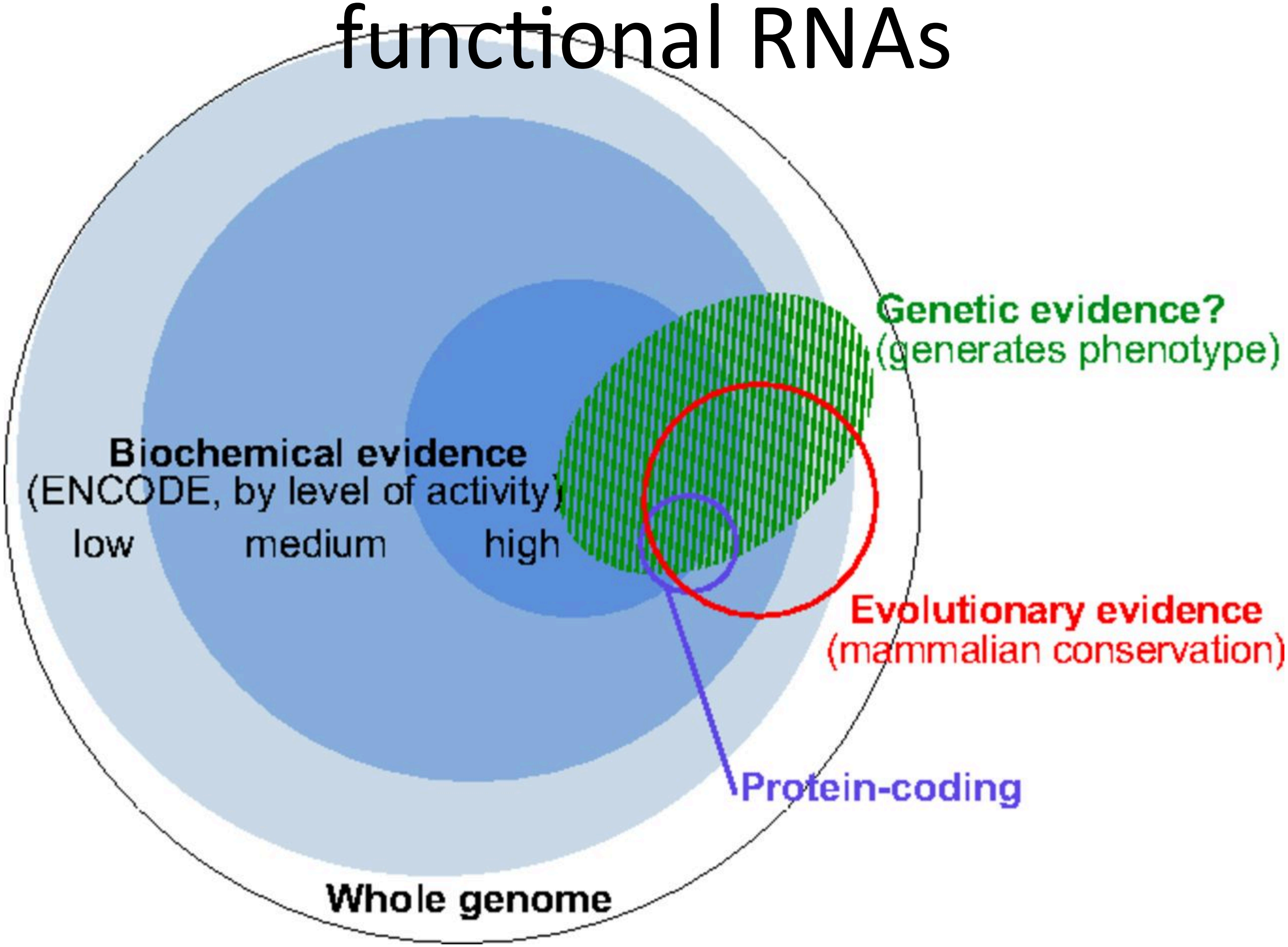
RNAseq analysis

-it's complicated

March 2017

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

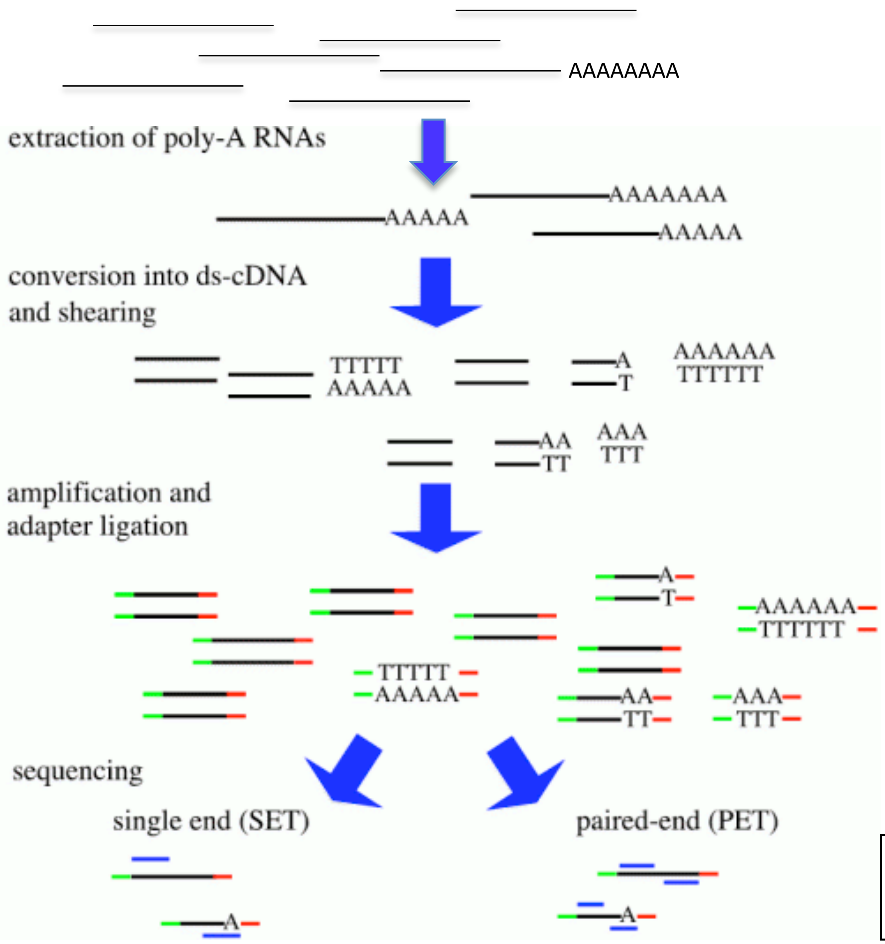
Depending on the steps from sample to RNA seq will give different results

RNA->

enrichments ->

library ->

reads ->



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

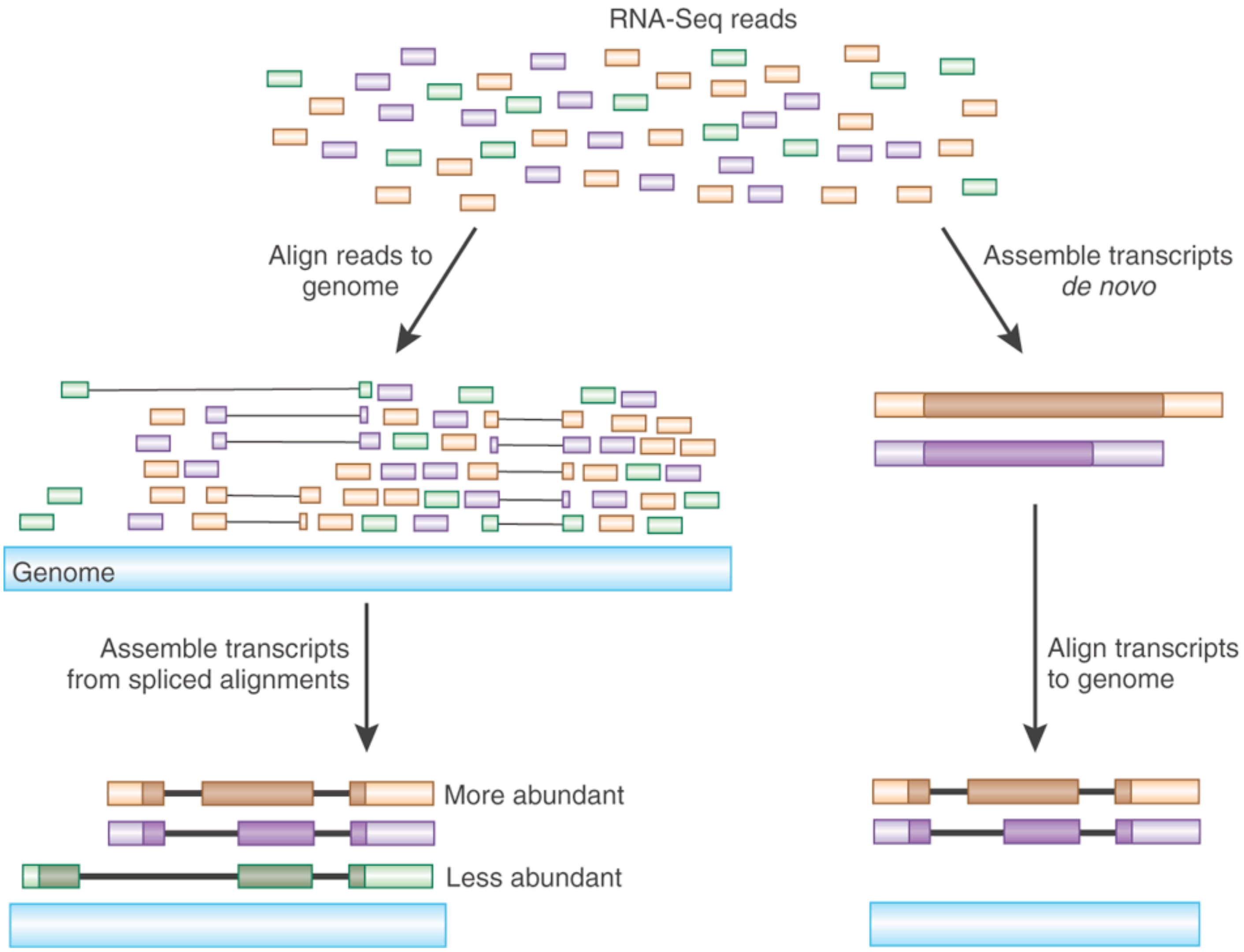
Size of fragment
Strand specific
5’ end specific
3’ end specific
…..

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
- For long (PacBio) reads, STAR, BLAT or GMAP can be used

Gene and Isoform detection


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Long reads might be the way to go

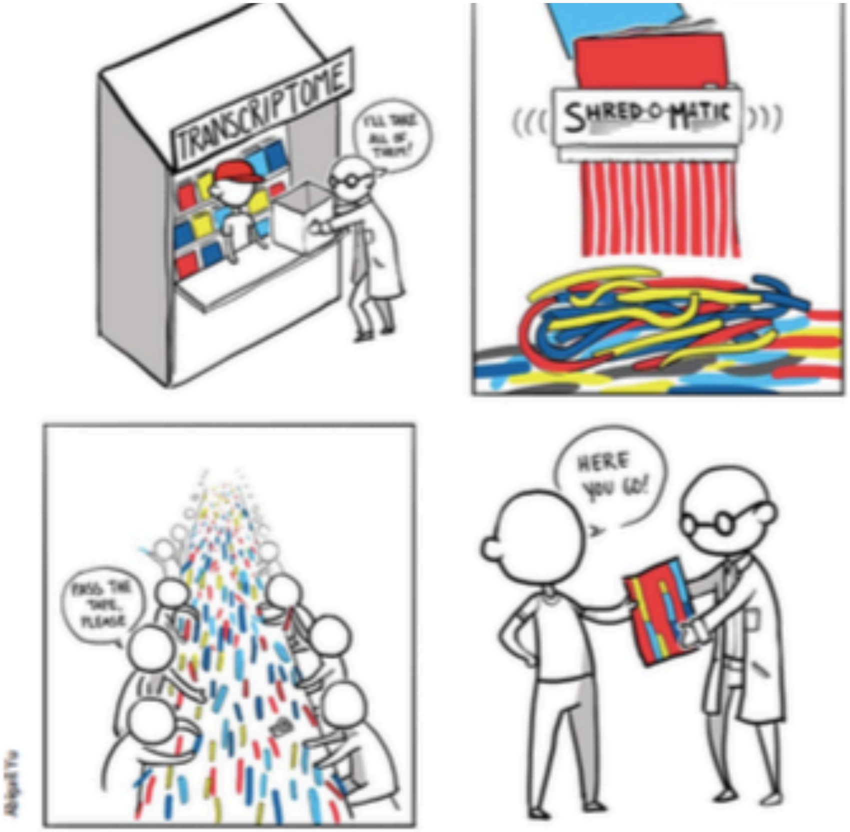


Figure 1 | Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.

Ian Korf (2013) Genomics: the state of the art in

"The way we do RNA-seq now is… you take the transcriptome, you **blow it up into pieces** and then you try to figure out **how they all go back together again…** If you think about it, it's kind of a **crazy way to do things."**

Michael Snyder
Stanford University

Tal Nawy (2013) End-to-end RNA sequencing, *Nature Methods* 10: 1144–1145

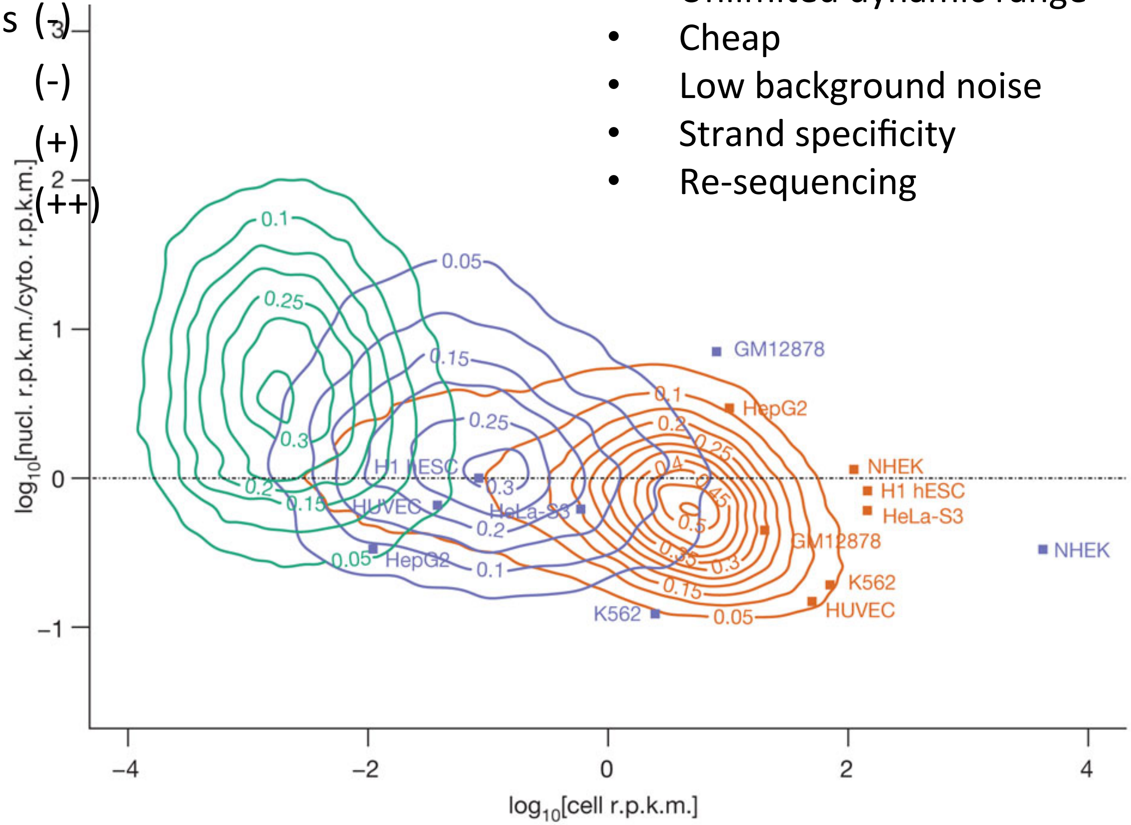
Promises and pitfalls

Long reads

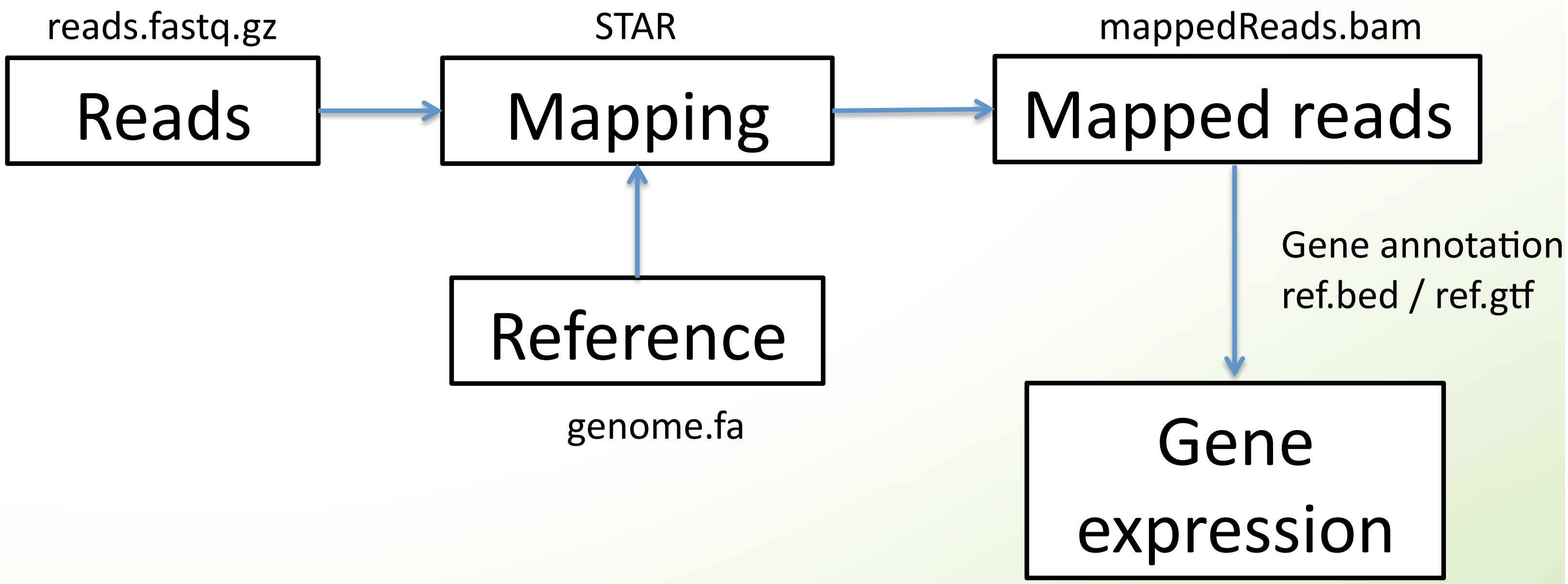
Low throughput	(-)
Complete transcripts	(++)
Not quantitative	(-)
Only highly expressed genes	(-)
Expensive	(-)
Low background noise	(+)
Easy downstream analysis	(++)

short reads

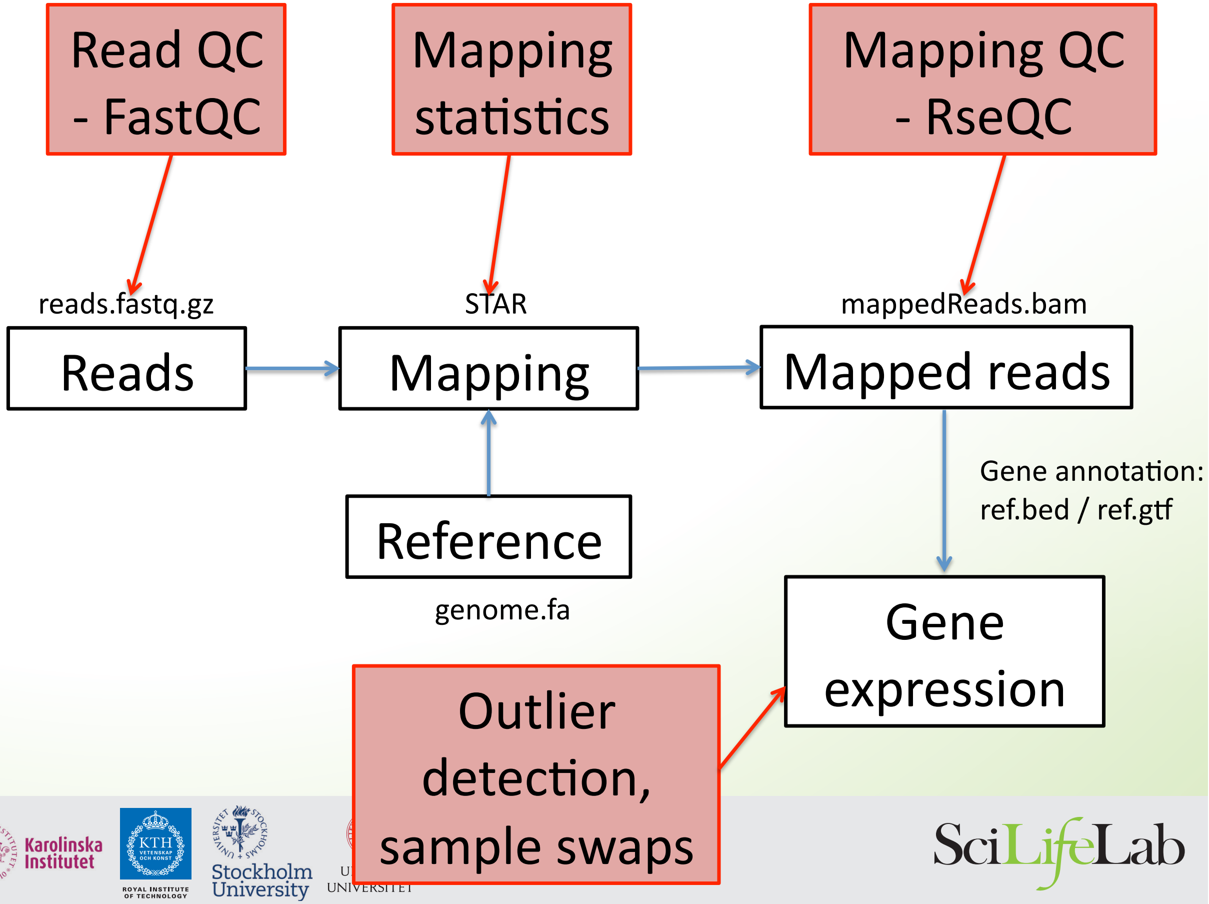
High throughput	(++)
Quantitative	(++)
Fractions of transcripts	(-)
Full dynamic range	(+-)
Unlimited dynamic range	(+)
Cheap	(+)
Low background noise	(+)
Strand specificity	(+)
Re-sequencing	(+)



RNA-seq analysis workflow

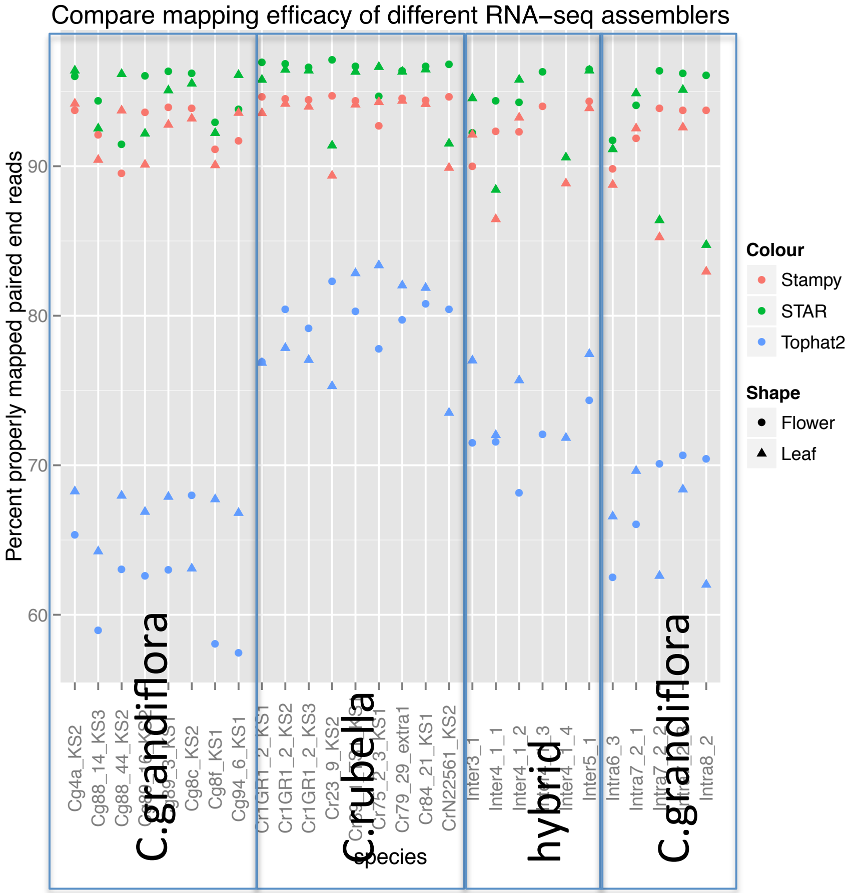


Do a lot of QC



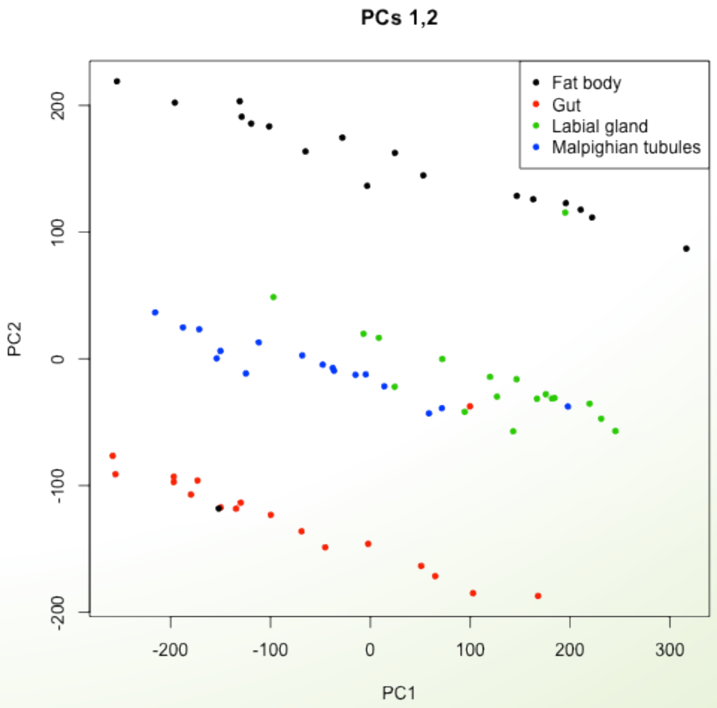
Read QC
- FastQC
Mapping statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene expression
Outlier detection, sample swaps

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

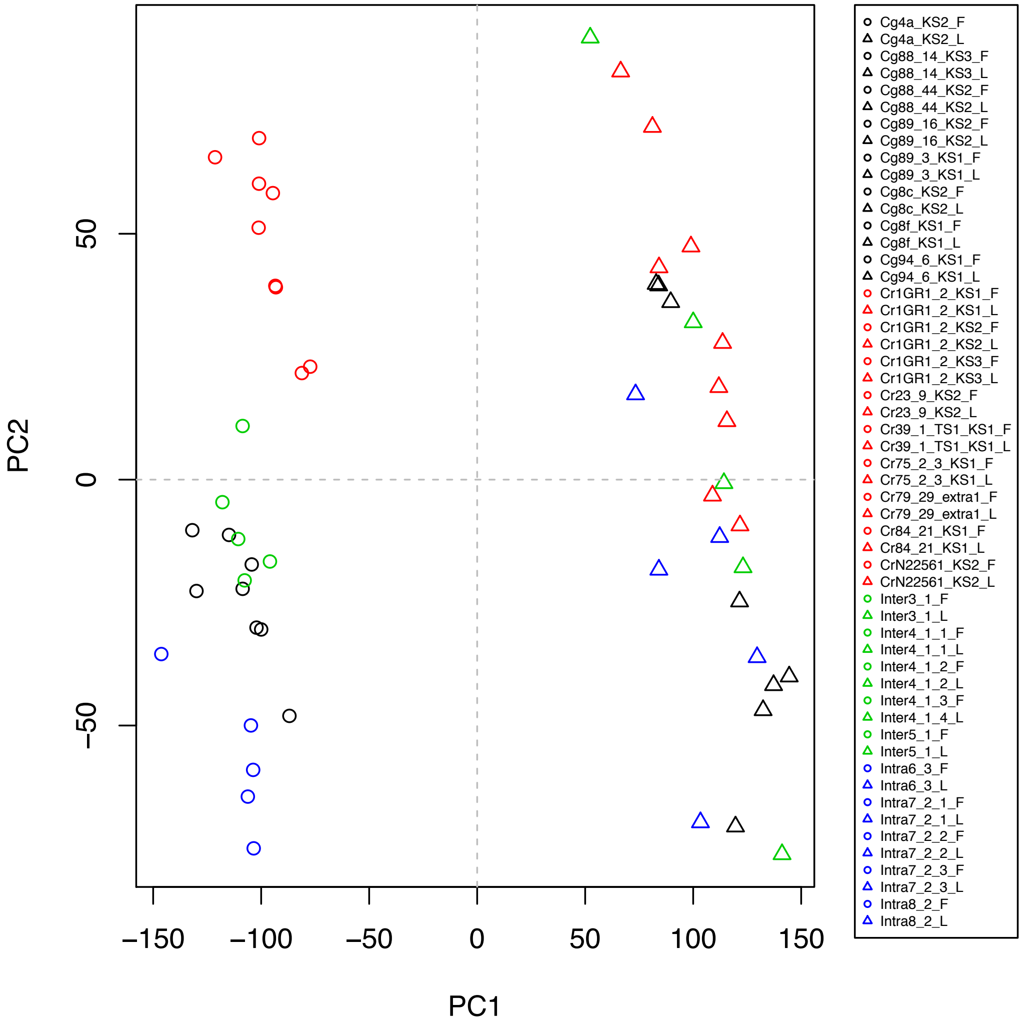


RNA QC

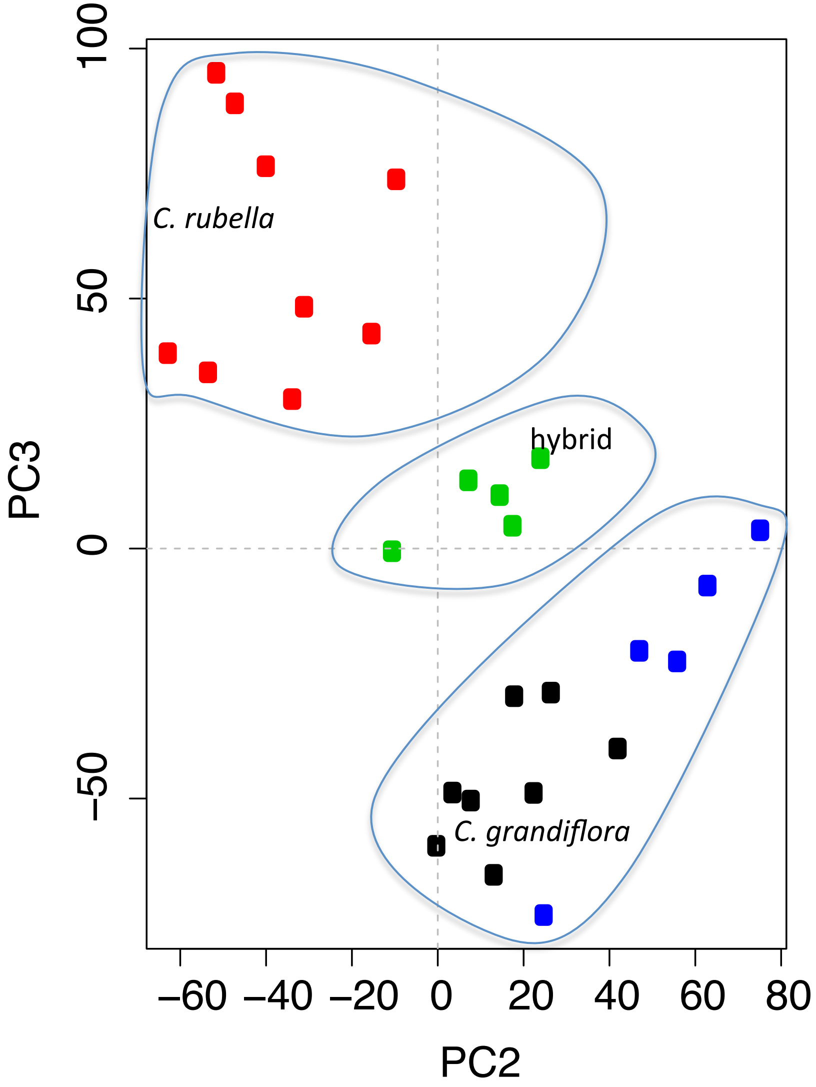
PCA analysis detected potential sample swaps



Principal component 1 separates samples from flowers and leaves



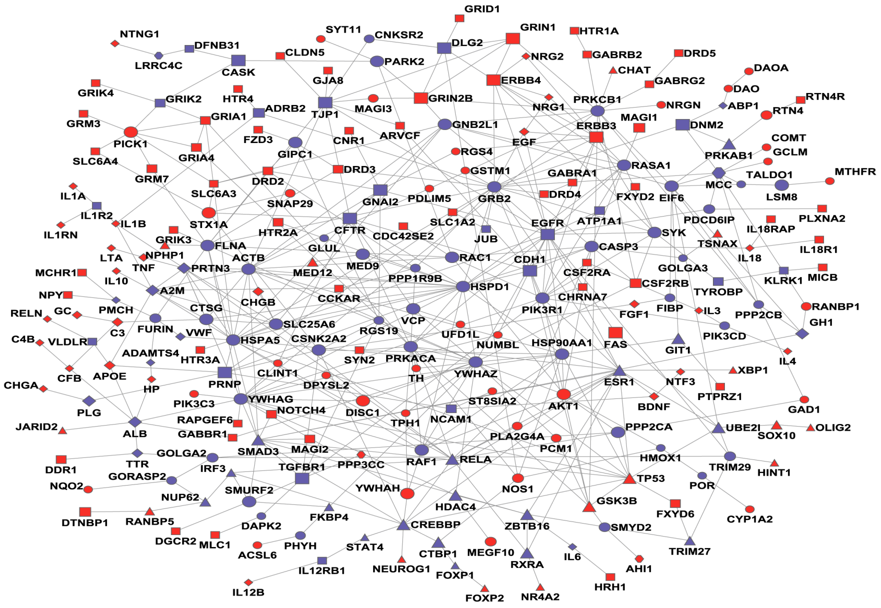
Principal component 2 and 3 separates the different species



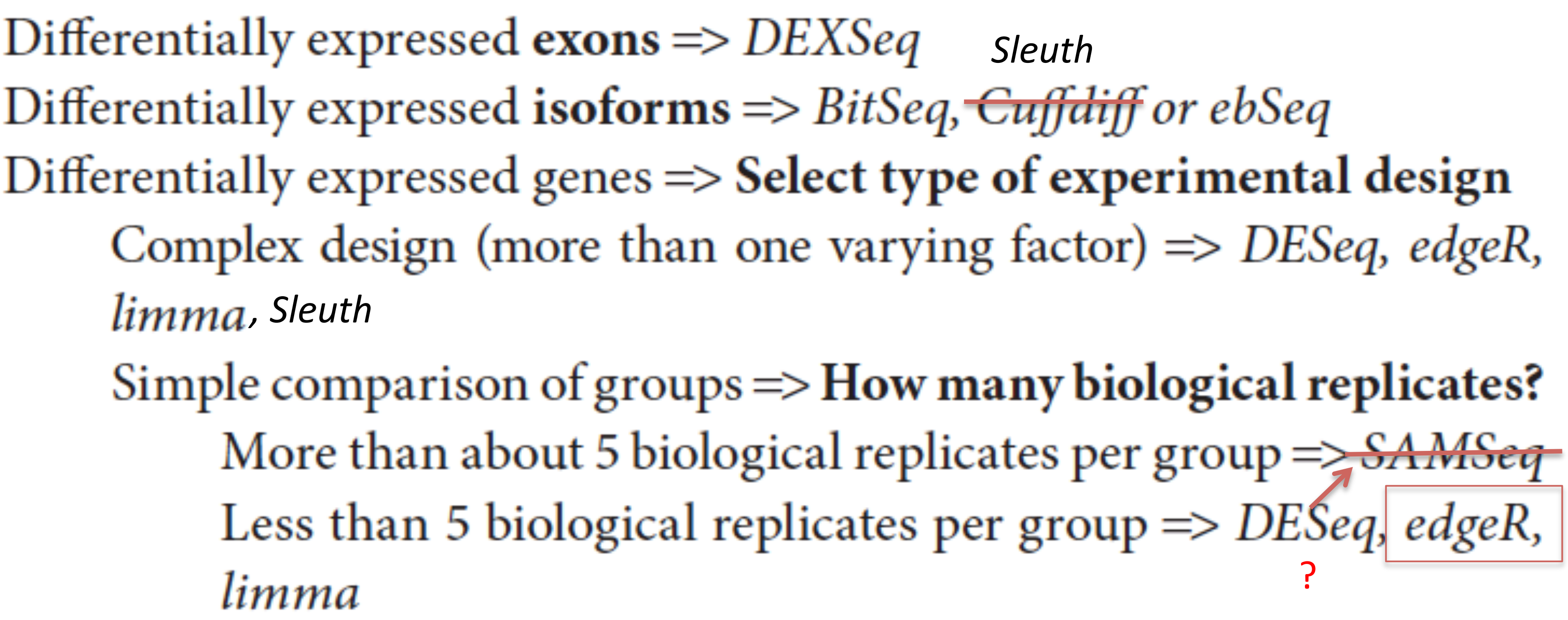
Differential expression analysis
Mikael Huss

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*

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

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

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

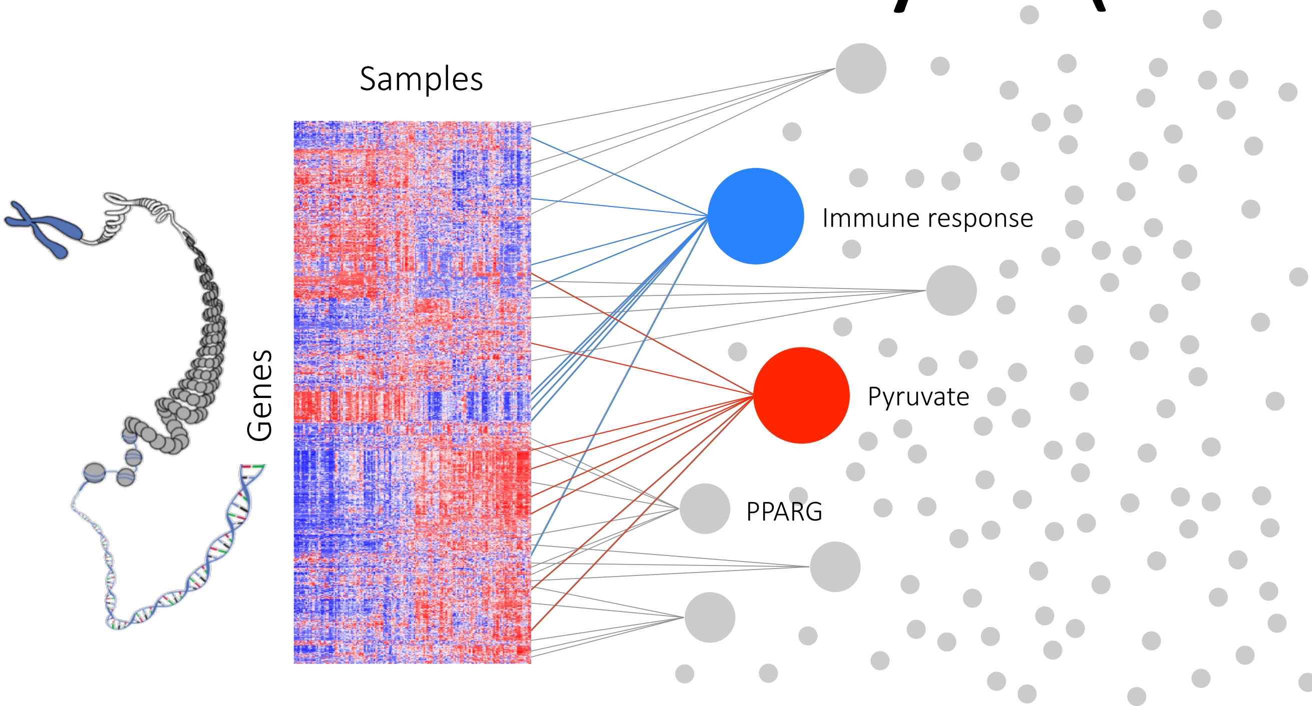
Simple comparison of groups => **How many biological replicates?**

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

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

?

Gene-set analysis (GSA)

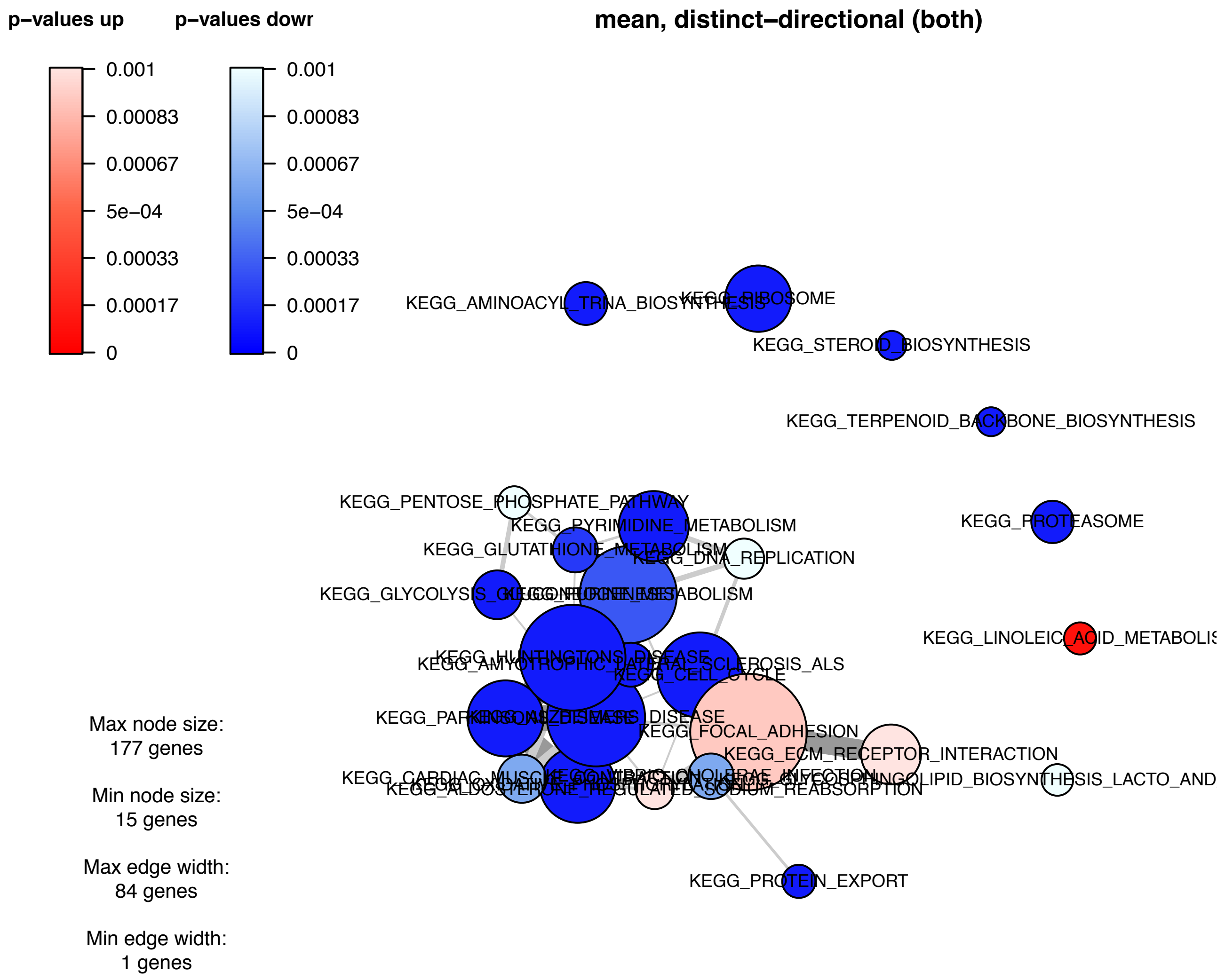


GO-terms
Pathways
Chromosomal locations
Transcription factors
Histone modifications
Diseases
etc…

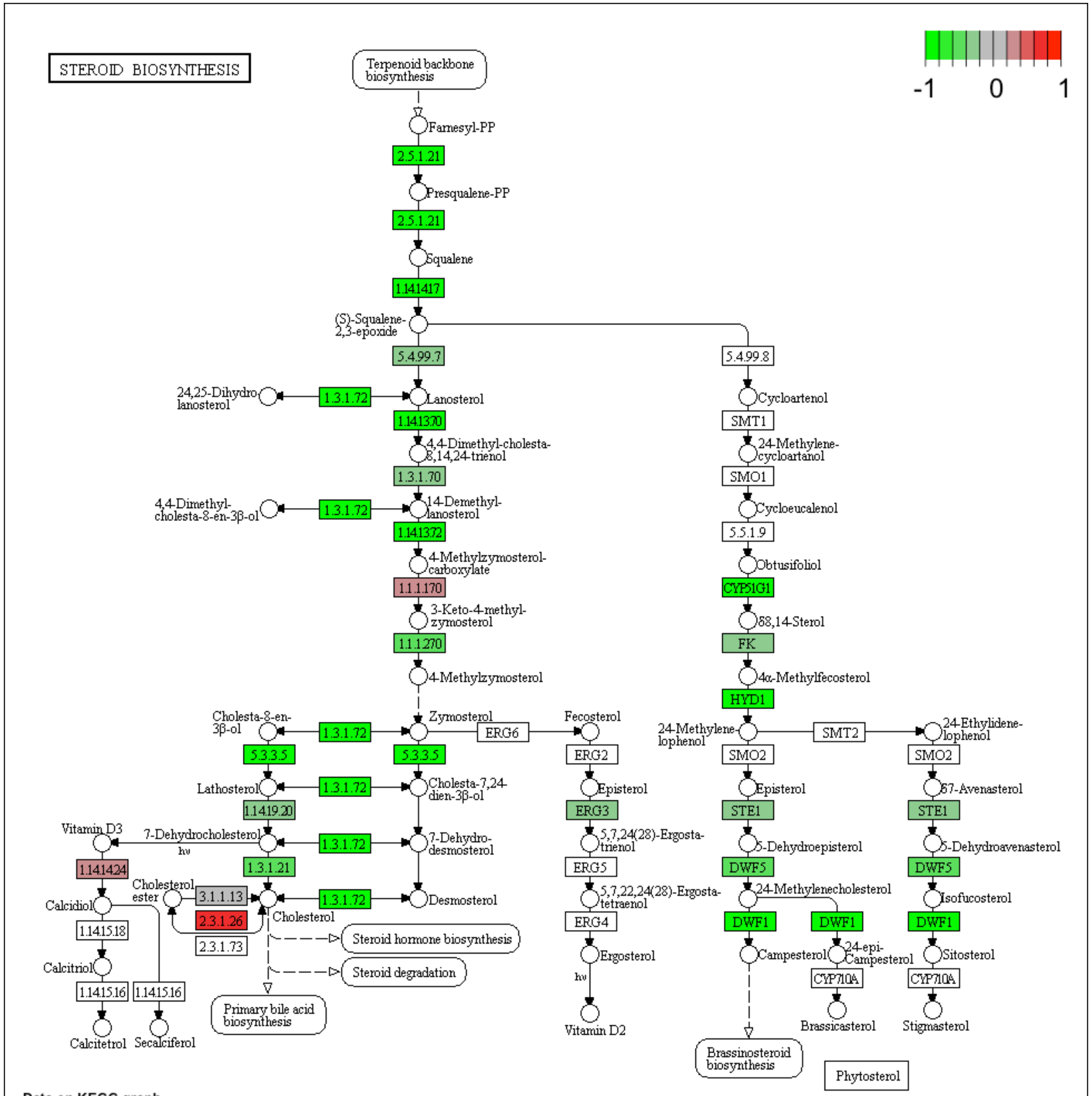
$$\text{Gene-level data} \xrightarrow{\text{Gene-set analysis}} \text{Gene-set data (results)}$$

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

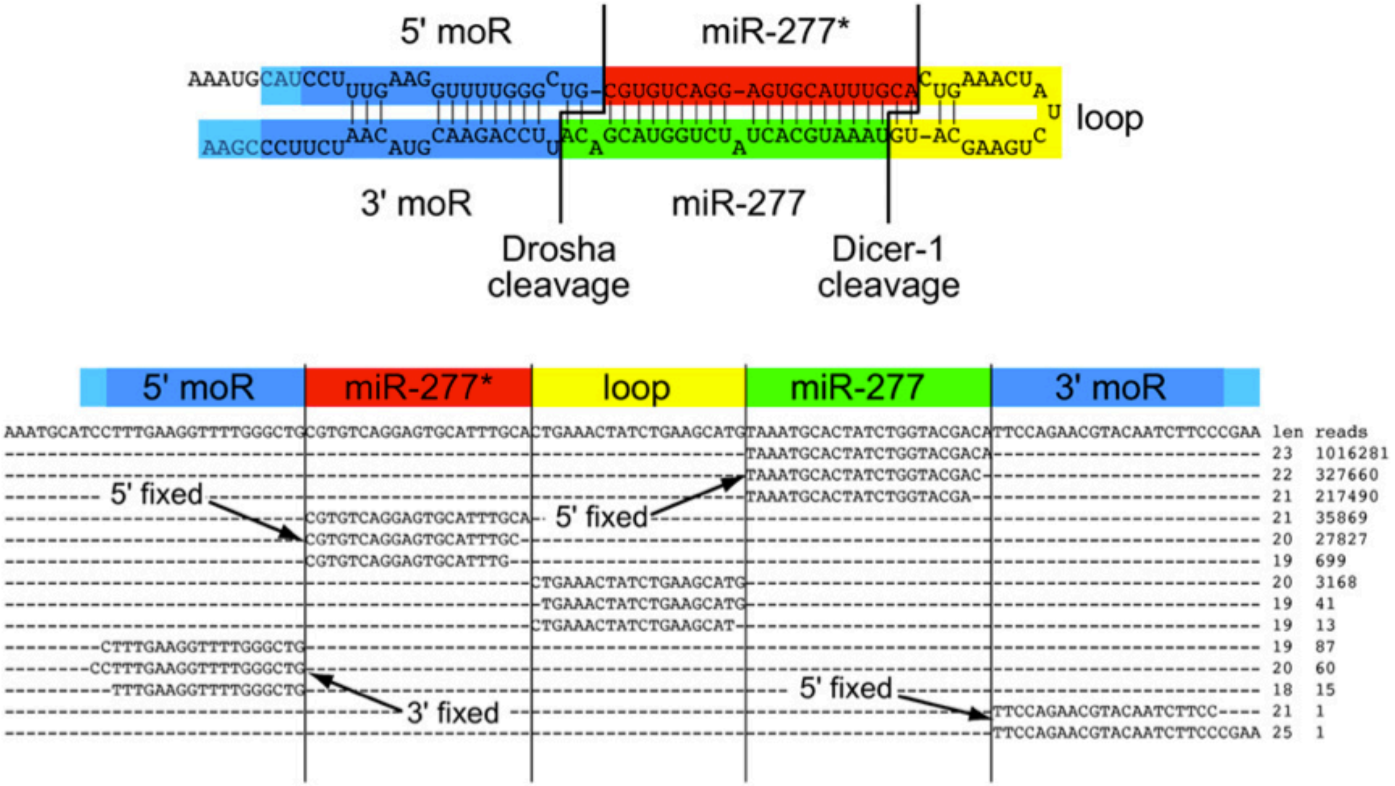
Analysis regarding Type II Diabetes


p−values up
p−values dowr
0.001
0.00083
0.00067
5e−04
0.00033
0.00017
0
mean, distinct−directional (both)
KEGG_AMINOACYL_TRNA_BIOSYNTHESIS
KEGG_RIBOSOME
KEGG_STEROID_BIOSYNTHESIS
KEGG_TERPENOID_BACKBONE_BIOSYNTHESIS
KEGG_PENTOSE_PHOSPHATE_PATHWAY
KEGG_PYRIMIDINE_METABOLISM
KEGG_GLUTATHIONE_METABOLISM
KEGG_DNA_REPLICATION
KEGG_PROTEASOME
KEGG_GLYCOLYSIS_GLUCONEOGENESIS
KEGG_PURINE_METABOLISM
KEGG_LINOLEIC_ACID_METABOLIS
KEGG_HUNTINGTONS_DISEASE
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS
KEGG_CELL_CYCLE
KEGG_PARKINSONS_DISEASE
KEGG_ALZHEIMERS_DISEASE
KEGG_FOCAL_ADHESION
KEGG_ECM_RECEPTOR_INTERACTION
KEGG_CARDIAC_MUSCLE_CONTRACTION
KEGG_OXIDATIVE_PHOSPHORYLATION
KEGG_VIBRIO_CHOLERAE_INFECTION
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND
KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION
KEGG_PROTEIN_EXPORT
Max node size:
177 genes
Min node size:
15 genes
Max edge width:
84 genes
Min edge width:
1 genes

Expression of genes on pathway



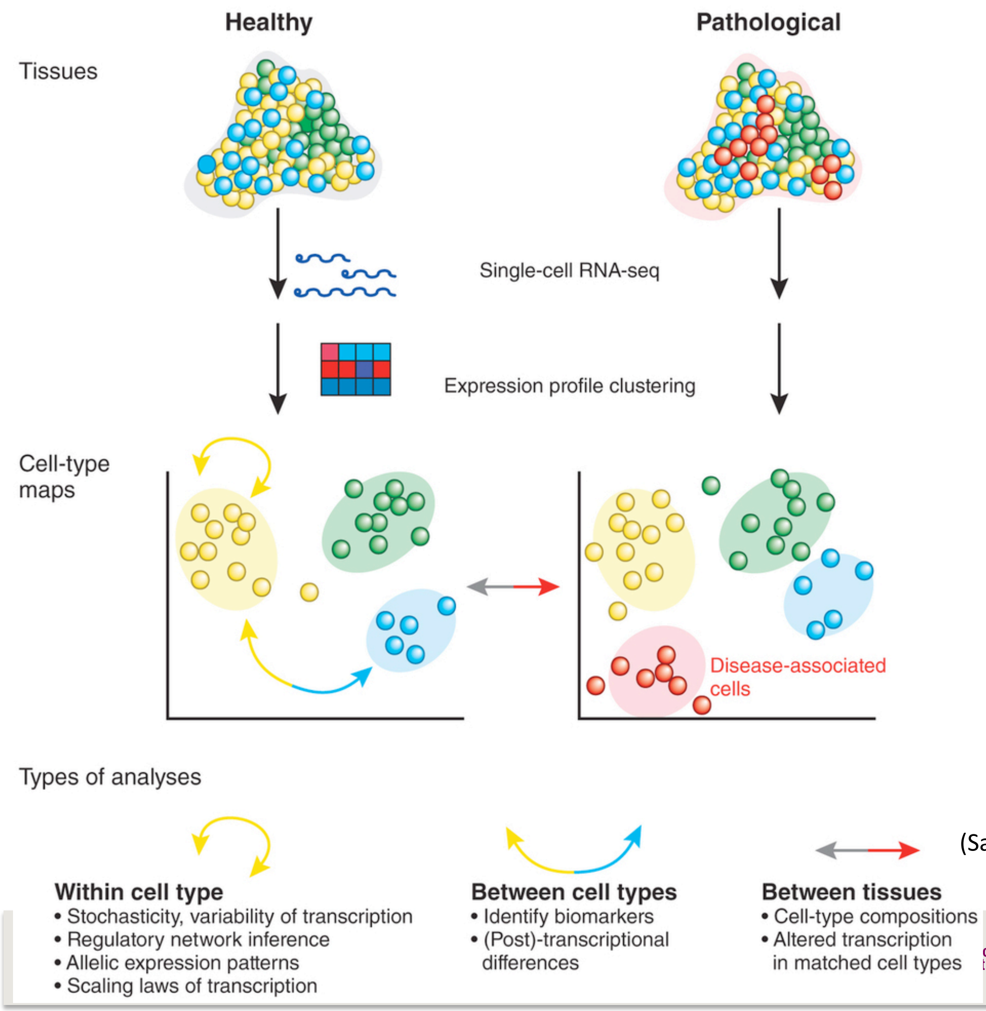
miRNA seq analysis



```
AAATGCATCCTTTGAAGGTTTTGGGCTGCGTGTCAGGAGTGCATTTGCACTGAAACTATCTGAAGCATGTAAATGCACTATCTGGTACGACATTCCAGAACGTACAATCTTCCCGAA len reads
---------------------------------------------------------------------TAAATGCACTATCTGGTACGACA------------------------- 23  1016281
---------------------------------------------------------------------TAAATGCACTATCTGGTACGAC-------------------------- 22  327660
---------------------------------------------------------------------TAAATGCACTATCTGGTACGA--------------------------- 21  217490
----------------------------CGTGTCAGGAGTGCATTTGCA-------------------------------------------------------------------- 21  35869
----------------------------CGTGTCAGGAGTGCATTTGC--------------------------------------------------------------------- 20  27827
----------------------------CGTGTCAGGAGTGCATTTG---------------------------------------------------------------------- 19  699
-------------------------------------------------CTGAAACTATCTGAAGCATG------------------------------------------------ 20  3168
--------------------------------------------------TGAAACTATCTGAAGCATG------------------------------------------------ 19  41
-------------------------------------------------CTGAAACTATCTGAAGCAT------------------------------------------------- 19  13
---------CTTTGAAGGTTTTGGGCTG----------------------------------------------------------------------------------------- 19  87
--------CCTTTGAAGGTTTTGGGCTG----------------------------------------------------------------------------------------- 20  60
----------TTTGAAGGTTTTGGGCTG----------------------------------------------------------------------------------------- 18  15
--------------------------------------------------------------------------------------------TTCCAGAACGTACAATCTTCC---- 21  1
--------------------------------------------------------------------------------------------TTCCAGAACGTACAATCTTCCCGAA 25  1
```

(Berezikov et al. Genome Research, 2011.)

Single cell sequencing



Within cell type
- Stochasticity, variability of transcription
- Regulatory network inference
- Allelic expression patterns
- Scaling laws of transcription

Between cell types
- Identify biomarkers
- (Post)-transcriptional differences

Between tissues
- Cell-type compositions
- Altered transcription in matched cell types

(Sandberg, Nature Methods 2014)

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.