Gene Expression and RNA-Seq

Cassie Raker

Overview

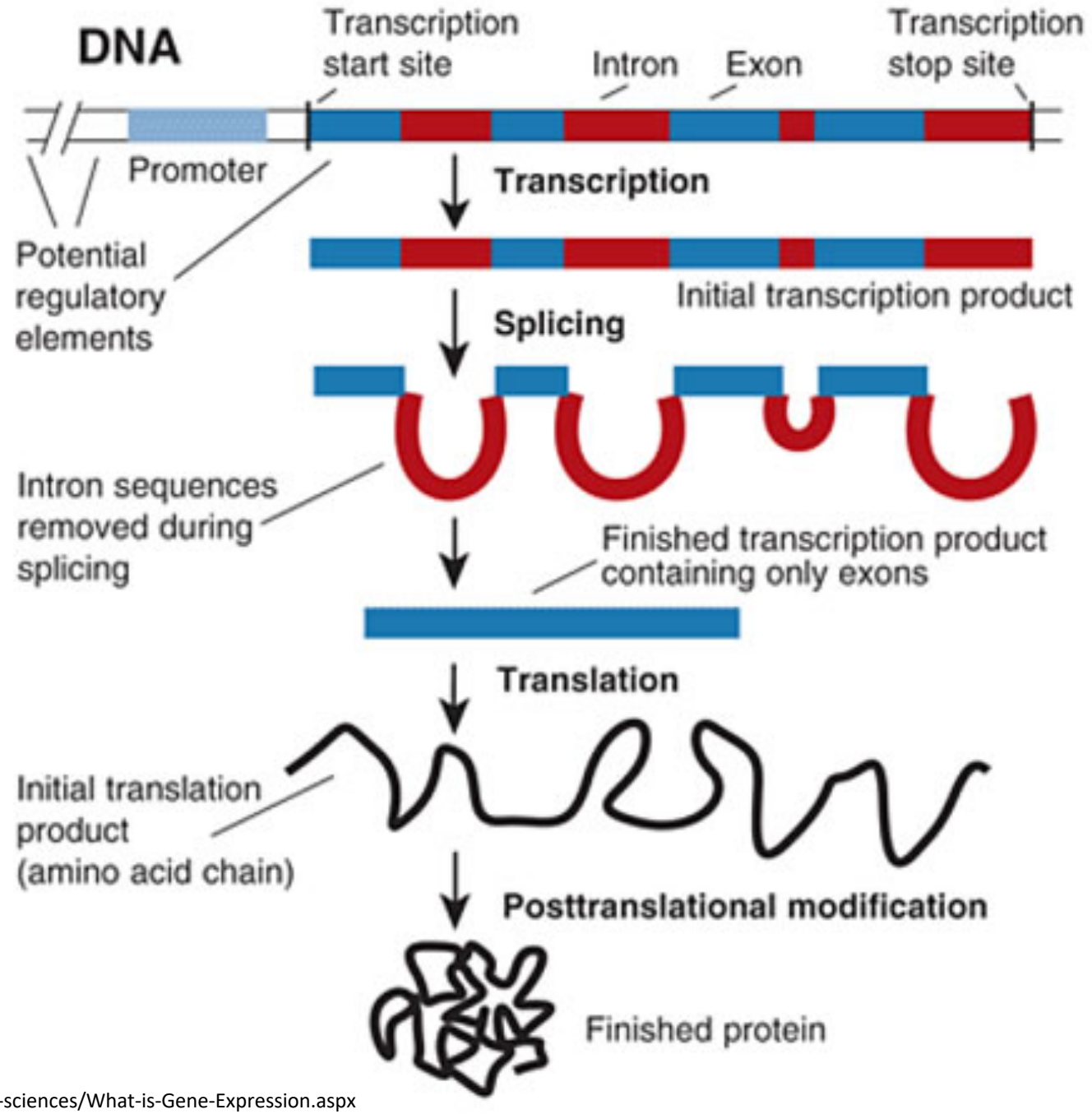
- Gene Expression
- RNA-seq
- Data integration
- Moving forward

Overview

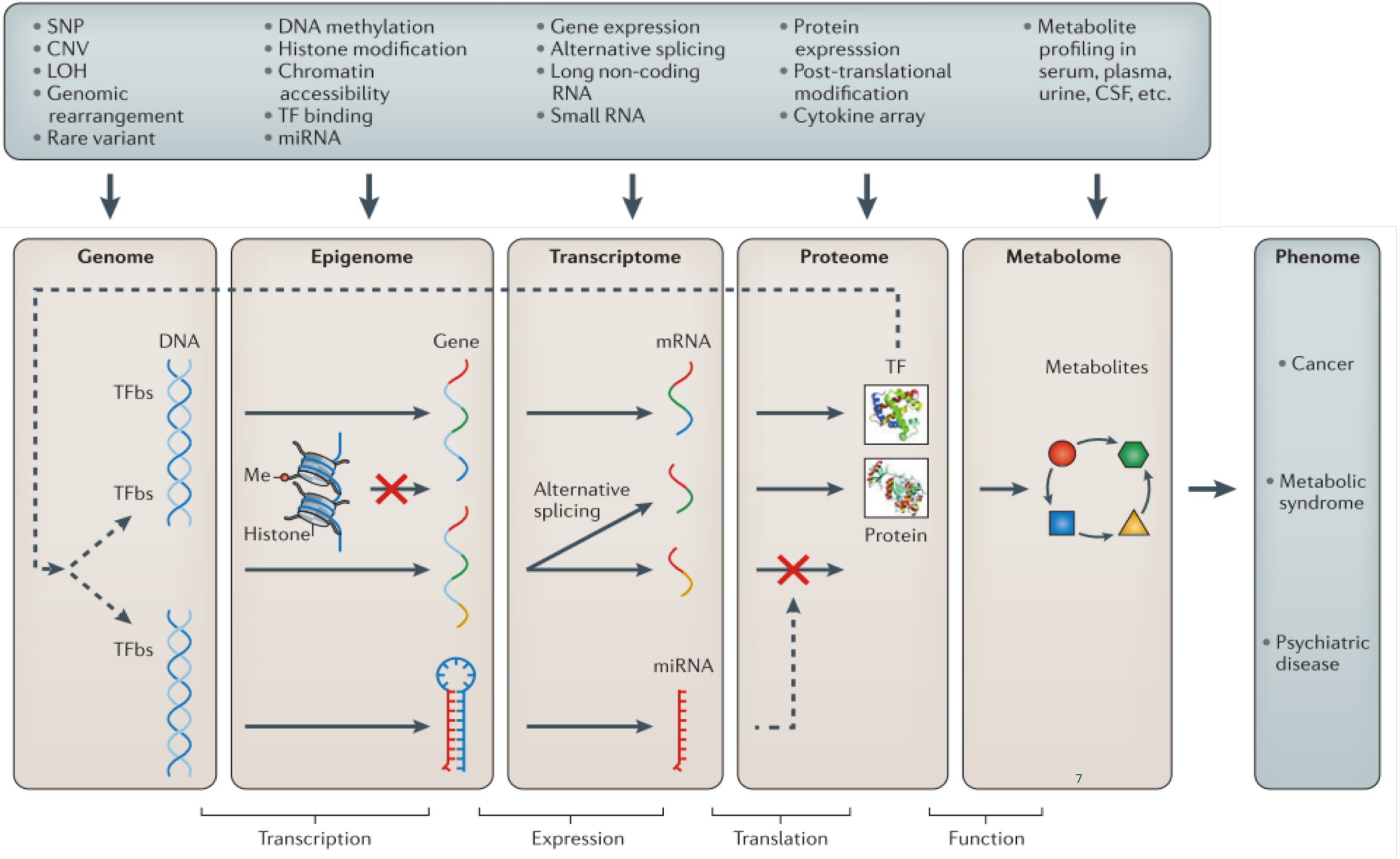
- **Gene Expression**
- RNA-seq
- Data integration
- Moving forward

Gene Expression

Structural variation -> gene expression



https://www.news-medical.net/life-sciences/What-is-Gene-Expression.aspx

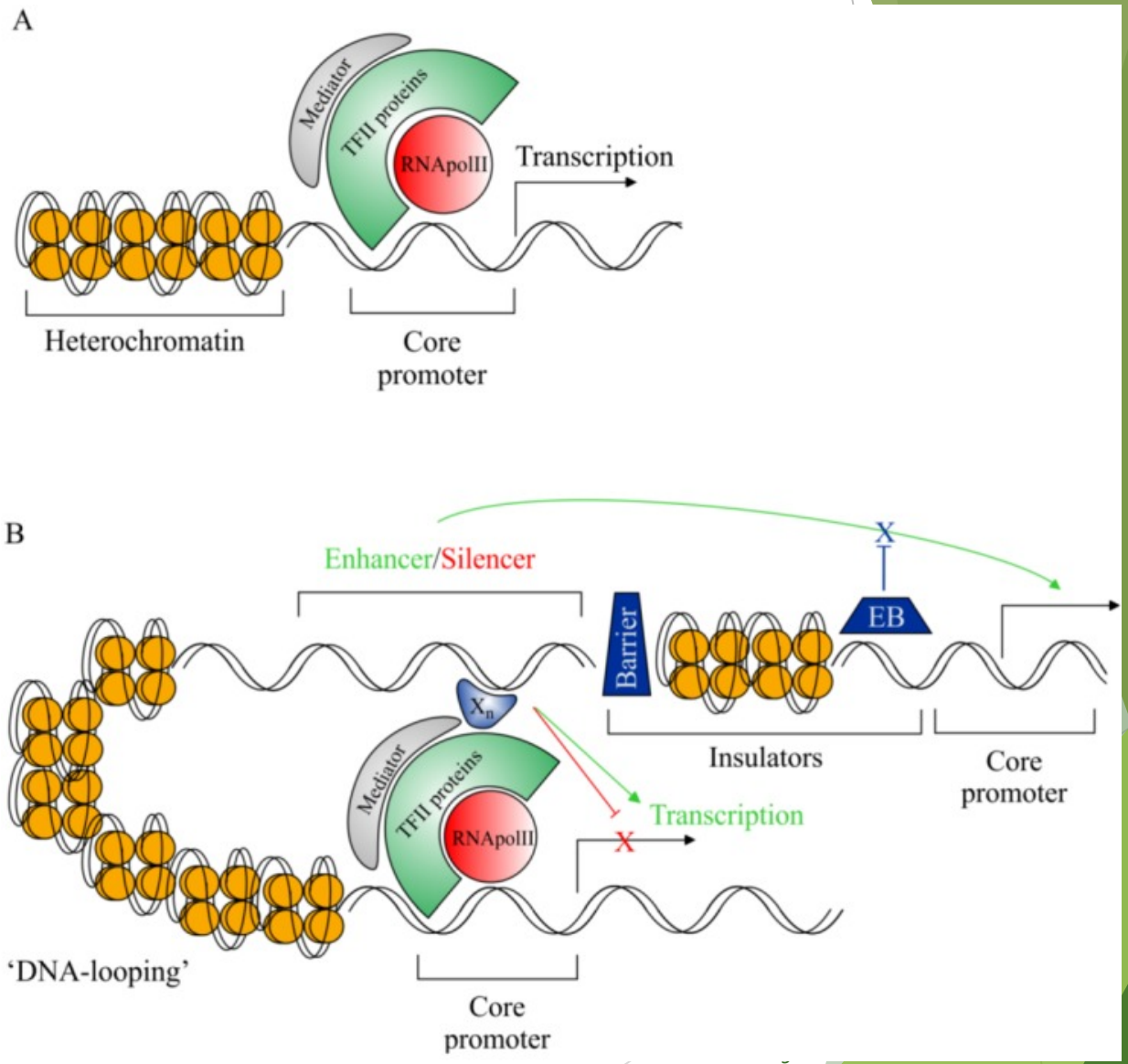

• SNP
• CNV
• LOH
• Genomic rearrangement
• Rare variant
• DNA methylation
• Histone modification
• Chromatin accessibility
• TF binding
• miRNA
• Gene expression
• Alternative splicing
• Long non-coding RNA
• Small RNA
• Protein expresssion
• Post-translational modification
• Cytokine array
• Metabolite profiling in serum, plasma, urine, CSF, etc.
Genome
DNA
TFbs
TFbs
TFbs
Epigenome
Gene
Me
Histone
Transcriptome
mRNA
Alternative splicing
miRNA
Proteome
TF
Protein
Metabolome
Metabolites
Phenome
• Cancer
• Metabolic syndrome
• Psychiatric disease
Transcription
Expression
Translation
Function

Gene Expression

- Transcription
- Translation

Gene Expression

- **Transcription**
  - RNA polymerase
  - Promoters
  - Transcription factors
- Translation

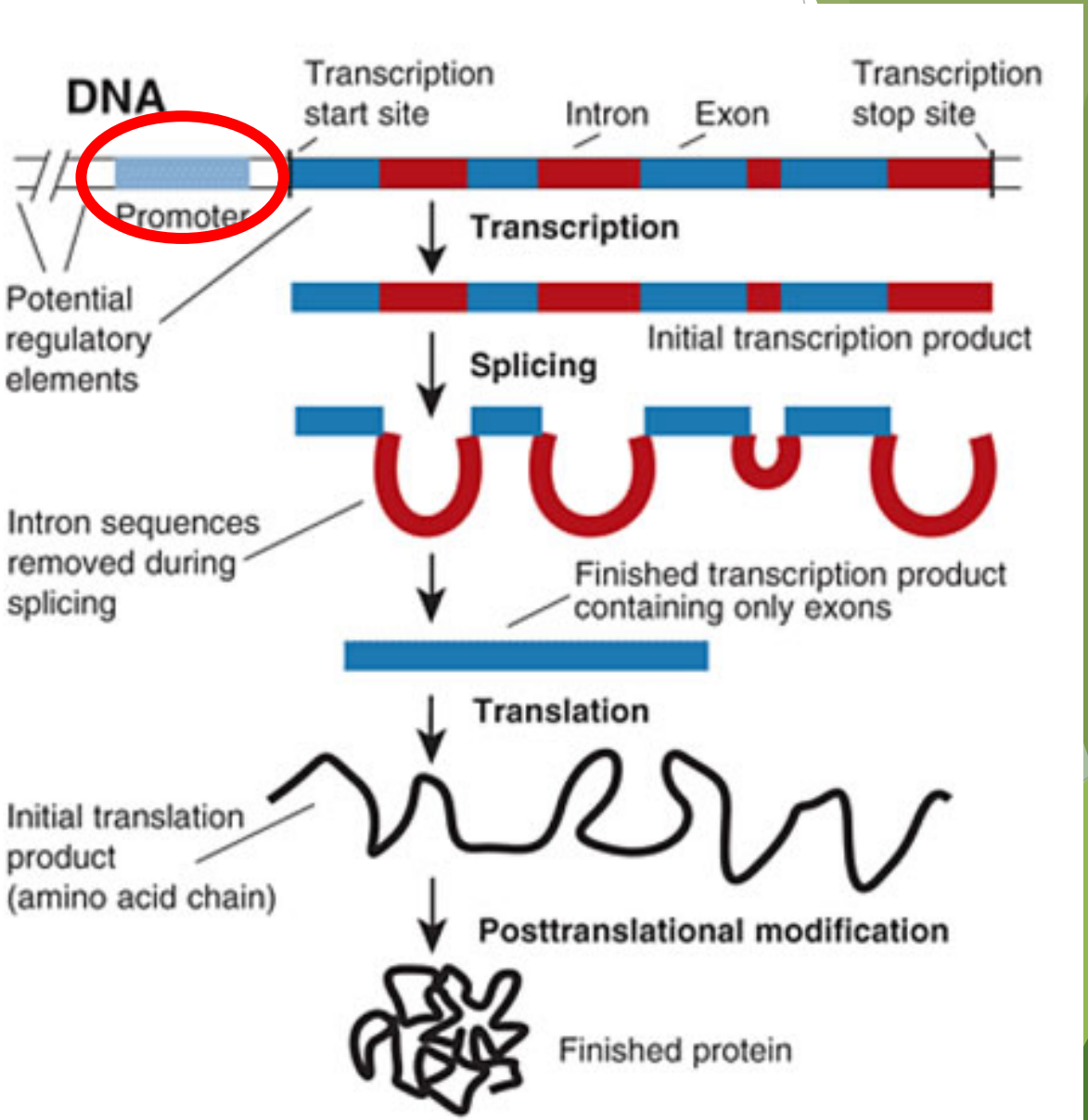


https://www.researchgate.net/figure/Graphic-representation-of-eukaryotic-transcriptional-machinery-A-Basal-eukaryotic_fig1_262384542

Promoters

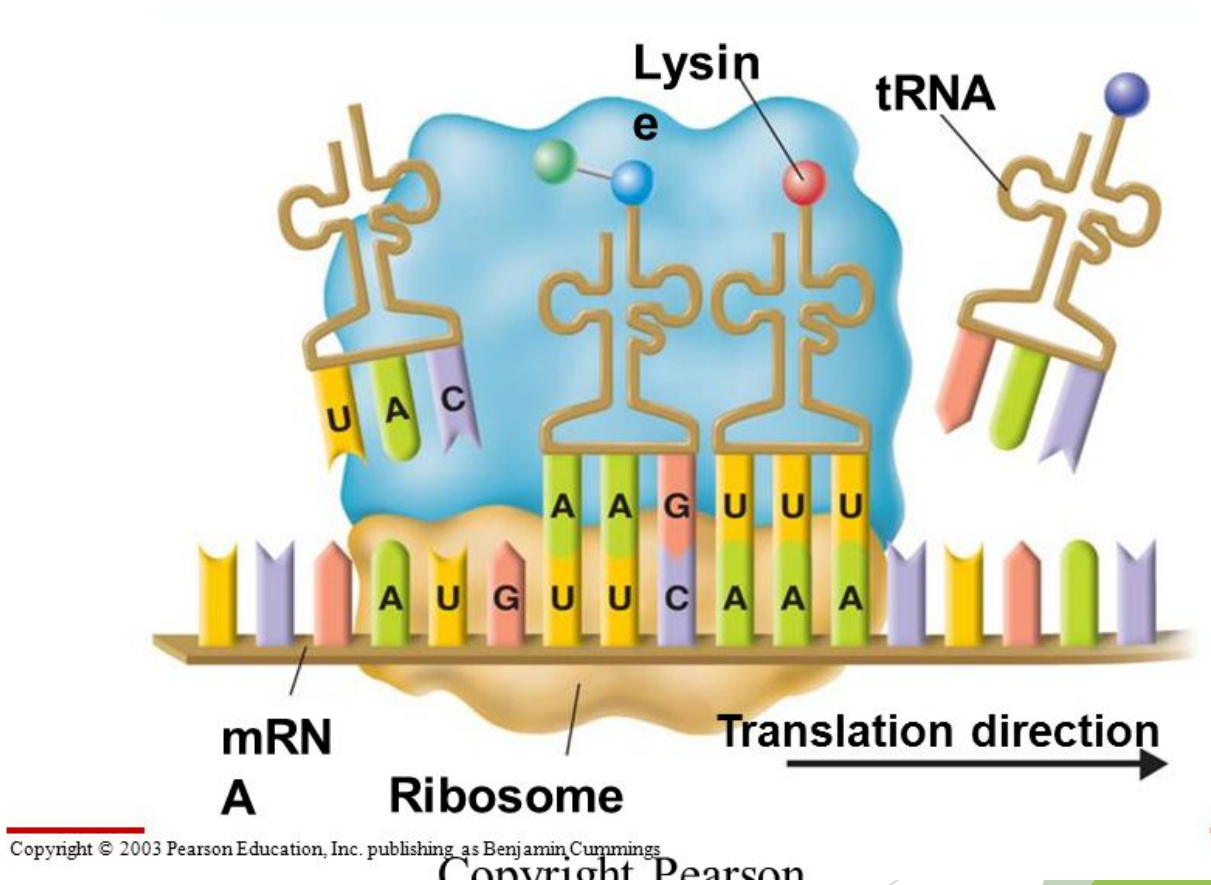
- Region of DNA
- Next to gene in question
- Initiates transcription

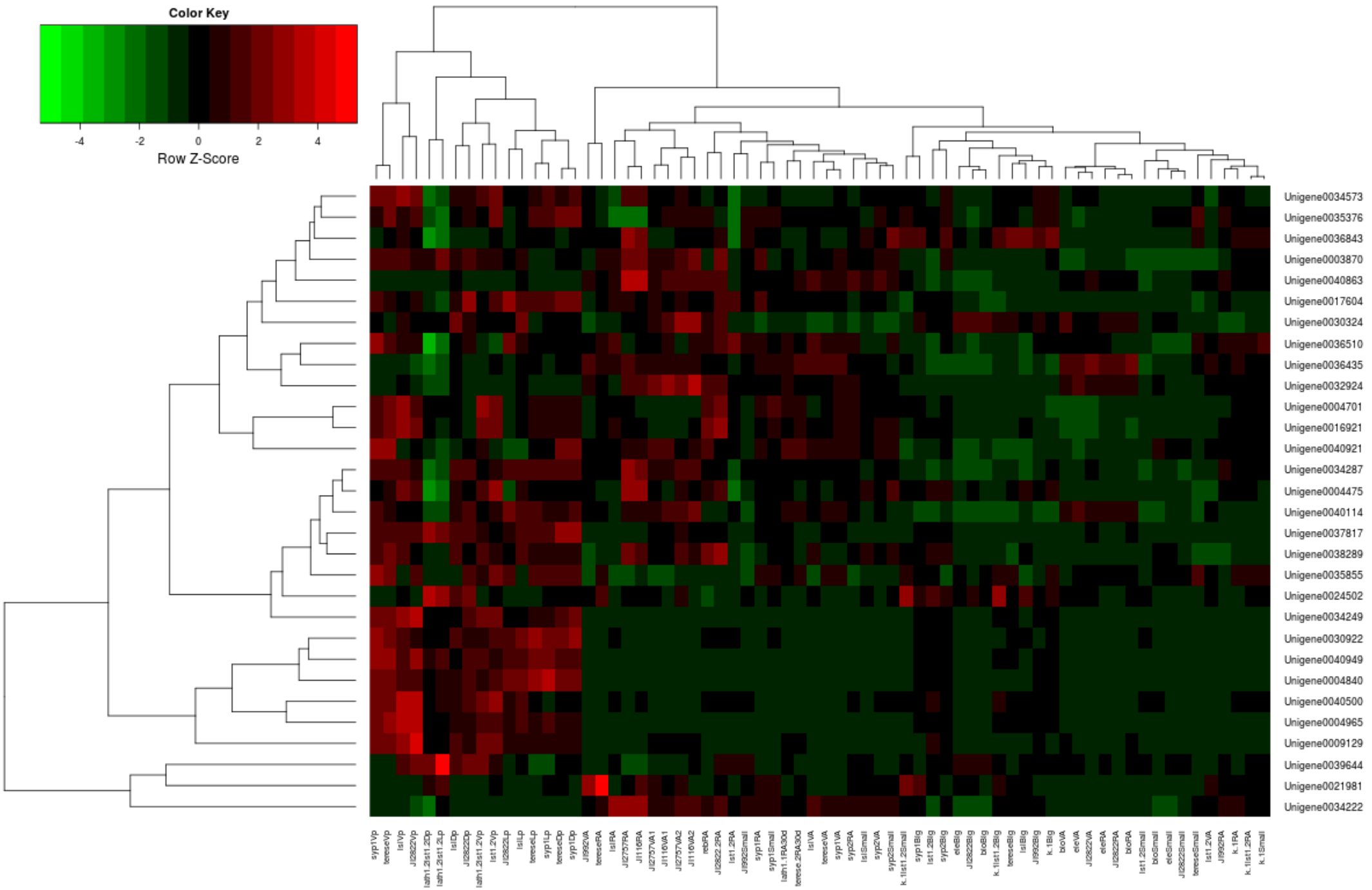
- Core promoter
- Proximal promoter
- Distal promoter



Gene Expression

- Transcription
- **Translation**
  - Initiation
  - Elongation
  - Termination





http://genedenovoweb.ticp.net:81/rsia/index.php?m=tools&f=heatmap

What can go wrong?

Overview

- Gene Expression
- **RNA-seq**
- Data integration
- Moving forward

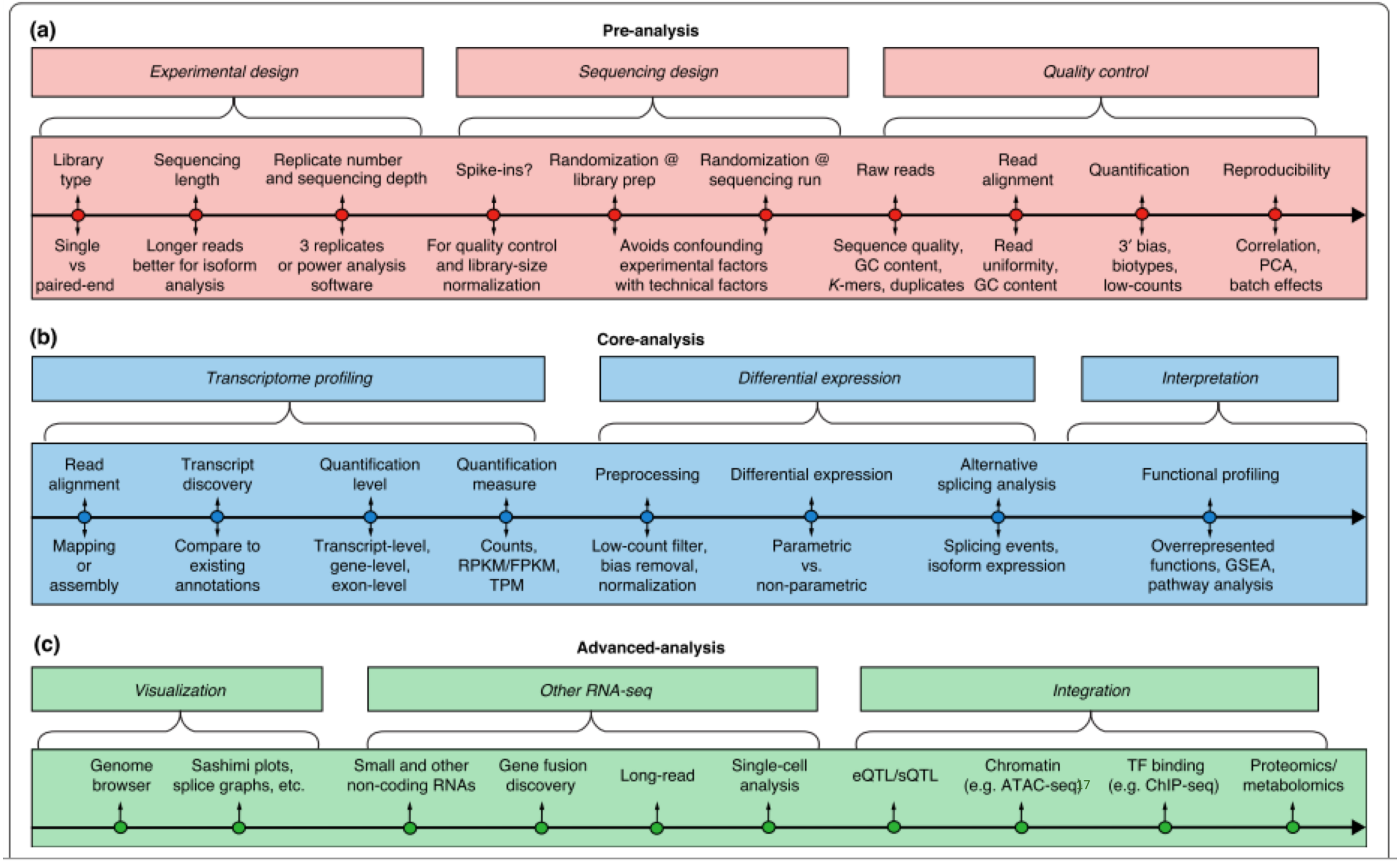
What is it anyway?

- What RNA is in this sample right now?

- Next-generation sequencing techniques
- Analyze gene expression!
  - Alternative to microarrays

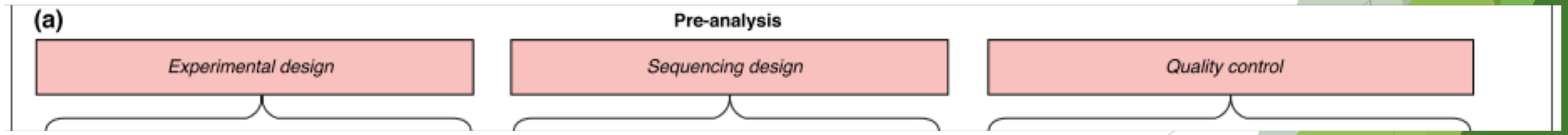
RNA-seq

- Pre-analysis
- Core-analysis
- Advanced-analysis


(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Sequencing length
Replicate number and sequencing depth
Spike-ins?
Randomization @ library prep
Randomization @ sequencing run
Raw reads
Read alignment
Quantification
Reproducibility
Single vs paired-end
Longer reads better for isoform analysis
3 replicates or power analysis software
For quality control and library-size normalization
Avoids confounding experimental factors with technical factors
Sequence quality, GC content, K-mers, duplicates
Read uniformity, GC content
3′ bias, biotypes, low-counts
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Transcript discovery
Quantification level
Quantification measure
Preprocessing
Differential expression
Alternative splicing analysis
Functional profiling
Mapping or assembly
Compare to existing annotations
Transcript-level, gene-level, exon-level
Counts, RPKM/FPKM, TPM
Low-count filter, bias removal, normalization
Parametric vs. non-parametric
Splicing events, isoform expression
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics

RNA-seq

- **Pre-analysis**
  - Experimental design
  - Sequencing design
  - Quality control
- Core analysis
- Advanced analysis

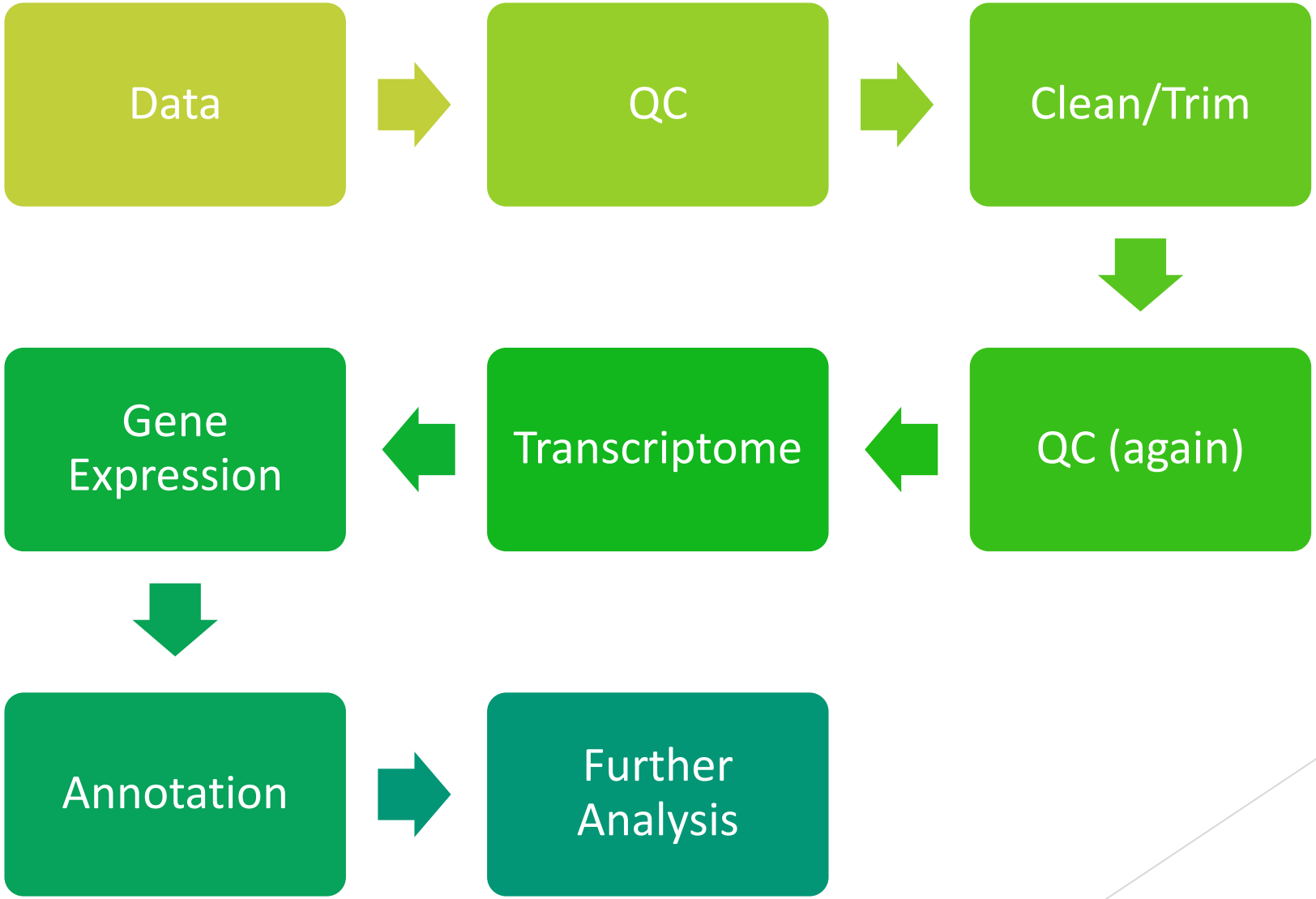


What do you need to consider?

- How will you extract RNA?
  - Interested in mRNA: enrich mRNA or deplete rRNA?
- Library type/sequencing depth
- Number of replicates
- Genome
- Transcriptome

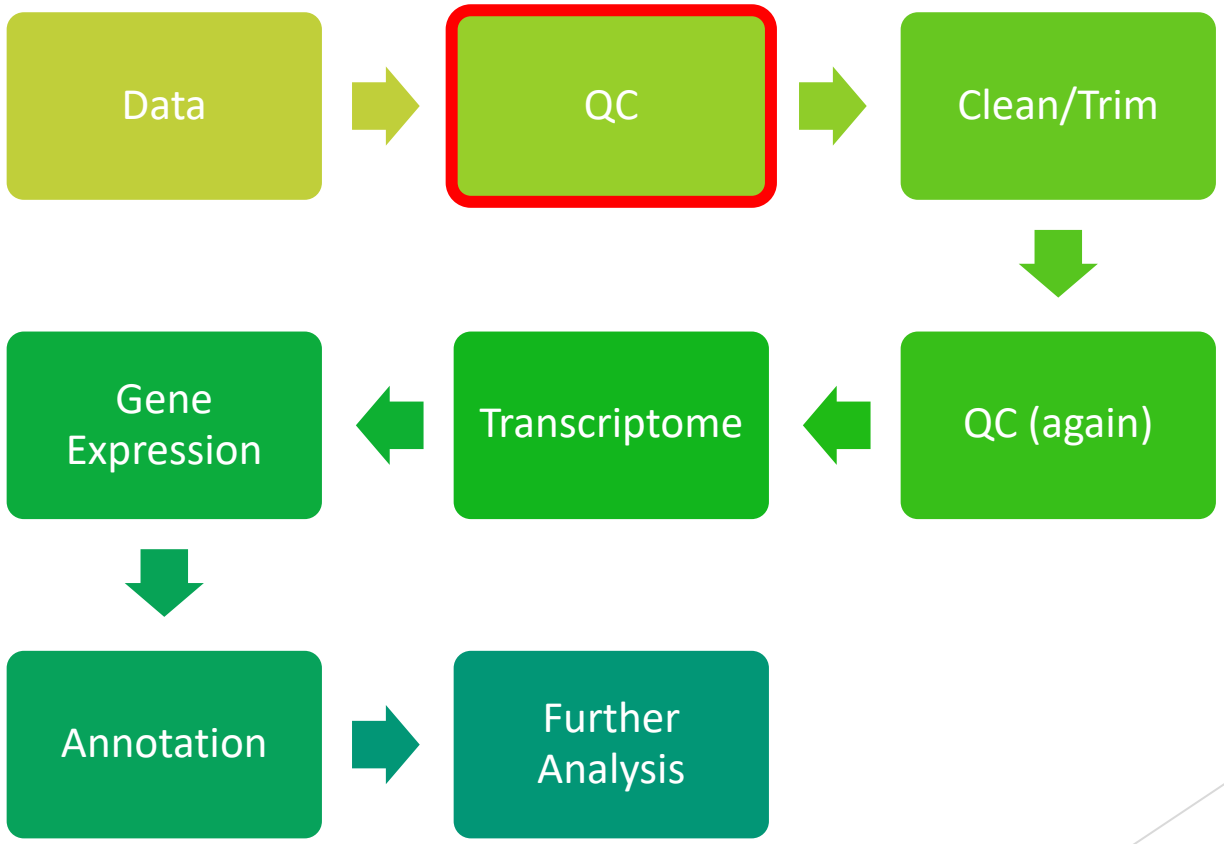
```
(base) [craker@KITT data]$ ls
backup
DA-HI-A_S79_L007_R1_001.fastq.gz
DA-HI-B_S80_L007_R1_001.fastq.gz    OADA0006_S16_L002_R1_001.fastq.gz
DA-HI-C_S81_L007_R1_001.fastq.gz    OADA0049_S25_L003_R1_001.fastq.gz
DA-HI-D_S82_L007_R1_001.fastq.gz    OADA0058_S19_L002_R1_001.fastq.gz
DA-LOW-A_S72_L007_R1_001.fastq.gz   OADA0071_S53_L005_R1_001.fastq.gz
DA-LOW-B_S73_L007_R1_001.fastq.gz   OADA0081_S67_L006_R1_001.fastq.gz
DA-LOW-C_S74_L007_R1_001.fastq.gz   OADA0085_S24_L002_R1_001.fastq.gz
DA-LOW-D_S75_L007_R1_001.fastq.gz   OADA0101_S18_L002_R1_001.fastq.gz
DA-MED-A_S76_L007_R1_001.fastq.gz   OADA0102_S33_L003_R1_001.fastq.gz
DA-MED-B_S77_L007_R1_001.fastq.gz   OADA0116_S26_L003_R1_001.fastq.gz
DA-MED-D_S78_L007_R1_001.fastq.gz   OADA0139_S17_L002_R1_001.fastq.gz
docs                                OADA0174_S20_L002_R1_001.fastq.gz
```

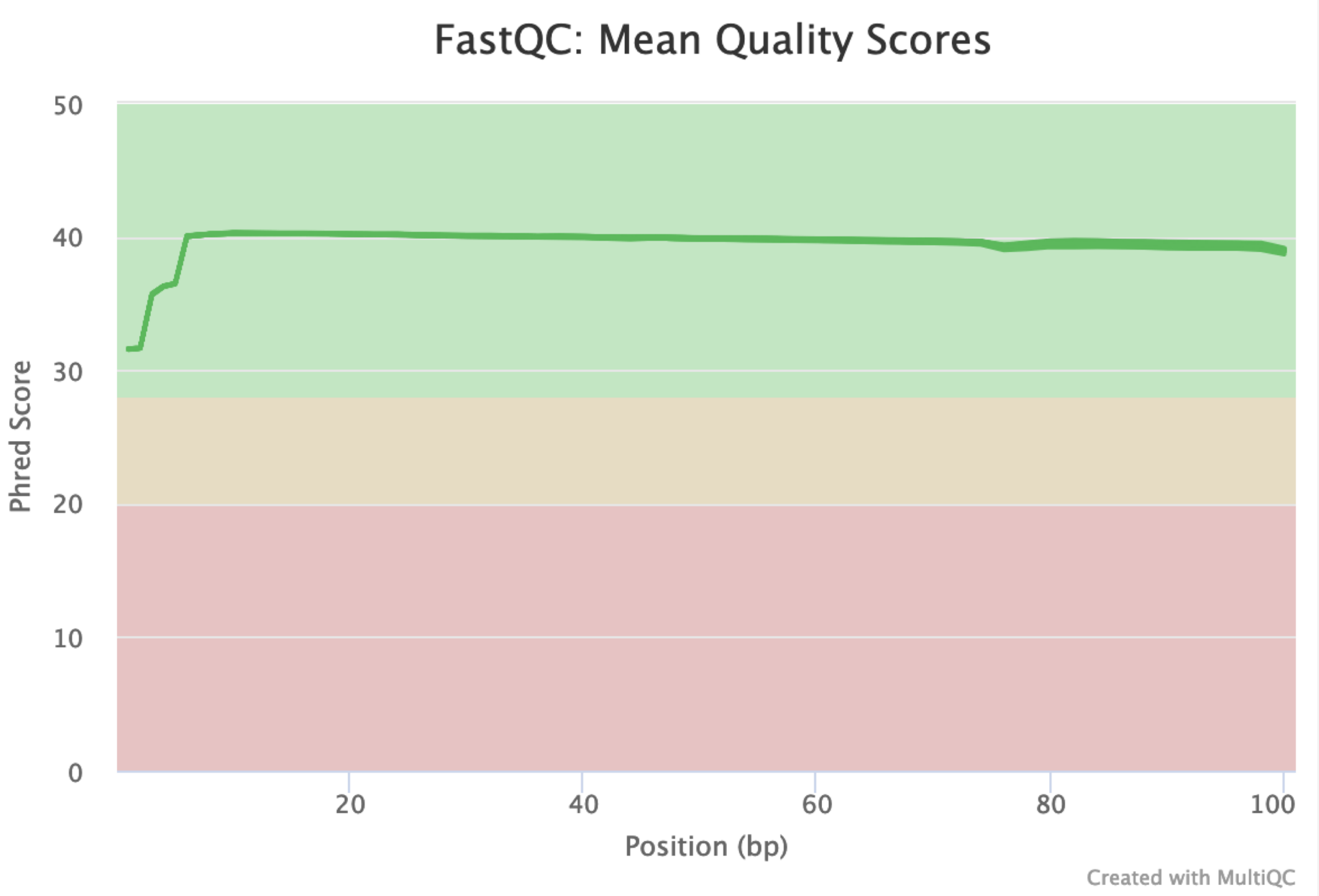
Analysis Pipeline

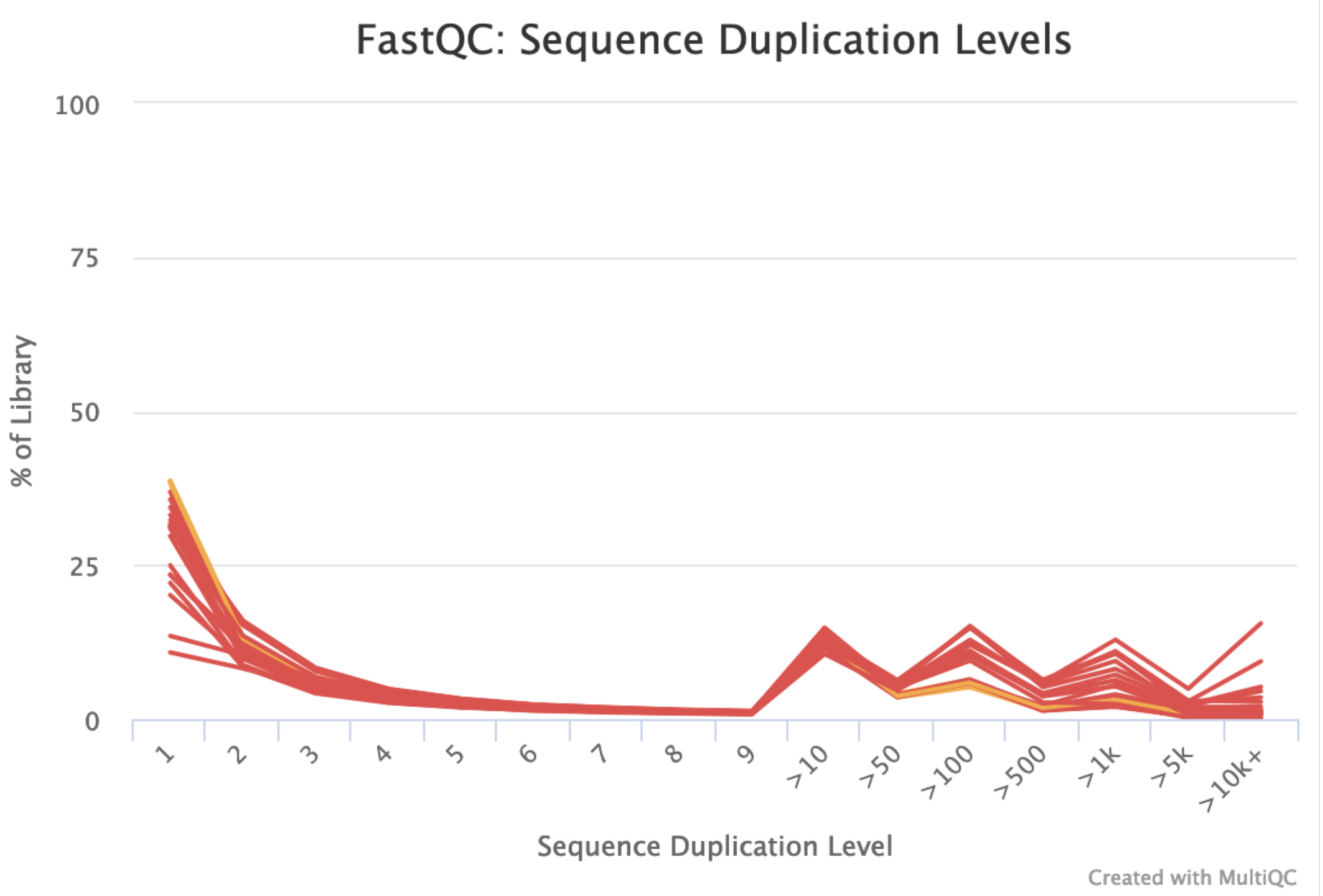


Quality Control

- Raw reads
  - FastQC
  - Trimmomatic
  - Clean rRNA

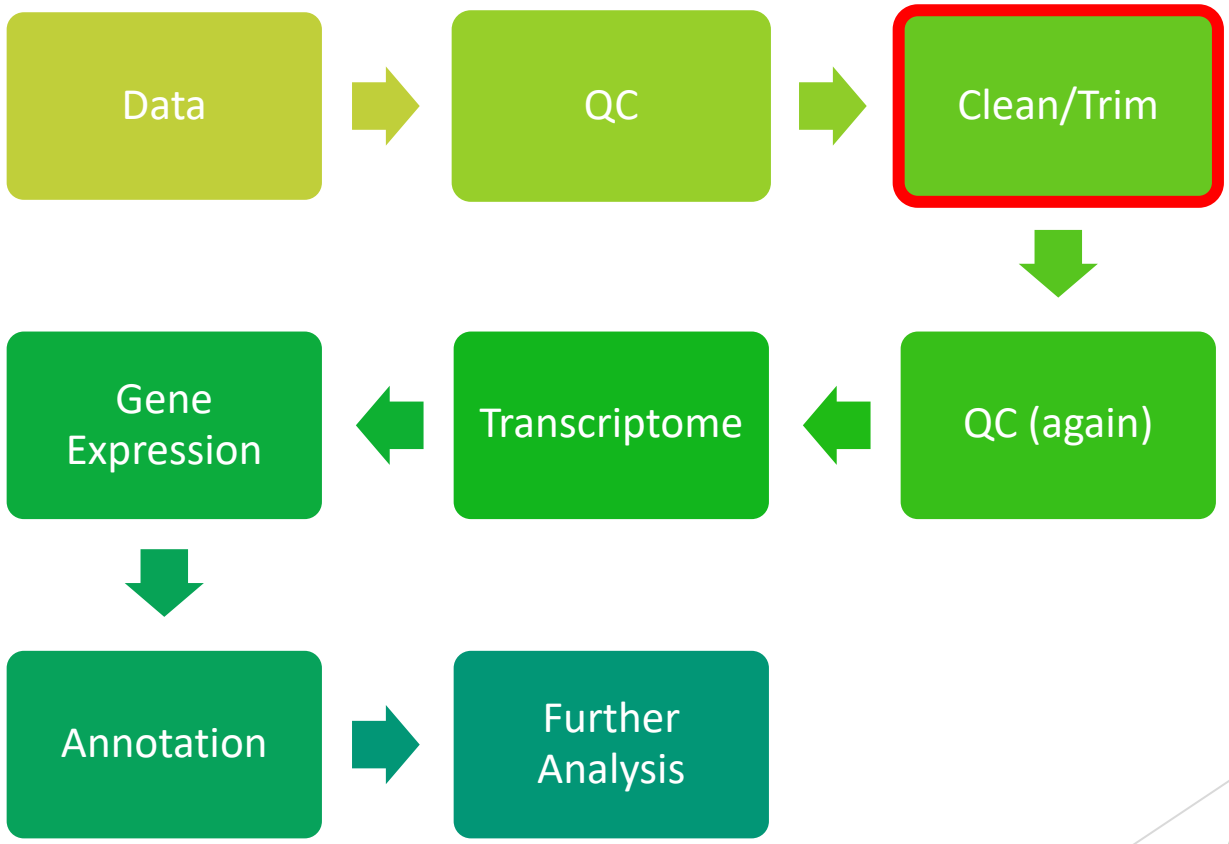



FastQC: Mean Quality Scores
Phred Score
50
40
30
20
10
0
20
40
60
80
100
Position (bp)
Created with MultiQC


FastQC: Sequence Duplication Levels
100
75
50
25
0
% of Library
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k+
Sequence Duplication Level
Created with MultiQC

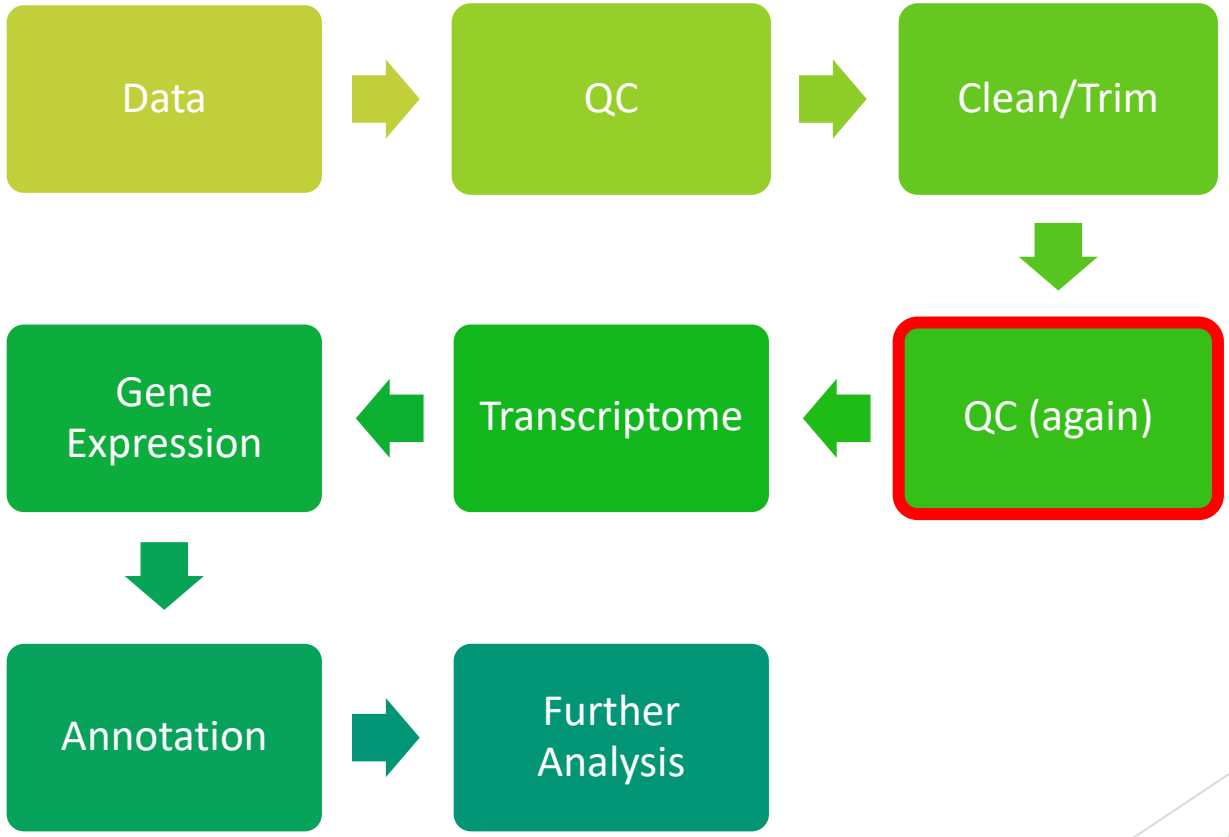
Quality Control

- Raw reads
  - FastQC
  - Trimmomatic
  - Clean rRNA



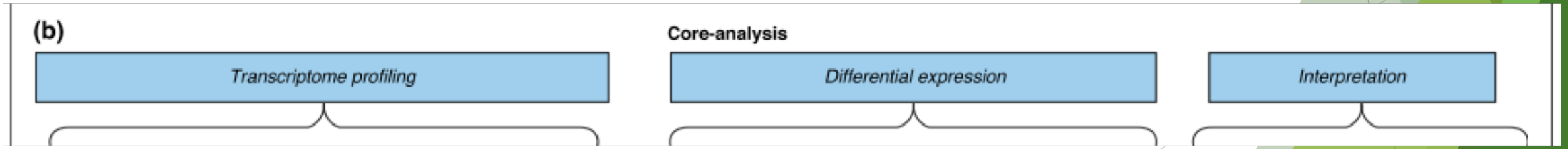
Quality Control

- Raw reads
  - FastQC
  - Trimmomatic
  - Clean rRNA
- Do QC again to check!



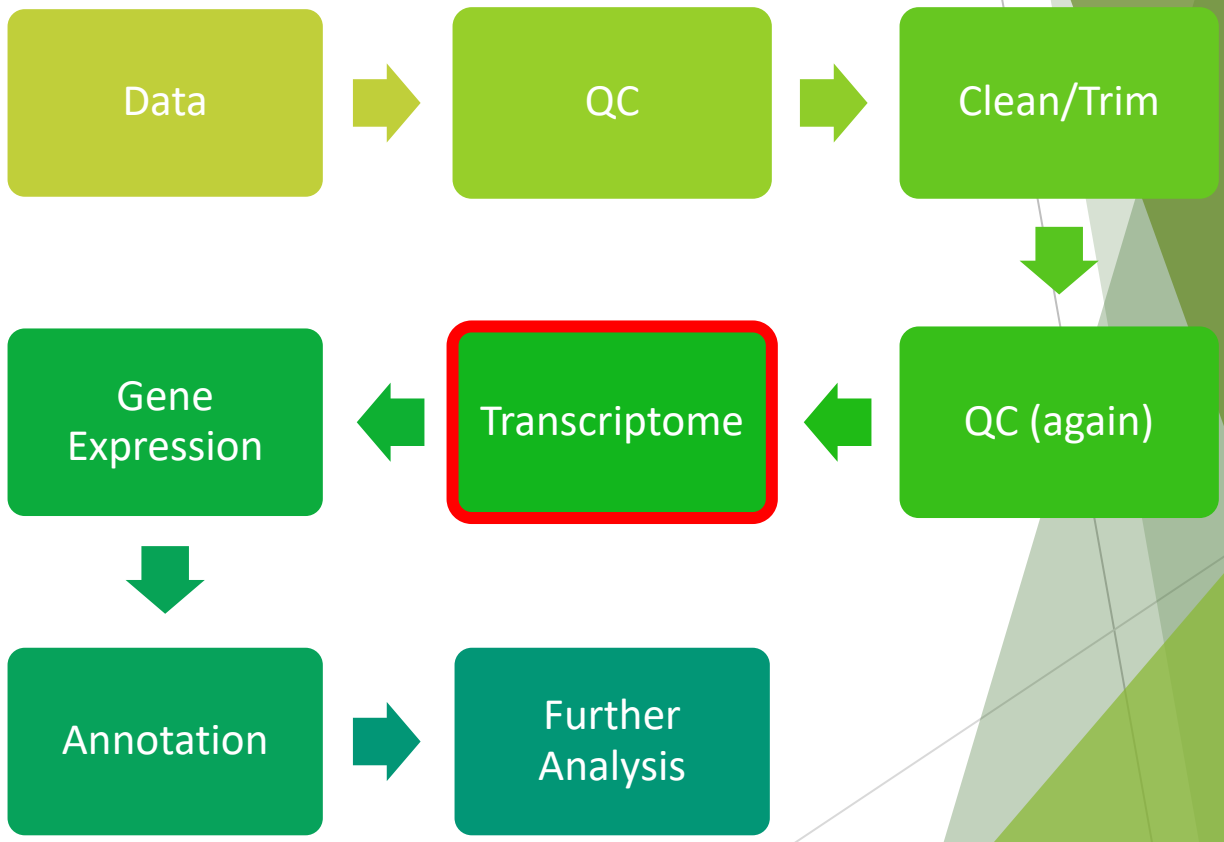
RNA-seq

- Pre-analysis
- **Core-analysis**
  - Transcriptome profiling
  - Differential expression
  - Interpretation
- Advanced-analysis



Transcriptome profiling: De novo transcript reconstruction

- No reference genome is available
- Can use Trinity (among other programs)
  - Trinotate
- Combine all reads from multiple samples
- End up with a matrix



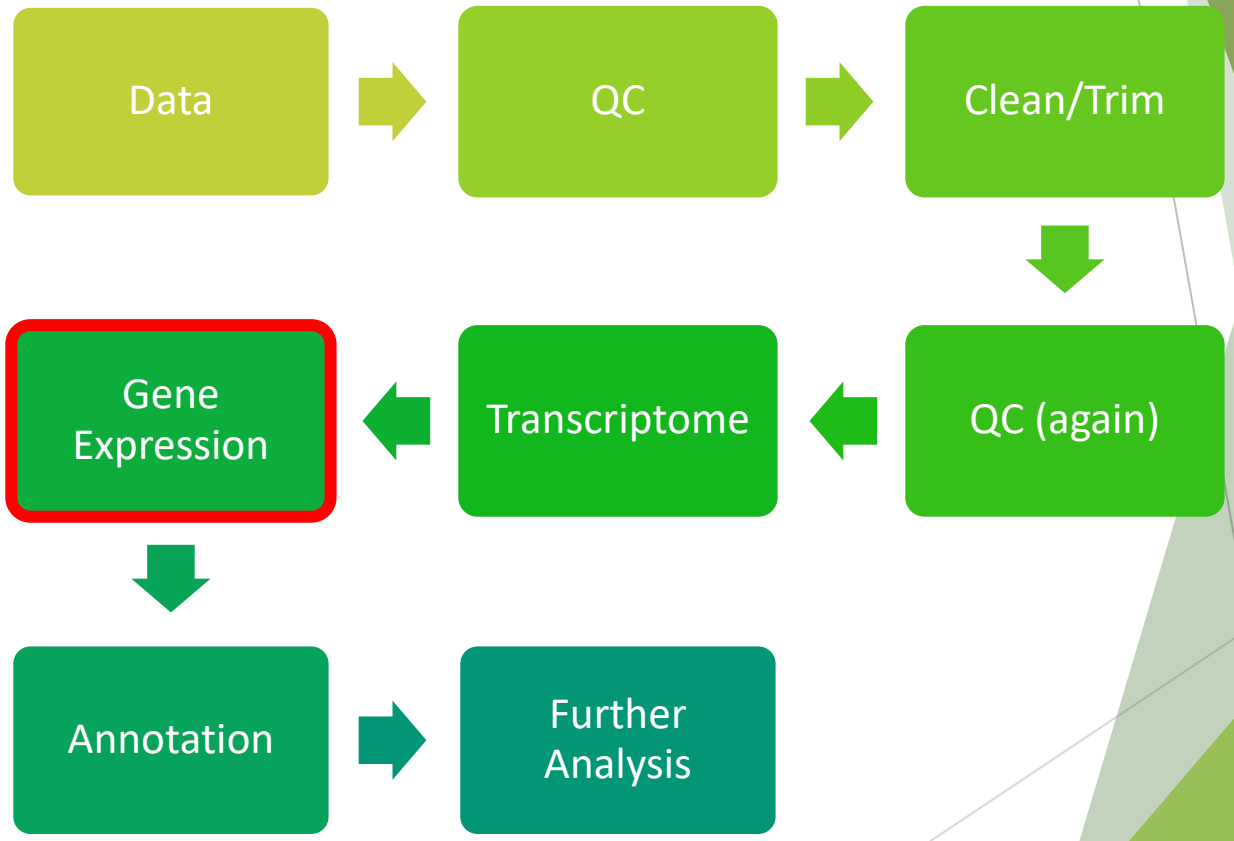
Differential gene expression analysis

Gene Expression

- Gene expression values compared among samples
- RPKM, FPKM, TPM
  - Normalize sequencing depth
  - Work best with samples with homogeneous transcript distributions
  - Rely on total or effective counts
- TMM, DESeq, UpperQuartile
  - Better for heterogeneous transcript distributions
- Remember to watch out for batch effects!
  - Use COMBAT or ARSyN

…time for R!

- Statistically determine which genes are differentially expressed
- EdgeR
- DSeq2



GENEONTOLOGY
Unifying Biology

Annotation

- Grouping genes by their functional characteristics
- Collapsing genes into clusters

- “[Provide a] computational representation of our current scientific knowledge about the functions of genes (or, more properly, the protein and non-coding RNA molecules produced by genes) from many different organisms, from humans to bacteria.”

The Gene Ontology

- Molecular function
- Cellular component
- Biological process

GO Term Elements

- Unique identifier and term name
  - Human-readable term name
  - 7 digit identifier
- Aspect
  - Molecular, cellular, or biological
- Definition
  - Textual description
  - References/sources
- Relationships to other terms

Term Information

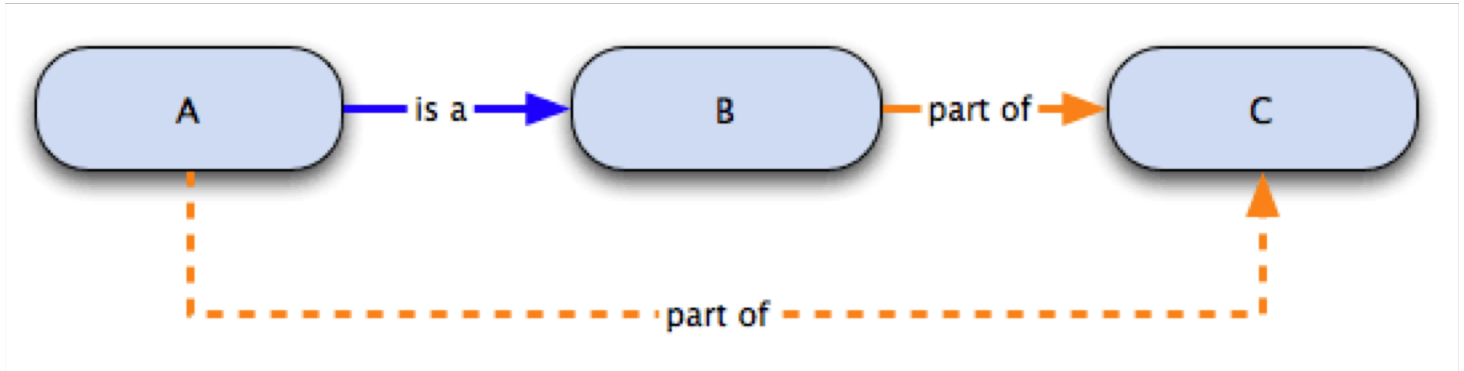
Accession	GO:1904659
Name	glucose transmembrane transport
Ontology	biological_process
Synonyms	glucose transport
Alternate IDs	GO:0015758
Definition	The process in which glucose is transported across a membrane. *Source:* PMID:9090050, GOC:TermGenie, GO_REF:0000069
Comment	None
History	See term history for GO:1904659 at QuickGO
Subset	None
Related	Link to all **genes and gene products** annotated to glucose transmembrane transport.
	Link to all direct and indirect **annotations** to glucose transmembrane transport.
	Link to all direct and indirect **annotations download** (limited to first 10,000) for glucose transmembrane transport.

http://amigo.geneontology.org/amigo/term/GO:1904659

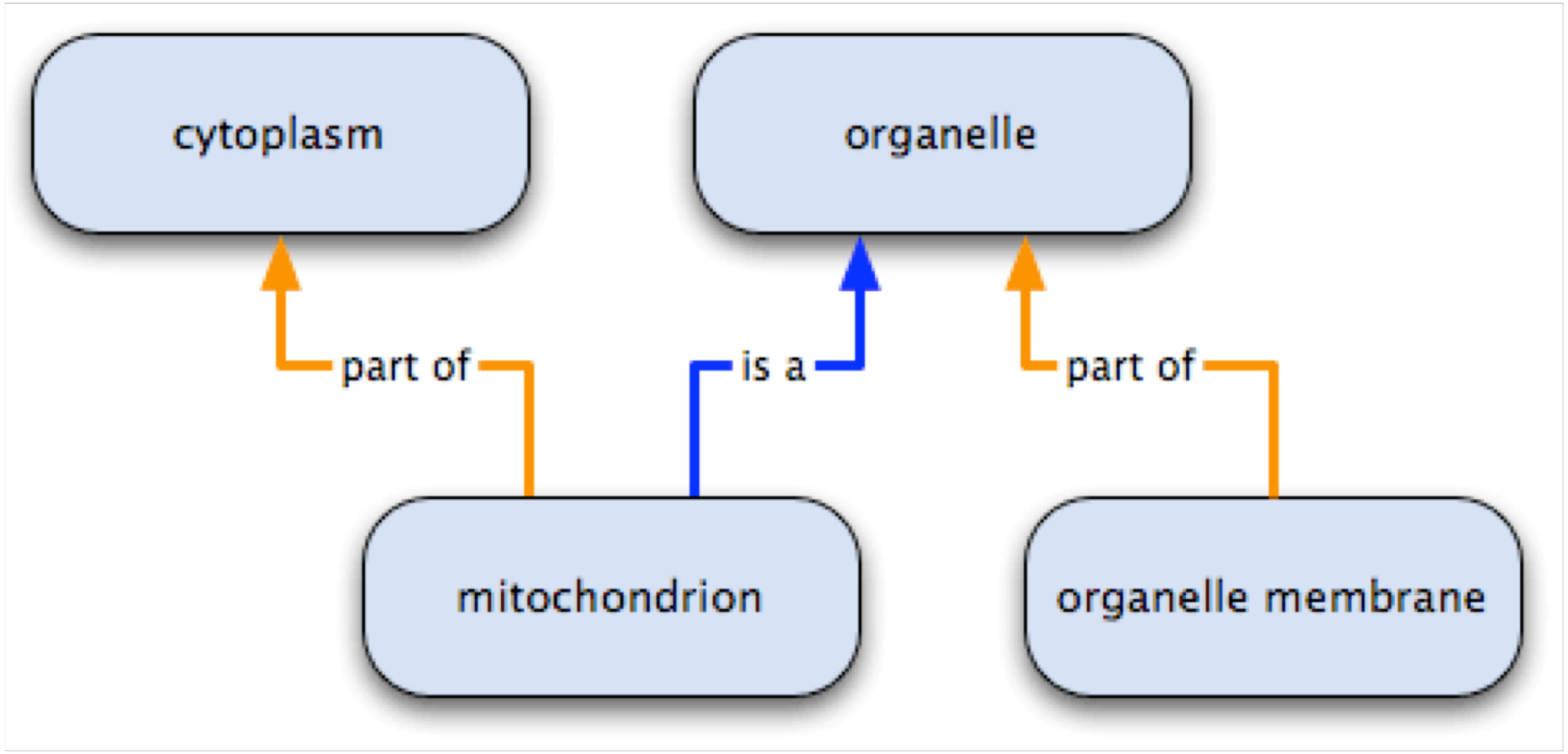
GO:1904659:glucose transport *is a* GO:0015749:monosaccharide transport.

GO:0031966:mitochondrial membrane is *part of* GO:0005740:mitochondrial envelope

GO:0006916:anti-apoptosis *regulates* GO:0012501:programmed cell death



http://geneontology.org/docs/ontology-relations/



http://geneontology.org/docs/ontology-relations/

GO Annotations

- Capture the biological role of genes or gene products
- Annotation of term implies annotation of all parents
- Unannotated = unknown

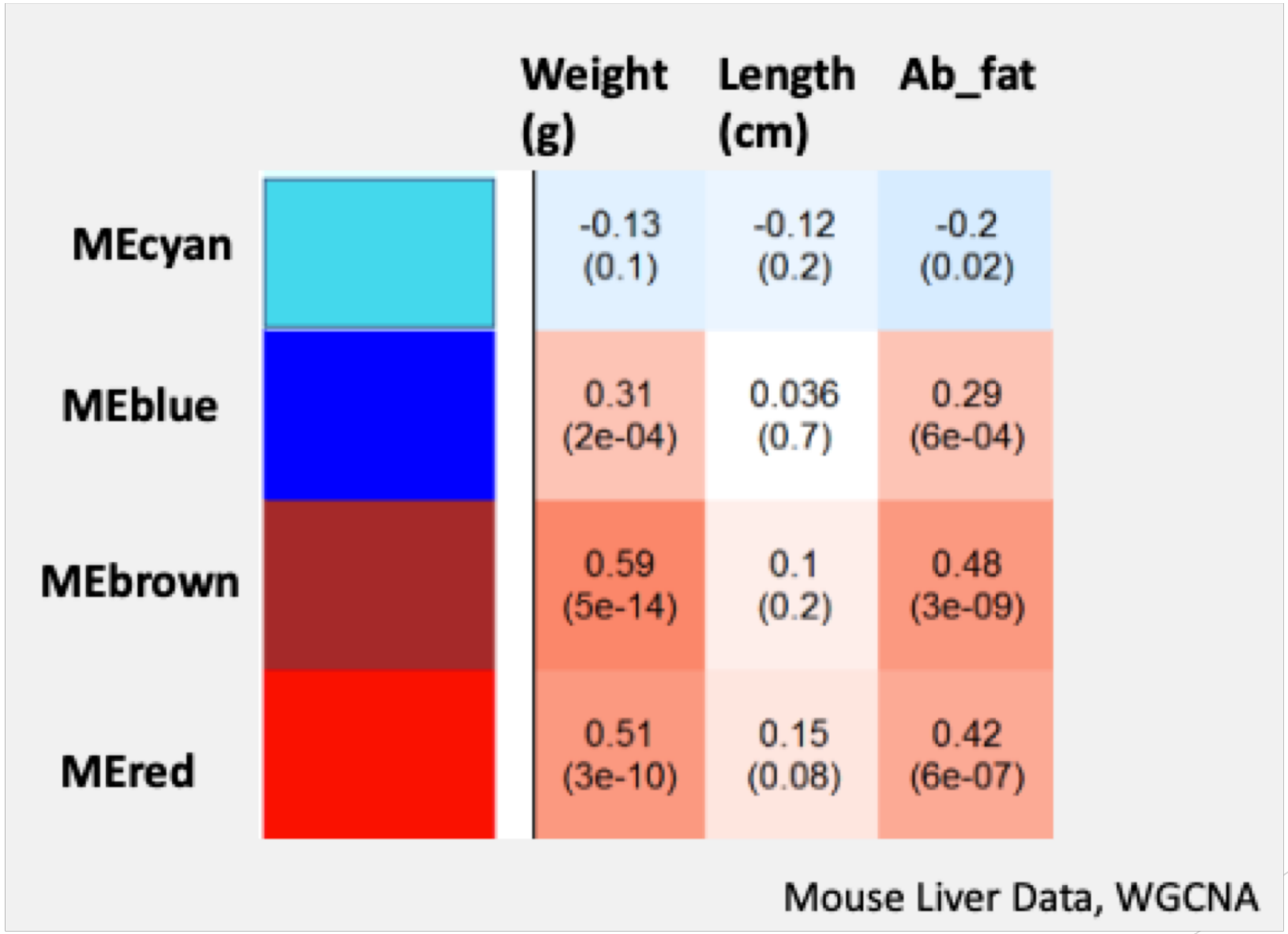
- Remember: we're always learning new things, so annotations can always change!

WGCNA: Weighted Gene Co-expression Network Analysis

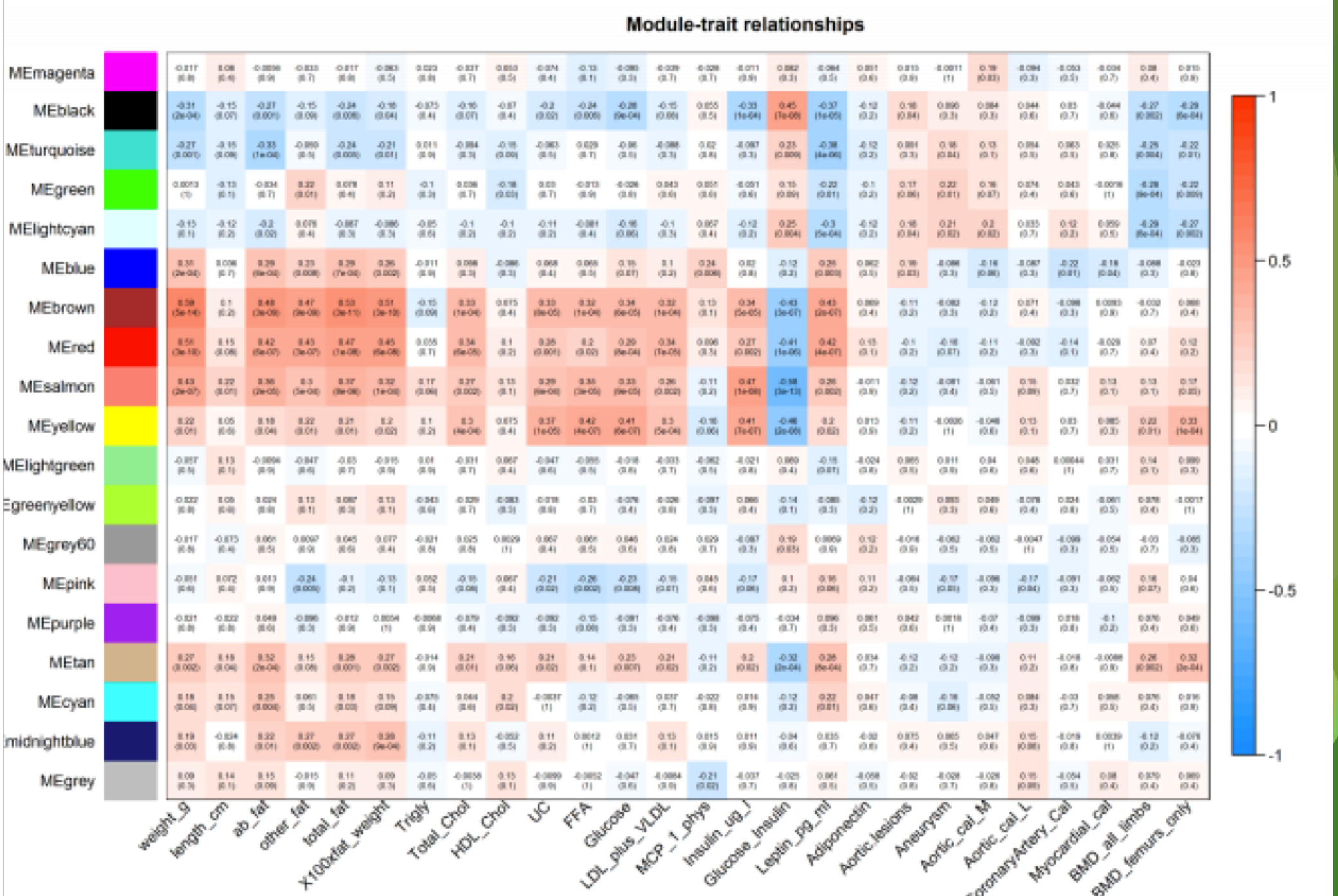
- Gene-gene similarity network
- Cluster genes by function
- Essentially guilt-by-association

- Good news: there's an R software package!

https://edu.isb-sib.ch/pluginfile.php/158/course/section/65/_01_SIB2016_wgcna.pdf



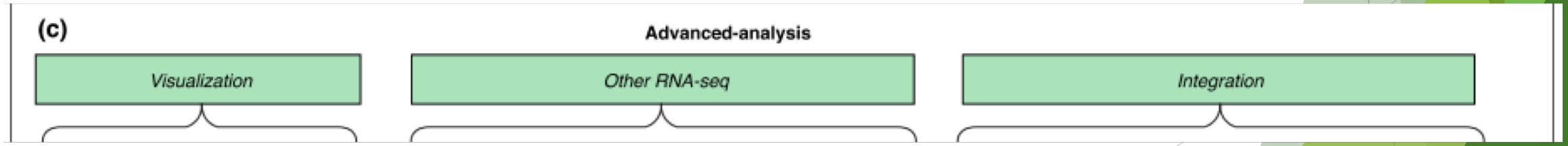
https://edu.isb-sib.ch/pluginfile.php/158/course/section/65/_01_SIB2016_wgcna.pdf



https://edu.isb-sib.ch/pluginfile.php/158/course/section/65/_01_SIB2016_wgcna.pdf

RNA-seq

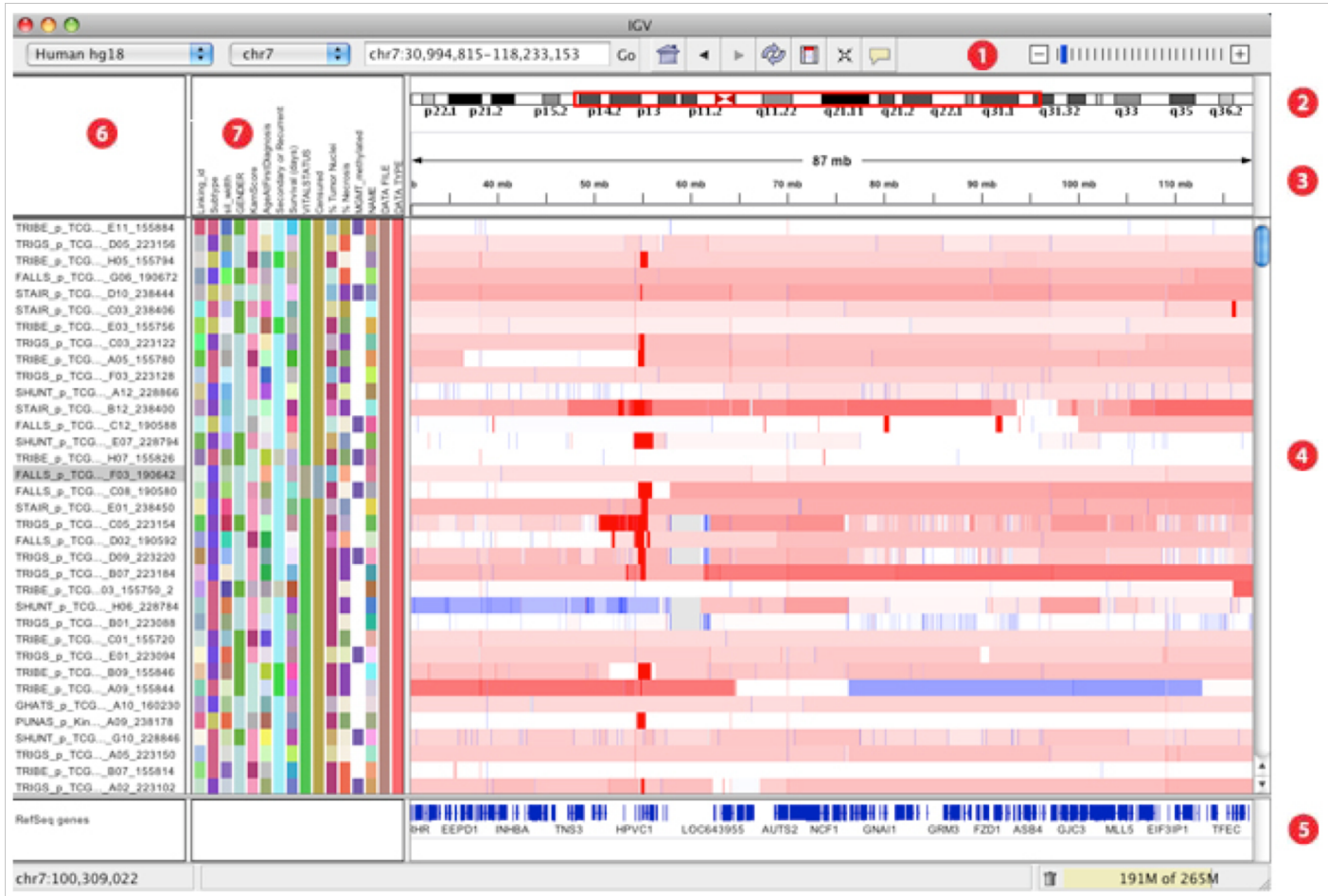
- Pre-analysis
- Core-analysis
- **Advanced-analysis**
  - Visualization
  - Other RNA-seq
  - Integration



Visualization

- Programs have functions solely for visualization!

- IGV: Integrative Genomics Viewer



1	The tool bar provides access to commonly used functions. The menu bar and pop-up menus (not shown) provide access to all other functions.
2	The red box on the chromosome ideogram indicates which portion of the chromosome is displayed. When zoomed out to display the full chromosome, the red box disappears from the ideogram.
3	The ruler reflects the visible portion of the chromosome. The tick marks indicate chromosome locations. The span lists the number of bases currently displayed.
4	IGV displays data in horizontal rows called *tracks*. Typically, each track represents one sample or experiment. This example shows methylation, gene expression, copy number, LOH, and mutation data.
5	IGV also displays features, such as genes, in tracks. By default, IGV displays data in one panel and features in another, as shown here. Drag-and-drop a track name to move a track from one panel to another. Combine data and feature panels by selecting that option on the General tab of the Preferences window.
6	Track names are listed in the far left panel. Legibility of the names depends on the height of the tracks; i.e., the smaller the track the less legible the name.
7	Attribute names are listed at the top of the attribute panel. Colored blocks represent attribute values, where each unique value is assigned a unique color. Hover over a colored block to see the attribute value. Click an attribute name to sort tracks based on that attribute value.

https://software.broadinstitute.org/software/igv/MainWindow

Other RNA-seq

- Gene fusion discovery
- Small RNAs
- Functional profiling

Overview

- Gene Expression
- RNA-seq
- **Data integration**
  - **Multi-staged analysis**
  - **Meta-dimensional analysis**
- Moving forward

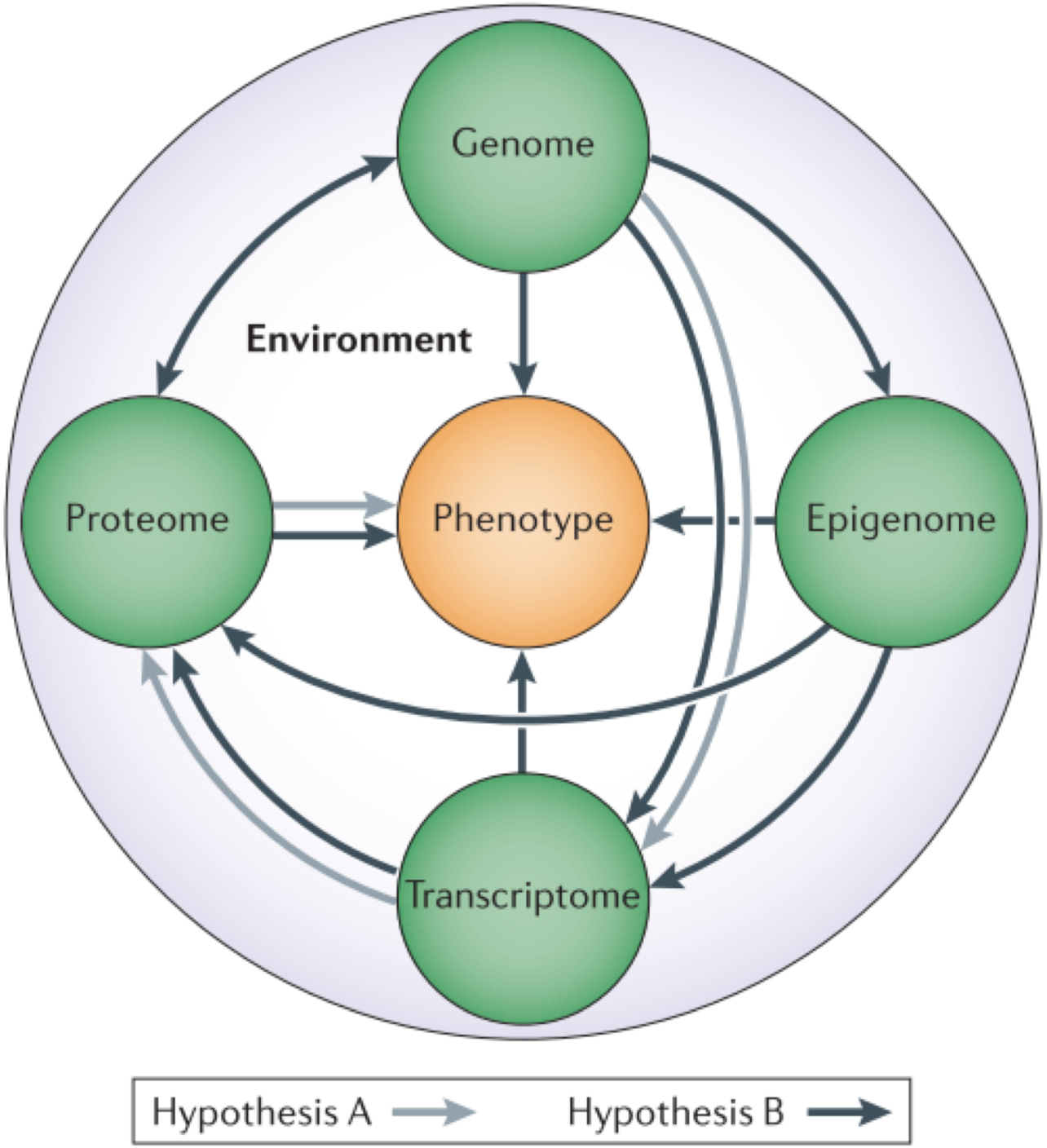
Why integrate data?

- Combine different types of omic data
- Better modeling of complex genotypes
- Predict biological outcomes

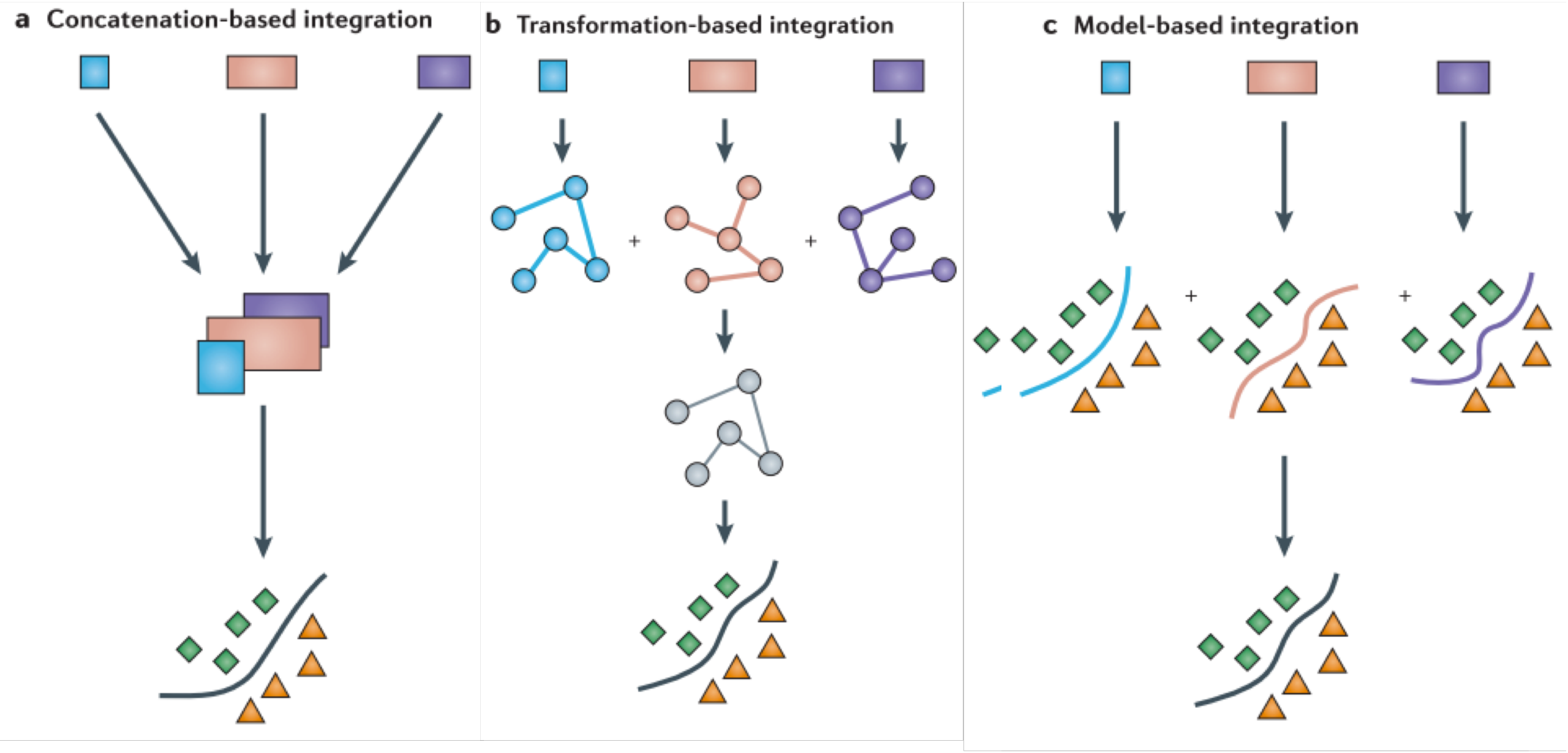
Challenges

- Evaluate each data type FIRST to avoid problems later!

- Quality assurance and quality control
- Data reduction
- Confounding



Ritchie et al. 2015



Ritchie et al. 2015

Moving Forward

- Single-cell RNA-seq
  - Amplify very small amounts of mRNA
- Long-read sequencing
  - Amplication free
  - Single-molecule sequencing of cDNAs
  - Nanopore GridION

And remember, there’s always Wikipedia

https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

Sources

- Conesa et al. 2016. A Survey of best practices for RNA-seq data analysis. Genome Biology. 17:13.
- Ritchie, M.D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nature Review Genetics. doi:10.1038/nrg3868.
- https://www.researchgate.net/figure/Graphic-representation-of-eukaryotic-transcriptional-machinery-A-Basal-eukaryotic_fig1_262384542
- https://www.news-medical.net/life-sciences/What-is-Gene-Expression.aspx
- https://slideplayer.com/slide/5255178/
- http://genedenovoweb.ticp.net:81/rsia/index.php?m=tools&f=heatmap
- http://amigo.geneontology.org/
- https://edu.isb-sib.ch/pluginfile.php/158/course/section/65/_01_SIB2016_wgcna.pdf
- https://edu.isb-sib.ch/pluginfile.php/158/course/section/65/_01_SIB2016_wgcna.pdf
- https://edu.isb-sib.ch/pluginfile.php/158/course/section/65/_01_SIB2016_wgcna.pdf
- https://software.broadinstitute.org/software/igv/MainWindow