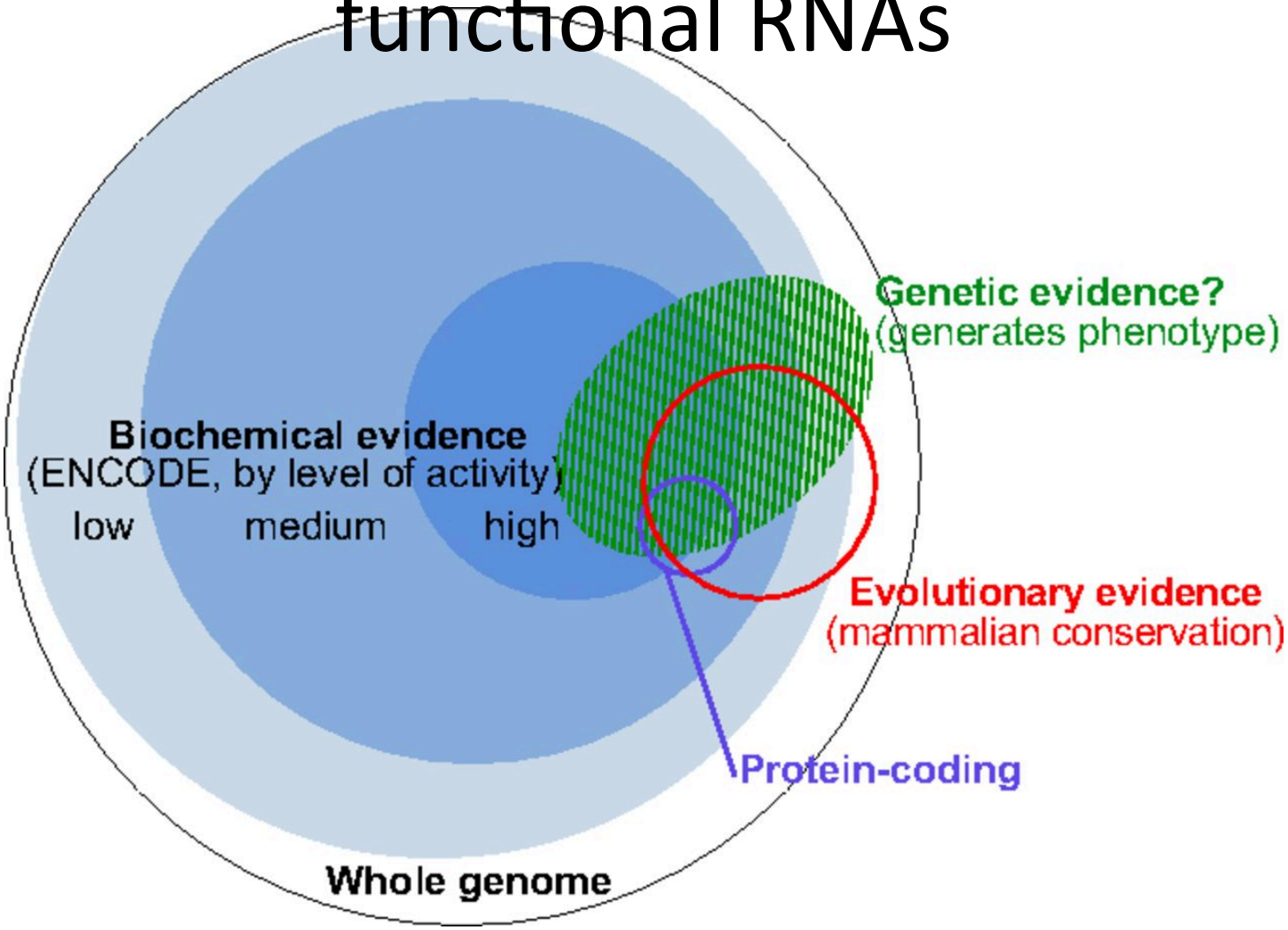
RNAseq analysis

-it's complicated

Oktober 2016

RNA reads are not enough to identify functional RNAs



Defining functional DNA elements in the human genome
Kellis M et al. PNAS 2014;111:6131-6138

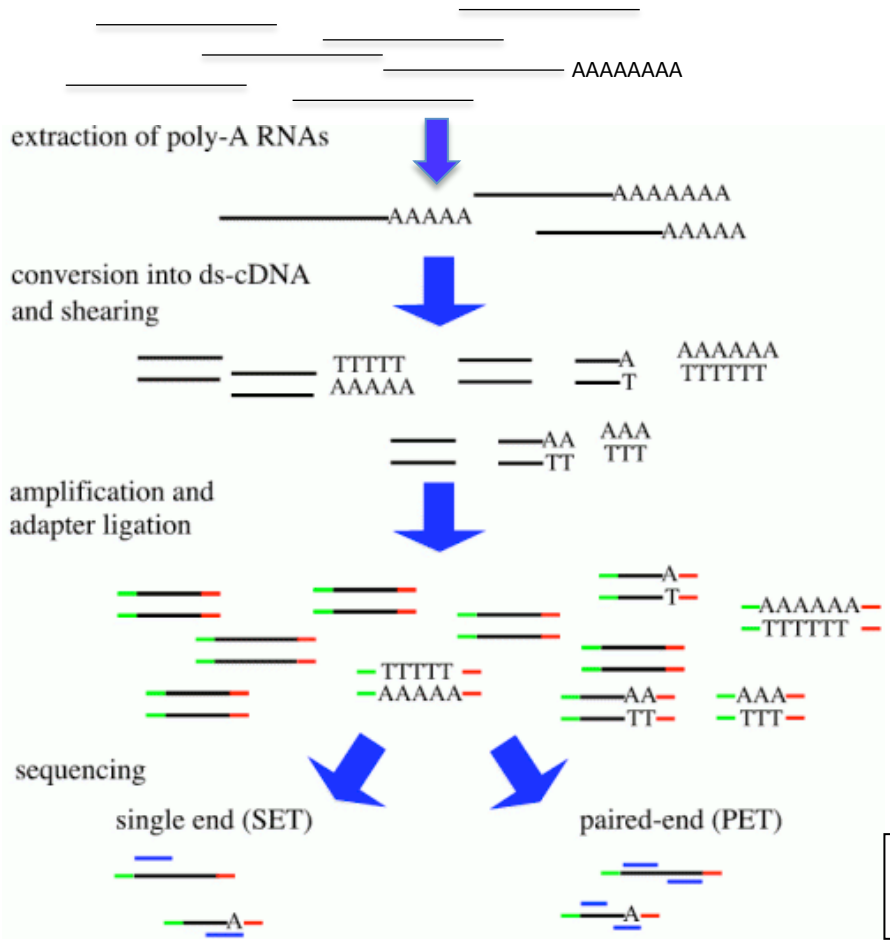
Depending on the steps from sample to RNA seq will give different results

RNA->

enrichments ->

library ->

reads ->



PolyA	(mRNA)
RiboMinus	(- rRNA)
Size <50 nt	(miRNA)
…..	

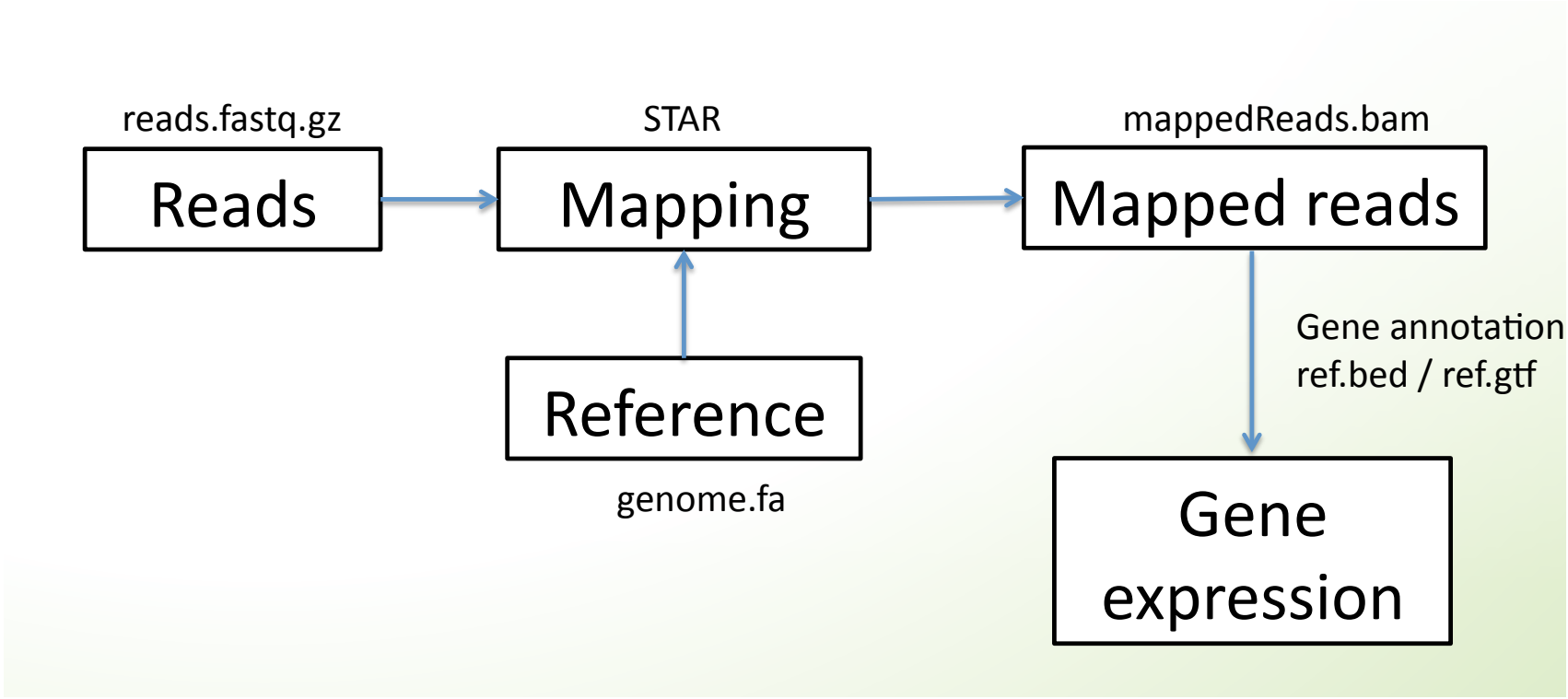
Size of fragment
Strand specific
5’ end specific
3’ end specific
…..

Single end (1 read per fragment)
Paired end (2 reads per fragment)

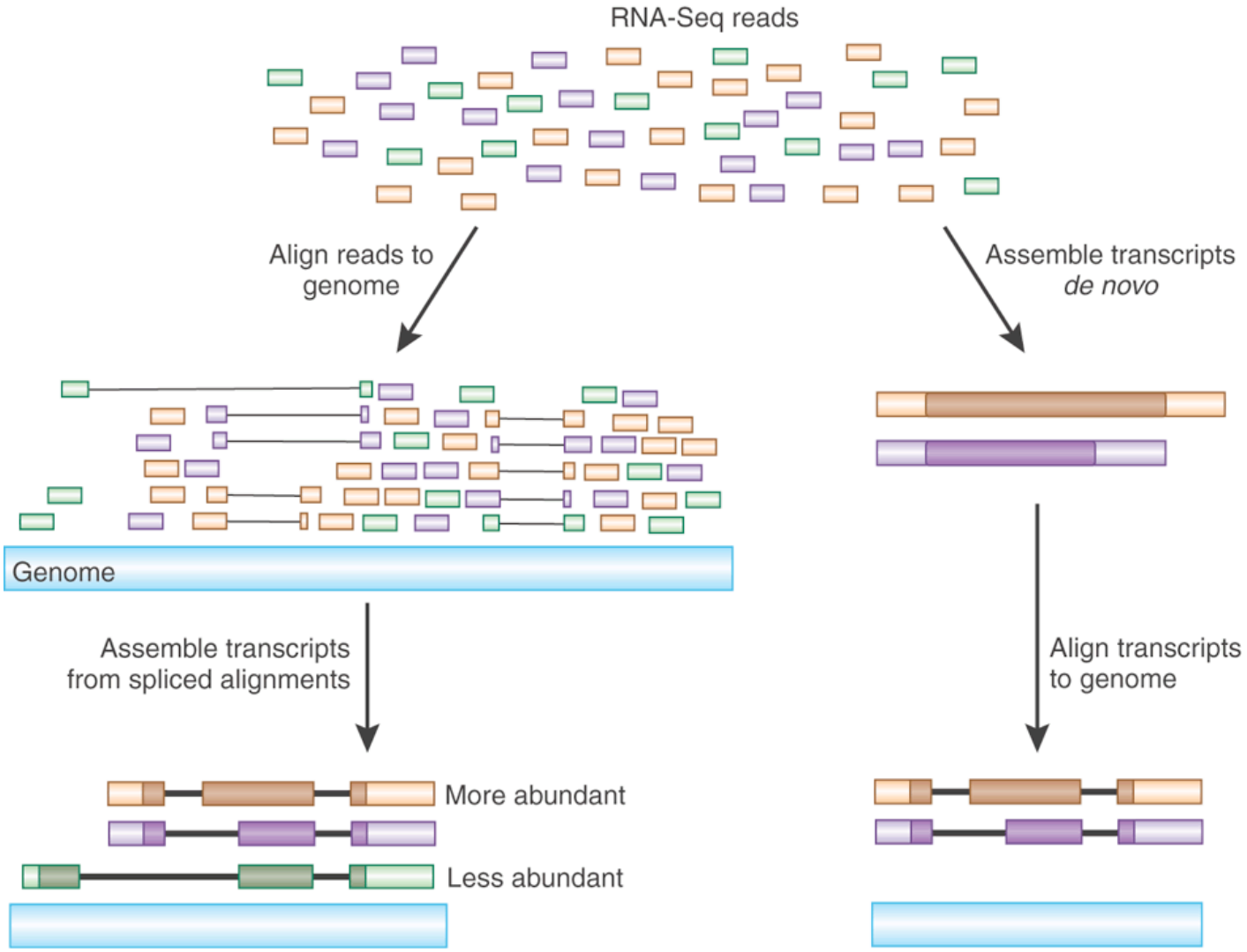
Mapping (Pär Engström)

- Use RNA specific mapper
- Use a two-pass workflow
- STAR or HISAT
- If you want to run Cufflinks, use TopHat or HISAT
- For long (PacBio) reads, STAR, BLAT or GMAP can be used

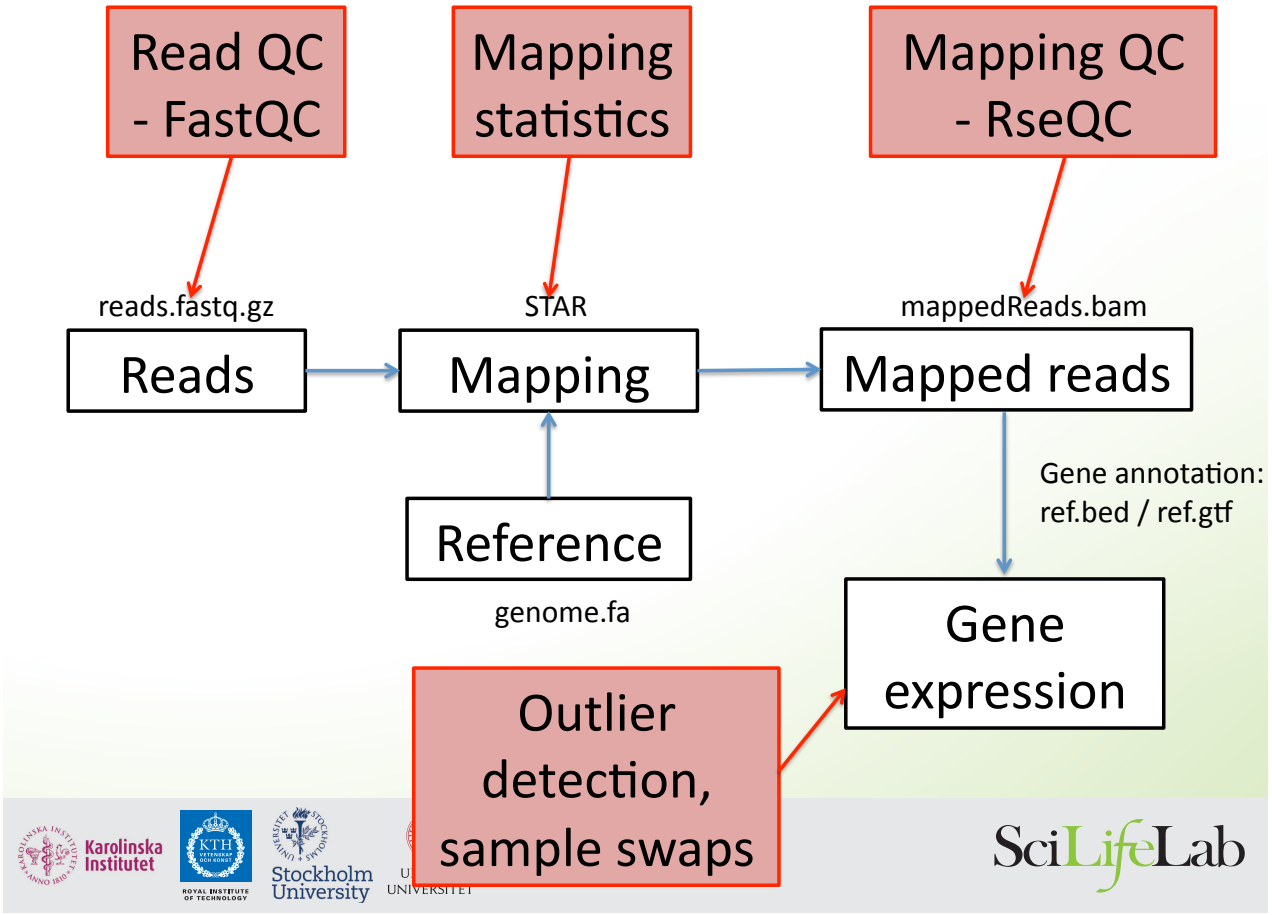
RNA-seq analysis workflow



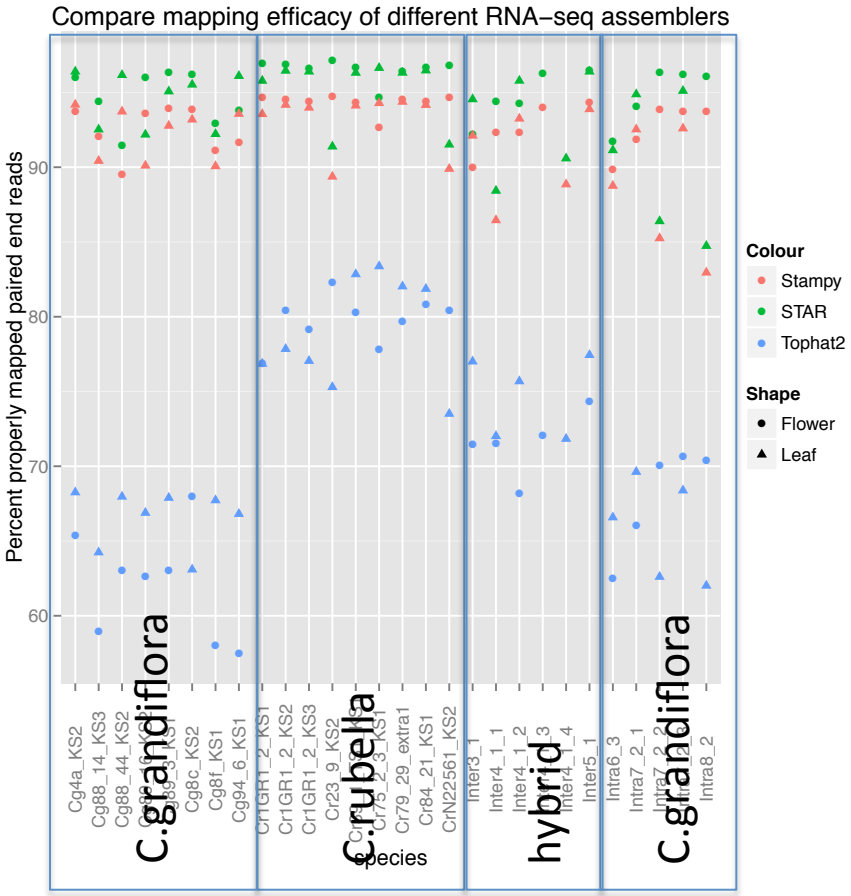
Gene and Isoform detection



Do a lot of QC


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

More variation when using top hat 2 with default settings than when using STAR or Stampy with default setting



RNA QC Åsa Björklund

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

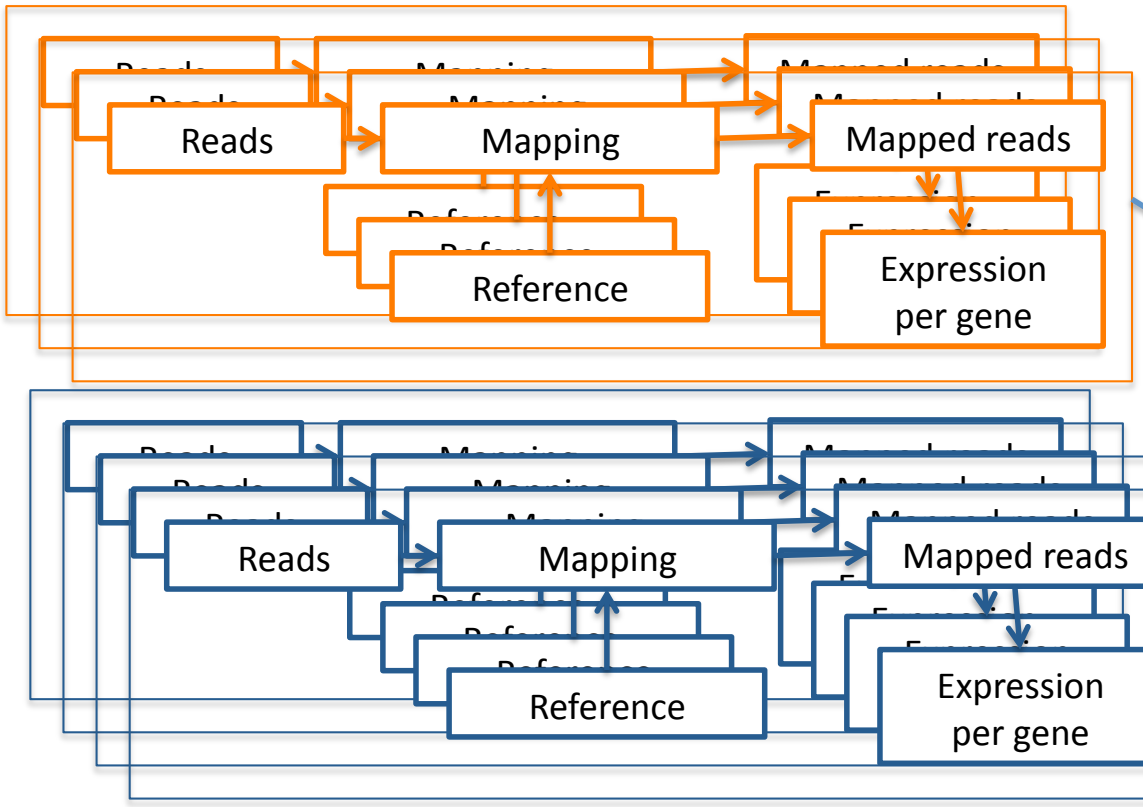
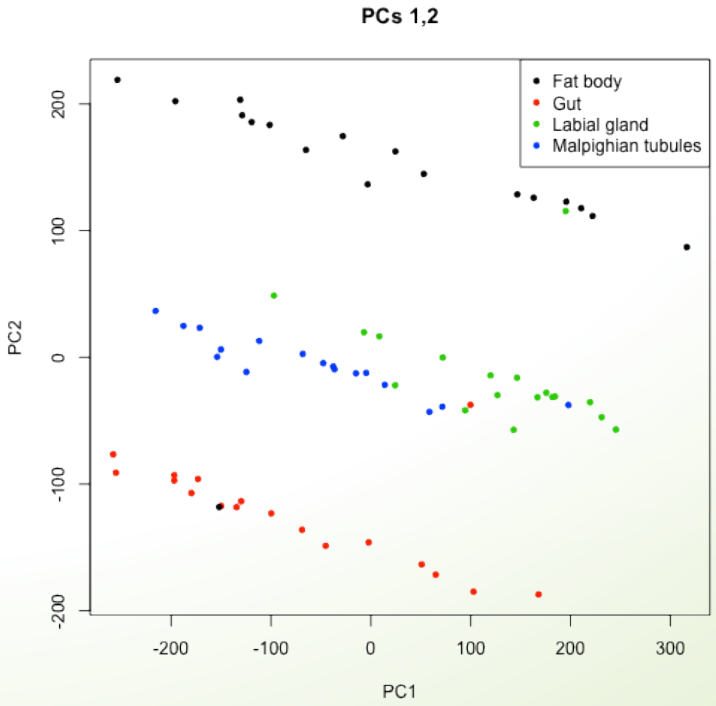


Table with counts, rpkms, fpkms or similar

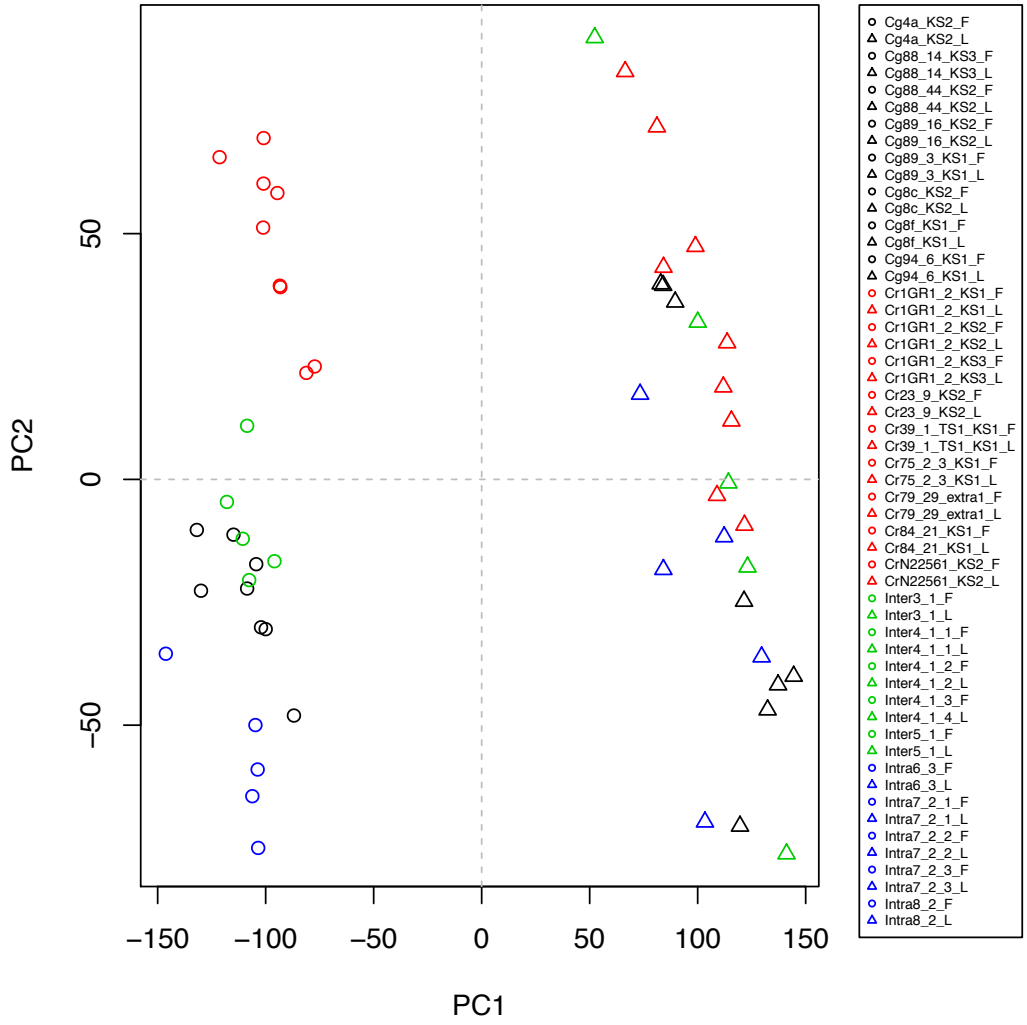
Sample swaps and outliers can be identified using PCA

RNA QC

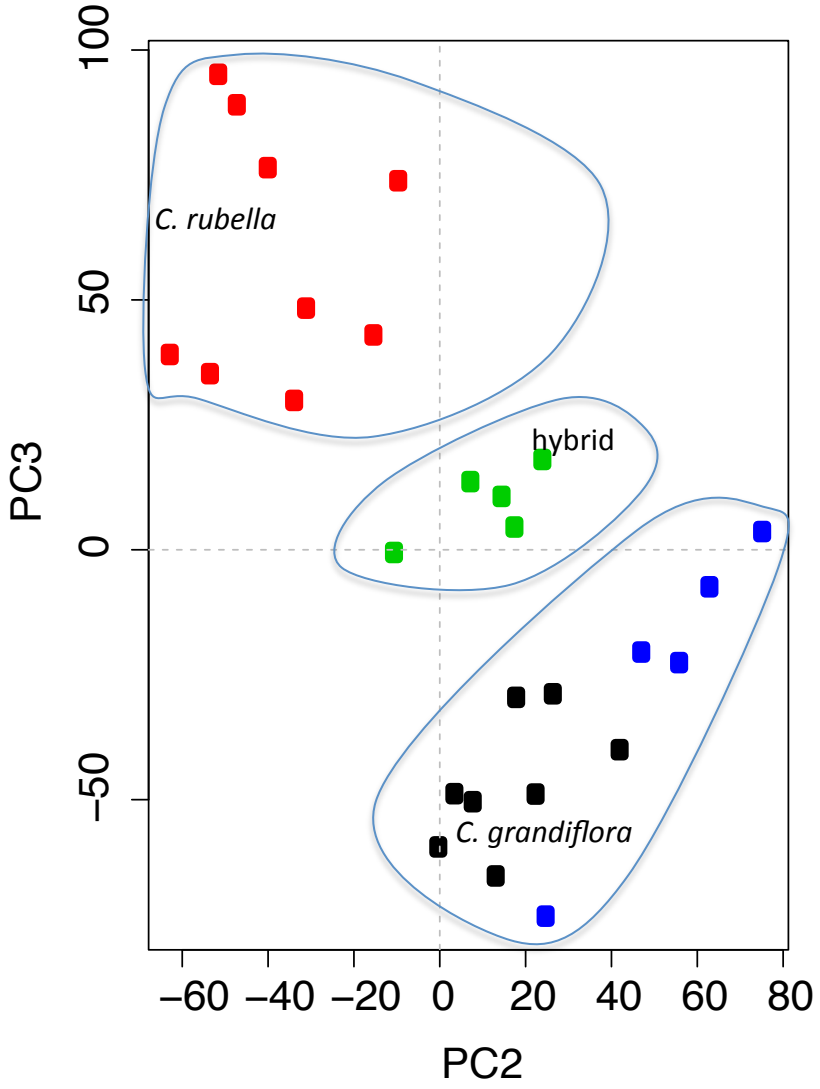
PCA analysis detected potential sample swaps



Principal component 1 separates samples from flowers and leaves



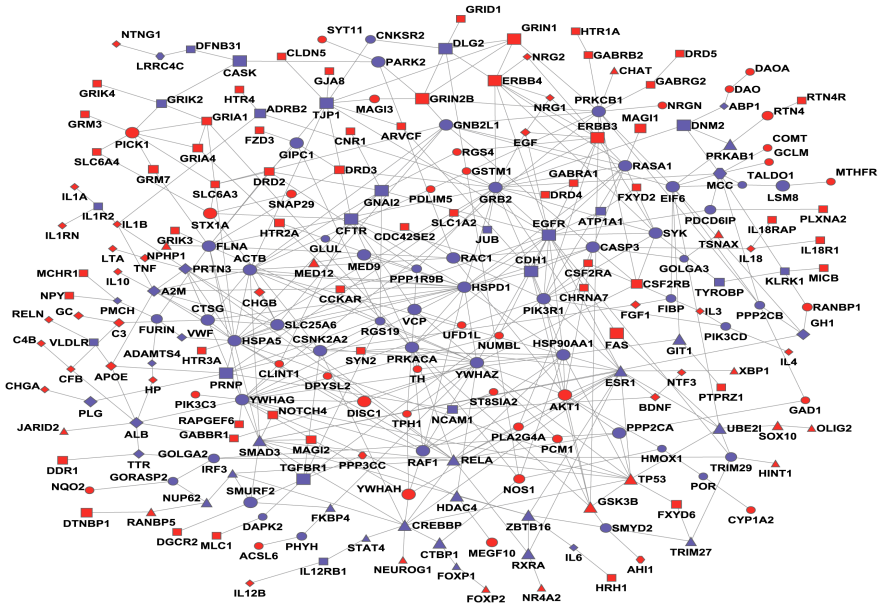
Principal component 2 and 3 separates the different species



Differential expression analysis
Mikael Huss

The identification of genes (or other types of genomic features, such as transcripts or exons) that are expressed in significantly different quantities in distinct groups of samples, be it biological conditions (drug-treated vs. controls), diseased vs. healthy individuals, different tissues, different stages of development, or something else.

Typically **univariate** analysis (one gene at a time) – even though we know that genes are not independent



Decision tree for software selection (2016)

Differentially expressed **exons** => *DEXSeq* *Sleuth*

Differentially expressed **isoforms** => *BitSeq*, ~~*Cuffdiff*~~ *or ebSeq*

Differentially expressed genes => **Select type of experimental design**

Complex design (more than one varying factor) => *DESeq, edgeR, limma*, *Sleuth*

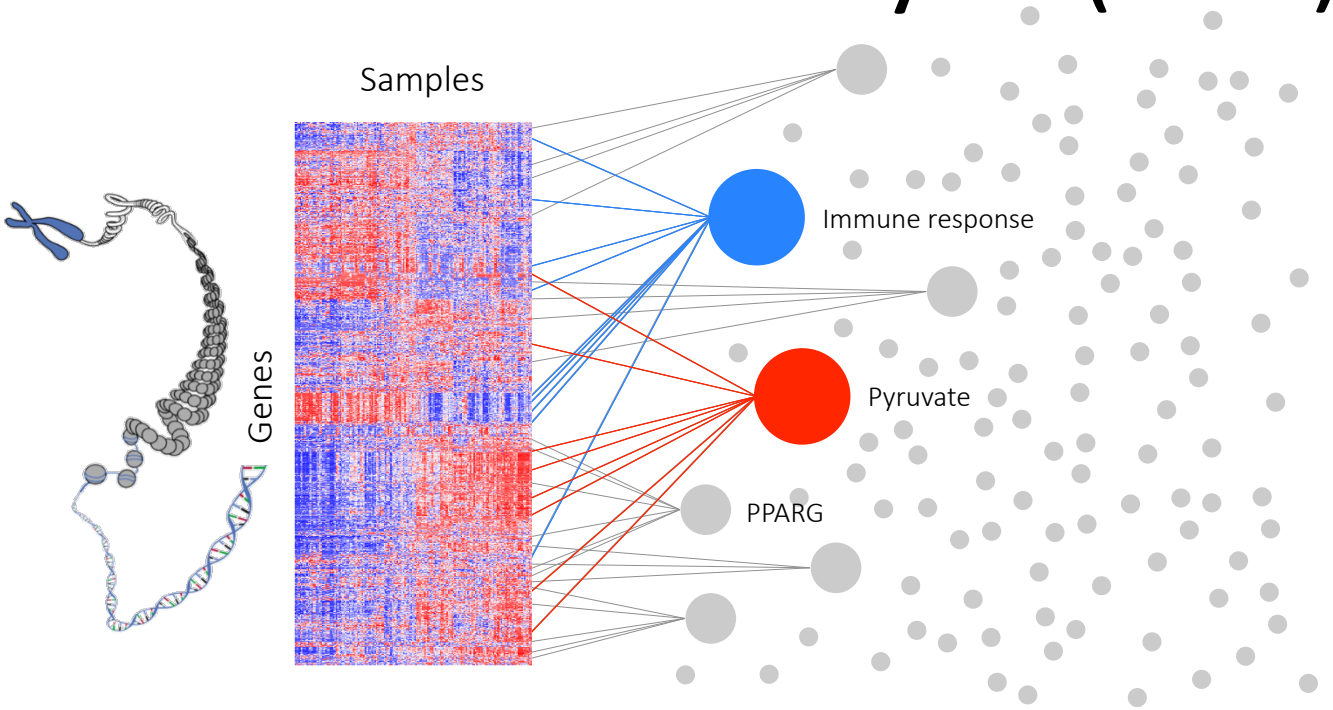
Simple comparison of groups => **How many biological replicates?**

More than about 5 biological replicates per group => ~~*SAMSeq*~~

Less than 5 biological replicates per group => *DESeq, edgeR, limma*

?

Gene-set analysis (GSA)

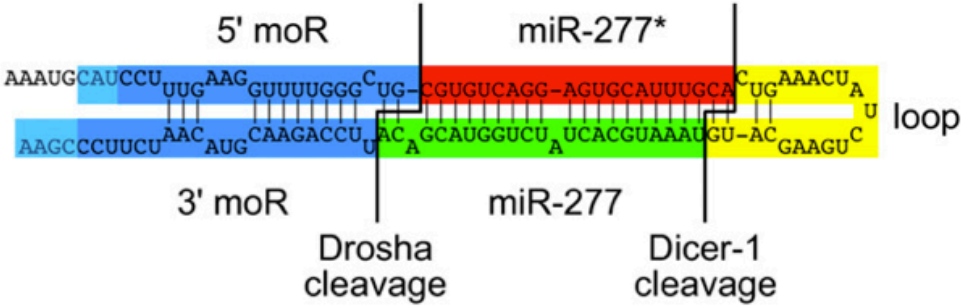


GO-terms
Pathways
Chromosomal locations
Transcription factors
Histone modifications
Diseases
etc…

Gene-level data → Gene-set analysis → Gene-set data (results)

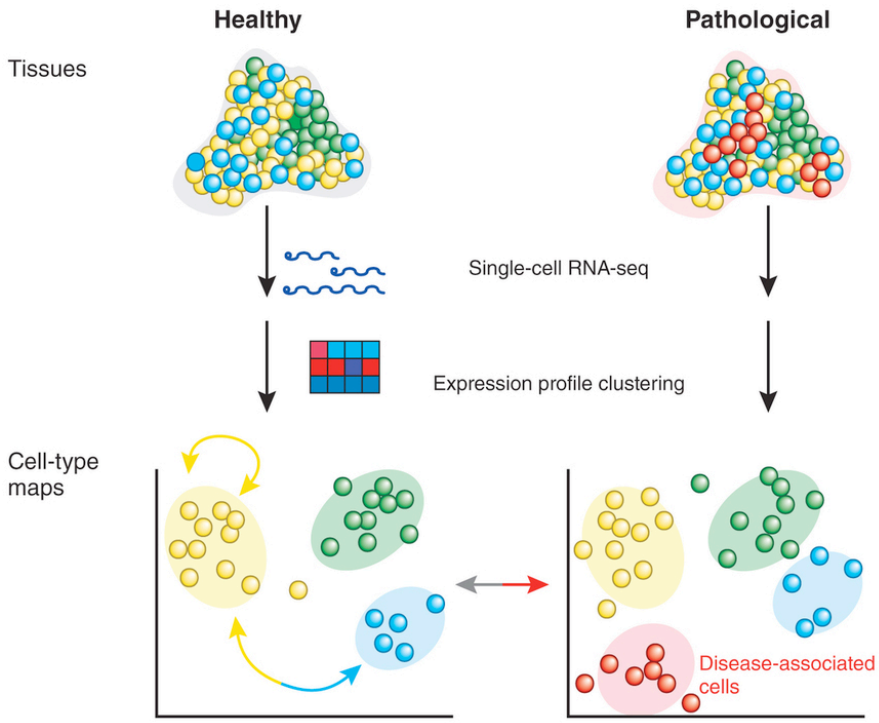
We will focus on transcriptomics and differential expression analysis
However, GSA can in principle be used on all types of genome-wide data.

miRNA seq analysis



(Berezikov et al. Genome Research, 2011.)

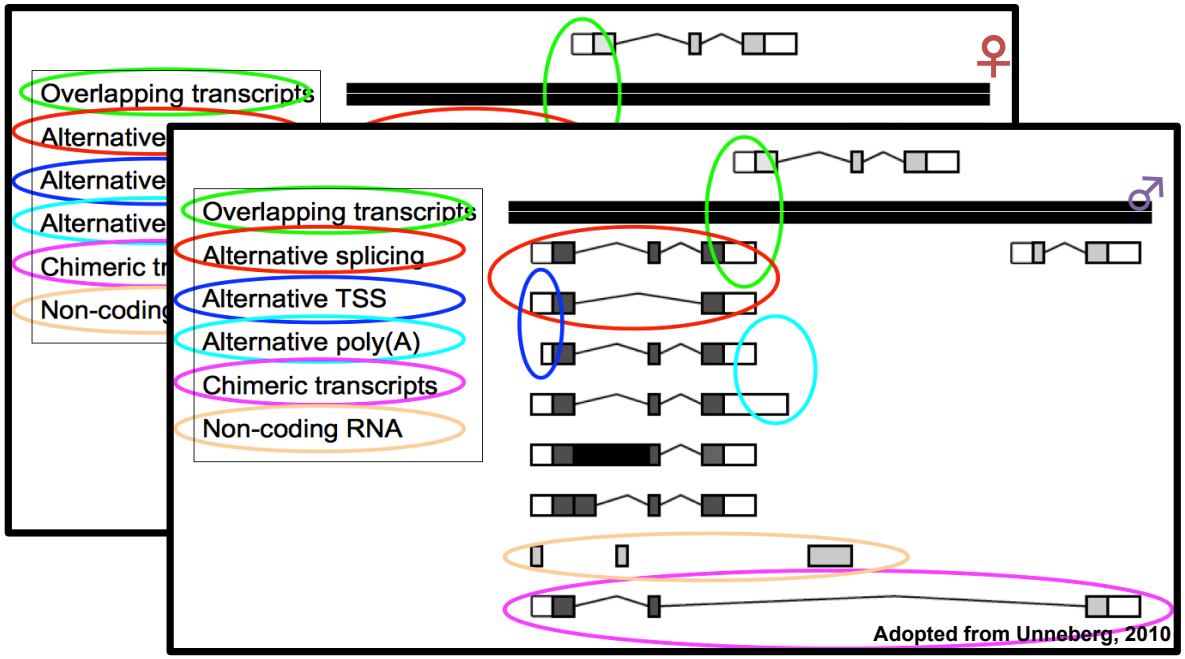
Single cell sequencing



(Sandberg, Nature Methods 2014)

Allele specific expression

Adding another layer to transcriptome complexity...



...and **each** gene is present on **two** chromosomes.

=> it has **two *alleles***

Exercises

- Mapping
  - STAR
  - HISAT2
- Tutorial for reference guided assembly
  - Cufflinks
  - Stringtie
- Tutorial for de novo assembly
  - Trinity
- Visualise mapped reads and assembled transcripts on reference
  - IGV
- RNA quality controll
  - Tutorial for RNA seq Quality Control
- Differential expression analysis
  - DEseq2
  - Calisto and Sleuth
  - multi variate analysis in SIMCA
- small RNA analysis
  - miRNA analysis

- **Introductory**
  - Introduction to the RNA seq data provided
  - Short introduction to R
  - Short introduction to IGV
- **Beta labs**
  - Single cell RNA PCA and clustering
  - Gene set analysis
- **UPPMAX**
  - sbatch script example

Need help??

- We are here for you. Apply for help.