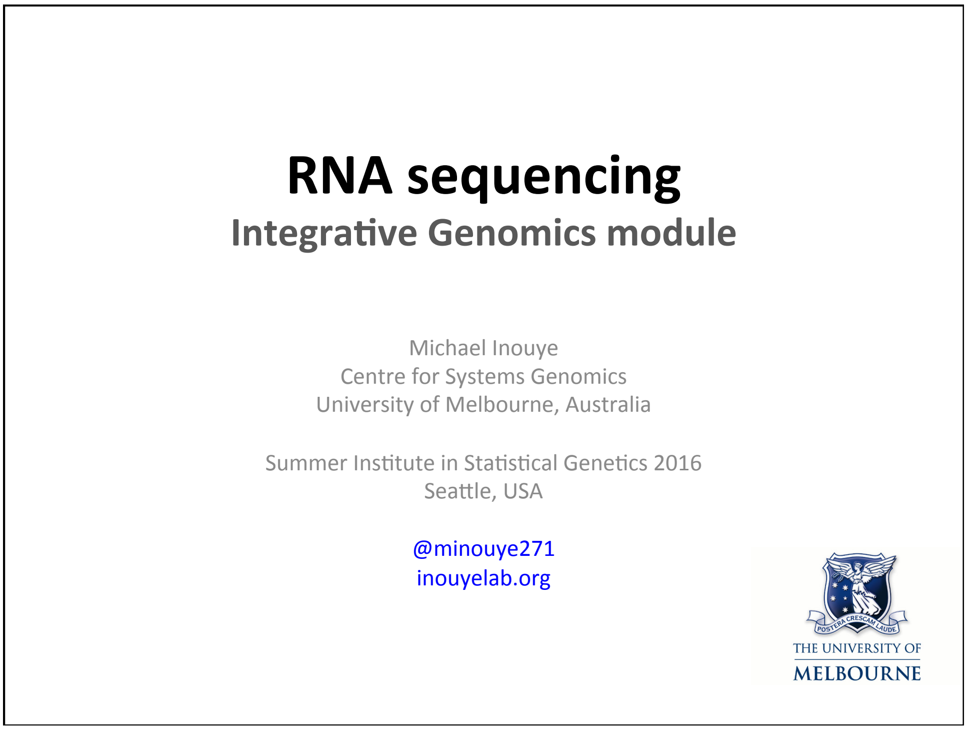


RNA sequencing

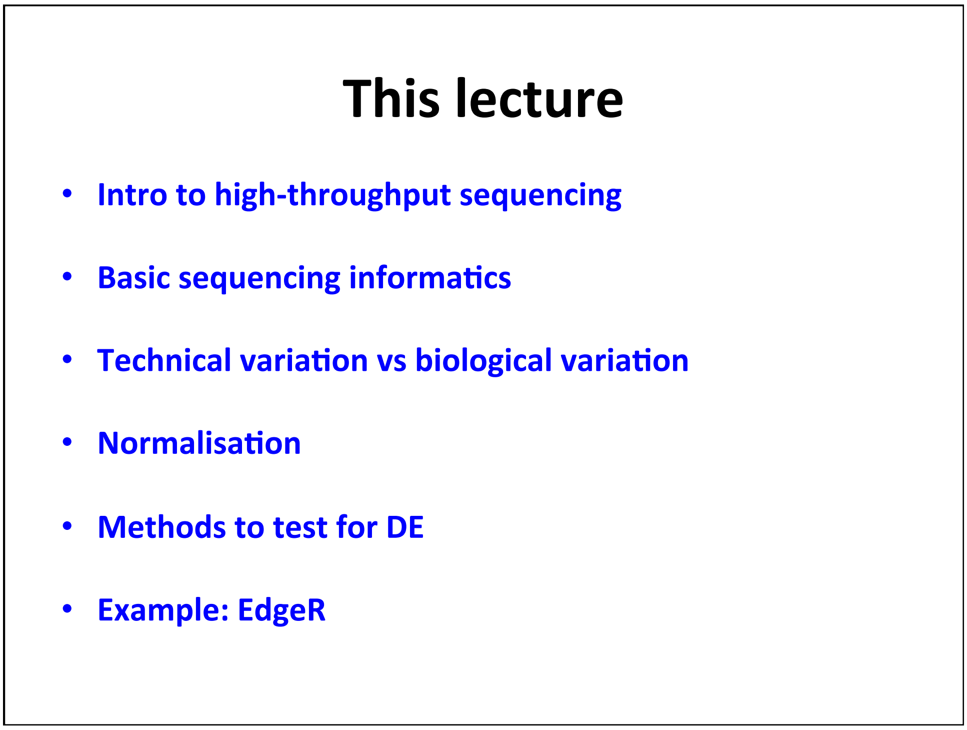
Integrative Genomics module

Michael Inouye
Centre for Systems Genomics
University of Melbourne, Australia

Summer Institute in Statistical Genetics 2016
Seattle, USA

@minouye271
inouyelab.org

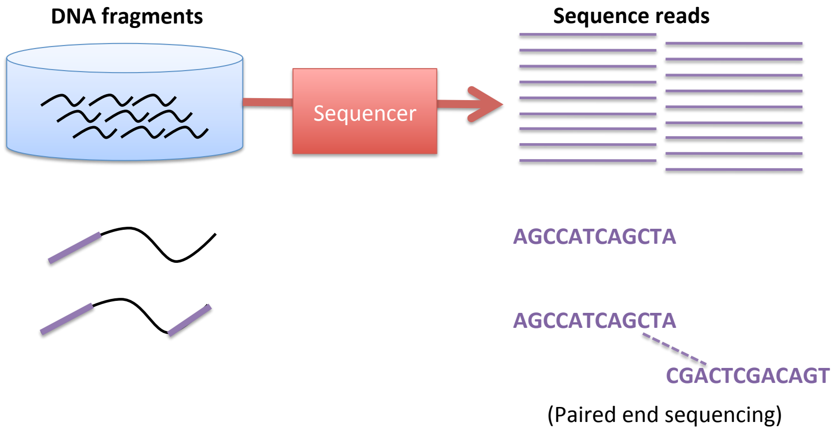
THE UNIVERSITY OF
MELBOURNE



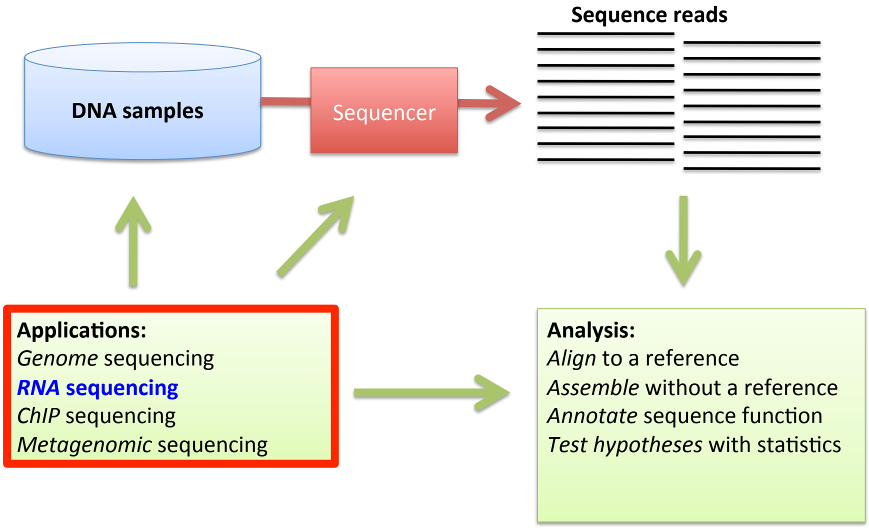
This lecture

- Intro to high-throughput sequencing
- Basic sequencing informatics
- Technical variation vs biological variation
- Normalisation
- Methods to test for DE
- Example: EdgeR

Sequencing experiments

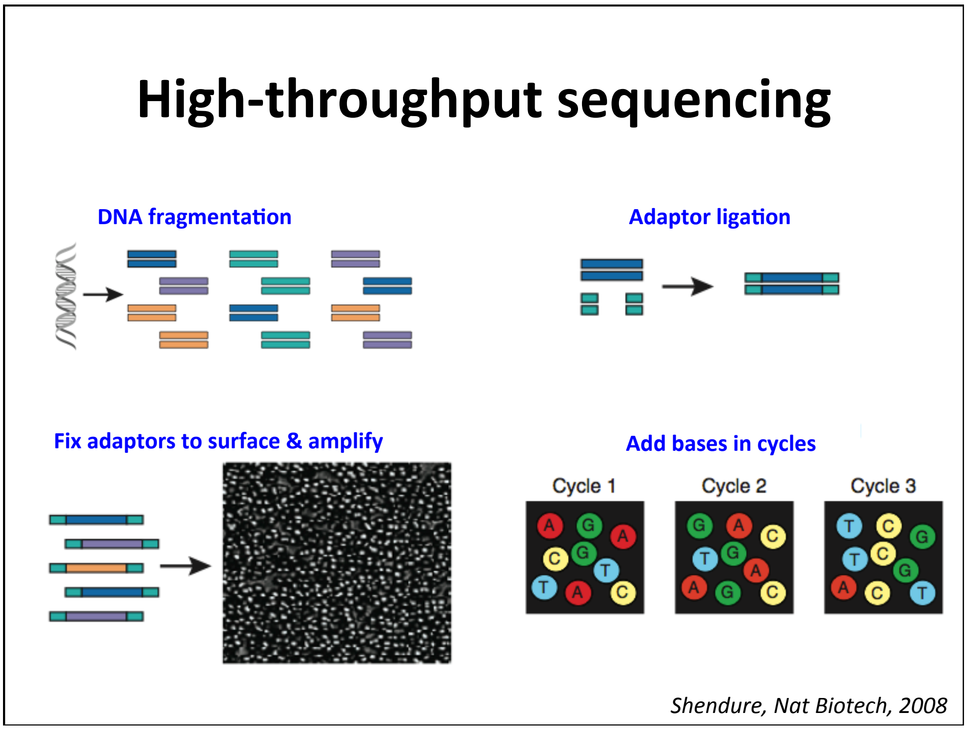


High-throughput sequencing experiments



Applications:
Genome sequencing
***RNA* sequencing**
ChIP sequencing
Metagenomic sequencing

Analysis:
Align to a reference
Assemble without a reference
Annotate sequence function
Test hypotheses with statistics



High-throughput sequencing

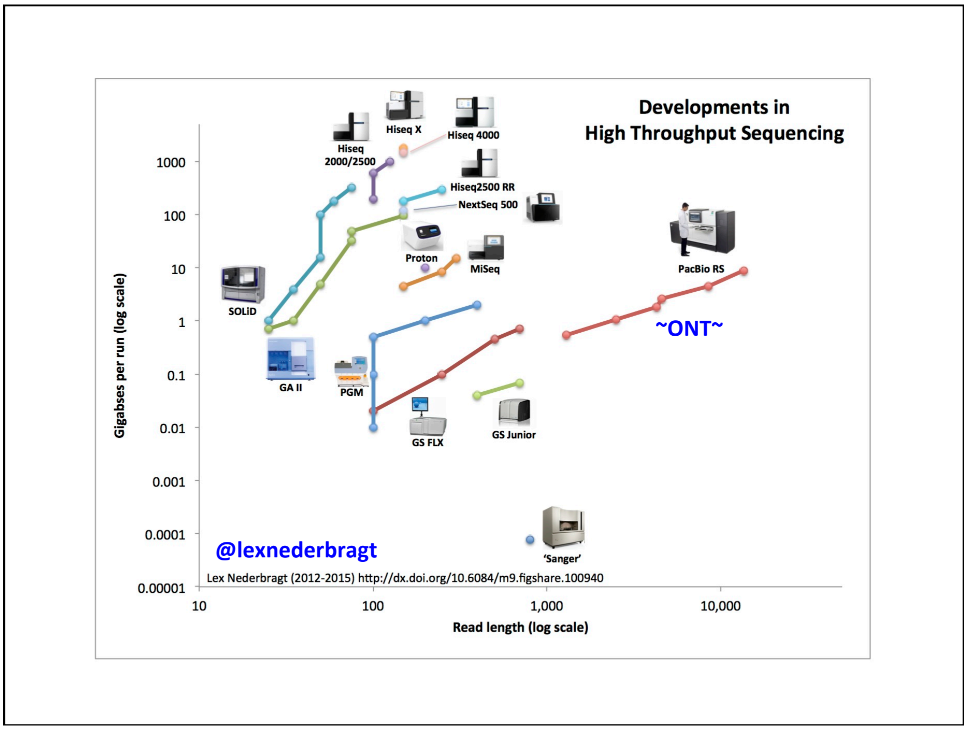
DNA fragmentation

Adaptor ligation

Fix adaptors to surface & amplify

Add bases in cycles

Shendure, Nat Biotech, 2008



Developments in
High Throughput Sequencing

~ONT~

@lexnederbragt

Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940

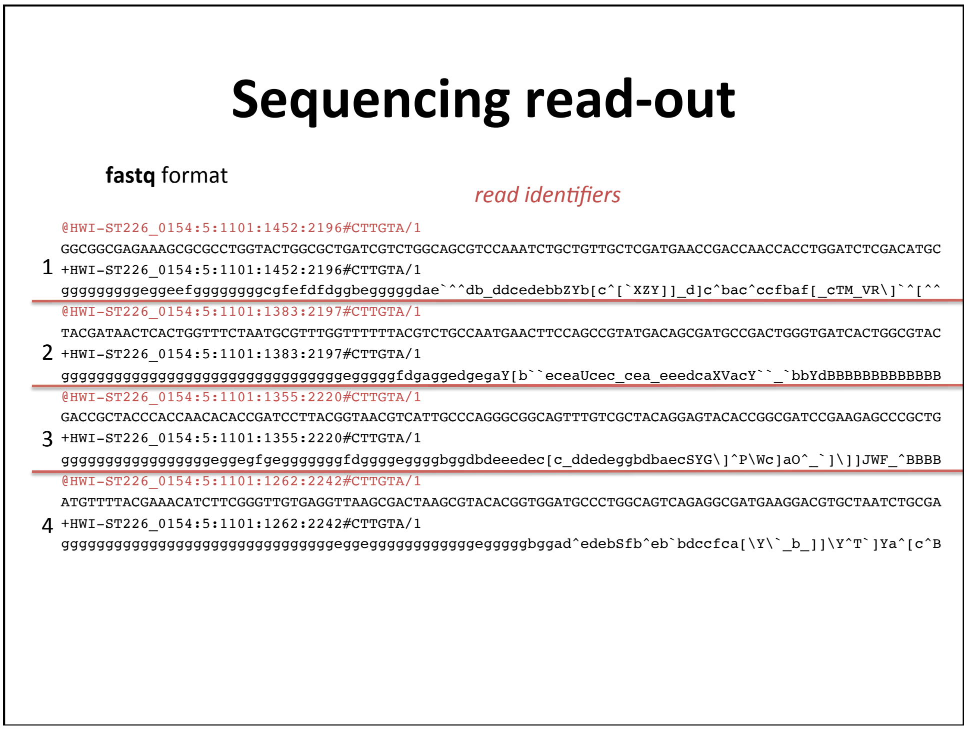
Watch this space

- Many new technologies emerging all the time
- Single cell
- Some day: Long read (1 read -> 1 transcript)
- Review of the latest sequencing technologies
    - Goodwin S et al, *Nat Rev Genetics* 2016. 17:333-351.

Sequencing read-out

fastq format

```
@HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
GGCGGCGAGAAAGCGCGCCTGGTACTGGCGCTGATCGTCTGGCAGCGTCCAAATCTGCTGTTGCTCGATGAACCGACCAACCACCTGGATCTCGACATGC
+HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
gggggggggeggeefggggggggcgfefdfdggbegggggdae`^^db_ddcedebbZYb[c^[`XZY]]_d]c^bac^ccfbaf[_cTM_VR\]`^[^^
@HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
TACGATAACTCACTGGTTTCTAATGCGTTTGGTTTTTTACGTCTGCCAATGAACTTCCAGCCGTATGACAGCGATGCCGACTGGGTGATCACTGGCGTAC
+HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
ggggggggggggggggggggggggggggggggegggggfdgaggedgegaY[b``eceaUcec_cea_eeedcaXVacY``_`bbYdBBBBBBBBBBBBB
@HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
GACCGCTACCCACCAACACACCGATCCTTACGGTAACGTCATTGCCCAGGGCGGCAGTTTGTCGCTACAGGAGTACACCGGCGATCCGAAGAGCCCGCTG
+HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
ggggggggggggggggeggegfgeggggggggfdggggeggggbggdbdeeedec[c_ddedeggbdbaecSYG\]^P\Wc]aO^_`]\]]JWF_^BBBB
@HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
ATGTTTTACGAAACATCTTCGGGTTGTGAGGTTAAGCGACTAAGCGTACACGGTGGATGCCCTGGCAGTCAGAGGCGATGAAGGACGTGCTAATCTGCGA
+HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
ggggggggggggggggggggggggggggggeggegggggggggggggeggggggbggad^edebSfb^eb`bdccfca[\Y\`_b_]]\Y^T`]Ya^[c^B
```

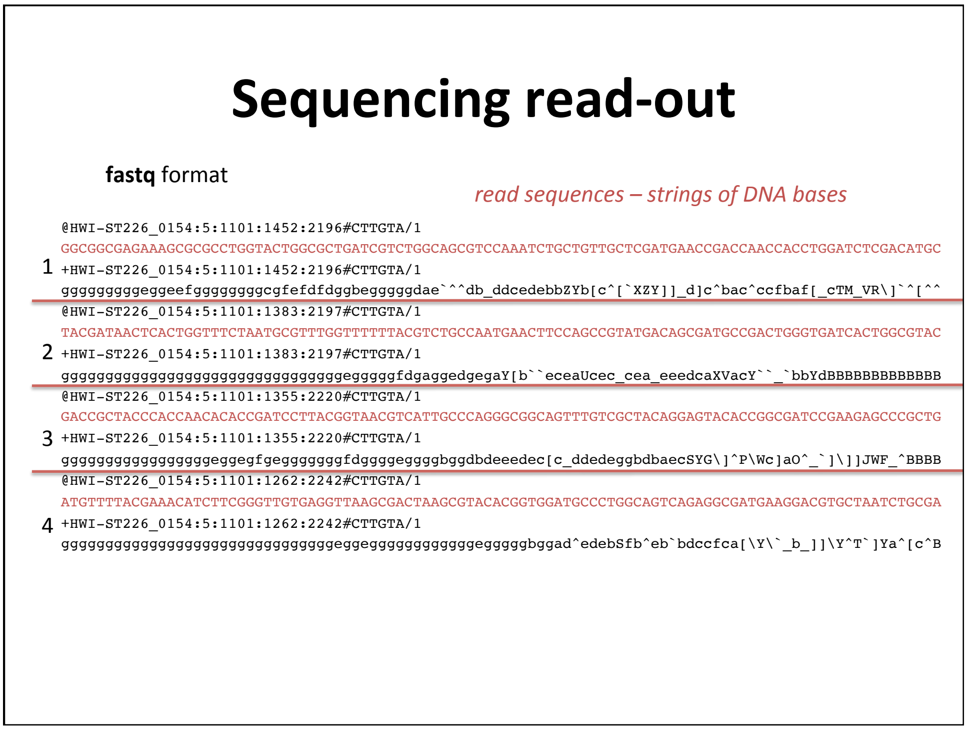


Sequencing read-out

fastq format

read identifiers

```
  @HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
  GGCGGCGAGAAAGCGCGCCTGGTACTGGCGCTGATCGTCTGGCAGCGTCCAAATCTGCTGTTGCTCGATGAACCGACCAACCACCTGGATCTCGACATGC
1 +HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
  gggggggggeggeefggggggggcgfefdfdggbegggggdae`^^db_ddcedebbZYb[c^[`XZY]]_d]c^bac^ccfbaf[_cTM_VR\]`^[^^
  @HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
  TACGATAACTCACTGGTTTCTAATGCGTTTGGTTTTTTACGTCTGCCAATGAACTTCCAGCCGTATGACAGCGATGCCGACTGGGTGATCACTGGCGTAC
2 +HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
  gggggggggggggggggggggggggggggggggegggggfdaggedgegaY[b``eceaUcec_cea_eeedcaXVacY``_`bbYdBBBBBBBBBBBBB
  @HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
  GACCGCTACCCACCAACACACCGATCCTTACGGTAACGTCATTGCCCAGGGCGGCAGTTTGTCGCTACAGGAGTACACCGGCGATCCGAAGAGCCCGCTG
3 +HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
  gggggggggggggggggeggegfgegggggggfdggggeggggbggdbdeeedec[c_ddedeggbdbaecSYG\]^P\Wc]aO^_`]\]]JWF_^BBBB
  @HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
  ATGTTTTACGAAACATCTTCGGGTTGTGAGGTTAAGCGACTAAGCGTACACGGTGGATGCCCTGGCAGTCAGAGGCGATGAAGGACGTGCTAATCTGCGA
4 +HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
  gggggggggggggggggggggggggggggggeggegggggggggggggegggggbggad^edebSfb^eb`bdccfca[\Y\`_b_]]\Y^T`]Ya^[c^B
```



Sequencing read-out

fastq format

read sequences – strings of DNA bases

```
  @HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
  GGCGGCGAGAAAGCGCGCCTGGTACTGGCGCTGATCGTCTGGCAGCGTCCAAATCTGCTGTTGCTCGATGAACCGACCAACCACCTGGATCTCGACATGC
1 +HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
  gggggggggeggeefggggggggcgfefdfdggbegggggdae`^^db_ddcedebbZYb[c^[`XZY]]_d]c^bac^ccfbaf[_cTM_VR\]`^[^^
  @HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
  TACGATAACTCACTGGTTTCTAATGCGTTTGGTTTTTTACGTCTGCCAATGAACTTCCAGCCGTATGACAGCGATGCCGACTGGGTGATCACTGGCGTAC
2 +HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
  gggggggggggggggggggggggggggggggggegggggfdaggedgegaY[b``eceaUcec_cea_eeedcaXVacY``_`bbYdBBBBBBBBBBBBB
  @HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
  GACCGCTACCCACCAACACACCGATCCTTACGGTAACGTCATTGCCCAGGGCGGCAGTTTGTCGCTACAGGAGTACACCGGCGATCCGAAGAGCCCGCTG
3 +HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
  gggggggggggggggggeggegfgegggggggfdggggeggggbggdbdeeedec[c_ddedeggbdbaecSYG\]^P\Wc]aO^_`]\]]JWF_^BBBB
  @HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
  ATGTTTTACGAAACATCTTCGGGTTGTGAGGTTAAGCGACTAAGCGTACACGGTGGATGCCCTGGCAGTCAGAGGCGATGAAGGACGTGCTAATCTGCGA
4 +HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
  gggggggggggggggggggggggggggggggeggegggggggggggggegggggbggad^edebSfb^eb`bdccfca[\Y\`_b_]]\Y^T`]Ya^[c^B
```

Sequencing read-out

fastq format

quality score for each DNA base

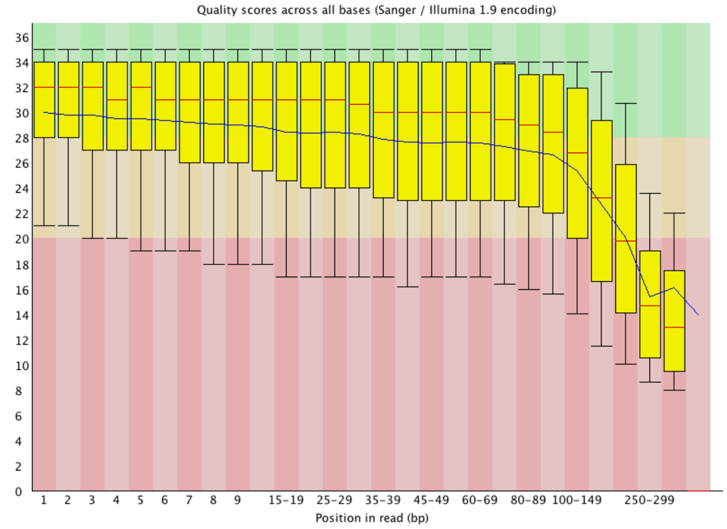
```
  @HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
  GGCGGCGAGAAAGCGCGCCTGGTACTGGCGCTGATCGTCTGGCAGCGTCCAAATCTGCTGTTGCTCGATGAACCGACCAACCACCTGGATCTCGACATGC
1 +HWI-ST226_0154:5:1101:1452:2196#CTTGTA/1
  gggggggggeggeefgggggggggcgfefdfdggbegggggdae`^^db_ddcedebbZYb[c^[`XZY]]_d]c^bac^ccfbaf[_cTM_VR\]`^[^^
  @HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
  TACGATAACTCACTGGTTTCTAATGCGTTTGGTTTTTTACGTCTGCCAATGAACTTCCAGCCGTATGACAGCGATGCCGACTGGGTGATCACTGGCGTAC
2 +HWI-ST226_0154:5:1101:1383:2197#CTTGTA/1
  gggggggggggggggggggggggggggggggggeggggfdgaggedgegaY[b``eceaUcec_cea_eeedcaXVacY``_`bbYdBBBBBBBBBBBBBB
  @HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
  GACCGCTACCCACCAACACACCGATCCTTACGGTAACGTCATTGCCCAGGGCGGCAGTTTGTCGCTACAGGAGTACACCGGCGATCCGAAGAGCCCGCTG
3 +HWI-ST226_0154:5:1101:1355:2220#CTTGTA/1
  ggggggggggggggggeggegfgeggggggggfdggggeggggbggdbdeeedec[c_ddedeggbdbaecSYG\]^P\Wc]aO^_`]\]]JWF_^BBBB
  @HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
  ATGTTTTACGAAACATCTTCGGGTTGTGAGGTTAAGCGACTAAGCGTACACGGTGGATGCCCTGGCAGTCAGAGGCGATGAAGGACGTGCTAATCTGCGA
4 +HWI-ST226_0154:5:1101:1262:2242#CTTGTA/1
  gggggggggggggggggggggggggggggggggggeggeggggggggggggegggggbggad^edebSfb^eb`bdccfca[\Y\`_b_]]\Y^T`]Ya^[c^B
```

Phred score: $Q = -10 \log_{10}P$

where P = probability of an error

Quality score	Prob. error	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

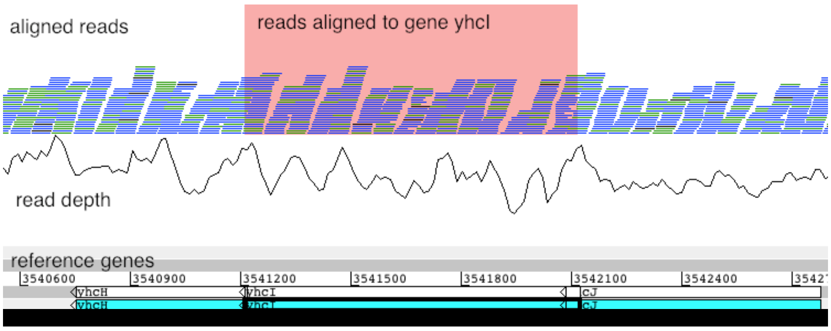
Phred vs read base position



Properties of sequence data to keep in mind

- **Data = Strings of bases + quality scores**
- **Read length**
  - Fixed or variable?
  - Short (e.g. 35bp SOLiD) or long (e.g. 500+ bp 454)
- **Errors**
  - Error rate: how frequent are errors? Phred score distribution?
  - Error profile: what kind of errors are most common?
- **Number of reads**
  - Millions? Hundreds of millions?
  - How much total sequence? How does that compare to genome size?

Read alignment

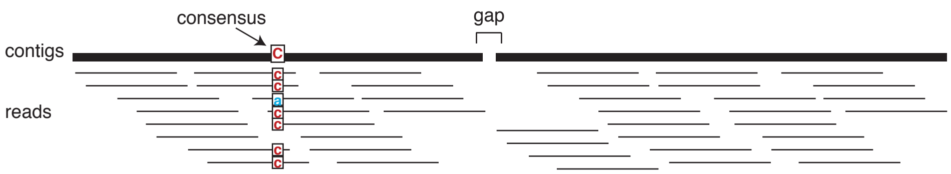


Reference sequence, *similar* to our DNA sample

Outputs:
• what reference sequences are present (e.g. genome variation, **RNA-seq**, ChIP-seq)
• how many copies are there?

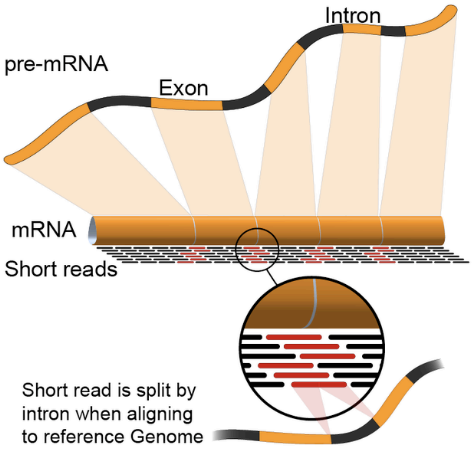
Read assembly

Reference-free, use the new reads alone (*de novo*)
to reconstruct what original DNA sample looked like



Genome sequencing: aim to assemble each chromosome
Metagenomics: aim to assemble DNA fragments from each member of the community
RNA-seq: aim to assemble each mRNA transcript

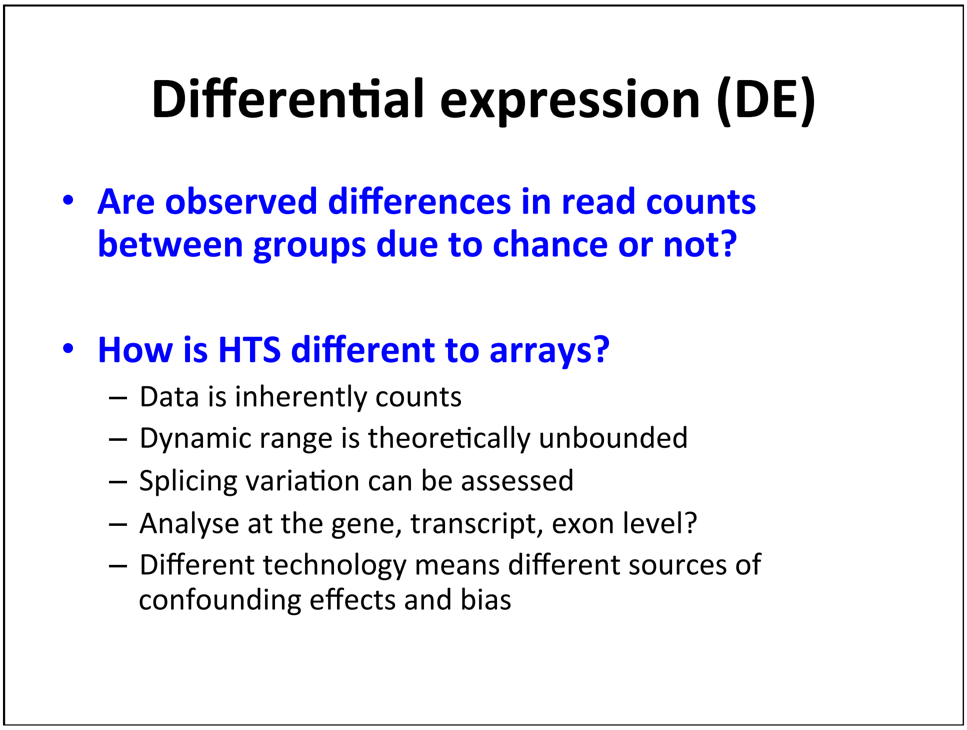
RNA sequencing (RNAseq)


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Input:
cDNA reverse transcribed from mRNA

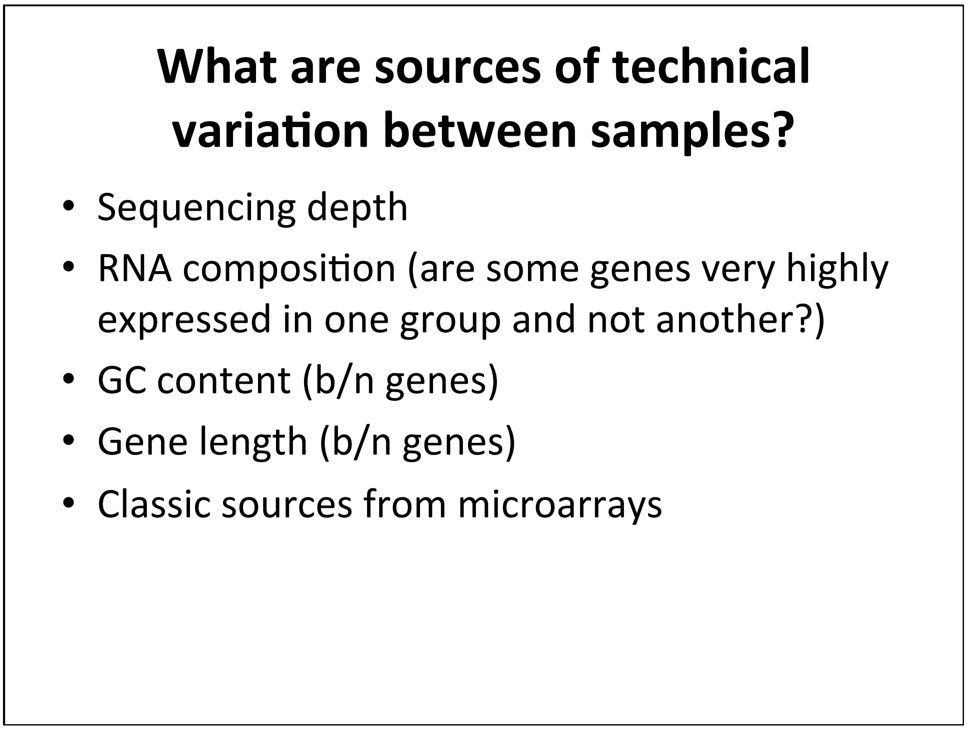
Represents:
all the messenger RNA transcripts present in a set of cells
(i.e. what is being expressed)

Image: Rgocs (Wikimedia Commons)



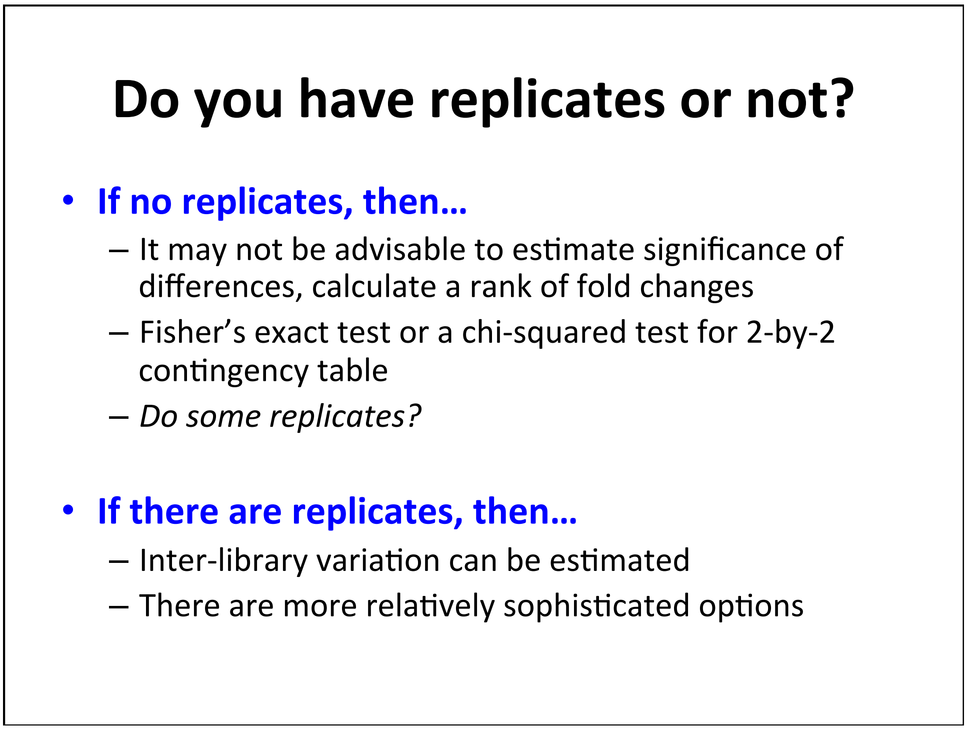
Differential expression (DE)

- **Are observed differences in read counts between groups due to chance or not?**
- **How is HTS different to arrays?**
  - Data is inherently counts
  - Dynamic range is theoretically unbounded
  - Splicing variation can be assessed
  - Analyse at the gene, transcript, exon level?
  - Different technology means different sources of confounding effects and bias



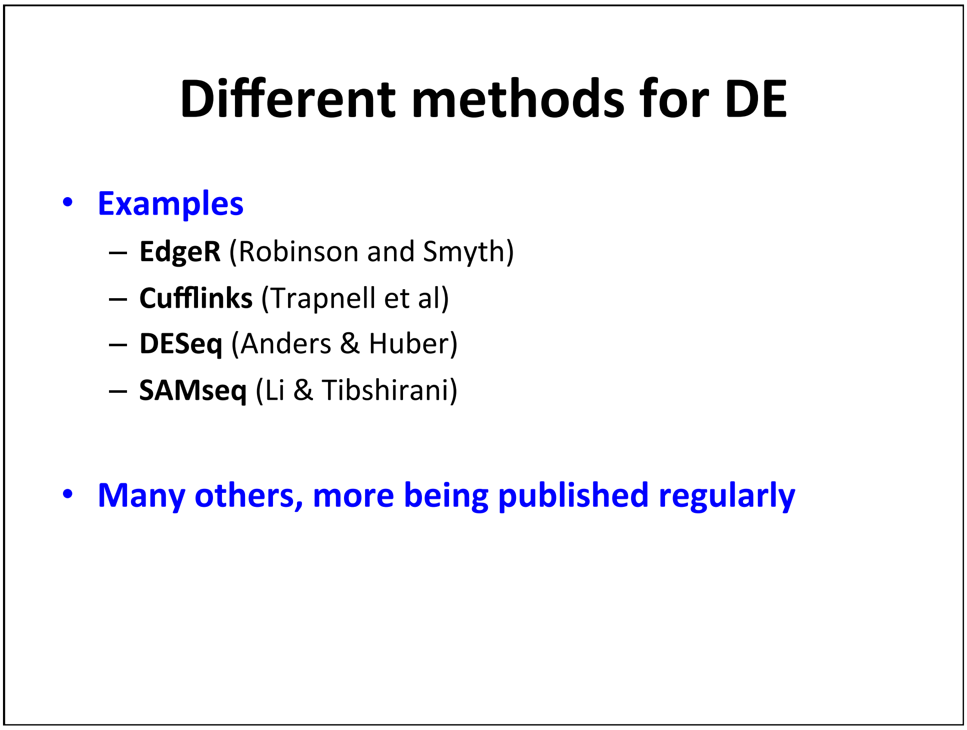
What are sources of technical variation between samples?

- Sequencing depth
- RNA composition (are some genes very highly expressed in one group and not another?)
- GC content (b/n genes)
- Gene length (b/n genes)
- Classic sources from microarrays



Do you have replicates or not?

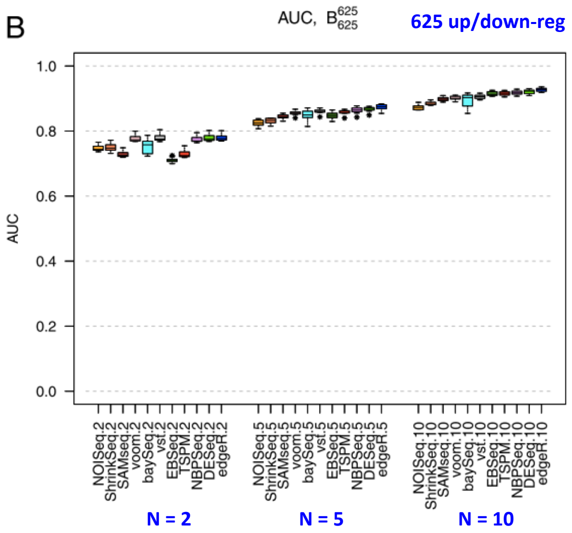
- **If no replicates, then…**
  - It may not be advisable to estimate significance of differences, calculate a rank of fold changes
  - Fisher's exact test or a chi-squared test for 2-by-2 contingency table
  - *Do some replicates?*
- **If there are replicates, then…**
  - Inter-library variation can be estimated
  - There are more relatively sophisticated options



Different methods for DE

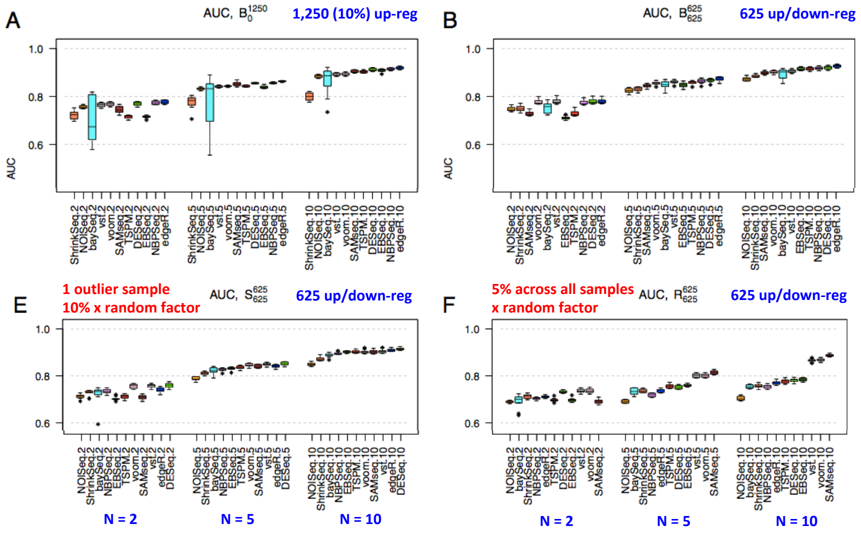
- **Examples**
  - **EdgeR** (Robinson and Smyth)
  - **Cufflinks** (Trapnell et al)
  - **DESeq** (Anders & Huber)
  - **SAMseq** (Li & Tibshirani)
- **Many others, more being published regularly**

How does one choose a method?



Modified from Soneson & Delorenzi, *BMC Bioinf* 2013

How does one choose a method?



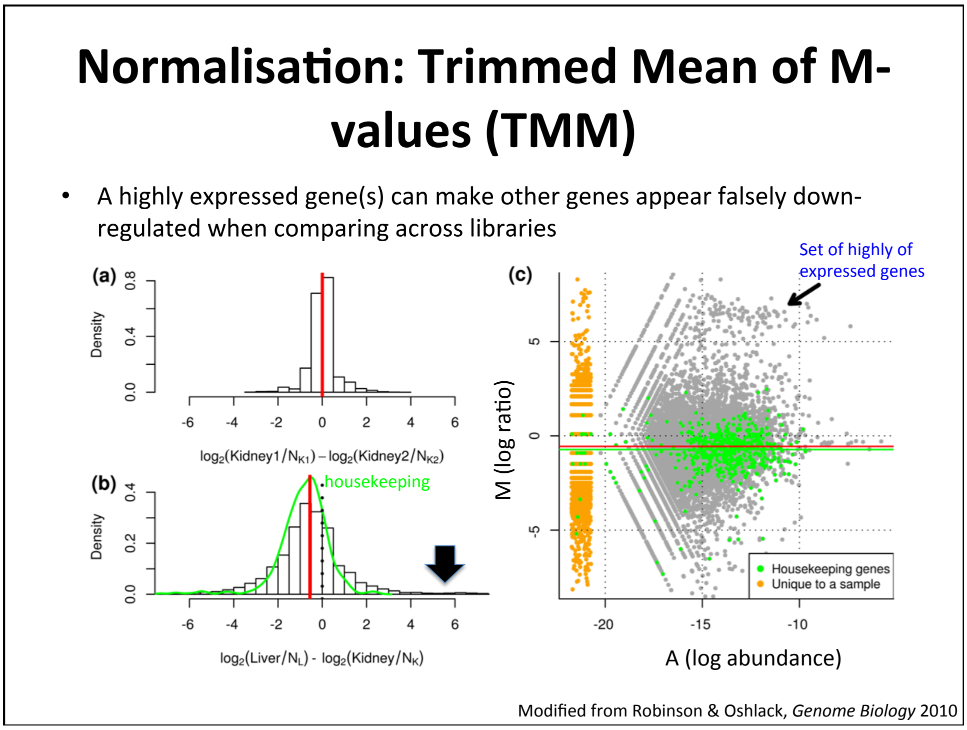
Modified from Soneson & Delorenzi, *BMC Bioinf* 2013

Example: EdgeR

- What are the inputs?
  - A table of counts (matrix)
    - Rows as 'genes'
    - Columns as samples (libraries)
  - A list of group assignments for each sample (vector)

Normalisation

- Explicit scaling by library size
  - TMM normalisation
- Other normalisation factors can be included in model

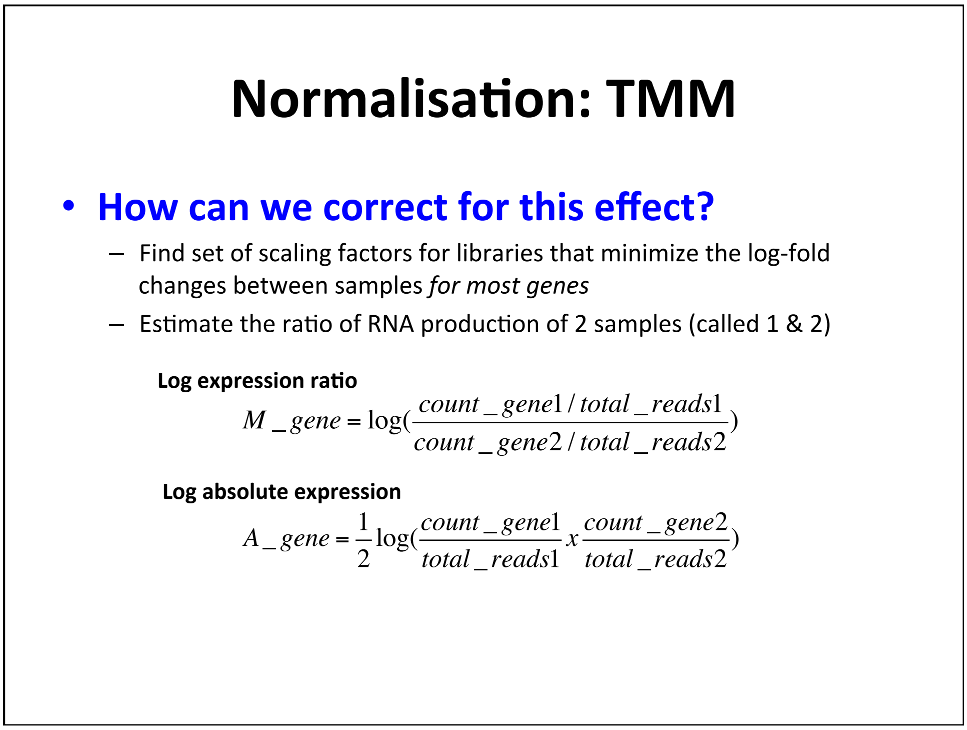


Normalisation: Trimmed Mean of M-values (TMM)

- A highly expressed gene(s) can make other genes appear falsely down-regulated when comparing across libraries

Set of highly of expressed genes

Modified from Robinson & Oshlack, *Genome Biology* 2010



Normalisation: TMM

- **How can we correct for this effect?**
  - Find set of scaling factors for libraries that minimize the log-fold changes between samples *for most genes*
  - Estimate the ratio of RNA production of 2 samples (called 1 & 2)

Log expression ratio

$$M\_gene = \log(\frac{count\_gene1 / total\_reads1}{count\_gene2 / total\_reads2})$$

Log absolute expression

$$A\_gene = \frac{1}{2}\log(\frac{count\_gene1}{total\_reads1} x \frac{count\_gene2}{total\_reads2})$$

Normalisation: TMM

- Trimmed Mean of the M values (TMM) is weighted average after removing the upper/lower N% of the data (typically 25% for M, 5% for A)
- Weight of a gene is the inverse of its estimated variance
- After trimming, calculate the scaling factor for library 1 (compared to library 2) as

$$\log(TMM) = \frac{\sum_{gene\_i \in G^*} (weight\_gene\_i)(M\_gene\_i)}{\sum_{gene\_i \in G^*} weight\_gene\_i}$$

If there's no RNA composition effect, then TMM = 1

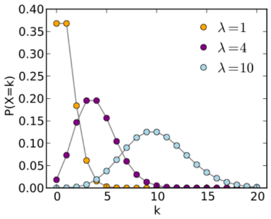
The *effective* library size (TMM x library_size) is then used in all downstream analysis

EdgeR model

- We're interested in read counts for a gene across replicates
- Variation in relative gene abundance is due to **biological causes + technical causes**
- Because the data is counts, we'll usually think it's Poisson distributed, and

Total CV^2 = Technical CV^2 + Biological CV^2

- What is a Poisson distribution?

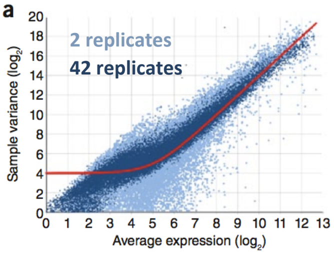


Wikipedia

Expected value = mean (λ) = variance

EdgeR model: Why not use a Poisson?

- **Assumption that mean = variance is strong**



- **In RNAseq, observed variation is typically greater than the mean**
  - That is, the data is 'overdispersed'

- **How can we handle overdispersion?**

Alternative: Negative binomial (gamma-Poisson)

- **Assume true expression level of a gene is a continuous variable with a gamma distribution across replicates**
  - Implies that the read counts follow a negative binomial distribution (a discrete analogue of gamma)

- **NB is parameterised by mean and r (dispersion parameter)**
  - Note the extra parameter (compared to Poisson) which handles variance independent of the mean
  - Biological CV is sqrt(r)

EdgeR model: Estimating the dispersion parameter

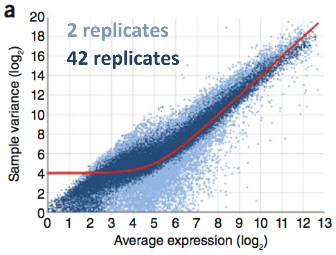
- **Why is this important?**
  - Overestimation likely means a conservative DE test
  - Underestimation likely means a liberal DE test

- **Many methods**
  - Maximum-likelihood (ML)
  - Pseudo-likelihood
  - Quasi-likelihood
  - Conditional ML (if libraries are equal size)
  - Quantile adjusted conditional ML (qCML)

- **Bottom line is a big simulation study was performed**
  - HTS data: many genes, means, variances, library sizes
  - qCML was most accurate across all scenarios
  - Robinson & Smyth *Biostatistics* 2008

EdgeR model

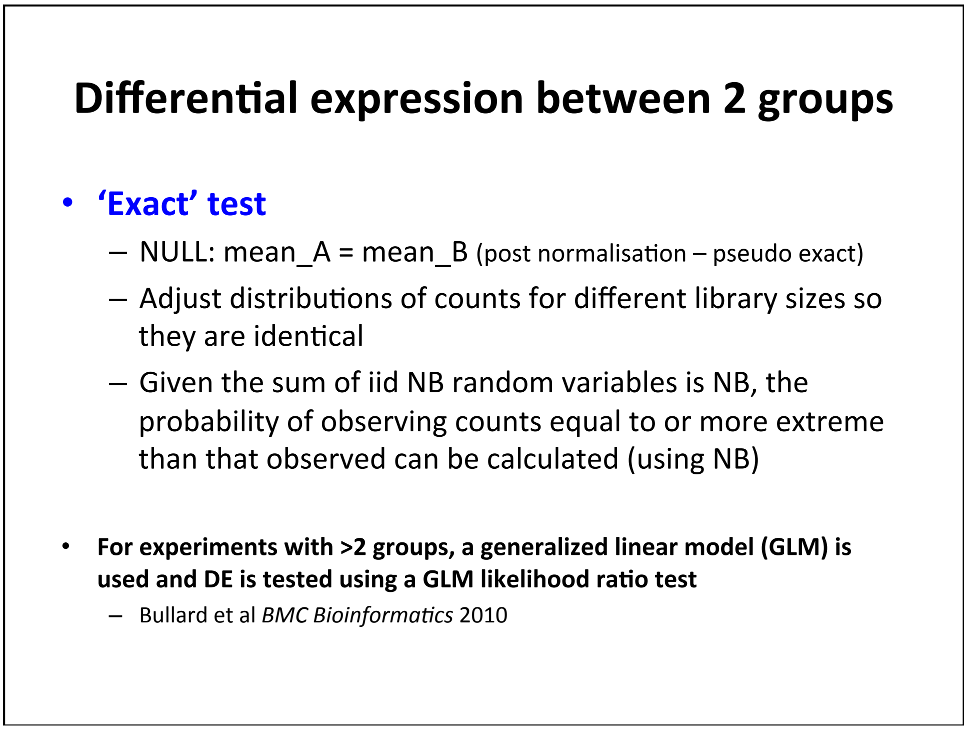
- Genes have different mean-variance relationships, so dispersion isn't same across genes



- Initially edgeR estimates 'common' dispersion across all genes then applies an empirical Bayes approach to shrink gene-specific dispersions toward the 'common'

- **Why do we care?**
  - Allows us to make weaker assumptions about mean-variance and thus **makes model more robust to outlier genes**

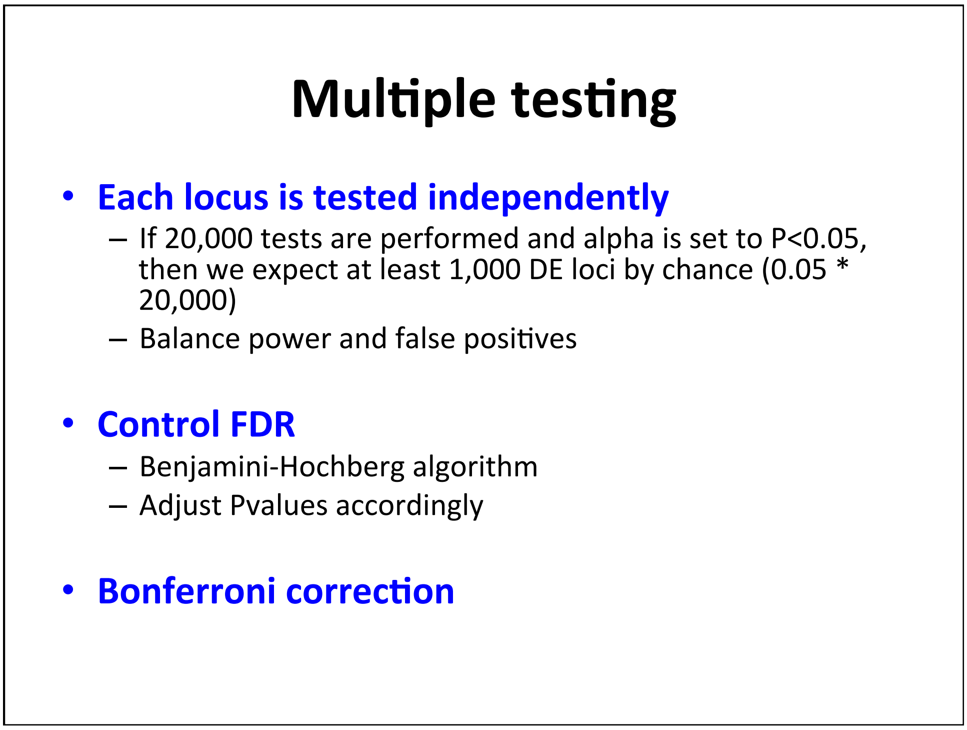
Subramaniam & Hsiao, *Nat Imm* 2012



Differential expression between 2 groups

- **'Exact' test**
  - NULL: mean_A = mean_B (post normalisation – pseudo exact)
  - Adjust distributions of counts for different library sizes so they are identical
  - Given the sum of iid NB random variables is NB, the probability of observing counts equal to or more extreme than that observed can be calculated (using NB)

- **For experiments with >2 groups, a generalized linear model (GLM) is used and DE is tested using a GLM likelihood ratio test**
  - Bullard et al *BMC Bioinformatics* 2010



Multiple testing

- **Each locus is tested independently**
  - If 20,000 tests are performed and alpha is set to P<0.05, then we expect at least 1,000 DE loci by chance (0.05 * 20,000)
  - Balance power and false positives

- **Control FDR**
  - Benjamini-Hochberg algorithm
  - Adjust Pvalues accordingly

- **Bonferroni correction**

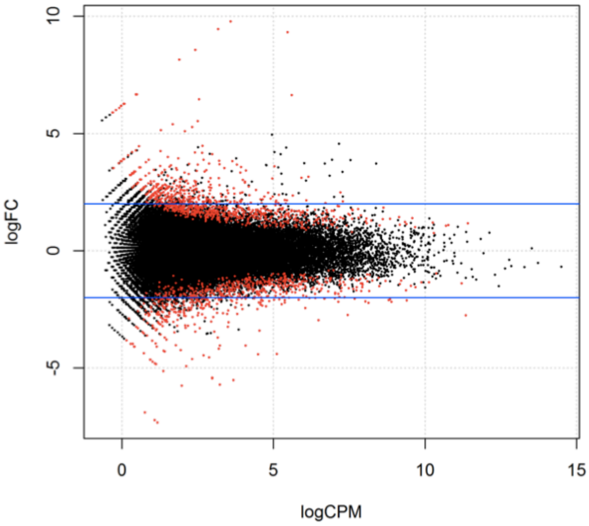
What output are we interested in?

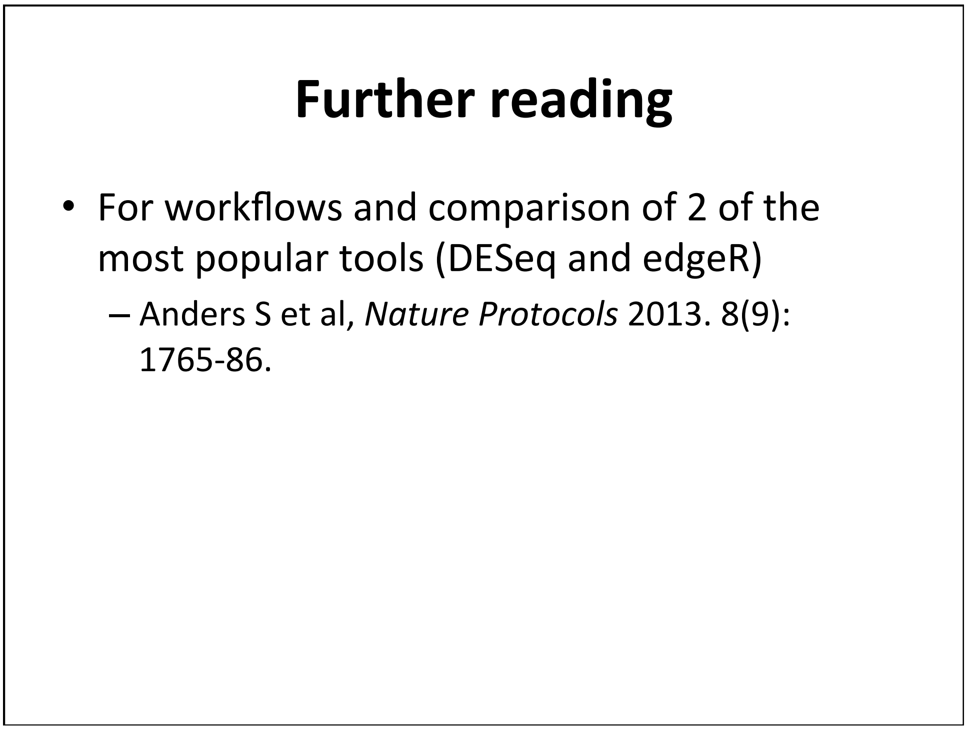
```
                 Length logFC logCPM     PValue        FDR
ENSG00000151503   5605  5.82   9.71  0.00e+00   0.00e+00
ENSG00000096060   4093  5.00   9.94  0.00e+00   0.00e+00
ENSG00000166451   1556  4.66   8.83 1.15e-228 6.31e-225
ENSG00000127954   3919  8.17   7.20 1.00e-209 4.14e-206
ENSG00000162772   1377  3.32   9.74 2.09e-182 6.91e-179
ENSG00000113594  10078  4.08   8.03 5.07e-153 1.39e-149
ENSG00000116133   4286  3.26   8.78 6.33e-148 1.49e-144
ENSG00000115648   2920  2.63  11.47 2.82e-139 5.81e-136
ENSG00000123983   4305  3.59   8.58 8.38e-138 1.54e-134
ENSG00000116285   3076  4.22   7.35 1.05e-135 1.73e-132
```

CPM – Counts per million (not formally used in edgeR DE)

FPKM (cufflinks) – Fragments Per Kb of transcript per Million mapped reads
inferred using a statistical model

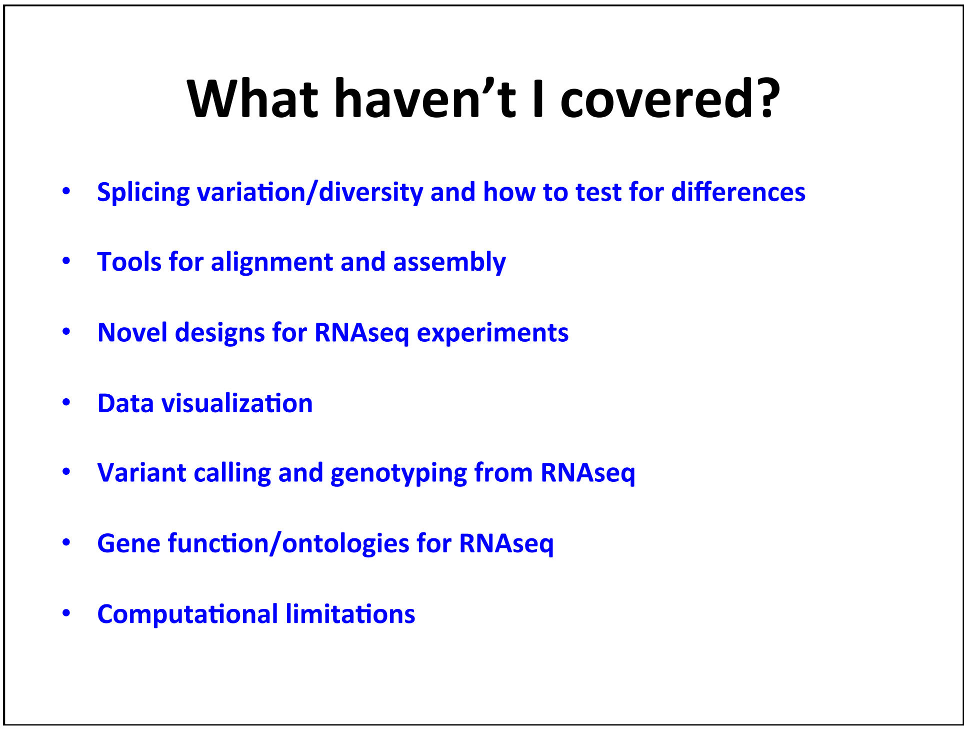
Smear plot





Further reading

- For workflows and comparison of 2 of the most popular tools (DESeq and edgeR)
  - Anders S et al, *Nature Protocols* 2013. 8(9): 1765-86.



What haven't I covered?

- **Splicing variation/diversity and how to test for differences**
- **Tools for alignment and assembly**
- **Novel designs for RNAseq experiments**
- **Data visualization**
- **Variant calling and genotyping from RNAseq**
- **Gene function/ontologies for RNAseq**
- **Computational limitations**