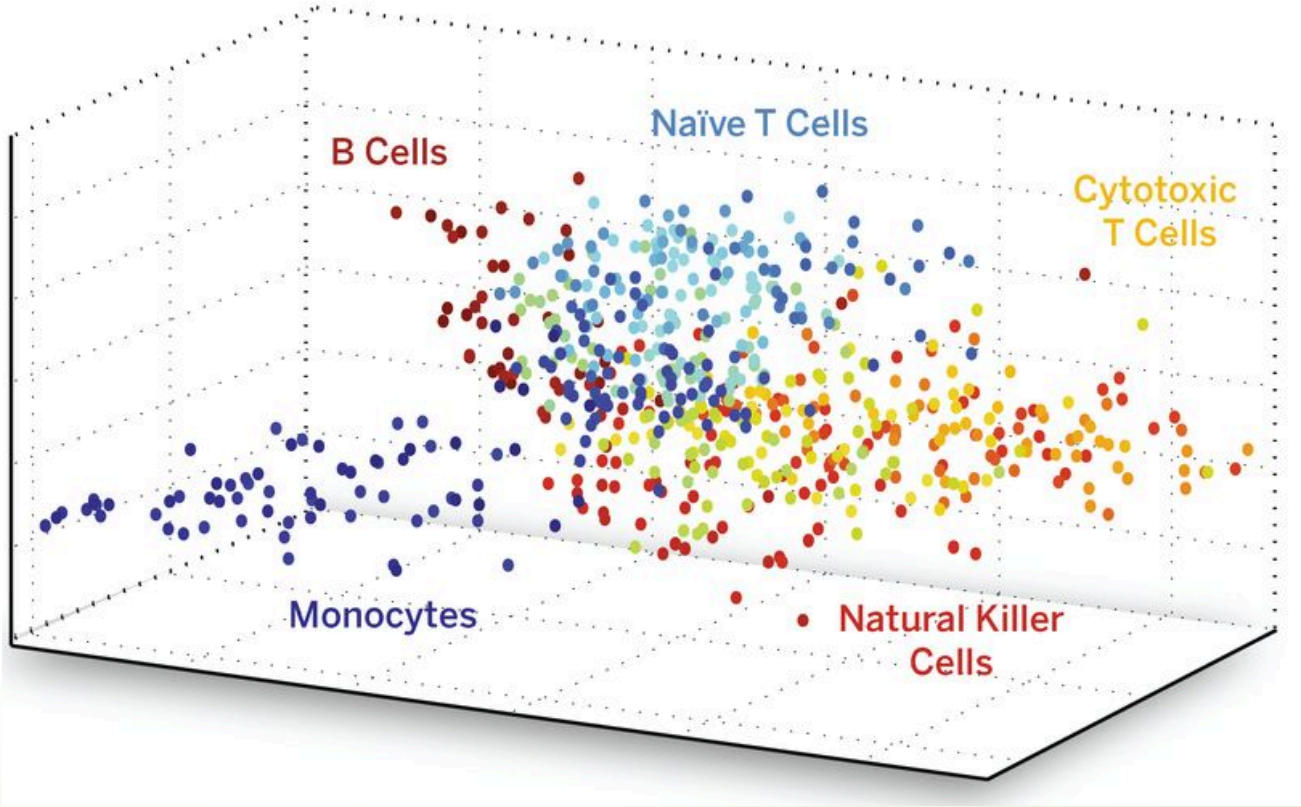
# RNA-seq Introduction

## Promises and pitfalls

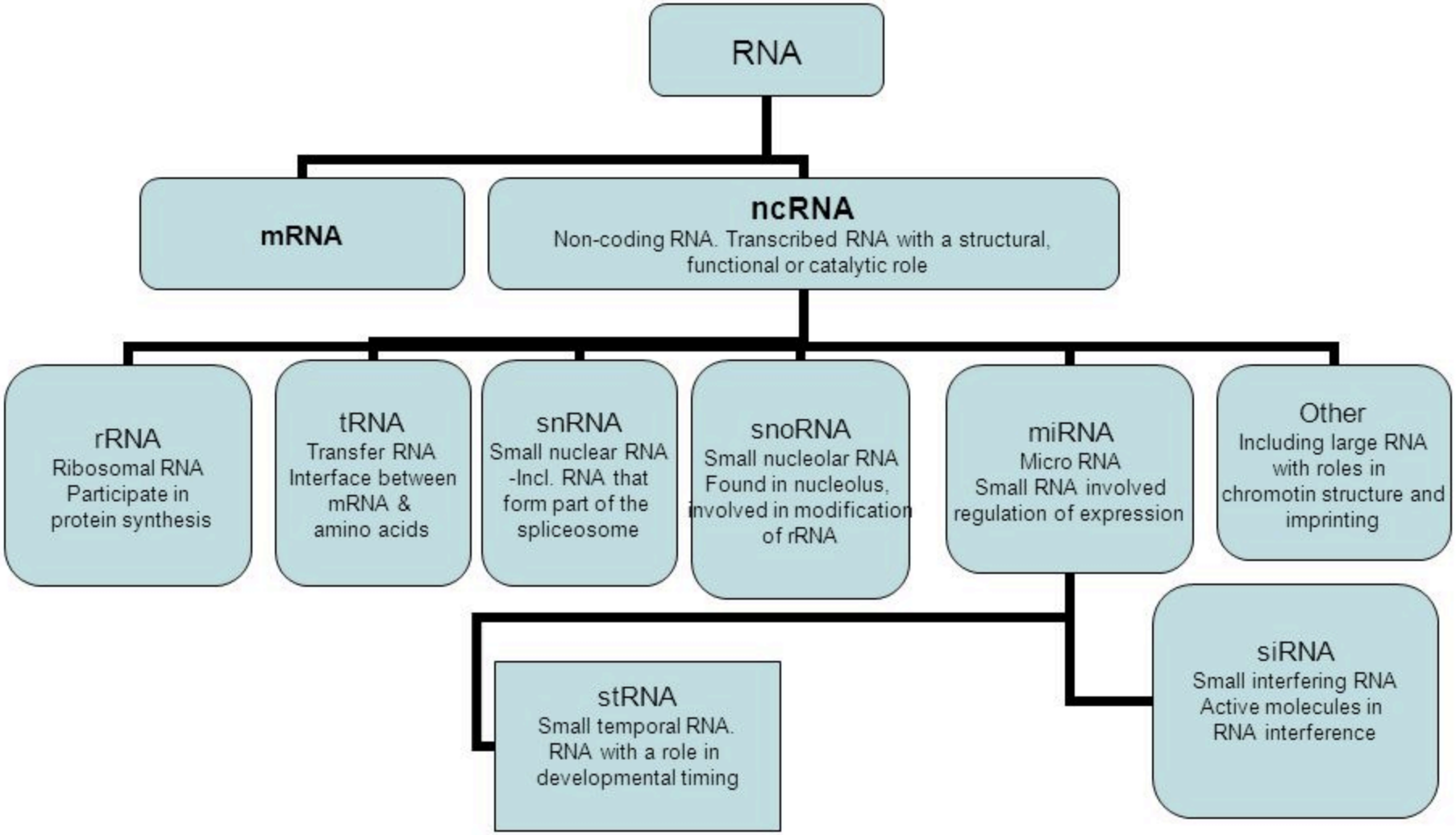
Enabler for Life Sciences

Karolinska Institutet
KTH Royal Institute of Technology
Stockholm University
Uppsala Universitet
SciLifeLab

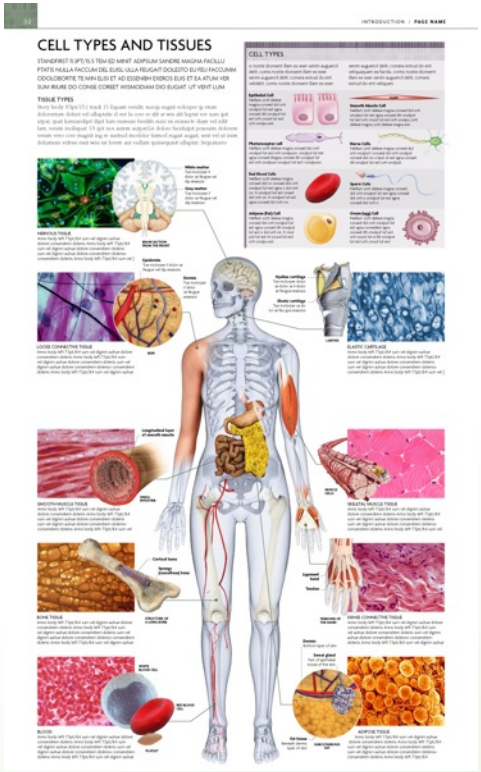
# DNA is the same in all cells but which RNAs that is present  is different in all cells



# There is a wide variety of different functional RNAs



# Which RNAs (and sometimes then translated to proteins) varies between samples




-Tissues

-Cell types

-Cell states

-Individuals

-Cells

# RNA gives information on which genes that are expressed

How DNA get transcribed to RNA (and sometimes then translated to proteins) varies between e. g.

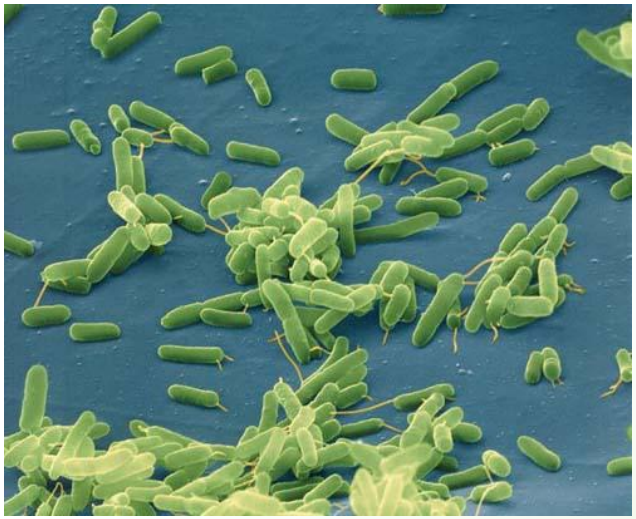
-Tissues

-Cell types

-Cell states

-Individuals

# RNA gives information on which genes that are expressed



How DNA get transcribed to RNA (and sometimes then translated to proteins) varies between e. g.

~~-Tissues~~

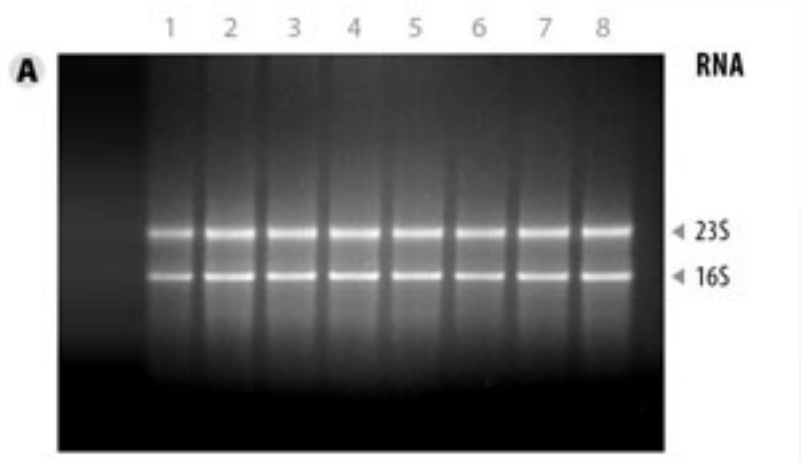
~~-Cell types~~

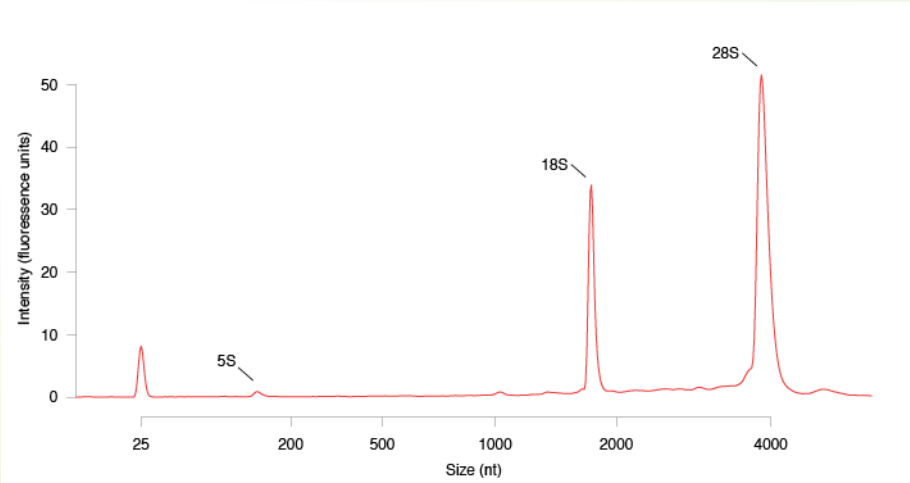
-Cell states

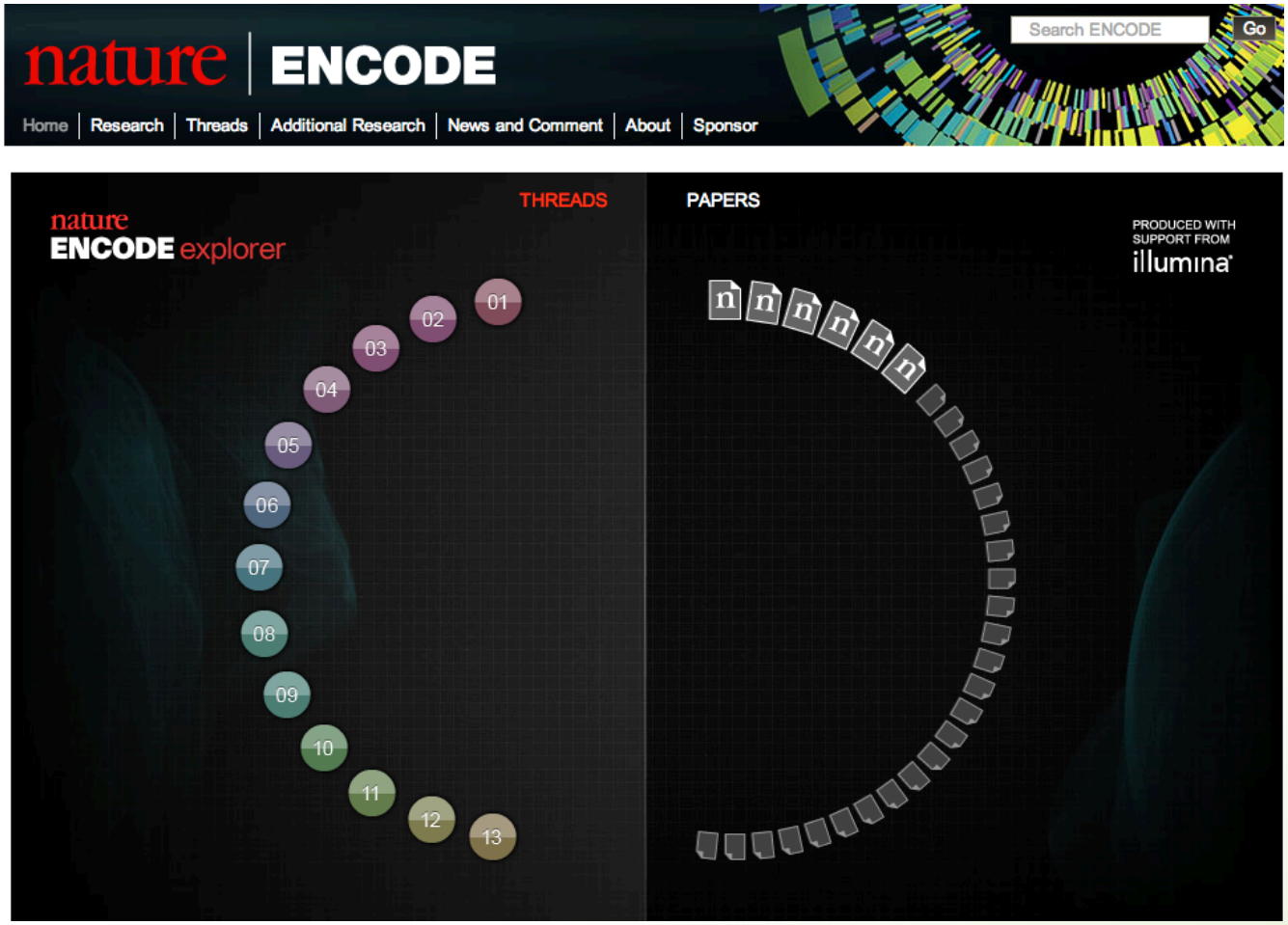
-Individuals

# RNA flavors (pre sequencing era)

- House keeping RNAs
  - rRNAs, tRNAs, snoRNAs, snRNAs, SRP RNAs, catalytic RNAs (RNAse E)
- Protein coding RNAs
  - (1 coding gene ~ 1 mRNA)
- Regulatory RNAs
  - Few rare examples

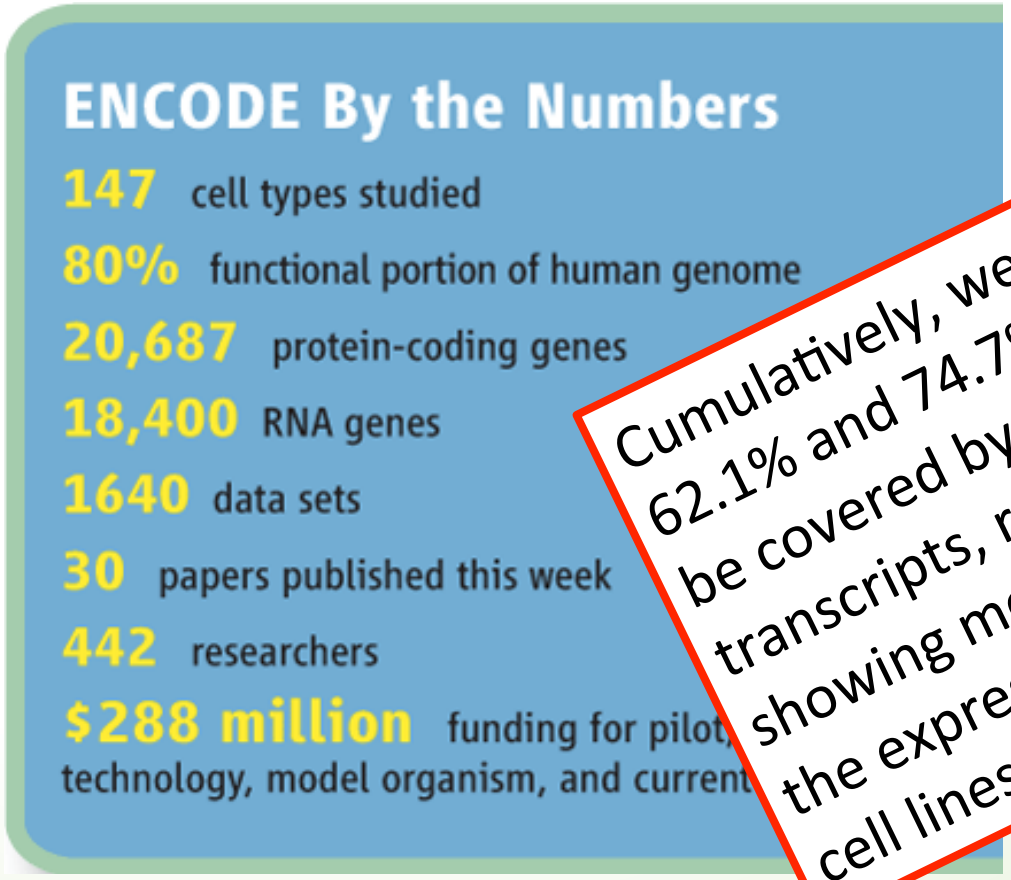





ENCODE, the Encyclopedia of DNA Elements, is a project funded by the National Human Genome Research Institute to identify all regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome sequence.

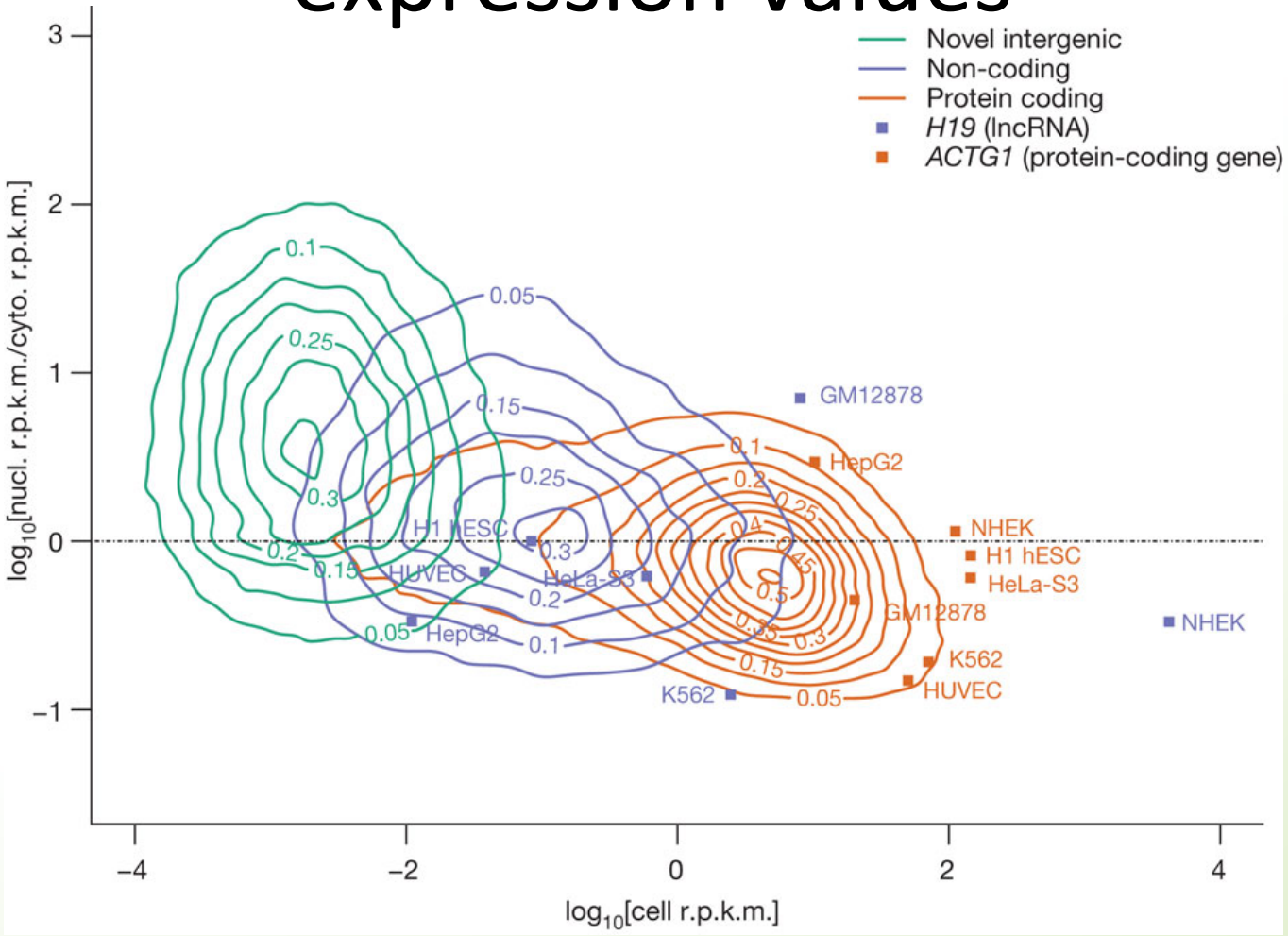
# ENCyclopedia Of Dna Elements




Cumulatively, we observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts, respectively, with no cell line showing more than 56.7% of the union of the expressed transcriptomes across all cell lines.

Karolinska Institutet · KTH Royal Institute of Technology · Stockholm University · Uppsala Universitet · SciLifeLab

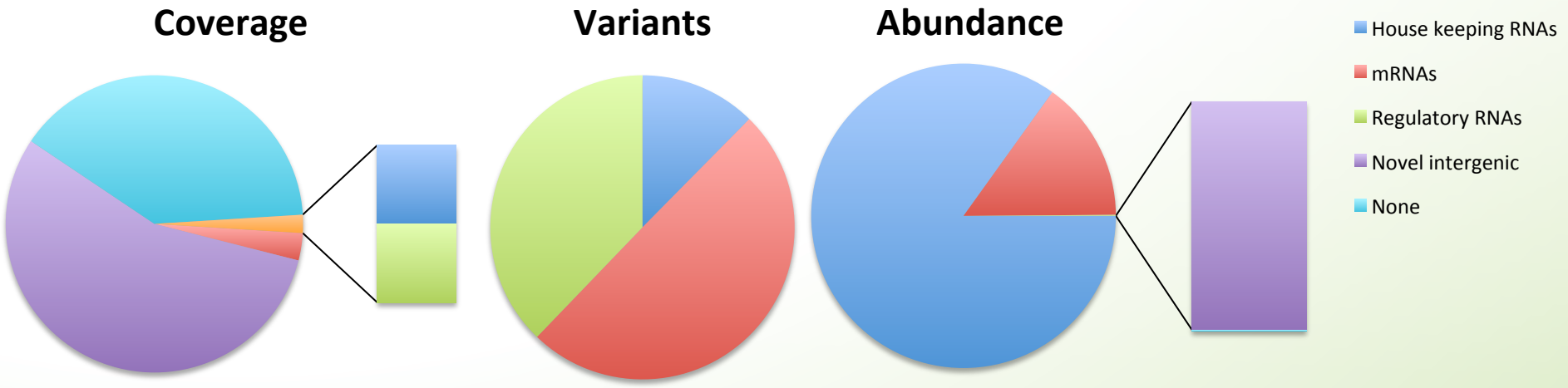
# Different kind of RNAs have different expression values



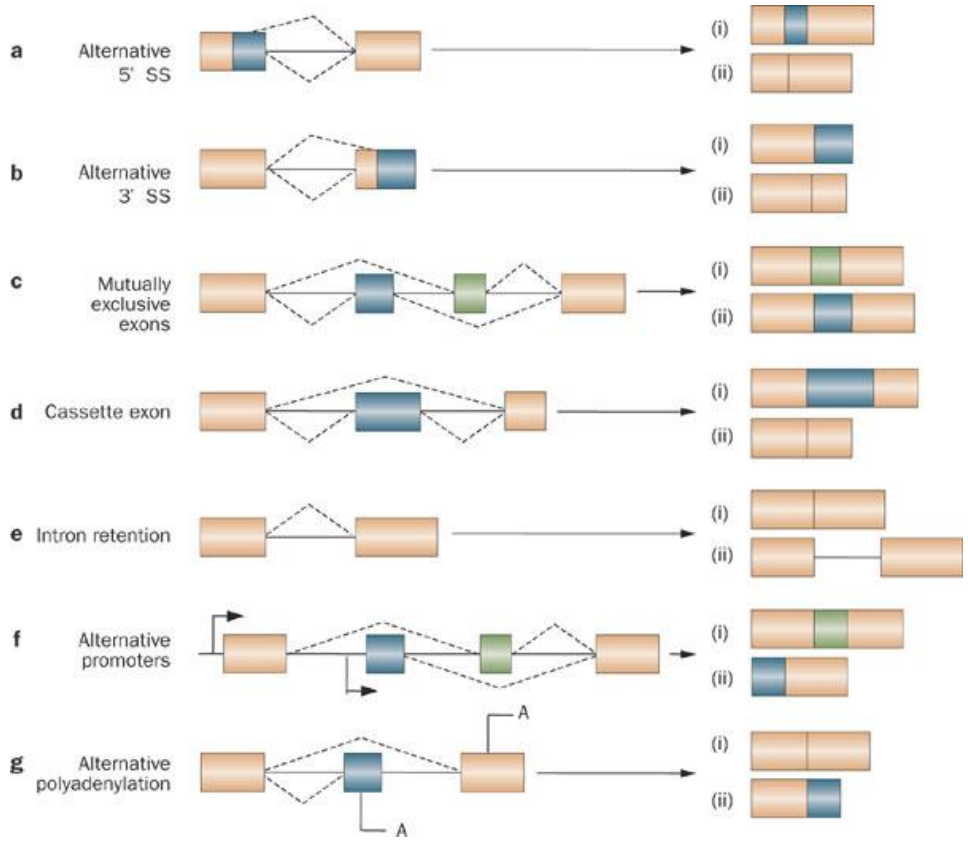



**Landscape of transcription in human cells,** S Djebali *et al.* ***Nature 2012***

# What defines RNA depends on how you look at it



# One gene many different mRNAs



# Defining functional DNA elements in the human genome

- Statement
  - A priori, we should not expect the transcriptome to consist exclusively of functional RNAs.
- Why is that
  - Zero tolerance for errant transcripts would come at high cost in the proofreading machinery needed to perfectly gate RNA polymerase and splicing activities, or to instantly eliminate spurious transcripts.
  - In general, sequences encoding RNAs transcribed by noisy transcriptional machinery are expected to be less constrained, which is consistent with data shown here for very low abundance RNA
- Consequence
  - Thus, one should have high confidence that the subset of the genome with large signals for RNA or chromatin signatures coupled with strong conservation is functional and will be supported by appropriate genetic tests.
  - In contrast, the larger proportion of genome with reproducible but low biochemical signal strength and less evolutionary conservation is challenging to parse between specific functions and biological noise.

# This is of course not without an debate

OPEN ACCESS Freely available online

PLoS BIOLOGY

## Most "Dark Matter" Transcripts Are Associated With K

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

page | discussion | view source | history

### Comments on van Bakel et al. (2011) Response to "The Reality of Pervasive Transcription"

Comments by Mike Clark

Van Bakel et al. 2011 (vB 11) have published their reply to our critique of their paper van Bakel et al. 2010 (vB 10).

Firstly lets briefly review some of our main criticisms of vB 10:

1. vB 10 didn't properly consider previous evidence for pervasive transcription (especially that from cDNA analysis in the mouse) when claiming the genome was not as tra previous evidence was unreliable due to false positives.

2. vB 10 incorrectly conflated pervasive transcription with the relative abundance of transcripts when the correct (and known) definition was the amount of the genome that

3. The tiling arrays vB 10 performed and then used to claim that previous array studies suffered from high false positives were atypical and lacked any validation of the fal

4. The RNA sequencing carried out by vB 10 was severely limited in its ability to address the question of pervasive transcription. The depth of sequencing was too shallow complex samples and then the assembly of what was found into transcripts was poor. Since it couldn't detect and/or characterize rare transcripts this meant it couldn't eva

5. vB 10 claimed that low level intergenic transcription may be due to "random initiation events" and/or transcriptional "byproducts" (ie: transcription noise), when the limitat differentiate properly between this and genuine transcripts under their detection threshold.

mattick

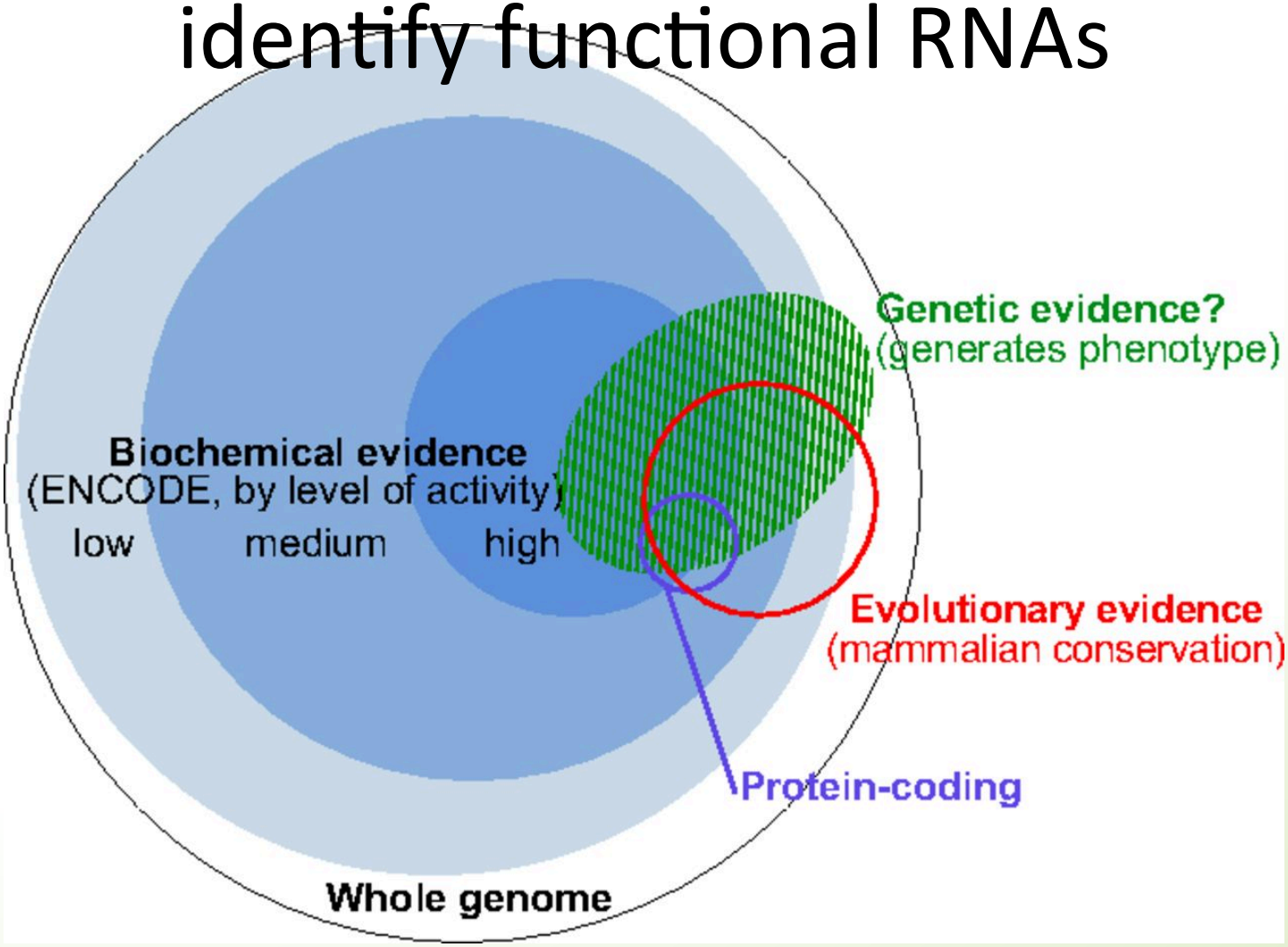
Navigation
- Home
- Research
- People
- Publications
- Links
- Internal Home
- Contact

Links
- QuARC
- NRED
- lncRNAdb

Novel intergenic
(protein-coding gene)

Karolinska Institutet | KTH Royal Institute of Technology | Stockholm University | Uppsala Universitet | SciLifeLab

# Biochemical evidence not enough to identify functional RNAs

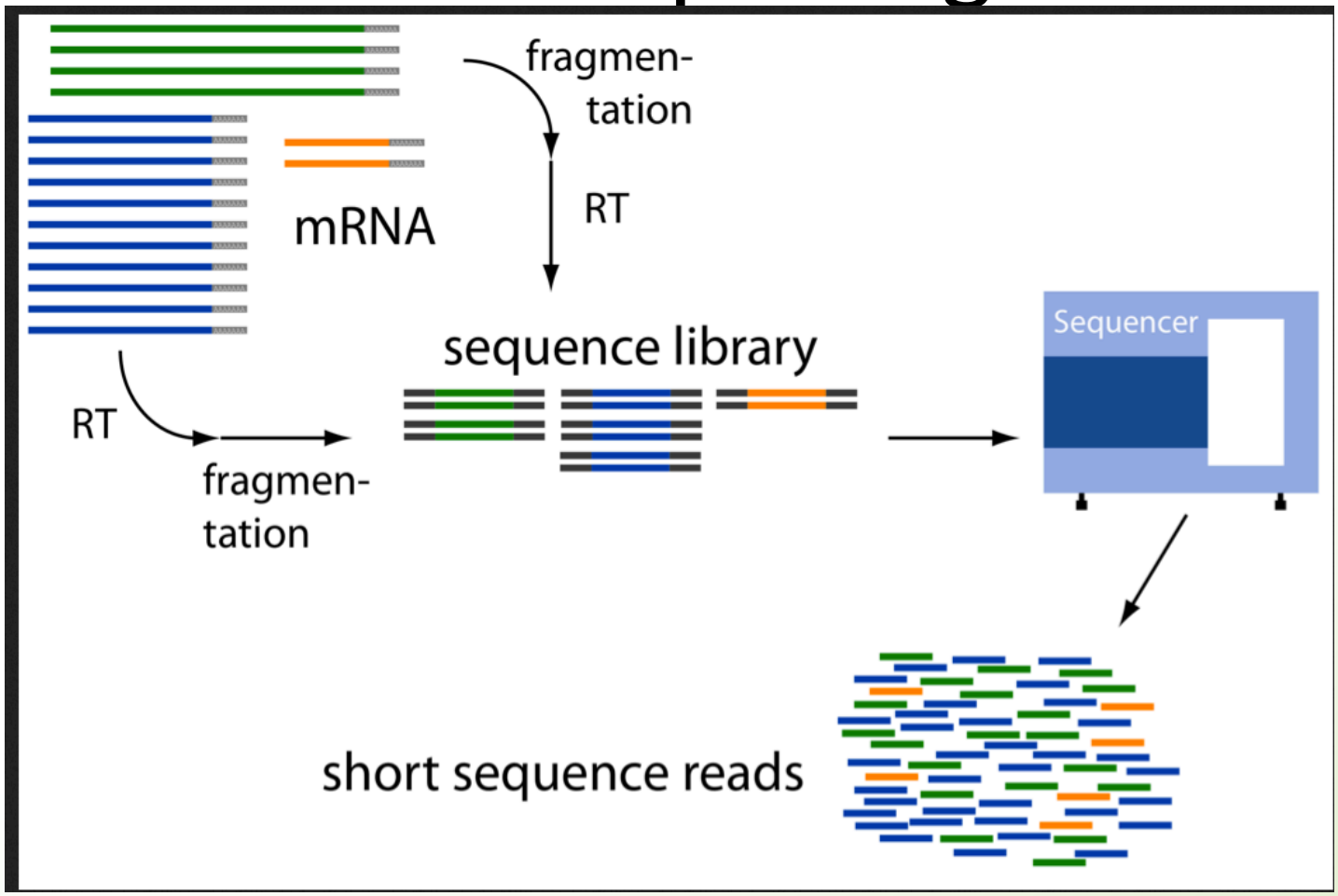


- RNA seq course

# The RNA seq course

- From RNA seq to reads (Introduction)
- Mapping reads programs (Tuesday)
- Transcriptome reconstruction using reference (Tuesday)
- Transcriptome reconstruction without reference (Tuesday)
- QC analysis (Wednesday)
- Differential expression analysis (Wednesday)
- Gene set analysis (Wednesday)
- miRNA analysis (Thursday)
- Allele specific analysis (Thursday)
- Single cell analysis  (Thursday)

# How are RNA-seq data generated?




**Sampling process**

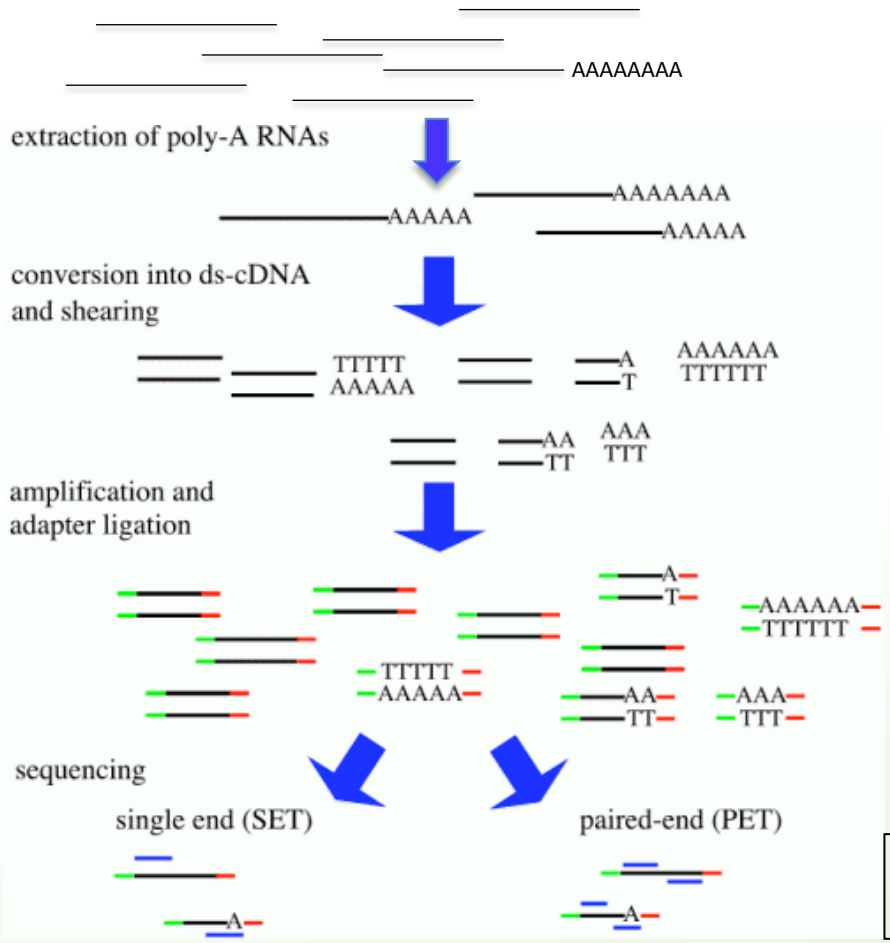
# Depending on the different steps you will get different results

RNA->

enrichments ->

library ->

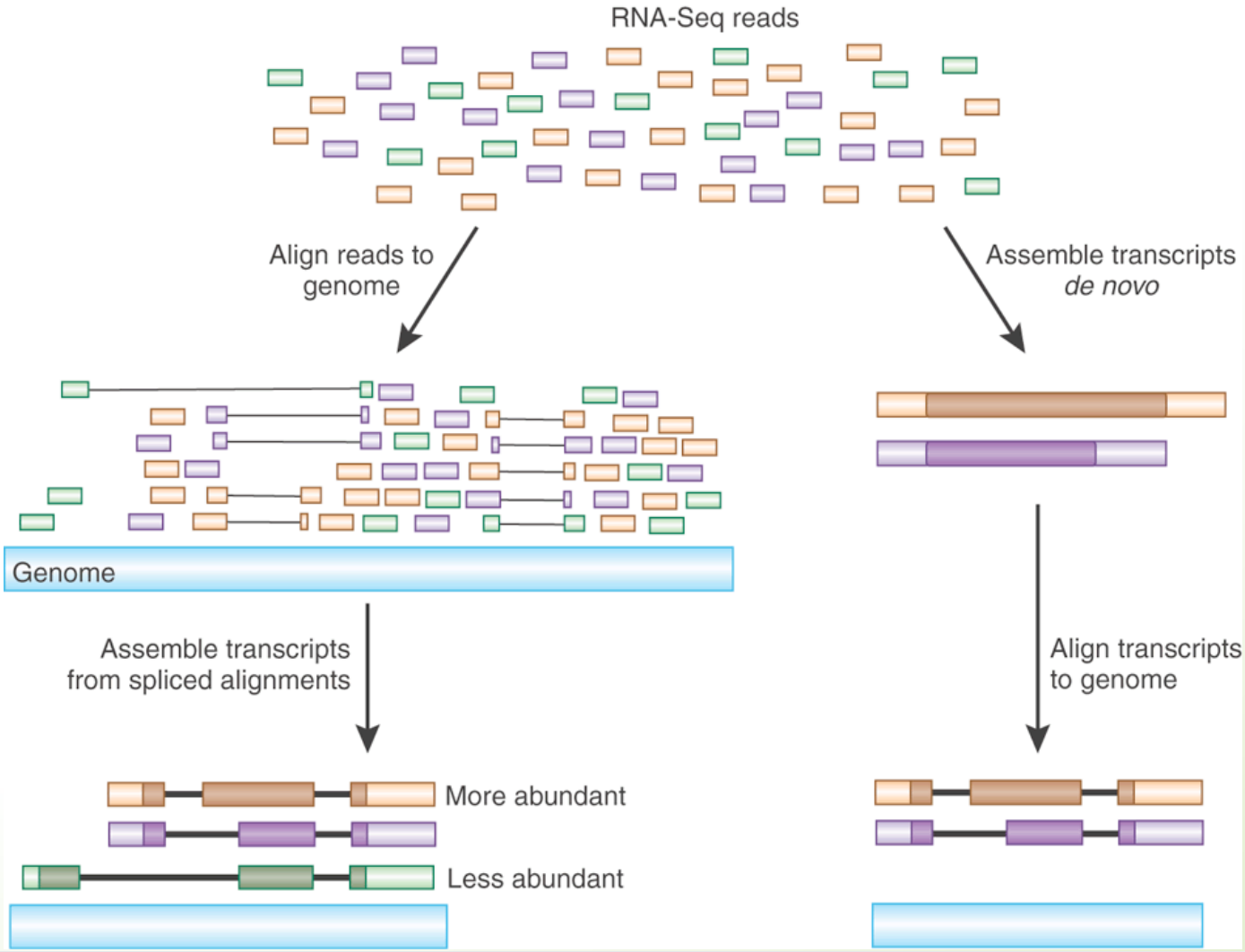
reads ->




| | |
|---|---|
| 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)


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

# Promises and pitfalls

## Long reads

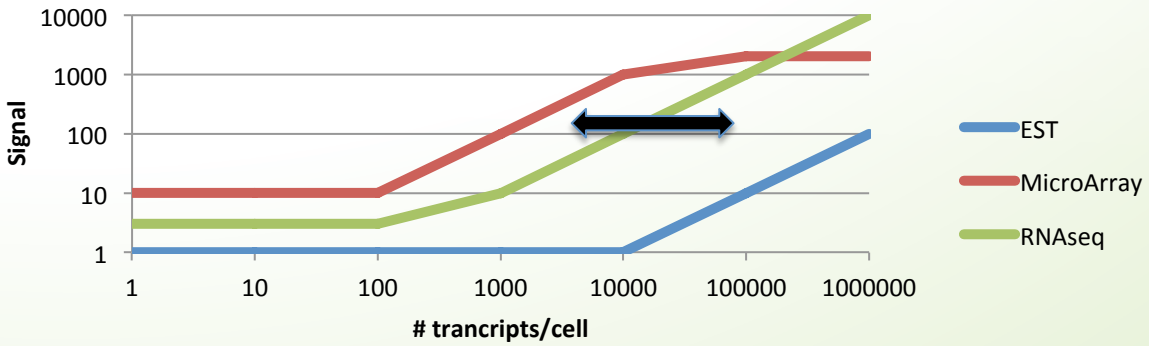
- Low throughput (-)
- Complete transcripts (+)
- Only highly expressed genes (--)
- Expensive (-)
- Low background noise (+)
- Easy downstream analysis (+)

## Micro Arrays

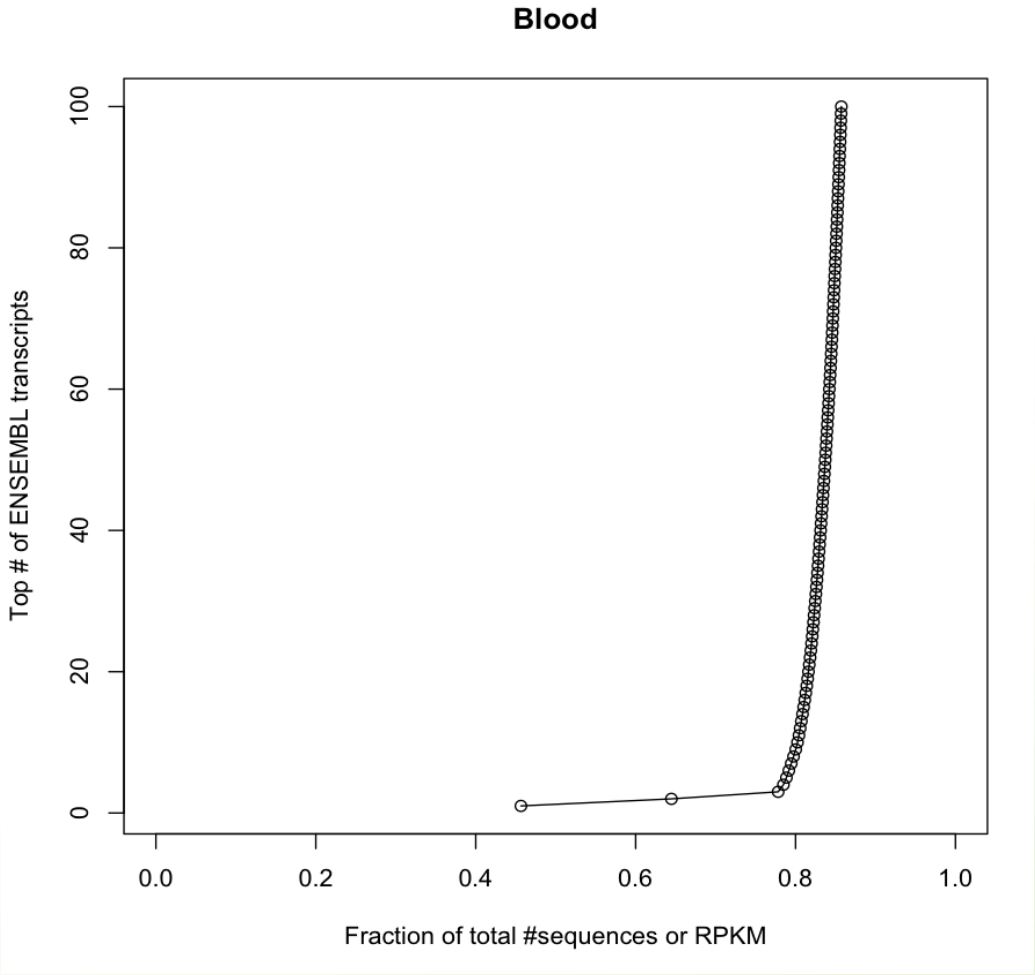
- High throughput (+)
- Only known sequences (-)
- Limited dynamic range (-)
- Cheap (+)
- High background noise (-)
- Not strand specific (-)
- Well established downstream methods (+)

## short reads

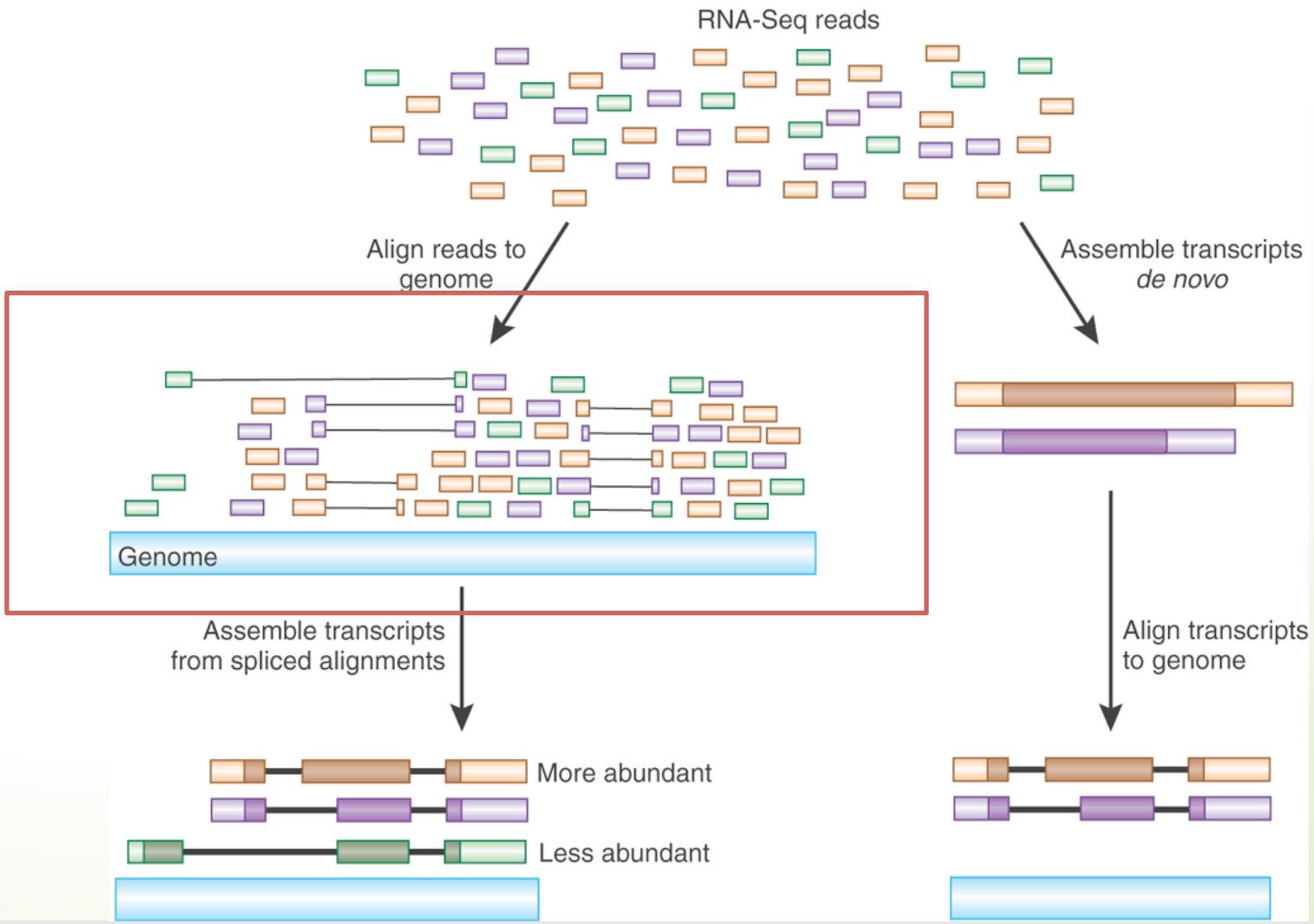
- High throughput (+)
- Fractions of transcripts (-)
- Full dynamic range (+-)
- Unlimited dynamic range (+)
- Cheap (+)
- Low background noise (+)
- Strand specificity (+)
- Re-sequencing (+)



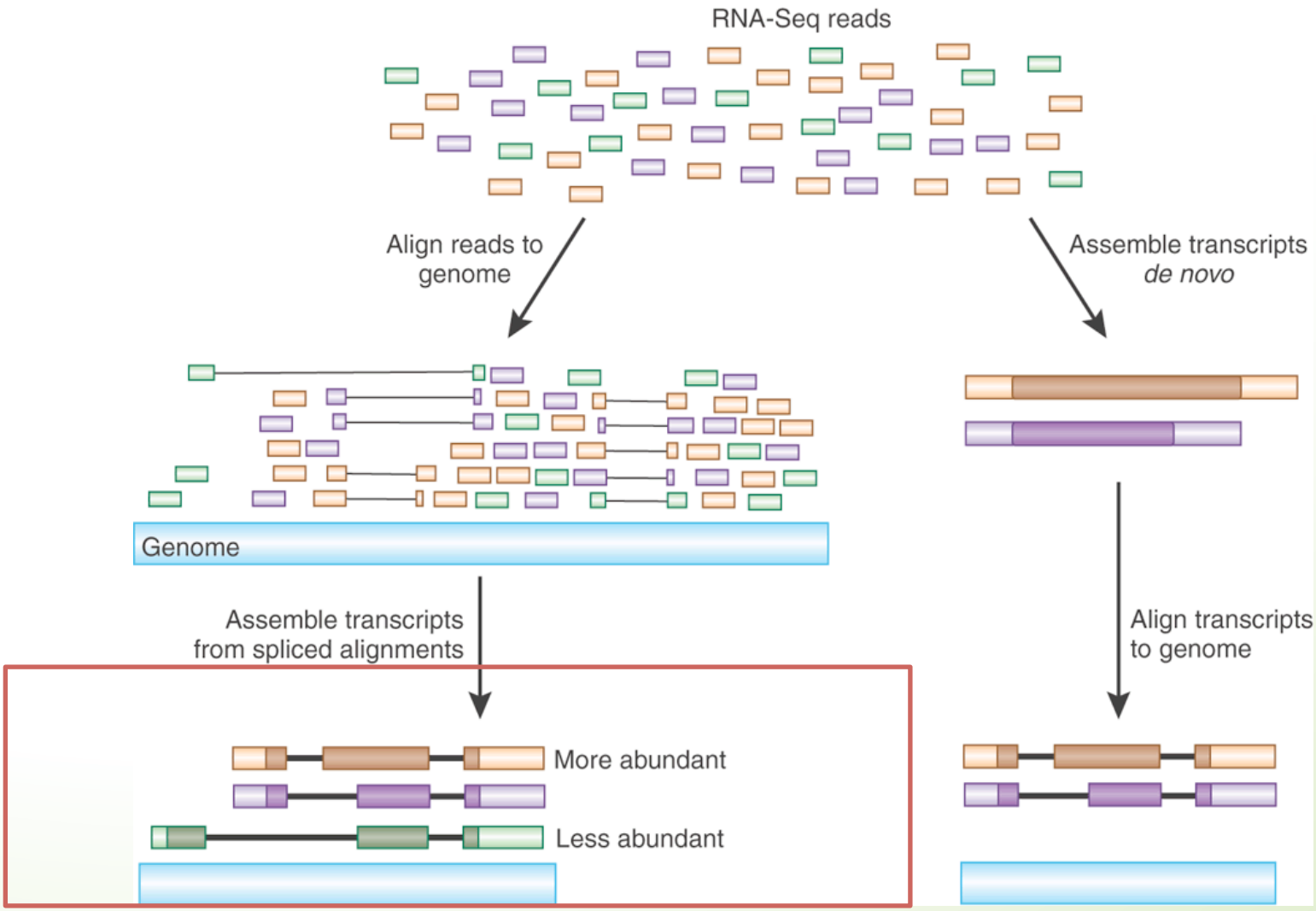
# RNA seq reads correspond directly to abundance of RNAs in the sample



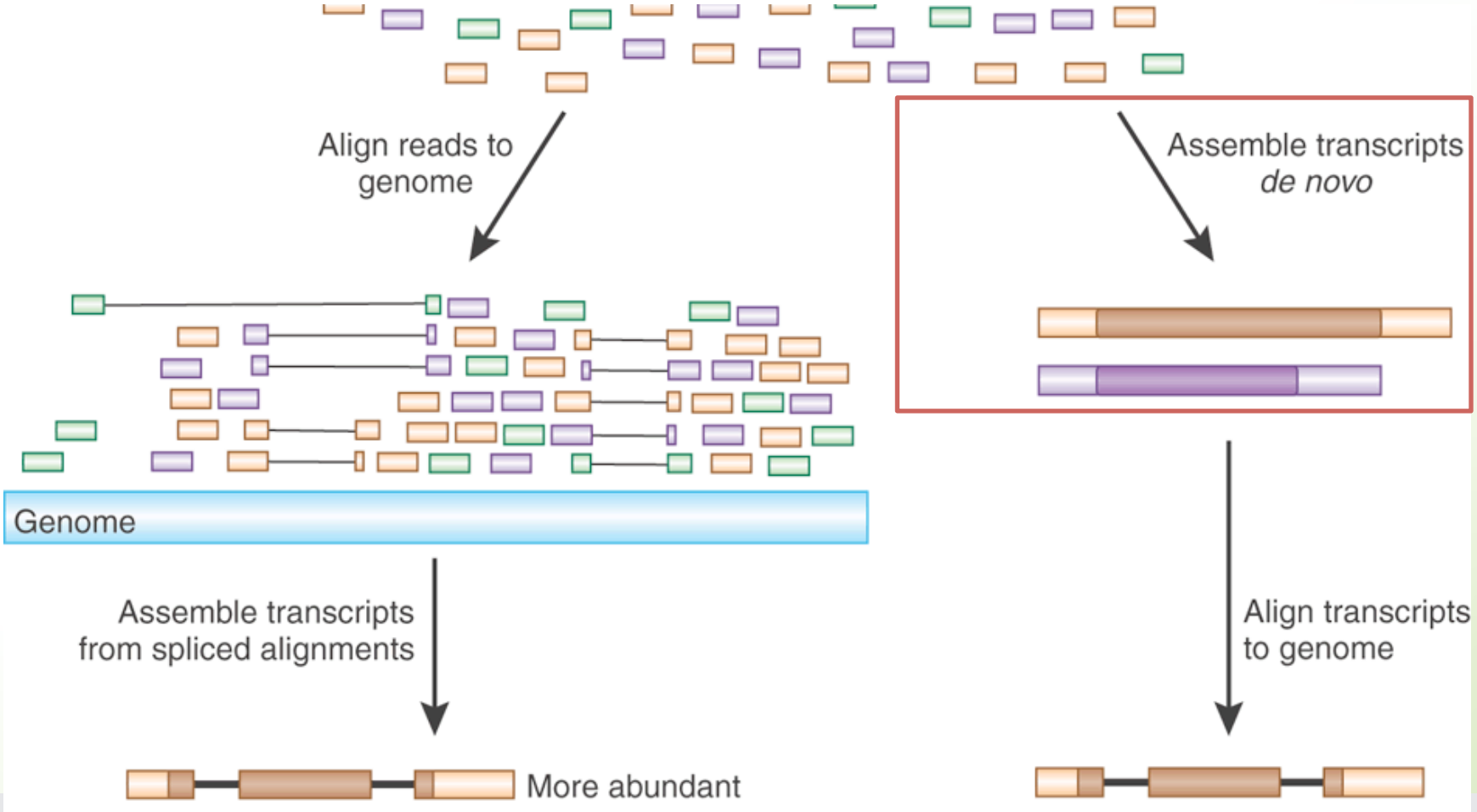
# Map reads to reference

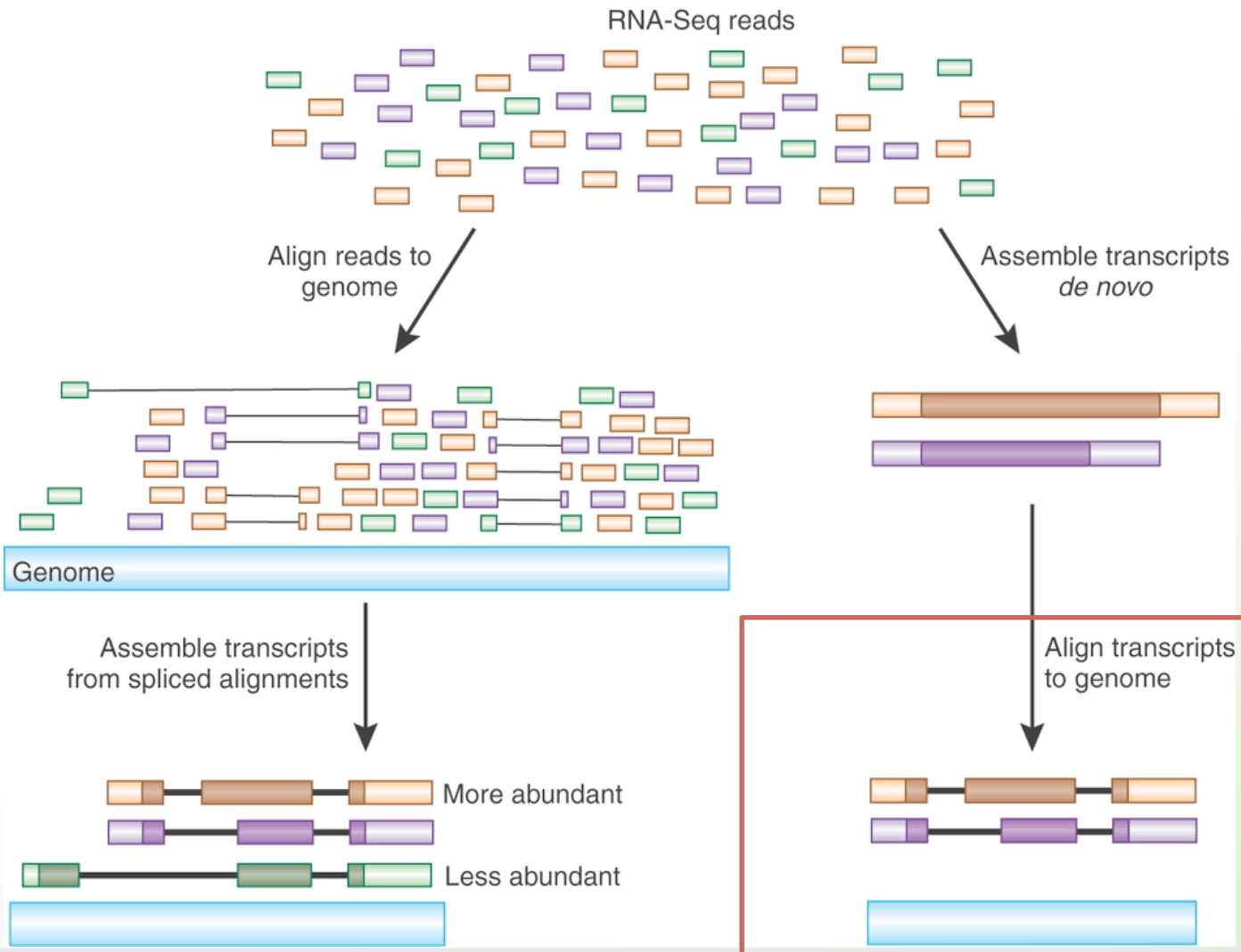


# Transcriptome assembly using reference



# Transcriptome assembly without reference




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

# Quality control

## -samples might not be what you think they are

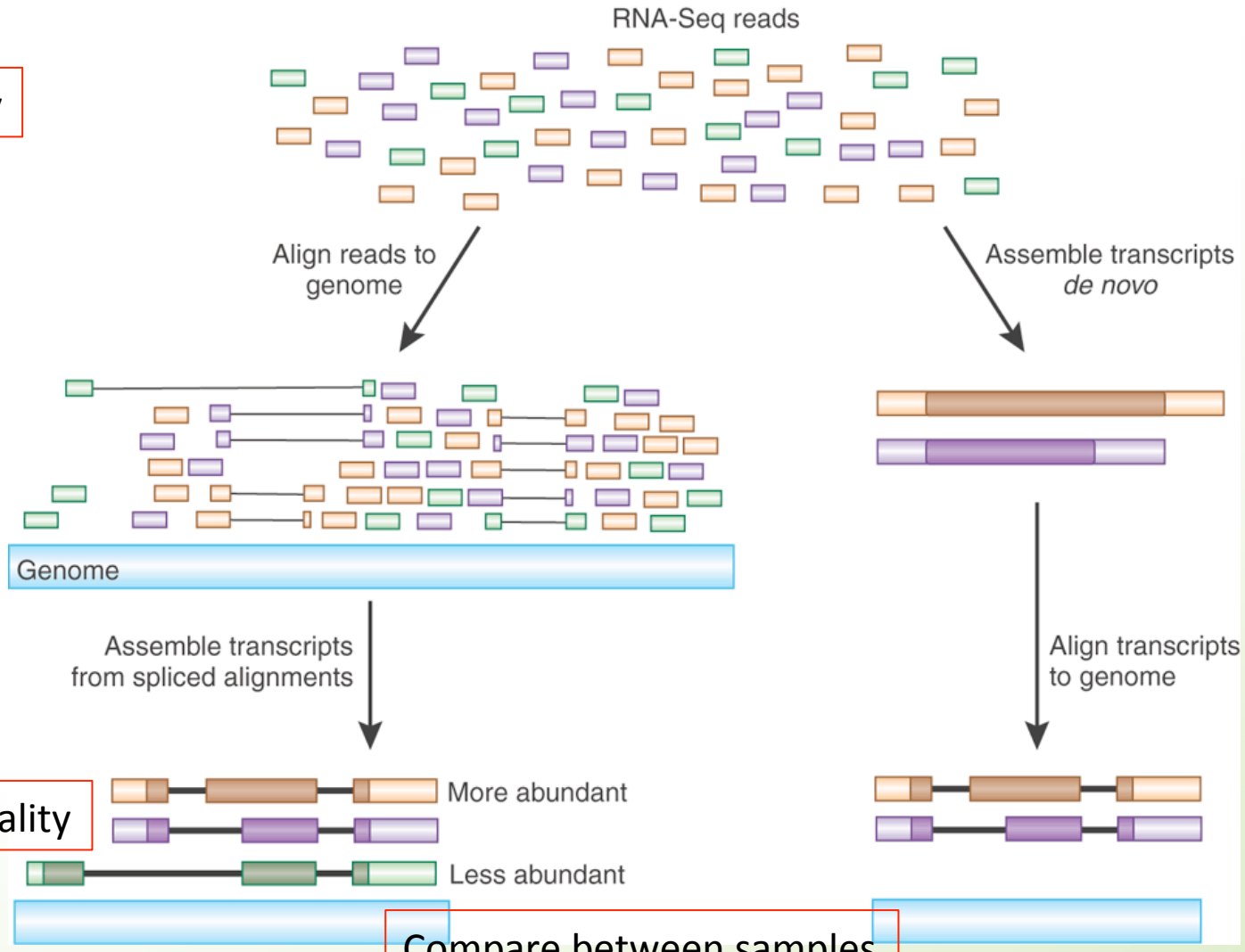
- Experiments go wrong
  - 30 samples with 5 steps from samples to reads has 150 potential steps for errors
  - Error rate 1/100 with 5 steps suggest that one of every 20 samples the reads does not represent the sample
- Mixing samples
  - 30 samples with 5 steps from samples to reads has ~24M potential mix ups of samples
  - Error rate 1/ 100 with 5 steps suggest that one of every 20 sample is mislabeled
- Combine the two steps and approximately one of every 10 samples are wrong

# RNA QC

Read quality

Mapping statistics

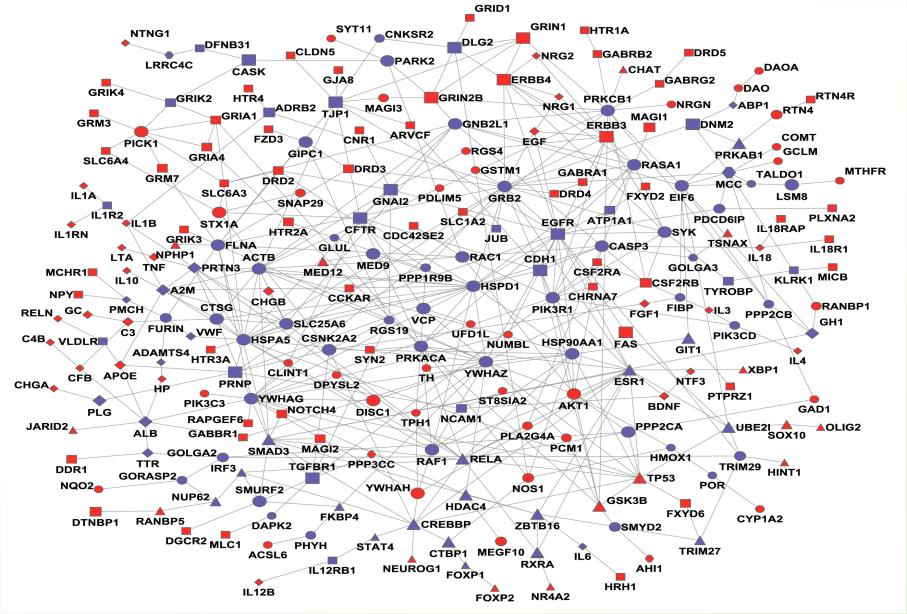
Transcript quality



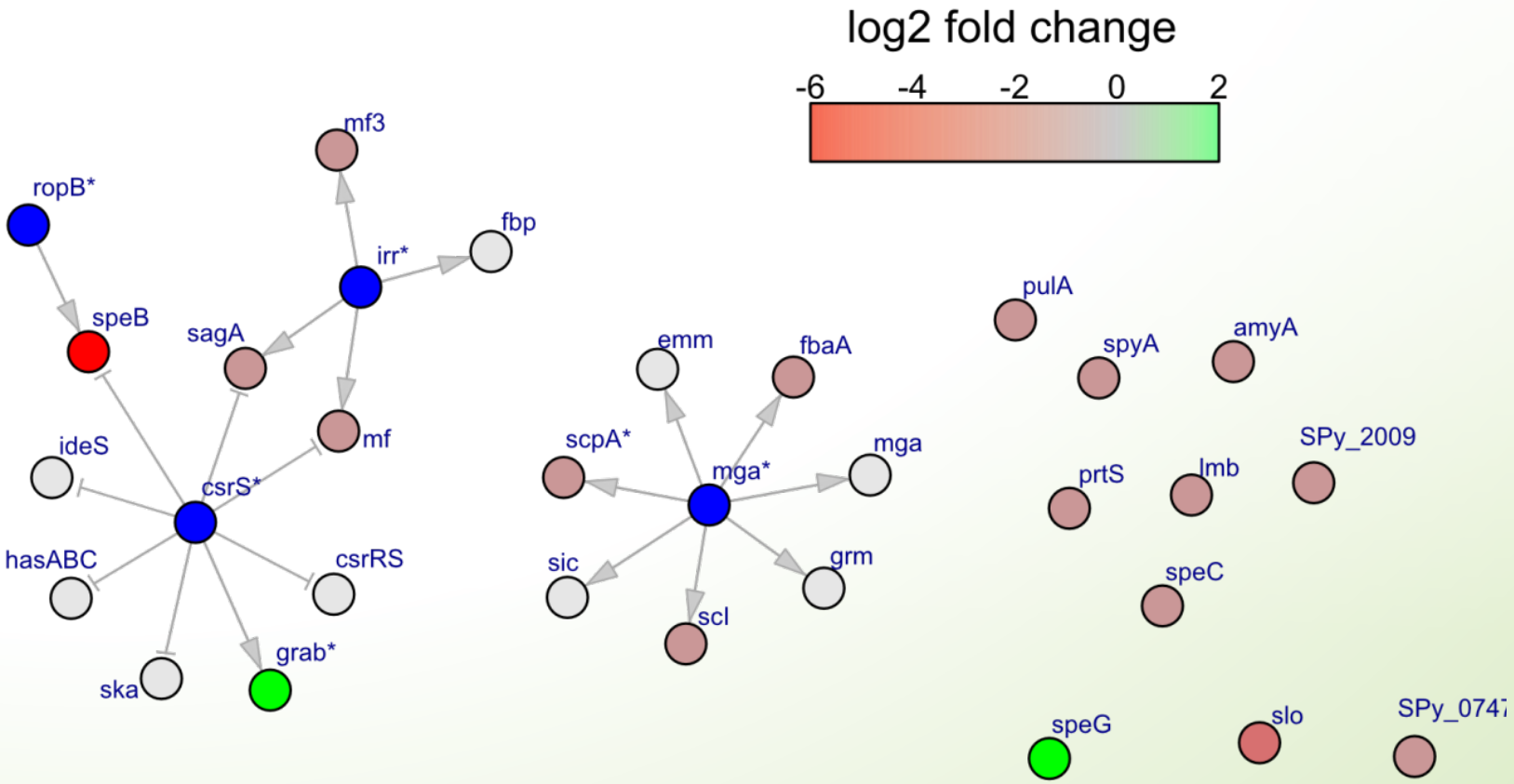

Compare between samples

# Differential expression analysis using univariate analysis

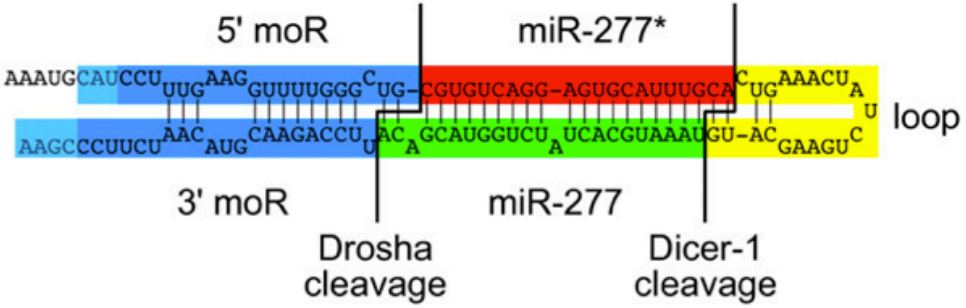
Typically **univariate** analysis (one gene at a time) – even though we know that genes are not independent

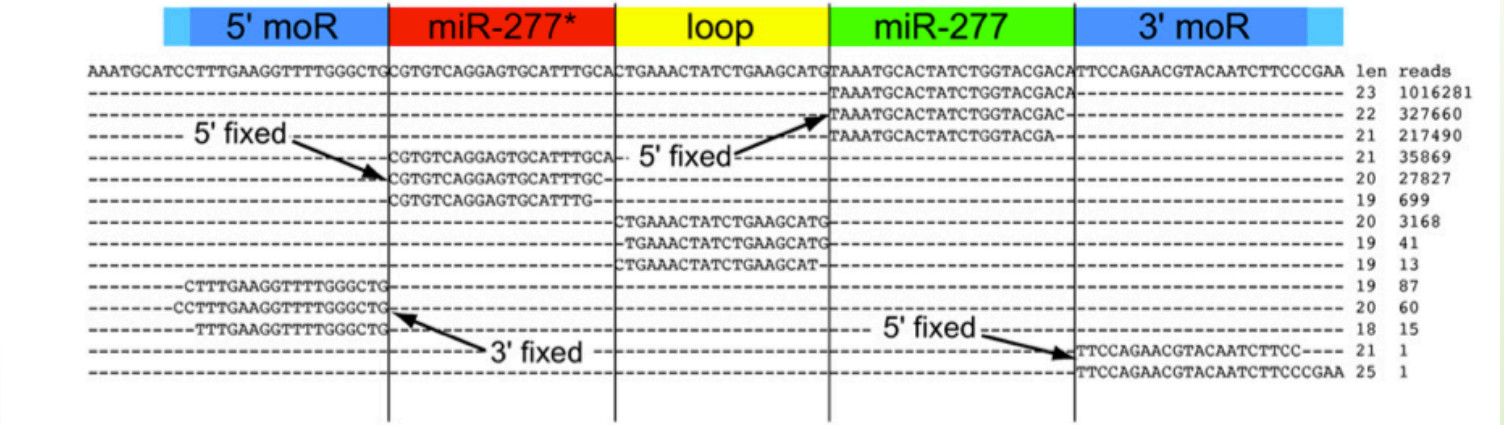


# Gene set analysis and data integration



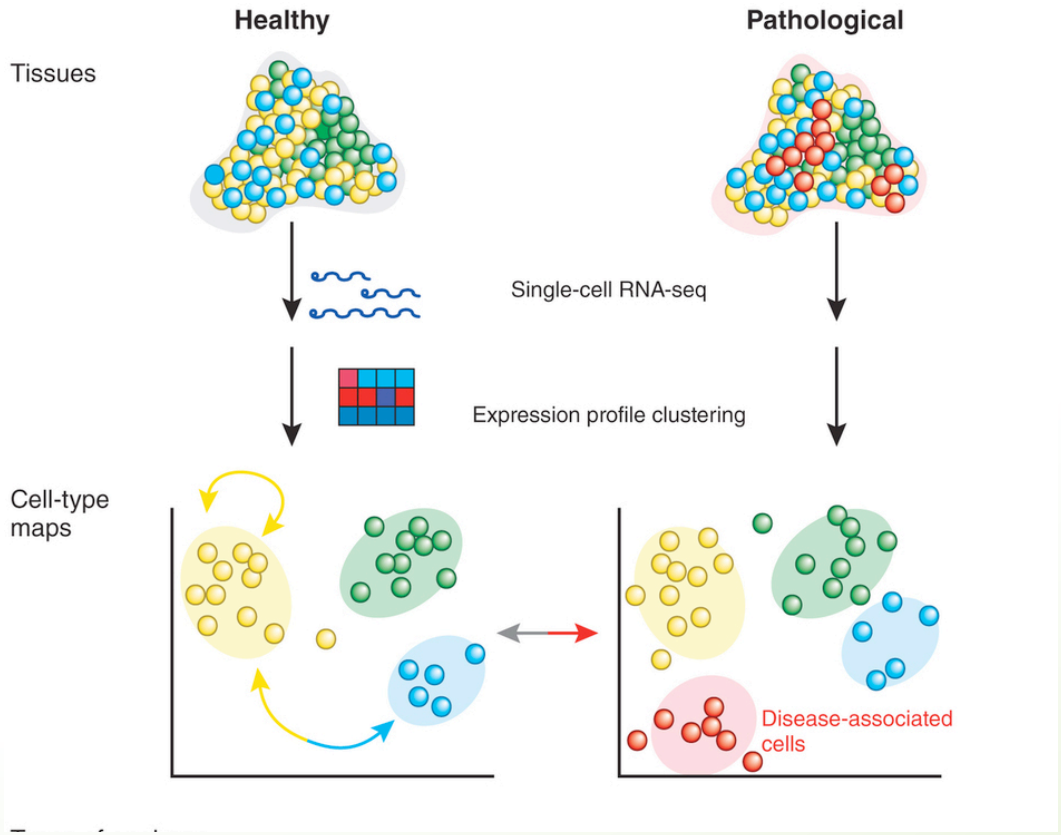
# microRNA analysis (Johan)




(Berezikov et al. Genome Research, 2011.)

# Single cell RNA-seq analysis




(Sandberg, Nature Methods 2014)