

# Characterizing transcriptomes using ngs data

T. Källman

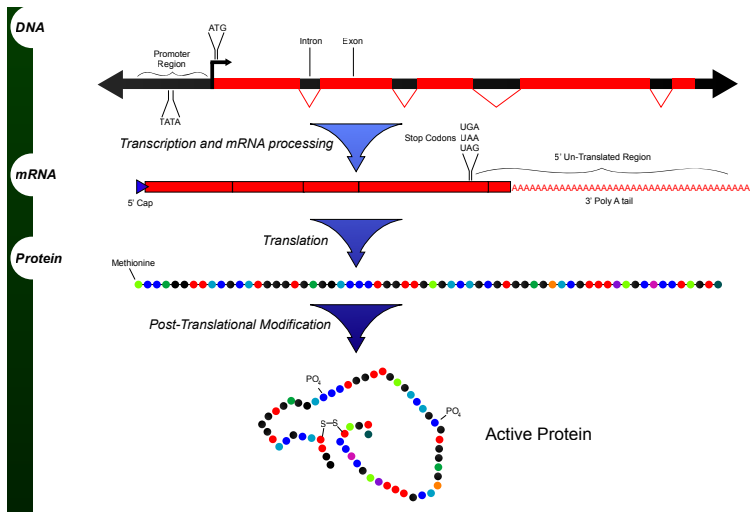
NBIS/Scilife Lab/Uppsala University

Sep. 2016

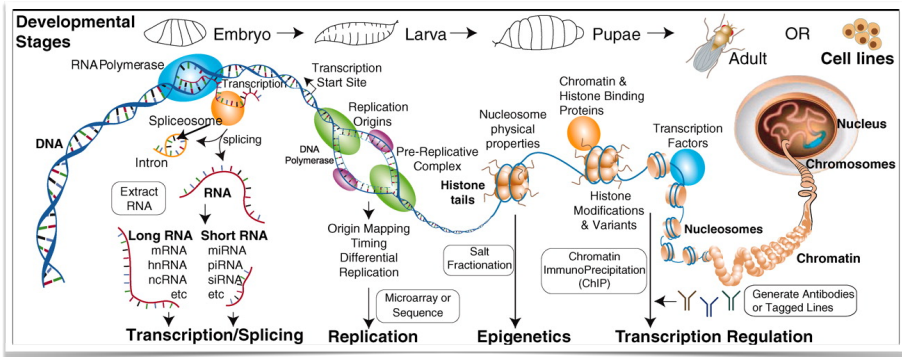
# Outline

- 1 Why study the transcriptome
- 2 RNA sequencing
- 3 RNA-seq analysis
  - Map based analysis
  - de-novo assembly

# The Central Dogma



# A more complex view



# Transcriptome data enables

- Differential gene expression
- Differential isoform usage, eg splicing patterns
- Identification of co-expressed genes (gene networks)
- Allele specific expression patterns
- Detection of fusion genes

# Transcriptomes vs genomes

- Dynamic, not the same over tissues and time points
- Smaller sequence space
- Less repetitive (but large gene families can be found)
- Fairly stable in size? (*eg.* 2-4 fold change among eukaryotes, whereas genome size can vary 1000-fold)
- Genes are often expressed in multiple different splice-variants
- RNA often from only one strand

# 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)

## Long read technologies

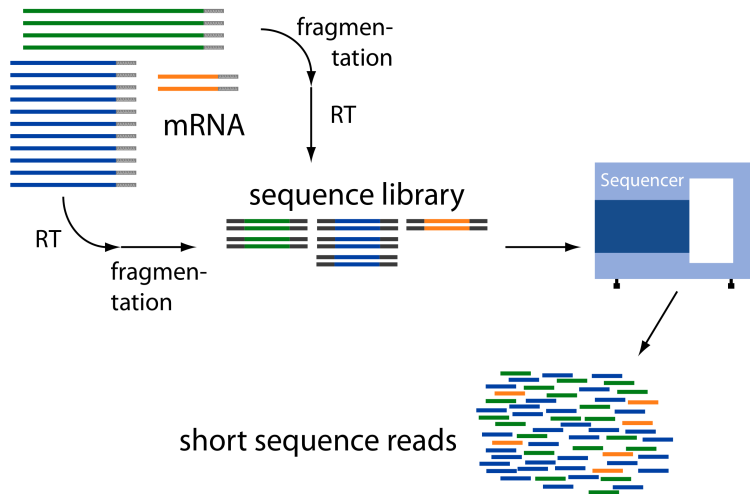
- With long reads (eg Pacific Bioscience) full transcripts can be directly sequenced
- Since transcripts can be “directly” observed all isoforms can be detected
- All current long read methods will give too little data (at reasonable cost) to estimate expression levels



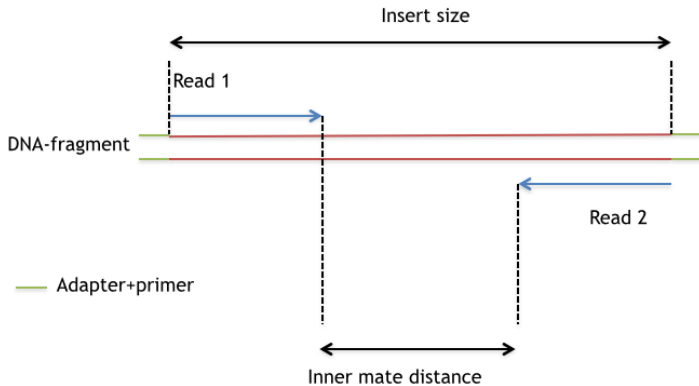
Mainly used for *de-novo* assembly of transcriptomes or isoform identification



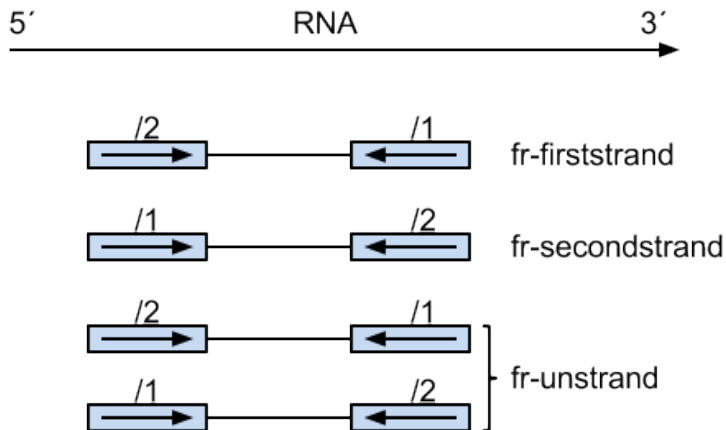
# Short read technologies



# Pair-end data

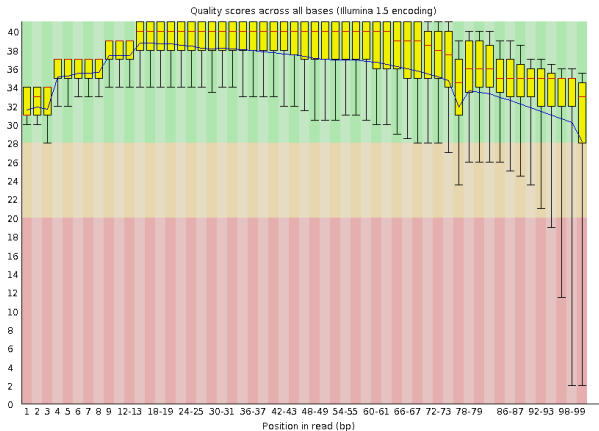


## Stranded or not



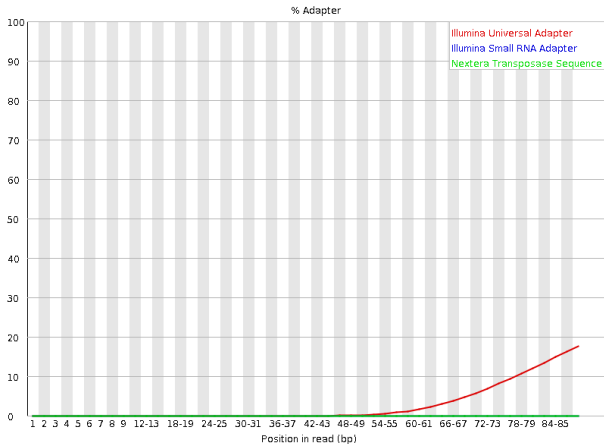
# Basic quality control of raw reads

- FastQC



# Basic quality control of raw reads

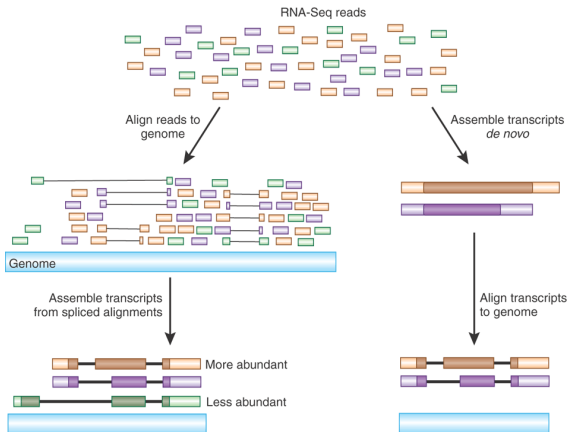
- FastQC



## Basic quality control of raw reads

- RNA-seq is not random sample from the genome eg. GC content might be different
- Highly expressed genes can be frequent and create warnings in quality controls that assumes whole genome data
- Random hexamer in cDNA synthesis might create 'biases' in base frequencies in the beginning of reads

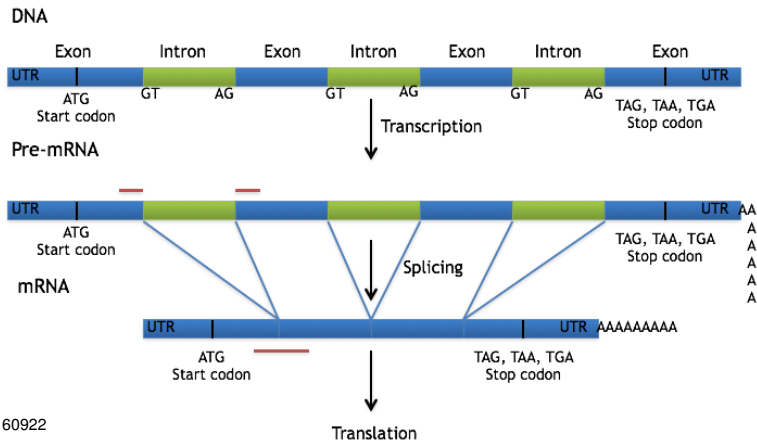
# Two main routes for analysis



Haas & Zody (2010), Nature Biotechnology 28, 421–423

# Aligning short reads from RNA to genomes

- Large number of programs available: Star, HiSat, Subreadalign etc
- Key feature compared to aligning DNA data: Allow for spliced mapping





# Aligning short reads from RNA to genomes

- If available map to the genome sequence
- If no genome sequence available map to transcriptome reference
- Make use of available genome annotation (GTF, GFF, BED files)

SeqnameSource		Feature	Start	End	ScoreStrandFrameAttributes
chr12	unknown	exon	87984	88017	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	88257	88392	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	88570	88771	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	88860	89018	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	89675	89827	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	90587	90655	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	90796	91263	-. + . gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS58200";
chr12	unknown	exon	147946	148509	-. + . gene_id "FAM1380"; gene_name "FAM1380"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	148612	148814	-. + . gene_id "FAM1380"; gene_name "FAM1380"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	149052	149412	-. + . gene_id "FAM1380"; gene_name "FAM1380"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	CDS	176049	176602	-. + 0 gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	176049	176602	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	start_codon	176049	176051	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	186542	186878	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	CDS	208312	208380	-. + 1 gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	208312	208380	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	208312	208380	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	234799	235078	-. + 1 gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	234799	235078	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	246577	246753	-. + . gene_id "LOC374538"; gene_name "LOC374538"; transcript_id "NR_033899"; tss_id "TSS17153";
chr12	unknown	CDS	247433	248520	-. + 0 gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	247433	248520	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	247433	248520	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	247439	248520	-. + 0 gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	start_codon	247439	247441	-. + . gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";

# Star

- 1 Index the genome as an uncompressed suffix array
- 2 Map reads in a two step fashion:
  - 1 Seed search that finds read or part of read that map without mismatch to genome
  - 2 Align complete reads by stitching seed mapping results using a local alignment procedure

# QC of mapped reads

## Reads should mostly map to known genes

```
read_distribution.py -i Pairend_StrandSpecific_51mer_Human_hg19.bam -r hg19.refseq.bed12
```

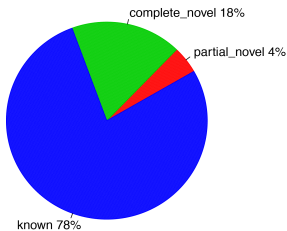
Output:

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

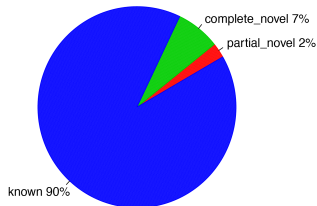
# QC of mapped reads

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

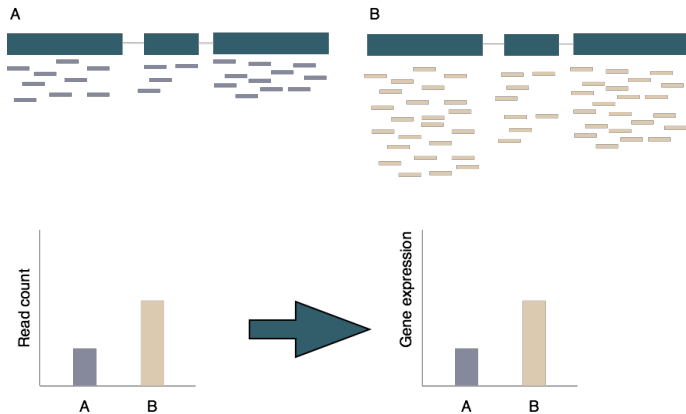
splicing junctions



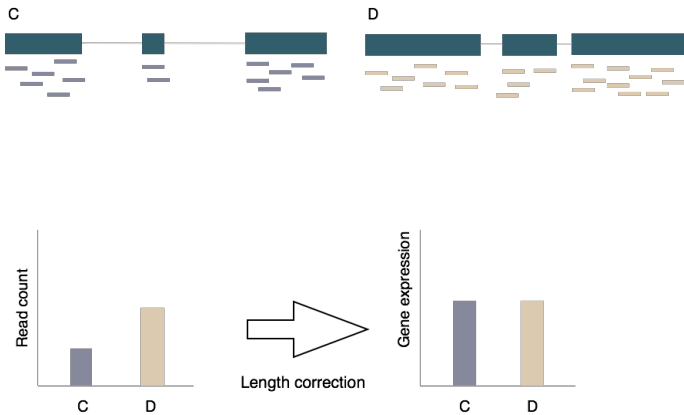
splicing events



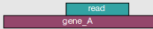
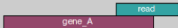
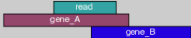
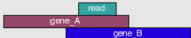
# From counts to gene expression



# From counts to gene expression



# Not all reads are the same

	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

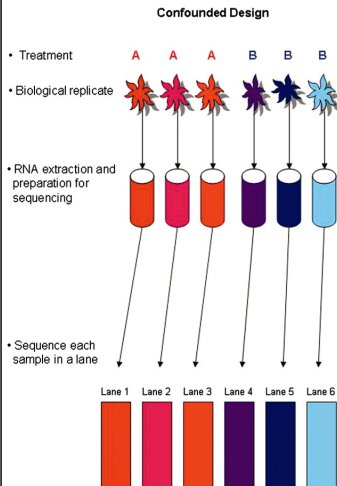
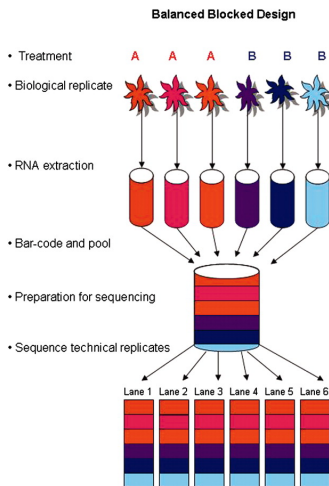
from: <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>

# Normalized expression Values

- Mapped read counts are normalized for both length of the transcript they map to and total depth of sequencing.
- Count data is hence converted to: Reads/Fragments per kb of transcript length and million mapped reads (RPKM or FPKM)



# Experimental design

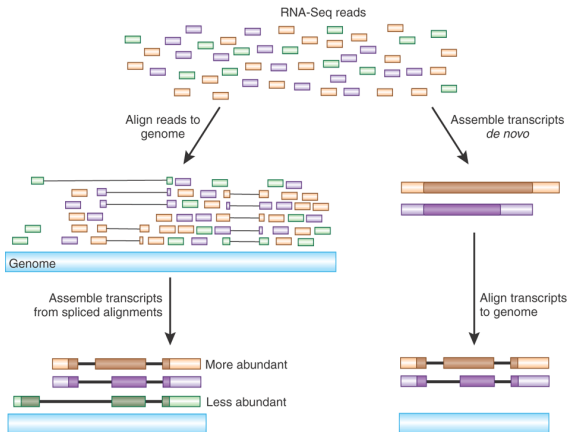


# Experimental design

- 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

# Two main routes for analysis



Haas & Zody (2010), Nature Biotechnology 28, 421–423

# Major challenges in relation to genome assembly

- Genes show different levels of gene expression, hence uneven coverage among genes
- Many genes are expressed in different isoforms
- As sequence depth increase detected number of loci increase. (What is actually expressed?)
- Sequence error from highly expressed genes might be seen more often than "true" sequences from lowly expressed genes

## Several programs available

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

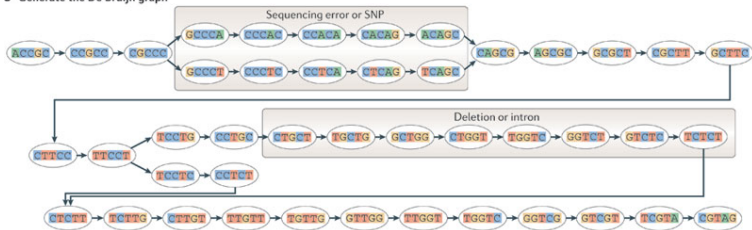
All of them uses de Bruijn graphs to cope with the data and many of them are based on genome assembly programs

## Trinity

## a Generate all substrings of length k from the reads



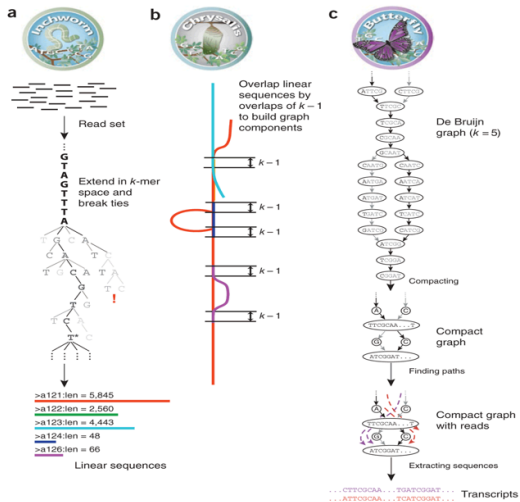
## b Generate the De Bruijn graph



## c Collapse the De Bruijn graph



## Trinity



## Summary - with ref.

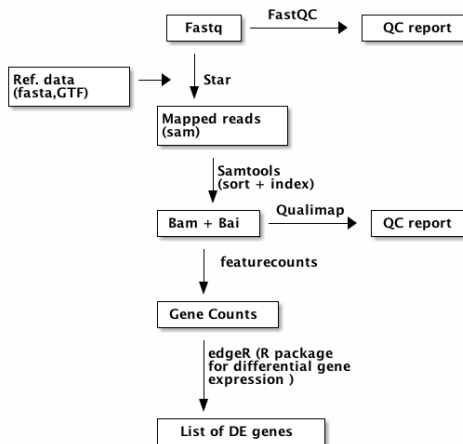
- Map to genome allow for spliced alignment
- If novel transcripts of interest: use method that can re-create transcripts from mapped reads (Cufflinks, Scripture or Bayesemblem)  
NB! In well annotated genomes most reads should map to known genes
- If interest is expression of known genes/exons: Use available annotation for analysis
- Spend time on experimental design and more replicates gives more power in gene expression analysis



## Summary - without ref.

- Assemble using your favourite assembler
- Spend lots of time in assessing the results (compare to related species, look for ORFs etc)
- Often large number of partial transcripts (hence often large number of contigs).
- Merge with other data from transcripts?

# Main exercise - RNA seq pipeline



# Bonus exercises

- 1 **Functional annotations** Making sense of DE results using GO terms etc
- 2 **Exon usage** Look at alternative splicing
- 3 **Visualisation** View bam files and create plots from DE data
- 4 **De novo assembly of transcriptomes** Analyse data without reference