# RNA-seq: A practical guide to the analysis of differential gene expression

December 1st, 2011

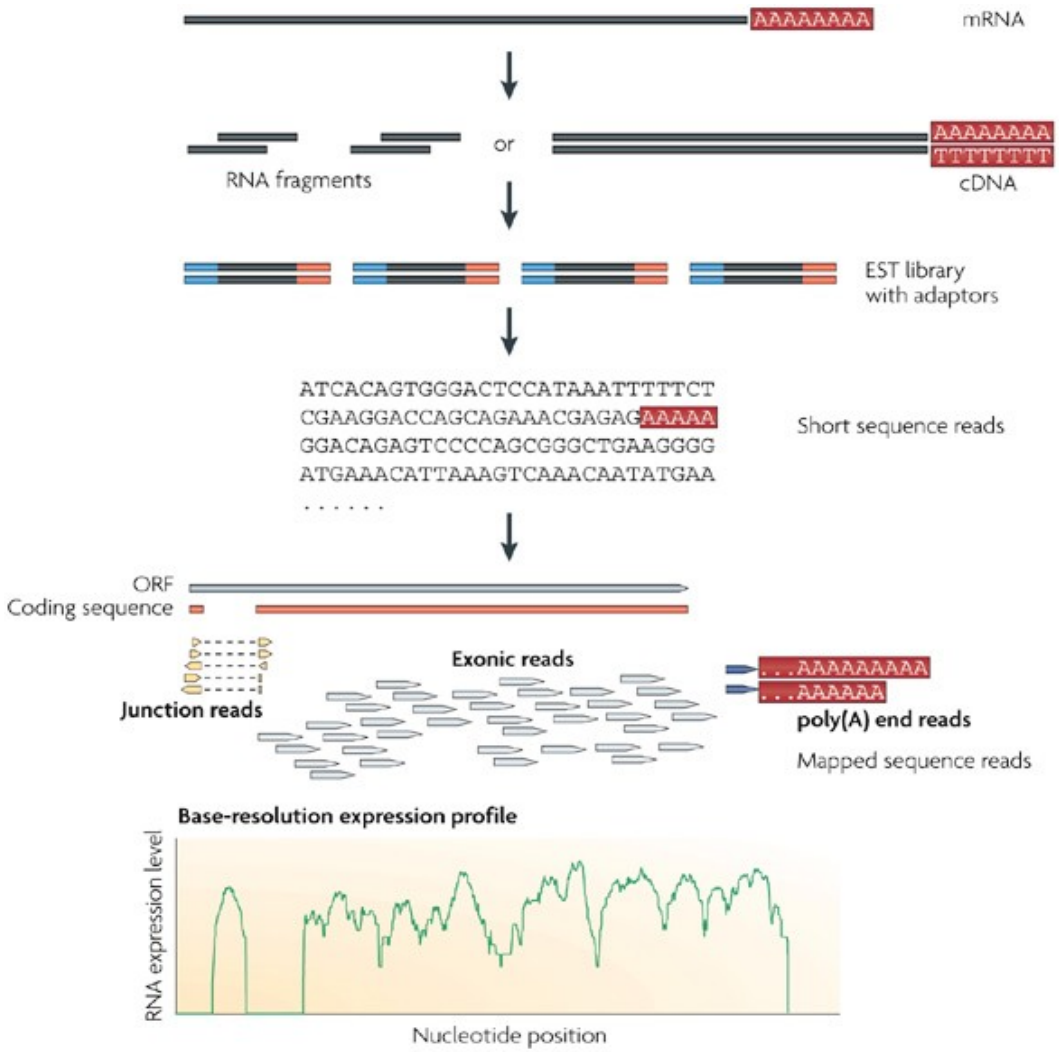
# RNA-seq Applications

- Annotation

  Identify novel genes, transcripts, exons, splicing events, ncRNAs.

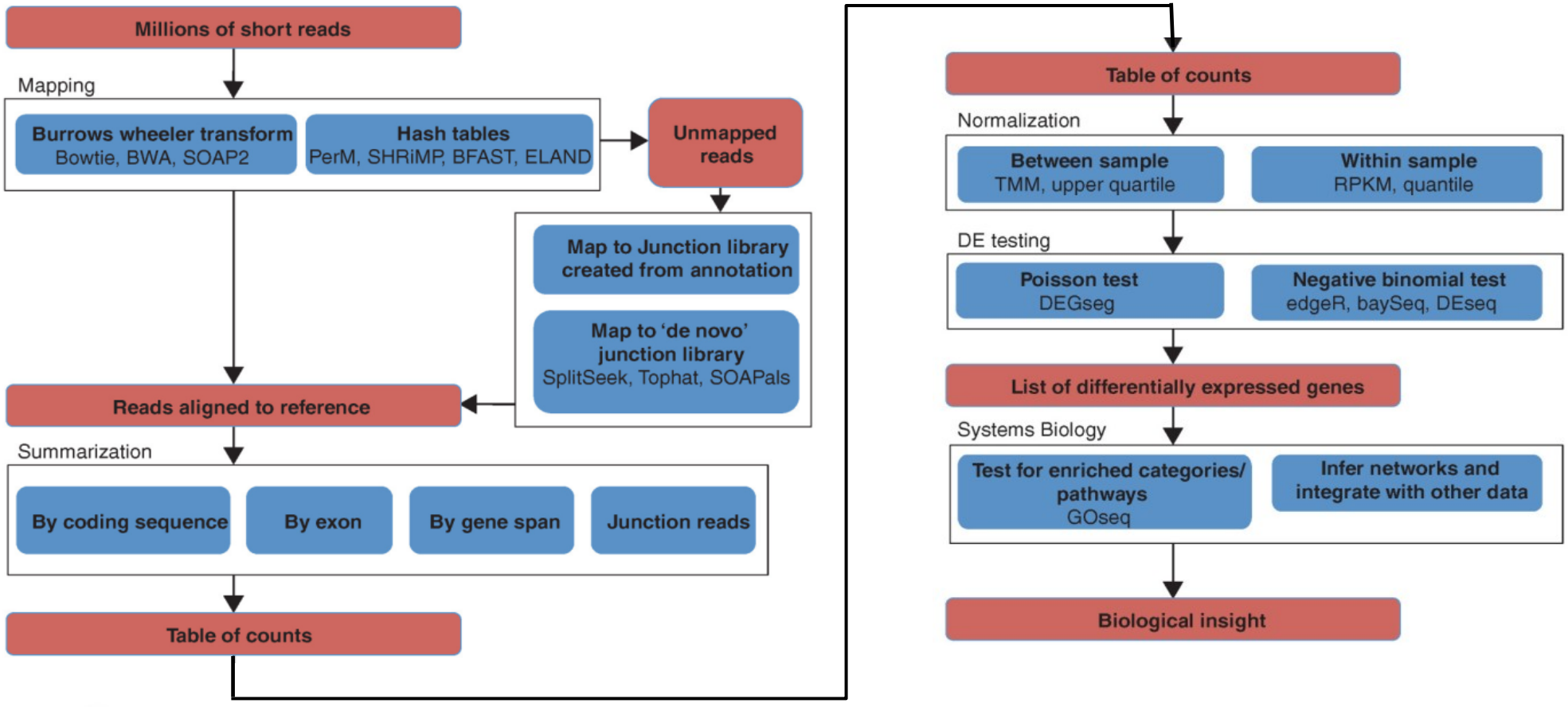
- Detecting RNA editing and SNPs.
- **Measurements: RNA quantification and differential gene expression**
  - ➢ Abundance of transcripts between different conditions

# RNA-Seq Experiment




Wang, Z. et al. *RNA-Seq: a revolutionary tool for transcriptomics Nature Reviews Genetics (2009)*

# Overview of the RNA-seq analysis pipeline for detecting differential expression




Oshlack et al., *From RNA-seq reads to differential expression results, Genome Biology 2010*.

# Steps involved on RNA-seq analysis for detecting differential expression

- Experimental design
- Preprocess
  - Split by barcodes
  - Quality control and removal of poor-quality reads
  - Remove adapters and linkers
- Map the reads
- Count how many reads fall within each feature of interest (gene, transcript, exon etc).
- Remove absent genes
- Add offset (such as 1)
  - Prevent dividing by 0
  - Moderate fold change of low-count genes
- Identify differentially expressed genes.

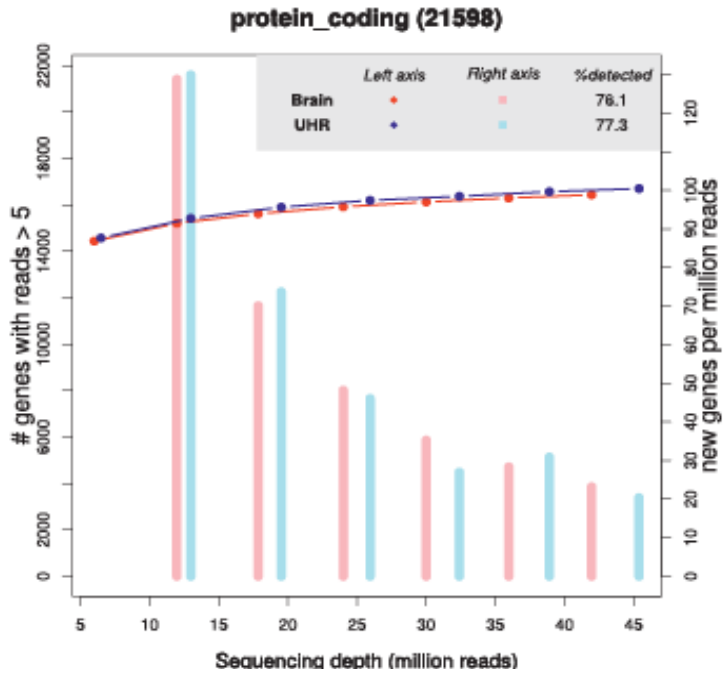
# Experimental design

- **Include replicas in your experiment.!!!!!!!** Conclusions drawn from a single RNA-seq experiment can be very misleading.
- Number of reads needed for an experiment.

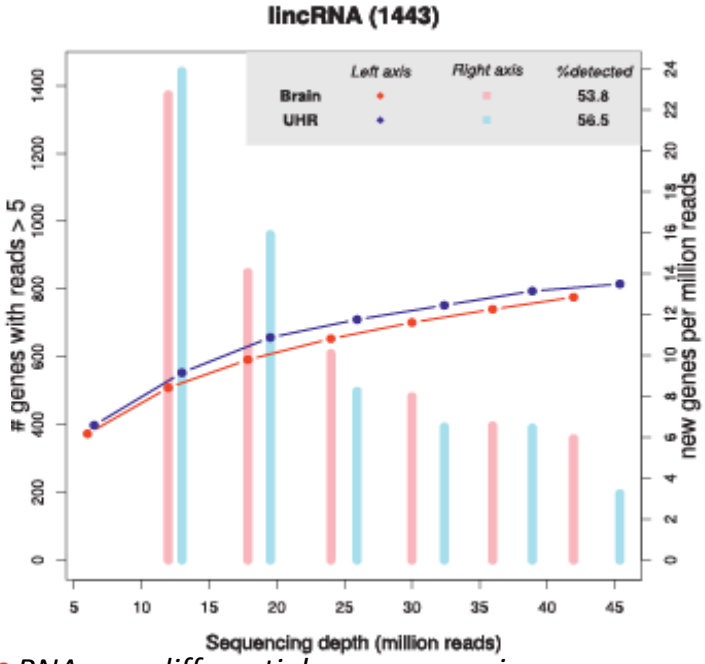
  Depends on the organism and the level of the differences you want to detect.

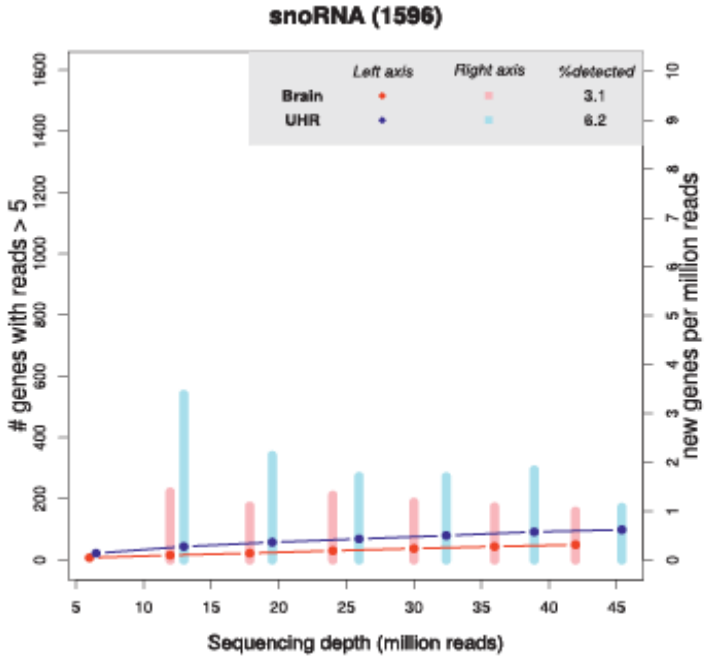
  - *Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. Wang et al. BMC Bioinformatics 2011, 12(Suppl 10):S5* http://www.biomedcentral.com/1471-2105/12/S10/S5
  - *Differential expression in RNA-seq: A matter of depth. Genome Res. 2011*. PMID: 21903743.

# Number of reads needed for an experiment




*Differential expression in RNA-seq: A matter of depth. Genome Res. 2011*. PMID: 21903743.



# Preprocess

- **Useful tools for preprocessing**

Fastx Toolkit : http://hannonlab.cshl.edu/fastx_toolkit

FastQC: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

## 1. Split by barcodes

- If the barcode is still on the sequence use "*fastx_barcode_splitter.pl*"

```
cat sequence.txt |   fastx_barcode_splitter.pl --bcfile mybarcodes.txt --bol –
  mismatches 2
```

--bcfile: File containing the barcodes
--bol: beginning of the sequence

- If the facility has removed the barcode and added to the header or the sequence like

@HWI-ST333_0165_FC:4:1101:1494:2176#ACCTGAAT/1
ATACATTGTTTCCTTTTTAGAAATATTCTGTTACTATTAT

use "*splitReadsByBarcodesInDescriptionLines.pl*" script in /nfs/BaRC_Public/BaRC_code/Perl

```
splitReadsByBarcodesInDescriptionLines.pl sequence.txt Barcodes.txt  splittedSeq
```

# Preprocess

## 2. Quality control and removal of poor-quality reads

(http://iona.wi.mit.edu/bio/education/hot_topics/QC_HTP/QC_HTP.pdf)

```
fastqc s_1_seq.txt s_2_seq.txt
```

```
fastq_quality_filter -q 20 -p 80 -i s_1_seq.txt   -o
s_1_seq.fastx_trim
```

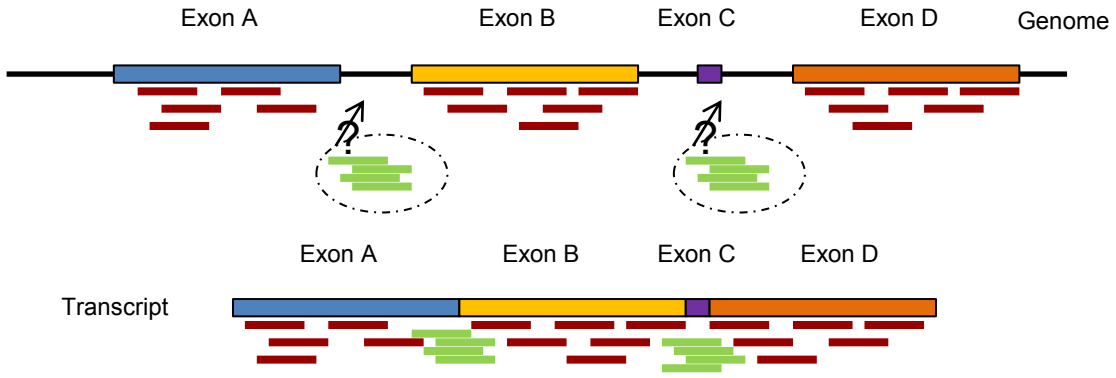
**-q = Minimum quality score to keep**

**-p = Minimum percent of bases that must have [-q] quality**

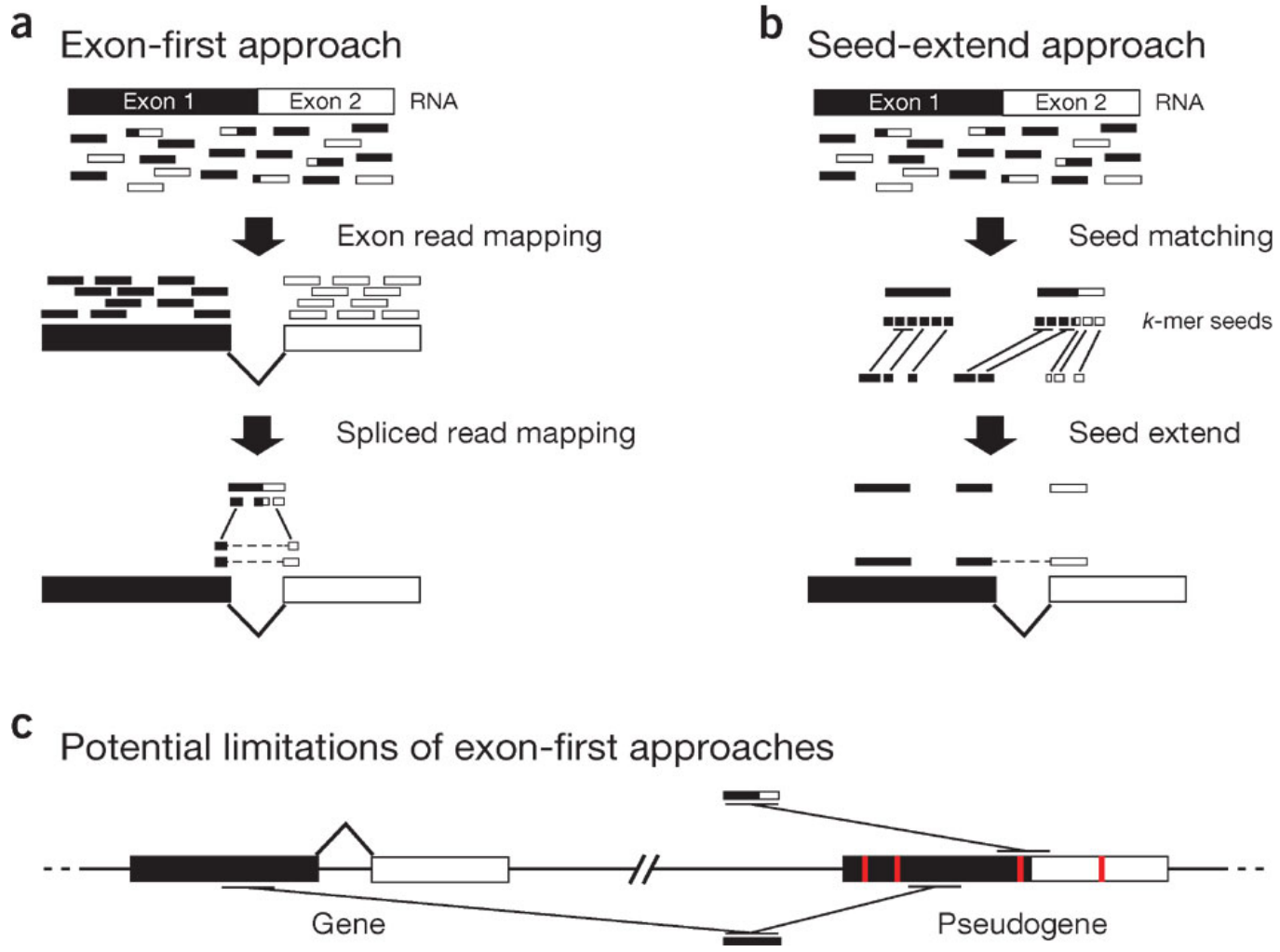
## 3. Remove adapters and linkers

```
fastx_clipper -a GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -i
s_1_seq.fastx_trim -l 22 -o s_1_seq.fastx_trim_clippedAdap
```

# Map the reads to the genome



# Map the reads to the genome




*Garber et al. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods 8, 469–477 (2011)*

# Map the reads to the genome

- Tophat (http://tophat.cbcb.umd.edu/)
  - include gff file that will be used in the counting
  - allow mapping to several places, the redundancy can be removed later.
- Sample commands:

1. Run tophat

```
tophat -G
/nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf  -o
OutputName --solexa-quals
/nfs/genomes/human_gp_feb_09_no_random/bowtie/hg19
seq.fastq
```

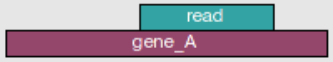
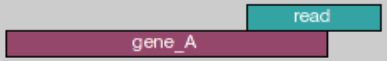
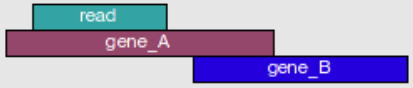
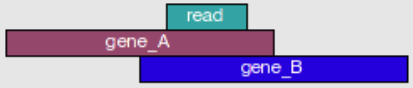
2. Convert BAM to SAM output (since SAM is required for htseq-count).

```
samtools view -h -o accepted_hits.sam accepted_hits.bam
```

# Count reads with Htseq-count

# Htseq-count

(http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

| | union | intersection_strict | intersection_nonempty |
|---|---|---|---|
|  | gene_A | gene_A | gene_A |
|  | gene_A | no_feature | gene_A |
|  | gene_A | no_feature | gene_A |
|  | gene_A | gene_A | gene_A |
|  | gene_A | gene_A | gene_A |
|  | ambiguous | gene_A | gene_A |
|  | ambiguous | ambiguous | ambiguous |

- ✓ Options
- ✓ Quantifying genes versus quantifying transcripts

# Count reads with Htseq-count

(http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

htseq-count is run for each sample :

```
bsub  "htseq-count  -m intersection-strict -s no
UHR1_accepted_hits.sam
/nfs7genomes/human_gp_feb_09/gtf/hg19.refgene.gtf >brain1_htseq-
count.out "
```

-s: whether the data is from a strand-specific assay (default: yes)

-m: mode to handle reads overlapping more than one feature (default: union)

```
head UHR1_htseq-count.out
 A1BG    67
 A1CF    31
 A2BP1   3
 …
head UHR2_htseq-count.out
 A1BG    67
 A1CF    39
 A2BP1   1
 …
```

and then counts are combined (it can also be done with excel):

```
paste UHR1_htseq-count.out UHR2_htseq-count.out brain1_htseq-
count.out brain2_htseq-count.out | awk -F "\t" '{print
$1"\t"$2"\t"$4"\t"$6"\t"$8  }' >  All_htseqCounts.txt
```

```
A1BG    67      67      20      53
A1CF    31      39      0       0
A2BP1   3       1       542     532
```

# Remove absent genes, add 1 pseudocount

- Remove absent genes (zero counts in all samples). It reduces the number of tests and the false discovery rate correction.
- Add 1 pseudocount (prevent dividing by 0).
- Remove the rows at the bottom with descriptions like no_feature, ambiguous, etc.

Update: Pseudocounts can be added for display purposes or other manual processing, but a counts matrix used for statistics should NOT include pseudocounts.

```
awk -F "\t" '{if ($2>0 || $3>0 ||  $4>0 || $5>0  ) print
$1"\t"$2+1"\t"$3+1"\t"$4+1"\t"$5+1  }' All_htseqCounts.txt |

 grep -v no_feature  | grep -v ambiguous | grep -v
too_low_aQual | grep -v not_aligned | grep -v
alignment_not_unique > All_Counts_nozero_1pseudocount.txt
```

Note that these steps can be done with excel.

# Add header

```
cat header.txt  All_Counts_no0_1pc.txt >
   All_Counts_no0_1pc_Header.txt
```

```
head All_Counts_no0_1pc_Header.txt
```

| ID | UHR_1 | UHR_2 | brain_1 | brain_2 |
|---|---|---|---|---|
| A1BG | 68 | 68 | 21 | 53 |
| A1CF | 32 | 40 | 1 | 1 |
| A2BP1 | 4 | 2 | 543 | 533 |
| A2LD1 | 11 | 11 | 4 | 7 |
| A2M | 1772 | 1901 | 407 | 450 |

File ready to use for running software to find DE genes

# Finding differentially expressed genes: Need for Normalization

- More reads mapped to a transcript if it is

  i) long

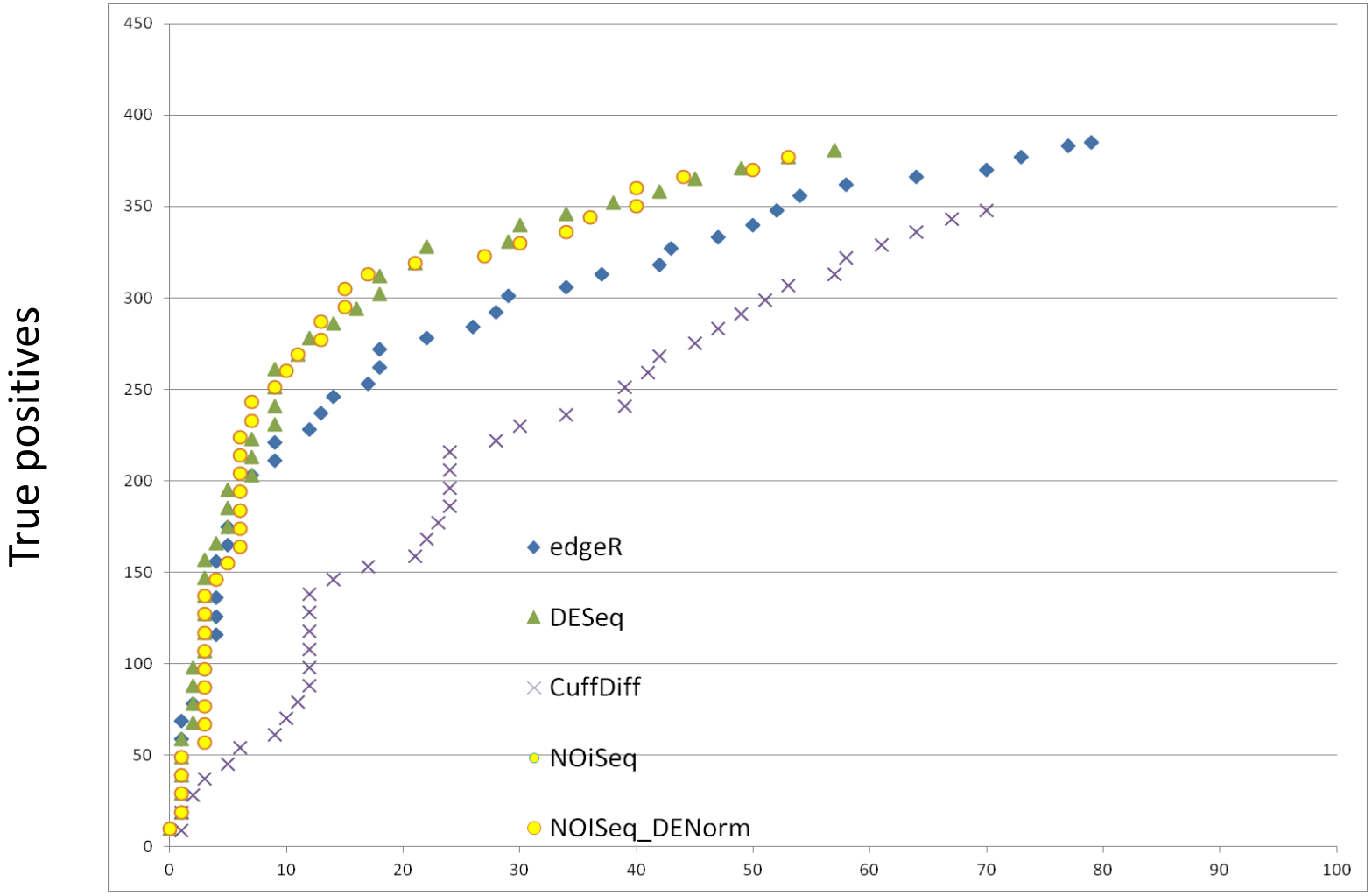
  ii) at higher depth of coverage
- Normalize such that i) *features* of different lengths and ii) total sequence from different conditions can be compared

# Finding differentially expressed genes

| Method | Normalization | Needs replicas | Input | Statistics for DE | Other |
|---|---|---|---|---|---|
| EdgeR | Library size and TMM (trimmed mean of M values) | Yes | Raw counts | Empirical Bayes estimation and exact tests based on the negative binomial distribution. | PMID: 19910308 |
| DESeq | Library size | No | Raw counts | Negative binomial distribution. | Uses different coefficients of variation for different expression strengths. PMID: 20979621 |
| baySeq | Library size | Yes | Raw counts | Empirical Bayesian methods using the negative binomial distribution. | PMID: 20698981 |
| NOISeq | Several options: counts per million read, RPKM, Upper Quartile. | No | Raw or normalized counts | Compares replicates within the same condition to estimate noise distribution of M (log-ratio ) and D (absolute value of the difference). A feature is considered to be differentially expressed if its corresponding M and D values are likely to be higher than noise values. | PMID: 21903743 |

# BaRC RNA-seq bakeoff

Differential expression between brain and human universal reference RNA (UHR) from the MicroArray Quality Control (MAQC) Project. (*BMC Bioinformatics. 2010 Feb 18;11:94, Nat Biotechnol. 2006 Sep;24(9):1151-61.*)



# Finding differentially expressed genes with NOISeq

- Sample code for running NOISeq

```
source("NOISeq.r")
mydata <- readData(file = "All_Counts_no0_1pc_Header.txt", cond1 = c(2:3), cond2 = c(4:5), header = TRUE)
myresults <- noiseq(mydata[[1]], mydata[[2]], repl = "bio“, q = 0.9, nss = 0)
write.table(cbind (myresults$Ms[myresults$deg],myresults$probab[myresults$deg] ),
    file="genesDE_FCUHRvbrain.txt", quote=F, sep="\t")
```

**nss = 0** If the experiment didn't include replicas the number of replicates to be simulated is provided by nss parameter

**repl = "bio“** indicates that the experiment includes biological replicates

**q = 0.9** indicates that the probability cut off for considering a gene differentially expressed is 0.9

- Sample code for DESeq, EdgeR and NOISeq:

  \wi-files1\BaRC_Public\BaRC_code\R

# References


- Tophat: http://tophat.cbcb.umd.edu/
- Htseq-Count: http://www-huber.embl.de/users/anders/HTSeq/doc/count.html
- EdgeR: PMID: 19910308
  http://www.bioconductor.org/packages/release/bioc/html/edgeR.html
- DESeq: PMID: 20979621
  http://www.bioconductor.org/packages/release/bioc/html/DESeq.html
- baySeq: PMID: 20698981
  http://www.bioconductor.org/packages/release/bioc/html/baySeq.html
- NOISeq: PMID: 21903743 http://bioinfo.cipf.es/noiseq/doku.php?id=tutorial
- *From RNA-seq to differential expression results*. Oshlack A et al. Genome Biol. (2010). PMID: 21176179.

# Resources

- BaRC Standard Operating Procedures (SOPs)
  https://gir.wi.mit.edu/trac/wiki/barc/SOPs
  https://gir.wi.mit.edu/trac/wiki/barc/SOPs/rna-seq-diff-expressions
- BaRC R scripts
  \wi-files1\BaRC_Public\BaRC_code\R
- BaRC Hot Topic: Assessing Sequence and Microarray Data Quality
  http://jura.wi.mit.edu/bio/education/hot_topics/QC_HTP/QC_HTP.pdf
- BaRC Short Course: Introduction to Bioconductor - microarray and RNA-Seq analysis
  http://jura.wi.mit.edu/bio/education/R2011/slides/Intro_to_Bioconductor_HotTopics_Oct_2011.color.pdf
- BaRC Hot Topic: RNA-Seq: Methods and Applications
  http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

# Contact BaRC for help

- Please stop by with questions or if you would like to analyze your own RNA-seq data

  wibr-bioinformatics@wi.mit.edu