# RNA-seq data processing and analyses

## RNA sequencing, transcriptome and expression quantification

Olga Dethlefsen

NBIS, Stockholm University

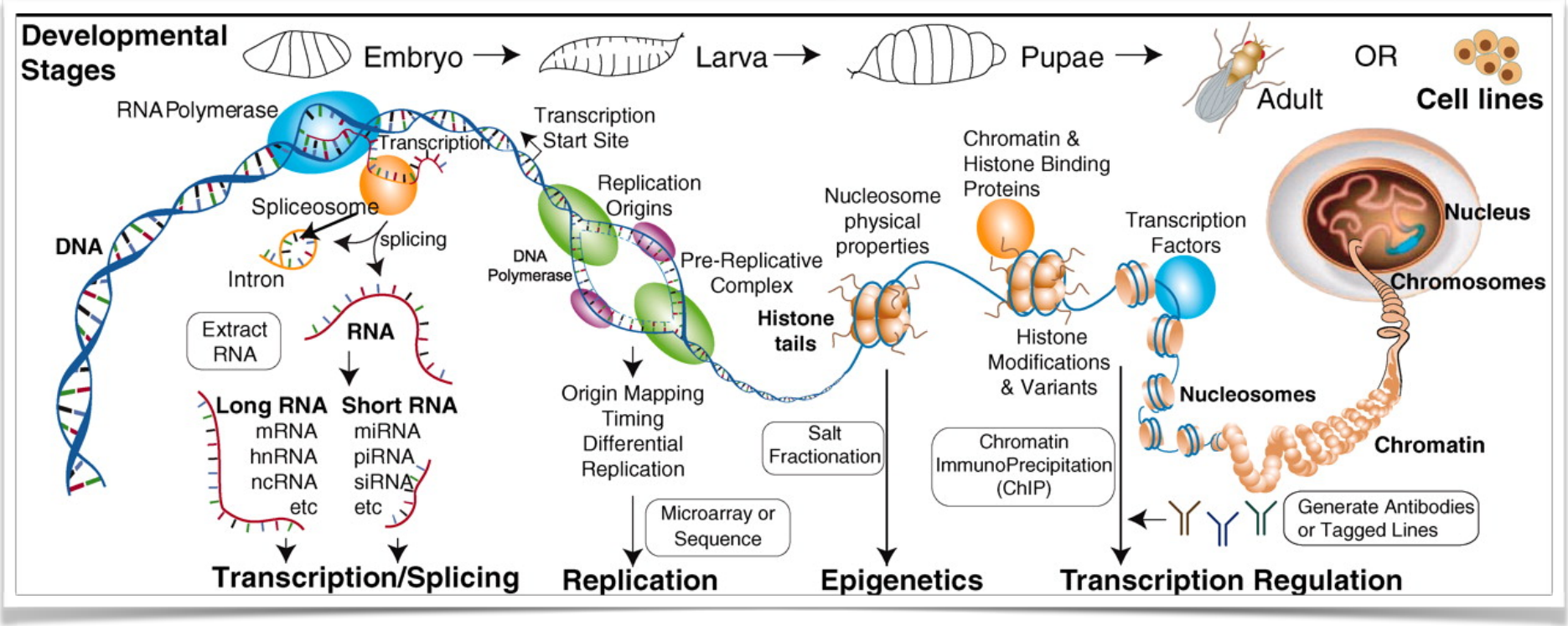
December 2016

# Outline

1. Why study transcriptome?
2. Overview of RNA-seq work flow
3. Understanding sequencing output
4. RNA-seq data analysis
   - Mapping based approach
   - Transcriptome assembly
5. Exercises

# Why study transcriptome?

# A complex view of central dogma of molecular biology



# Advantages and Applications

## Transcriptomes are

- **dynamic**, that is not the same over tissues and time points
- **directly derived** from functional genomics elements, mostly protein-coding genes, providing a useful functionally relevant subset of the genome, that is **smaller sequence space**

## Transcriptomes enable

- to investigate **differences in gene expression** patterns
- to distinguish different isoforms and allelic expression
- to explore gene functions
- to analyze single nucleotide variants, fusion genes and co-expression networks

# Advantages and Applications

## Transcriptomes can also help to

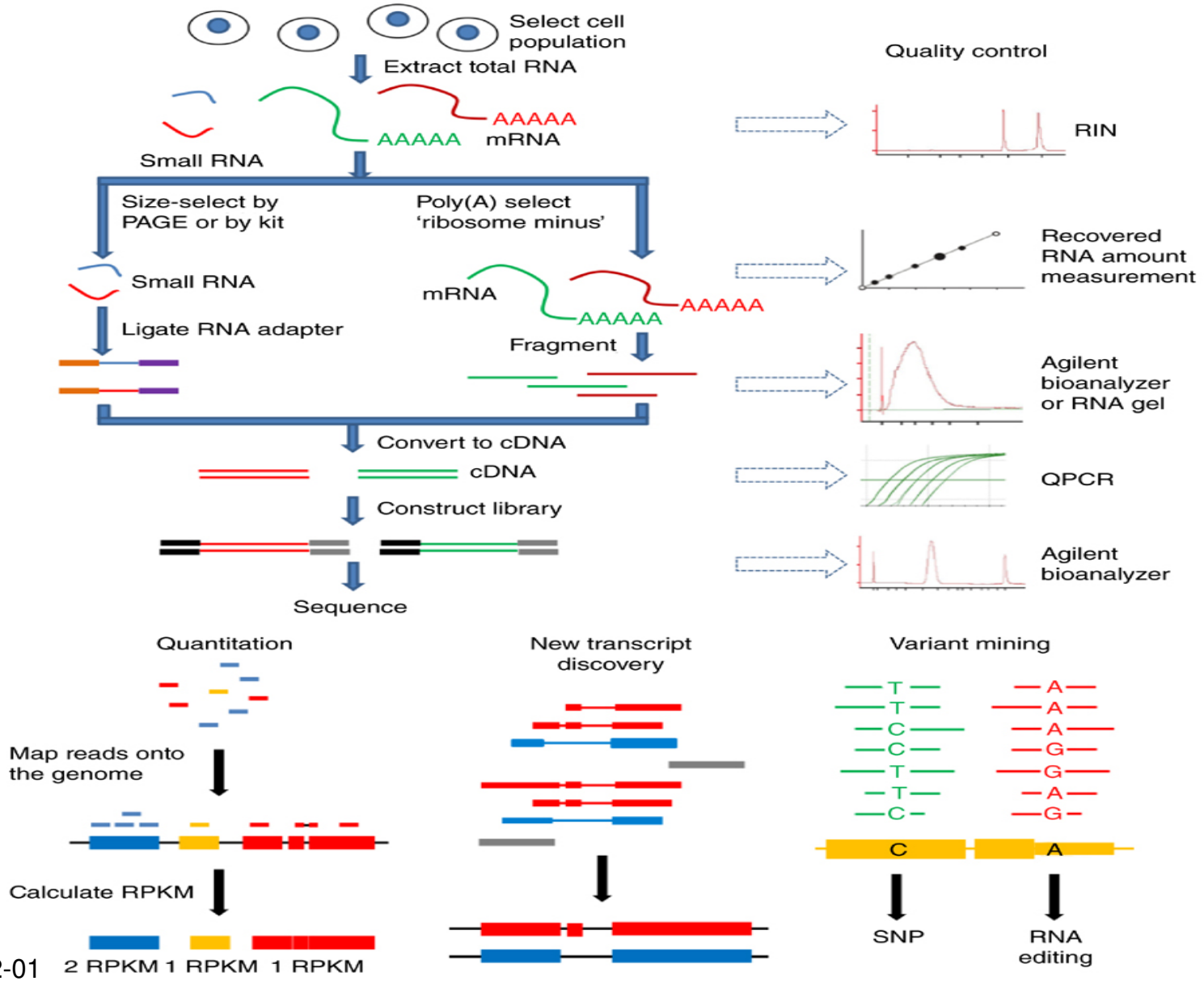
- understand host-pathogen immune interactions and predict resistance to specific antibiotics
- understand tumour classification and progression by determining which variants are expressed in cancer samples
- understand tumour heterogeneity and clonal evolution (scRNA-seq)
- understand complex tissues, e.g. neural (scRNA-seq)
- study other biological questions in which cell-specific changes in transcriptome are important, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression etc.

# Overview of RNA-seq work flow

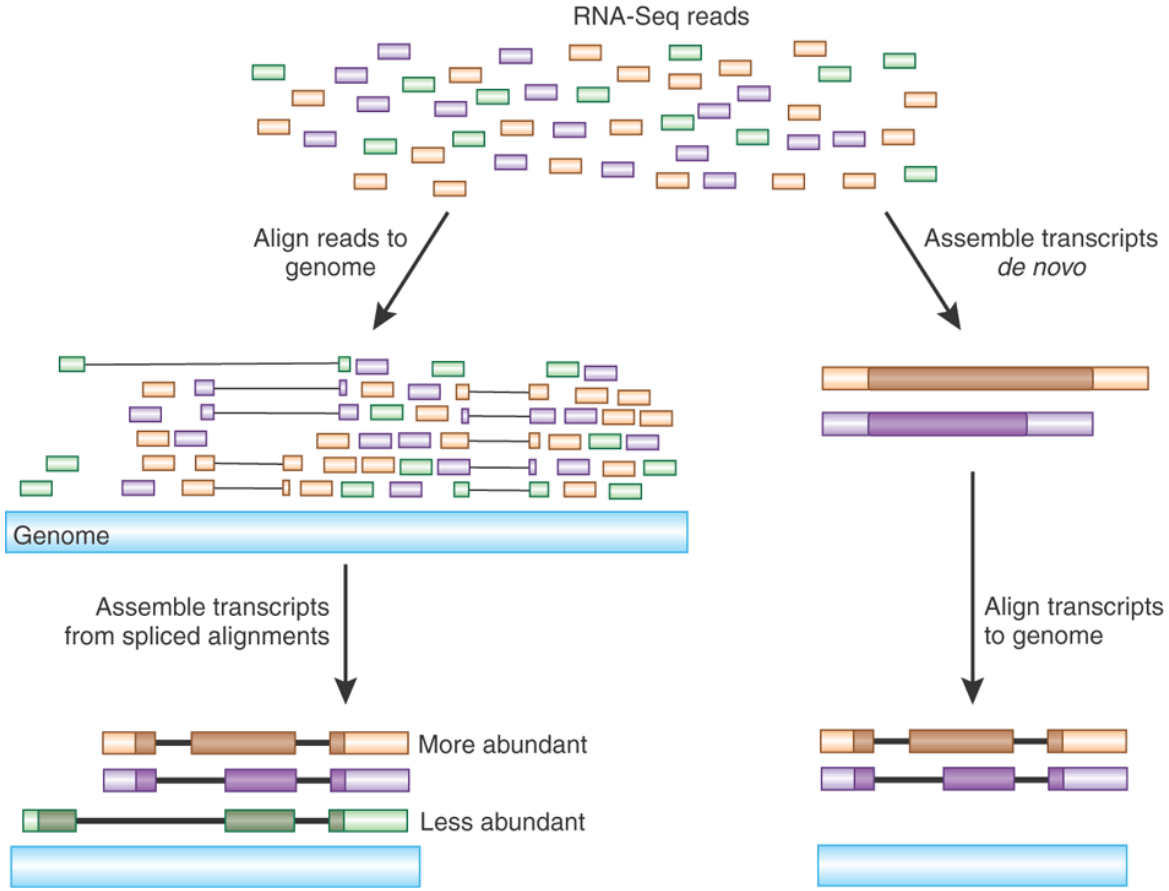
# High level work flow overview

- Experimental design (biology, medicine, statistics)
- RNA extraction (biology, biotechnology)
- Library preparation (biology, biotechnology)
- High throughput sequencing (engineering, biology, chemistry, biotechnology, bioinformatics)
- Data processing (bioinformatics)
- Data analysis (bioinformatics & biostatistics)

# More detailed work flow overview

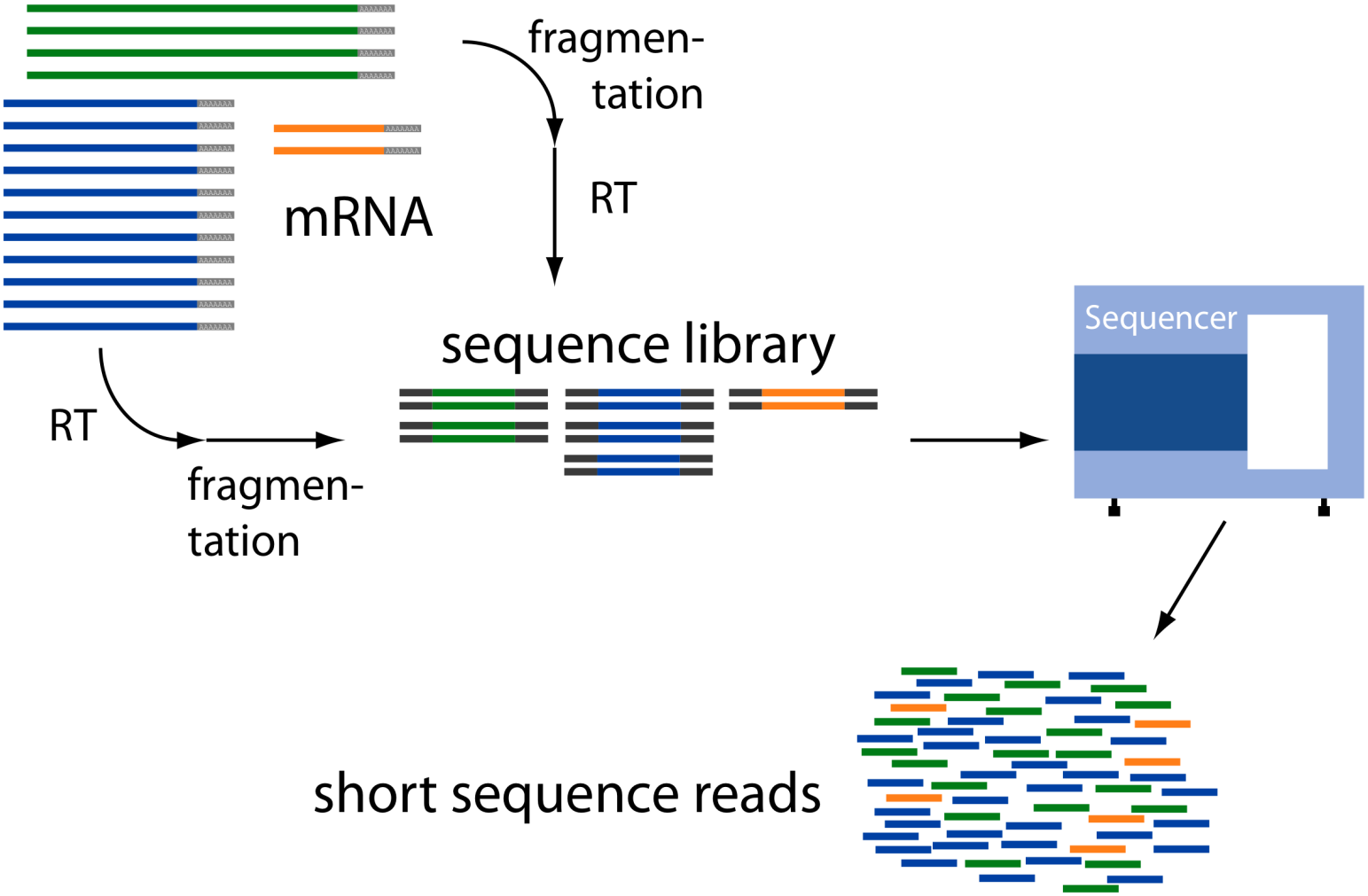


# Two main bioinformatics routes



# Understanding sequencing output

# NGS data


fragmen-
tation
RT
mRNA
sequence library
Sequencer
RT
fragmen-
tation
short sequence reads

# .fastq Machine output

fastq — less — 195×69

```
@HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
AGGCACTCCCTGCAGGTGTTGGACCACCTGGCTGAGCCACAGCGTCGCTTCCTGCTGCCAGGGCCTCGGAGAGGGTGGCTGTGGAGACACTGTGGGAGCA
+HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
^_P\`ccceeceeeee[b[beedaae_fddddc_cfhheedfeeh__`aeadd`d]baccc\[TKT\]_\ZQT^a[W[^^aW`^`aX^X^`_Y]^aBBBB
@HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
TCTTTATTGGCATCAGGCATCACCACACCATGGTTCTTGGCTCCCATGTTGGCCTGGACTCTCTTGCCATTCCGGGATCCTCTCTCATAGATGTACTCGC
+HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
__P`cccegge]eghhhhdfhhhhhhhhhfhhefghffffhfhhfheg^eeffgfegf`fghhhffhhggadcX[`bbbbbbbbbcbbbcbR]aabaa
@HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
CCGAAACCCCGAAAGCACCCCAAAATCCCTGTGGGGAACCCCGAAAATCCCGAAATTACCCCAAAATACCTGTGGGATACCCTGAAAACCCGAAAGCACC
+HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
__[VJ\`\e[eefdgbafagfffagfd`Rc[cac`a_ef[a_N`]aced]\X]Z^RGYYYXa^_``bb_YYYbBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
GCTCAAGTCCCGGAGGAGGTCAGAGCTGGCATCTCTTCCCCAGCTGCTGCTCAGGAGTGTAAGCACCTGCAAACAGCTGCCAGCCAGGGAGCTGTGACTT
+HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
__J\`accccec[eagag`gggedbffhhffgfhhhhheaaefaghhhfdghhhdfd`ddgbd]_^abbbb__^ababbGXY_[aa^`aOOT[`bbGYYS
@HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
CAGCNGCTCTGGGCACCCTGTGCCAGGGCNTGNCCACCCTCCCAGCCAAGAATTCCTTCCCNATATCTAACCCAAATTTCTTCCCNGTAGGAGCAGGATG
+HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
Z_[aBQQQ`ccace_d_Y`a_Xd^ecc]fBPYBOYacedeZeVRbWVW\_\bcS\bdde`VBKKT^accab]GT\Z_YY``_]YGBKKWWO]`_W^[W_R
@HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
CAGATGAGGACTTTTGCTCCAAATGGGAAAGGAGAAAACCTCAGTCCGTAGAGATGCTCCCAGAGATAGTCCTAAAGATGAAGATGATGAAGACCCTTTG
+HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
abbeeeeeggggiihiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiihiiiihighiihiihichifggdgeebaceebbddbbdcecdcccbccbccc
@HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATACTAAAATGAAAGGAAACAAAGTGAATGAAGTACTGAATAGATTACACT
+HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^adghWWaff_H_cbdbbd\dbddV^_ZRMHHZGUZ_b_YRTGTT]`__b[
@HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
CTCACGATGGTCCCCAGGCTGTCCACAGTTGCCACACACTGATAATATCCTTCATCAGGTTTATTATGCTTGGAATGCACCACACTGTTAATTAATAAAG
+HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
^__ccc\ccY`^ae`Z_`bR`b]fa]dec^ceeeffc^fcdceXc]cehehaebefd`eW\\b]bebeeedde`R\_aa_c]b\baaZ`accdc[`]a`a
@HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
GCCAGTACAGCTGTAGTAGTCTGTCCTTCCCATCCGTGCCCATGTGACACAGCAGGTTCACAGCATGGTGACCAGTTTGAAGCTTCCTACCTCTGTGGTA
+HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
bbbeeeeegfggghfhhhihiiiihiiiiiiiiiighiiiiiiiiiiihiihghiiiiiihhhbdghgggggeeceeebddcddcccccbcccc
@HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
GTTCAGACAAGTTCGATCTCTTGTGCATCGACTGTGCTGGATGATAGTTTTTCAGTGAGTATTATGGTTAGTAGATATAGTACCAGGCTGCAAATAGCTA
+HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
a_P\cceeggggiighihiiiighhiiihhihiiiiiihihiiehhiiifhhibfaedfhiiifghih____dgeeddgeeeeeddc_bbccccbb
@HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
CTCGAAATCCAGGGCAACGTAGCACAGCTTCTCCTTGATGTCACGCACAATTTCTCTCTCAGCTGTGGTGGTGAAGCTGTAGCCTCTCTCTGTCAGGATC
+HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
_aaeceeeggggdfgfihghffhhhiiihffgiiiiiihihhfiigghdgdhffhiiifdhihd`^bV^aabb__bdc_]bZ`bccccccccb]`bcc
@HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
TAACAGTCCCCCTGGTATGAAATGGCACCTTGGTTACACTGAGGGAGGGGTGAGGTTACAGGGAGTAATTTTCATGTGTAACTGGGGTTAAAAAAAAAAA
+HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
_bPaceeegfggcghfgfhhiiiiiiffh_fghghhfhhhgfghfcghhiT_bddddddeeeeaac`bbcccb`cb`cbbc`ccbcccaacbbbbcaaccc
@HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
CCTGGATGGTGGGCTGATCAACTTTGAGAAGAGAAGGAAGGAGTTCGAAGTCATCGCGCAGATCAAGCTGCTCCAGTCGGCCTGCAACAACTACAGCTTC
+HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
_bbeeeeegeggiiiiiiiiiiiiihidghhihiiiiiiiiggghhhhihiiiiiihigeeddddddcccccccbcccccccaccccccccccccbbccc
@HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
CAGTACTCTTTGTACCGCTCATCTGCATCTCCAAACACTTTGTACCTGCTGCCTTTTATTTTGTATGTTTACCTGTGTCAGAGAGTCGCCAAGTTTGTTC
+HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
abbeeeeefgggghifiihiiihigaghefhhdgheghhhhf^afgffhfhiiihidfhhihhhiiggbg`gfdg`bdddeade`]`_Z^Y[]bcbccb
@HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1
AACAAAACGGGCTGTTTTAGGACCCTTGGTCCCAAGGGGTAATGGCCCTCAGCACCCACTATCCCTGCTCTCCAGGGCTCTCTAGGGATTTAGTGCTGAT
+HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1
_b_ceeeeggggiiiiihiiiiiiiiiiiiiiiiiiiefhiihiiiihihhhhgggeeee_dddcddccccbcaccccccccbcbcccccccccdccc
@HWI-ST0866_0110:5:1101:2517:2226#GATCAG/1
TATCAATTTGCGCTTGATTACTAGTGCTACCTTCCCATACATTGCAGAGAGCGCGTGTCCATAGTGTATGGTACAGTACAACCAGCACCACAGCTTAGAG
+HWI-ST0866_0110:5:1101:2517:2226#GATCAG/1
__P`^ceeggff`fhiiiighfhgghihghhhghhdfdffh__agffhghcddgggggedeeea^^acddZ_bRU]`]bcZZ[`ab[`^c_aaccccbb
@HWI-ST0866_0110:5:1101:2659:2245#GATCAG/1
TGCCTATGGAATTACGTTAATTTACACAAACAAATTCCATATTAGCTTTAAAAAATAAACCTACTTCTAACTAGAAGTGAAAGAAGTTTAAAAGTGCTGC
+HWI-ST0866_0110:5:1101:2659:2245#GATCAG/1
^YYccacc]`ae^affhhhbc^dgeghfeeafg`dfhhgf]fhhSYcgfd]cae_fghhRW\eb]ffhZV\`d]dR\V^Z_dced`bGZ`b`bbZ]]_Y`
:
```

# .fastq Machine output

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description
- Line 2 is the raw sequence letters
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

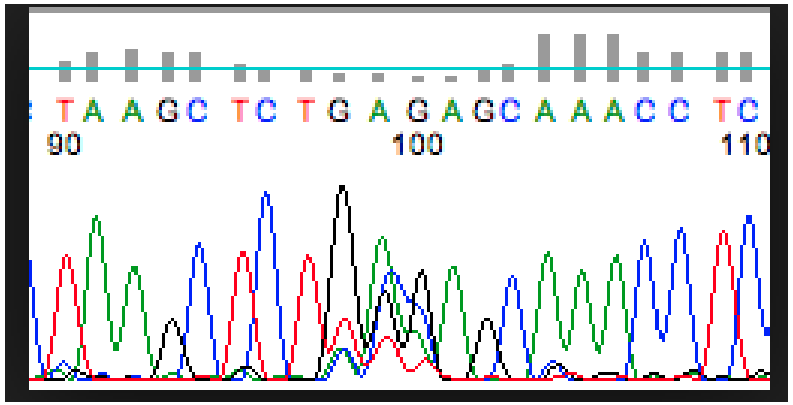
.fastq

```
@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFGGGGGGGGGGHFFFHGHHGFFHHHHHAG
```

# Sequence quality

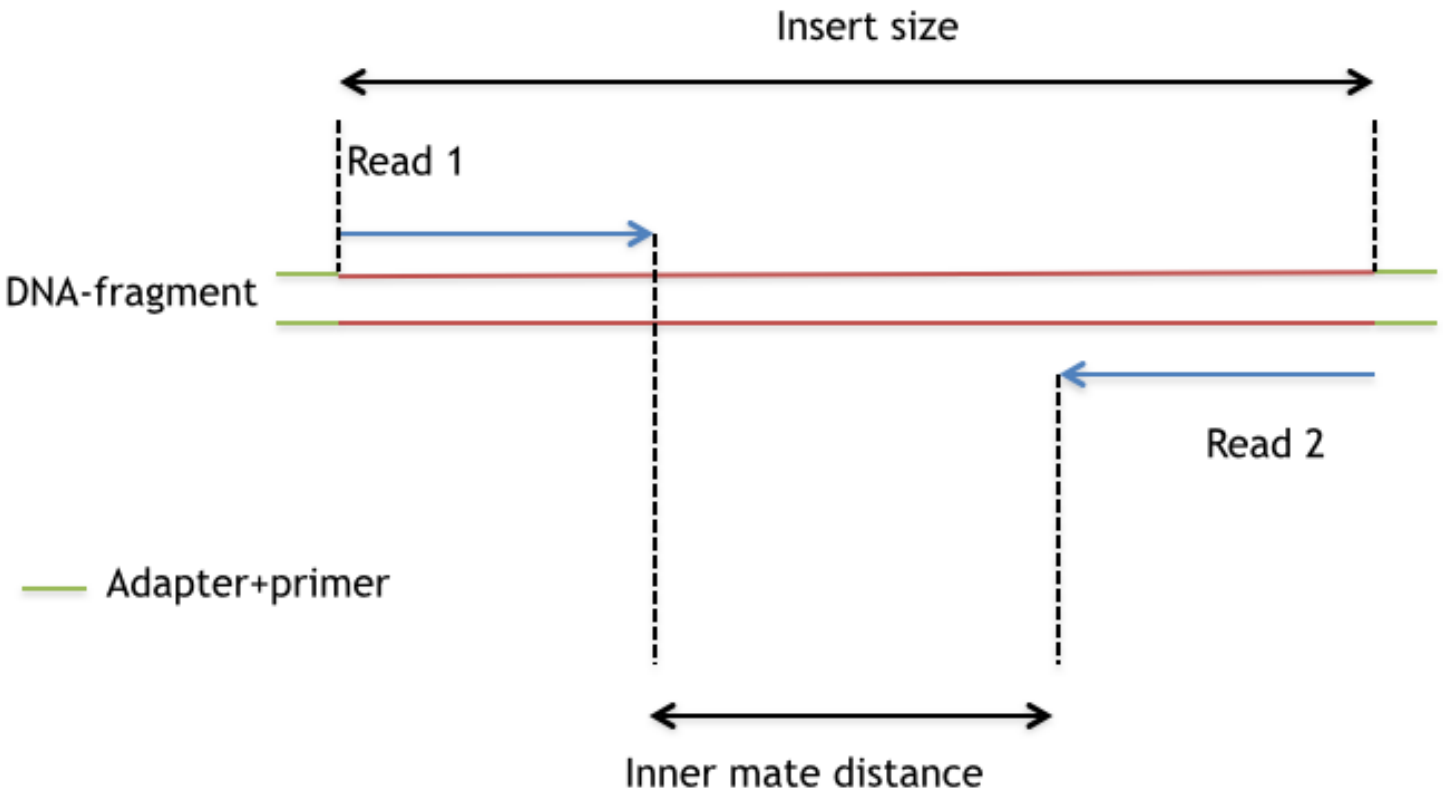
## Phred Quality Score

- Q = -10 x log P
- where:
  - P, probability of base calling being incorrect
  - High Q = high probability of the base being correct




- A Phred quality score of 10 to a base, means that the base is called incorrectly in 1 out of 10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of 100 times.
- A Phred quality score of 30 to a base, means that the base is called incorrectly in 1 out of 1000 times.
- etc.

# Paired-end (PE) reads



# Paired-end (PE) reads

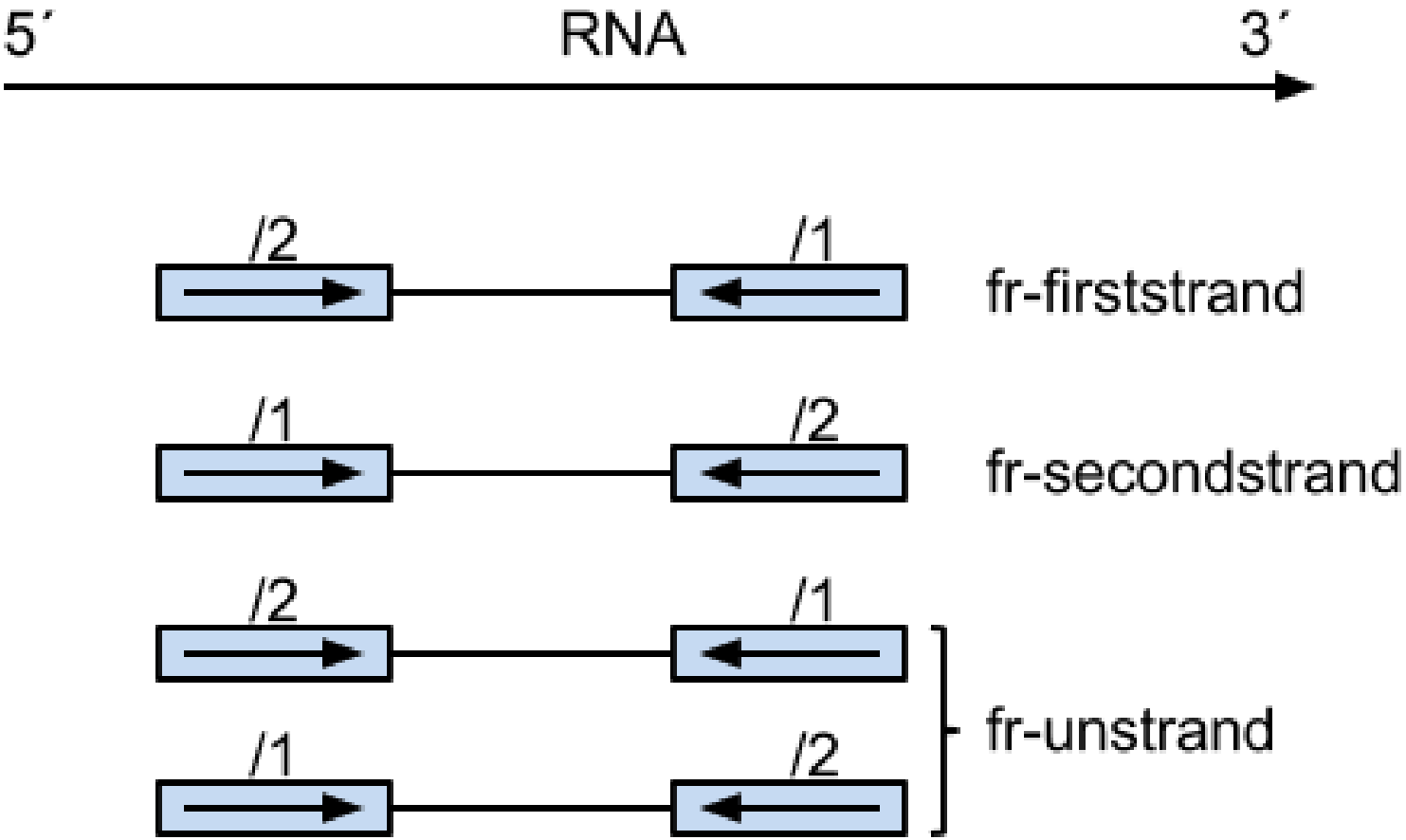
## File format

- Two files are created
- The order in files identical and naming of reads are the same with the exception of the end
- The way of naming reads are changing over time so the read names depend on software version

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

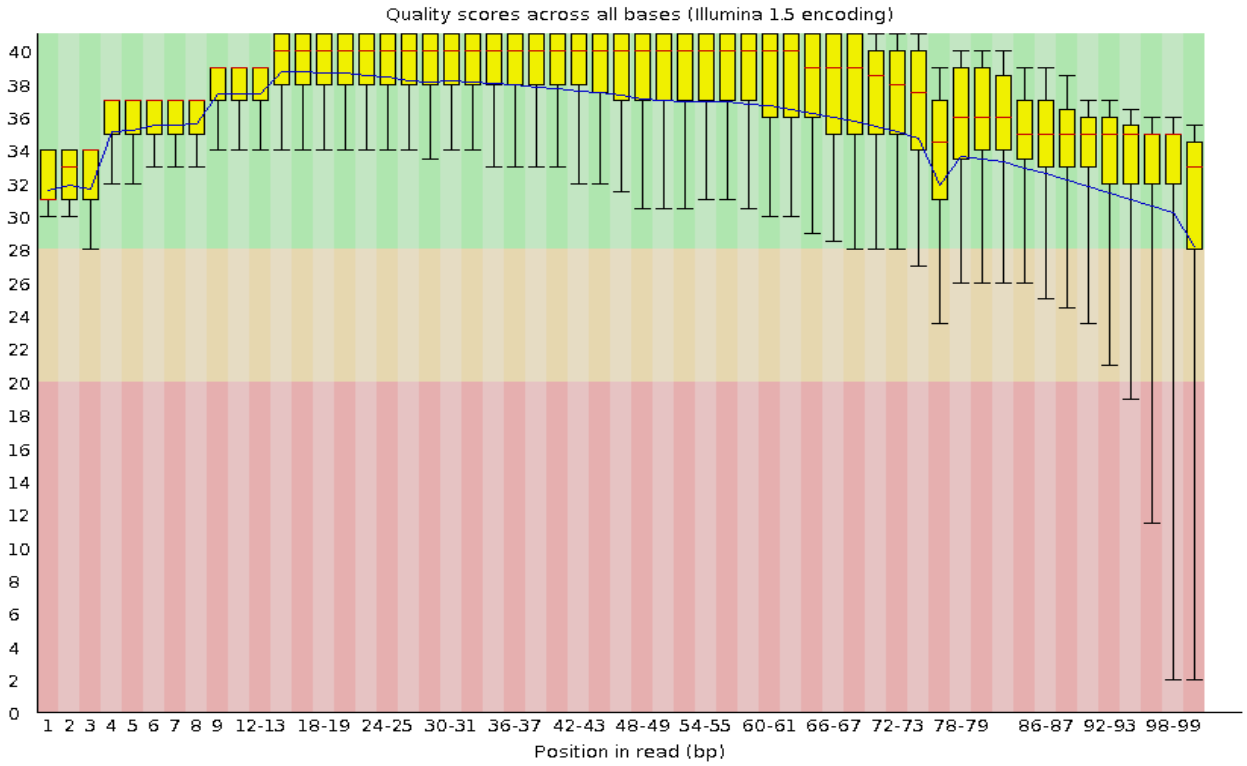
```
@61DFRAAXX100204:1:100:10494:3070/2
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATA
+
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^ad
```

# Strandness


5´
RNA
3´
/2
/1
fr-firststrand
/1
/2
fr-secondstrand
/2
/1
/1
/2
fr-unstrand

# RNA-seq data analysis
## Mapping based approach

# Quality control of raw reads




## Available tools

FastQC, PRINSEQ

# Raw reads filtering and trimming

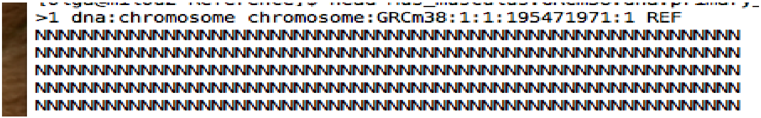

- filtering reads for quality score, e.g. with avg. quality below 20 defined within 4-base wide sliding window
- filtering reads for read length, e.g. reads shorter than 36 bases
- removing artificial sequences, e.g. adapters

**Available tools**

TRIMMOMATIC, FastX, PRINSEQ, Cutadapt

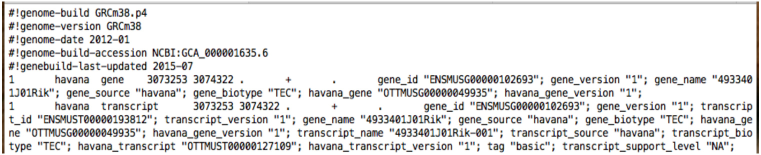
# Reference genome and annotations

.fasta (download reference genome FASTA file)



```
>1 dna:chromosome chromosome:GRCm38:1:1:195471971:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

.gtf (download the corresponding genome annotation in GTF or GFF)



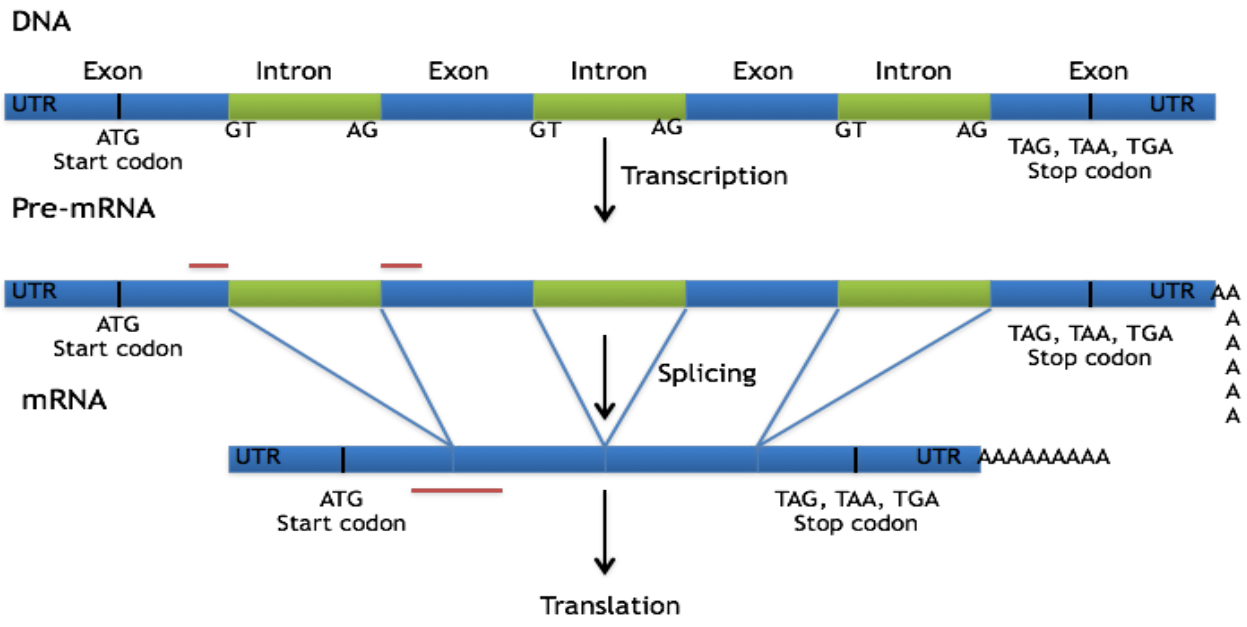
```
#!genome-build GRCm38.p4
#!genome-version GRCm38
#!genome-date 2012-01
#!genome-build-accession NCBI:GCA_000001635.6
#!genebuild-last-updated 2015-07
1       havana  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693"; gene_version "1"; gene_name "4933401J01Rik"; gene_source "havana"; gene_biotype "TEC"; havana_gene "OTTMUSG00000049935"; havana_gene_version "1";
1       havana  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693"; gene_version "1"; transcript_id "ENSMUST00000193812"; transcript_version "1"; gene_name "4933401J01Rik"; gene_source "havana"; gene_biotype "TEC"; havana_gene "OTTMUSG00000049935"; havana_gene_version "1"; transcript_name "4933401J01Rik-001"; transcript_source "havana"; transcript_biotype "TEC"; havana_transcript "OTTMUST00000127109"; havana_transcript_version "1"; tag "basic"; transcript_support_level "NA";
```

**Source**

ENSEMBL, NCBI

# Mapping reads to the genome

- Choose adequate aligner
- Use annotations and allow for spliced mapping

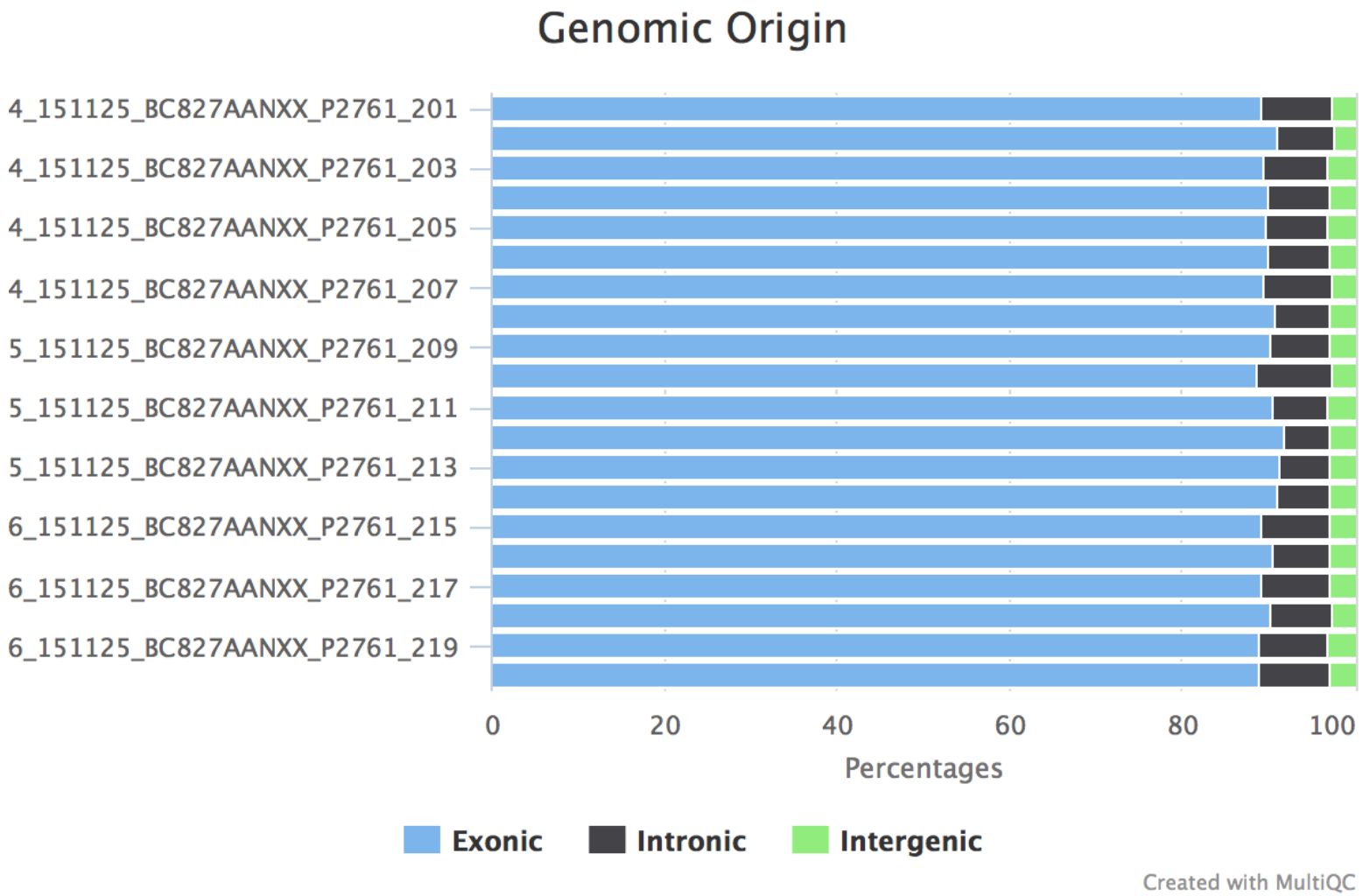



**Available tools**

Star, Tophat, Subread and many more...

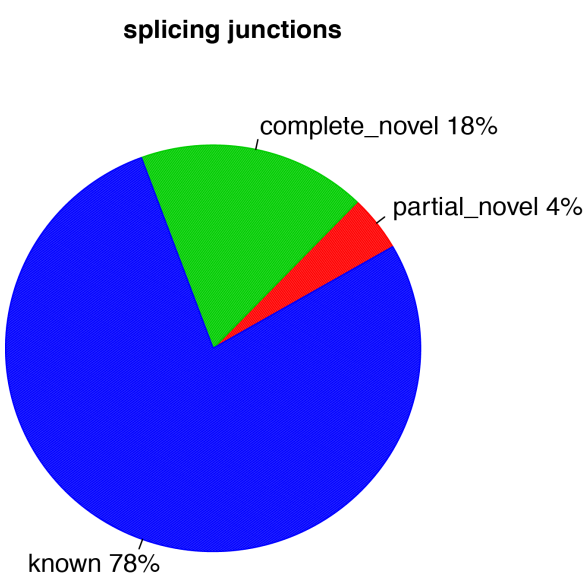
# Mapping reads to the genome: QC

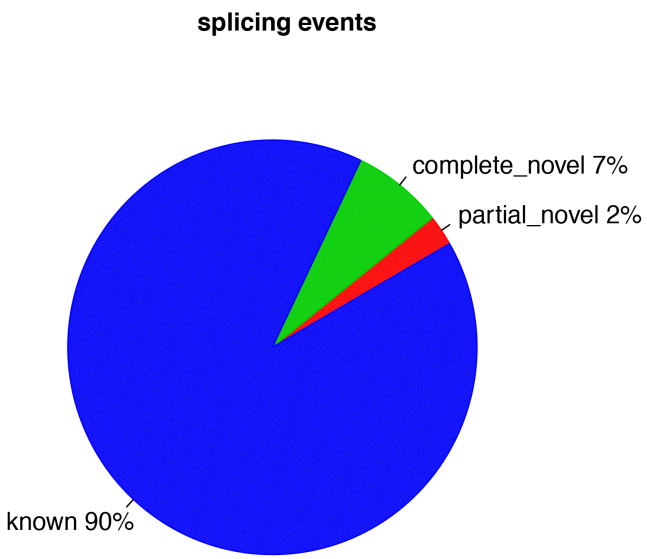
Reads should mostly map to known genes



# Mapping reads to the genome: QC

Most splice event should be known and canonical (GU-AG)

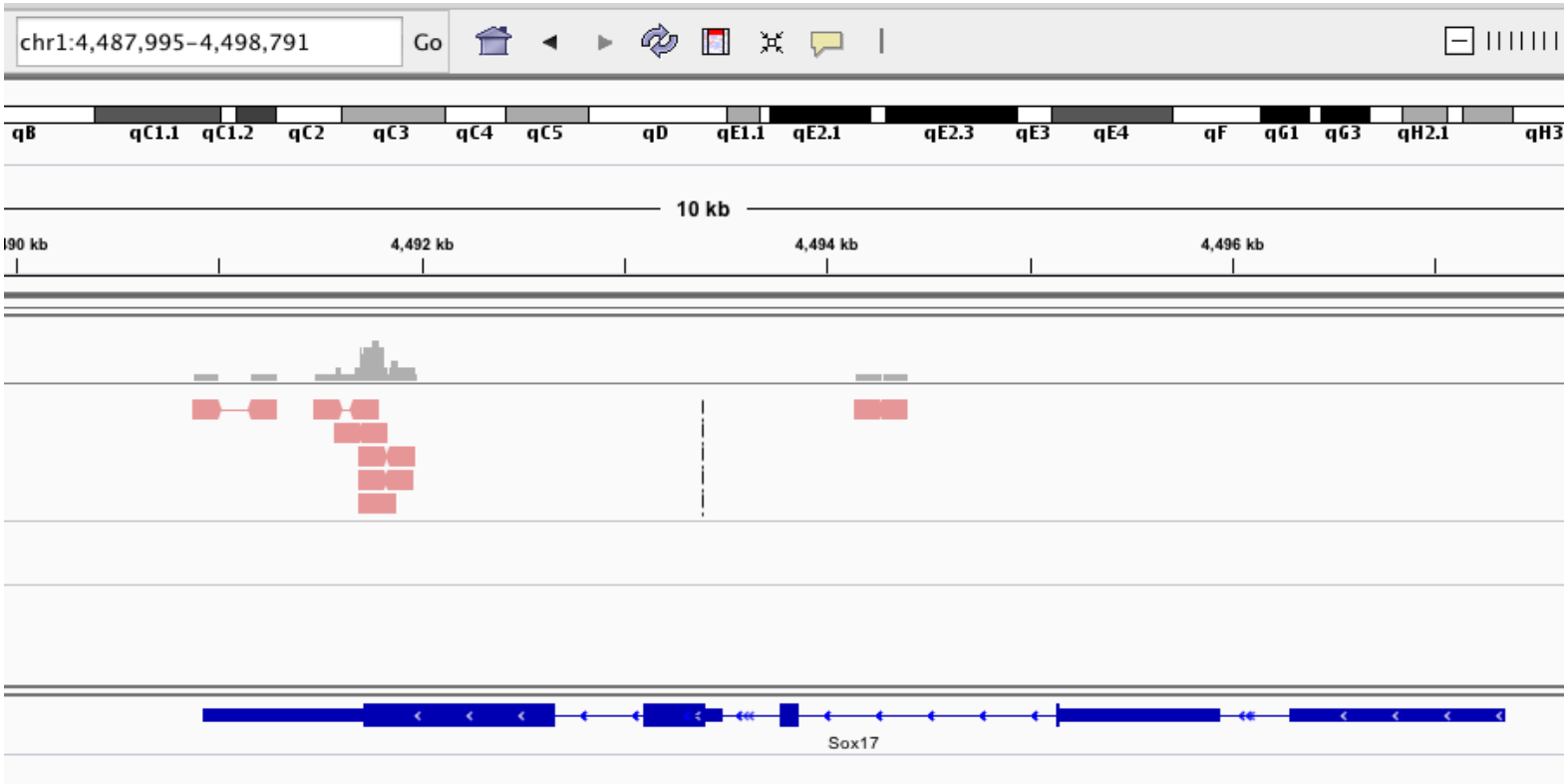



**Available tools**

RseQC, Picard, QualiMap

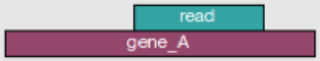
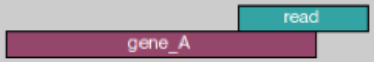
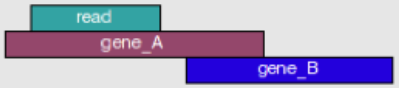
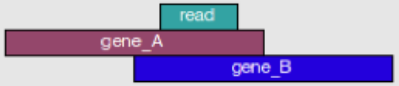
# Counting reads




## Available tools

HTSeq, featureCounts, R

# Counting reads

| | 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 |

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

# Counting reads

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Transcript | P1822_1 | P1822_2 | P1822_3 | P1822_4 | P1822_5 | P1822_6 | P1822_7 | P1822_8 | P1822_9 | P1822_10 | P1822_11 | P1822_12 | P1822_13 | P1822_14 | P1822_15 | P1822_16 |
| 2 | ENSMUSG00000102693 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | ENSMUSG00000088000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | ENSMUSG00000103265 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | ENSMUSG00000103922 | 7 | 7 | 7 | 4 | 1 | 12 | 3 | 6 | 14 | 3 | 9 | 3 | 9 | 7 | 9 | 7 |
| 33 | ENSMUSG00000033845 | 972 | 860 | 878 | 1085 | 1058 | 1009 | 992 | 1143 | 947 | 1059 | 970 | 1147 | 801 | 837 | 1042 | 927 |
| 34 | ENSMUSG00000102275 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 35 | ENSMUSG00000025903 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 36 | ENSMUSG00000104217 | 16 | 13 | 17 | 16 | 22 | 17 | 12 | 27 | 11 | 5 | 12 | 15 | 8 | 9 | 9 | 12 |
| 37 | ENSMUSG00000033813 | 2560 | 2581 | 2937 | 3904 | 2975 | 3100 | 3027 | 3417 | 2272 | 2801 | 2266 | 3294 | 2491 | 2578 | 2554 | 2806 |
| 38 | ENSMUSG00000062588 | 3 | 1 | 1 | 1 | 0 | 1 | 0 | 3 | 3 | 0 | 4 | 0 | 2 | 1 | 0 | 0 |
| 39 | ENSMUSG00000103280 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 40 | ENSMUSG00000002459 | 7 | 10 | 5 | 7 | 4 | 6 | 3 | 8 | 2 | 5 | 7 | 8 | 1 | 5 | 4 | 1 |
| 41 | ENSMUSG00000091305 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 42 | ENSMUSG00000102653 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 43 | ENSMUSG00000085623 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 44 | ENSMUSG00000091665 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | ENSMUSG00000033793 | 3682 | 3757 | 4414 | 5978 | 3774 | 4102 | 3815 | 4250 | 4193 | 4962 | 4240 | 5694 | 3565 | 3757 | 3849 | 4094 |
| 46 | ENSMUSG00000104352 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47 | ENSMUSG00000104046 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 48 | ENSMUSG00000102907 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 49 | ENSMUSG00000025905 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 50 | ENSMUSG00000103936 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 51 | ENSMUSG00000093015 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 52 | ENSMUSG00000103519 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 53 | ENSMUSG00000033774 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 54 | ENSMUSG00000103090 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 55 | ENSMUSG00000025907 | 1816 | 2087 | 2088 | 2820 | 2012 | 2236 | 2065 | 2727 | 2586 | 2931 | 2813 | 3667 | 2410 | 2739 | 2479 | 2745 |
| 56 | ENSMUSG00000090031 | 43 | 58 | 55 | 73 | 38 | 38 | 57 | 96 | 89 | 107 | 98 | 123 | 76 | 93 | 66 | 69 |

## Available tools

HTSeq, featureCounts, R

# Normalization: from counts to gene expression

## RPKM & FPKM

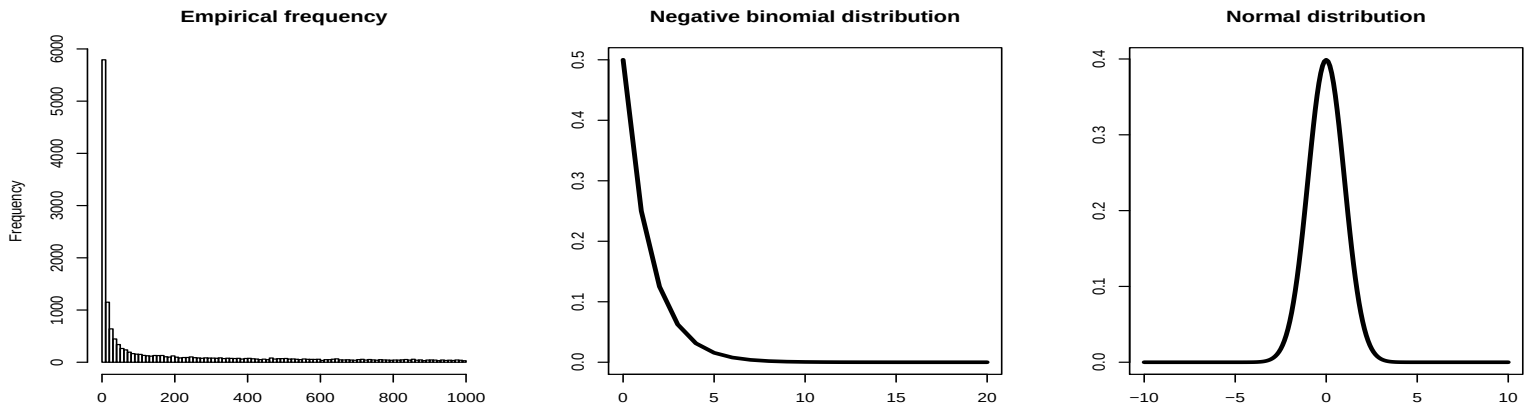
- Reads or Fragments Per Kilobase per Milion
- Correct for: differences in sequencing depth and transcript length

## Other

- TMM: correct for differences in transcript pool composition
- TPM: correct for transcript length distribution in RNA pool
- Voom: removing dependence of variance on the mean

# Differential expression

- Reads counts do not follow normal distribution
- Typically low number of replicates or samples per group
- Recommend to use statistical packages prepared specifically for the statistical analyses of count data




**Available tools**

Cuffdiff, edgeR (R), limma (R), DESeq (R)

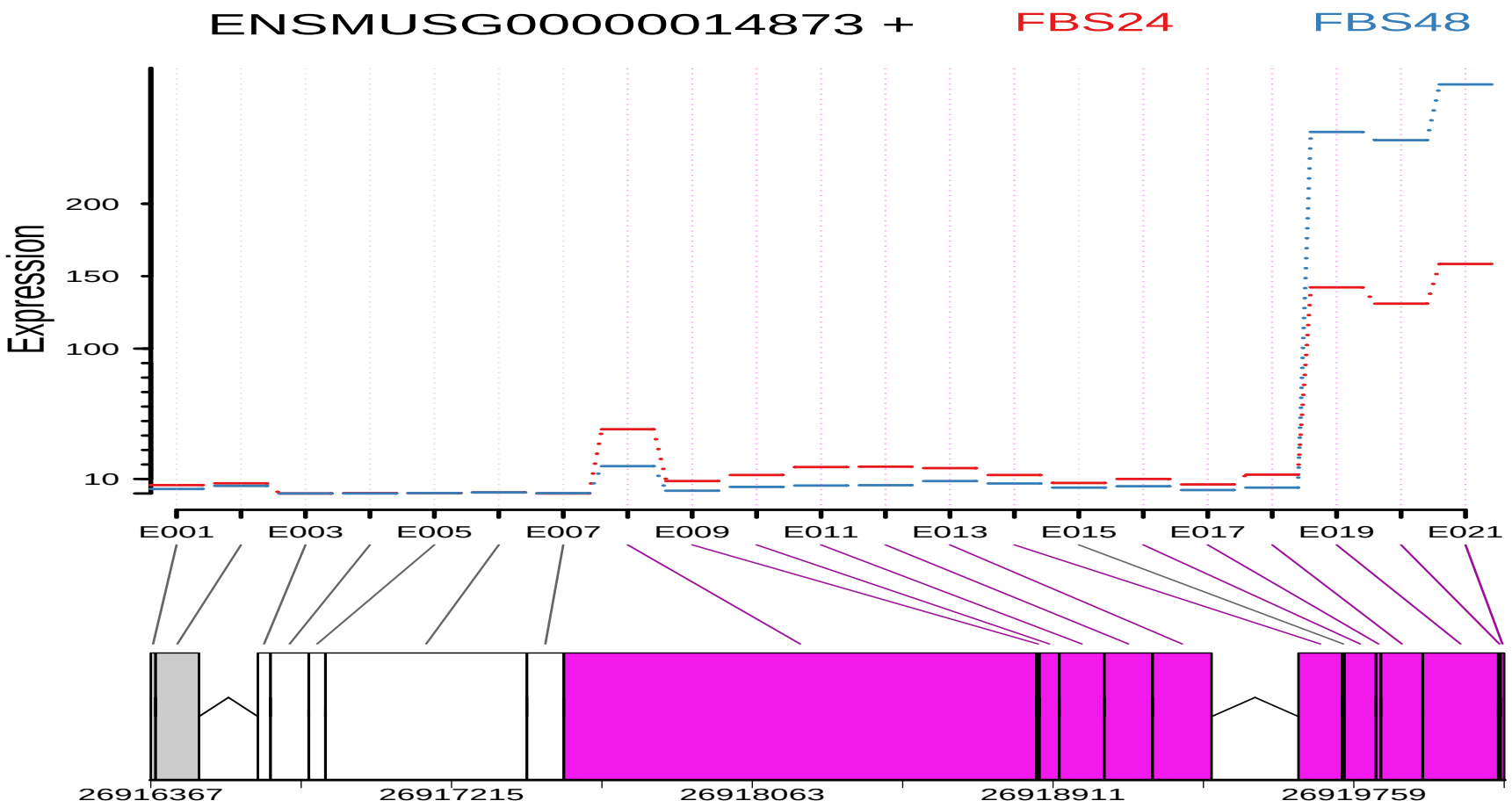
# Differential expression

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ensembl_gene_id | ensembl_transcript_id | chromosome_name | mgi_symbol | description | logFC | logCPM | LR | PValue | FDR |
| 2 | ENSMUSG00000028328 | ENSMUST00000107773 | 4 | Tmod1 | tropomodulin 1 [Source:MGI Symbol;Acc:MGI:98775] | 1.971089 | 5.958225 | 581.2916 | 1.96E-128 | 2.79E-124 |
| 3 | ENSMUSG00000066705 | ENSMUST00000085939 | 9 | Fxyd6 | FXYD domain-containing ion transport regulator 6 [Source:MGI Symbol;Acc | 3.18062 | 5.916499 | 553.8787 | 1.80E-122 | 1.28E-118 |
| 4 | ENSMUSG00000049112 | ENSMUST00000053306 | 6 | Oxtr | oxytocin receptor [Source:MGI Symbol;Acc:MGI:109147] | 3.820952 | 3.423774 | 375.1689 | 1.40E-83 | 6.65E-80 |
| 5 | ENSMUSG00000017446 | ENSMUST00000124861 | 11 | C1qtnf1 | C1q and tumor necrosis factor related protein 1 [Source:MGI Symbol;Acc:M | 1.484213 | 7.145099 | 345.7577 | 3.56E-77 | 1.26E-73 |
| 6 | ENSMUSG00000029123 | ENSMUST00000094836 | 5 | Stk32b | serine/threonine kinase 32B [Source:MGI Symbol;Acc:MGI:1927552] | 3.453001 | 2.321613 | 338.7155 | 1.22E-75 | 3.46E-72 |
| 7 | ENSMUSG00000009378 | ENSMUST00000009522 | 19 | Slc16a12 | solute carrier family 16 (monocarboxylic acid transporters), member 12 [Sc | 4.173029 | 3.89466 | 335.706 | 5.50E-75 | 1.30E-71 |
| 8 | ENSMUSG00000025355 | ENSMUST00000026411 | 10 | Mmp19 | matrix metallopeptidase 19 [Source:MGI Symbol;Acc:MGI:1927899] | 1.940915 | 8.973932 | 328.4969 | 2.04E-73 | 4.15E-70 |
| 9 | ENSMUSG00000029671 | ENSMUST00000128245 | 6 | Wnt16 | wingless-type MMTV integration site family, member 16 [Source:MGI Syml | 2.339149 | 5.673738 | 315.6779 | 1.27E-70 | 2.25E-67 |
| 10 | ENSMUSG00000042190 | ENSMUST00000047936 | 5 | Cmklr1 | chemokine-like receptor 1 [Source:MGI Symbol;Acc:MGI:109603] | 2.518748 | 3.540638 | 305.0157 | 2.66E-68 | 4.20E-65 |
| 11 | ENSMUSG00000028035 | ENSMUST00000134701 | 3 | Dnajb4 | DnaJ (Hsp40) homolog, subfamily B, member 4 [Source:MGI Symbol;Acc:M | 1.417856 | 7.292192 | 297.1316 | 1.39E-66 | 1.98E-63 |
| 12 | ENSMUSG00000048960 | ENSMUST00000027056 | 1 | Prex2 | phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 2 | 1.706461 | 6.676335 | 283.7963 | 1.12E-63 | 1.44E-60 |
| 13 | ENSMUSG00000002289 | ENSMUST00000002360 | 17 | Angptl4 | angiopoietin-like 4 [Source:MGI Symbol;Acc:MGI:1888999] | -1.73049 | 7.972378 | 282.7705 | 1.87E-63 | 2.22E-60 |

**Available tools**

Cuffdiff, edgeR (R), limma (R), DESeq2 (R)

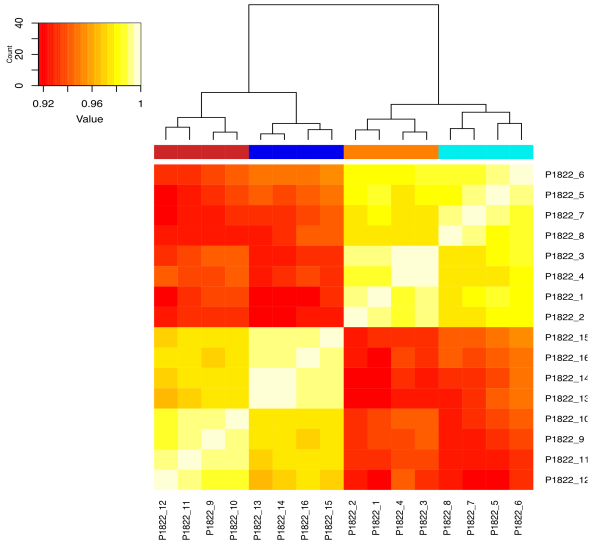
# Differential expression

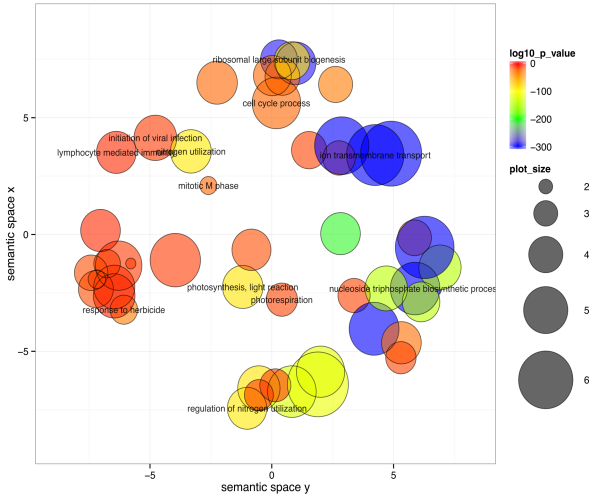



**Available tools**

Cuffdiff, edgeR (R), limma (R), DESeq2 (R)

# Beyond differential expression



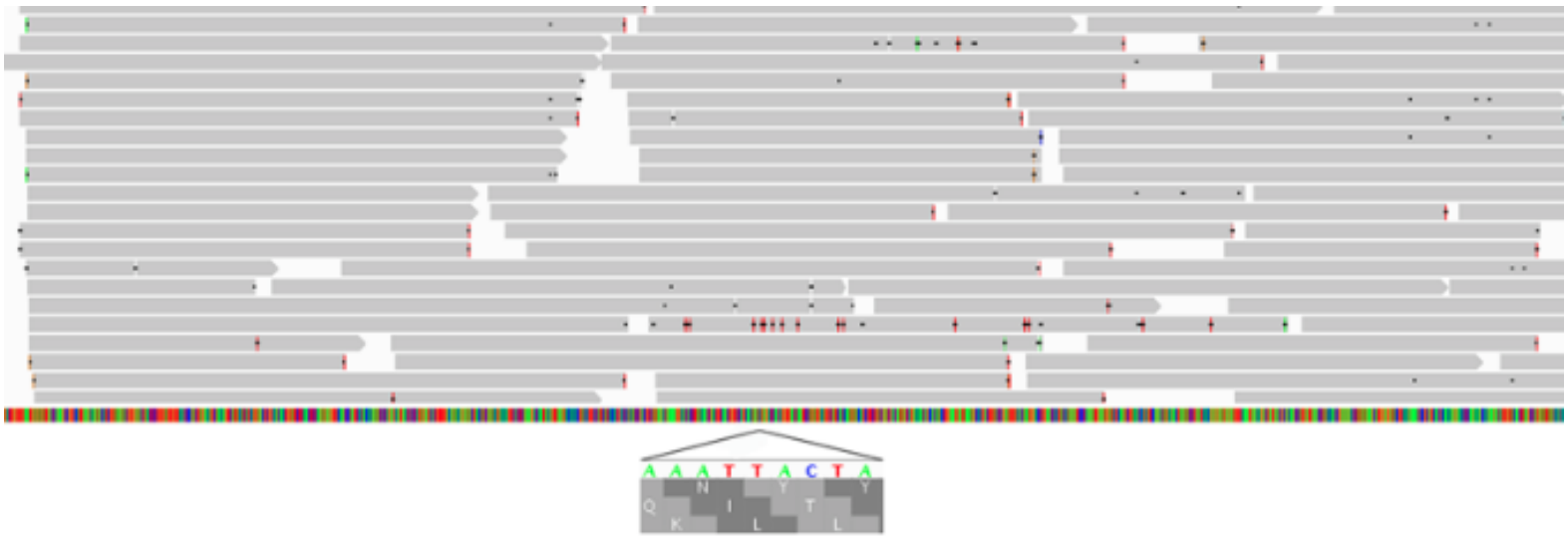

- Annotating the results e.g. with gene symbols, GO terms
- Visualising the results, e.g. Volcano plots
- Gene set analysis etc...

**Available tools**

bioMart (R), DAVID, GOrilla, REVIGO, ClustVis...

# RNA-seq data analysis
## Transcriptome assembly

# Transcriptome assembly




- The goals is to reconstruct full-length transcript based on the sequence reads
- This is done via algorithms using the small overlapping reads fragments
  - If the reference genome is known: genome guided
  - If the reference genome is unknown *de novo* assembly

# Transcriptome assembly

## Challenges

- Genes show different levels of gene expression, hence uneven coverage among genes
- More sequencing depth is needed to represent less abundant genes and rare events
- In order to balance the abundance differences between genes, laboratory procedures for library normalisation

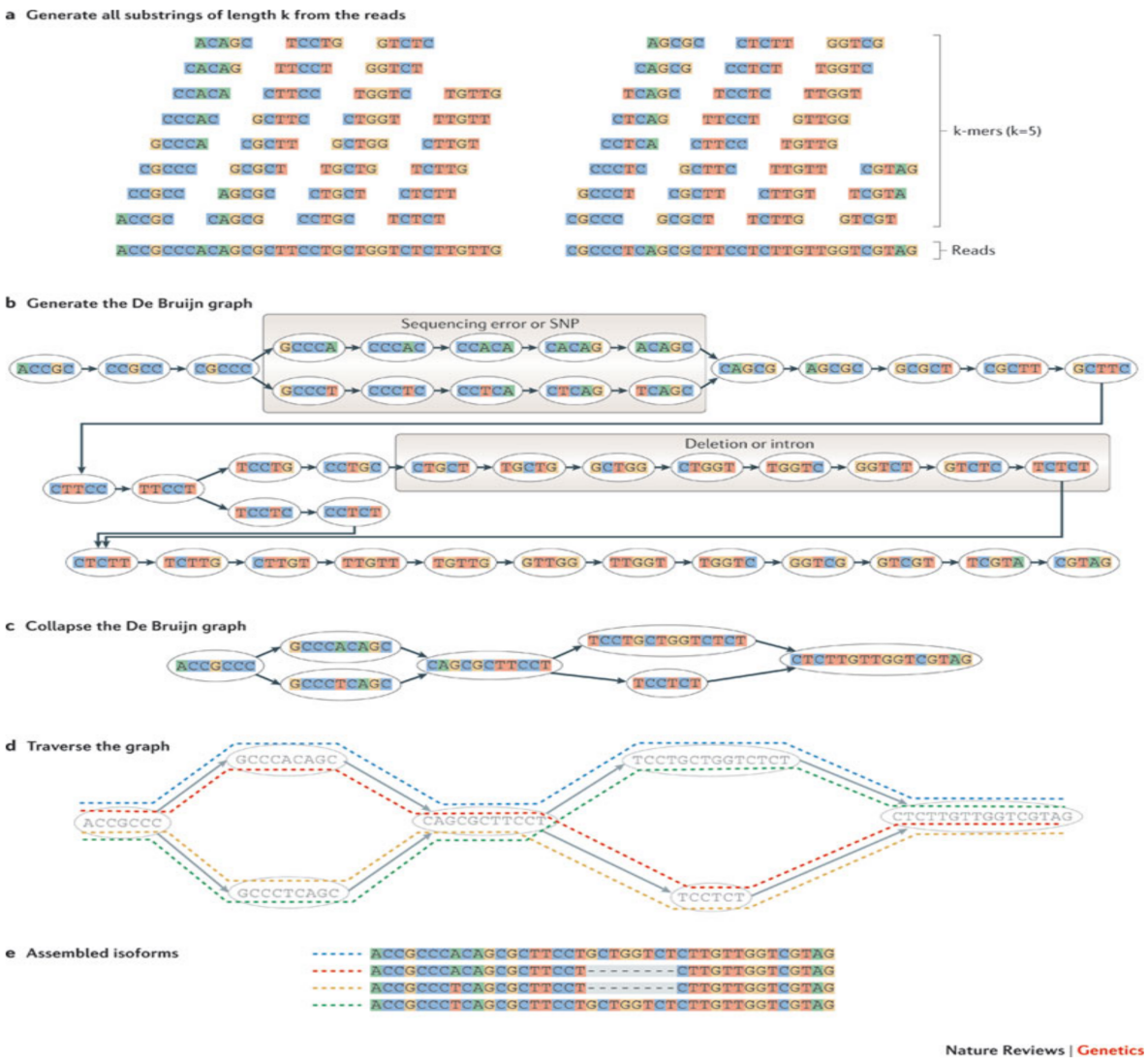
# Transcriptome assembly

## Available tools

- SOAP-denovo TRANS
- Oases
- Trans-ABYSS
- Trinity

All of them uses de Bruijn graphs to cope with the data and many of them have been developed from a genome assembly program

# Trinity


a Generate all substrings of length k from the reads
k-mers (k=5)
Reads
b Generate the De Bruijn graph
Sequencing error or SNP
Deletion or intron
c Collapse the De Bruijn graph
d Traverse the graph
e Assembled isoforms
Nature Reviews | Genetics

# Summary

- Many different expertises needed for RNA-seq experiment
  - Think ahead, plan wisely, ask for help
  - If your experimental design is wrong nothing will help
- Assess and try to improve the quality of raw reads
  - use QC tools and talk to sequencing centre
- If reference genome is available
  - get a corresponding genome annotation
  - align your reads using spliced alignment
  - in well-annotated genomes most reads should map to known genes
  - use tools designed for statistical analyses of sequencing count data (bioconductor)

# Summary

- If interested in transcriptome assembly
  - use reference genome to guide it, if available
  - send lots of time in assessing the results e.g. by comparing related species, looking at ORFs
  - consider merging with other data sources
  - consider trying different assembler
- Ensure that your experimental design allows addressing the question of interest
  - More replicates translates into more power for differential gene expression and easier publication process

# Exercises

# Exercises

## Main exercise

- checking the quality of the raw reads with FastQC
- mapping the reads to the reference genome using STAR
- converting between SAM and BAM files format using Samtools
- assessing the post-alignment reads quality using QualiMap
- counting reads overlapping with genes regions using featureCounts
- building statistical model to find DE genes using edgeR called from a prepared R script

## Bonus exercises

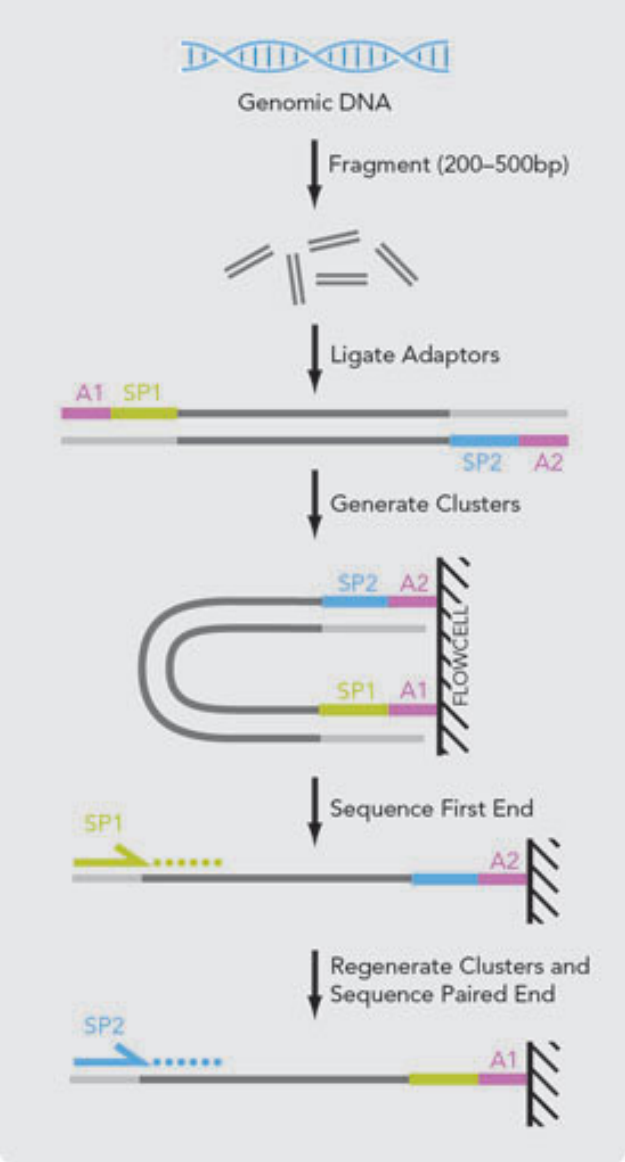
- functional annotation, putting DE genes in the biological context
- exon usage, studying the alternative splicing
- data visualisation and graphics
- de novo transcriptome assembly

# Questions?

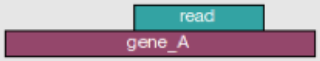
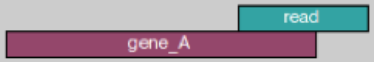
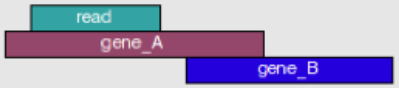
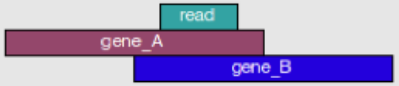
# .fastq Machine output

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

# .fastq Machine output: Paired-end (PE) sequencing

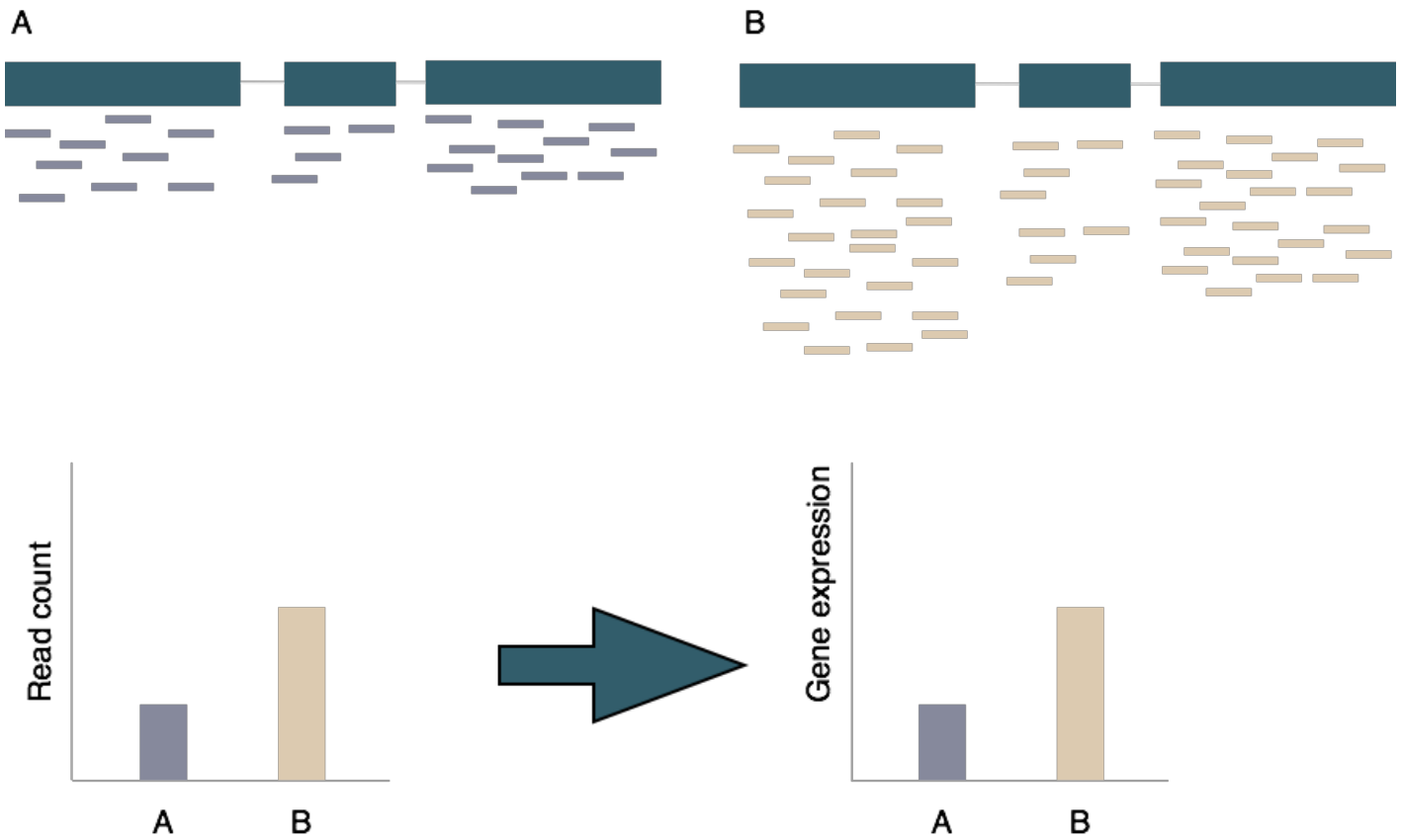


# Counting reads

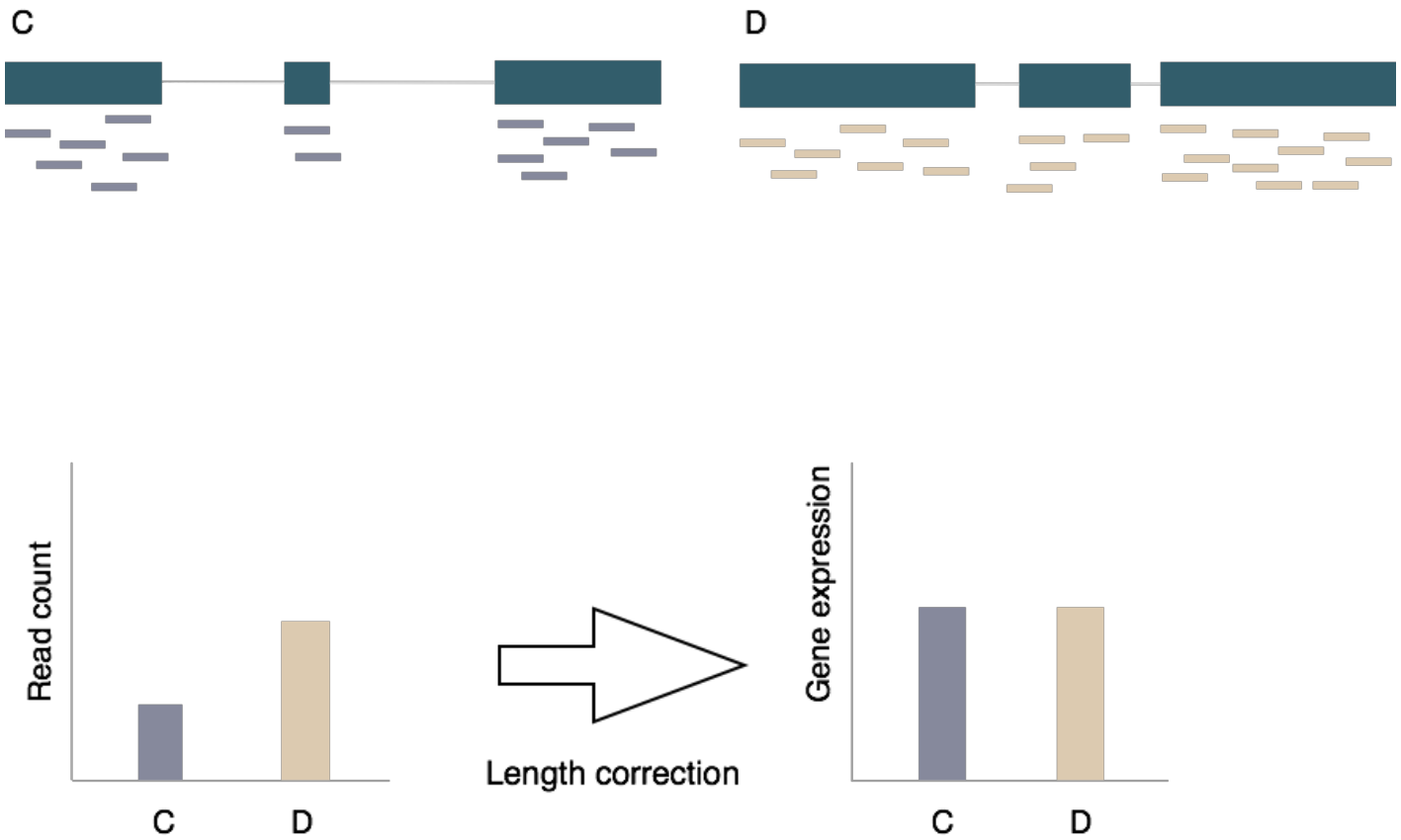
| | union | intersection _strict | intersection _nonempty |
|---|---|---|---|
|   | gene_A | gene_A | gene_A |
|   | gene_A | no_feature | gene_A |
|   | gene_A | no_feature | gene_A |
|   | gene_A | gene_A | gene_A |
|   | gene_A | gene_A | gene_A |
|   | ambiguous | gene_A | gene_A |
|   | ambiguous | ambiguous | ambiguous |

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

# From counts to gene expression



# From counts to gene expression



# Experimental design

- Count reads (convert to RPKM/FPKM?)
- Small number of reads (= low RPKM/FPKM values) often non-significant
- Remember that Fold change is not the same as significance

| | Condition 1 | Condition 2 | Fold_Change | Significant? |
|---|---|---|---|---|
| **Gene A** | 1 | 2 | 2-fold | No |
| **Gene B** | 100 | 200 | 2-fold | Yes |