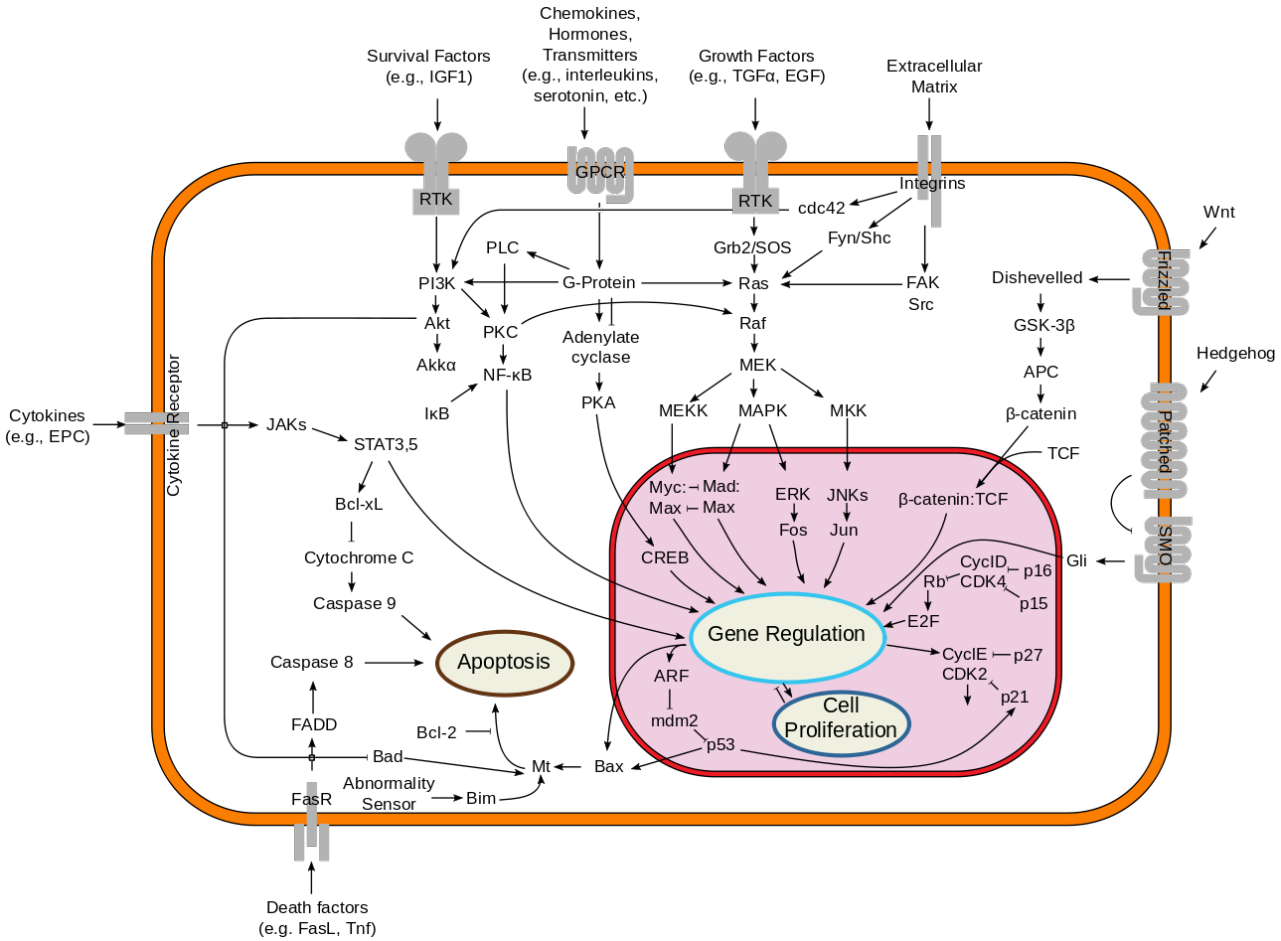
RNA-Seq Motivation

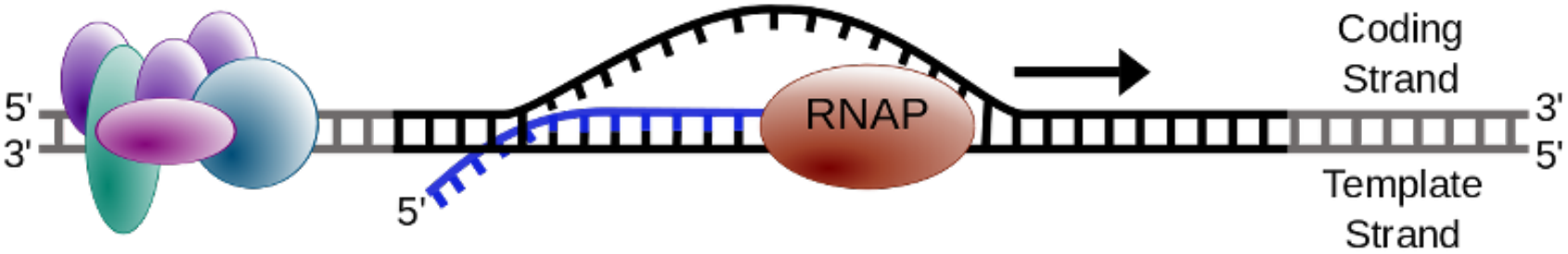
"Living systems are dynamic and complex, and their behavior may be hard to predict from the properties of individual parts."

(Website of the Department of Systems Biology, Harvard Medical School)



Systems biology aims to understand complete pathways/networks of interaction of biological molecule, and how they will react when manipulated

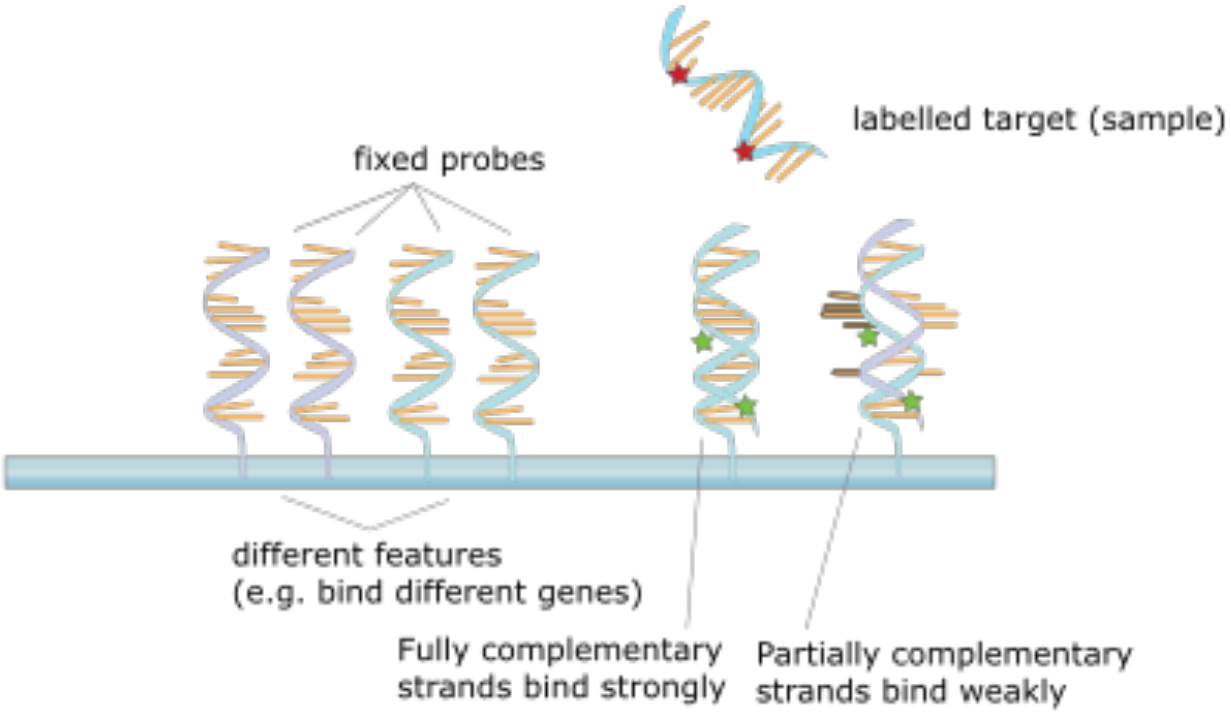
Image http://en.wikipedia.org/wiki/File:Signal_transduction_v1.png



genetic DNA -> transcribed to RNA (transcripts)

RNA-seq is an application of high throughput sequencing that attempts to analyse what composition of RNA molecules is present in a sample

Image Wikipedia



RNA-seq is replacing microArrays

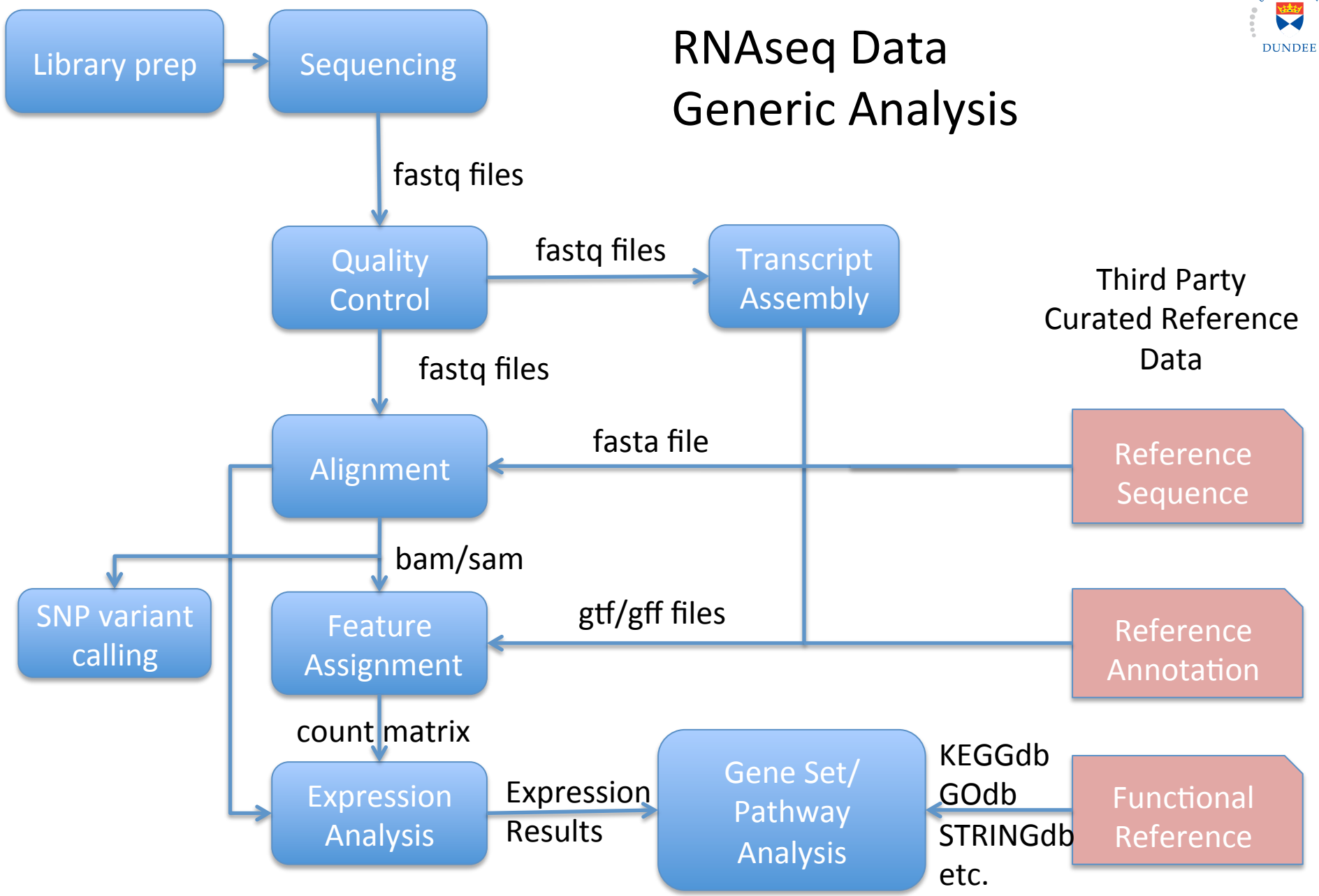
cDNA molecules tagged with fluorophores hybidise to pre-defined array of complementary probes.

Image Wikipedia

- Whole transcriptome
  - but can be targetted at specific regions
- Disadvantages of microArrays
  - probes decided on in advance, can only detect what you look for
  - complicated normalisation require to account for with and between array technical variation
- Percieved advantages of RNA-seq
  - probe free, whole transcriptome discovery
  - "simpler" more direct measure of transcripts counts
  - novel organisms with no existing arrays

- Transcriptome Assembly
  - Identify the sequences of the RNA's in the sample
    - Identify rare non-canonical transcripts alternate splicing
  - Quantify relative amounts of these transcripts
  - SNP variant calling
- Differential expression
  - Identify which gene/transcript change between different conditions
    - mutant (disease) vs.. control
    - treatment vs. control
    - over a time course (cell cycle)

RNAseq Data Generic Analysis



RNA-Seq Library Prep

"For measurements to be meaningful, however, they must retain their connection to the theoretical and instrumental context from which they were derived. "

(Houle et. al. 2011 Q. Rev. Biol.)

RNA

- ~75% of RNA is 18S & 28S ribosomal RNAs (rRNA)
- ~10% of RNA is transfer RNAs (tRNA)
- ~2% of RNA is protein coding messenger RNAs mRNA
- Remainder is mix of small and long non-coding mirRNA, snoRNA, lncRNA etc.

(Jackson et. al. FASEB J. 2000)

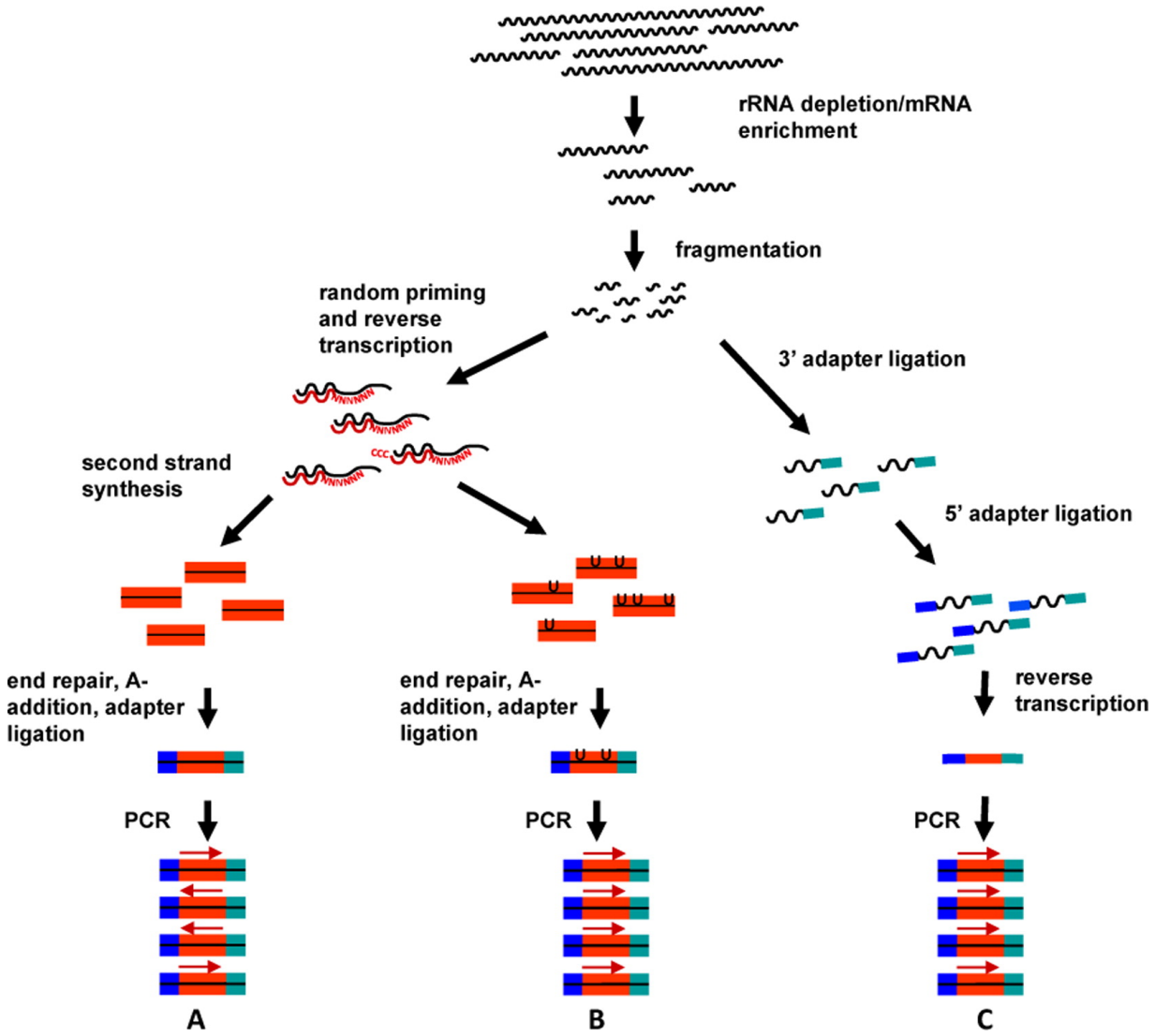
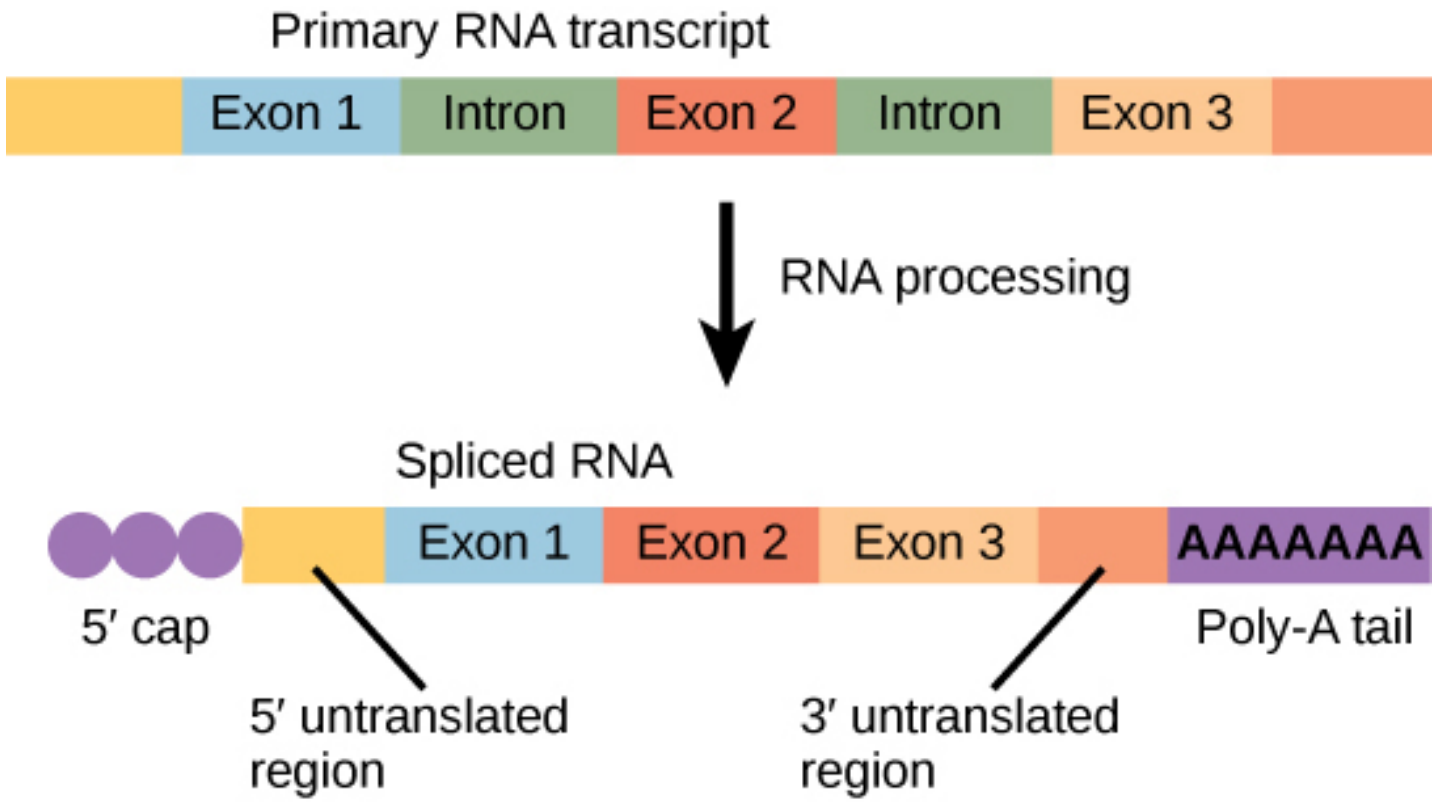


Image van Dijk et. al. (2014) Exp Cell Res



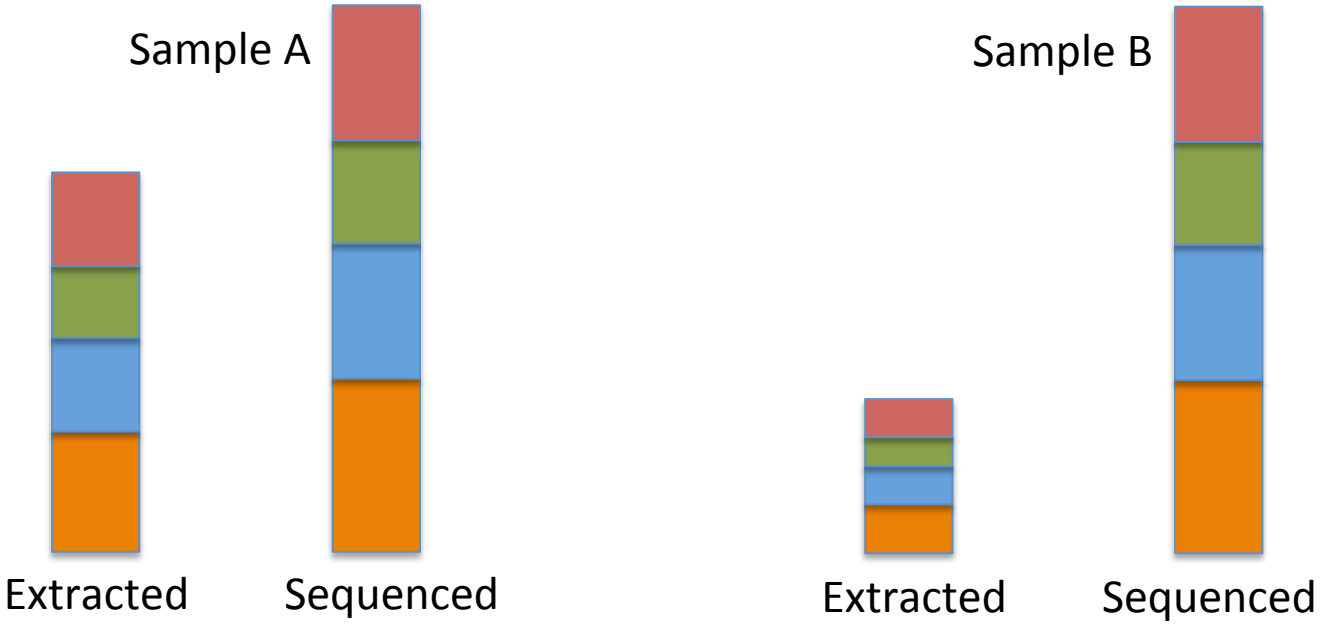
most RNA is processed after/during transcription

mRNA has a 5'cap and a 3' poly-A tail added

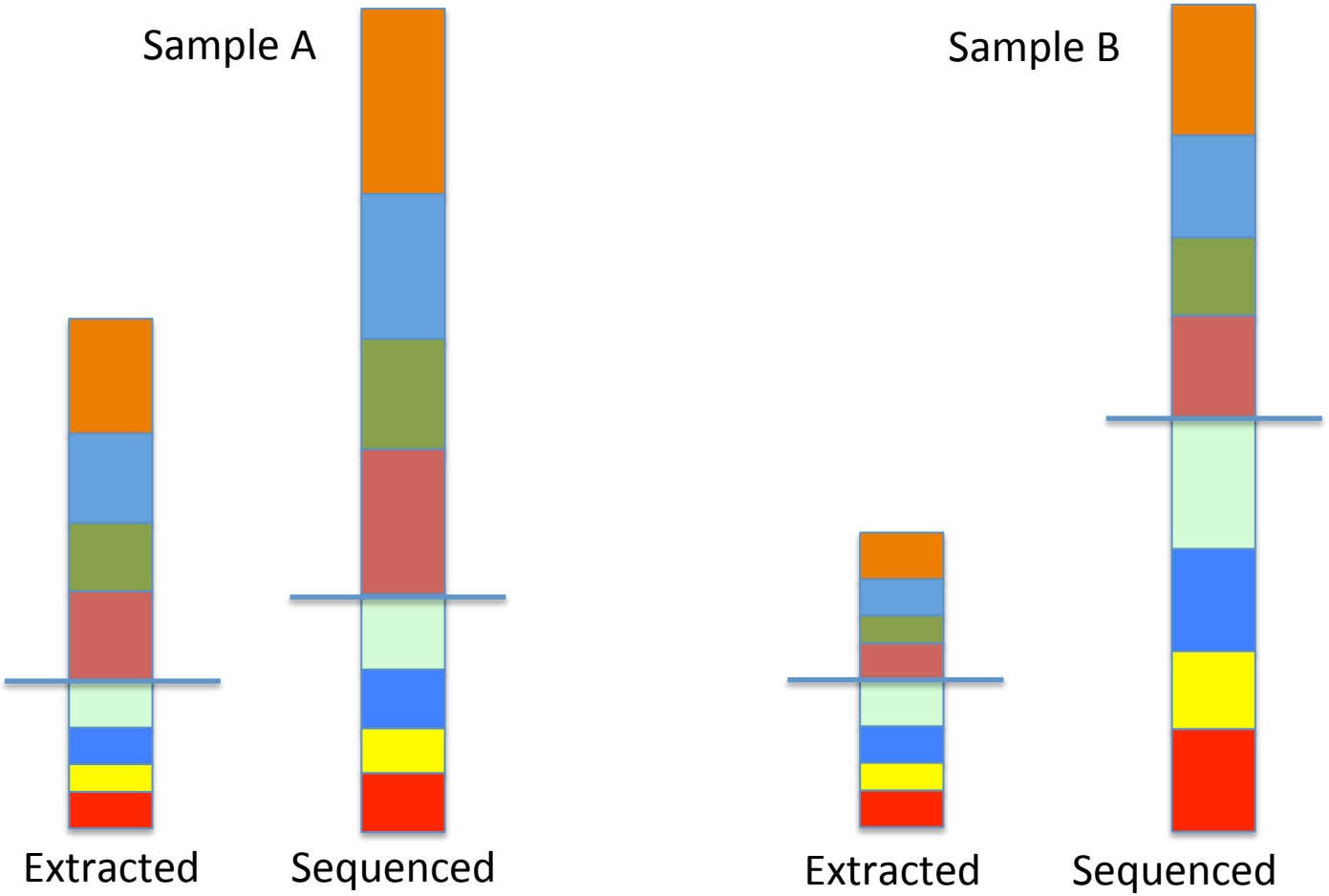
Image https://cnx.org/contents/TkuNUJis@3/Transcription

- Enrich for non ribosomal RNA
  - polyA pull down
  - ribosomal RNA depletion
    - ribo-minus, ribo-zero kits
- Fragment
  - Add primers
- Reverse transcribe
  - Can produce strand specific cDNA's
- Amplify what is left
  - PCR can introduce bias
- Then sequence
  - more often than not these days with Illumina

- But generally this is not quantities it is proportional
  - scale up to a fix volume



Add (spike-in) fixed amounts artificial calibration RNA's



External RNA Controls Consortium (ERCC)

RNA-Seq Experimental Design

"Although the principles of good design are straightforward, their proper implementation often requires significant planning and statistical expertise."

(Auer & Doerge 2010 Genetics)

- replication
  - the more times something is repeated, the greater the confidence of ending up with a genuine result.
    - biological replicates – different biological samples
    - technical replicates – same biological unit processed multiple times
- randomization
  - experimental subjects must be allocated to treatment groups at random
- record covariates,
  - including processing day – likely 'batch effects'

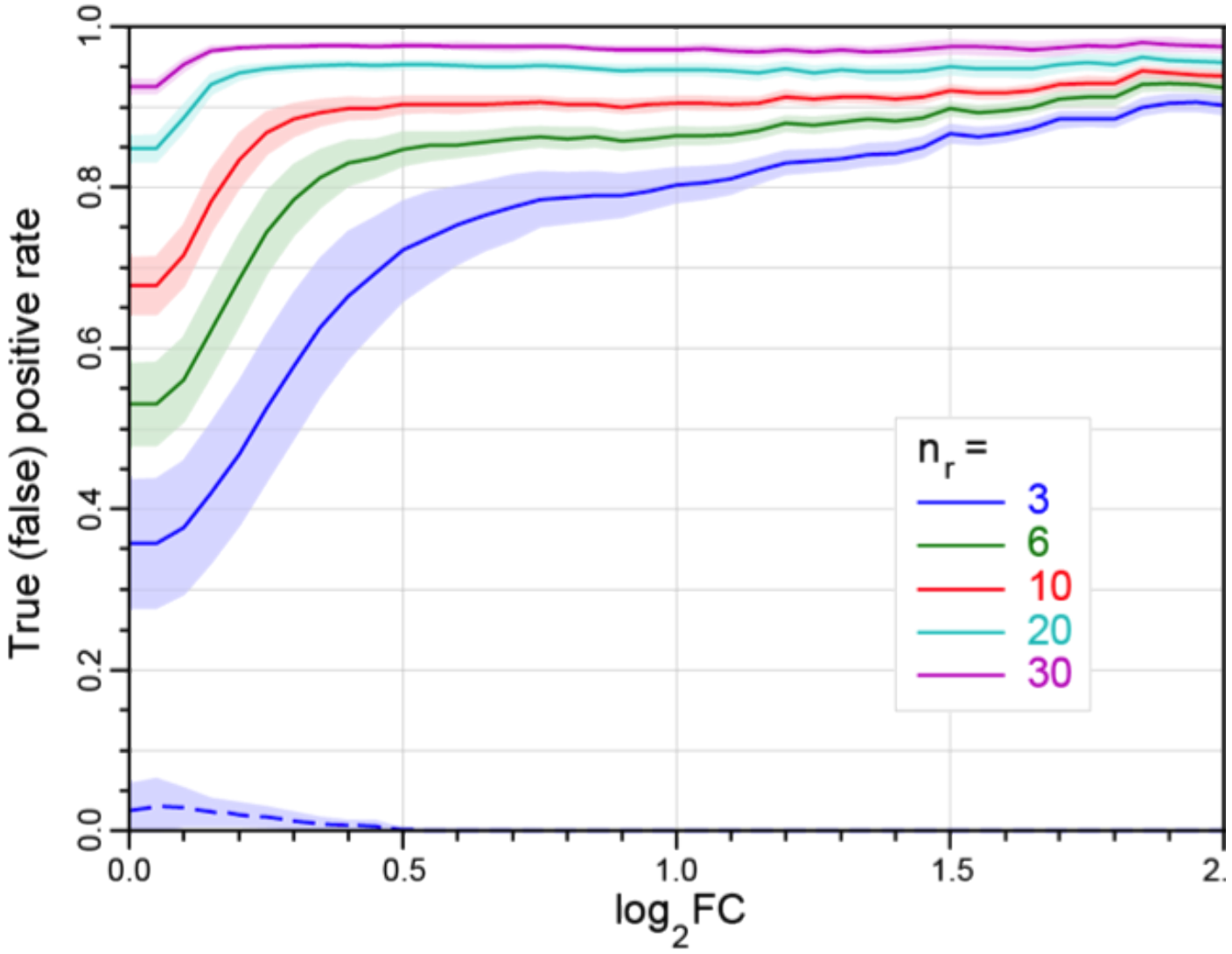


image Schurch et. al. 2015

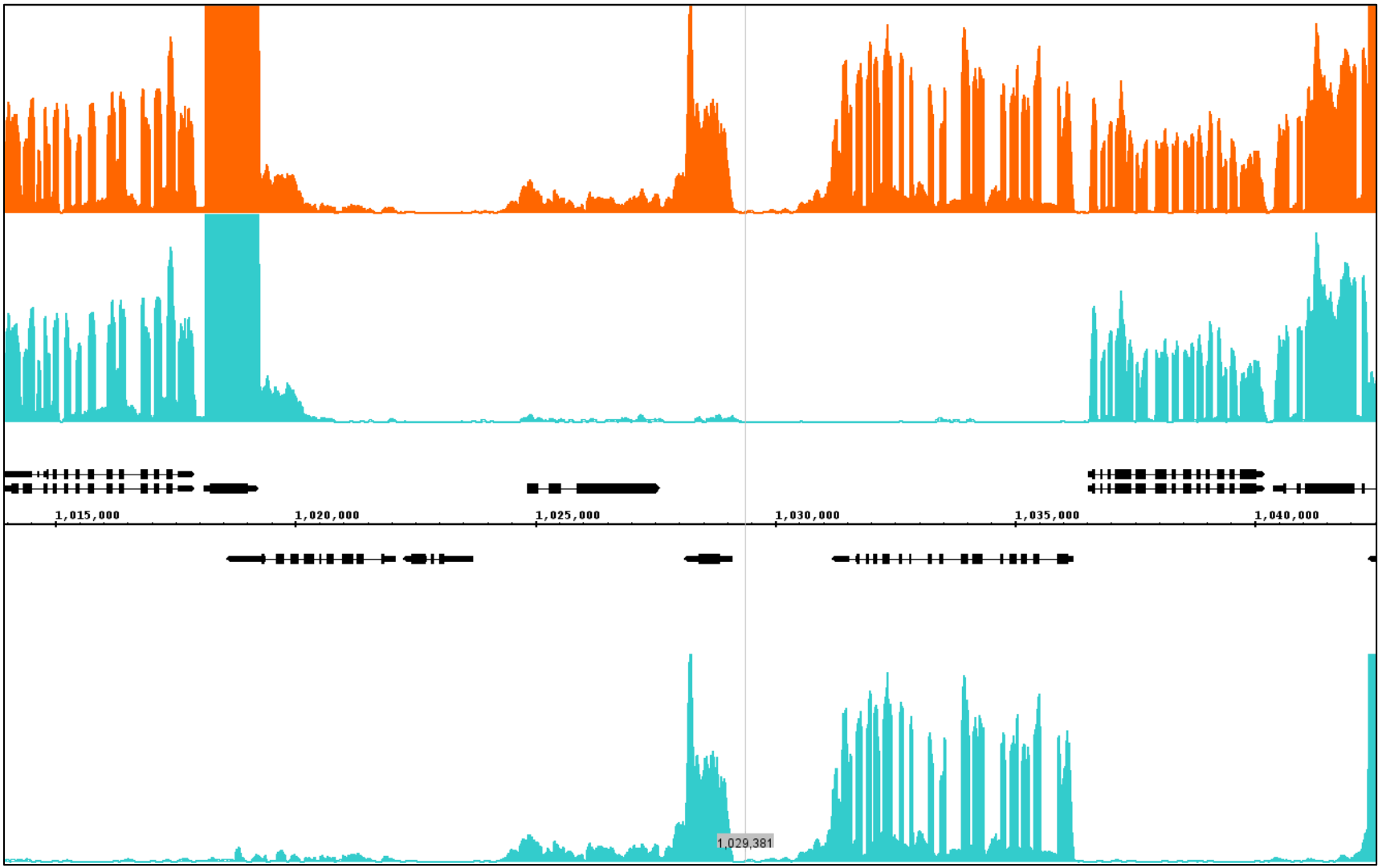


image Schurch 2014

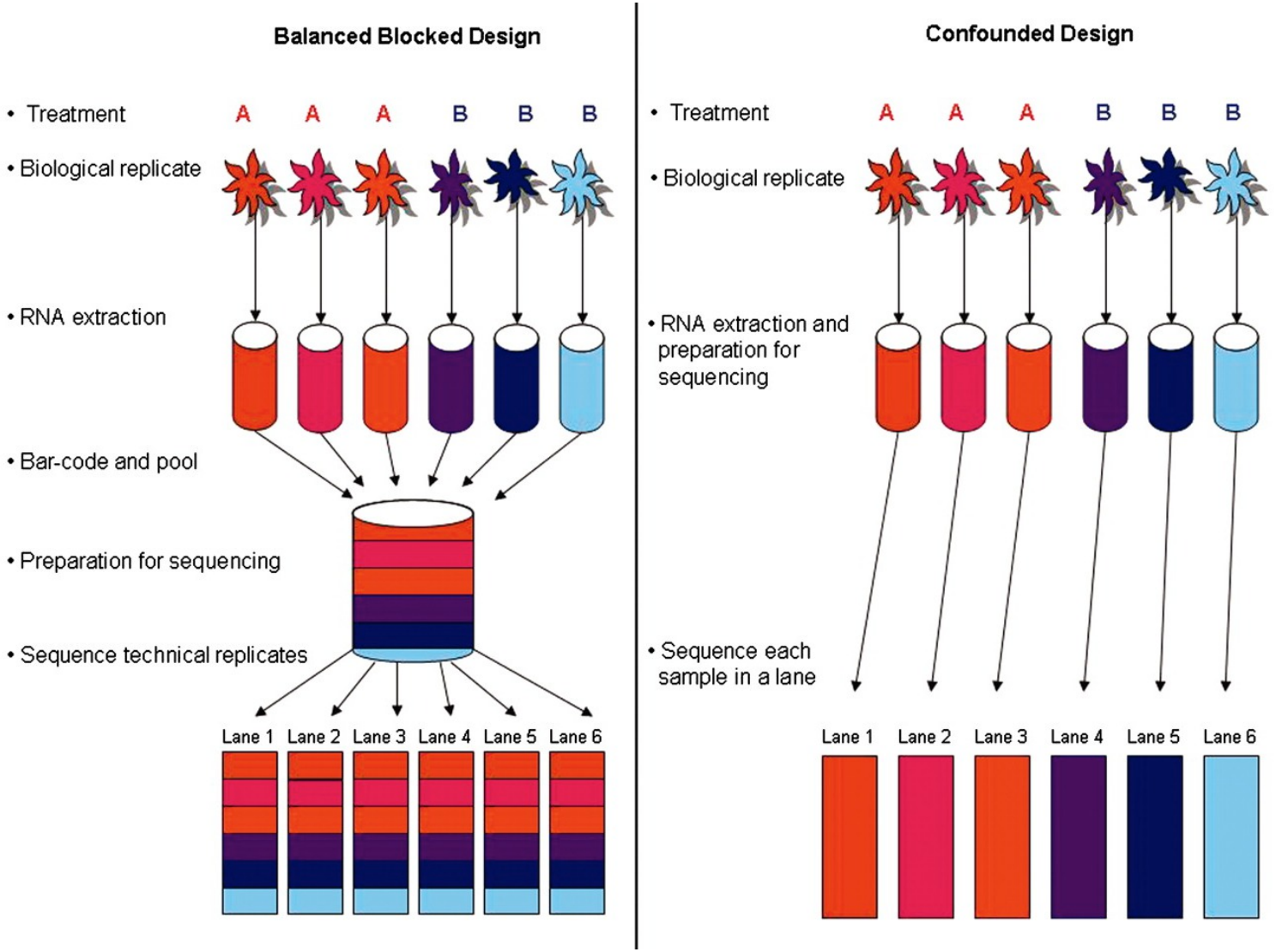


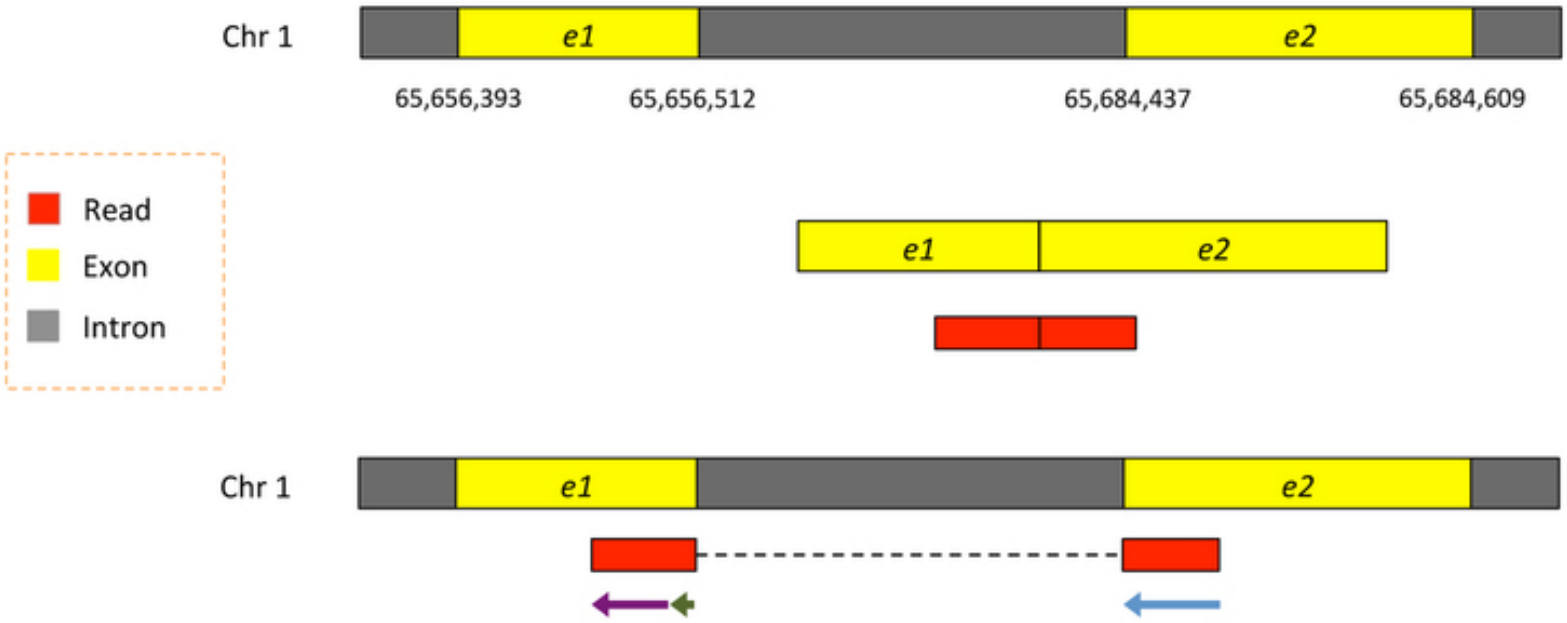
Image Auer & Dourge 2010 Genetics

RNA-seq Data Pre-processing

On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

(Babbage 1864. *Passages from the Life of a Philosopher*.)

- Fastq file.
  - Raw read quality control (e.g. FastQC)
    - base prediction qualities
    - GC content
    - over represented sequences
    - k-mer content
  - Reads may be spliced



mRNA sequence tags (reads)

- Use reads to quantify proportions of RNA transcripts
  - Use a genomic reference (whole genome sequence one sequence per chromosome)
    - align the reads to the reference
    - for list of regions associated with features of interest (typically genes) find how many reads map to those regions
  - Use a transcriptome reference (sequences of the transcripts of interest one sequence per transcript)
    - align reads to the reference
  - Assemble transcripts from the reads de novo
    - align reads back to new de novo reference

de novo assembly

- advantages
  - you quantify the transcripts that are present
  - novel gene models, alternative splicing
  - not limited to existing annotation quality
- disadvantages
  - quality transcriptome assembly is non-trivial
  - requires more depth sequence for than for simple DE
  - introduce error
- if your organism has not been sequenced this is your only option

Genomic reference alignment

- spliced reads
  - require splice aware aligner
- relies on an existing genomic reference and feature annotation
  - but if you look you might discover unannotated features and gene models
- requires second processing step to quantify reads per feature

Transcriptome alignment

- will only discover what you look for
- faster
  - get away with non-splice aware aligner
  - no second mapping step

Reads that map to multiple features

- multi mapping reads
  - repetitive regions
  - copy number, pseudogenes
  - often the approach is to randomly select one of the possible locations
- over lapping genes on opposite strands
  - strand specific libraries can help here
- alternative transcript models

RNA-Seq exploration

"You can't fix by analysis what you bungled by design."

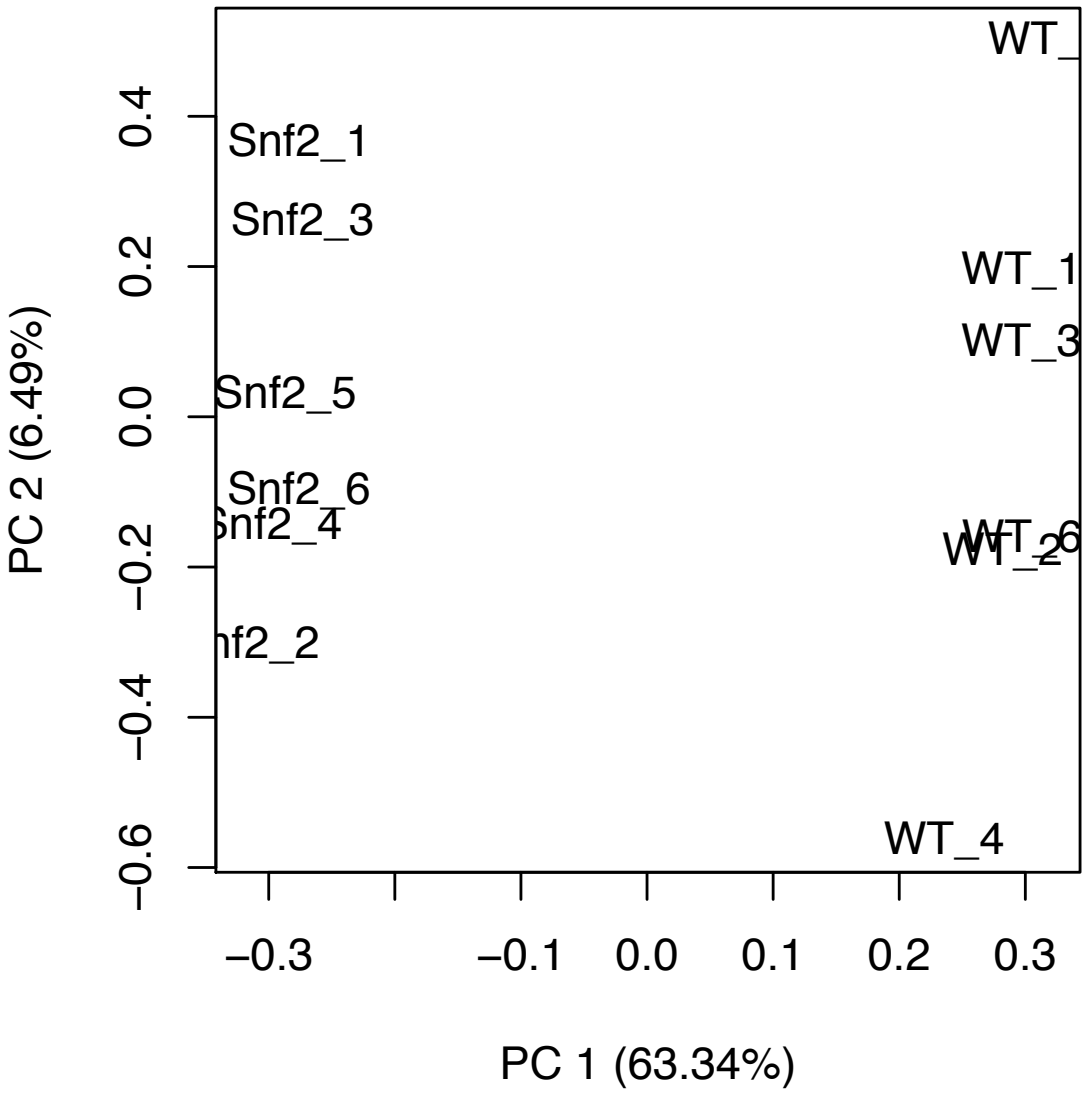
Light, Singer and Willett (1990)

- TPM vs. RPKM/FPKM vs. Normalised Counts
- Look for unwanted variance
- Normalisation
  - library size
  - total RNA composition
  - gene length
  - GC content

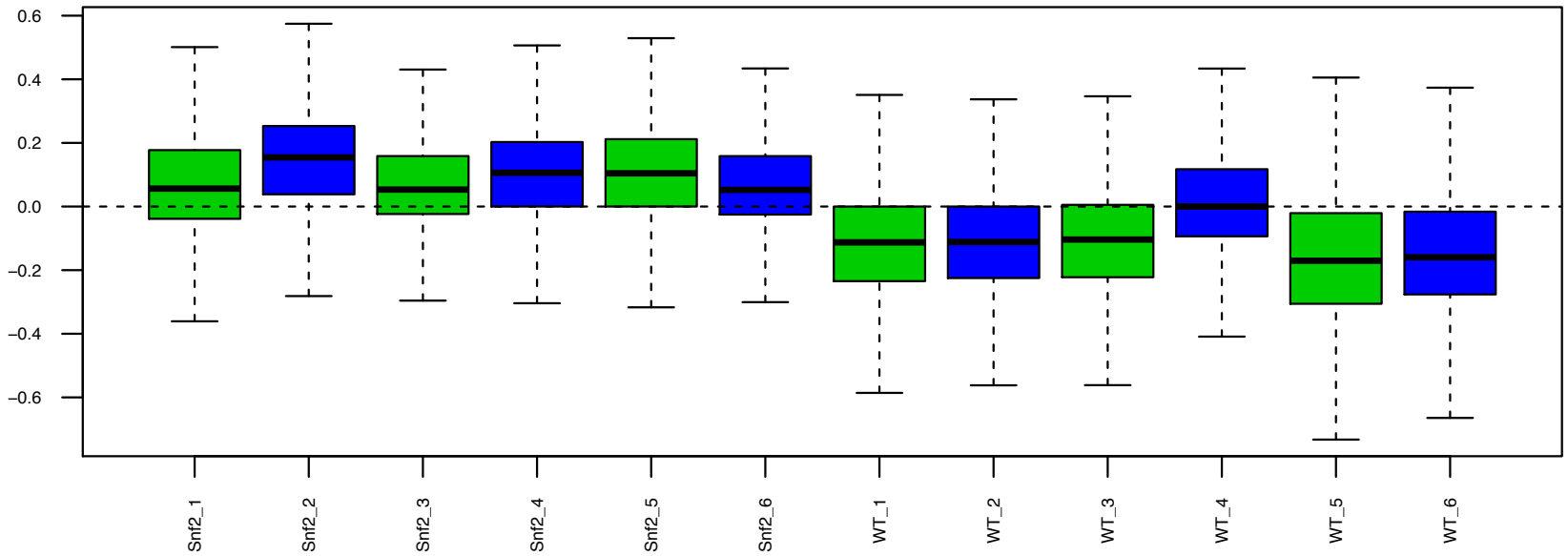
RNA-seq Visualisation

"Numerical quantities focus on expected values, graphical summaries on unexpected values."

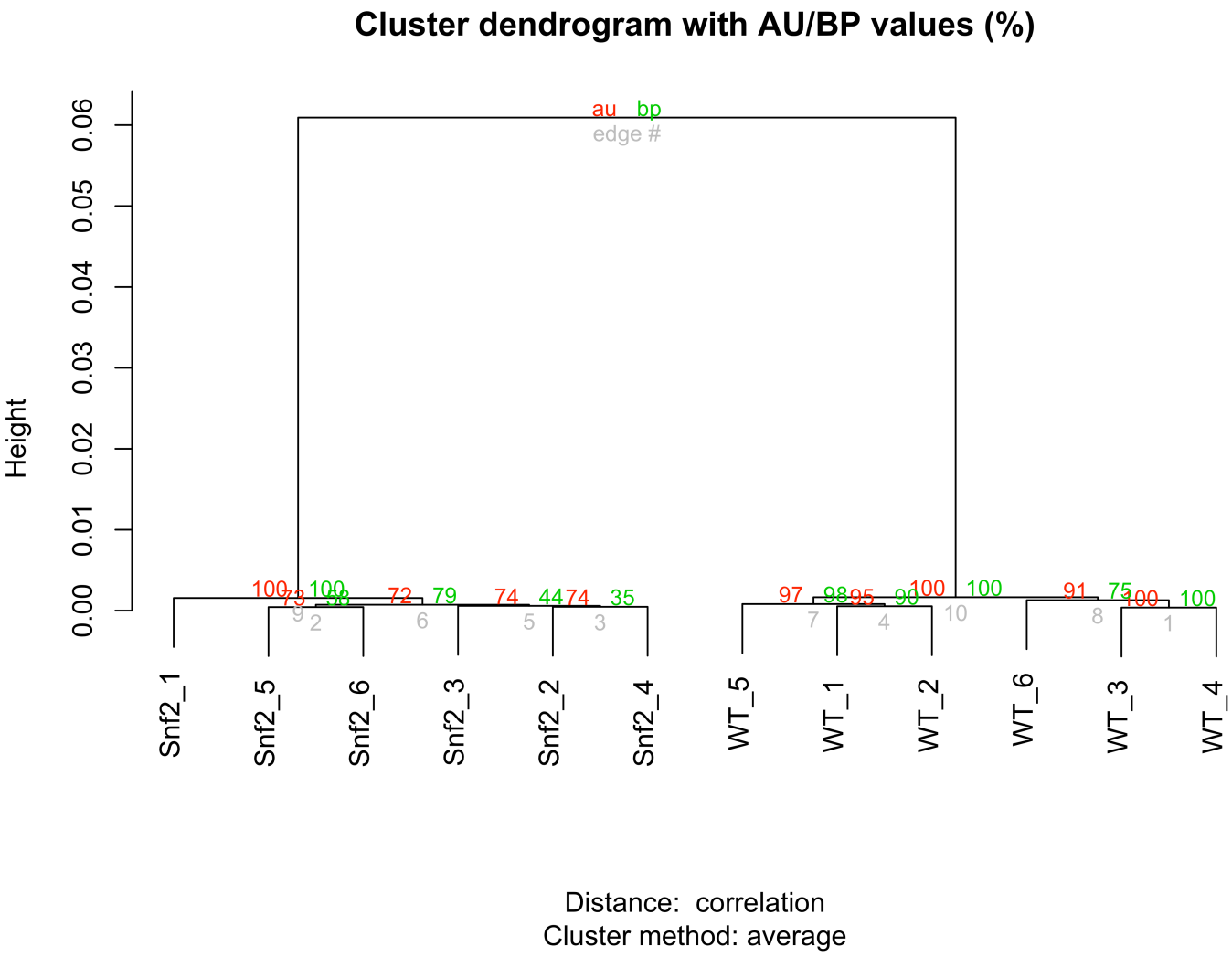
John Tukey



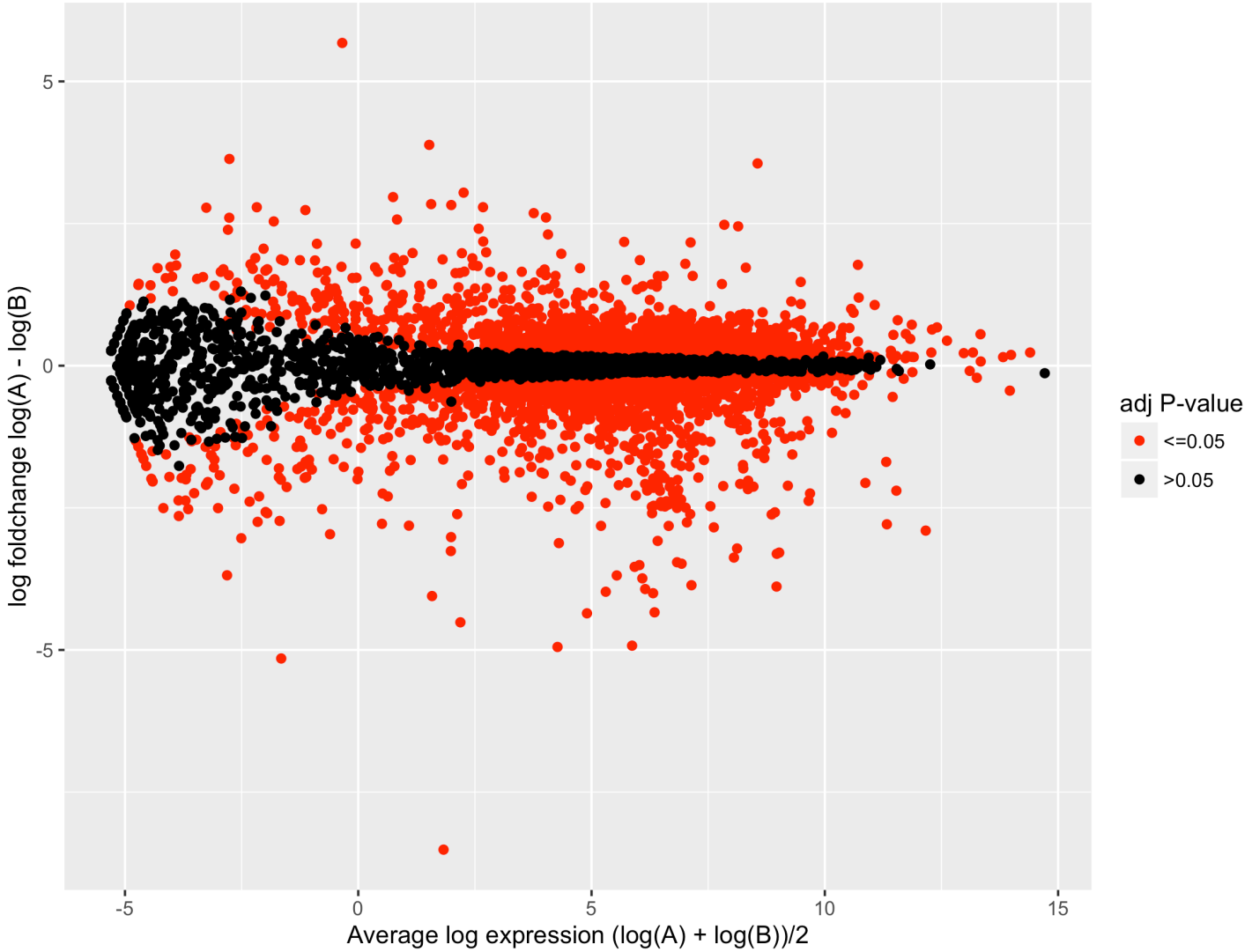
PCA Plot: first two principal components (cf. multi-dimensional scaling MDS Plot)



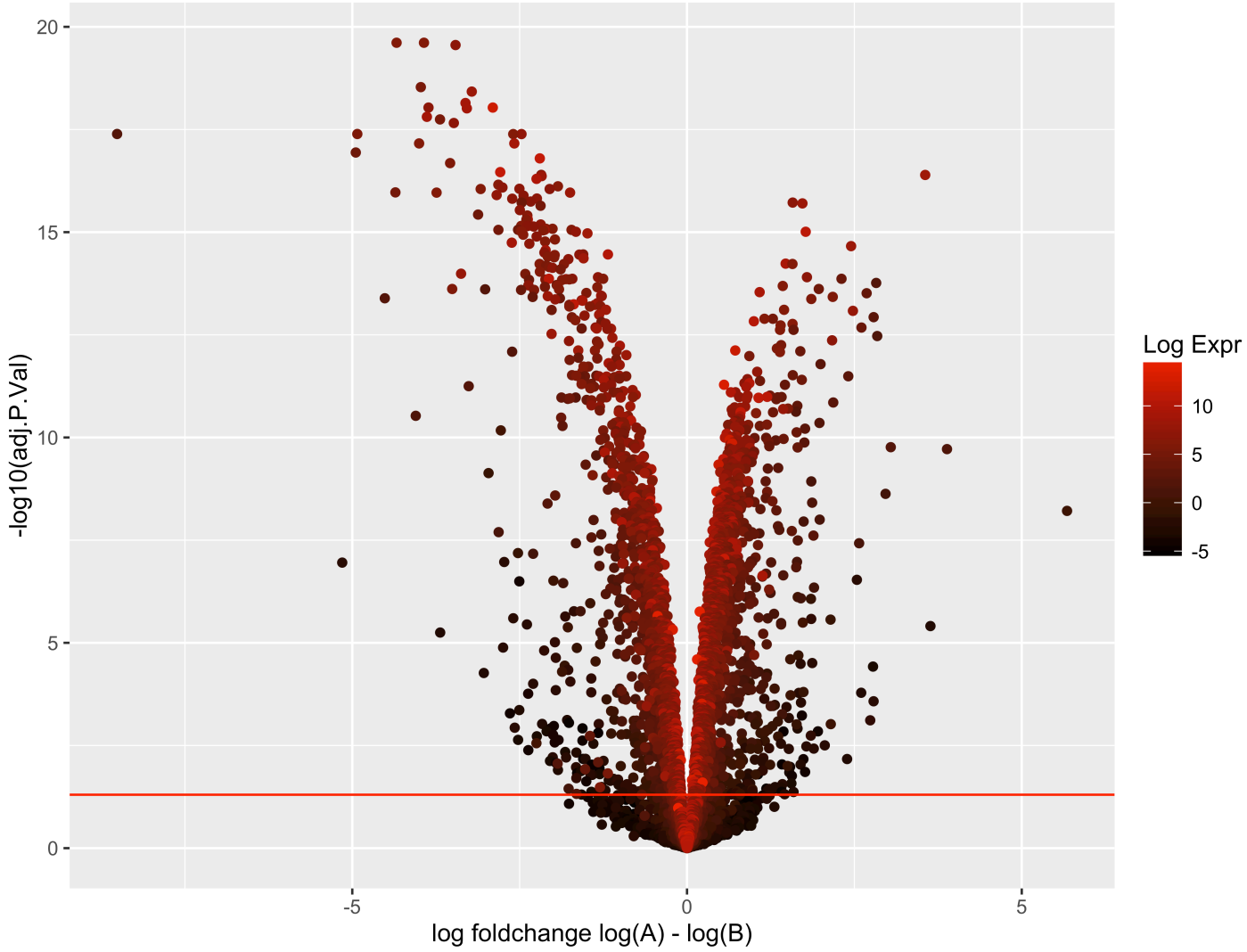
Boxplots and relative log expression (RLE) plots



Dendrogram (cluster analysis)



MA Plot: difference of the log expressions versus the mean log expression (Bland-Altman or Tukey Mean-Difference Plot)



Volcano Plot

RNA-Seq downstream processing

"Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination"

R. A. Fisher

- Gene Set Enrichment
  - Pathway analysis
  - Gene Ontology functional annotation analysis
  - Gene Network analysis

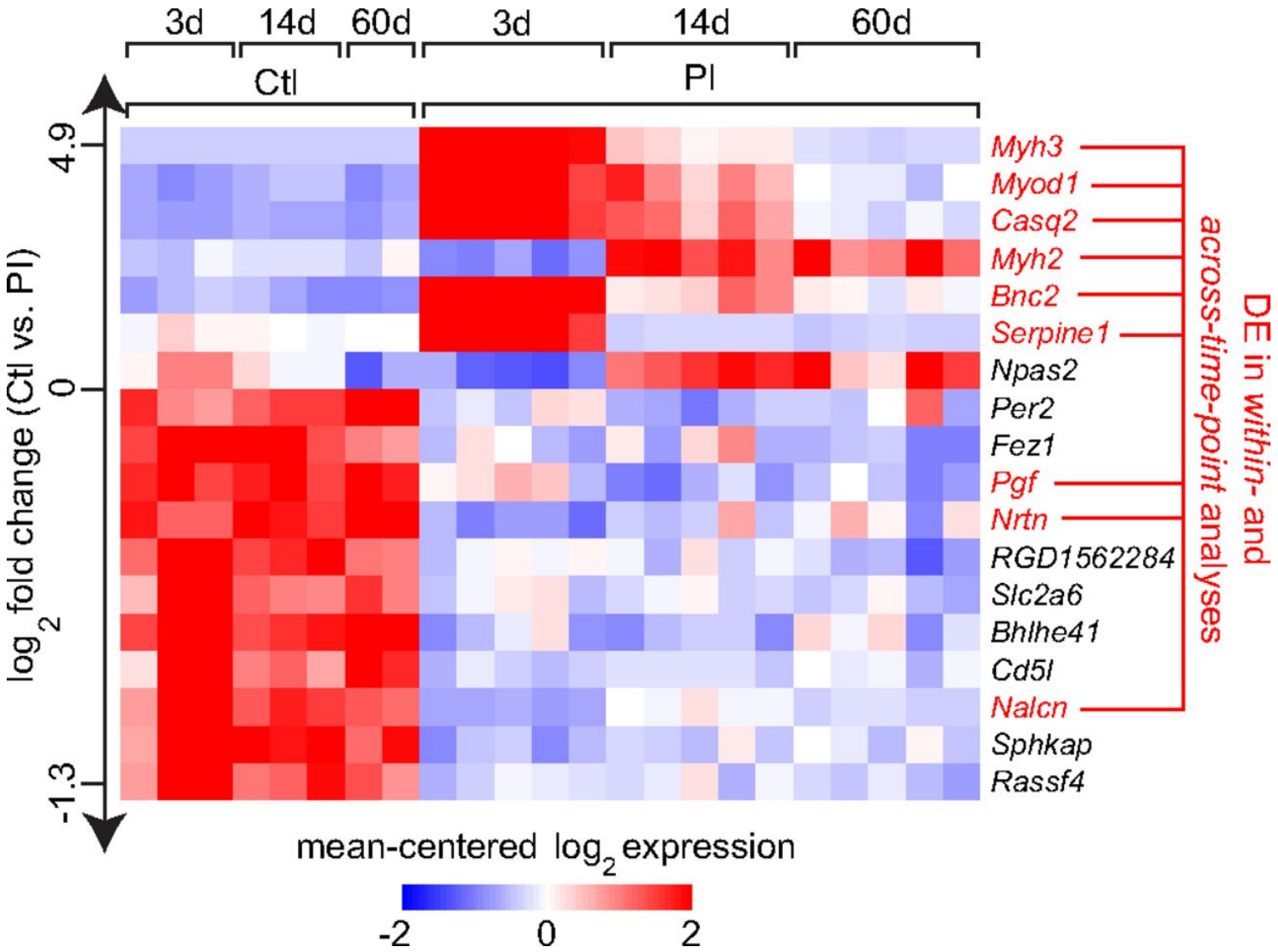


Image Welham et. al. 2015 Dis. Mod. Mech.

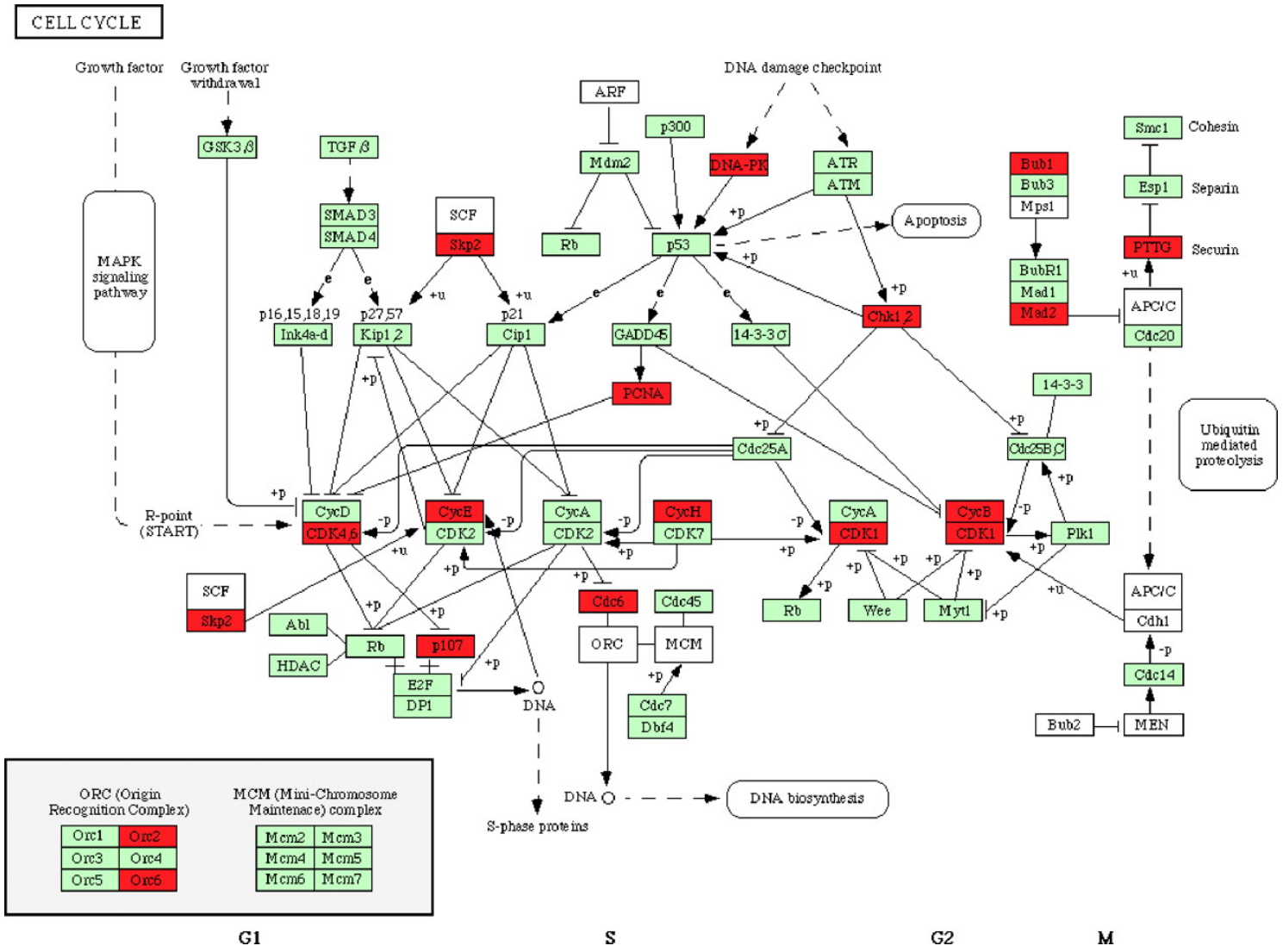


Image King et. al. 2005 Physio. Gen.

Analysis Software

"Reinventing the wheel is sometimes the right thing, when the result is the radial tire."

Jonathan Gilbert

- Quality Control
  - FastQC, FASTX-Toolkit, trimmomatic, seqtk...
  - R Bioconductor
- Assembly
  - trinity, SOAP denovo-Trans...
  - cufflinks
- Aligners, optimal depends on downstream task
  - bowtie, bwa-mem, subread....
  - tophat, subjunt, STAR,...
  - kallisto, salmon
  - R Bioconductor

https://en.wikipedia.org/wiki/List_of_sequence_alignment_software

- Alignment Processing
  - Samtools, sambamba, picard
  - R Bioconductor
- Data Exploration
  - R Bioconductor
    - RUVSeq, NOISeq, EDASeq...
- Differential Expression
  - RSEM...
  - cuffdiff...
  - R Bioconductor
    - edgeR, limma, DESeq...
    - sleuth

- GSEA
  - Ingenuity Pathway Analysis
  - R Bioconductor
  - Panther
- SNP variant calling
  - GATK
  - R Bioconductor

- Integreated Analysis Environments
  - CLC Bio
  - Galaxy
  - UGENE
  - Pipeline Pilot

R Bioconductor

- Suite of packages for the R Statistical Computing environment
- pros
  - an attempt at standards of documentation
  - scriptable (extensible)
  - reproducible (dynamic document production)
  - extensive pre-existing libraries
  - free
- cons
  - command line not point and click
  - scripting language is R
  - unfriendly (learning curve)

"The NIH (Not Invented Here) Syndrom is a disease"
Linus Torvalds
https://en.wikipedia.org/wiki/Not_invented_here