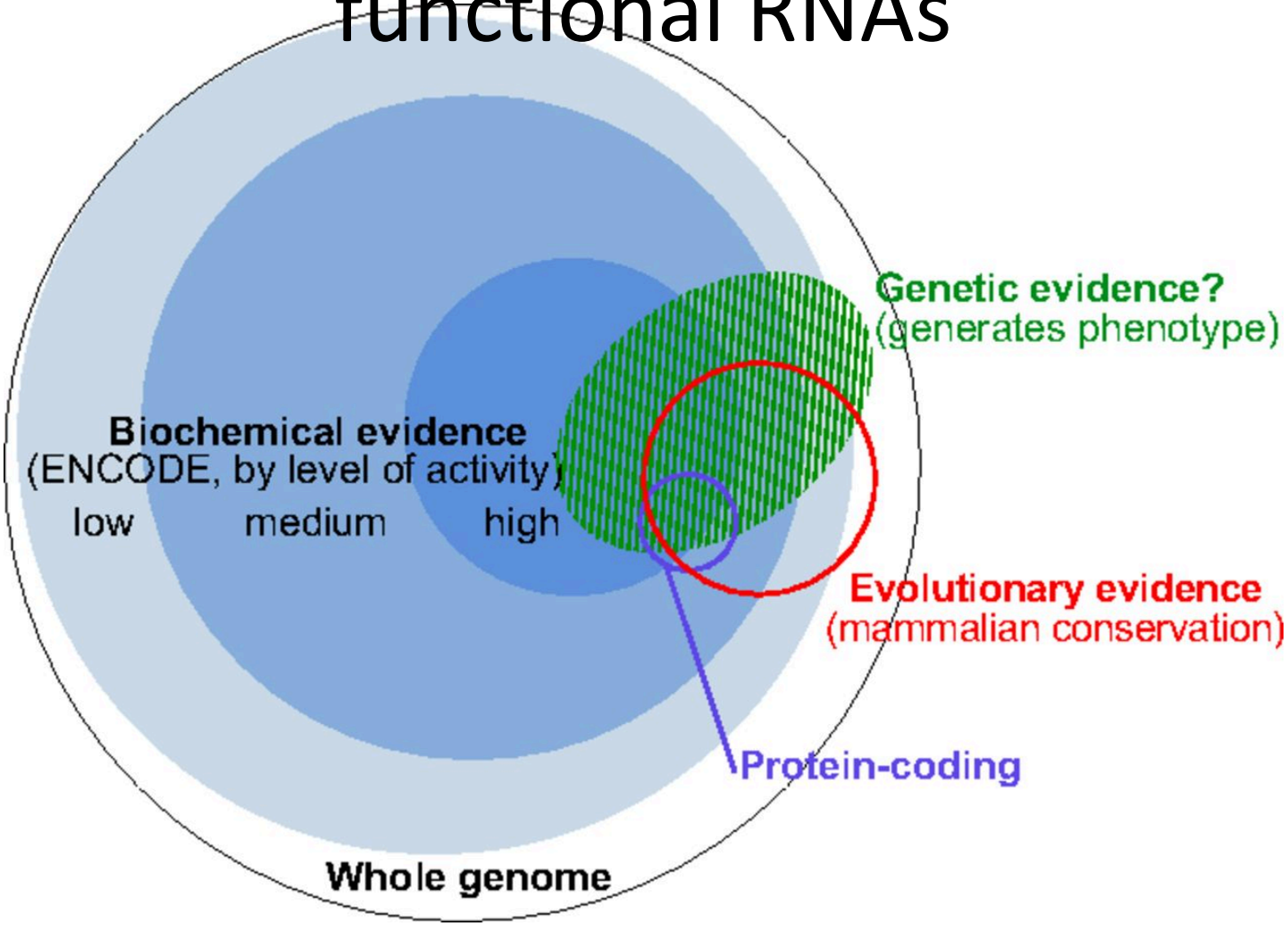
RNAseq analysis

-it's complicated

November 2017

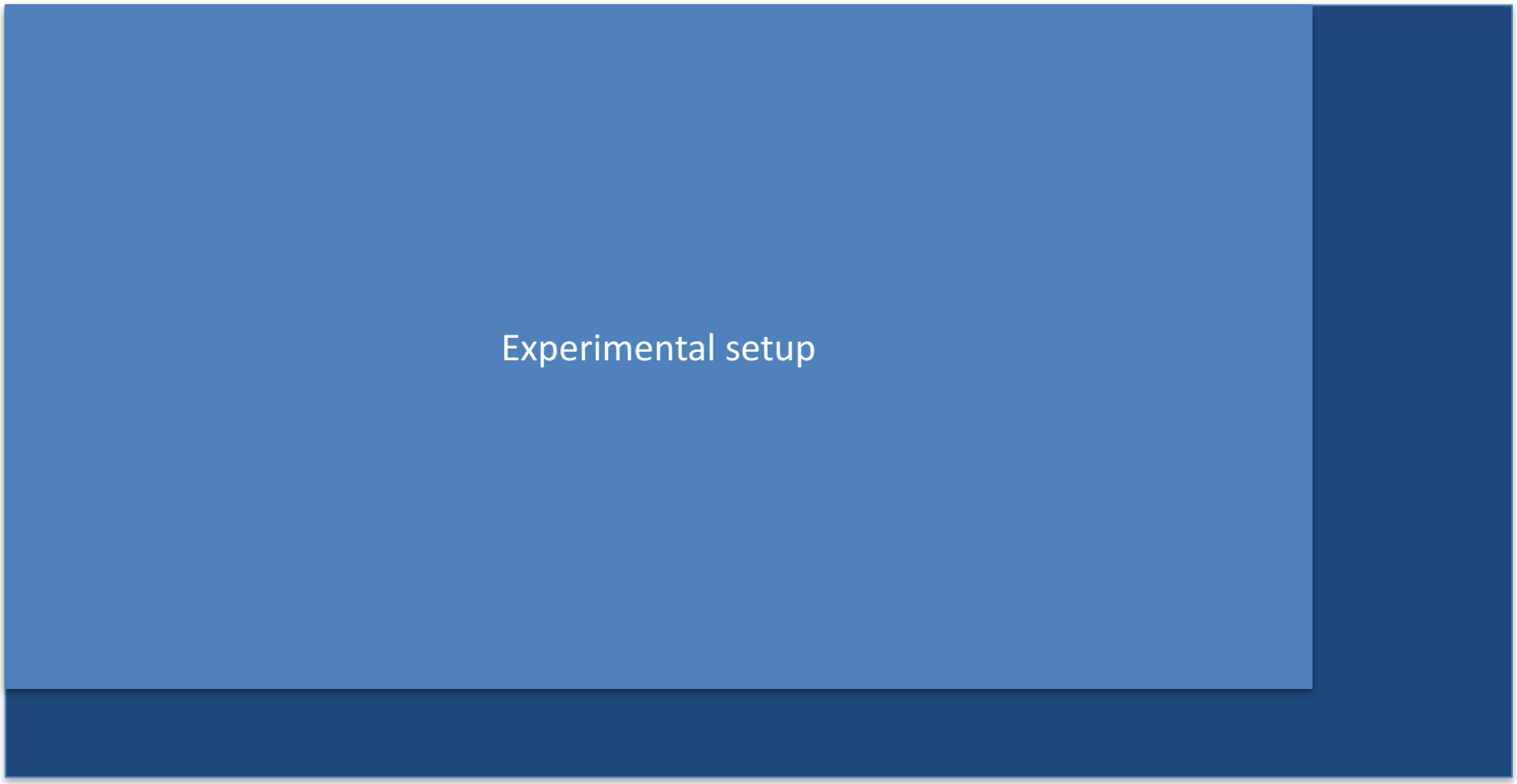
RNA reads are not enough to identify functional RNAs



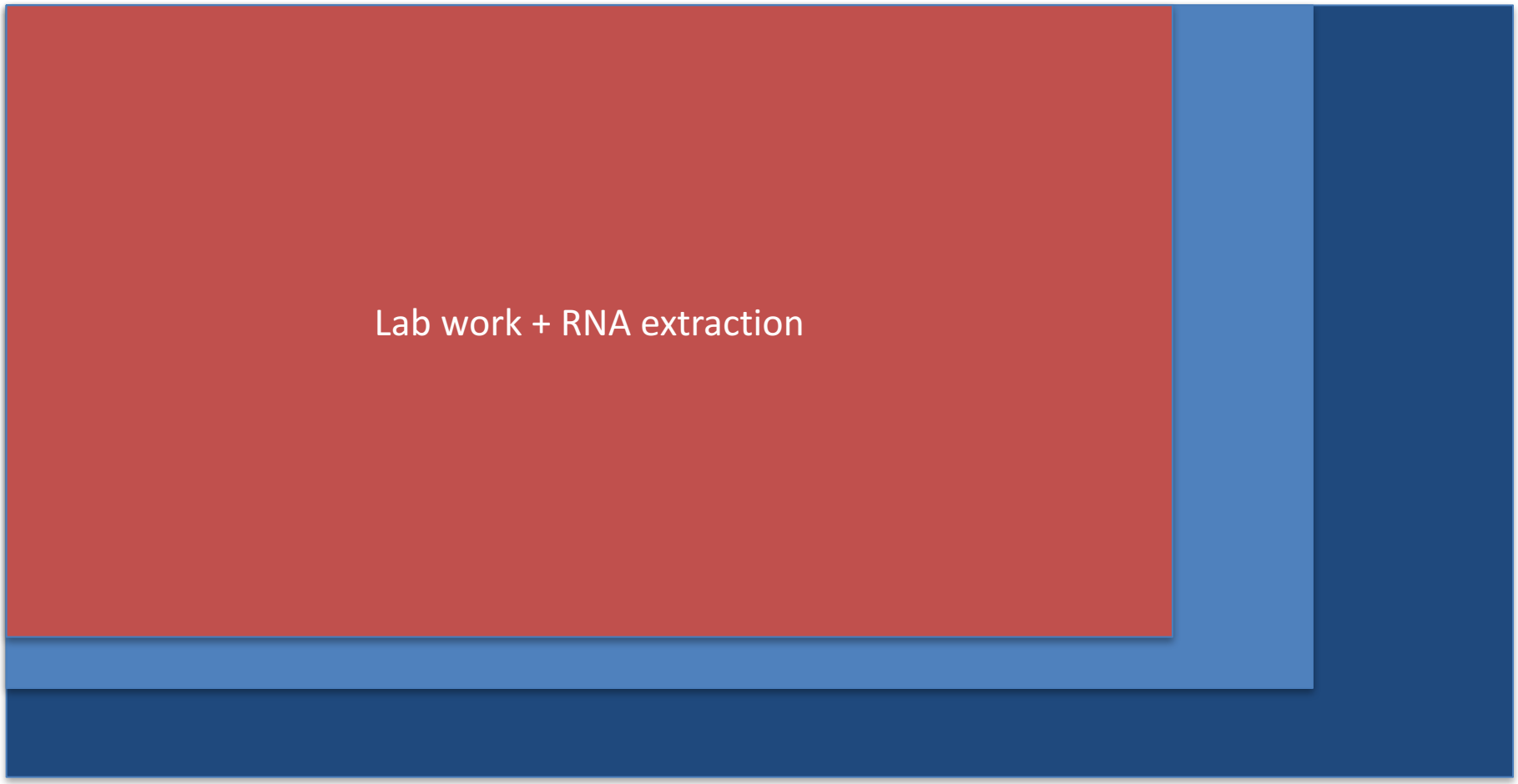
Defining functional DNA elements in the human genome
Kellis M et al. PNAS 2014;111:6131-6138

All the steps will affect the results

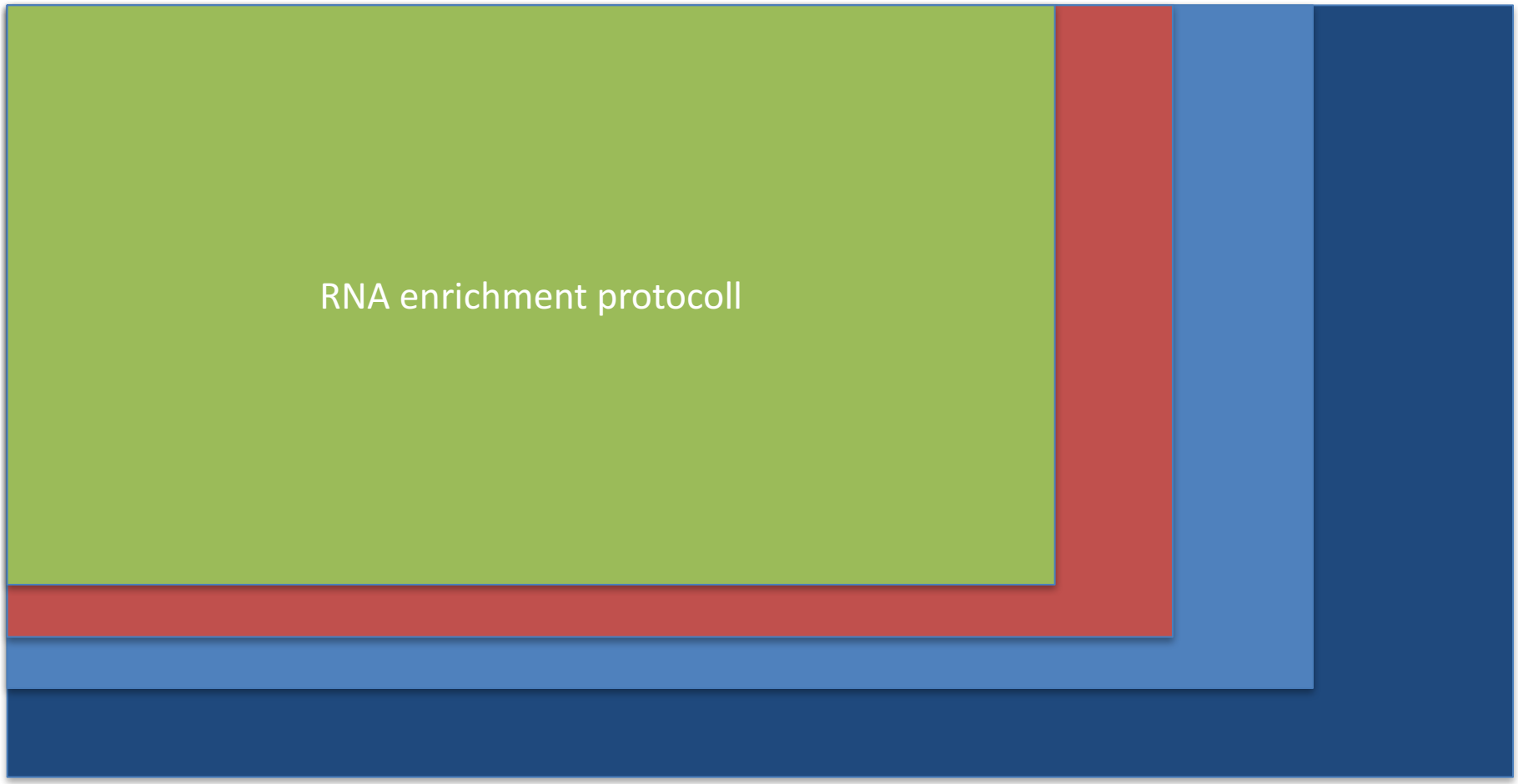
All the steps will affect the results



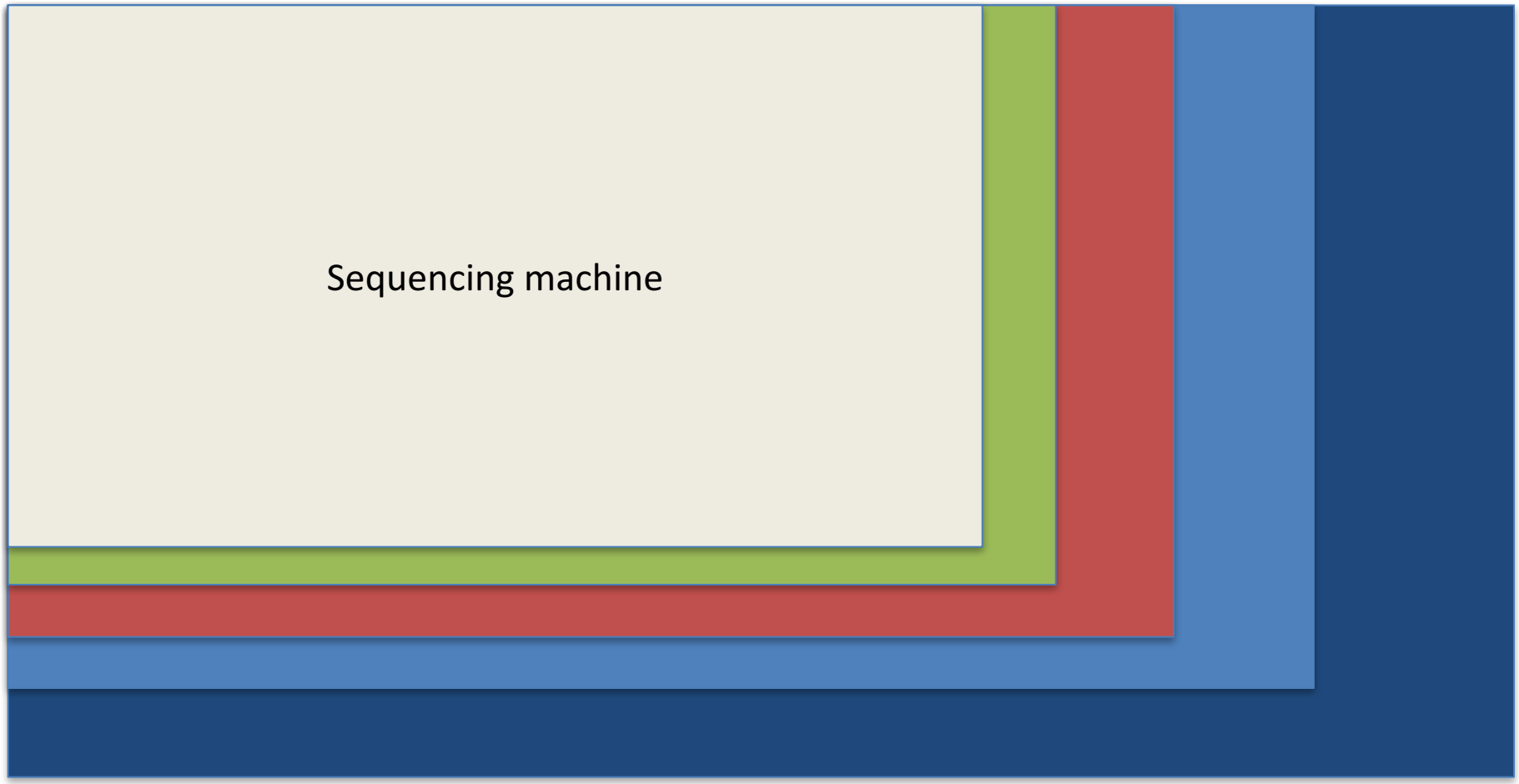
All the steps will affect the results



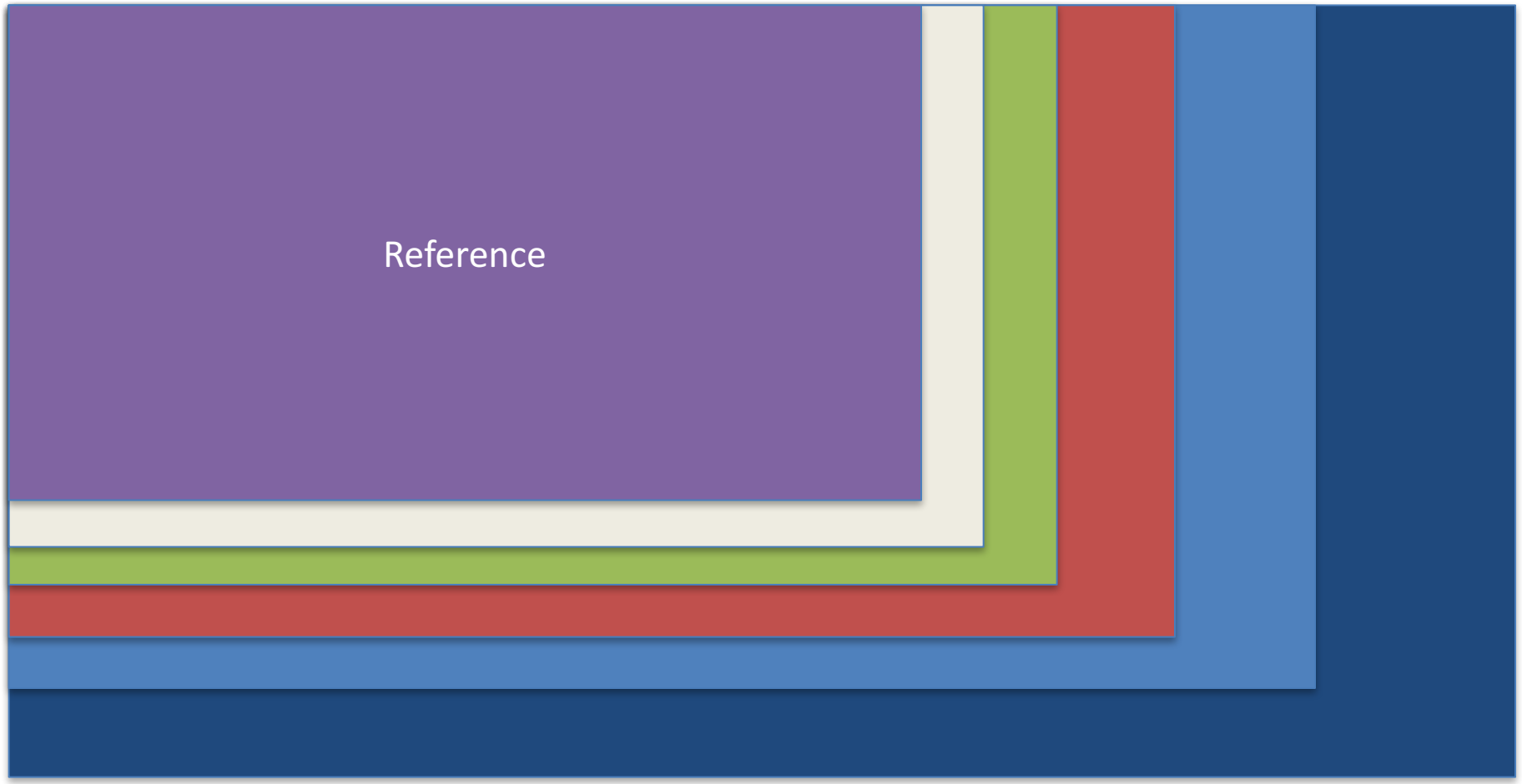
All the steps will affect the results



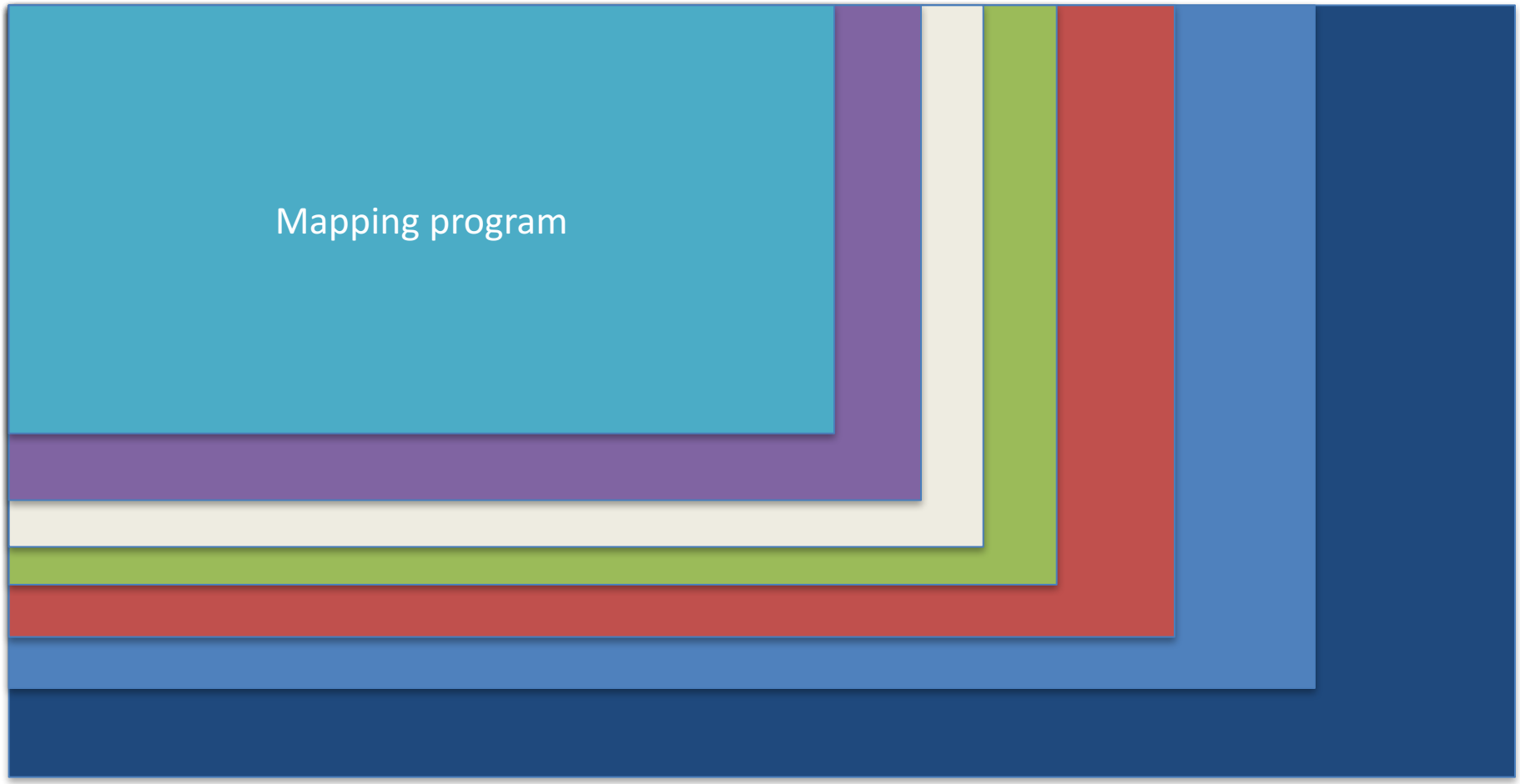
All the steps will affect the results



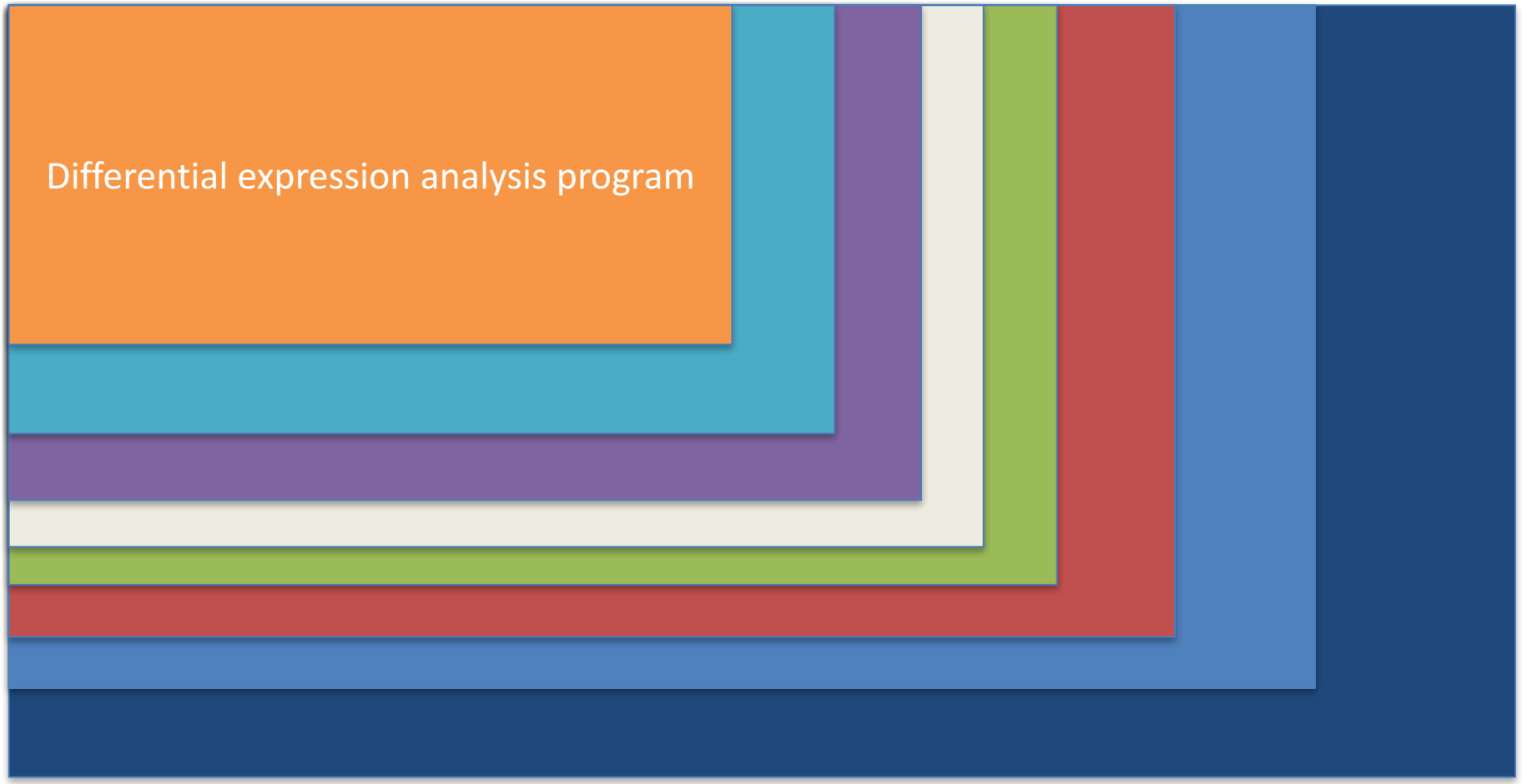
All the steps will affect the results



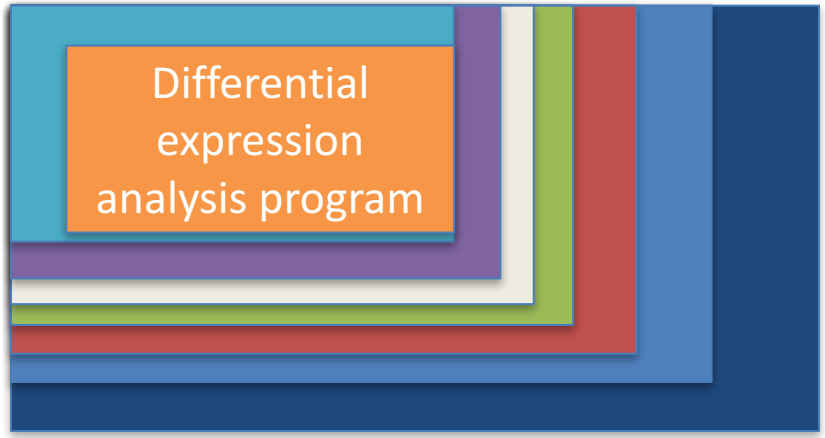
All the steps will affect the results

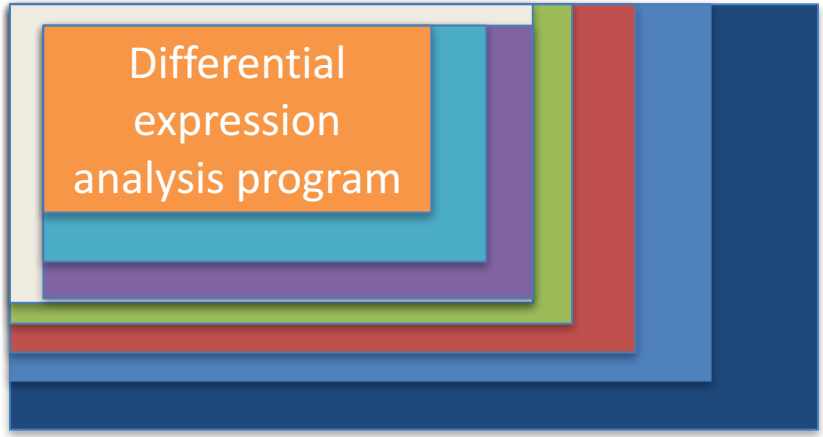


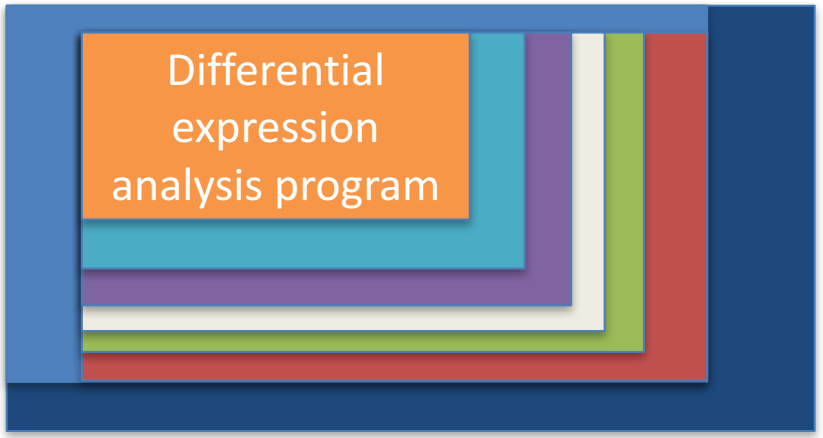
All the steps will affect the results



Try to be as consistent as possible

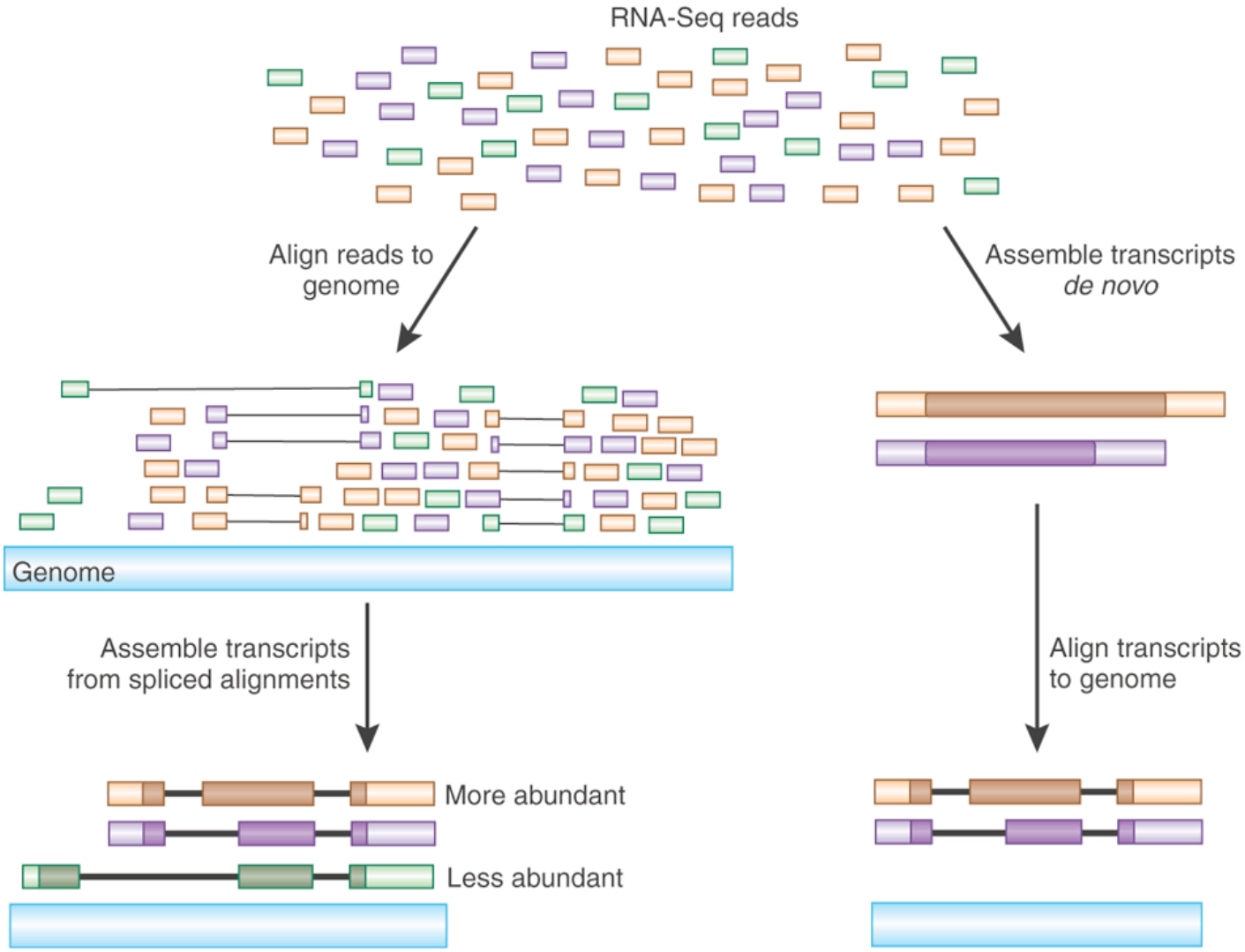






Use a pipeline!

Gene and Isoform detection



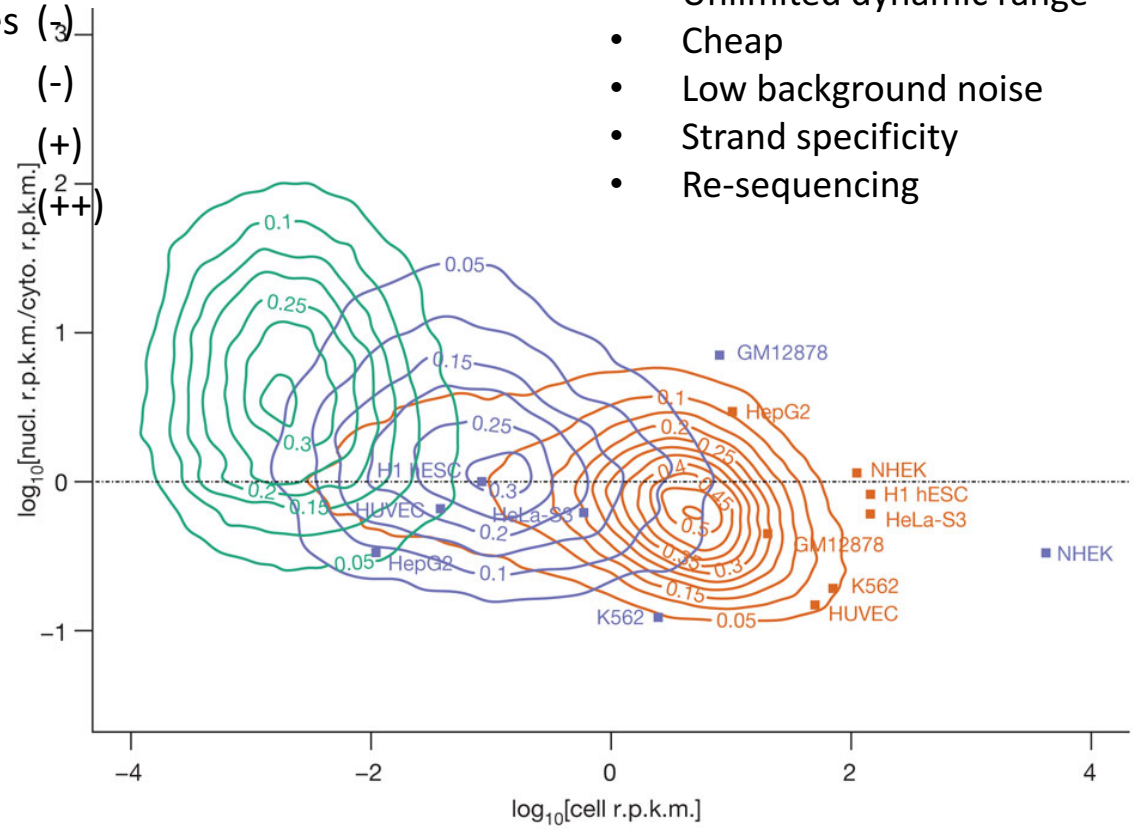
Promises and pitfalls

Long reads

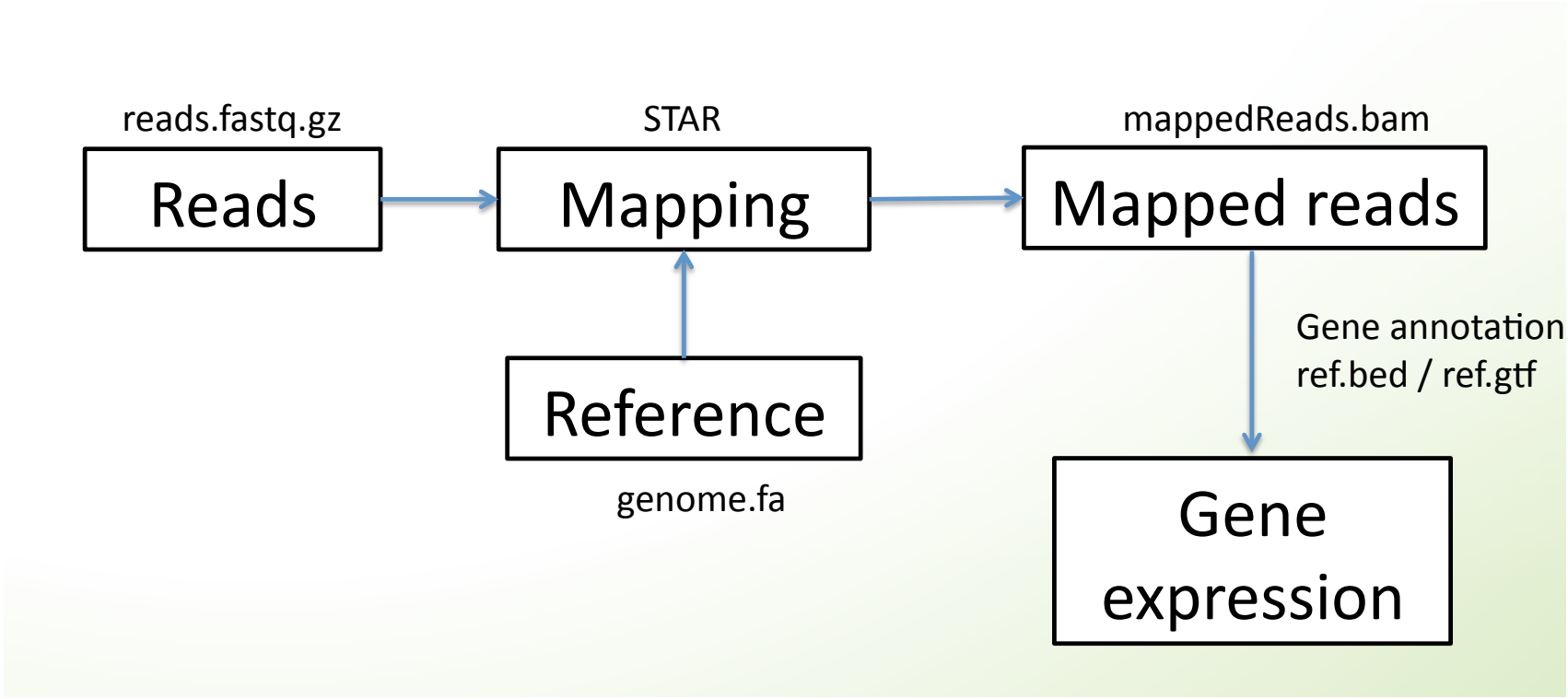
- Low throughput (-)
- Complete transcripts (++)
- Not quantitative (-)
- Only highly expressed genes (-)
- Expensive (-)
- Low background noise (+)
- Easy downstream analysis (++)

short reads

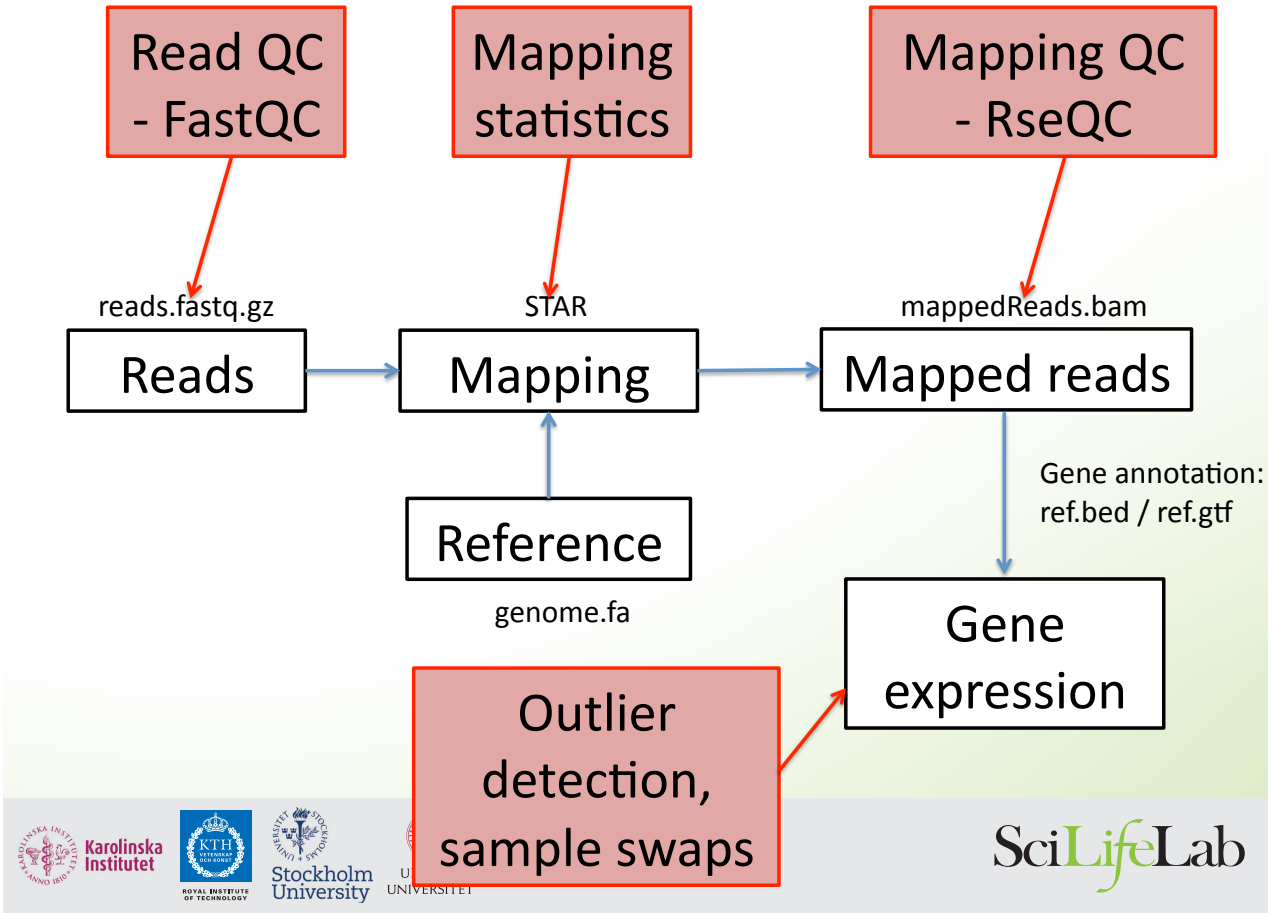
- High throughput (++)
- Quantitative (++)
- Fractions of transcripts (-)
- Full dynamic range (+-)
- Unlimited dynamic range (+)
- Cheap (+)
- Low background noise (+)
- Strand specificity (+)
- Re-sequencing (+)



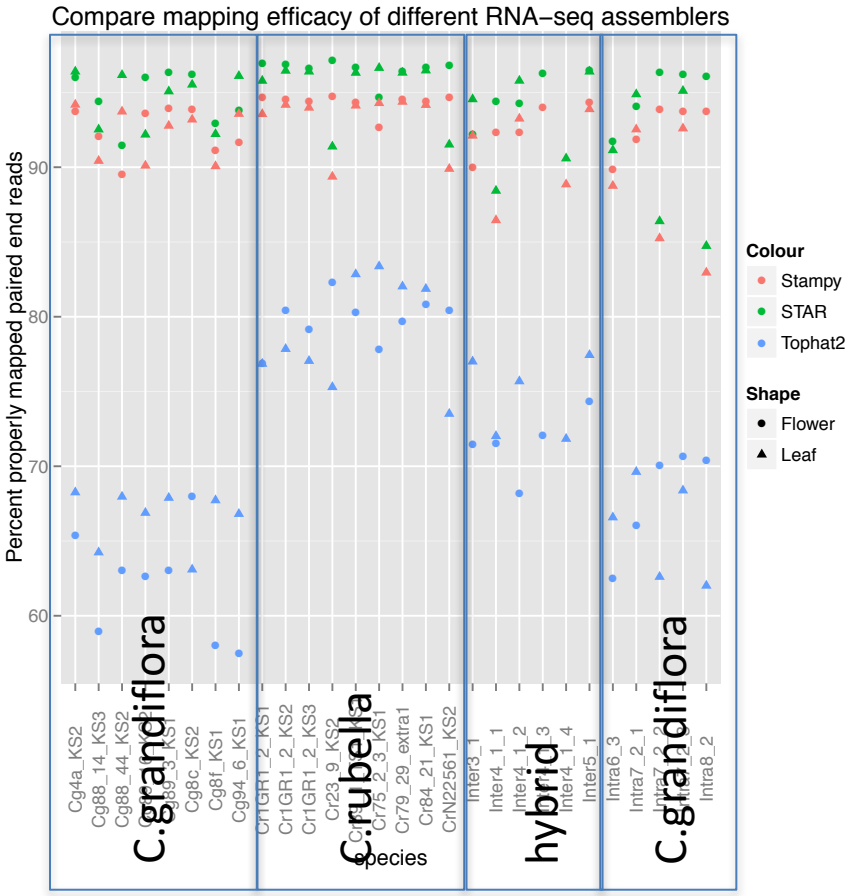
RNA-seq analysis workflow



Do a lot of QC

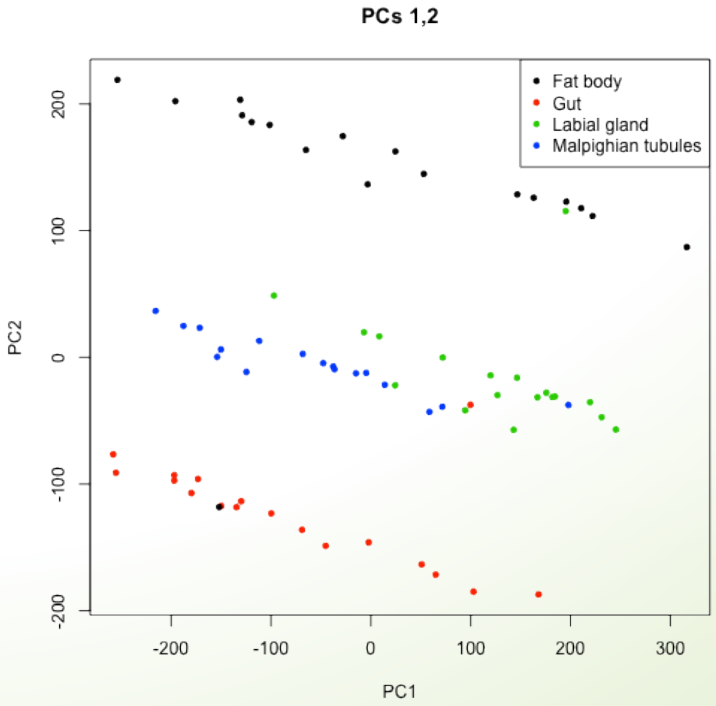

Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

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

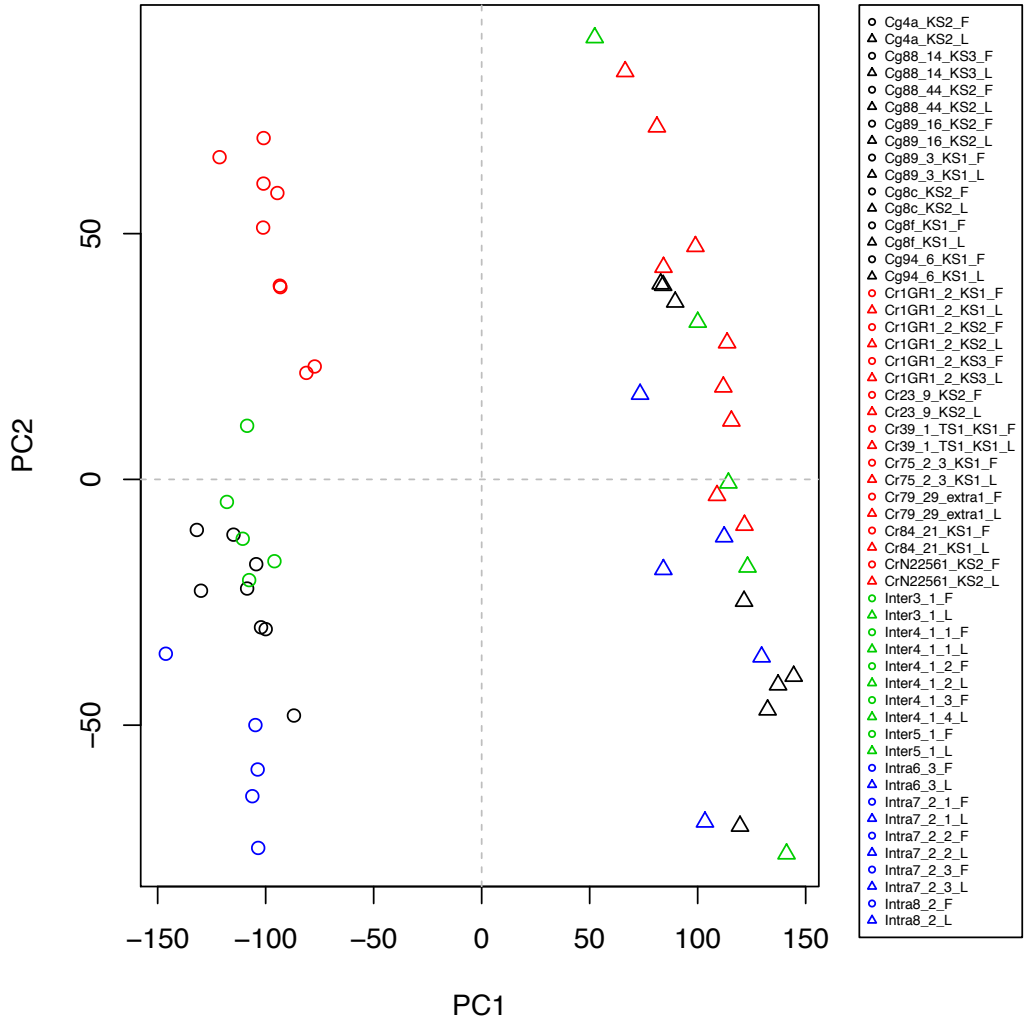


RNA QC

PCA analysis detected potential sample swaps



Principal component 1 separates samples from flowers and leaves



Expression levels are similar between RT-qPCR and RNA-seq data

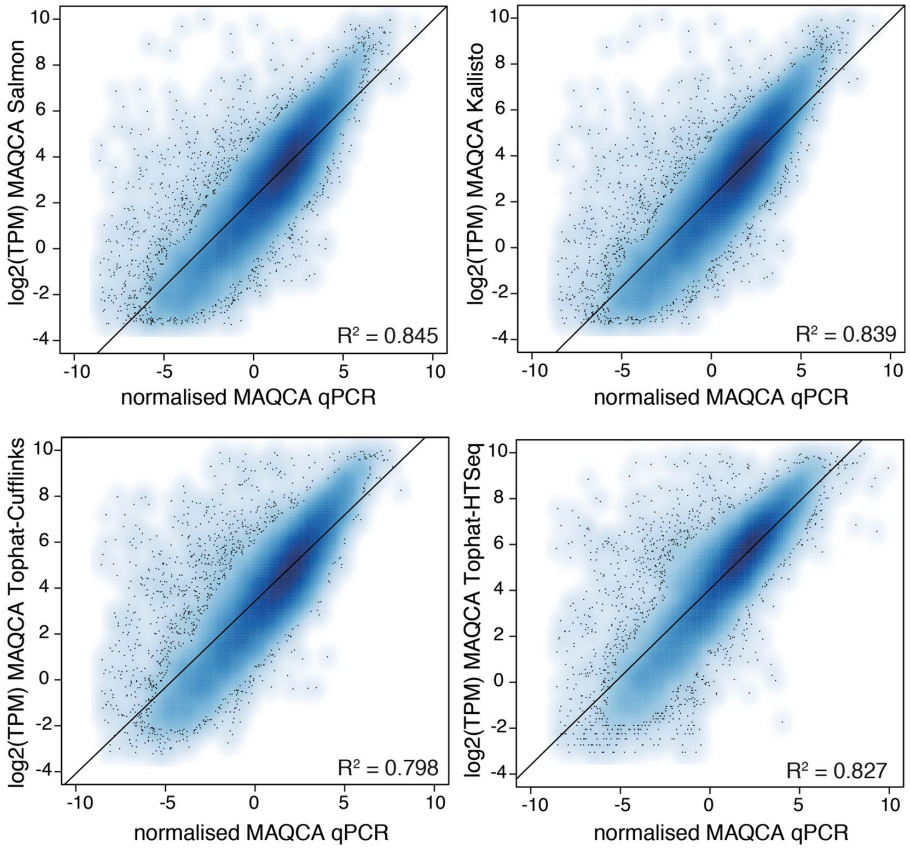
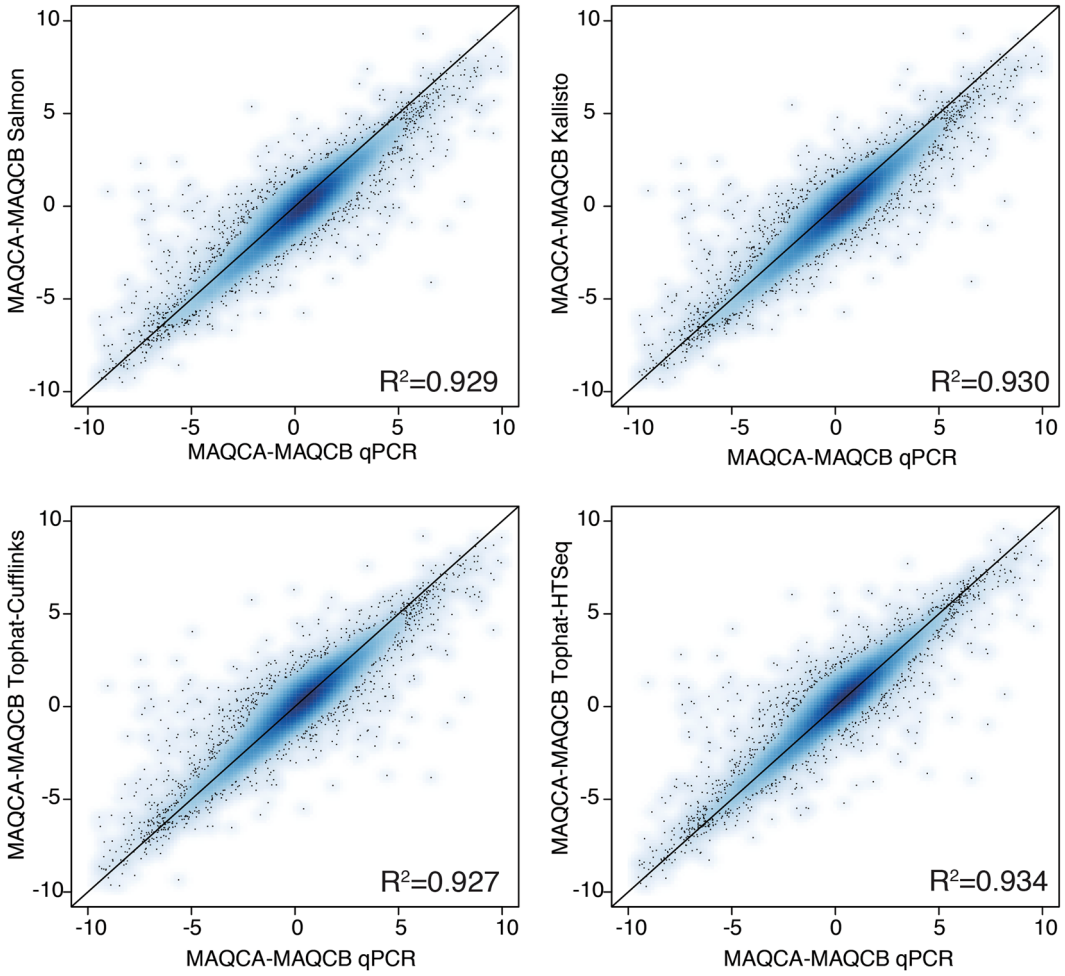
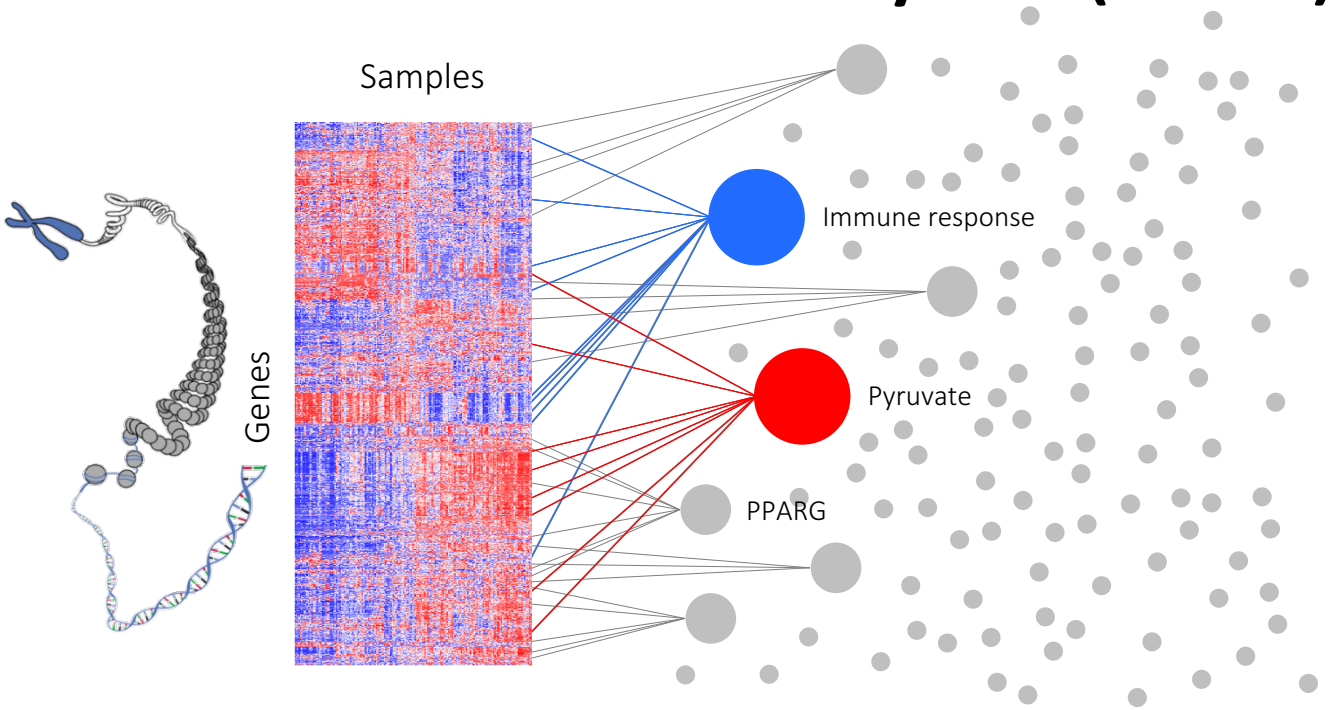


Figure 1. Gene expression correlation between RT-qPCR and RNA-seq data. The Pearson correlation coefficients and linear regression line are indicated. Results are based on RNA-seq data from dataset 1.

Most problems are consistent so they disappear when you do diff-exp analysis



Gene-set analysis (GSA)

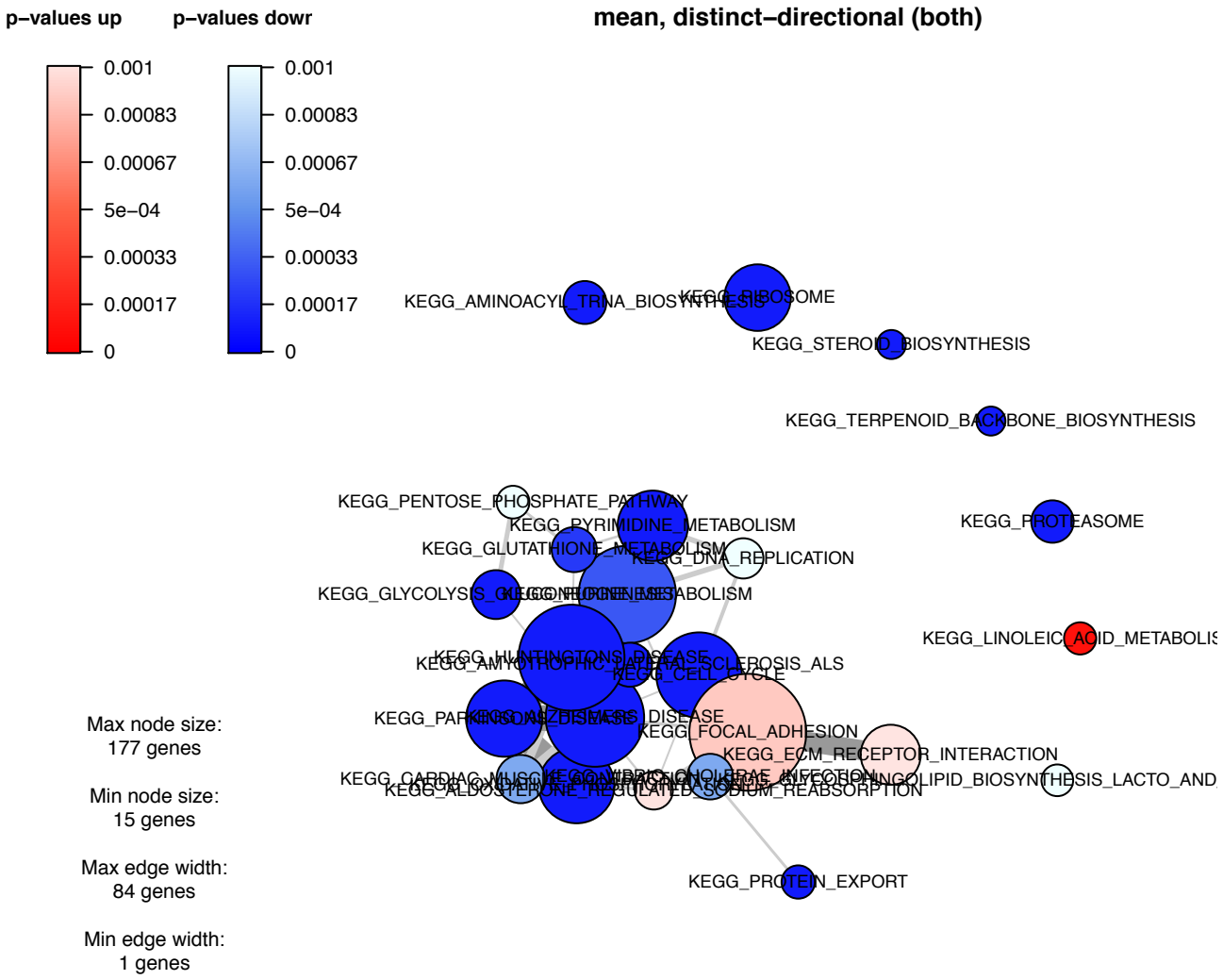


GO-terms
Pathways
Chromosomal locations
Transcription factors
Histone modifications
Diseases
etc…

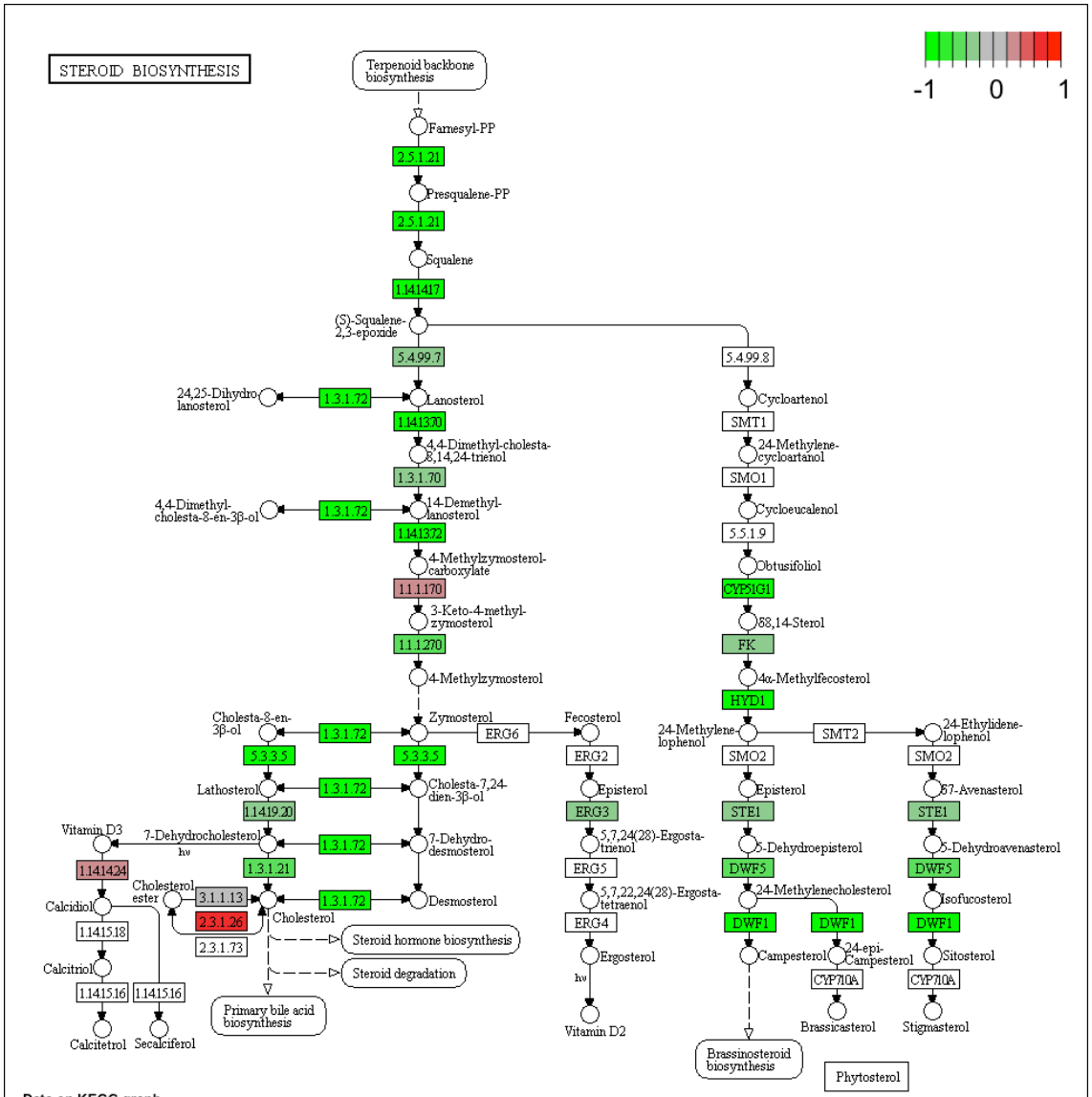
Gene-level data → Gene-set analysis → Gene-set data (results)

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

Analysis regarding Type II Diabetes



Expression of genes on pathway



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.