
# Differential expression analysis



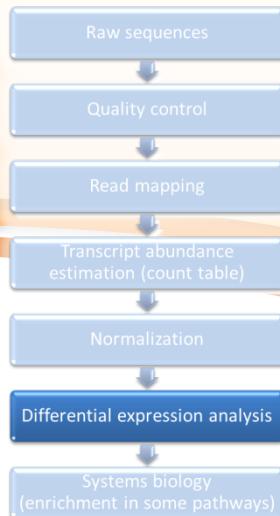
- Microarray data follow Poisson distribution
- RNAseq
  - Genes with high mean counts (longer or highly expressed) tend to show more variance  
→ Fit negative-binomial distribution better



# DESeq

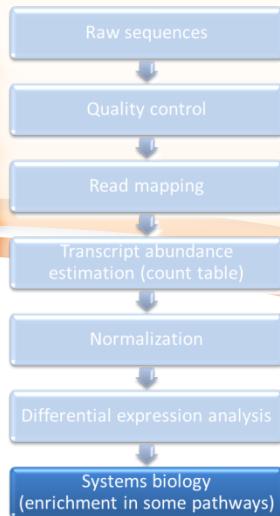


# Multiple testing

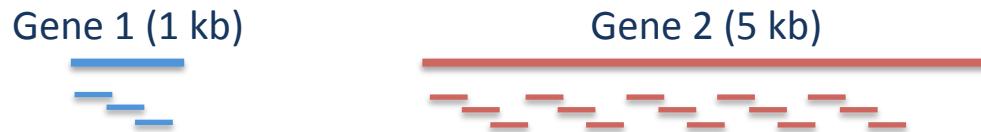


- Multiple testing
  - We test for differential expression simultaneously for n number of genes
  - Suppose we have 10,000 genes, just by chance we expect that  $10,000 * 0.05 = 500$  genes have a p-value < 0.05  
→ p-values for each gene no longer correspond to significant findings
- Bonferroni Correction:
  - $p\text{-value}' = p\text{-value}/n$
  - Problem: very conservative.
- False Discovery Procedure (FDR) (Benjamini and Hochberg, 1995)
  - order p-values in increasing order and assign a rank (smallest: rank 1, second smallest: rank 2...)
  - $FDR = p\text{-value} * n/rank$
  - expected proportion of Type I errors among the rejected hypotheses
  - If we find 40 genes significant differential expressed at a 5% FDR, we expect 2 false discoveries

# GO enrichment analysis



- Gene Ontology categories are tested for over representation amongst differentially expressed genes



- Problem: length bias
  - genes with same expression level  
→ longer genes will have more reads
  - More information for longer transcripts
  - Longer genes have higher power to detect differential expression
- GOSeq (Young et al. 2010)
  - Correct length bias (probability weighting function) in null distribution
  - Random samples of genes are created by selecting a subset of genes from the experiment → null distribution

