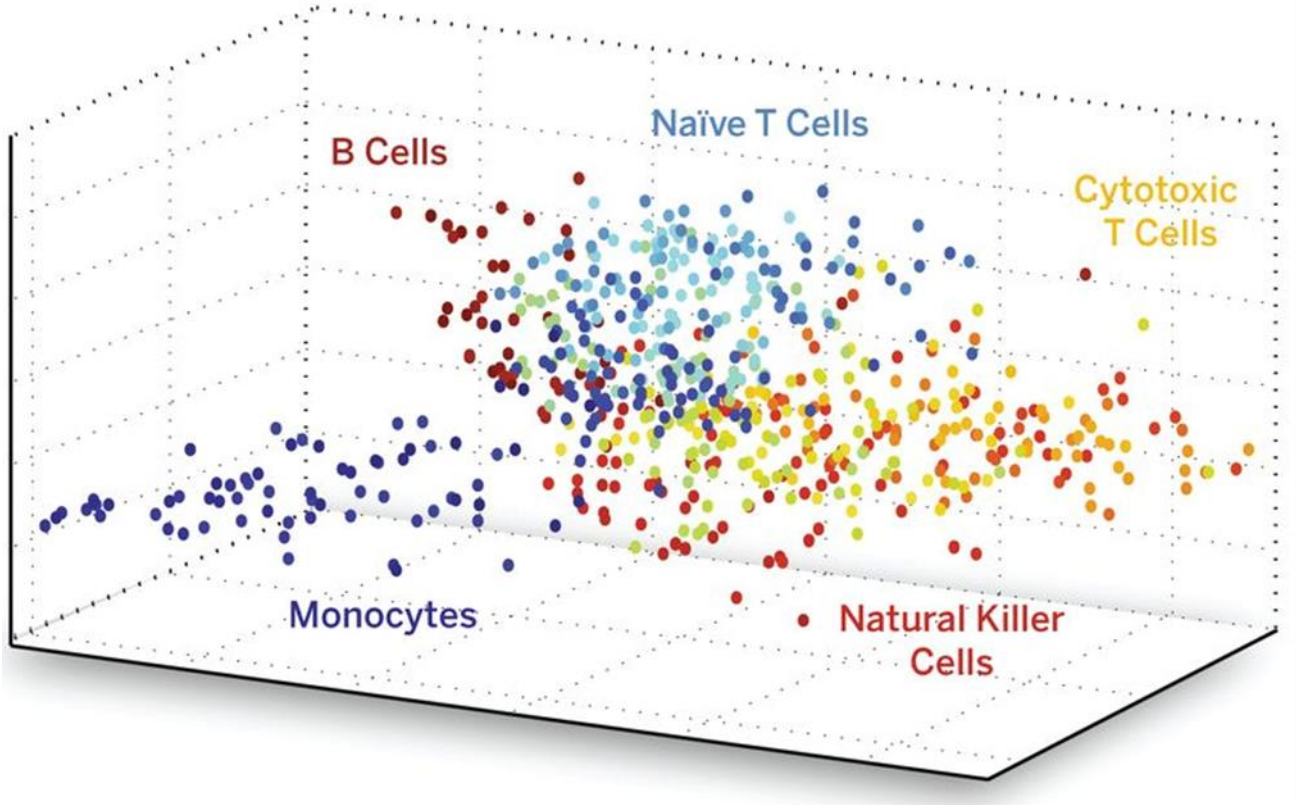
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

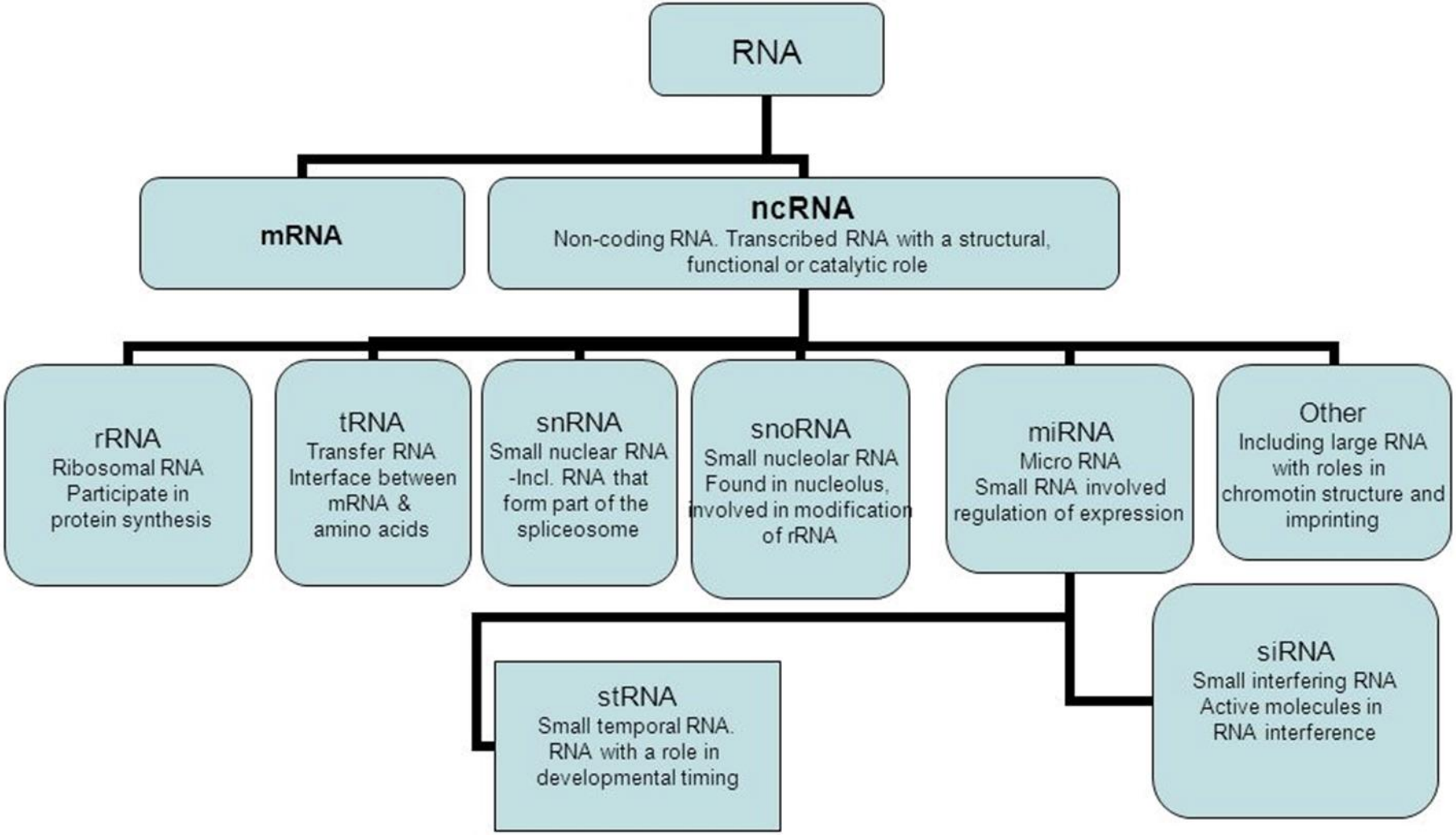
RNAseq Introduction

DNA is the same in all cells…

… but which RNAs that are present is different

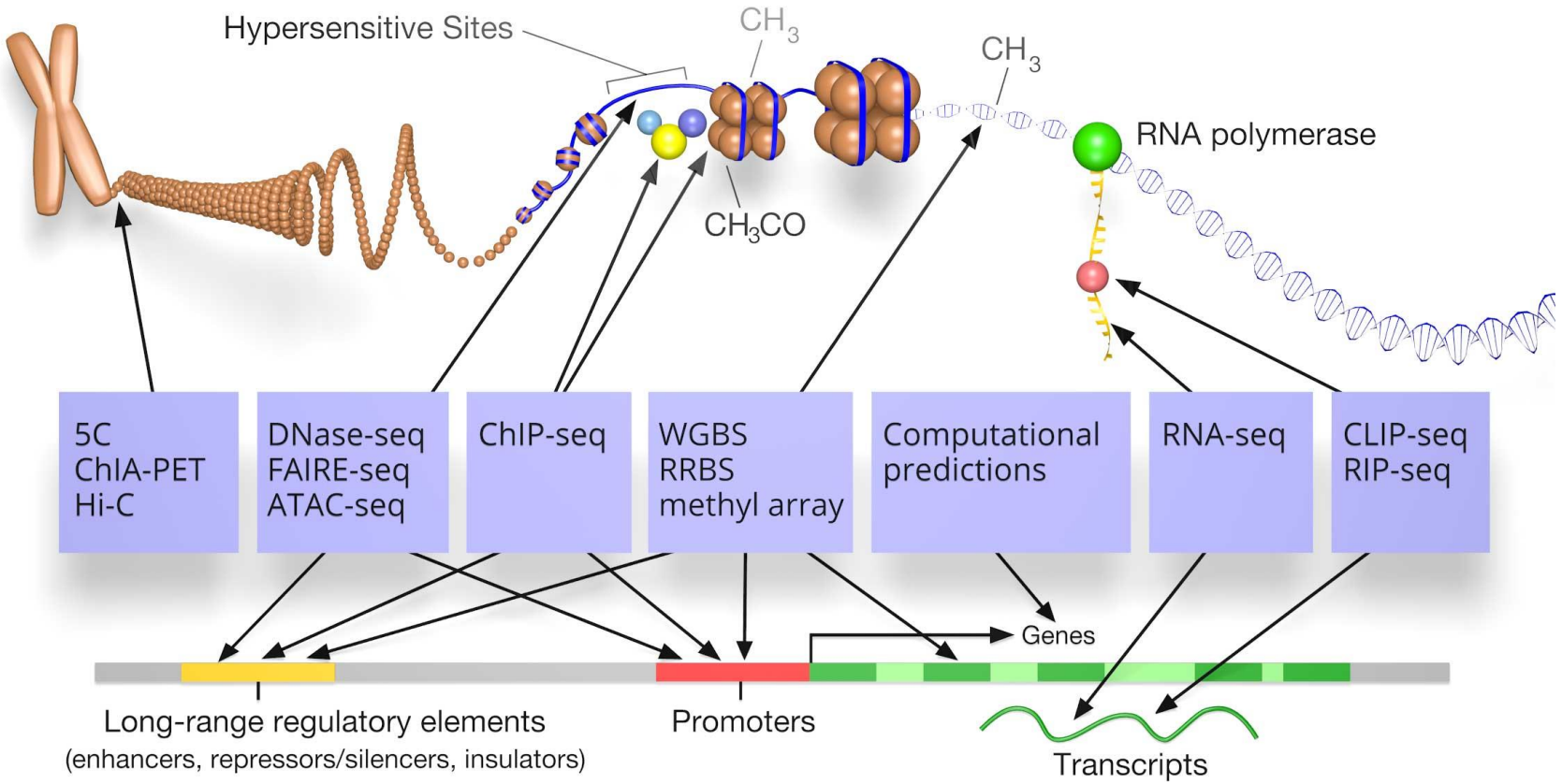


A wide variety of functional RNAs



ENCODE

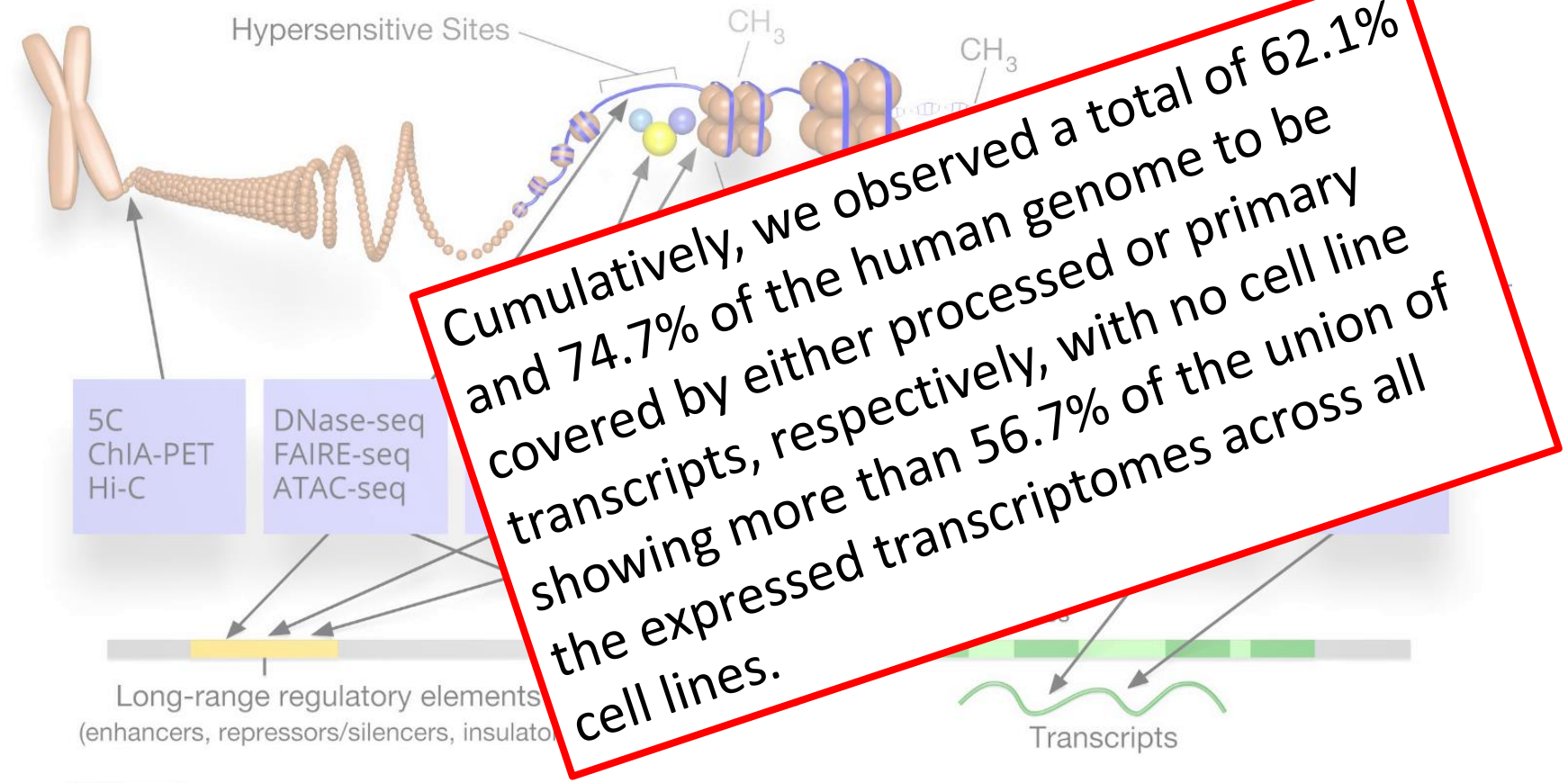
Encyclopedia **O**f **D**NA **E**lements - identify all functional elements in the human and mouse genomes.



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

ENCODE

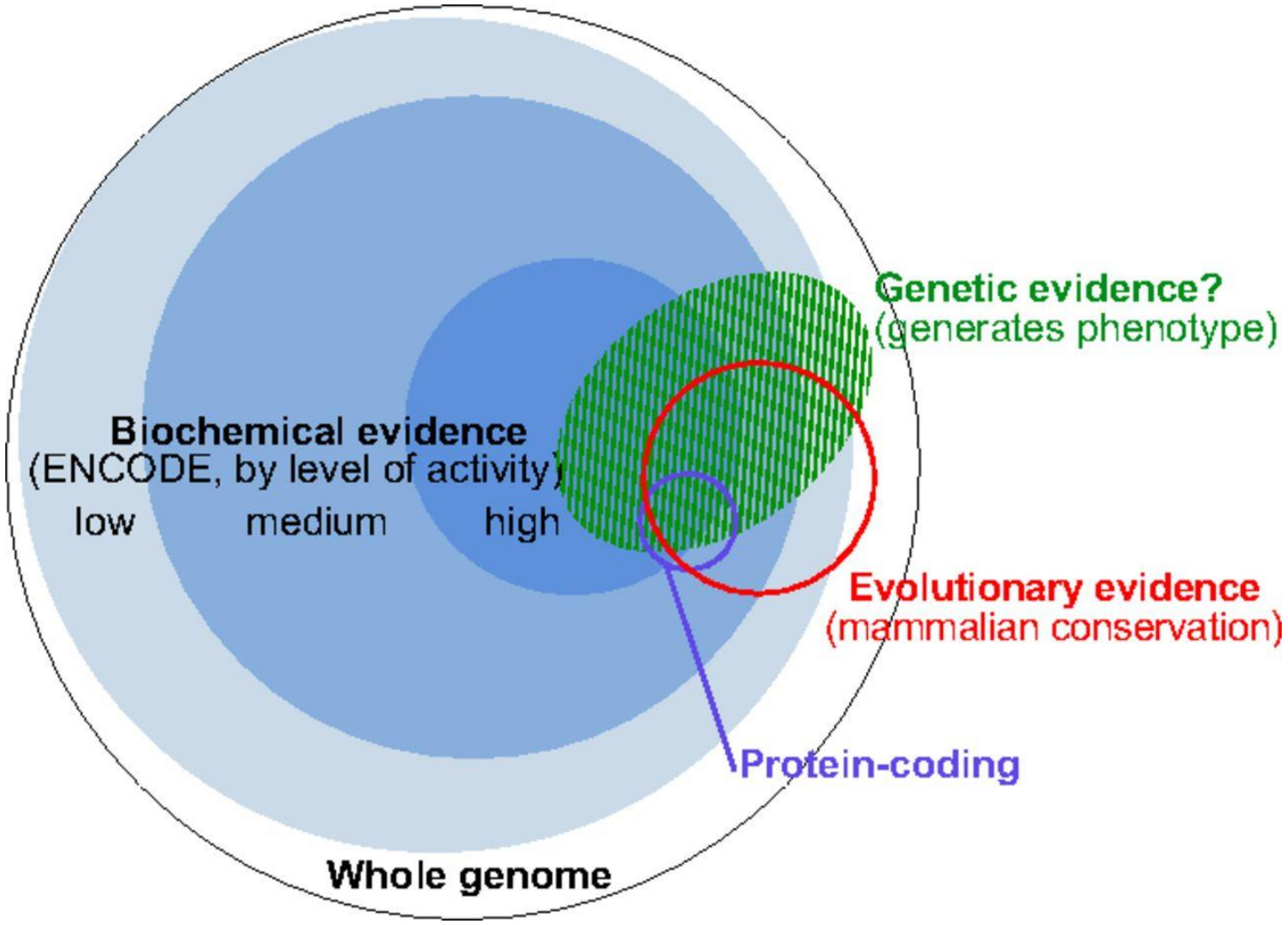
Encyclopedia **O**f **D**NA **E**lements - identify all functional elements in the human and mouse genomes.



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.

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Biochemical evidence not enough to identify functional RNAs

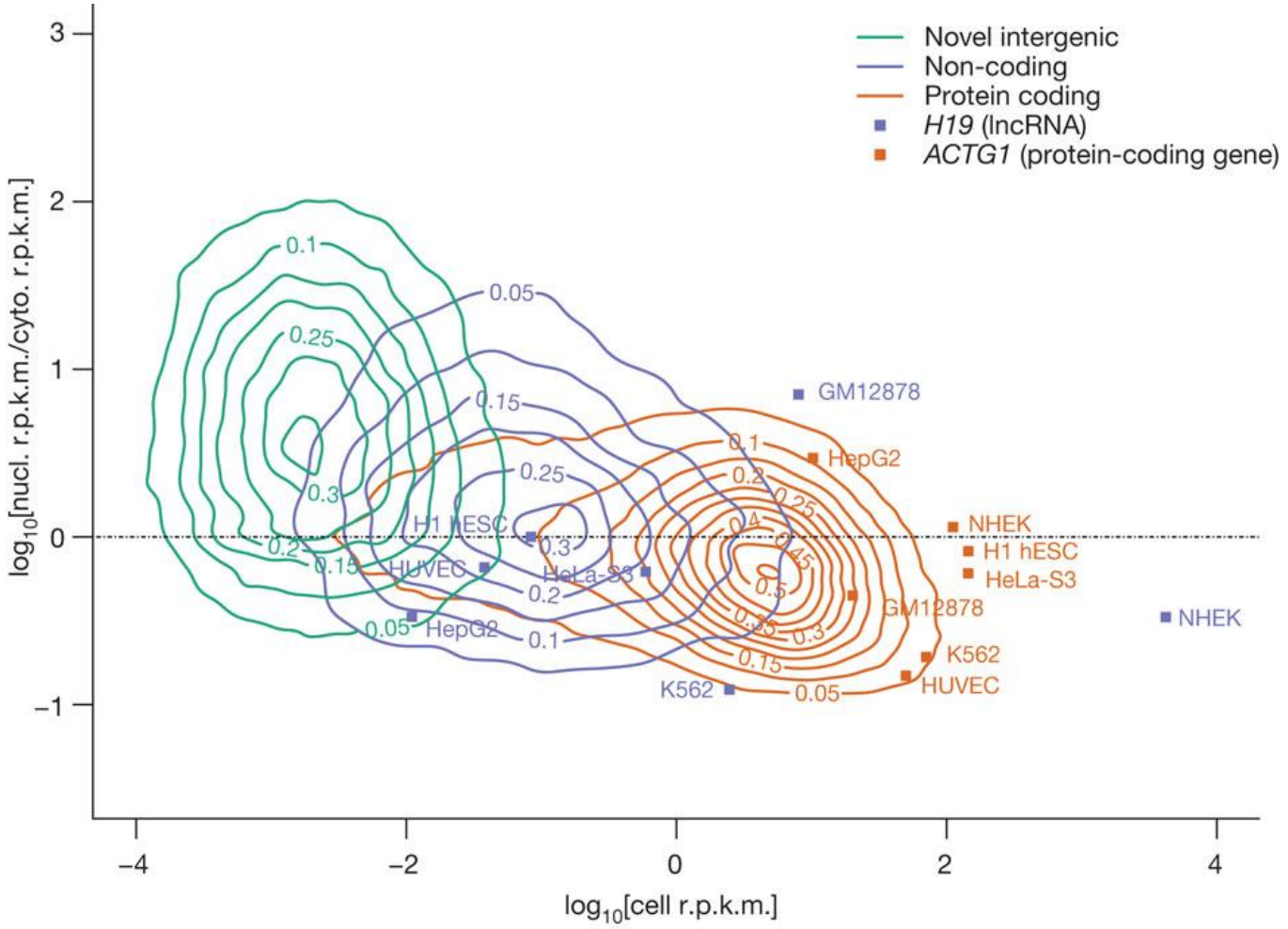


Defining functional DNA elements in the human genome
Kellis M et al. PNAS 2014;111:6131-6138

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 (perfectly gate RNA polymerase and splicing activities, instantly eliminate spurious transcripts)
  - In general, sequences encoding RNAs transcribed by noisy transcriptional machinery are expected to be less constrained, which is also shown by data
- Consequence
  - 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.

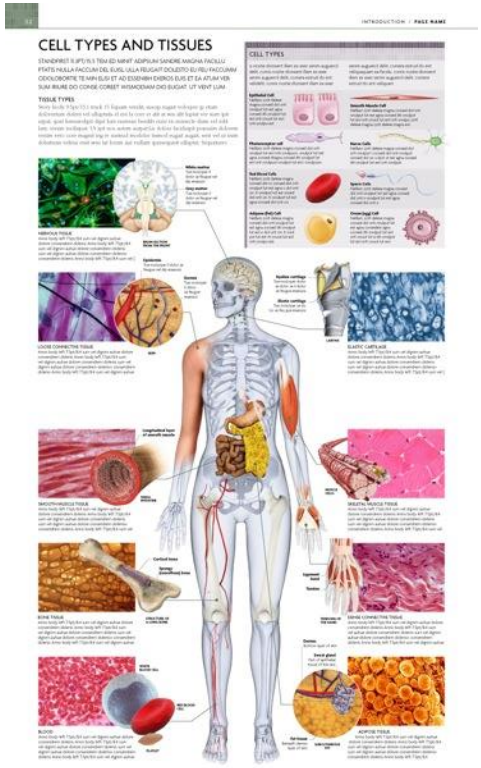
Different kind of RNAs have different expression values



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

Gene expression

- Which RNAs are expressed, and sometimes then translated to proteins, varies between samples
- RNA gives information on which genes that are expressed



-Tissues

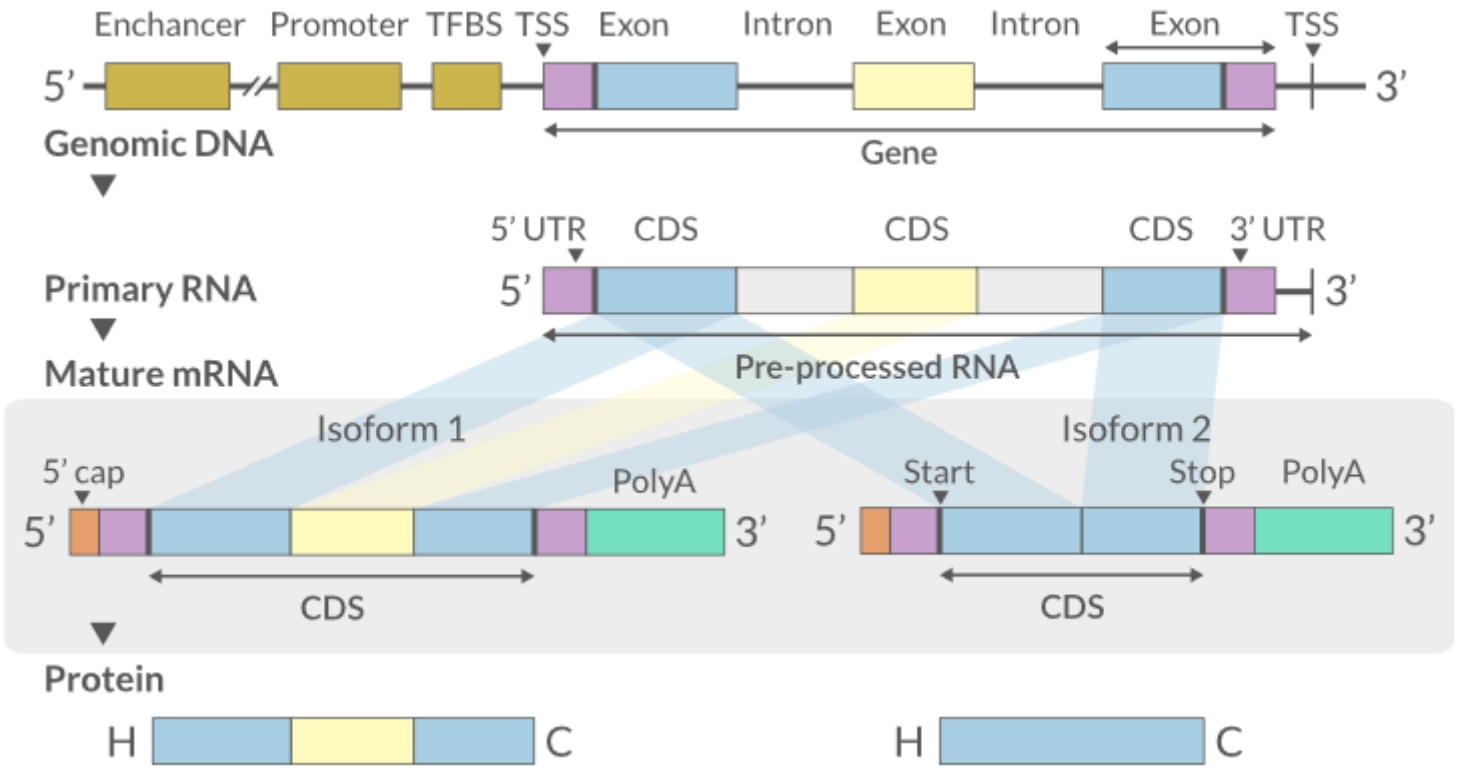
-Cell types

-Cell states

-Individuals

-Cells

Transcriptome

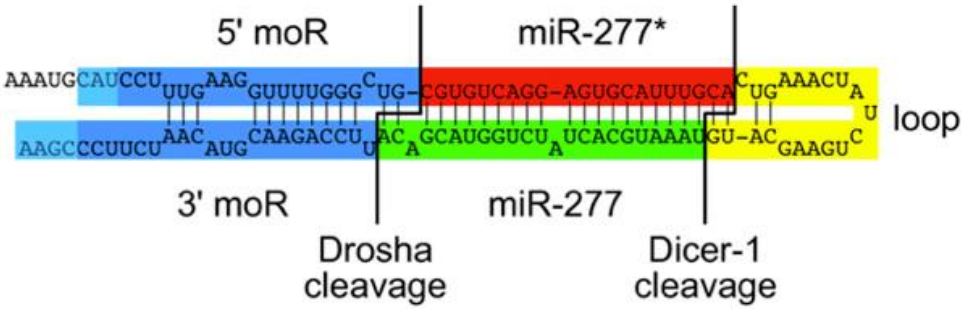


- The transcriptome is spatially and temporally dynamic
- Data comes from functional units (coding regions)
- Only a tiny fraction of the genome
- One gene, many different mRNAs

Applications of RNAseq

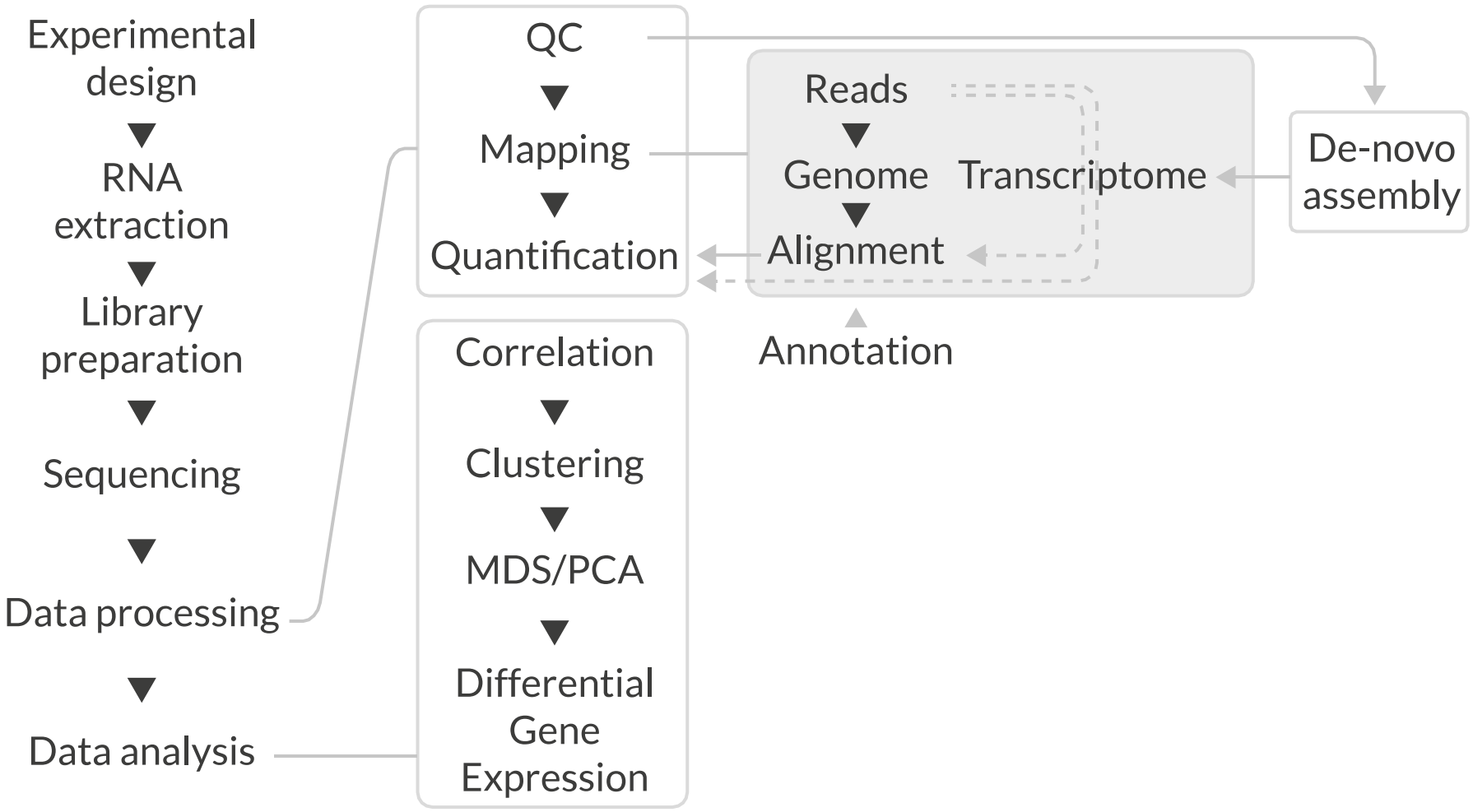
- Identify gene sequences in genomes
  - Novel gene identification/transcriptome assembly
- Differential gene expression
  - Different conditions
  - Transcriptional profiling (e.g. tissue specific expression)
- Explore isoform and allelic expression
- Understand co-expression, pathways and networks
- Gene fusion
- Identification of splice variants
- RNA editing
- SNP finding
- miRNAs

microRNA analysis



(Berezikov et al. Genome Research, 2011.)

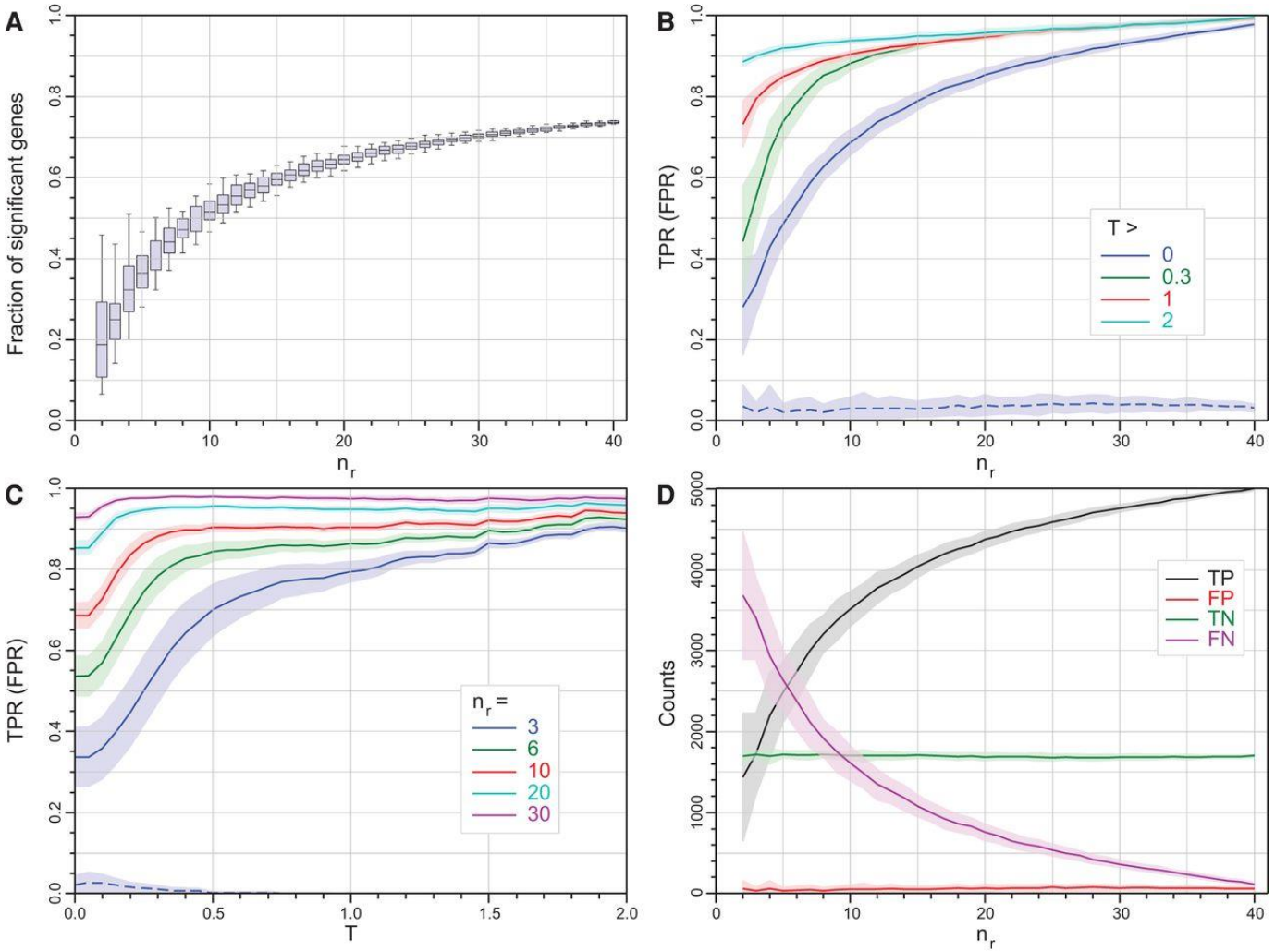
Workflow (for DGE)



Experimental design

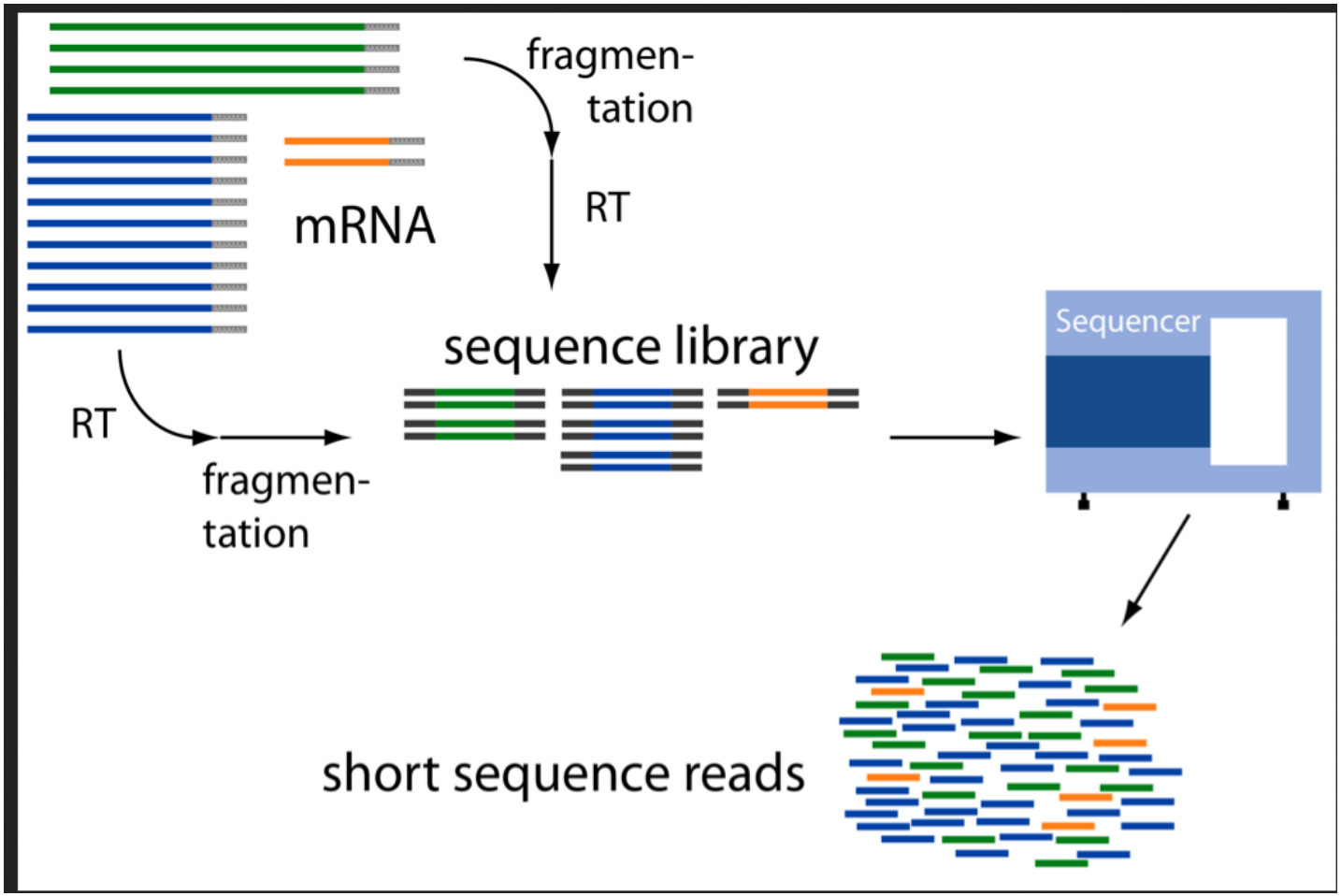
- Avoid technical biases
  - Batch effects
- Balanced design
- Number of replicates
  - Depends on amount of variability (technical, biological) as well as on desired statistical power
  - Technical replicates (most often) not necessary
  - Biological replication required if inference on the population is to be made, three replicates is the minimum for any inferential analysis
  - Biological replicates: 6 - 12 (Schurch *et al.*, 2016)

Statistical properties of edgeR (exact) as a function of threshold, T, and the number of replicates, nr.



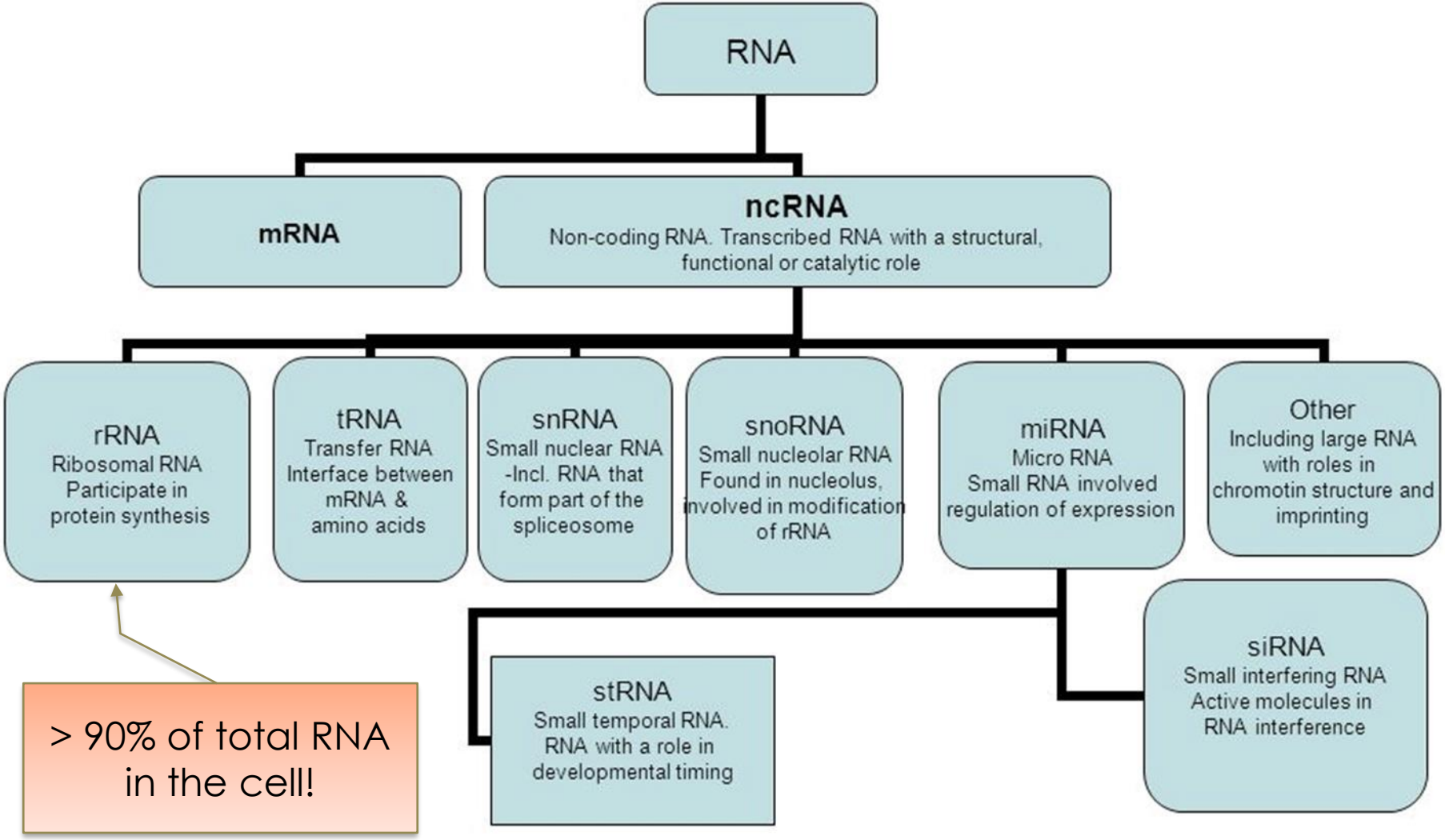
Nicholas J. Schurch et al. RNA 2016;22:839-851

How is RNA-seq data generated?



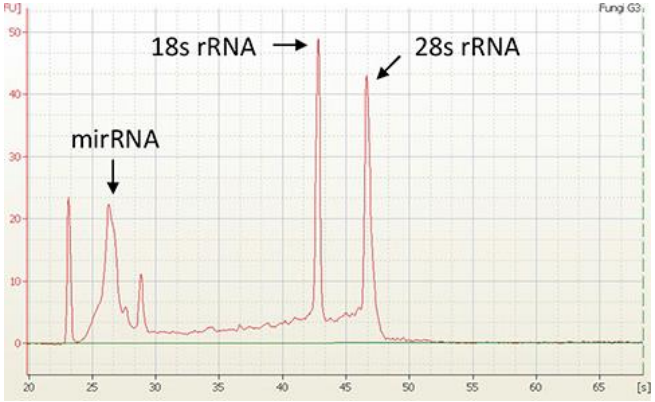
Sampling process

A wide variety of functional RNAs



rRNA depletion vs. polyA selection

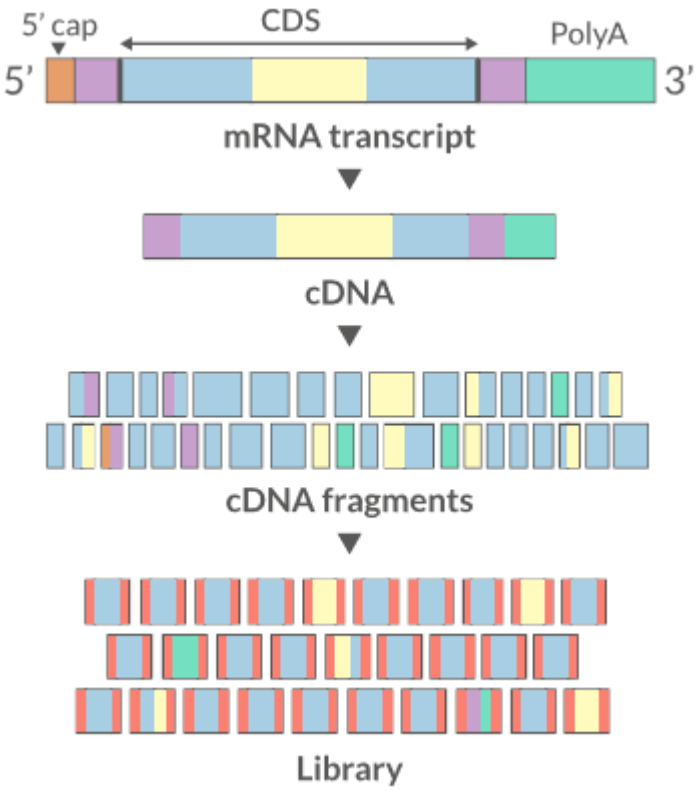
Method	Pros	Cons	Recommended
rRNA depletion	•Captures on-going transcription •Picks up non-coding RNA	•Does not get rid of all rRNA •Messy diff.ex. profile	20-40 mln reads (single or PE)
polyA selection	•Gives a clean diff.ex. profile	•Does not pick non-coding RNA •Does not work on prokaryotes	5-20 mln reads



Alternative for **human** RNA-seq:
AmpliSeq Human Transcriptome panel:
• faster, cheaper, works fine with FFPE
• input: 50 ng **total** RNA
• diff.ex. ONLY

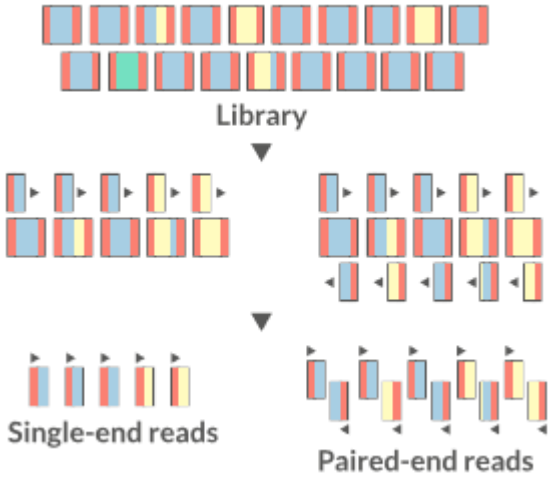
Library preparation

- RNA selection
  - rRNA depletion/polyA selection
  - Size selection (miRNA)
  - Exome capture
- Generation of cDNA
  - Strand preserving library?
- Fragmentation and size selection



Sequencing

- Sequencer (Illumina/PacBio)
- Read length
- Pooling samples
- Sequencing depth (coverage)
- Single-end reads
  - Cheaper
- Paired-end reads
  - Easier to map correctly; higher accuracy for DGE
  - Better assemblies
  - Better for structural variation and isoforms



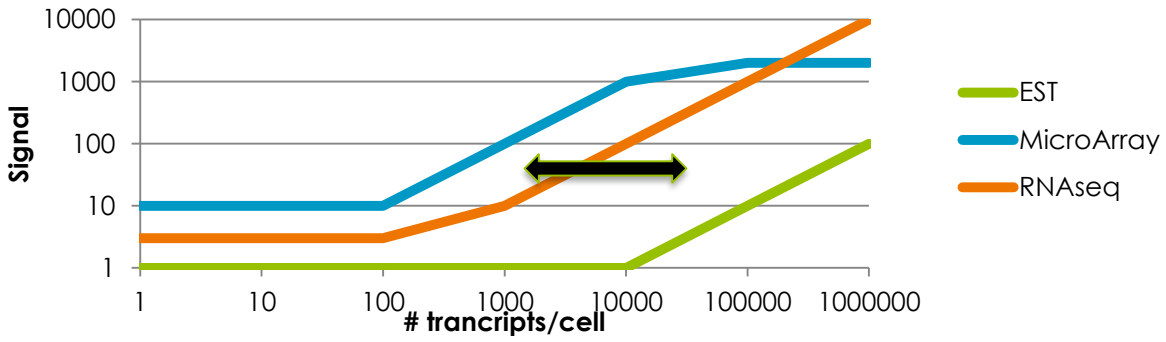
Promises and pitfalls

Long reads

- Low throughput (-)
- Complete transcripts (+)
- Only highly expressed genes (--)
- Expensive (-)
- Easy downstream analysis (+)
- Better for isoforms (+)

Short reads

- High throughput (+)
- Fractions of transcripts (-)
- Full dynamic range (+-)
- Cheap (+)
- Strand specificity (+)
- Greater than 50bp does not improve DGE



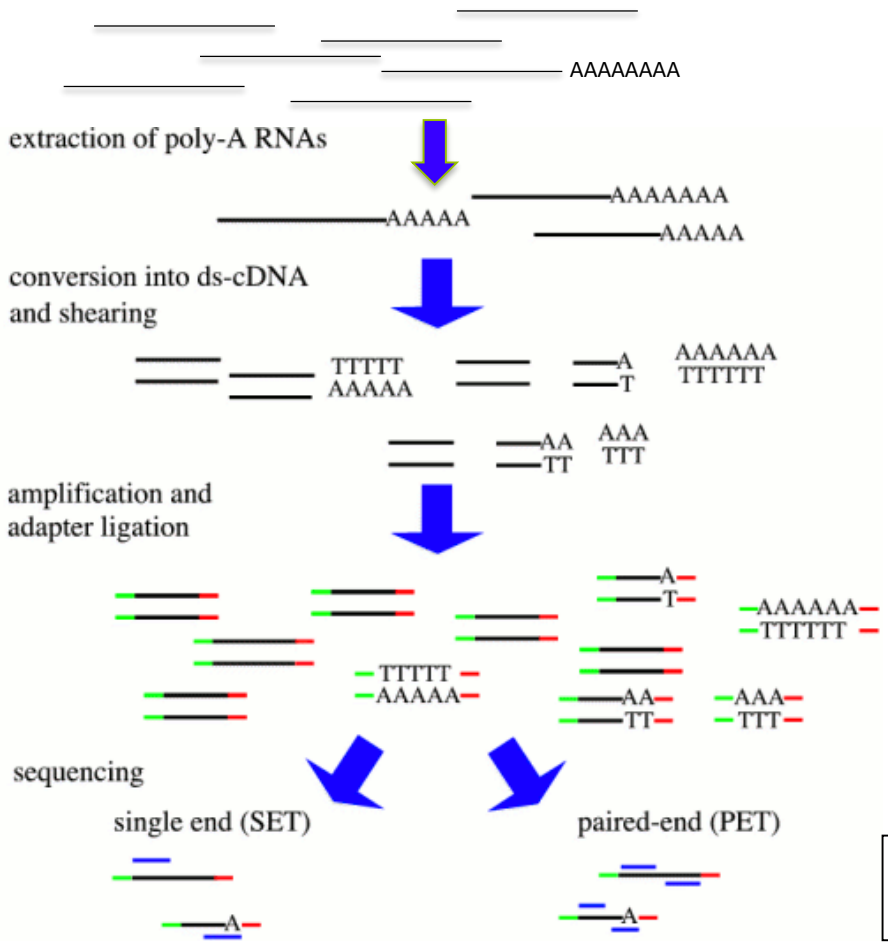
Depending on the different steps you will get different results

RNA->

enrichments ->

library ->

reads ->



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

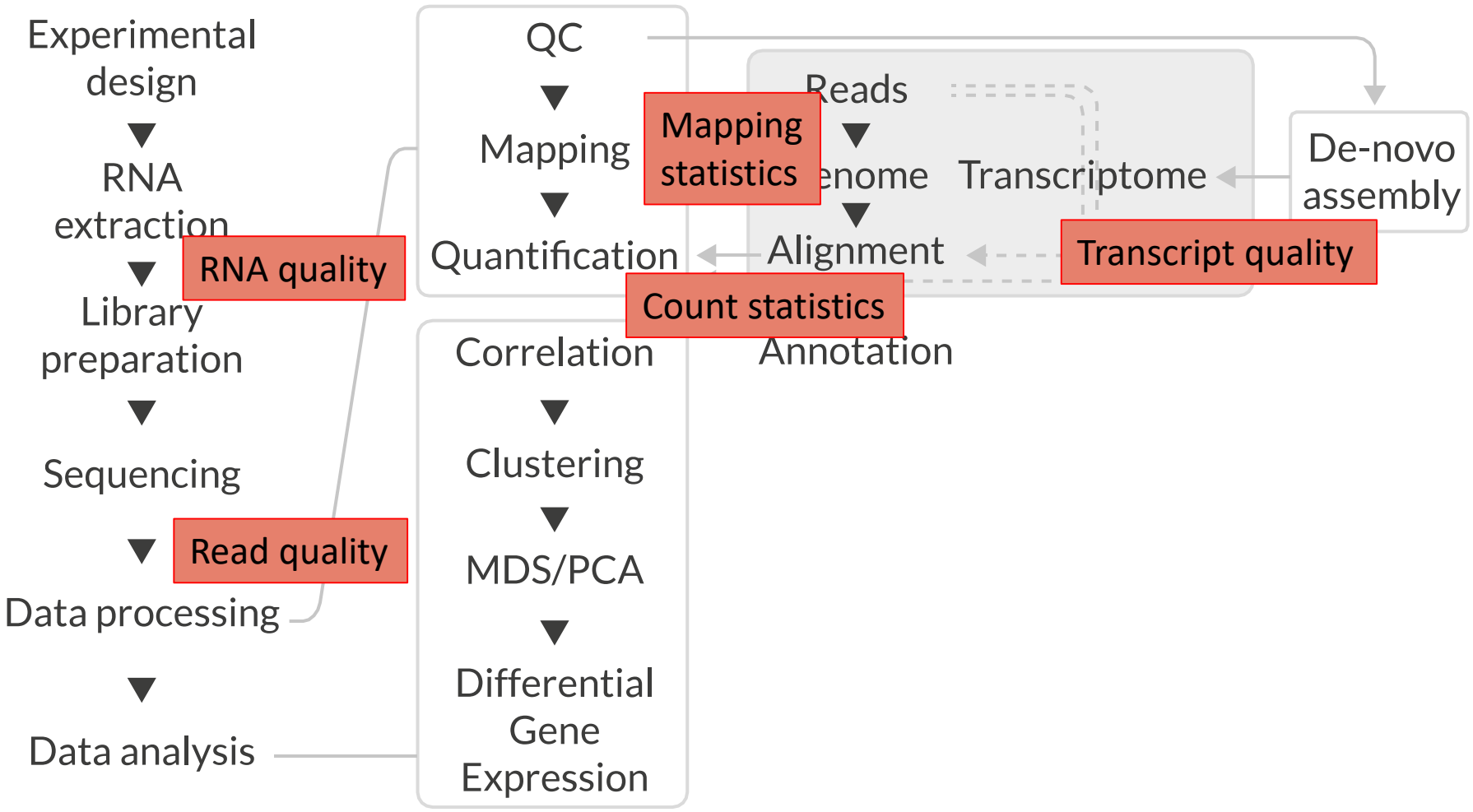
Size of fragment
Strand specific
5’ end specific
3’ end specific
…..

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

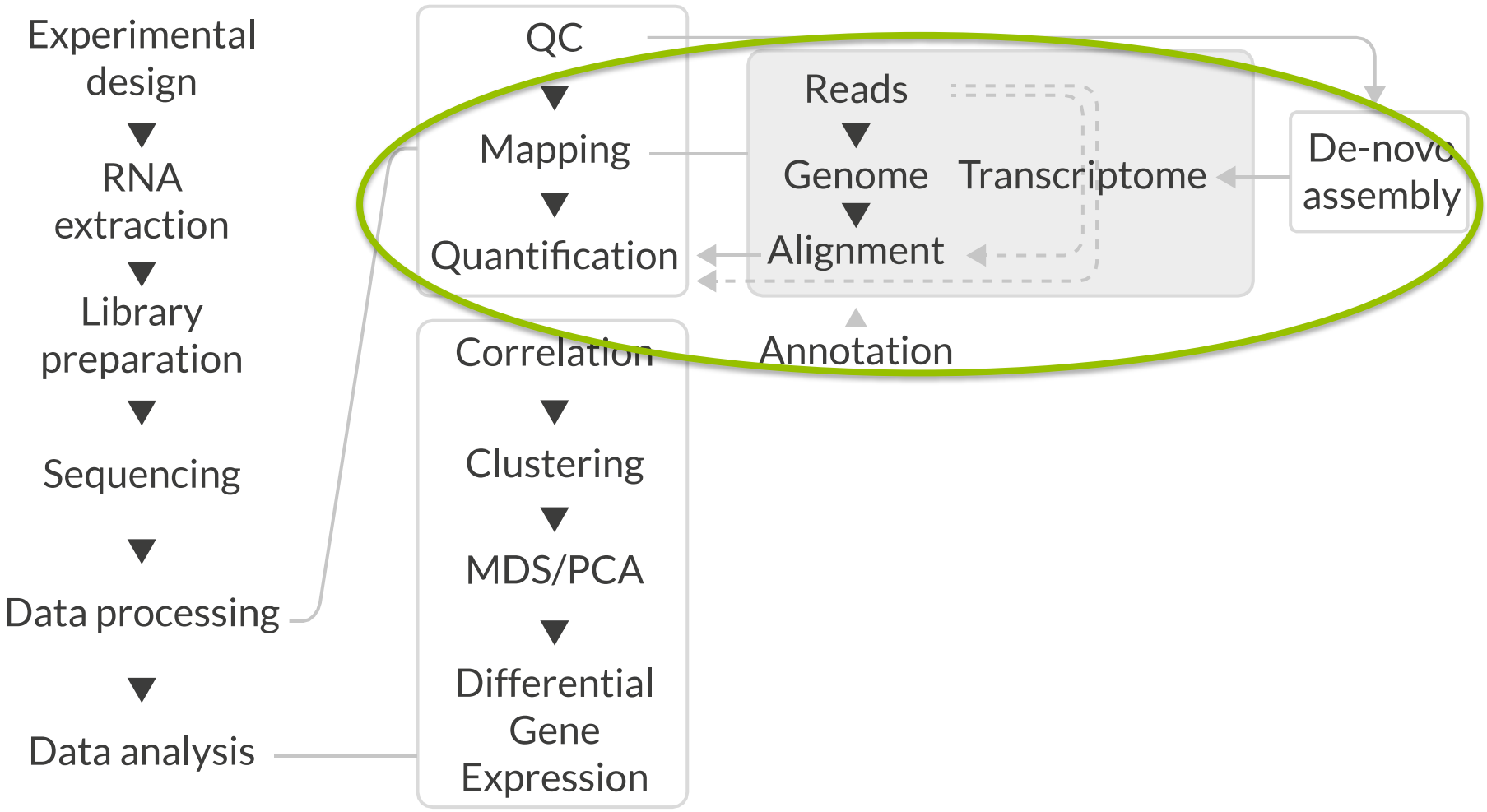
Quality control -samples might not be what you think they are

- 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 for one of every 20 samples the reads do 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

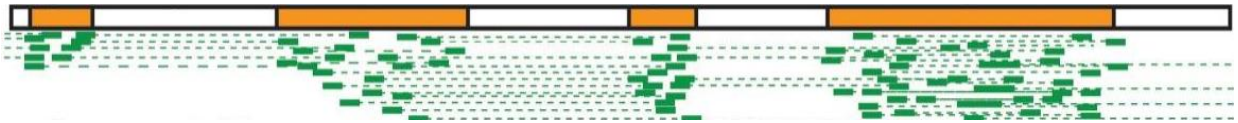
Workflow


Experimental design
RNA extraction
Library preparation
Sequencing
Data processing
Data analysis
RNA quality
Read quality
QC
Mapping
Quantification
Mapping statistics
Count statistics
Correlation
Clustering
MDS/PCA
Differential Gene Expression
Reads
enome
Transcriptome
Alignment
Annotation
De-novo assembly
Transcript quality

Workflow

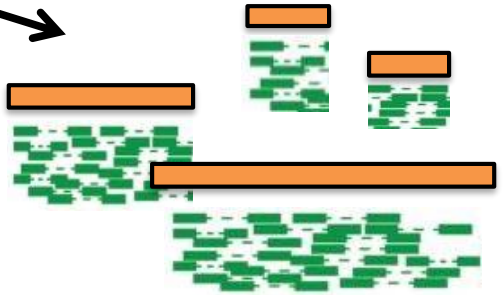

Experimental design
RNA extraction
Library preparation
Sequencing
Data processing
Data analysis
QC
Mapping
Quantification
Reads
Genome
Transcriptome
Alignment
Annotation
De-novo assembly
Correlation
Clustering
MDS/PCA
Differential Gene Expression

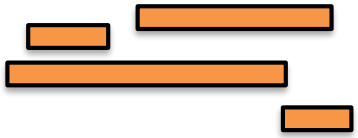
From reads to counts



Map reads to reference genome

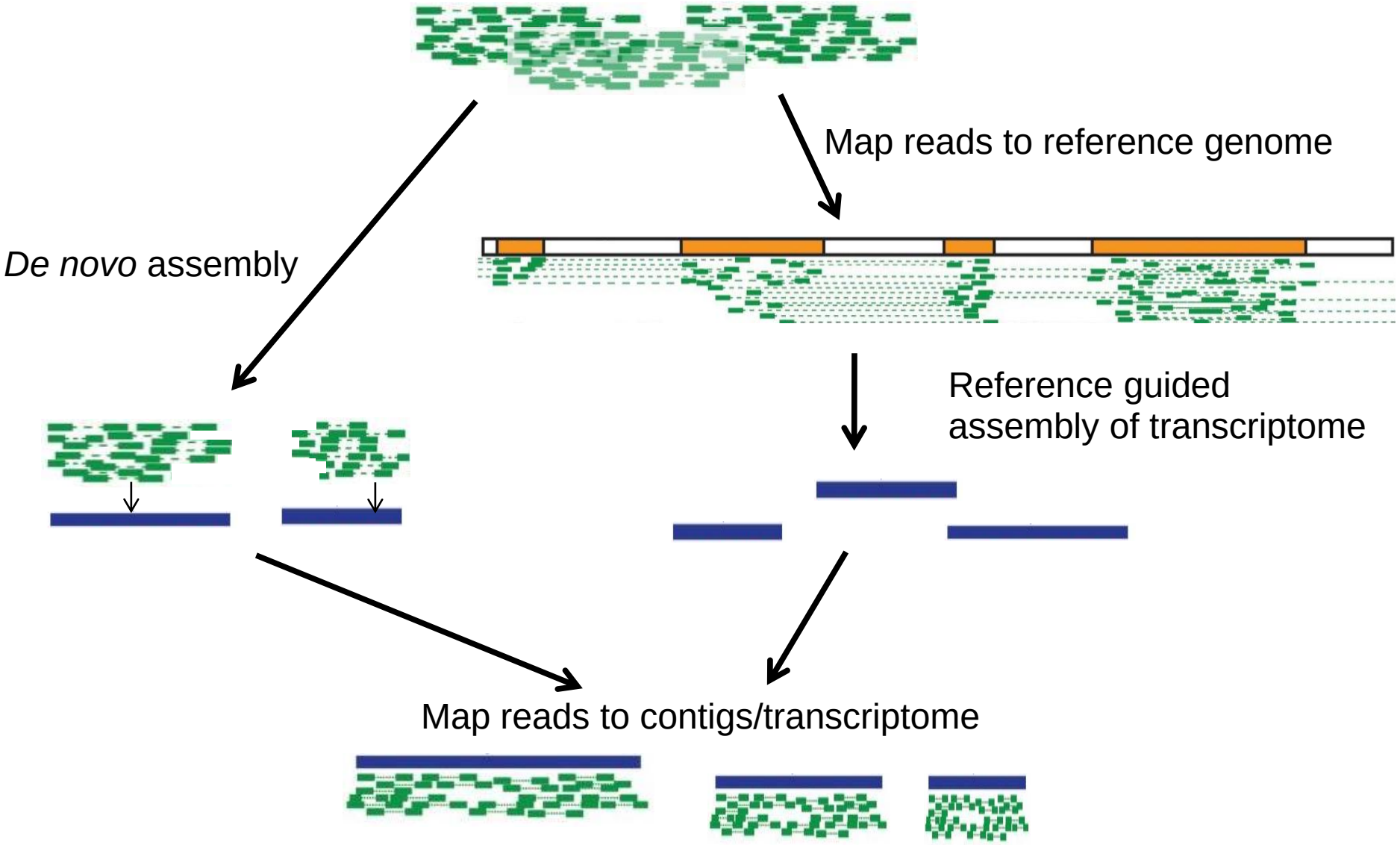
Map reads to transcriptome





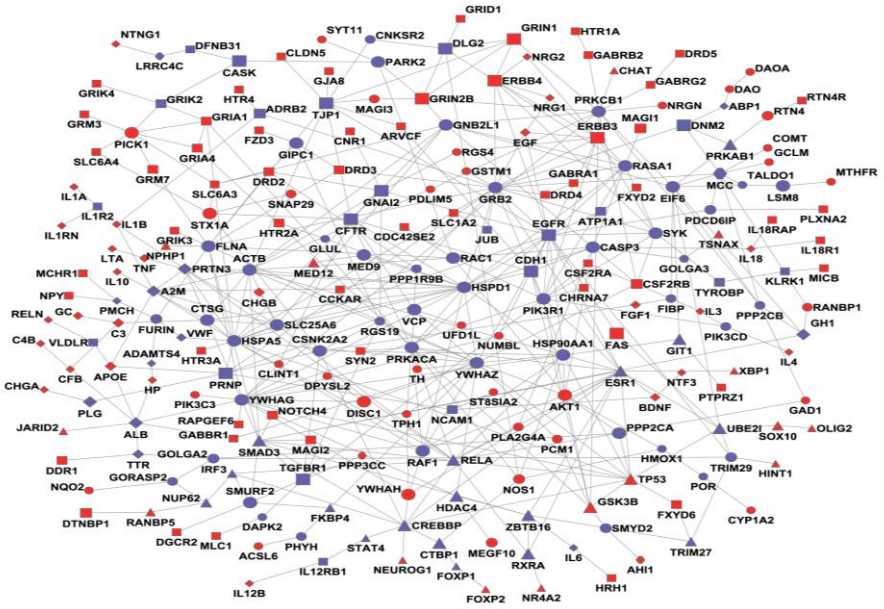
Reads to counts using transcriptome, alignment free

From reads to counts

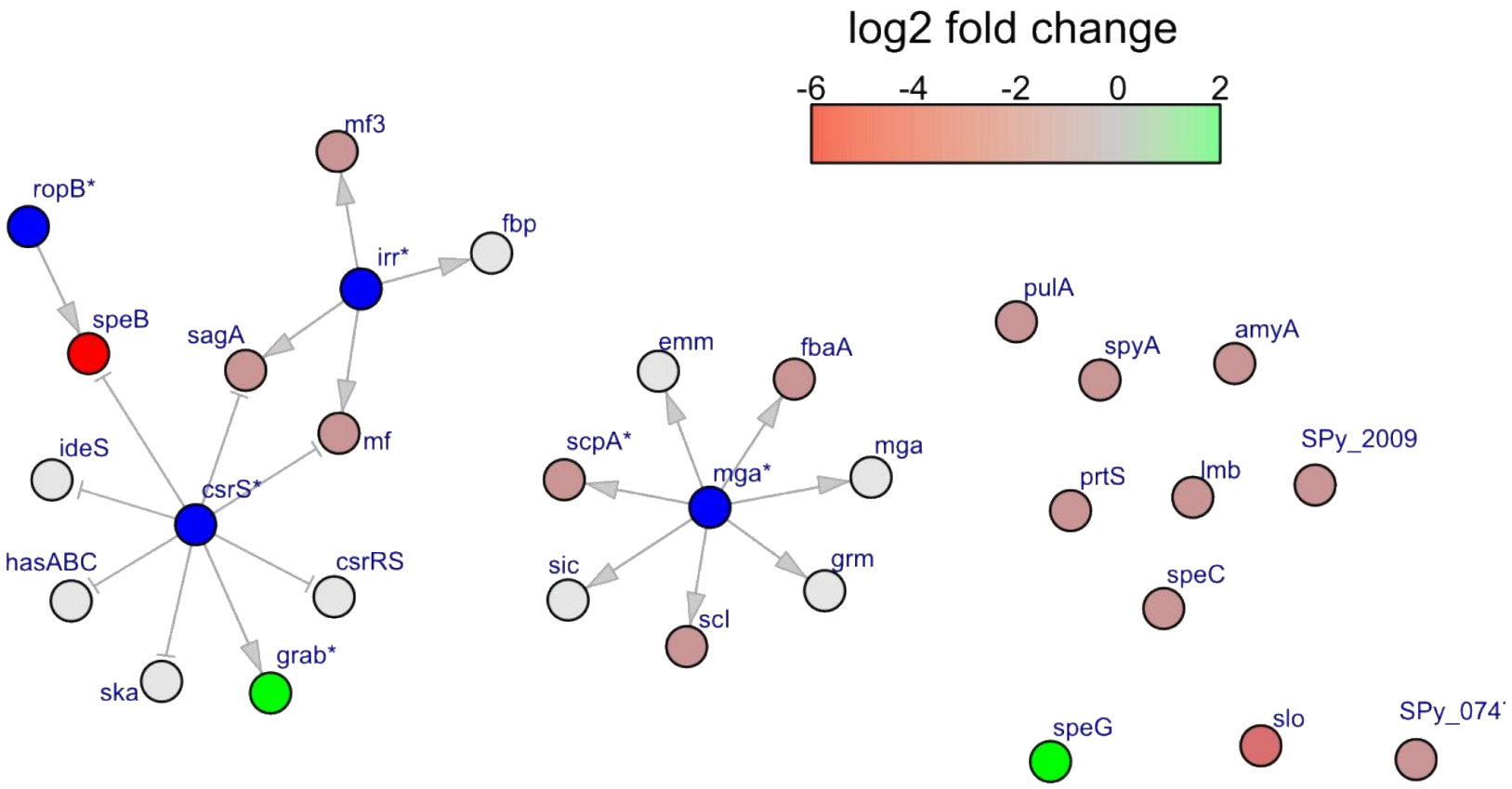


Differential expression analysis using univariate analysis

- Which genes are up-/down-regulated in one group compared the other(s)?
- Typically univariate analysis (one gene at a time) – even though we know that genes are not independent



Gene set analysis and data integration



The RNAseq workshop

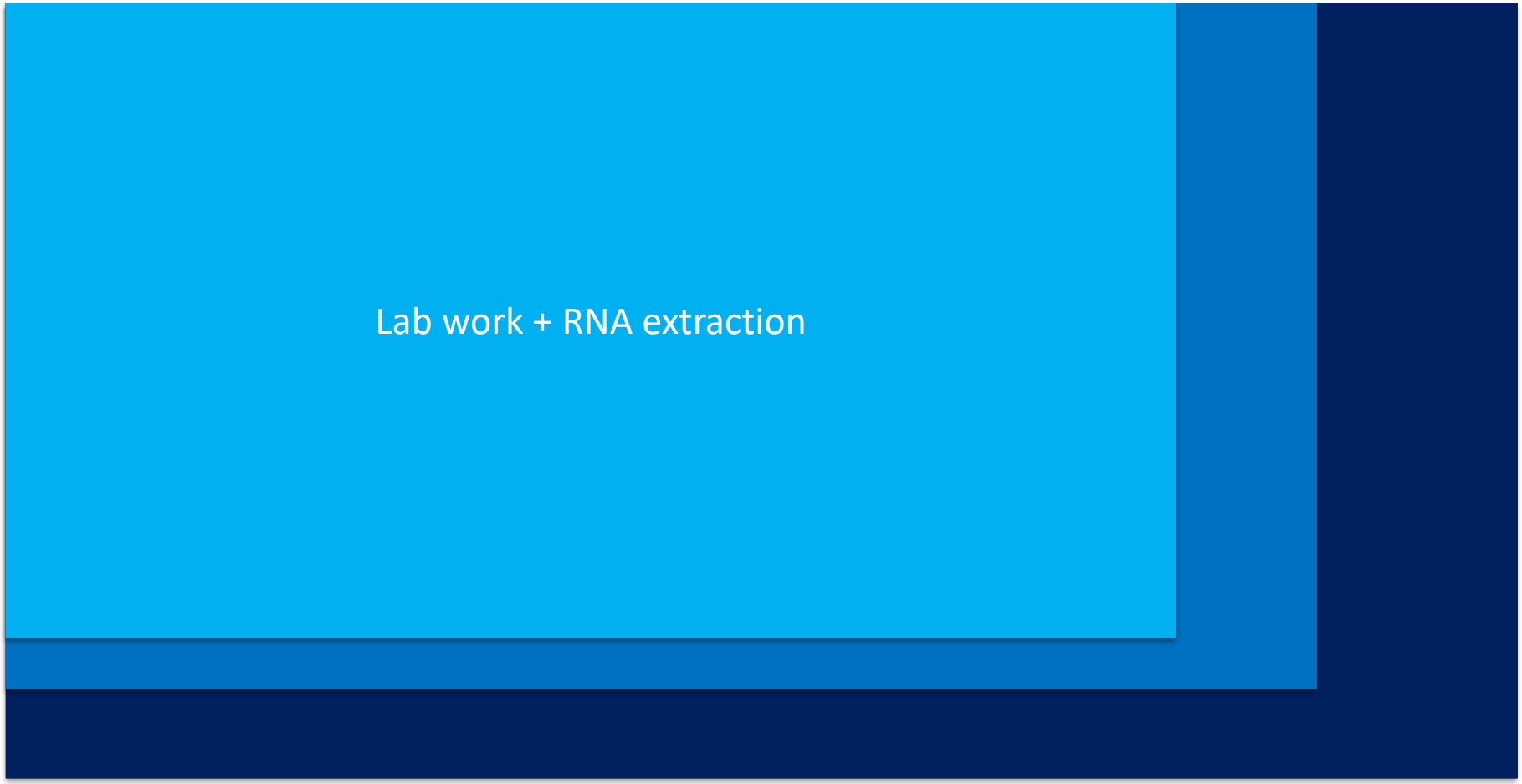
- From RNA to seq to reads (Introduction)
- Quality control (Wednesday)
- Mapping reads programs (Wednesday)
- Data management (Wednesday)
- Differential expression analysis (Thursday)
- Gene set analysis (Thursday)
- RNAseq pipeline (Thursday)
- Transcriptome assembly, with and without reference(Friday)
- Functional annotation of transcripts (Friday)

All steps will affect the results

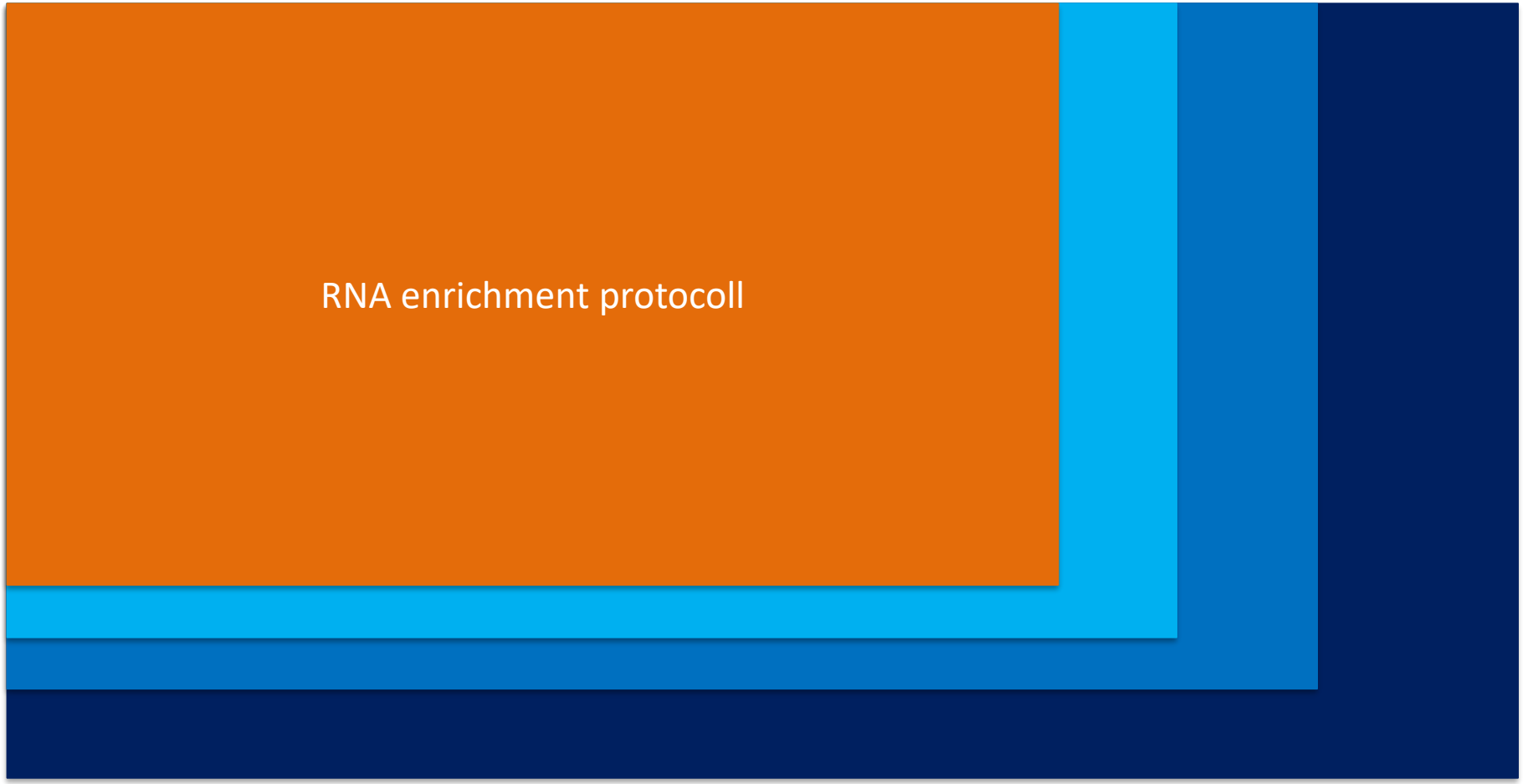
All steps will affect the results

Experimental setup

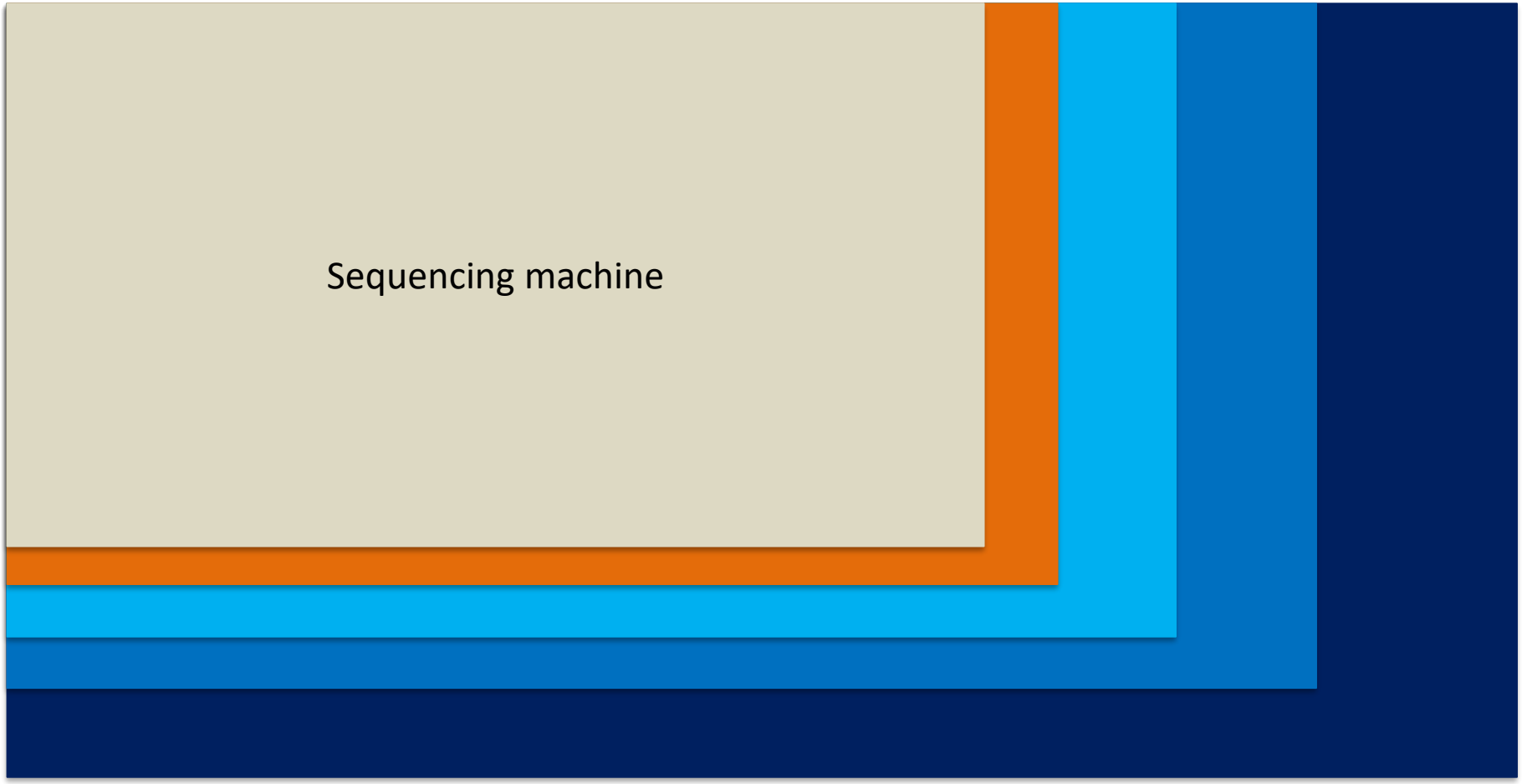
All steps will affect the results



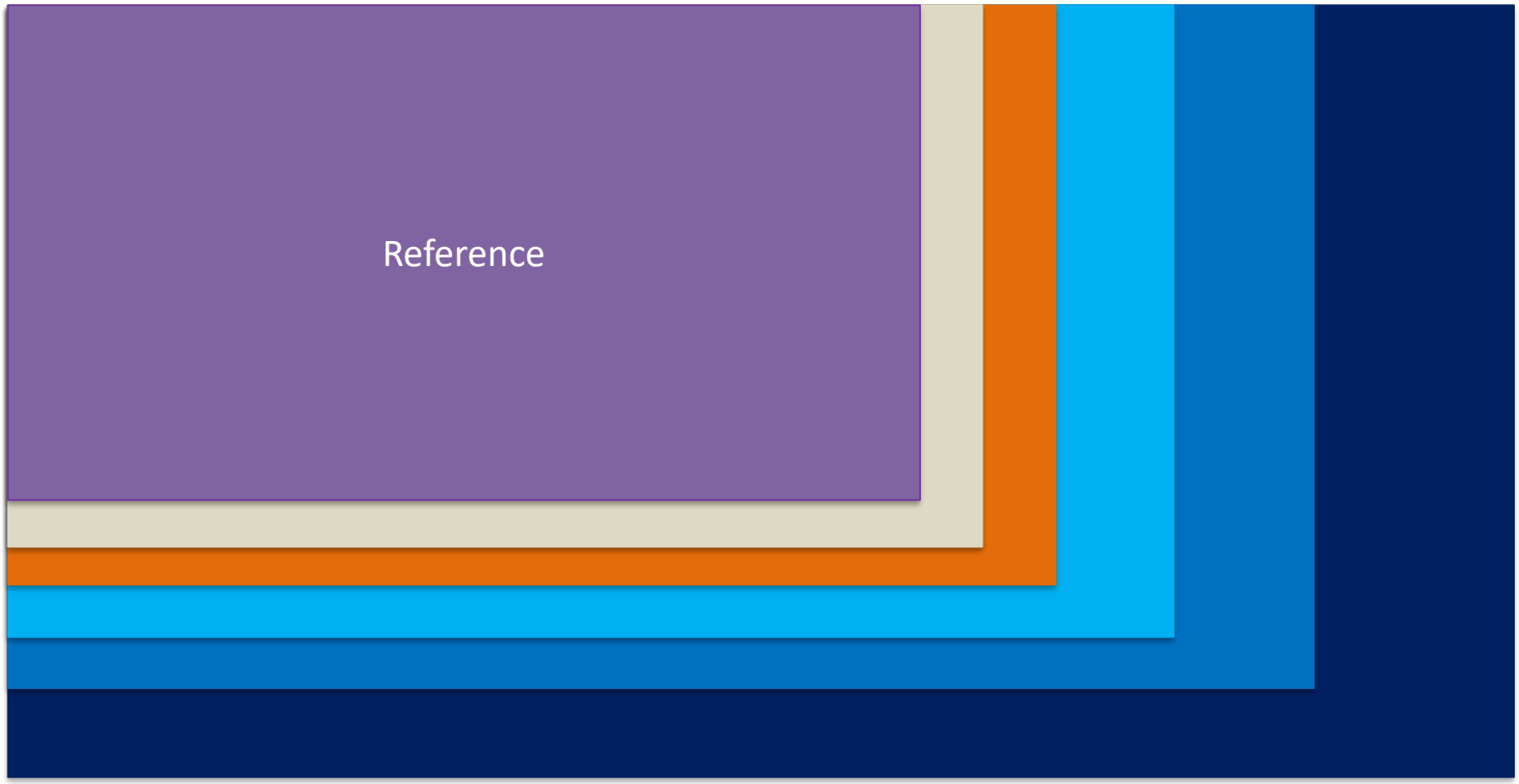
All steps will affect the results



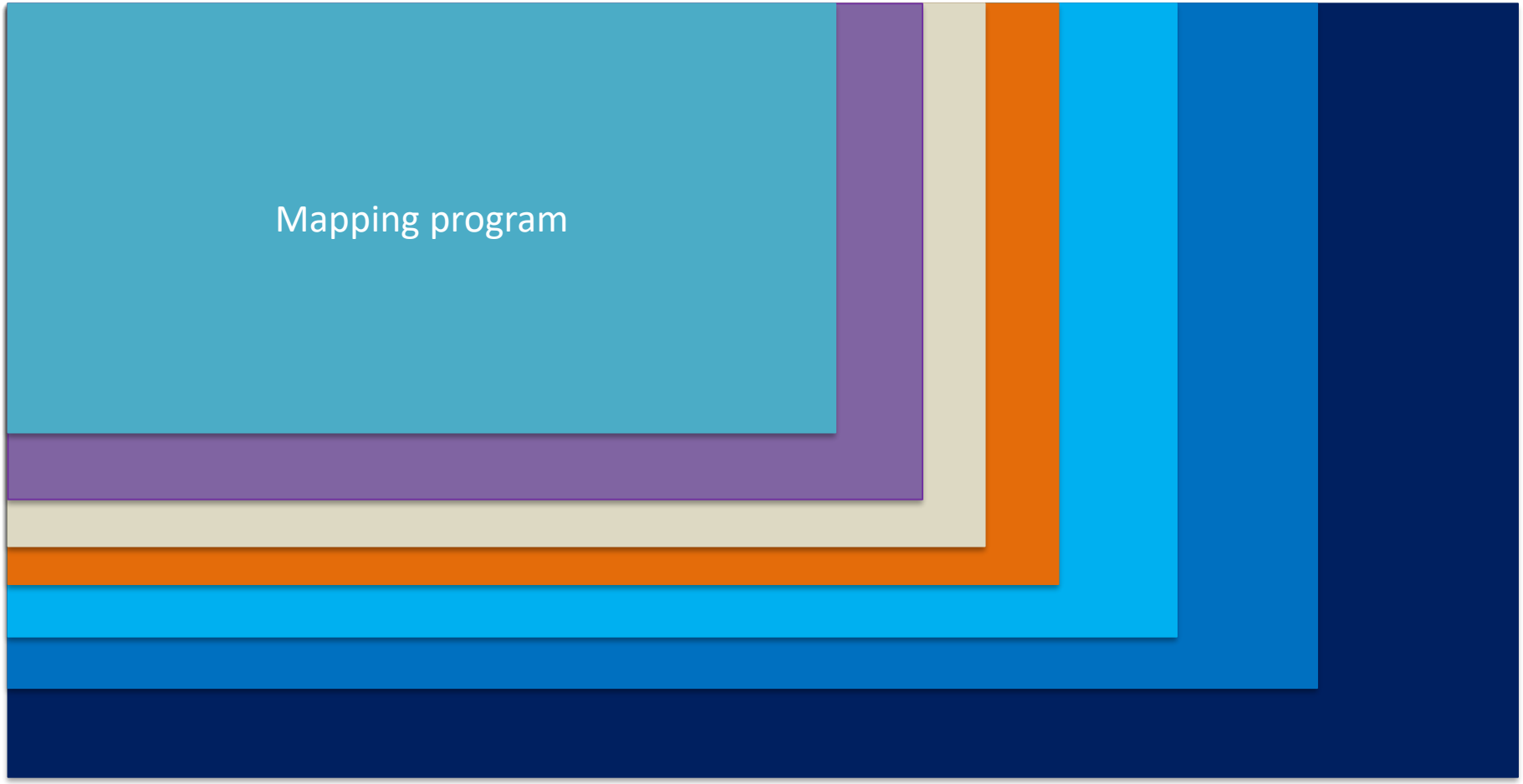
All steps will affect the results



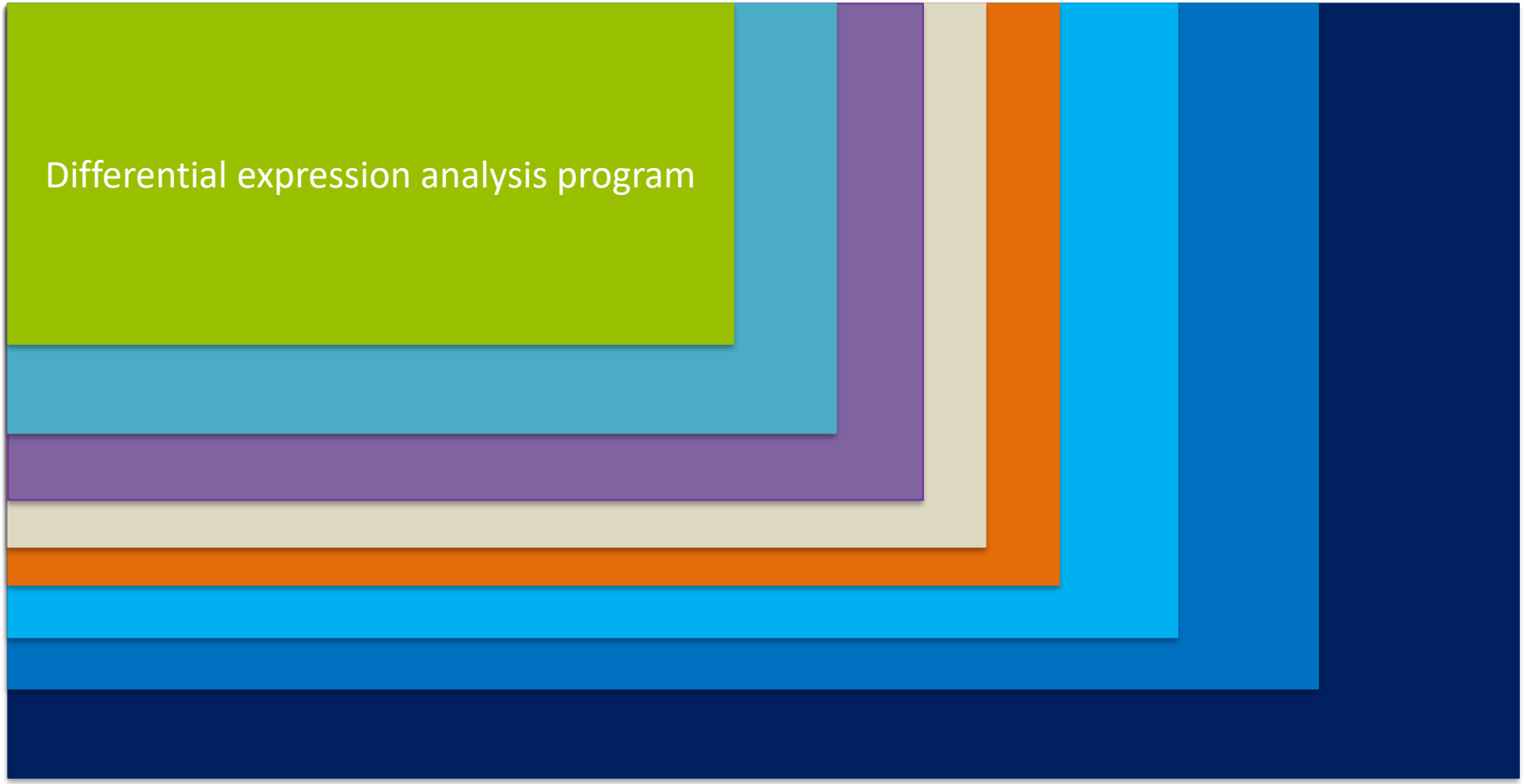
All steps will affect the results



All steps will affect the results



All steps will affect the results



Try to be as consistent as possible