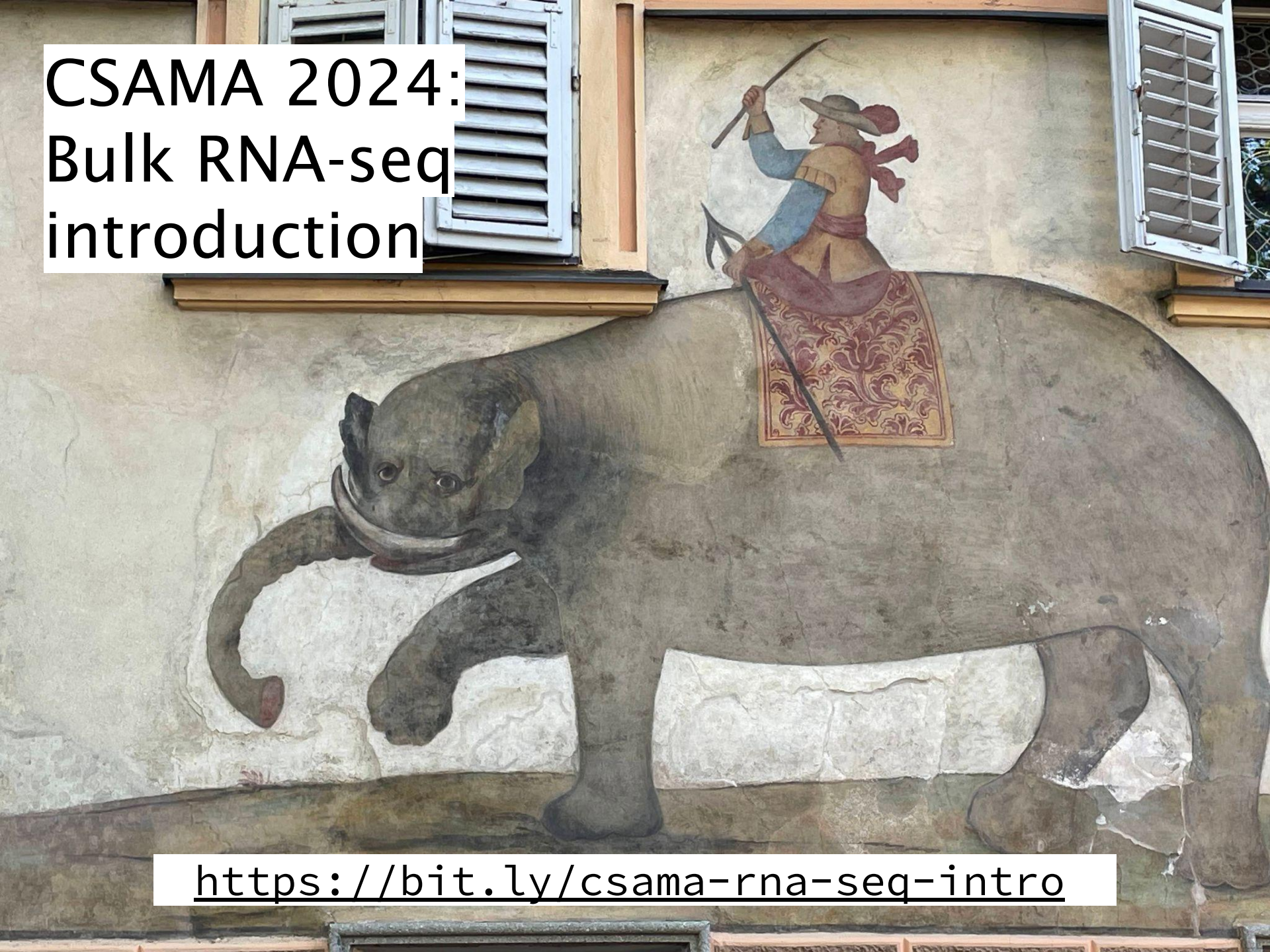
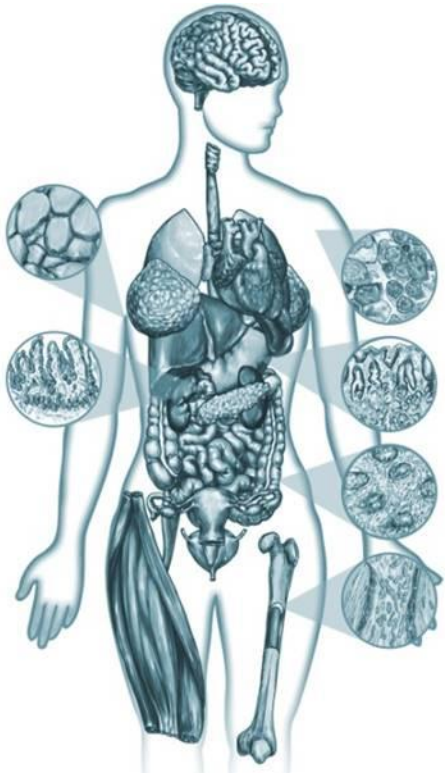
CSAMA 2024: Bulk RNA-seq introduction



https://bit.ly/csama-rna-seq-intro

Gene expression

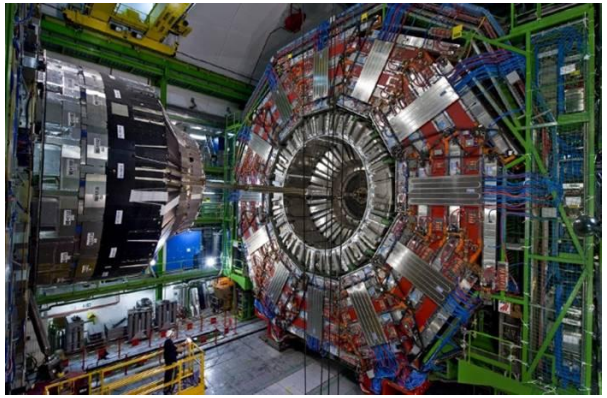
- Dynamic across time, tissue, individuals
- Measurement is harder than for the genome



[Roadmap Epigenomics Project]

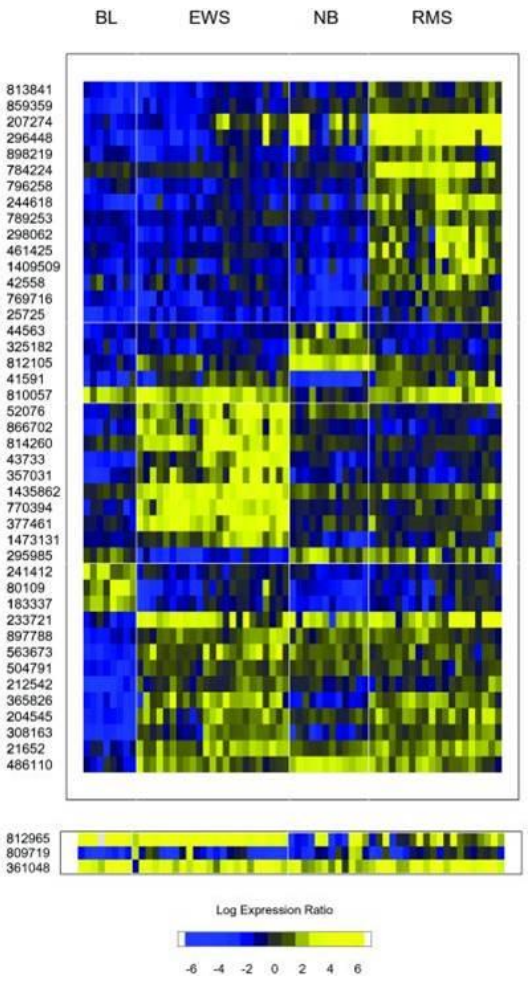
Why gene expression?

- Basic biology: transcription, translation, RNA enzyme
- As a phenotype: easier to measure than proteins
- Research: find interesting genetic loci
- Diagnostic: classify cancer subtypes ← FDA (2007)
- New and better measuring devices drive discovery



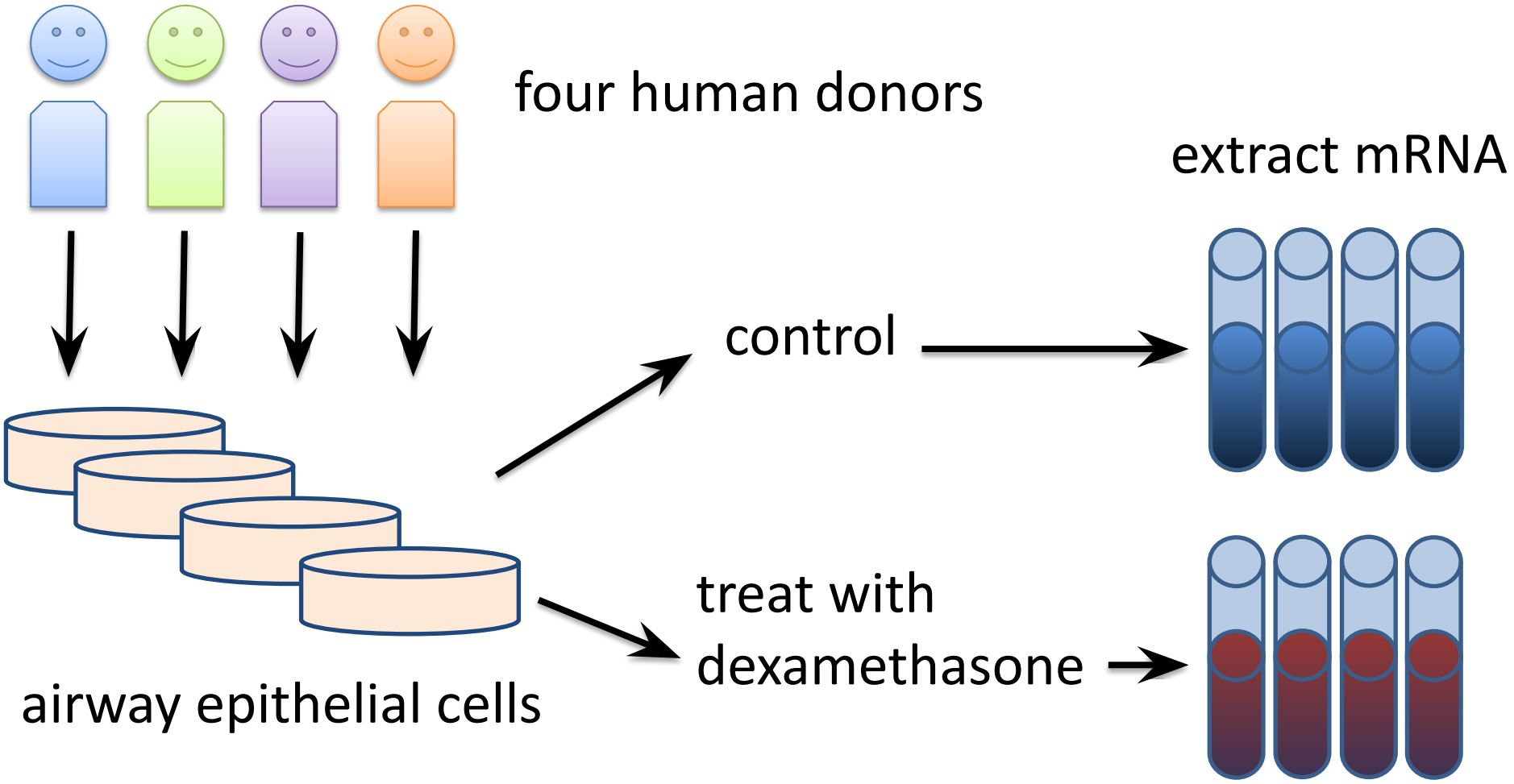
Statistical analysis of gene expression

- Alternatively: Northern blot, qPCR
- Era of microarrays, now sequence
- Clustering, differential expression
- Marginal testing, simple enough
- Statistical methods offer benefit
- Key insight about expression data:
  - Costly, often few replicates (n=3-5)
  - Many genes over which to estimate parameters

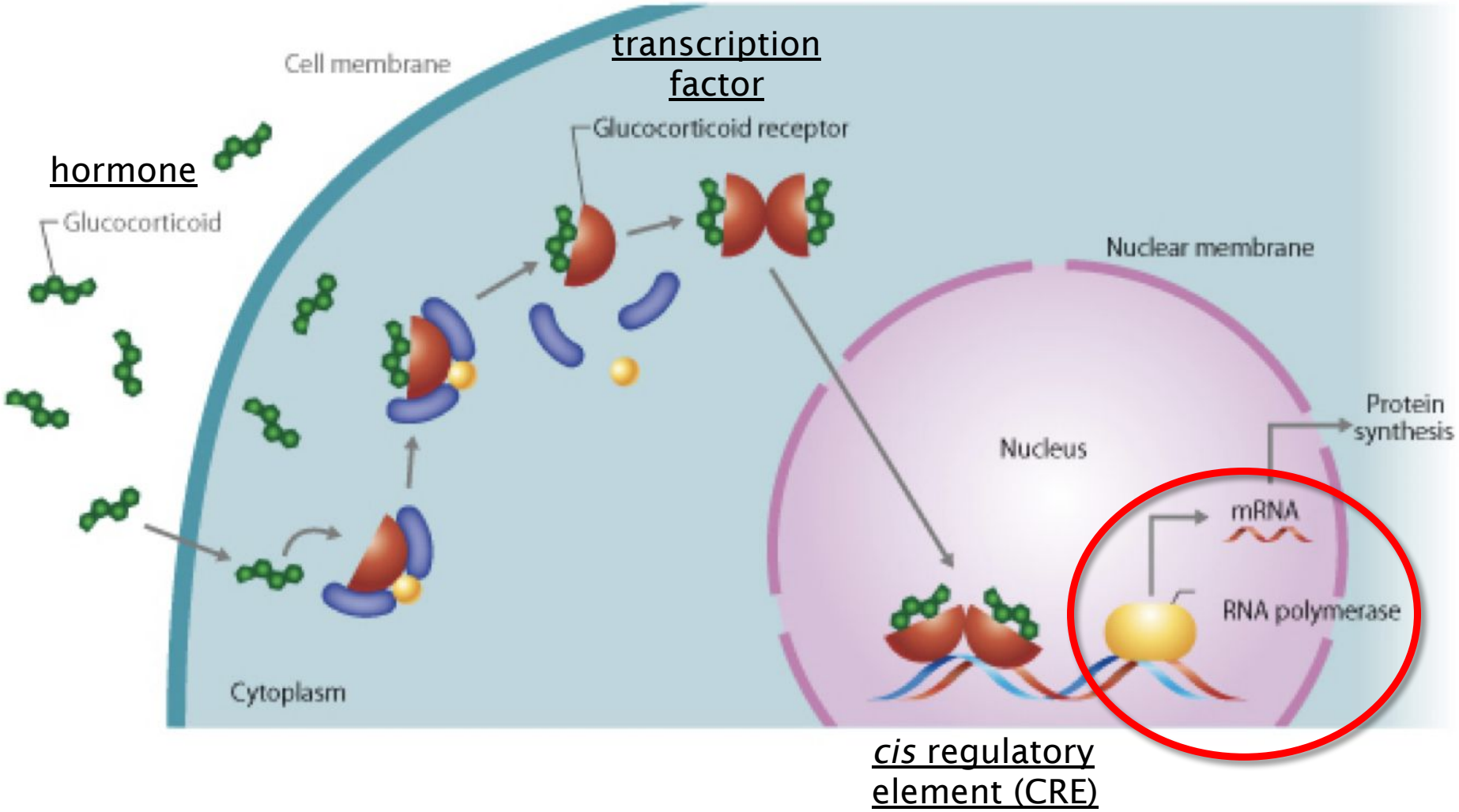


Tibshirani et al (2002)

Lab: what is airway transcriptome response to glucocorticoid hormone?

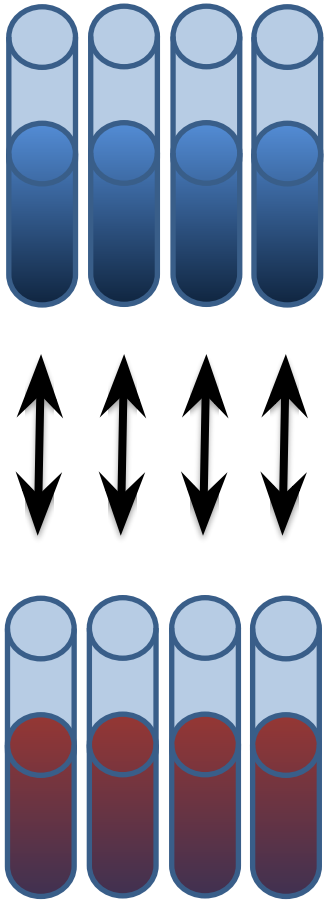


Glucocorticoid mechanism of action



Compare gene expression across treatment, within cell line

cDNA libraries



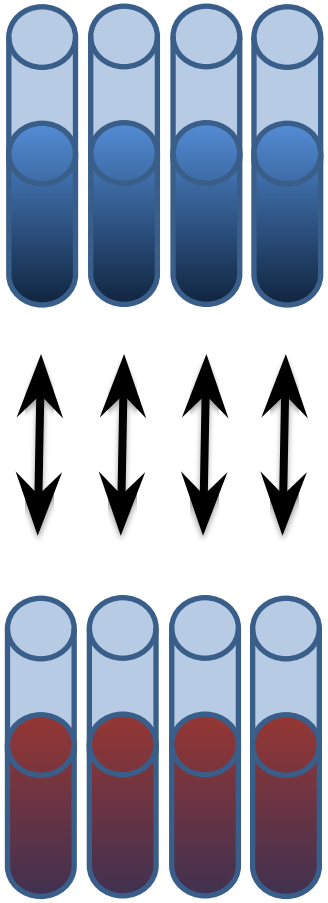
control

treated with dexamethasone

- ✔ Visualize differences between samples
- ✔ Test for differences in gene expression, one gene at a time
- ✔ Visualize differences across all genes

Compare gene expression across treatment, within cell line

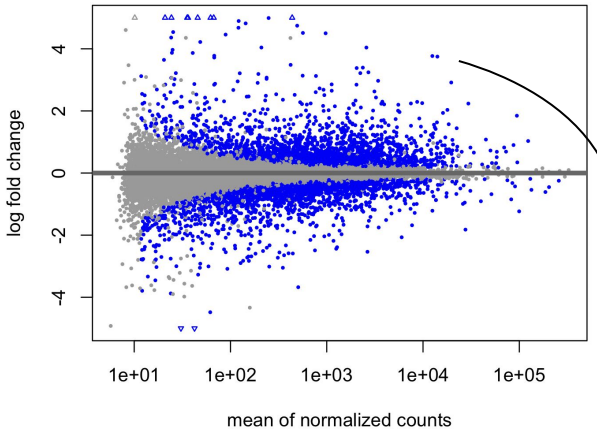
cDNA libraries



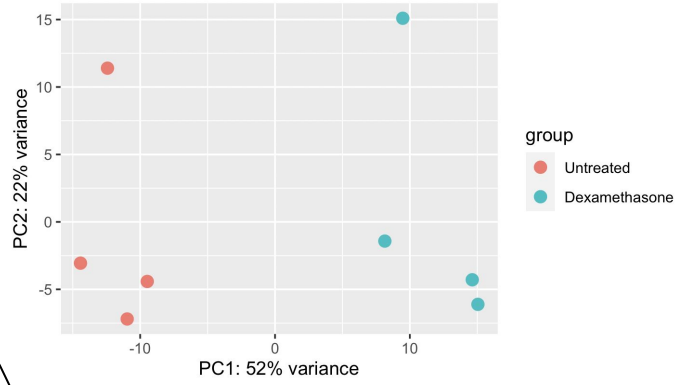
control

treated with dex.

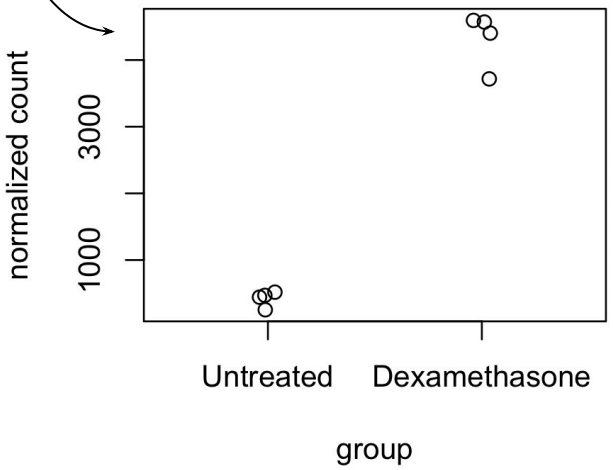
MA plot



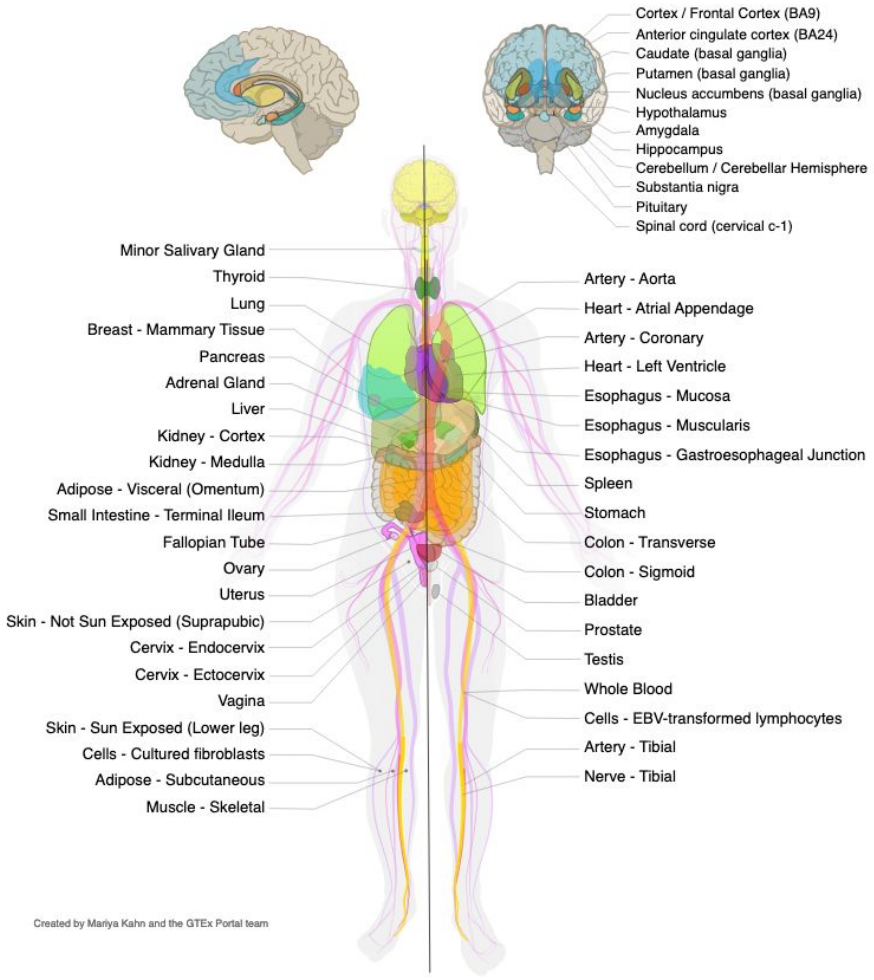
PCA plot



Counts plot



Does it make sense to do DE across tissues?



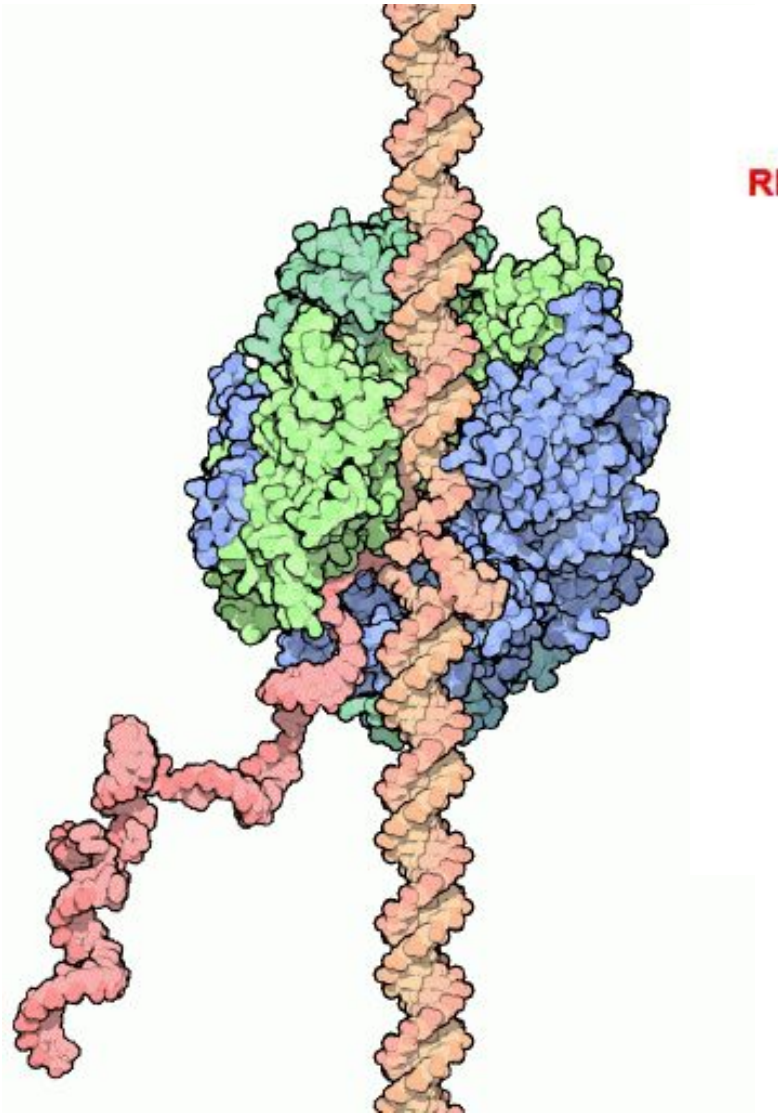
GTEx Portal

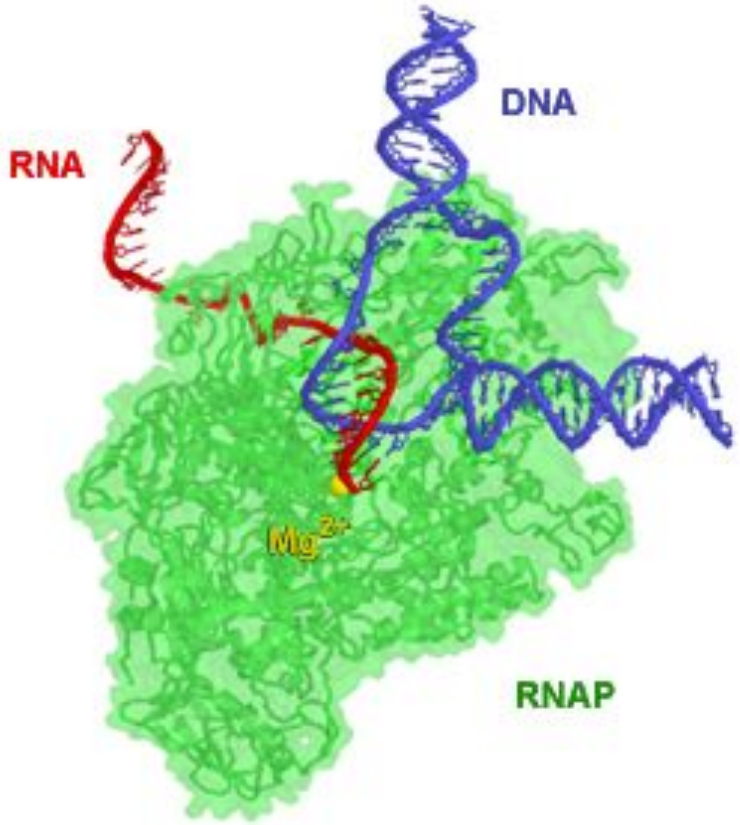
- Biology intro
- RNA-seq count table
- Batch effects and QC
- Quantification (reads → count table)
- Import into Bioconductor

Who has what?

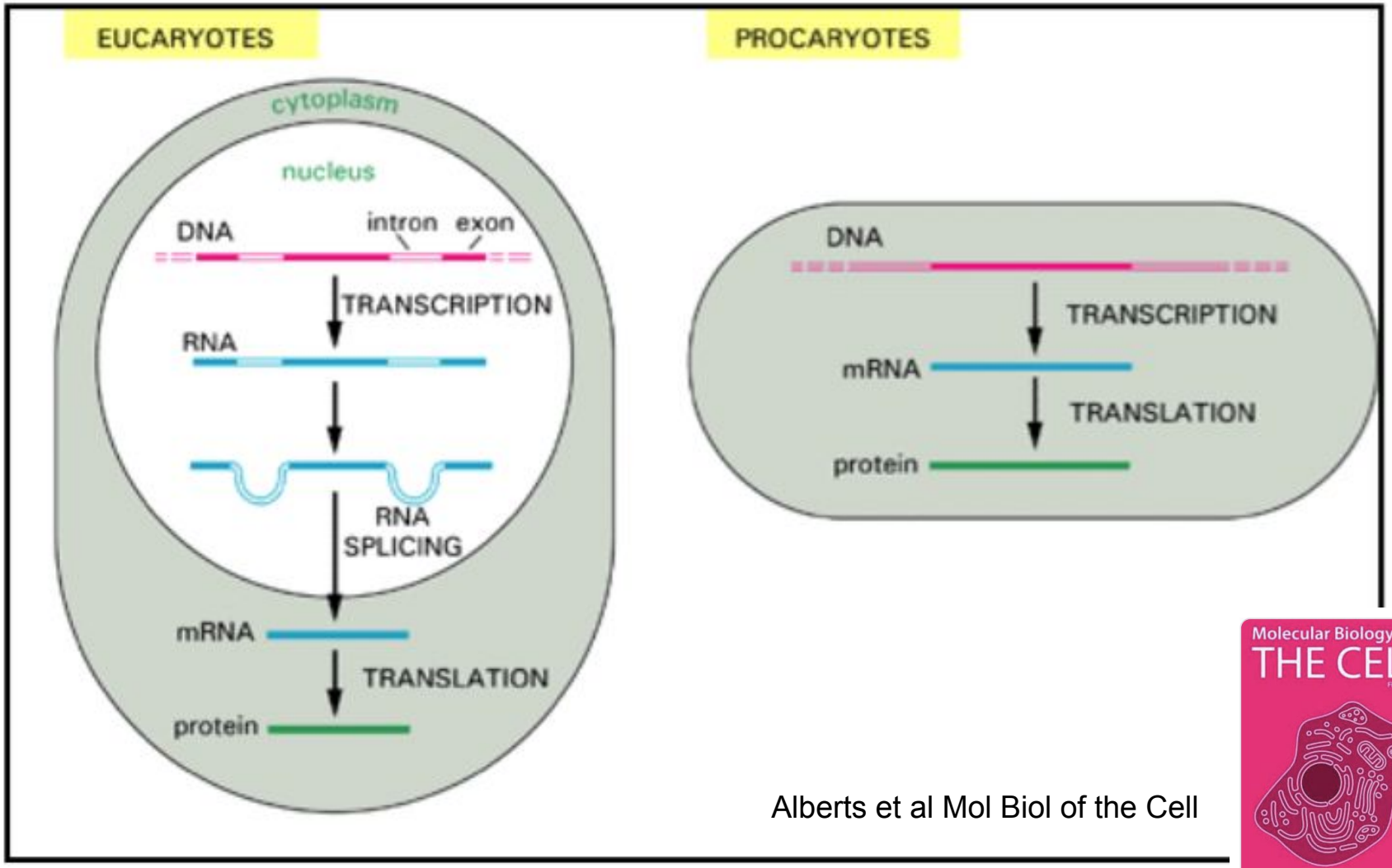
	prokaryote	eukaryote		
	Bacteria	Single cell fungus	Plant	Animal
Cell wall	✅	✅	✅ (rigid)	✅
Nucleus	?			
Splicing	🚫	🚫	✅	✅
Mitochondria	🚫	✅	✅	✅
Photosynthesis	?			
Chromosomes	circular	linear	linear	linear
Meiosis (sex)	?			
Horizontal gene transfer	✅	🚫	🚫	🚫

RNA polymerase = the DNA copier



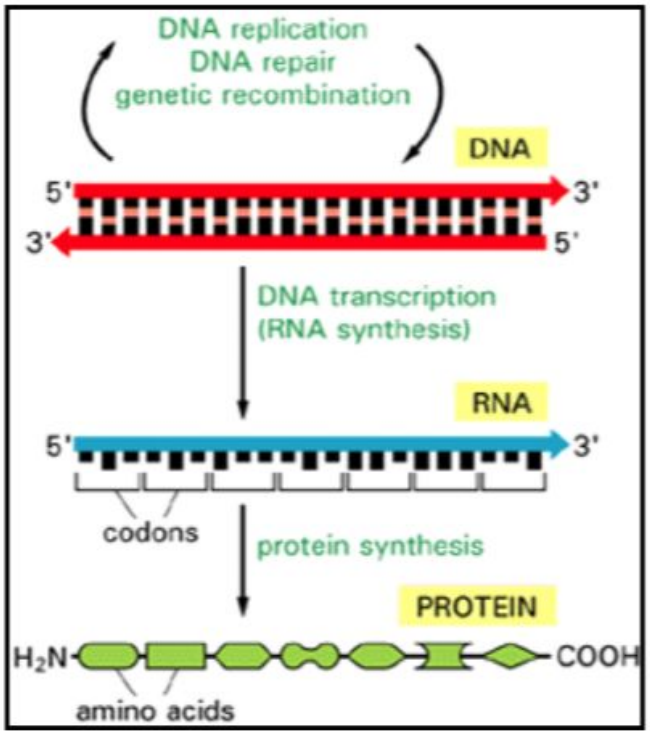


Simple diagram: transcription, splicing, & translation



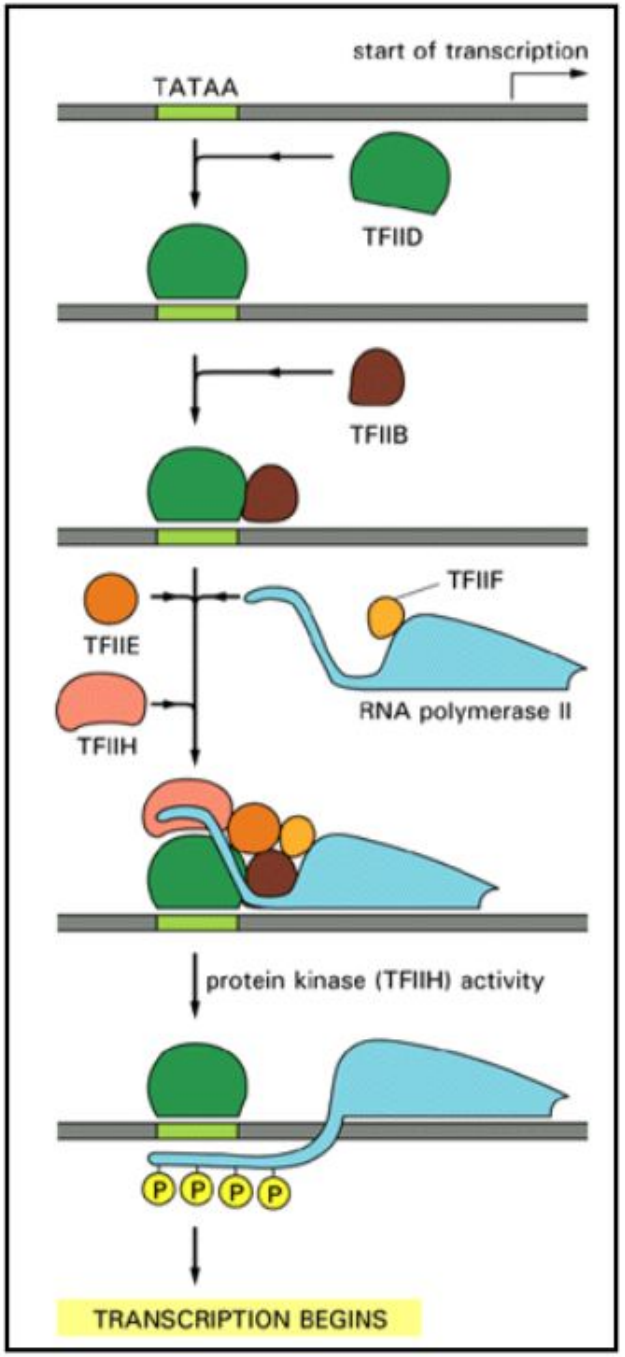
Alberts et al Mol Biol of the Cell

DNA → RNA → protein (usually)

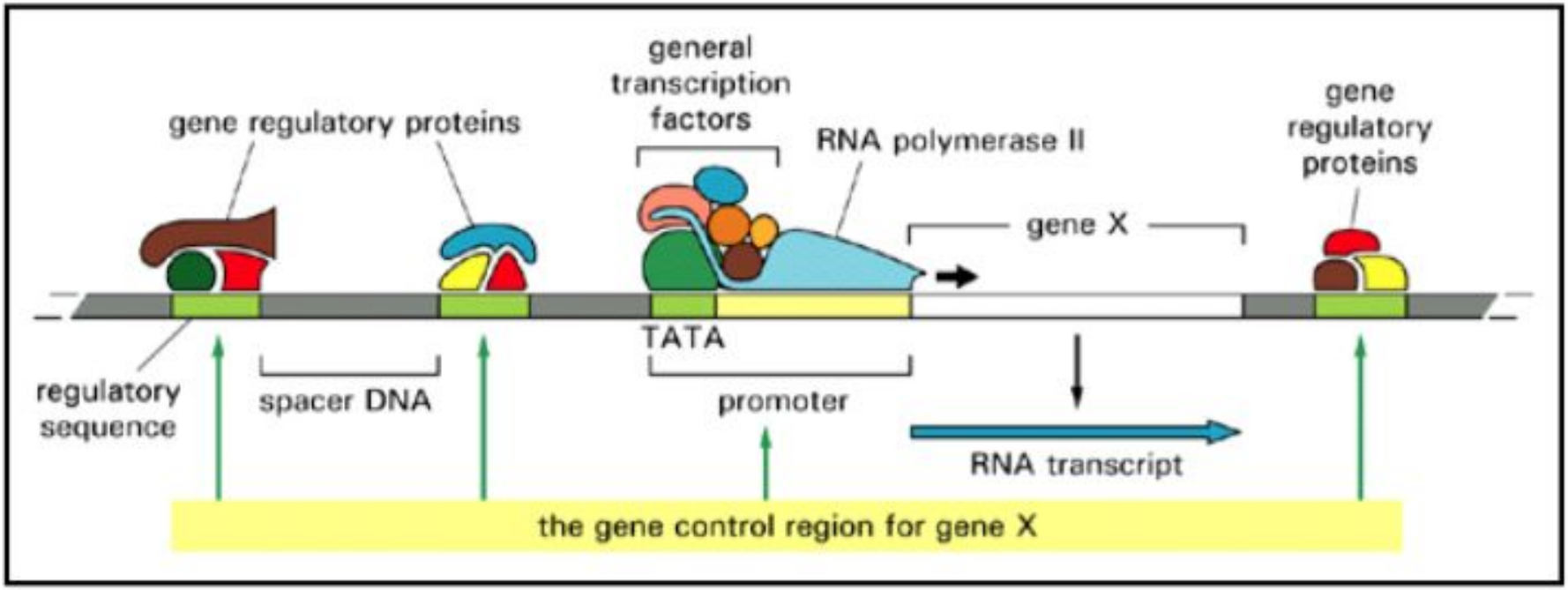


- By making more RNA, the cell makes more protein (usually)
- Proteins can then loop back and bind to, or change the properties of DNA
- This forms a regulatory network, kind of like a computer
- The system can have positive and negative feedback loops, logic, memory

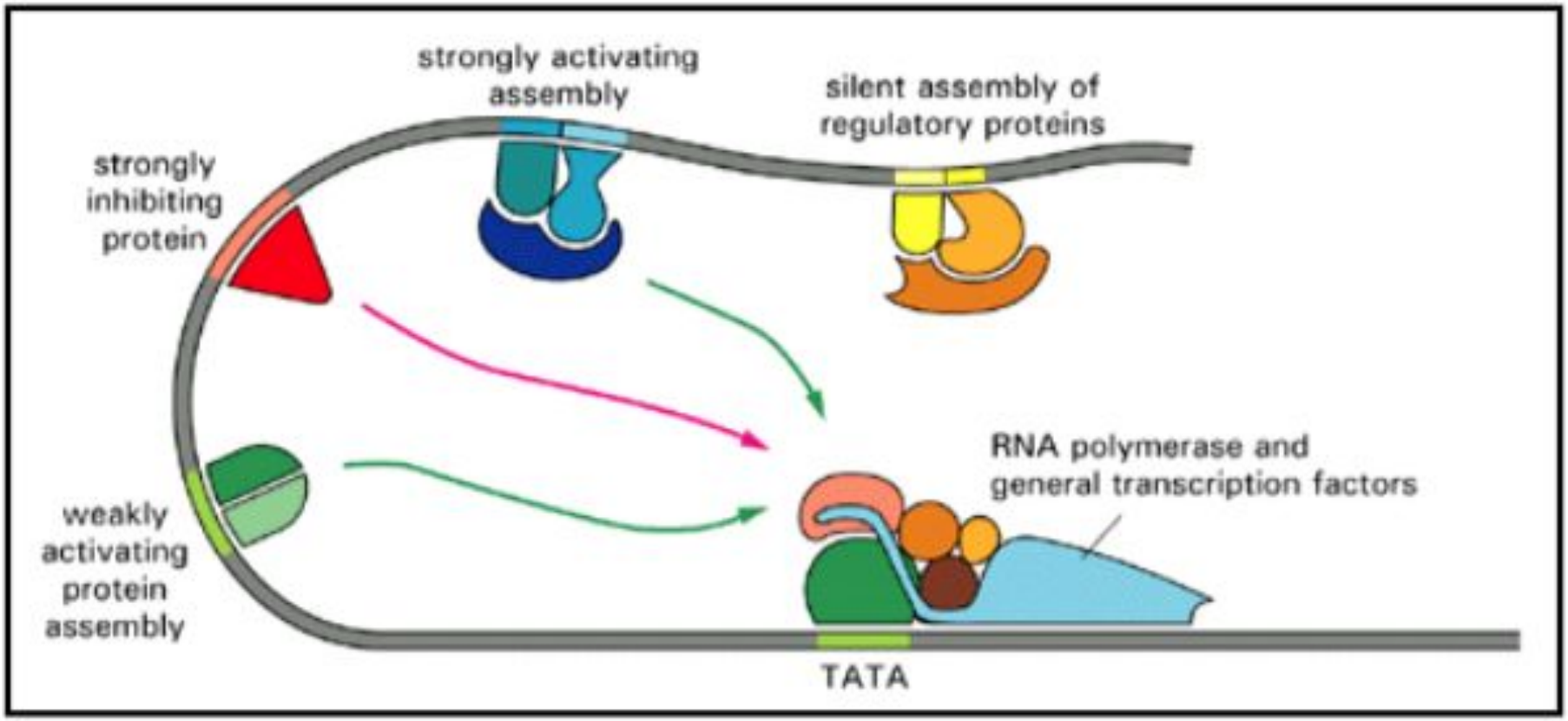
Transcription initiation involves a lot of proteins



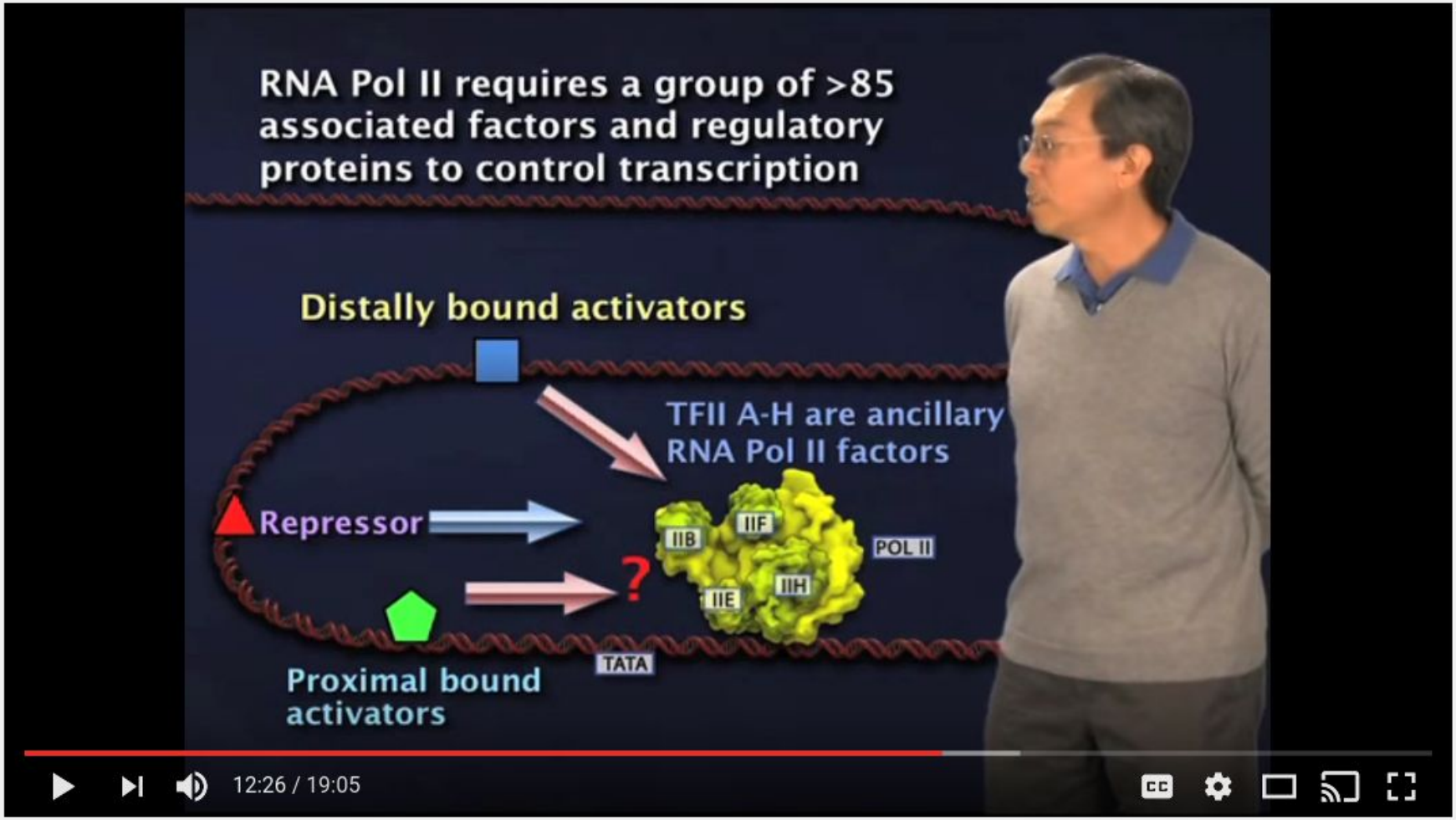
The whole picture: lots of *transcription factors*



Regulatory signals are integrated

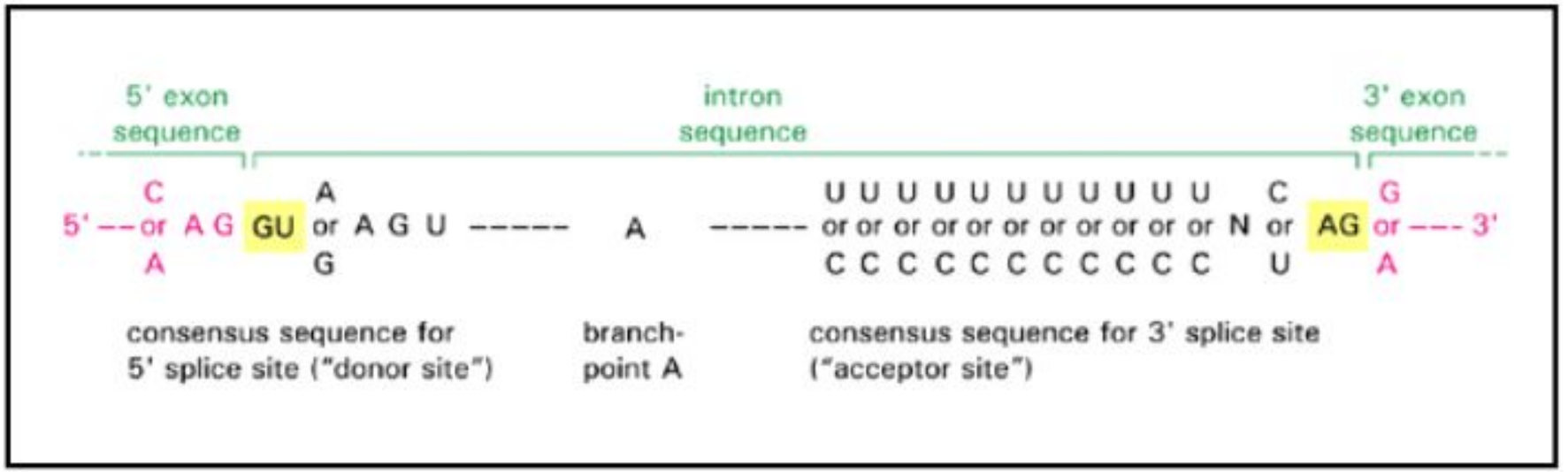


Robert Tjian (HHMI) video

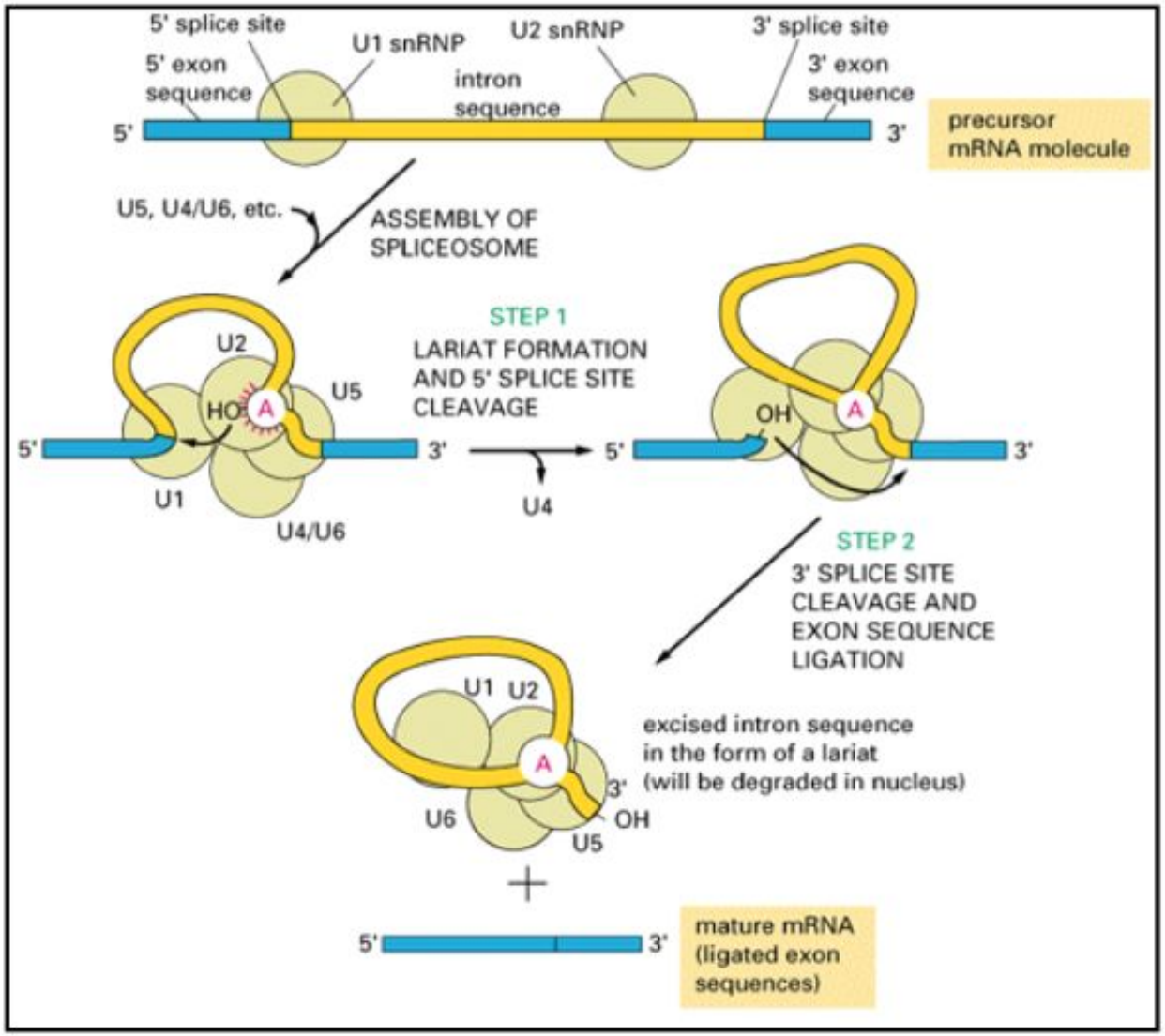


Introduction to Transcription - Robert Tjian (Berkeley/HHMI)

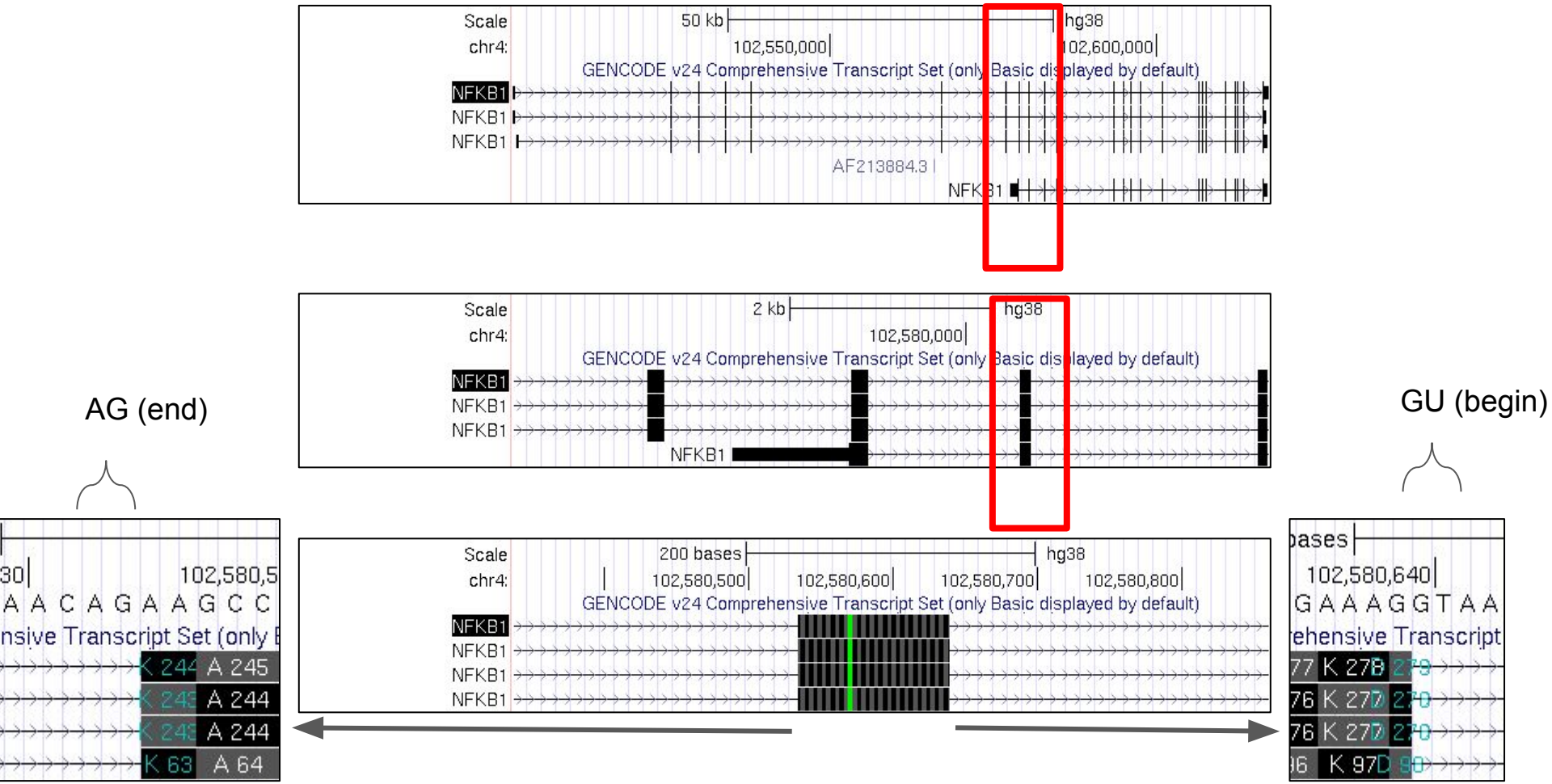
Splicing signal in the DNA sequence



Splicing proteins (spliceosome)



Typical gene

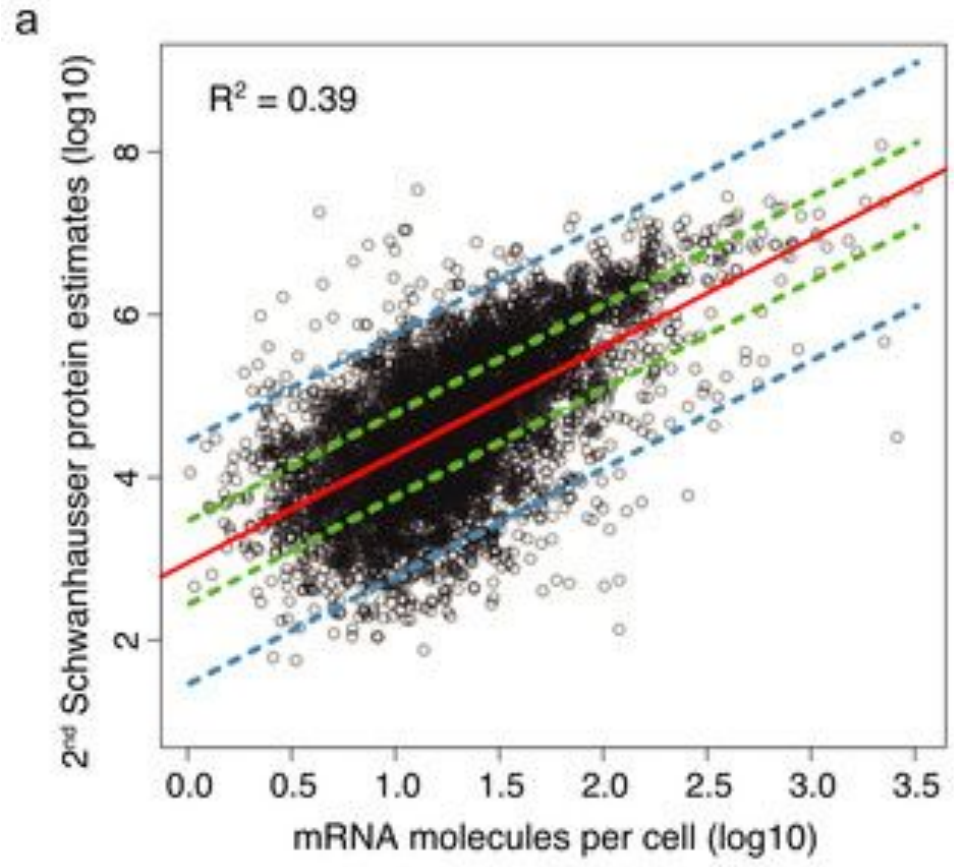


Typical gene is ~100kb in genome. Only ~4kb of *coding* sequence.
Also often *untranslated* regions in beginning and end.

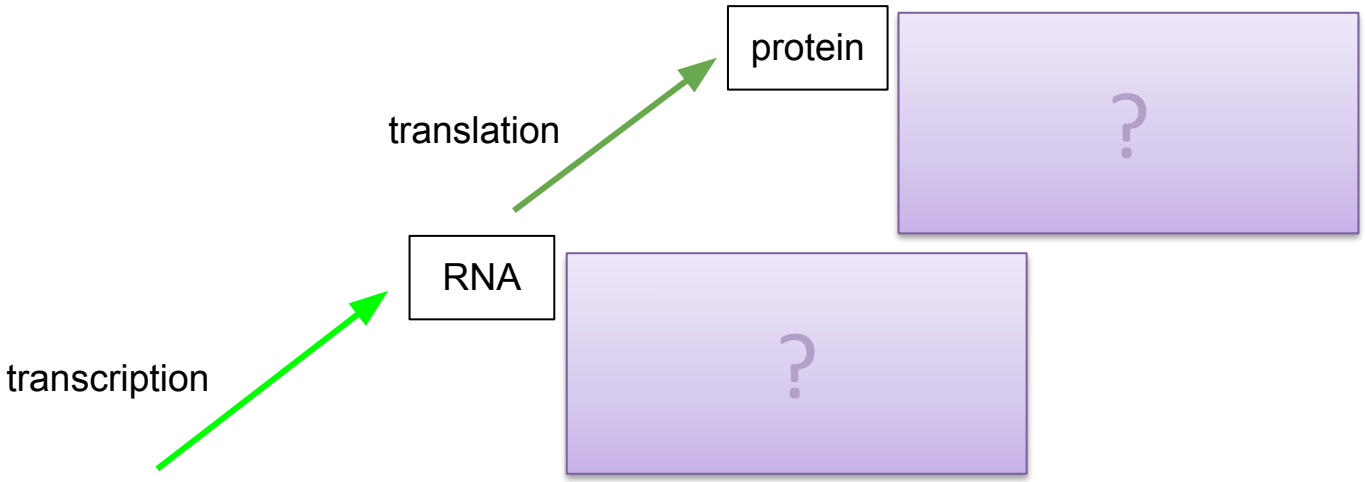
RNA and protein abundance?

RNA and protein abundance correlation is low?

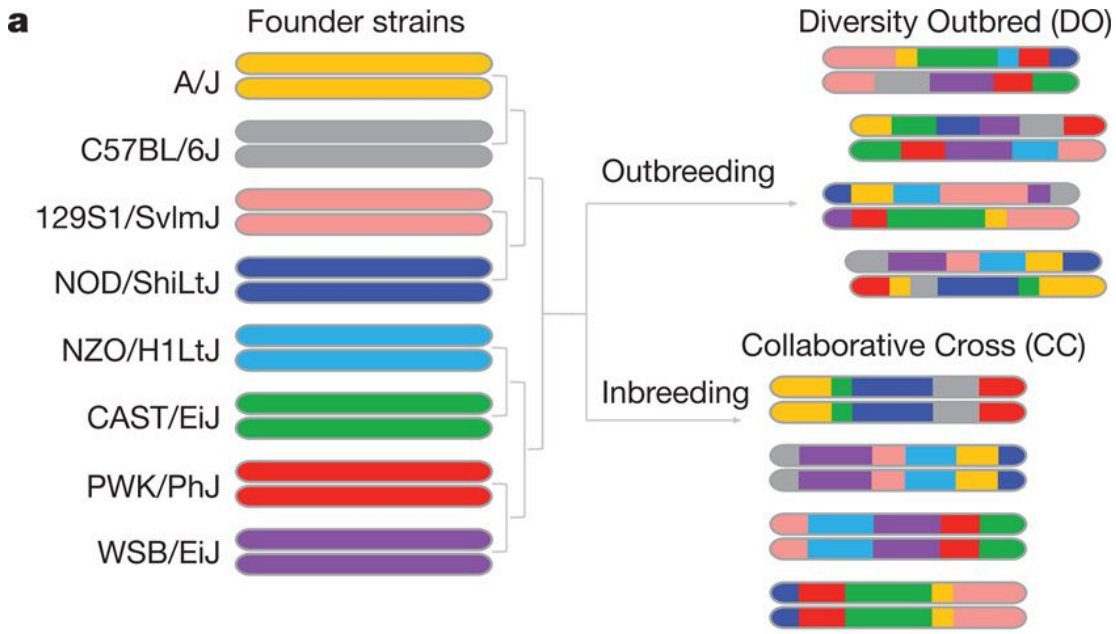
http://rajlaboratory.blogspot.com/2015/05/rna-doesnt-correlate-with-protein-huh.html

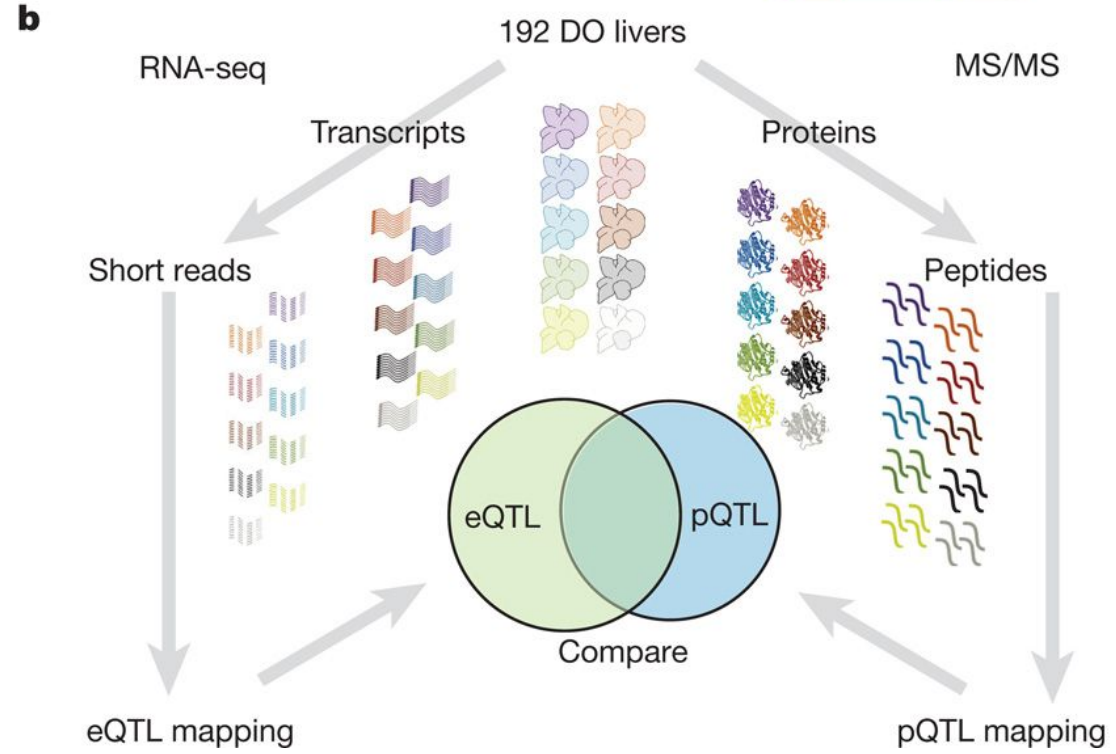


Have to think about steady state measurements



- “Do mRNA and protein levels correlate across all genes?”
- “If you increase the amount of mRNA, will you end up with more of the corresponding protein?”
- Answer: Often, yes.
  Look at fold change vs fold change.
  Or, look at abundance changes due to genotype ("QTL")

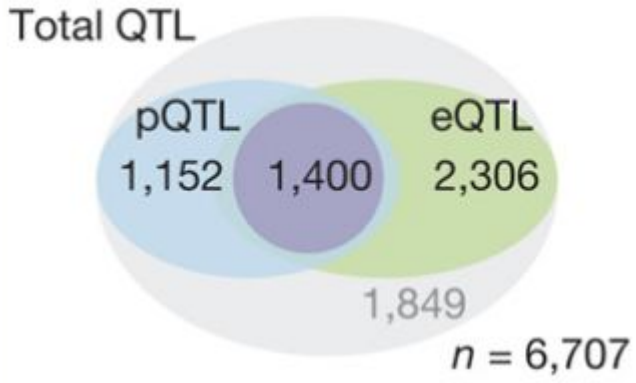


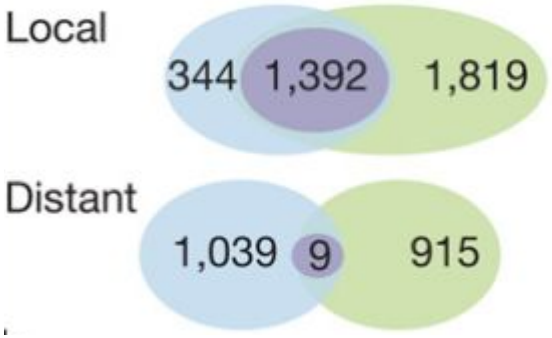


doi: 10.1038/nature18270
PMCID: PMC5292866

Defining the consequences of genetic variation on a proteome–wide scale

Joel M. Chick,* Steven C. Munger,* Petr Simecek, Edward L. Huttlin, Kwangbom Choi, Daniel M. Gatti, Narayanan Raghupathy, Karen L. Svenson, Gary A. Churchill, and Steven P. Gygi

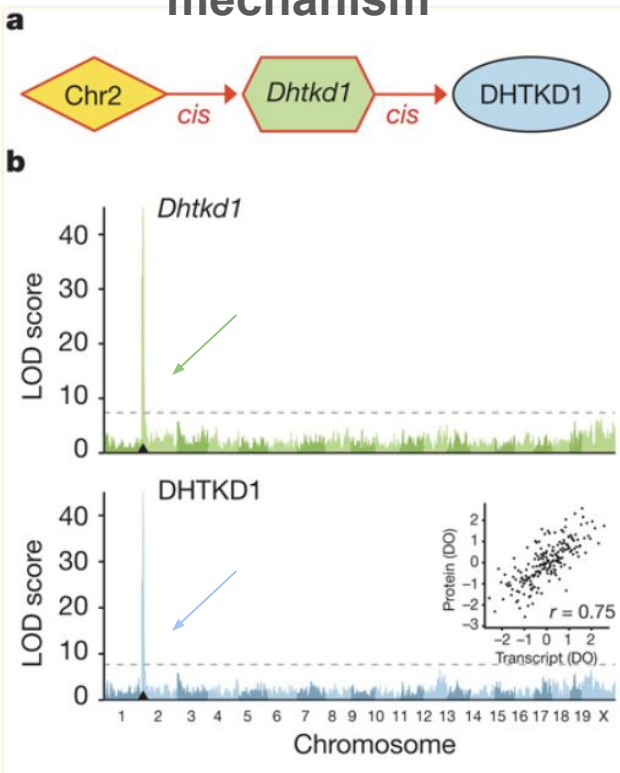




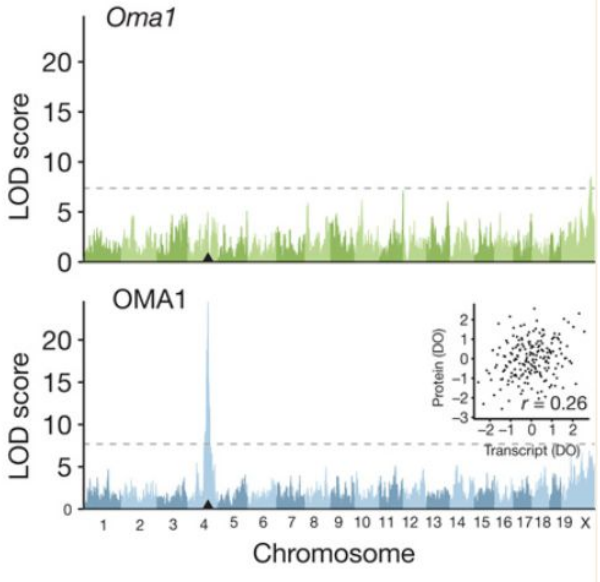
cor(DNA, RNA)

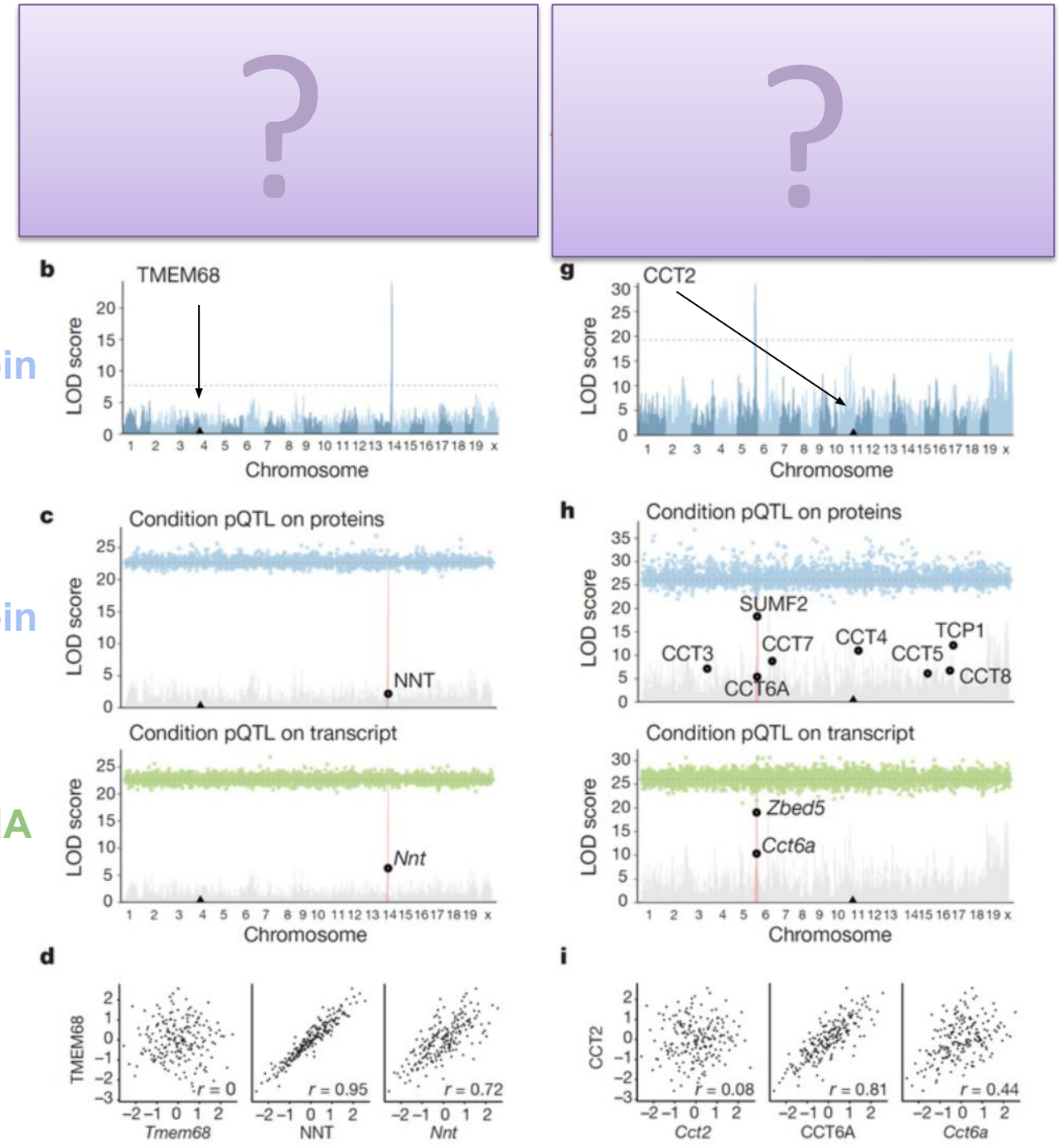
cor(DNA, protein)

mechanism



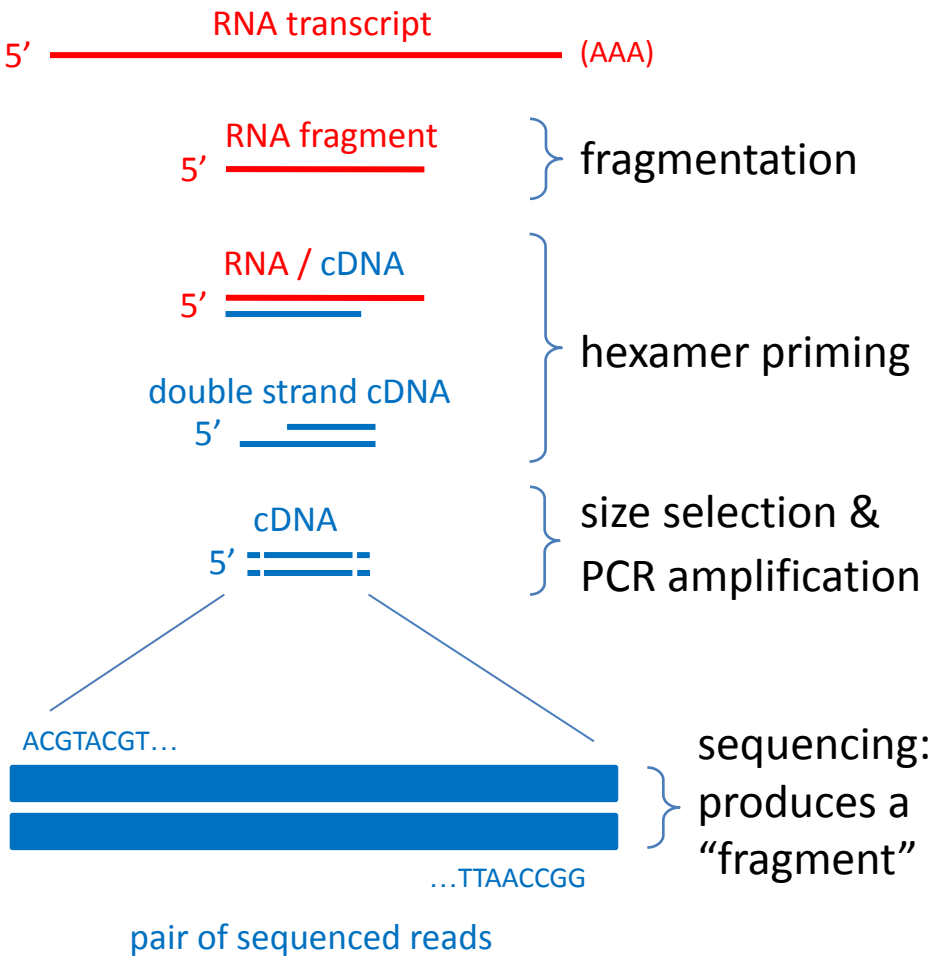
?



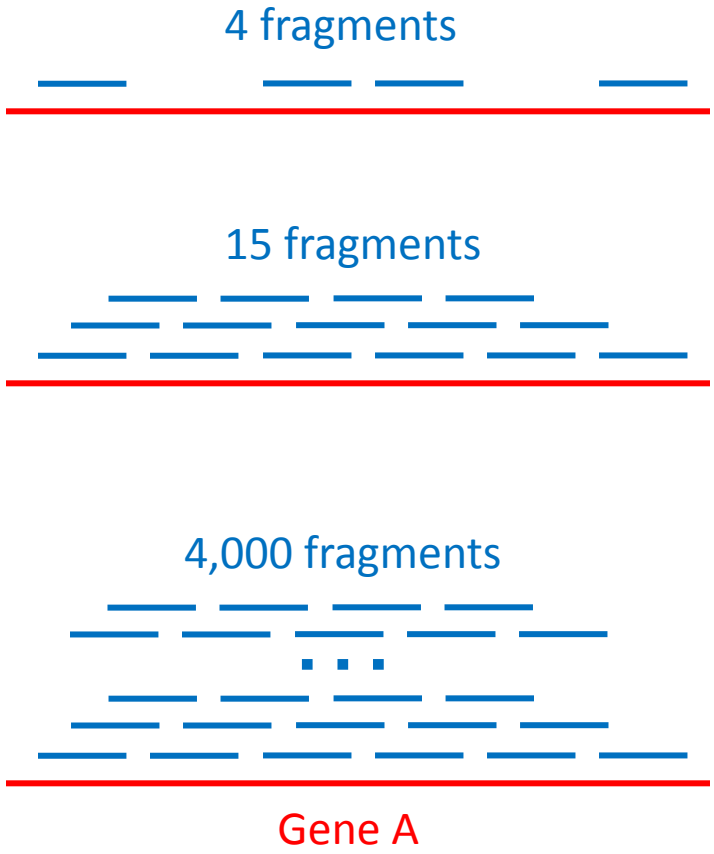

?
?
protein
protein
RNA
b
TMEM68
LOD score
Chromosome
g
CCT2
LOD score
Chromosome
c
Condition pQTL on proteins
NNT
Condition pQTL on transcript
Nnt
LOD score
Chromosome
h
Condition pQTL on proteins
SUMF2
CCT3
CCT7
CCT4
TCP1
CCT5
CCT8
CCT6A
Condition pQTL on transcript
Zbed5
Cct6a
LOD score
Chromosome
d
TMEM68
r = 0
Tmem68
r = 0.95
NNT
r = 0.72
Nnt
i
CCT2
r = 0.08
Cct2
r = 0.81
CCT6A
r = 0.44
Cct6a

- Biology intro
- RNA-seq count table
- Batch effects and QC
- Quantification (reads → count table)
- Import into Bioconductor

RNA (cDNA) short read sequencing

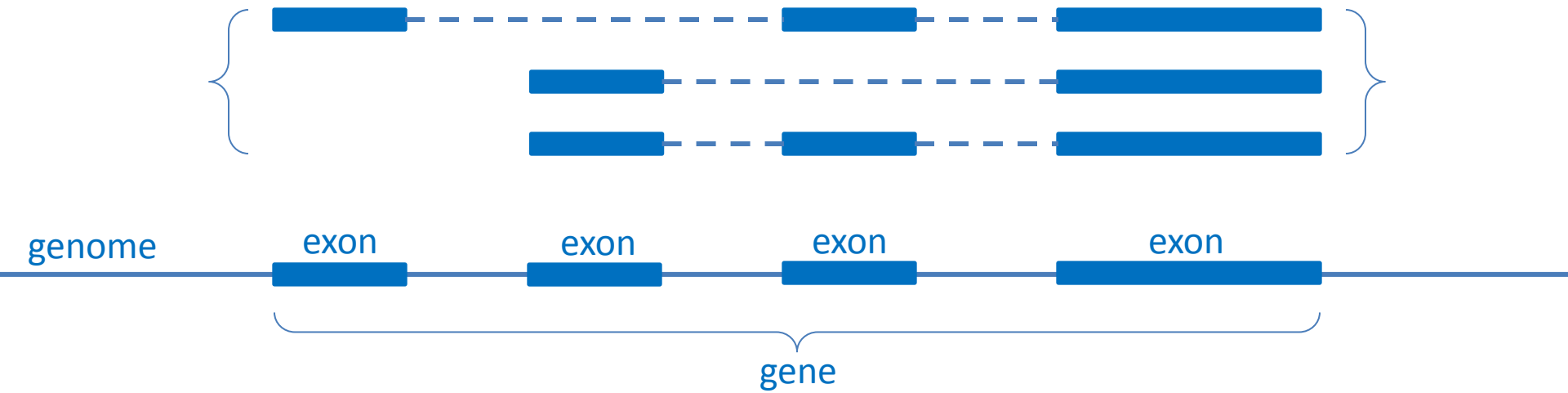


High dynamic range:



More complex

- Gene: region of genome
- Three “isoforms”, also called “transcripts”

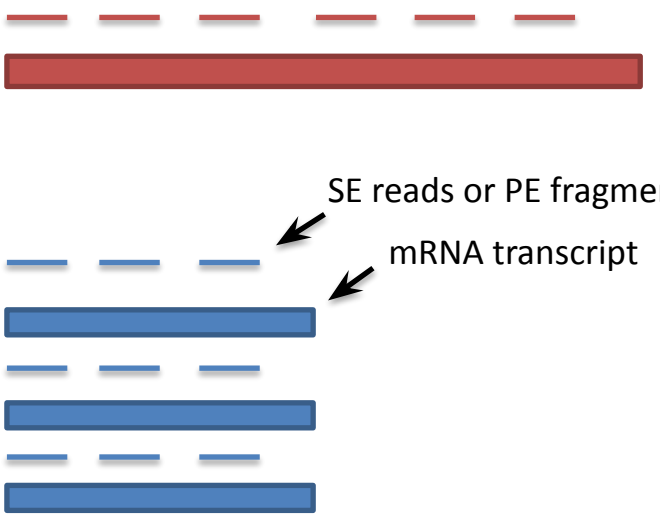


mRNAs to short read fragments

colors: different genes

C_{ij} = count of fragments aligned to gene i, sample j

is proportional to:



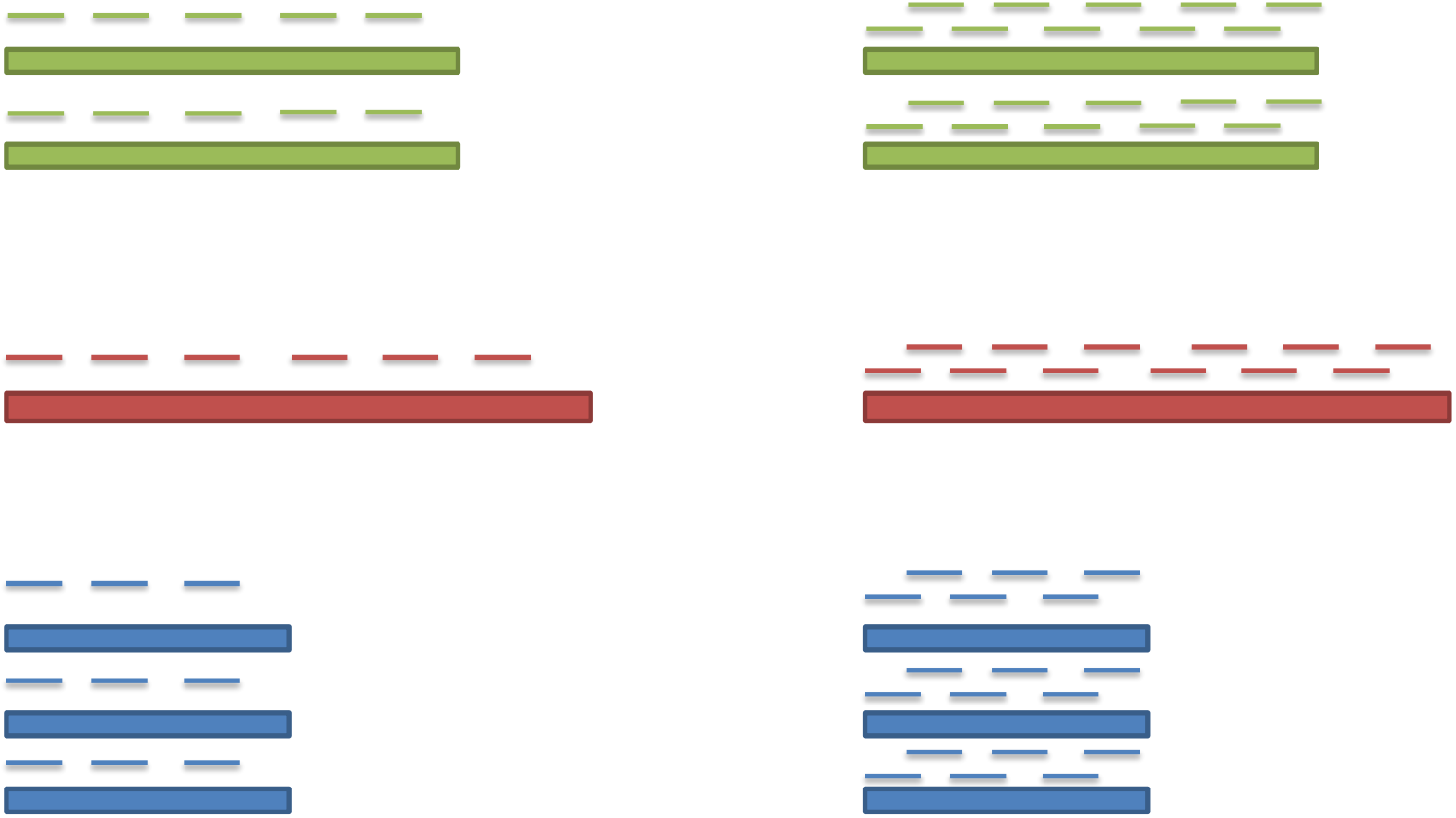
?

Sequencing depth

more tomorrow from Charlotte

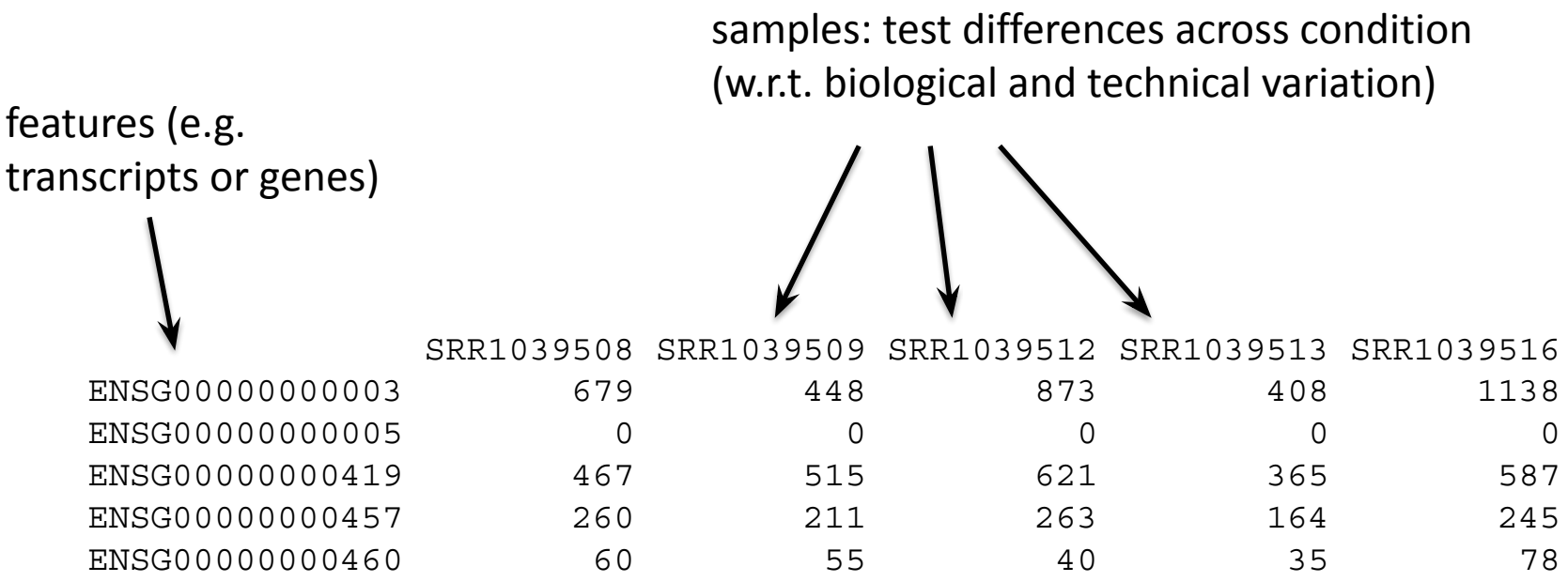
sample 1

sample 2



The cDNA count table

- observed data consists of *counts* of fragments across features (rows) and samples (columns)
- more about how we count later in this lecture (*htseq*, *featureCounts*, *Salmon*, etc.)



	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Ready to go?

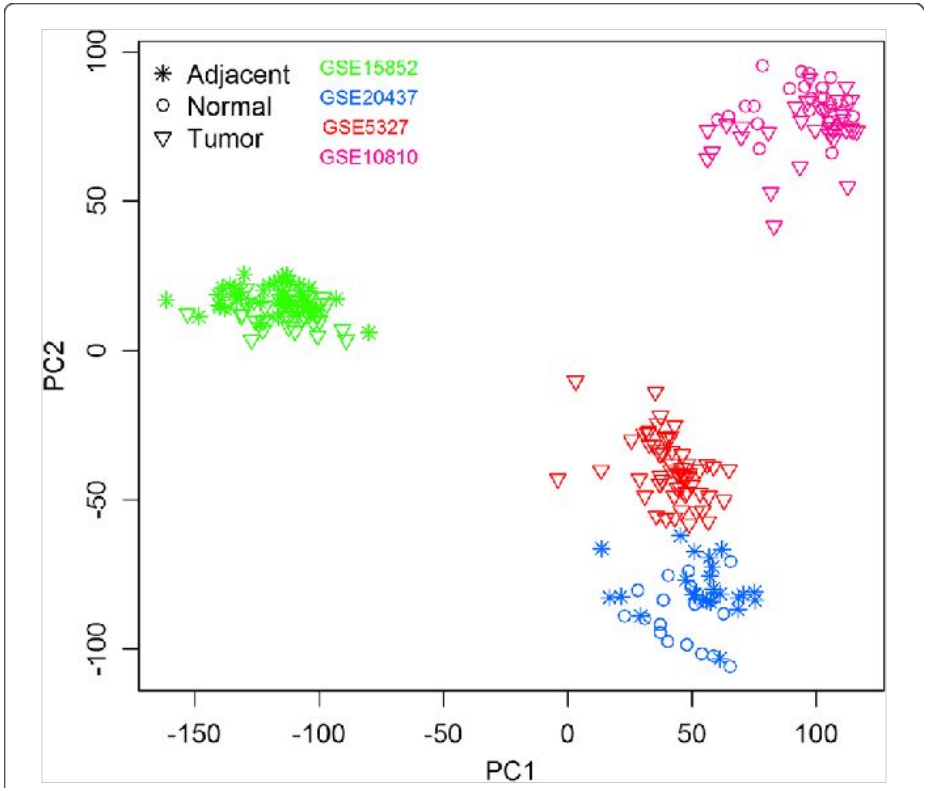
Comparing these counts across samples gives the DE genes?

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

- Biology intro
- RNA-seq count table
- Batch effects and QC
- Quantification (reads → count table)
- Import into Bioconductor

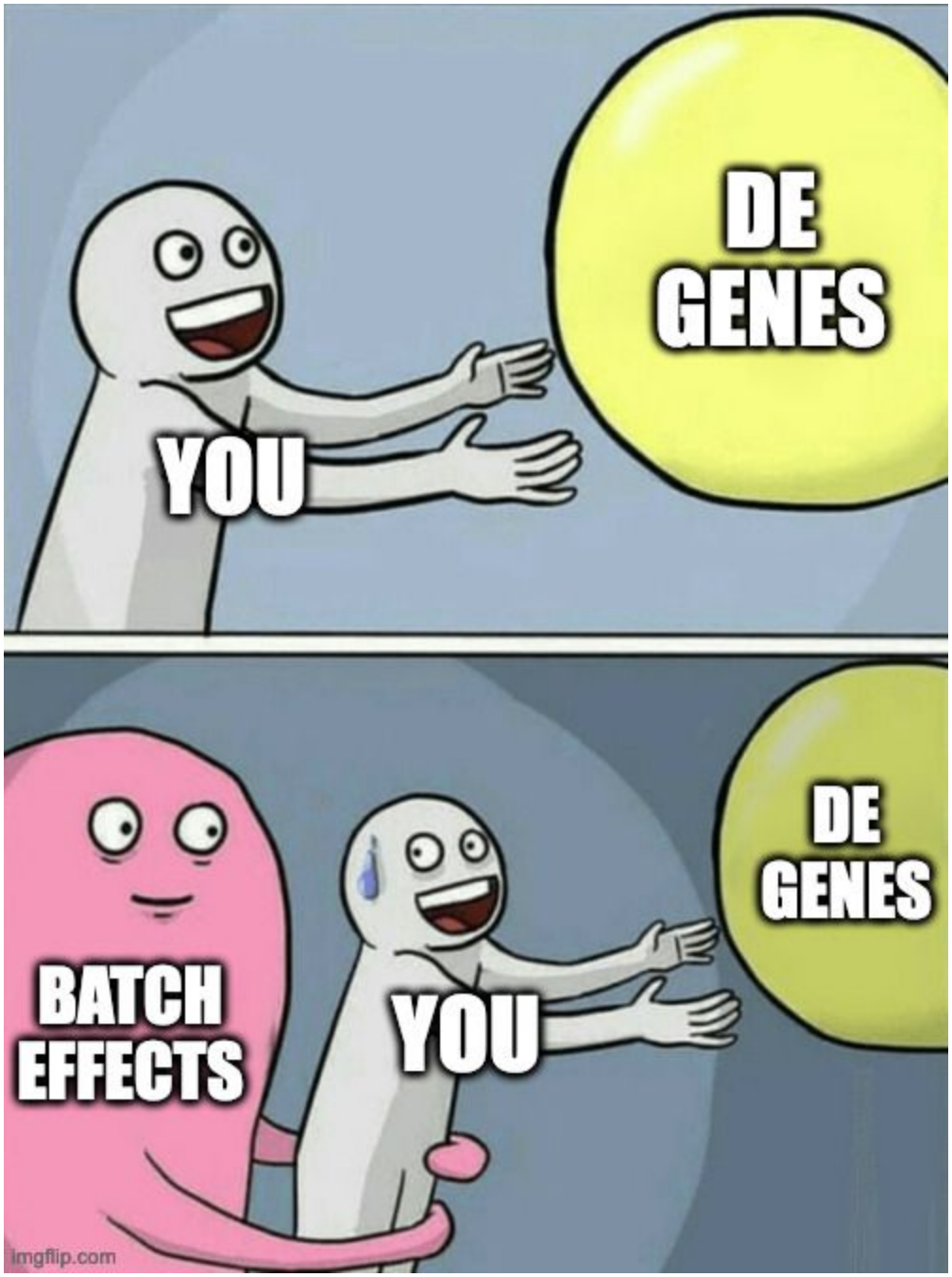
Batch effects!

Sample correlations =
Factors of unwanted variation (RUV) =
Surrogate variables (sva)

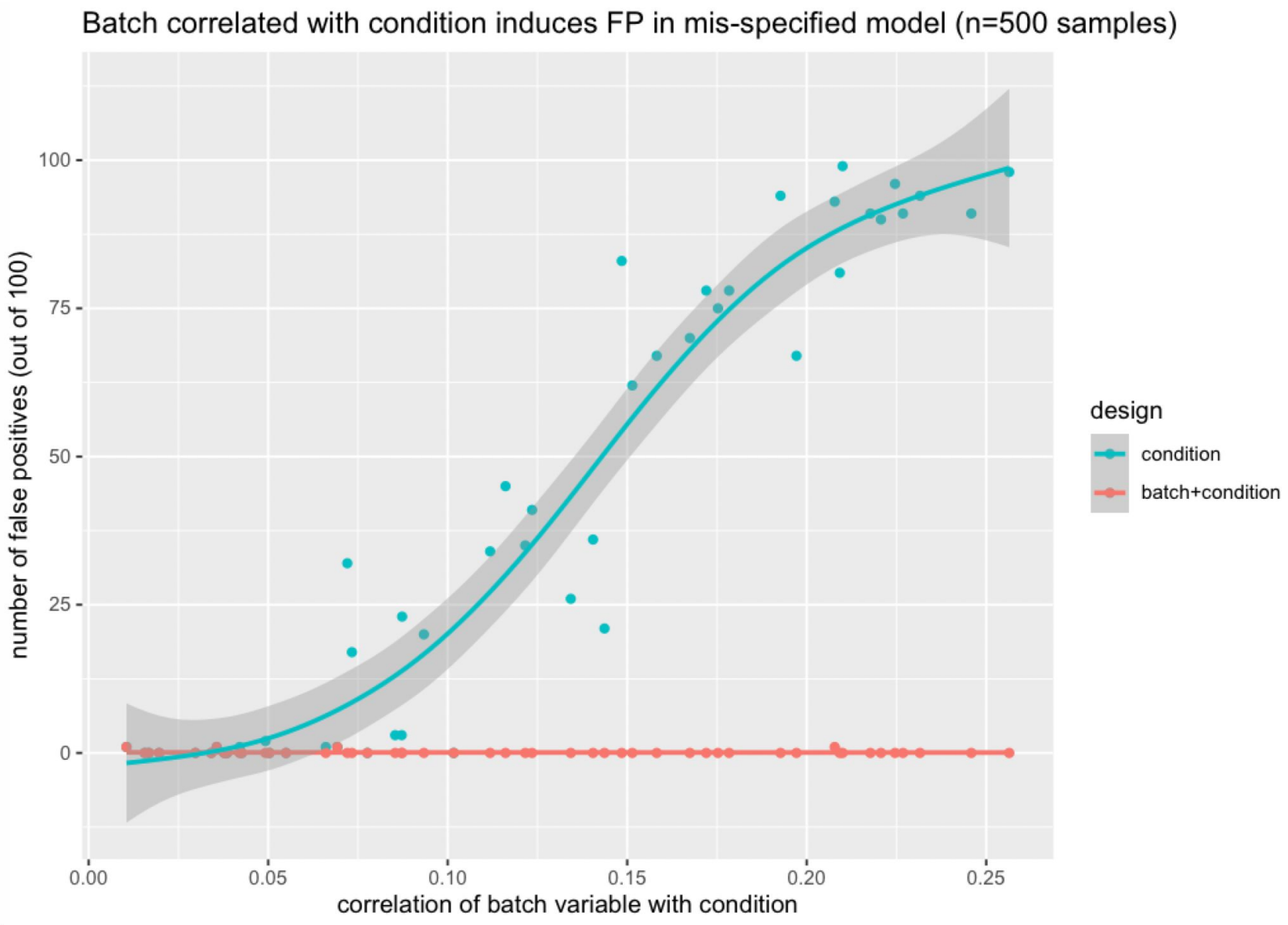


Impact on your results:

- sample swaps / mis-labelled data
- batch effects
- …
- …
- which DE tool
- colour vs color in ggplot


DE
GENES
YOU
DE
GENES
BATCH
EFFECTS
YOU
imgflip.com

Even low correlations are a problem for large *N*



More in the DESeq2 vignette

- DESeq2.html#multi-factor-designs
- mikelove/preNivolumabOnNivolumab

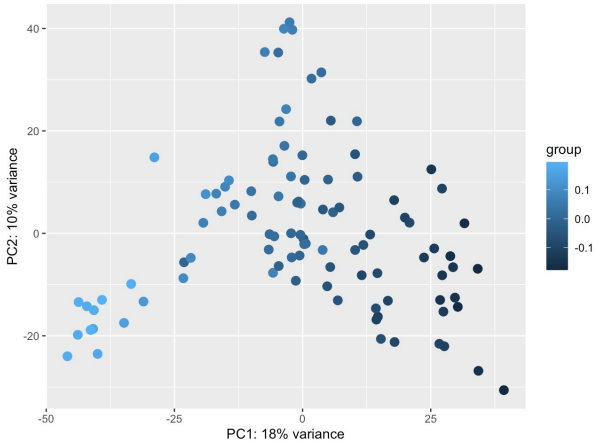
Multi-factor designs

Experiments with more than one factor influencing the counts can be analyzed using design formula that include the additional variables. In fact, DESeq2 can analyze any possible experimental design that can be expressed with fixed effects terms (multiple factors, designs with interactions, designs with continuous variables, splines, and so on are all possible).

By adding variables to the design, one can control for additional variation in the counts. For example, if the condition samples are balanced across experimental batches, by including the `batch` factor to the design, one can increase the sensitivity for finding differences due to `condition`. There are multiple ways to analyze experiments when the additional variables are of interest and not just controlling factors (see section on interactions).

Experiments with many samples: in experiments with many samples (e.g. 50, 100, etc.) it is highly likely that there will be technical variation affecting the observed counts. Failing to model this additional technical variation will lead to spurious results. Many methods exist that can be used to model technical variation, which can be easily included in the DESeq2 design to control for technical variation which estimating effects of interest. See the RNA-seq workflow for examples of using RUV or SVA in combination with DESeq2. For more details on why it is important to control for technical variation in large sample experiments, see the following thread, also archived here by Frederik Ziebell.

"don't ignore batch"



Differences across condition

i indexes genes

log2 of RNA abundance
(controlling for library size)
gene *i*, sample *j*

$$\log_2 q_{ij} = \sum_r x_{jr}\beta_{ir}$$

r indexes coefficients:

r = 0 Intercept

r = 1 condition B vs A

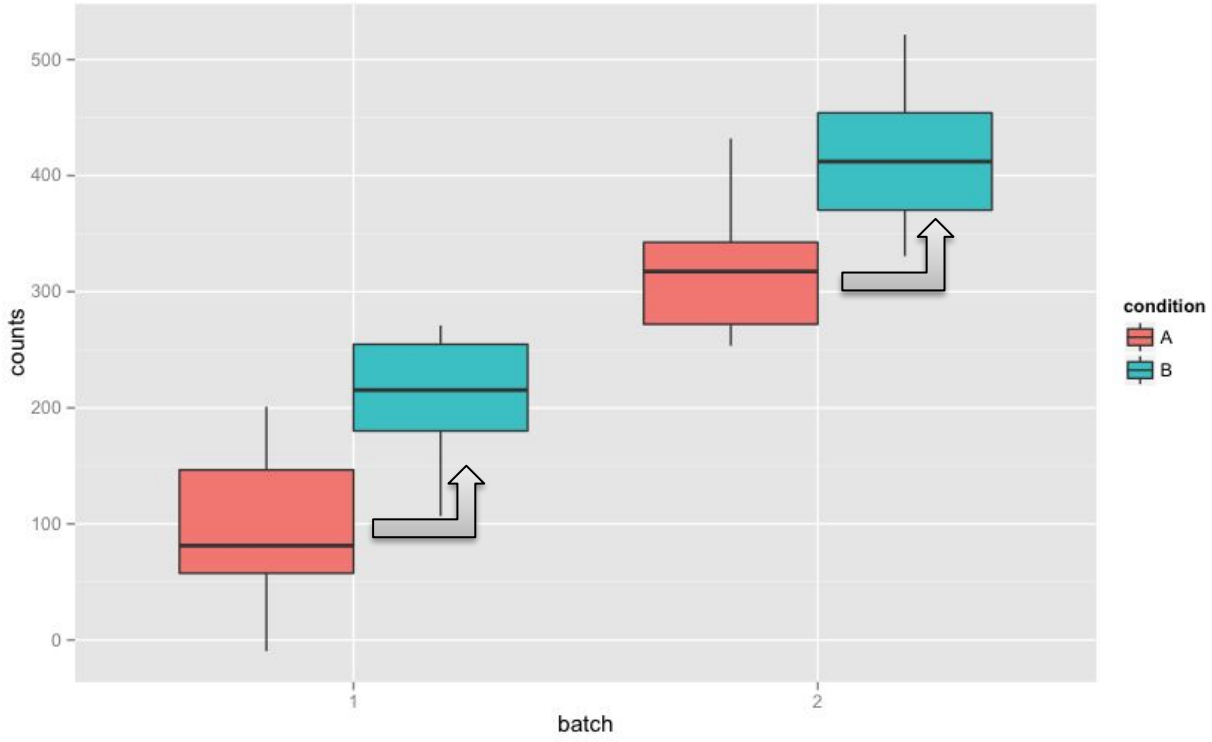
j indexes samples:

x_{11} = 0 : sample 1 is in condition A

x_{41} = 1 : sample 4 is in condition B

Controlling for different batches

- Using a design formula: ~batch + condition, adds coefficients that control for batch differences
- If batches are unknown, possible to detect these with other methods: svaseq, RUVSeq



Differences across condition

- Describe experiment with formula, e.g.: ~batch + condition
- Formulae are a convenient and powerful tool

$\log_2$ q1		1	0	0
$\log_2$ q2		1	0	0
$\log_2$ q3		1	1	0
$\log_2$ q4	=	1	1	0
$\log_2$ q5		1	0	1
$\log_2$ q6		1	0	1
$\log_2$ q5		1	1	1
$\log_2$ q6		1	1	1

Intercept
batch2
conditionB

more tomorrow from Charlotte

Differences across condition

- Describe experiment with formula, e.g.: ~batch + condition
- Formulae are a convenient and powerful tool

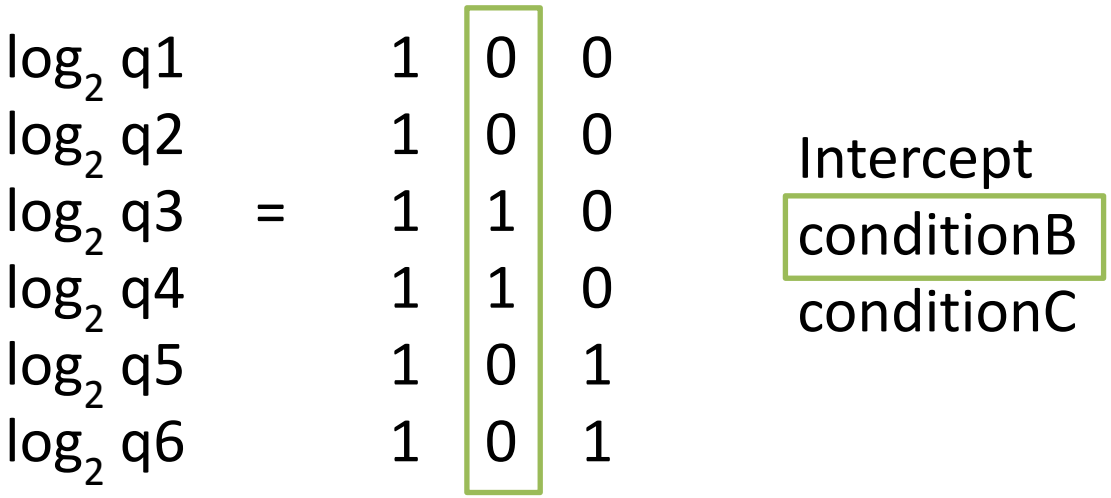
$\log_2$ q1		1	0	0
$\log_2$ q2		1	0	0
$\log_2$ q3		1	1	0
$\log_2$ q4	=	1	1	0
$\log_2$ q5		1	0	1
$\log_2$ q6		1	0	1
$\log_2$ q5		1	1	1
$\log_2$ q6		1	1	1

Intercept
batch2
conditionB

```
results(dds, name="condition_B_vs_A")
```

Differences across condition

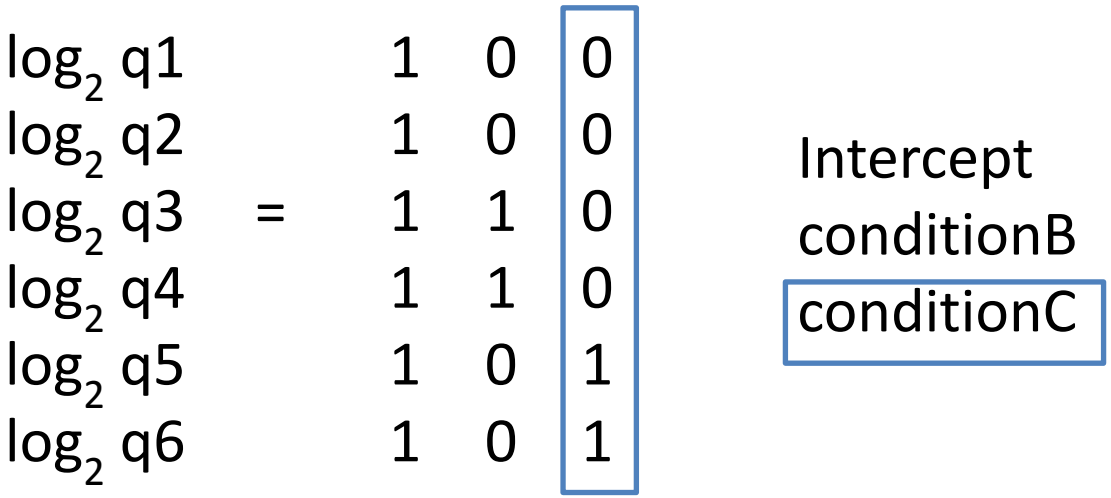
- Describe experiment with formula, e.g.: ~condition
- Formulae are a convenient and powerful tool




```
results(dds, contrast=c("condition","B","A"))
```

Differences across condition

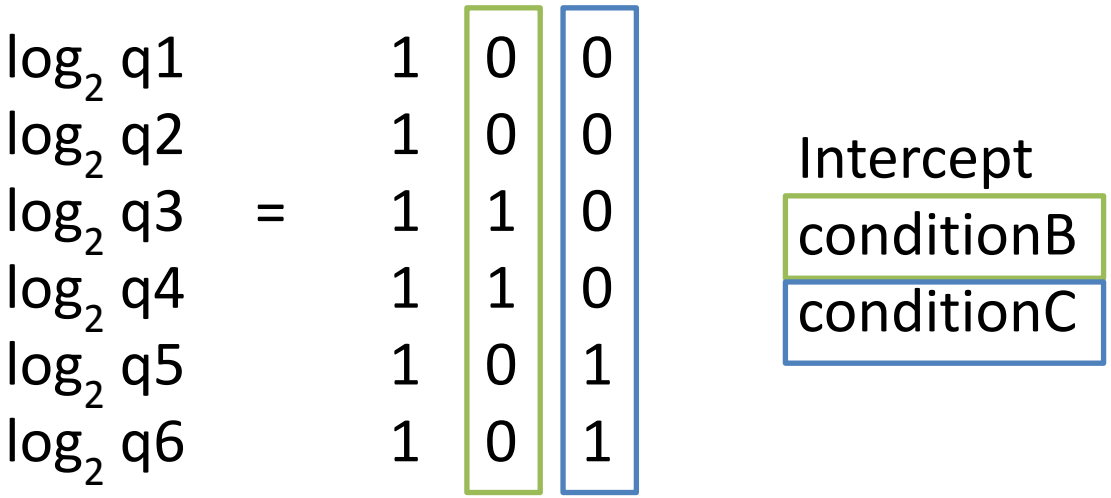
- Describe experiment with formula, e.g.: ~condition
- Formulae are a convenient and powerful tool




```
results(dds, contrast=c("condition","C","A"))
```

Differences across condition

- Describe experiment with formula, e.g.: ~condition
- Formulae are a convenient and powerful tool

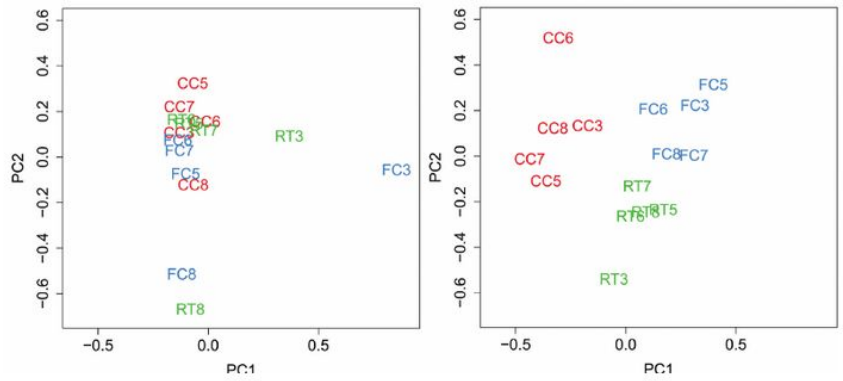



```
results(dds, contrast=c("condition","C","B"))
```

QC and remedies for batch effects

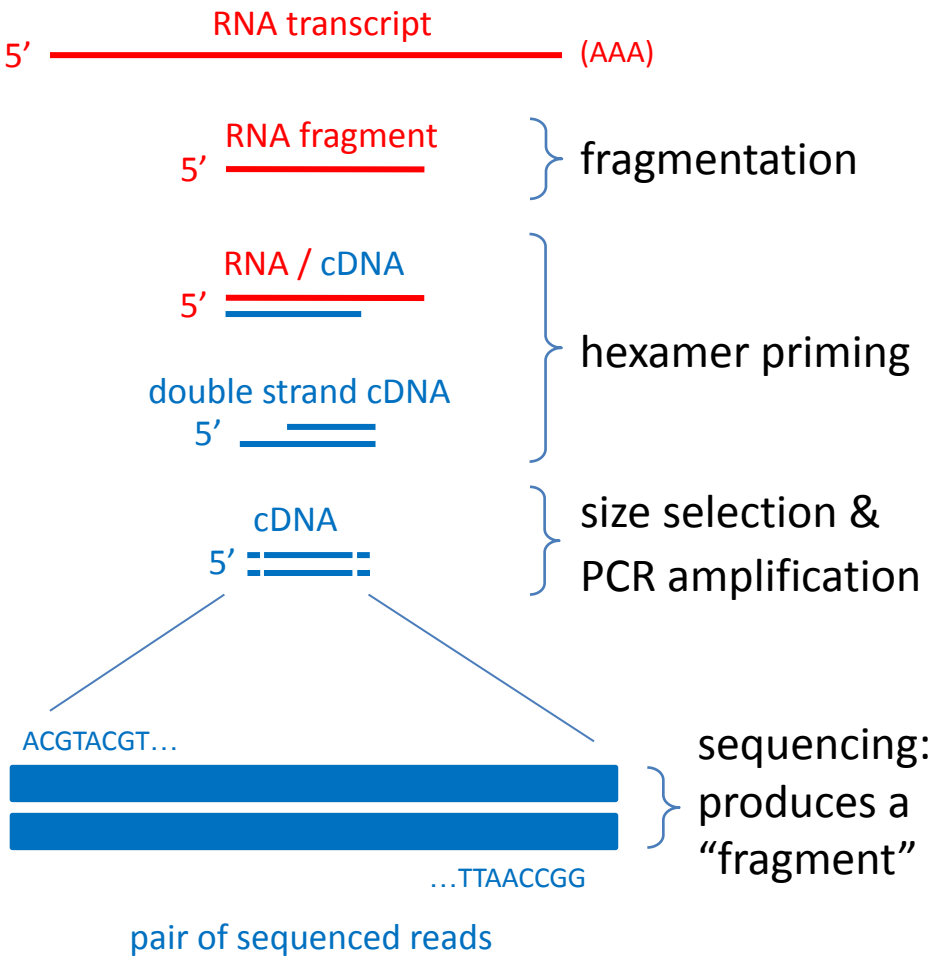
- ✓ Include known batch in design
- Estimate biases during quant
- MultiQC inspection
- Factor analysis (after quant)

RUV

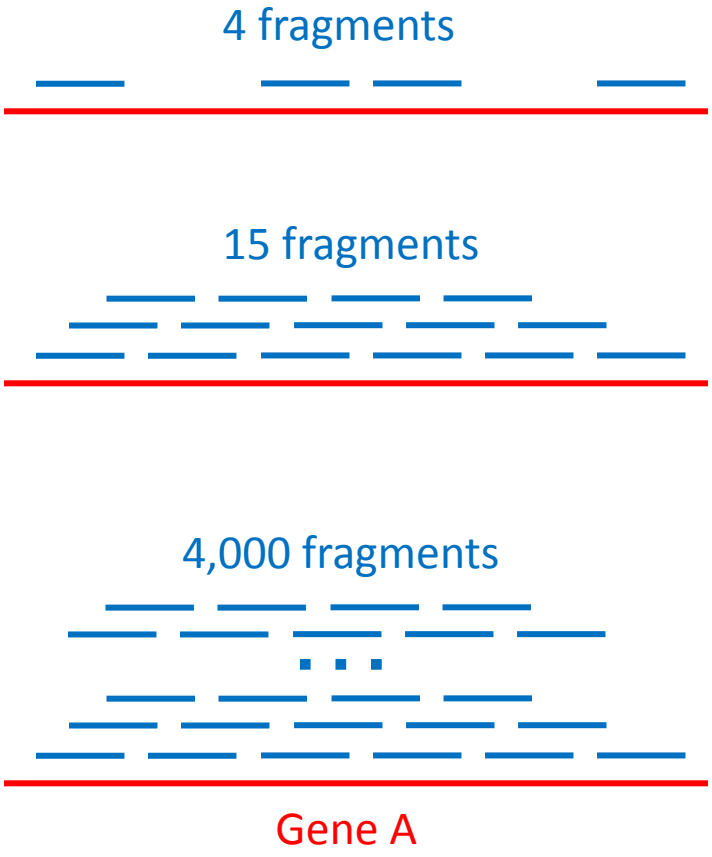


Salmon

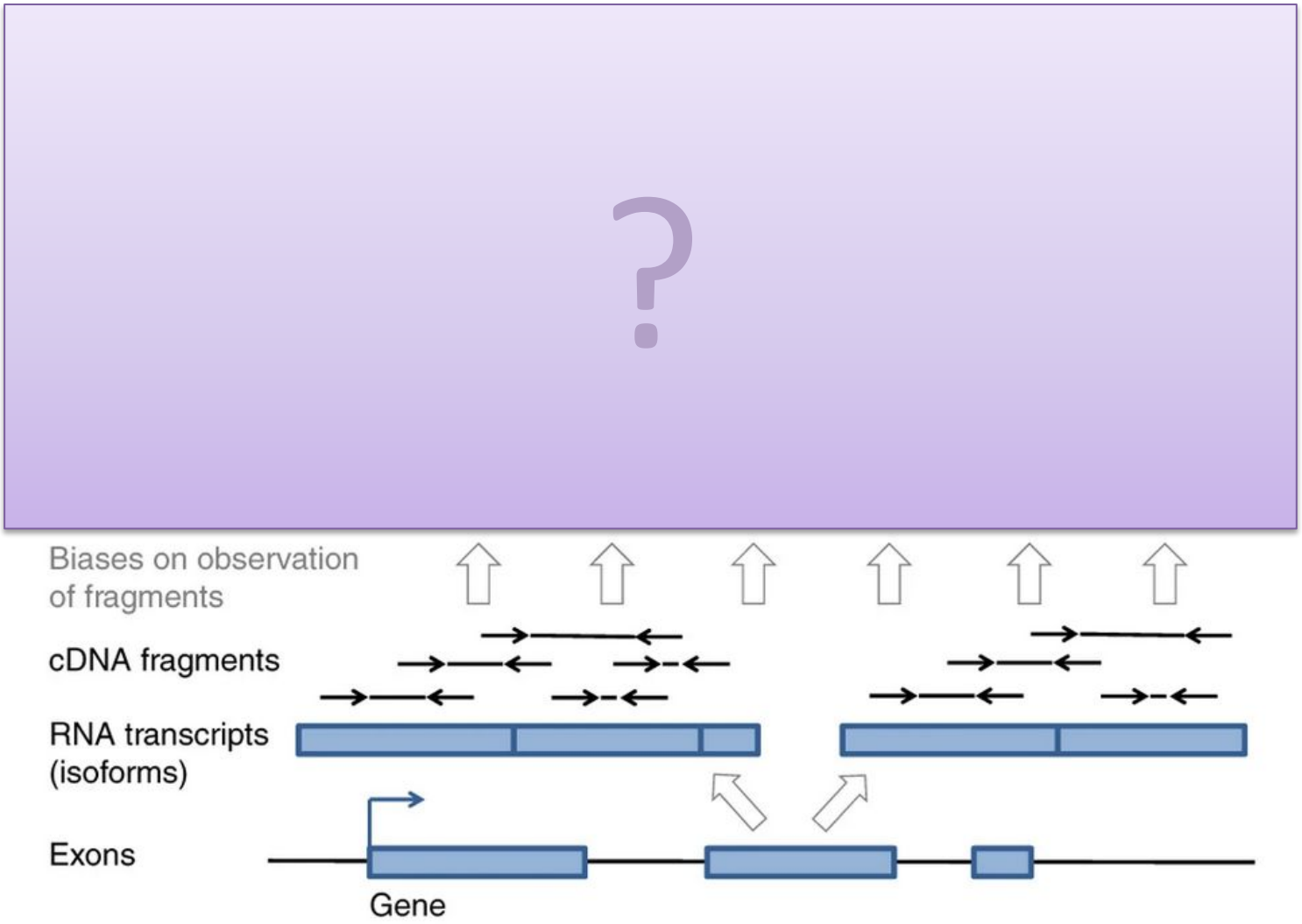
RNA (cDNA) short read sequencing



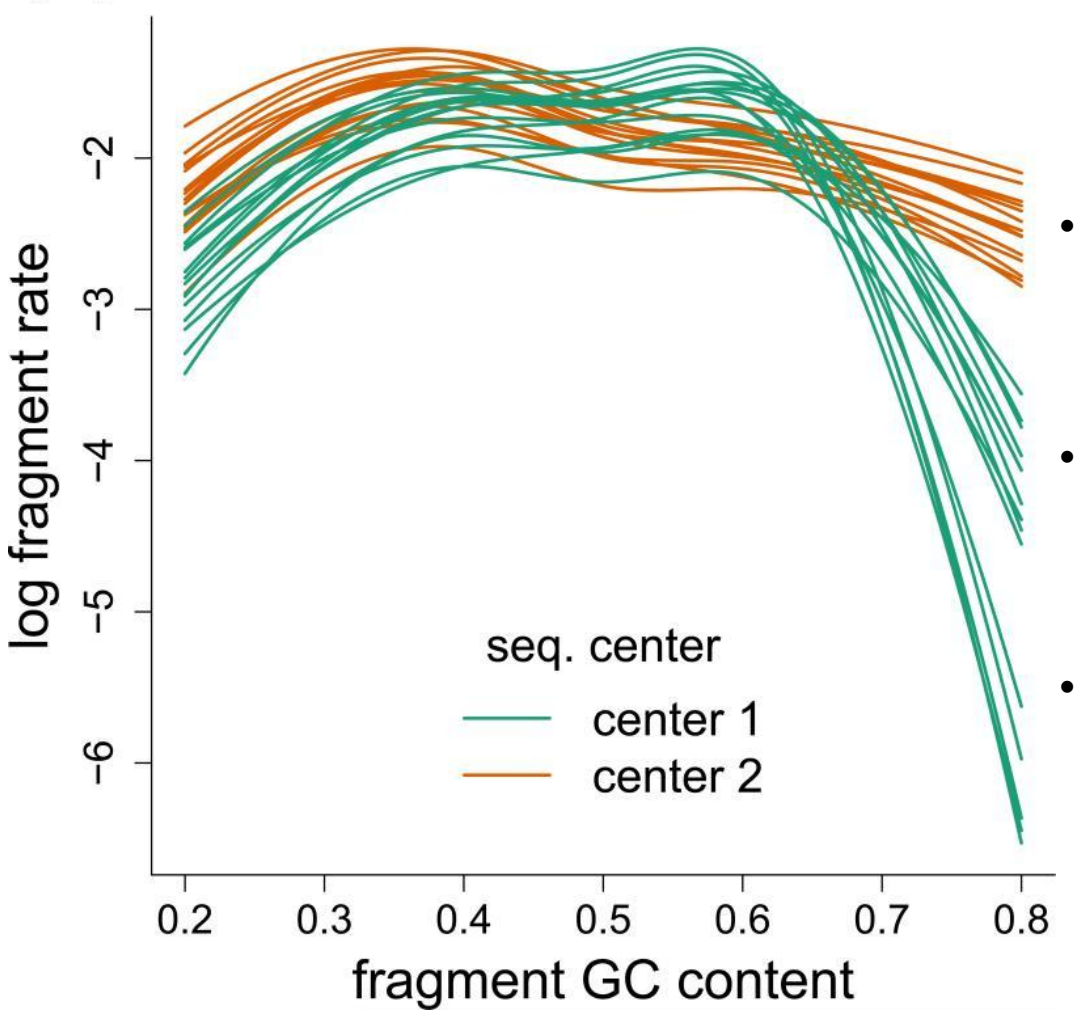
High dynamic range:



Technical biases corrected by Salmon



Fragment sequence bias



- Shown is *after* removing "hexamer bias" / "sequence bias"
- Can be attributed to differences in PCR amplification
- Sample- and batch-specific in comparison to "sequence bias" which doesn't vary much across samples

Love, Hogenesch, & Irizarry, "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation" (2016) PMC5143225

MultiQC

Home | Examples | Supported Tools | Docs | About

Example Report: RNA-Seq

MultiQC v1.14

- General Stats
- featureCounts
- STAR
- Cutadapt
  - Filtered Reads
  - Trimmed Sequence Lengths (3')
- FastQC
  - Sequence Counts
  - Sequence Quality Histograms
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Sequence Duplication Levels
  - Overrepresented sequences

Sequence Quality Histograms 8

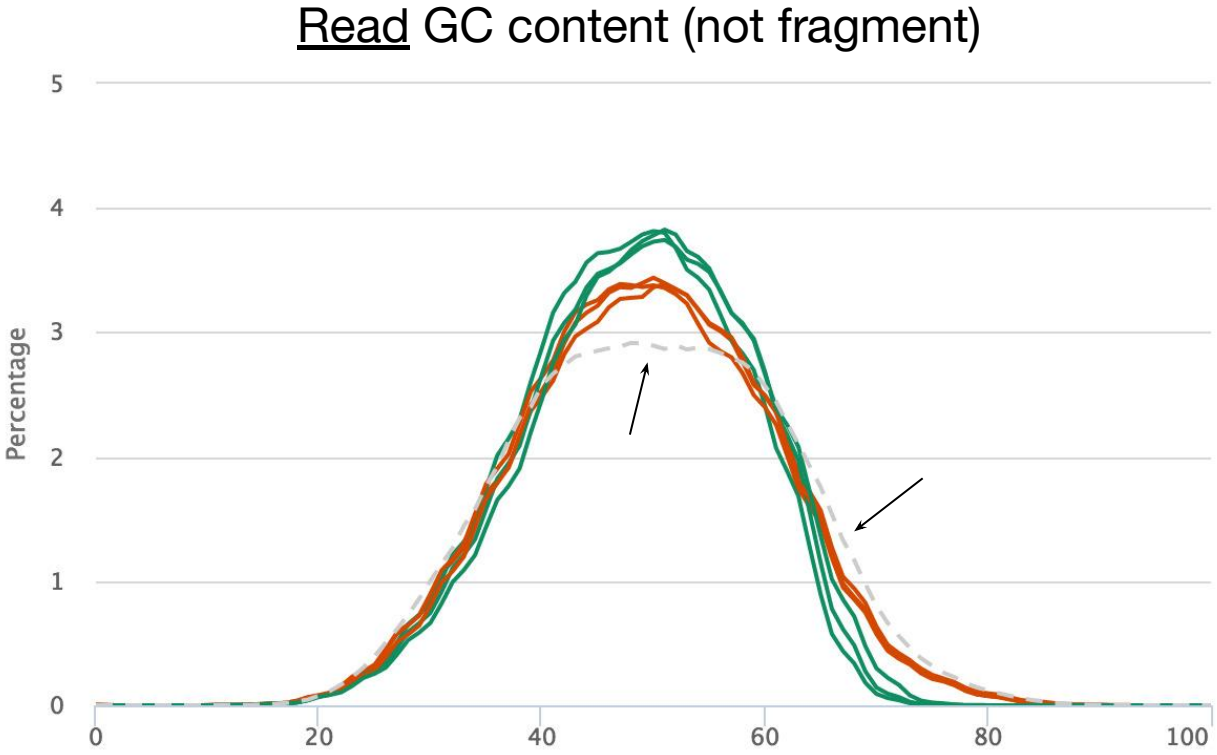
Help

The mean quality value across each base position in the read.

Y-Limits: on

Toolbox

MultiQC plugin



Theoretical GC Content

It is possible to plot a dashed line showing the theoretical GC content for a reference genome. MultiQC comes with genome and transcriptome guides for Human and Mouse. You can use these in your reports by adding the following MultiQC config keys (see Configuring MultiQC):

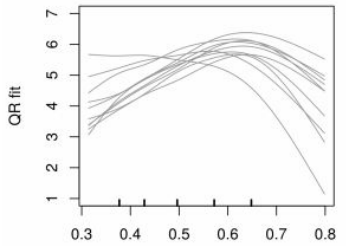
```
fastqc_config:
    fastqc_theoretical_gc: 'hg38_genome'
```

Only one theoretical distribution can be plotted. The following guides are available: `hg38_genome`, `hg38_txome`, `mm10_genome`, `mm10_txome` (txome = transcriptome).

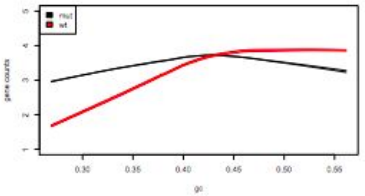
Different types of normalization

model counts on covariates (length, GC)

cqn



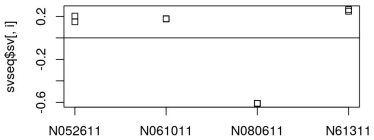
EDASeq



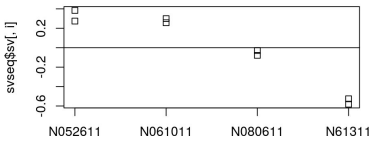
→ offsets

factor analysis on residuals of:
log counts ~ biological

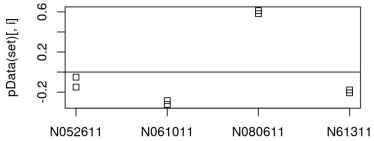
svaseq

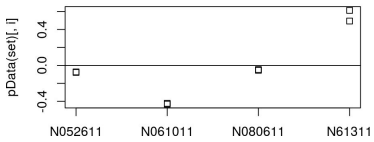


SV2



RUVSeq

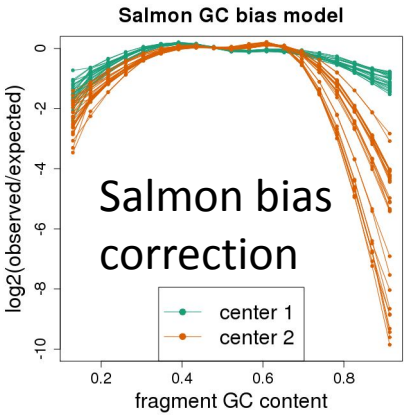




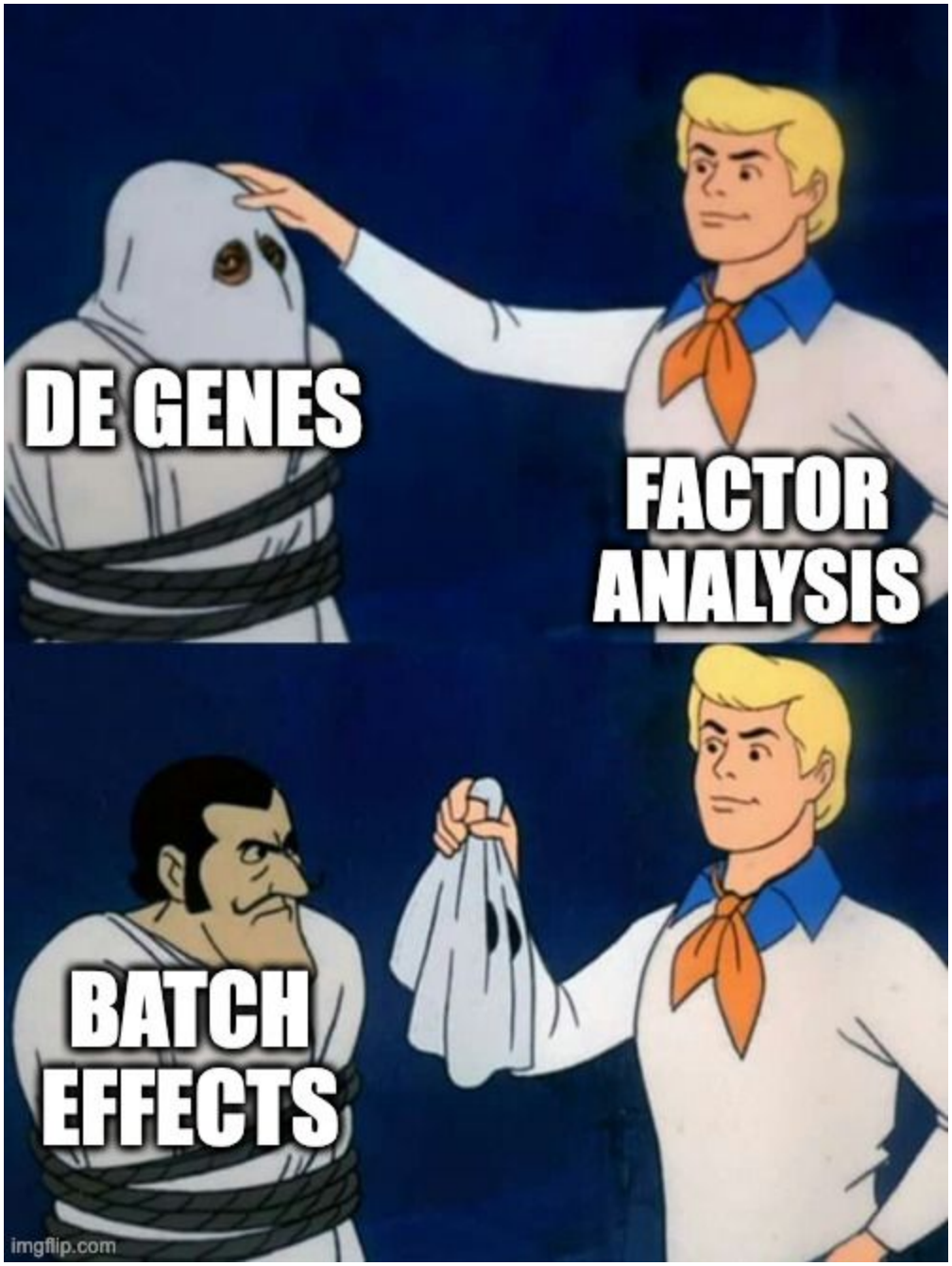
→ add to design

txiimport normalization

Length biases and…



→ offsets


DE GENES
FACTOR ANALYSIS
BATCH EFFECTS
imgflip.com

- Biology intro
- RNA-seq count table
- Batch effects and QC
- Quantification (reads → count table)
- Import into Bioconductor

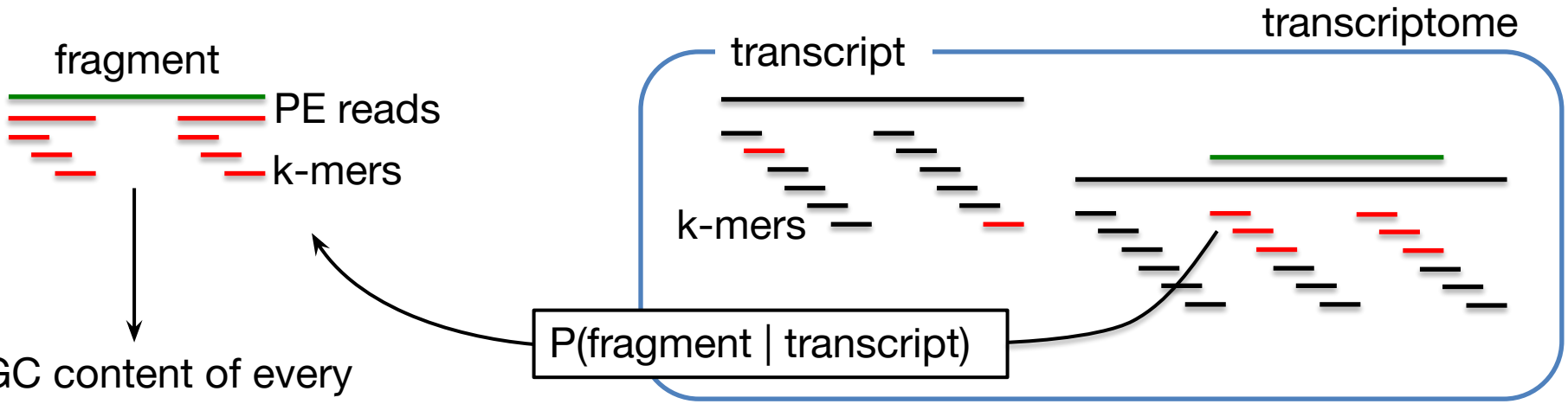
Lightweight quantifiers

Sailfish: Patro et al (2014), kallisto: Bray et al (2016), Salmon: Patro et al (2017)

Salmon maps reads to transcriptome with **RapMap:** Srivastava *et al* (2016)

Elements of mapping algorithm use ideas from **kallisto:** Bray *et al* (2016):
(1) skipping ahead to find Next Informative Position (NIP) and
(2) defining the consensus as ∩ of transcript sets from all hits

Salmon maps all PE reads, useful for estimating bias



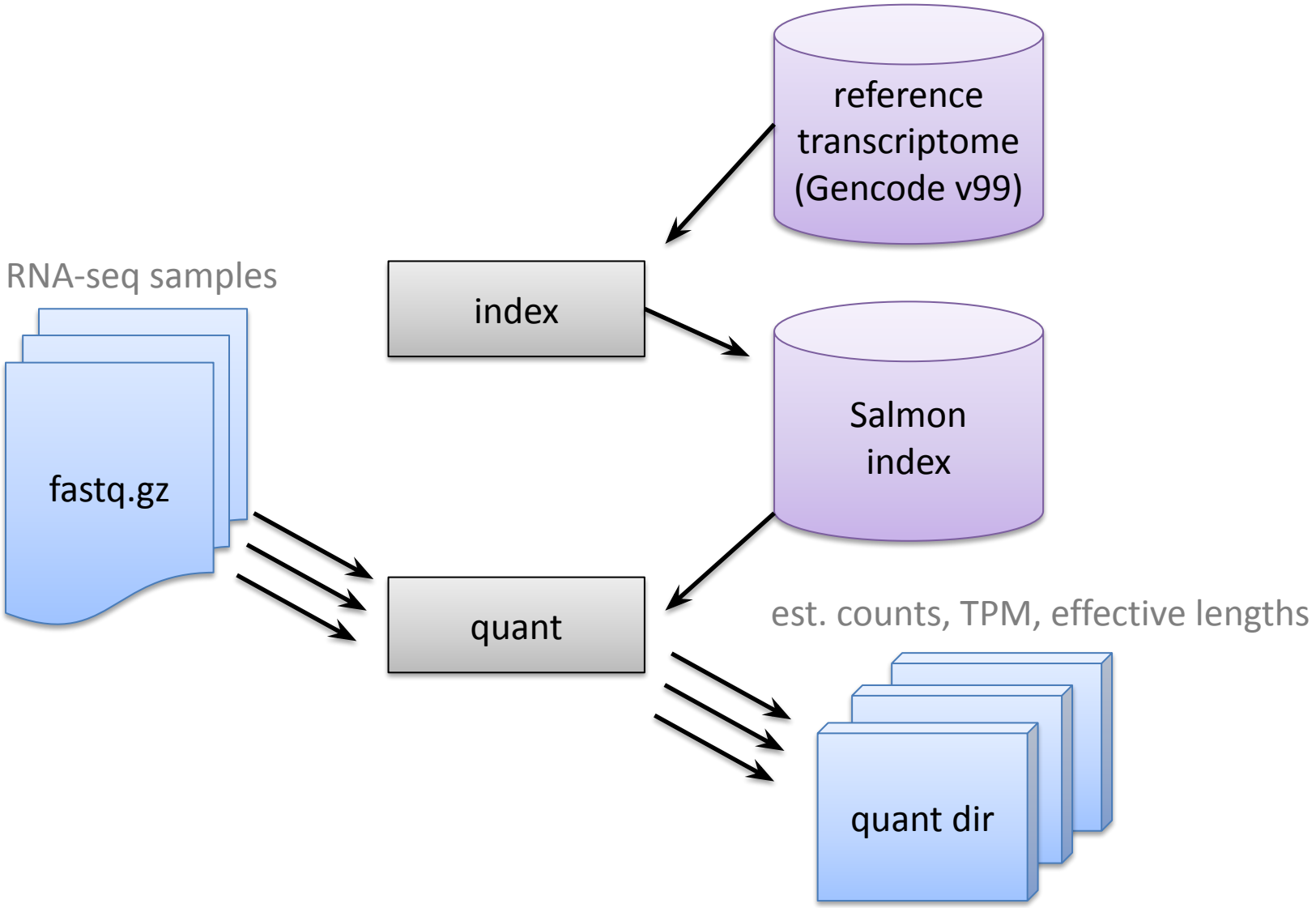
GC content of every inferred fragment, adjust *effective* transcript lengths, pass this information to DE tools as an *offset*

RESEARCH Open Access

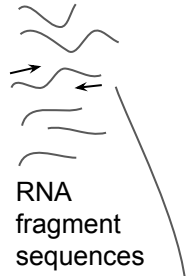
Alignment and mapping methodology influence transcript abundance estimation

Avi Srivastava[1†], Laraib Malik[1†], Hirak Sarkar[2], Mohsen Zakeri[2], Fatemeh Almodaresi[2], Charlotte Soneson[3,4], Michael I. Love[5,6], Carl Kingsford[7] and Rob Patro[2*]

Steps for running Salmon



Recommend *Salmon* with Snakemake




```
rule all:
 input: expand("quants/{run}/quant.sf", run=RUNS)


rule salmon_index:
    input: "{ANNO}/gencode.vXYZ.transcripts.fa.gz"
    output: directory("{ANNO}/gencode.vXYZ-salmon_1.3.0")
    shell: "{SALMON} index --gencode -p 8 -t {input} -i {output}"
```

Reference transcripts (GENCODE, Ensembl, etc.)

```
rule salmon_quant:
    input:
        r1 = "/pine/scr/m/i/milove/{sample}_1.fastq.gz",
        r2 = "/pine/scr/m/i/milove/{sample}_2.fastq.gz",
        index = "/proj/milovelab/anno/gencode.vXYZ-salmon_1.3.0"
    output:
        "quants/{sample}/quant.sf"
    params:
        dir = "quants/{sample}"
    shell:
        "{SALMON} quant -i {input.index} -l A -p 8 --gcBias "
        "--numGibbsSamples 30 --thinningFactor 100 "
        "-o {params.dir} -1 {input.r1} -2 {input.r2}"
```

Key:
Input Arguments
Output **Wildcard**

My Salmon Snakemake file (you can find this later on my website, under Software)

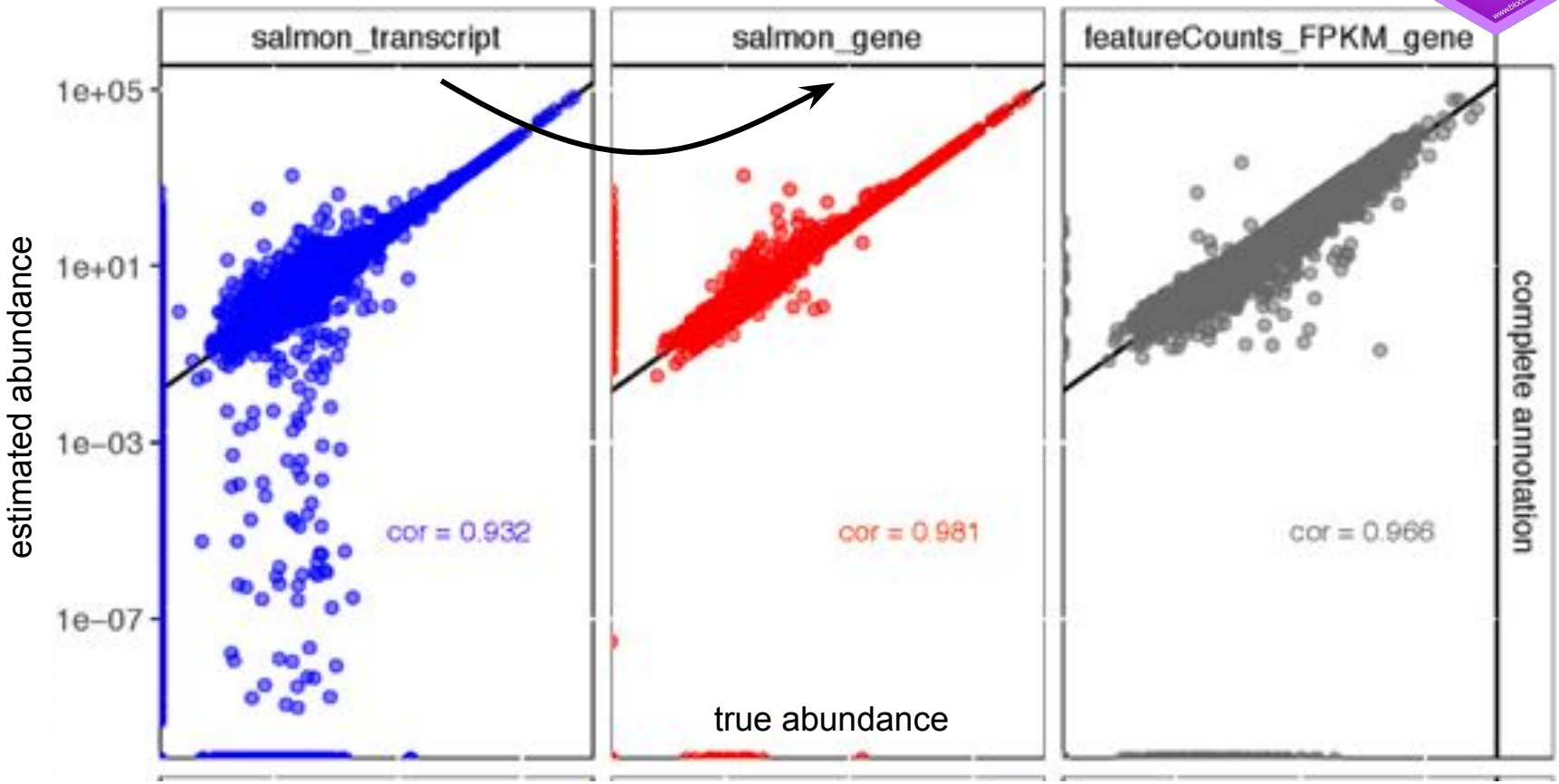
- Biology intro
- RNA-seq count table
- Batch effects and QC
- Quantification (reads → count table)
- Import into Bioconductor

Packages for data import

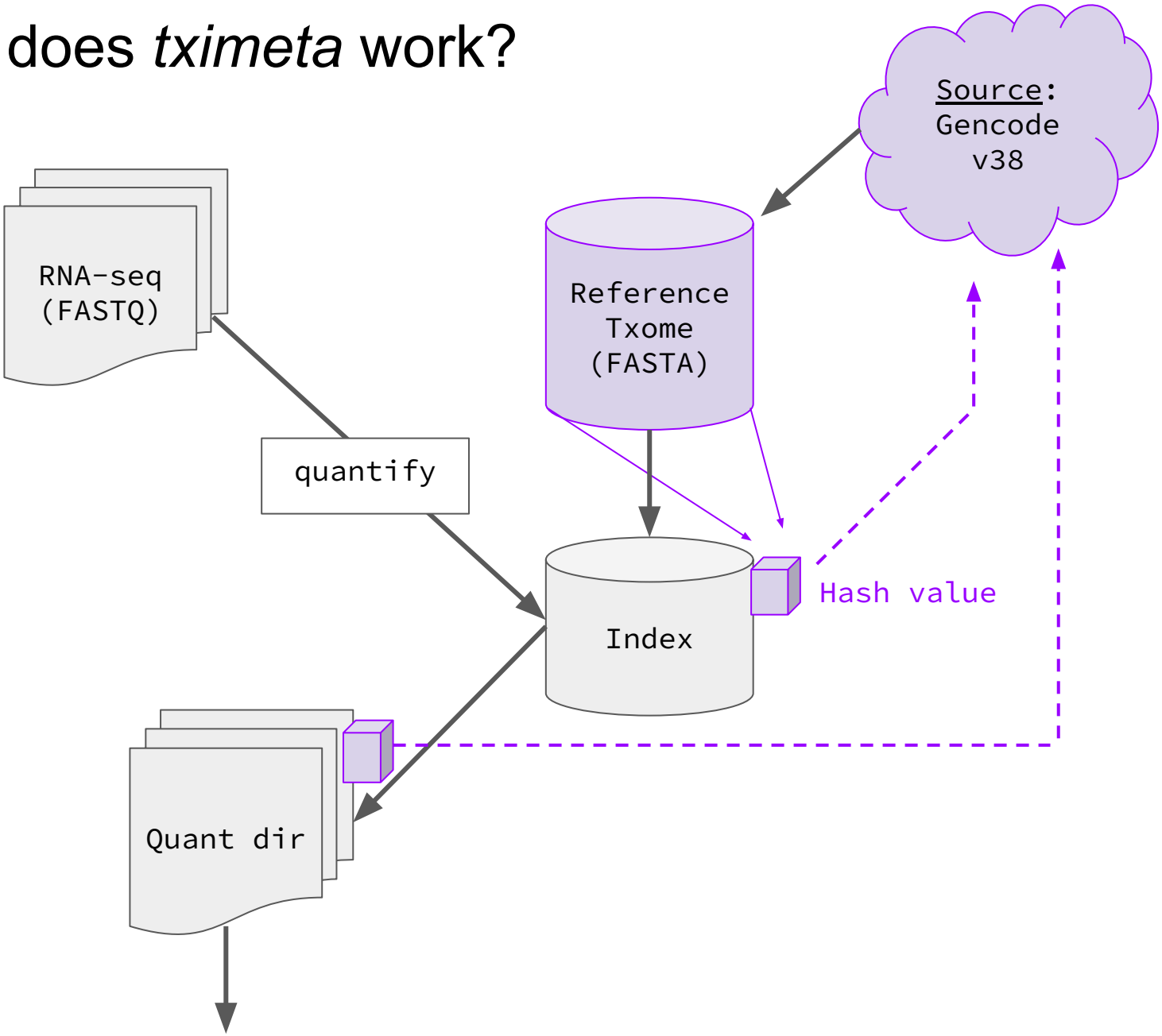
METHOD ARTICLE

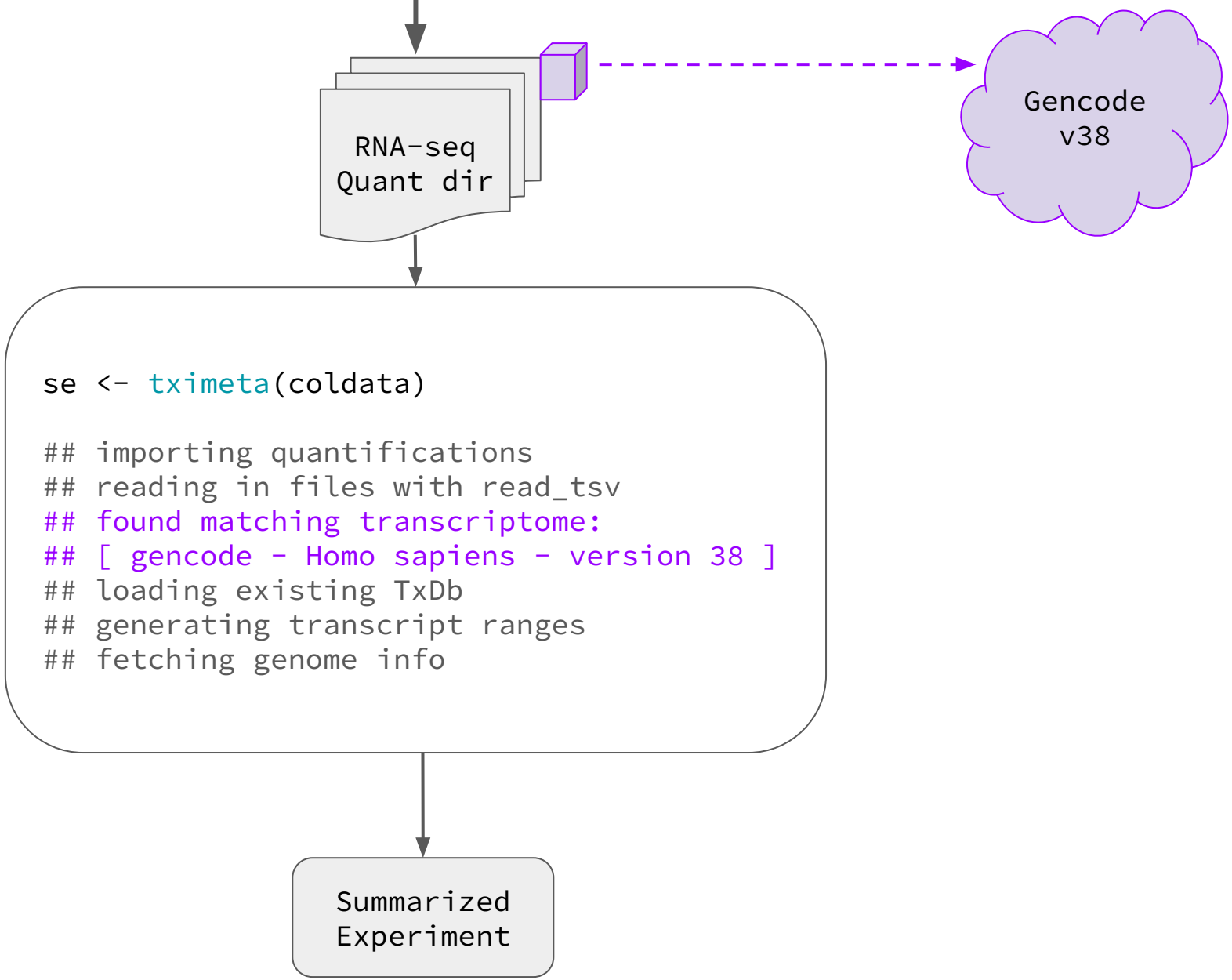
REVISED Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences

Charlotte Soneson [1,2], Michael I. Love [3,4], Mark D. Robinson [1,2]



How does *tximeta* work?




RNA-seq
Quant dir
Gencode
v38
se <- tximeta(coldata)
importing quantifications
reading in files with read_tsv
found matching transcriptome:
[gencode - Homo sapiens - version 38]
loading existing TxDb
generating transcript ranges
fetching genome info
Summarized
Experiment

Tximeta

Pre-computed digests

We plan to support a wide variety of sources and organisms for transcriptomes with pre-computed digests, though for now the software focuses on predominantly human and mouse transcriptomes

The following digests are supported in this version of `tximeta`:

source	organism	releases
GENCODE	Homo sapiens	23-46
GENCODE	Mus musculus	M6-M35
Ensembl	Homo sapiens	76-112
Ensembl	Mus musculus	76-112
Ensembl	Drosophila melanogaster	79-112
RefSeq	Homo sapiens	p1-p13
RefSeq	Mus musculus	p2-p6

← as of June 2024

OPEN ACCESS PEER-REVIEWED

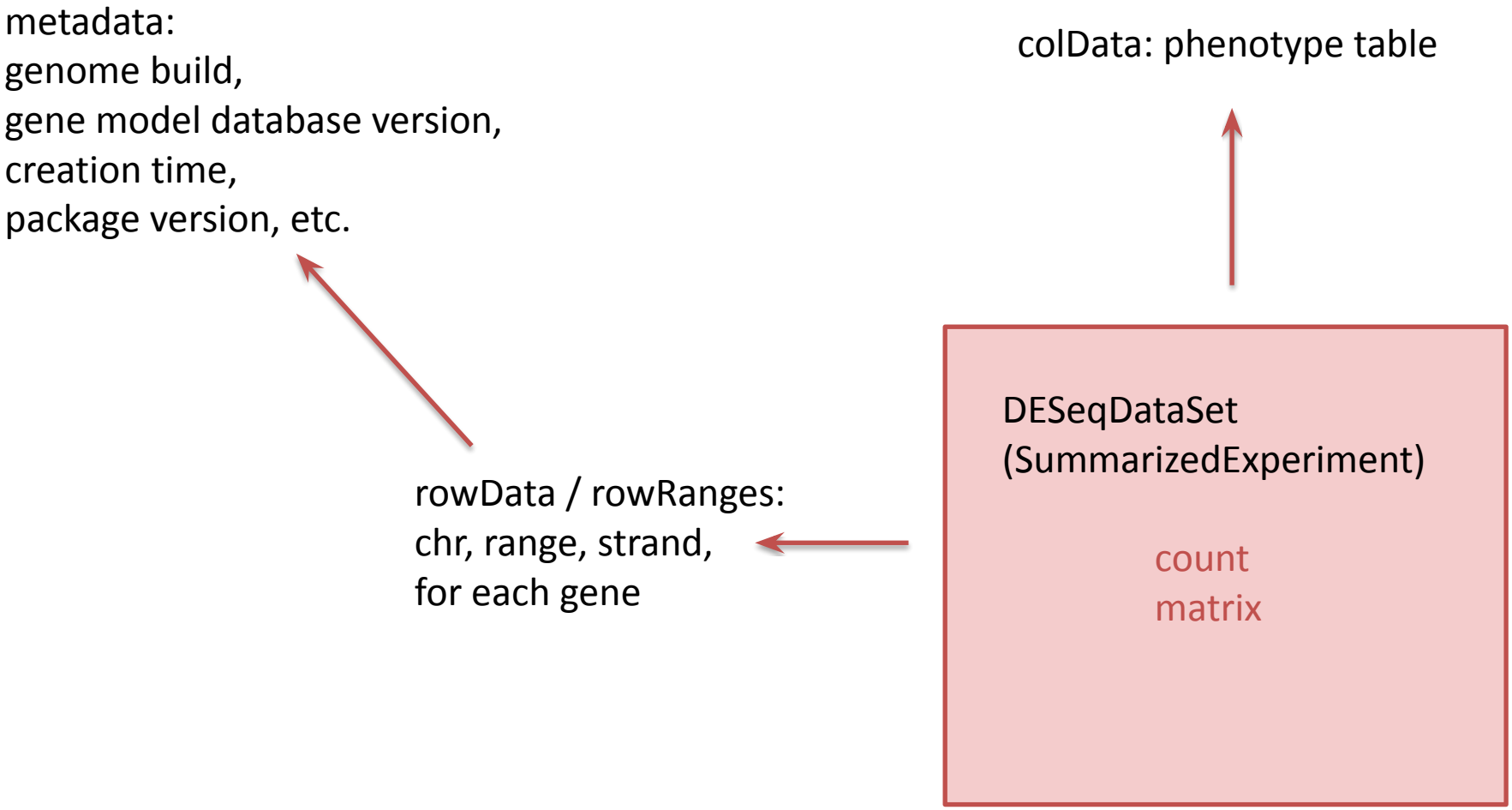
RESEARCH ARTICLE

Tximeta: Reference sequence checksums for provenance identification in RNA-seq

Michael I. Love, Charlotte Soneson, Peter F. Hickey, Lisa K. Johnson, N. Tessa Pierce, Lori Shepherd, Martin Morgan, Rob Patro

Version 2 Published: February 25, 2020 • https://doi.org/10.1371/journal.pcbi.1007664

Bioconductor has rich objects



Bioconductor help

- Vignettes:

```
> browseVignettes("DESeq2")
> vignette("DESeq2")
```

- Type ? then the function name:

```
> ?results
```

Bioconductor help

```
results                  package:DESeq2                  R Documentation

Extract results from a DESeq analysis

Description:

     'results' extracts a result table from a DESeq analysis giving
     base means across samples, log2 fold changes, standard errors,
     test statistics, p-values and adjusted p-values; 'resultsNames'
     returns the names of the estimated effects (coefficents) of the
     model; 'removeResults' returns a 'DESeqDataSet' object with
     results columns removed.

Usage:

     results(object, contrast, name, lfcThreshold = 0,
       altHypothesis = c("greaterAbs", "lessAbs", "greater", "less"),
       listValues = c(1, -1), cooksCutoff, independentFiltering = TRUE,
       alpha = 0.1, filter, theta, pAdjustMethod = "BH",
       format = c("DataFrame", "GRanges", "GRangesList"), test, addMLE = FALSE,
       tidy = FALSE, parallel = FALSE, BPPARAM = bpparam())

...

Arguments:

  object: a DESeqDataSet, on which one of the following functions has
          already been called: 'DESeq', 'nbinomWaldTest', or
          'nbinomLRT'

contrast: this argument specifies what comparison to extract from the
          'object' to build a results table. one of either:

            - a character vector with exactly three elements: the name
              of a factor in the design formula, the name of the
              numerator level for the fold change, and the name of the
              denominator level for the fold change (simplest case)
```

Bioconductor help

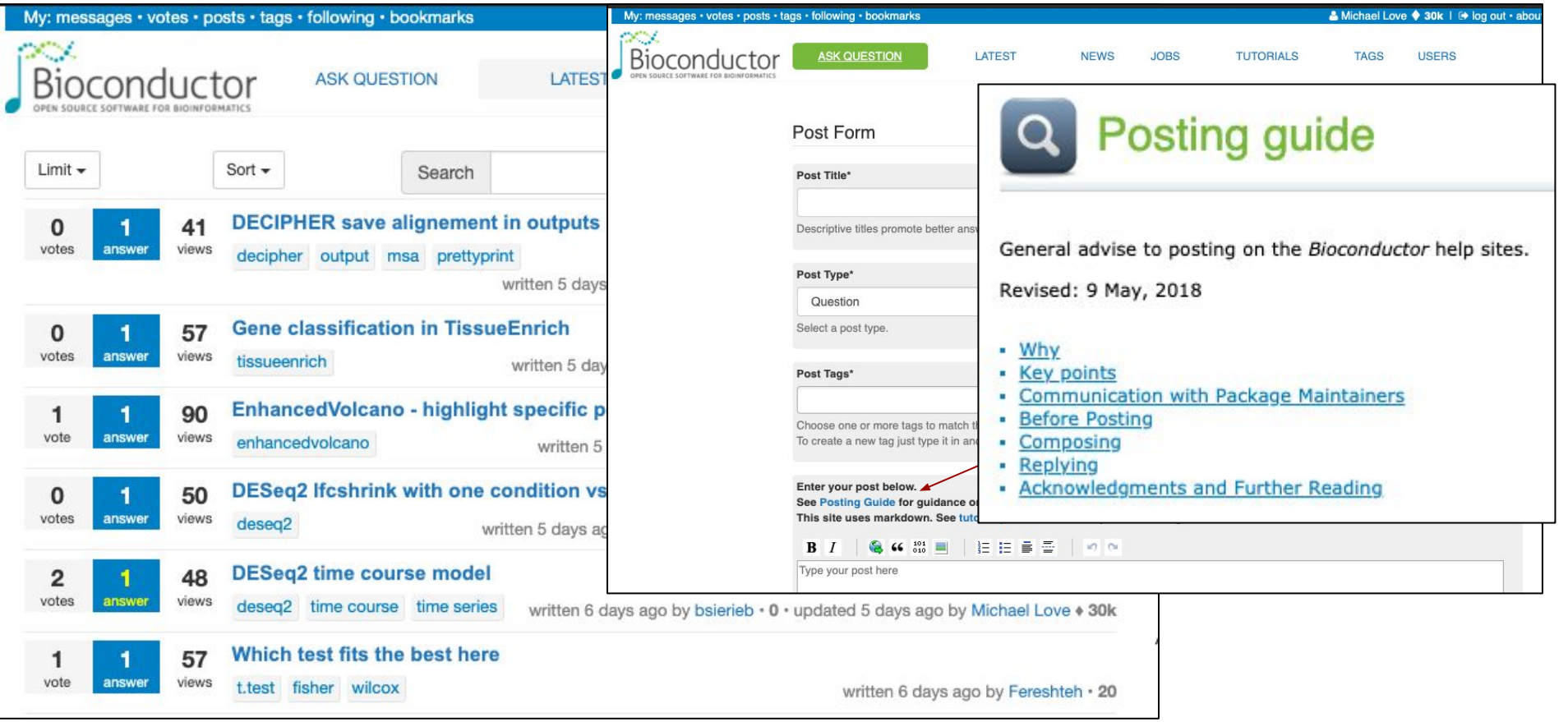
```
Value:

     For 'results': a 'DESeqResults' object, which is a simple subclass
     of DataFrame. This object contains the results columns:
     'baseMean', 'log2FoldChange', 'lfcSE', 'stat', 'pvalue' and
     'padj', and also includes metadata columns of variable
     information....

...

References:

     Richard Bourgon, Robert Gentleman, Wolfgang Huber: Independent
     filtering increases detection power for high-throughput
     experiments. PNAS (2010), <URL:
     http://dx.doi.org/10.1073/pnas.0914005107>

See Also:

     'DESeq'

Examples:

     ## Example 1: simple two-group comparison

     dds <- makeExampleDESeqDataSet(m=4)

...
```

Looking up help for objects

```
> class(dds)
[1] "DESeqDataSet"
attr(,"package")
[1] "DESeq2"

> ?DESeqDataSet

> help(package="DESeq2", help_type="html")

> methods(class = "DESeqDataSet")
```

How to get help

https://support.bioconductor.org



How to get help

https://support.bioconductor.org

Key points

- Use the latest *Bioconductor* release version. Ensure that your packages are up-to-date.
- Post all of your *R* code.
- Include a copy of any error or warning messages that appeared in *R*.
- If your question involves experimental data, include an example of the sample-level covariate data (one row per sample, one column per covariate). If it would help answer your technical question, and can be shared, explain the motivation behind your experiment.
- Always paste the output of `sessionInfo()` at the end of your post.
- If possible, provide a minimal, self-contained example that someone else can cut-and-paste into a new *R* session to reproduce your problem.
- If the example produces an error, provide the output of `traceback()` after the error occurs.

Extra slides

Statistical power review

- False positive rate (α under H_0):
  of the ? how many ?
- False discovery rate:
  of the ? how many ?
- Power (sensitivity):
  of the ? how many ?

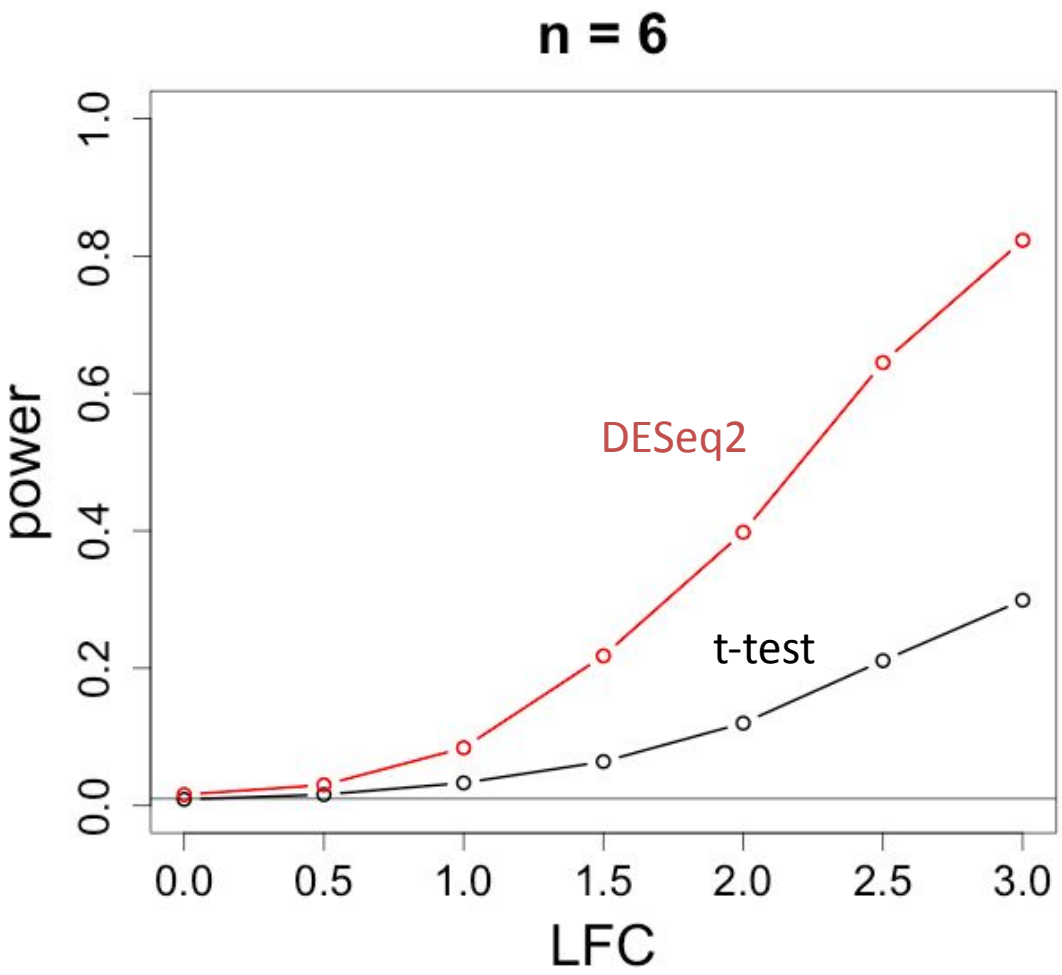
	true:	
test:	Nulls	Alternatives
Negative	true negative	false negative
Positive	false positive	true positive

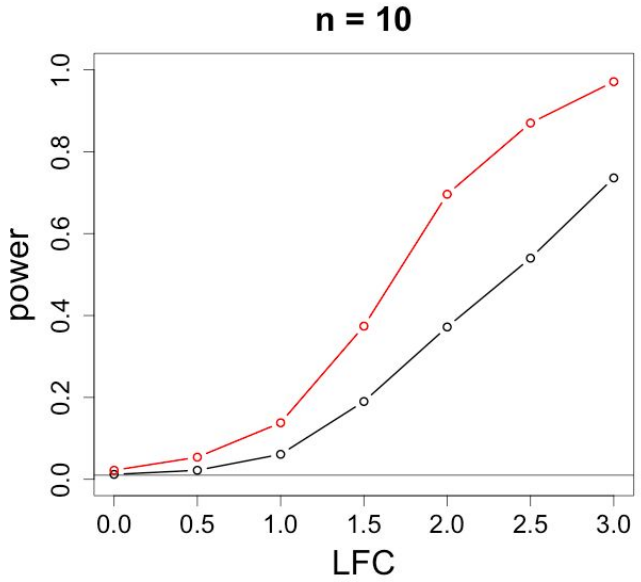
seek to avoid

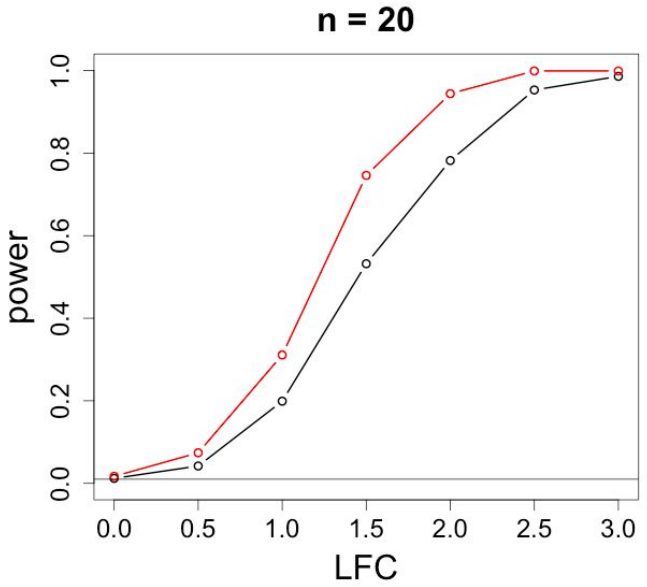
seek to avoid

Statistical power

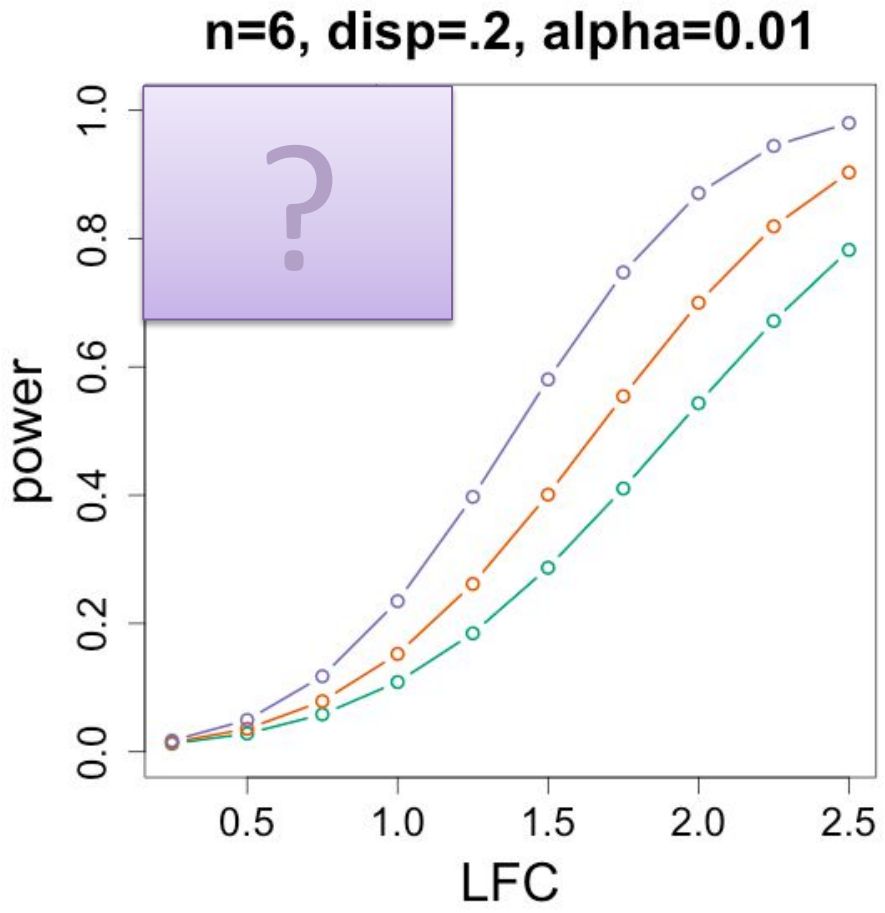
Why not use a (simple) t-test on log scaled counts?



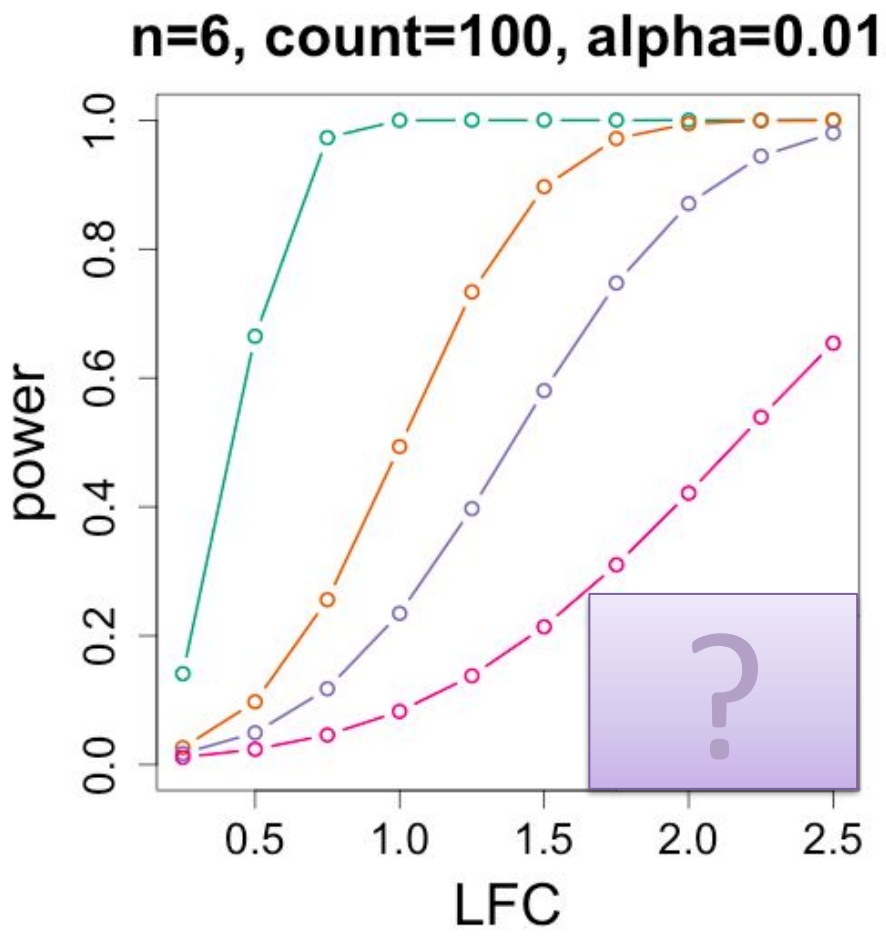




Bioc pkg: RNASeqPower



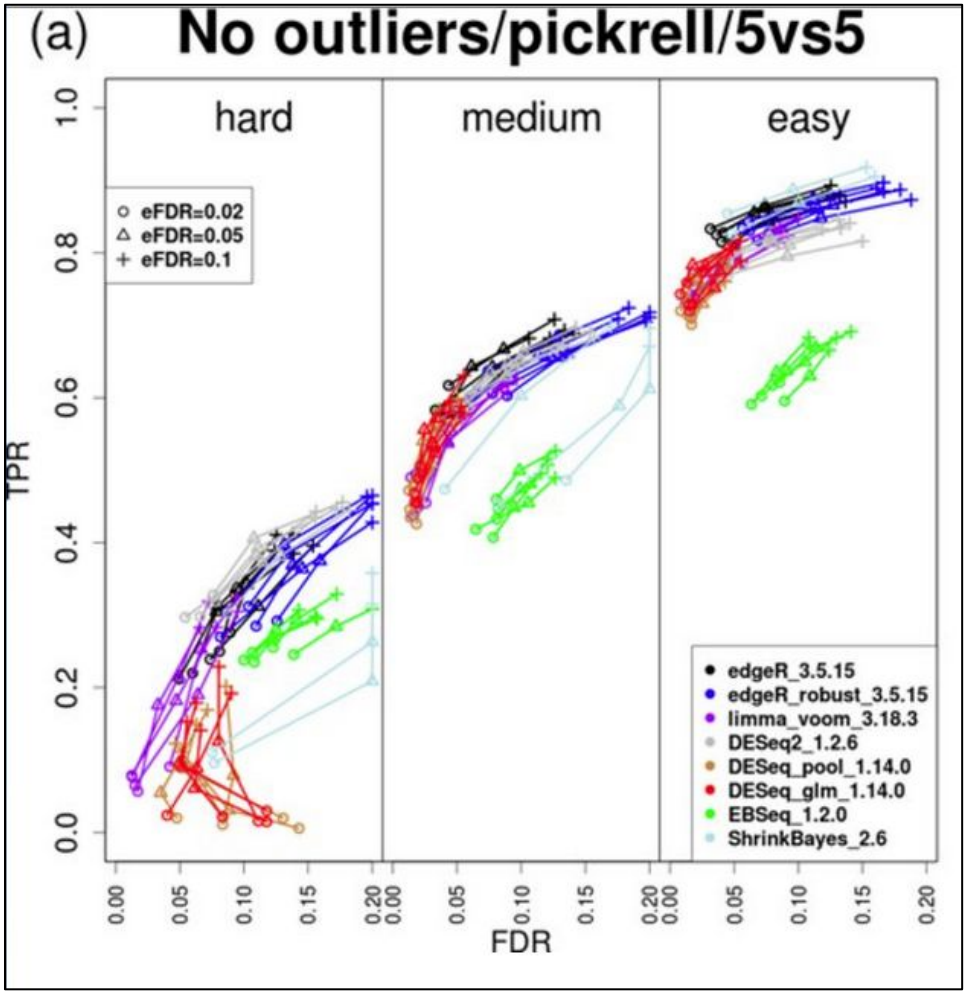
varying the count



varying the dispersion

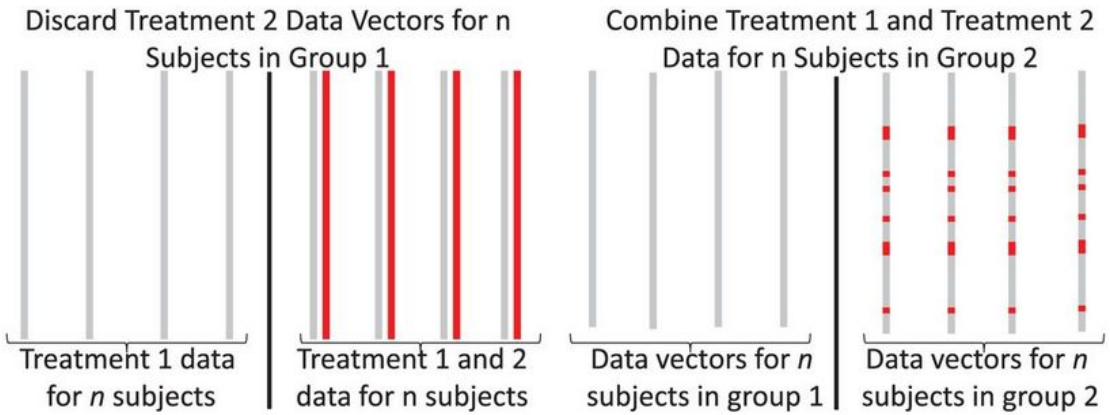
Brief comparison of DE methods

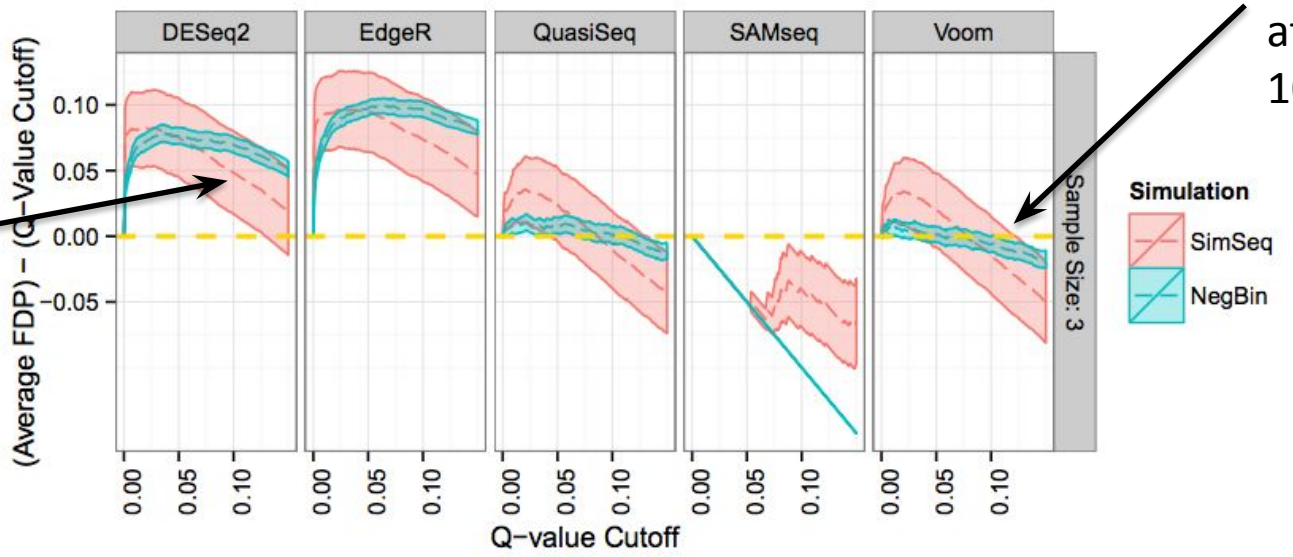
NB simulation



"Robustly detecting differential expression in RNA sequencing data using observation weights"
Xiaobei Zhou, Helen Lindsay and Mark D. Robinson (2014)

Nonparametric simulation





15% FDR at reported 10% FDR

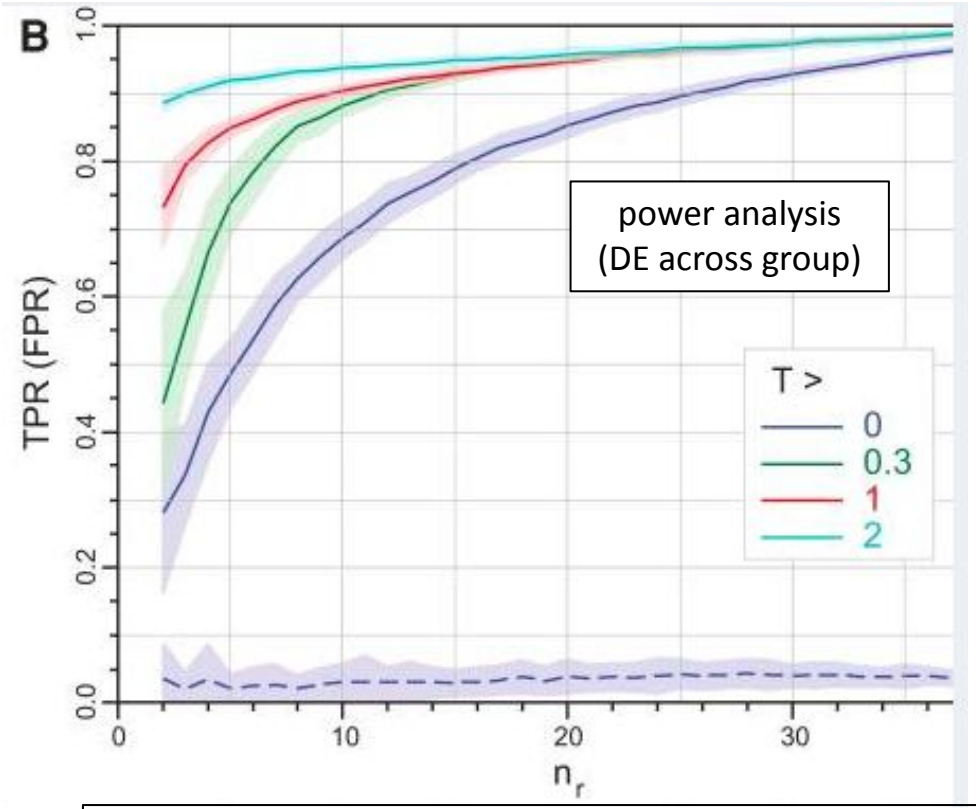
10% FDR at reported 10% FDRD

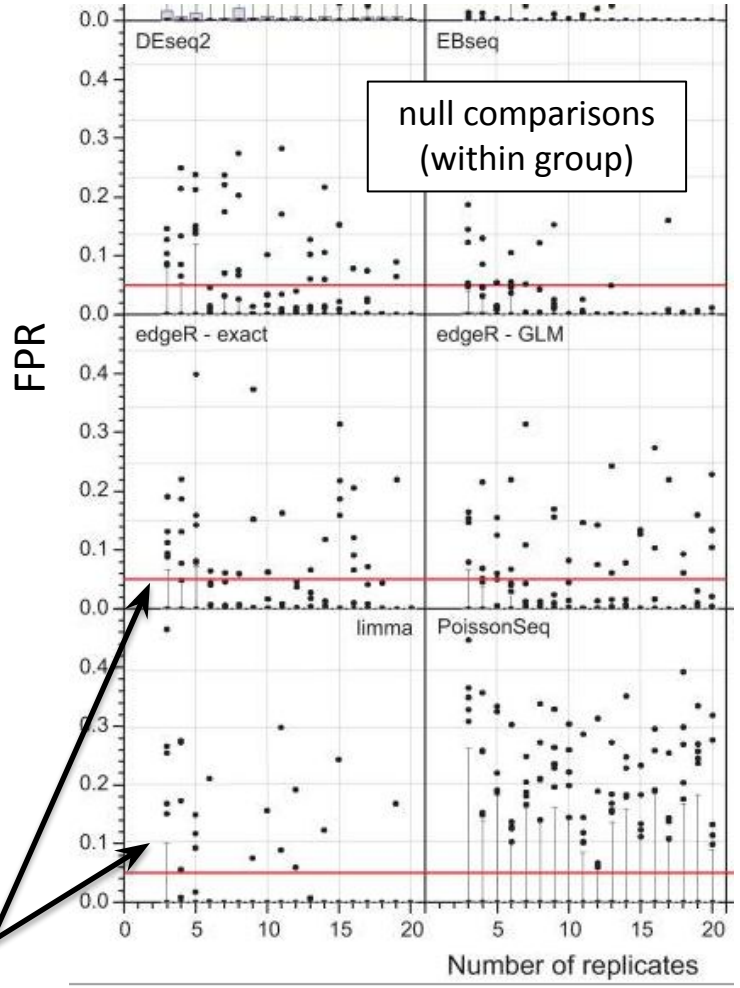
Benidt & Nettleton (2015)

PMC4878611

Schurch *et al* 2016: "How many biological replicates are needed in an RNA-seq experiment?" Generated a "highly replicated" yeast RNA-seq (n=42)

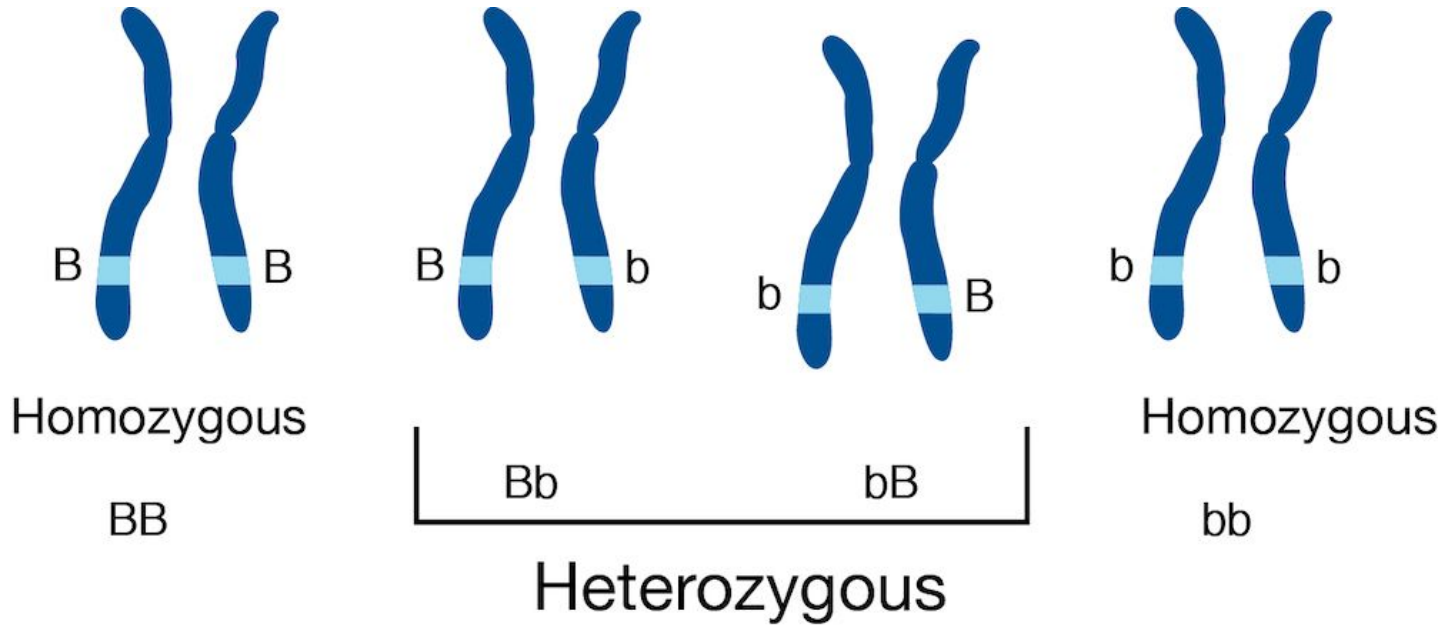
edgeR TPR over per-group sample size



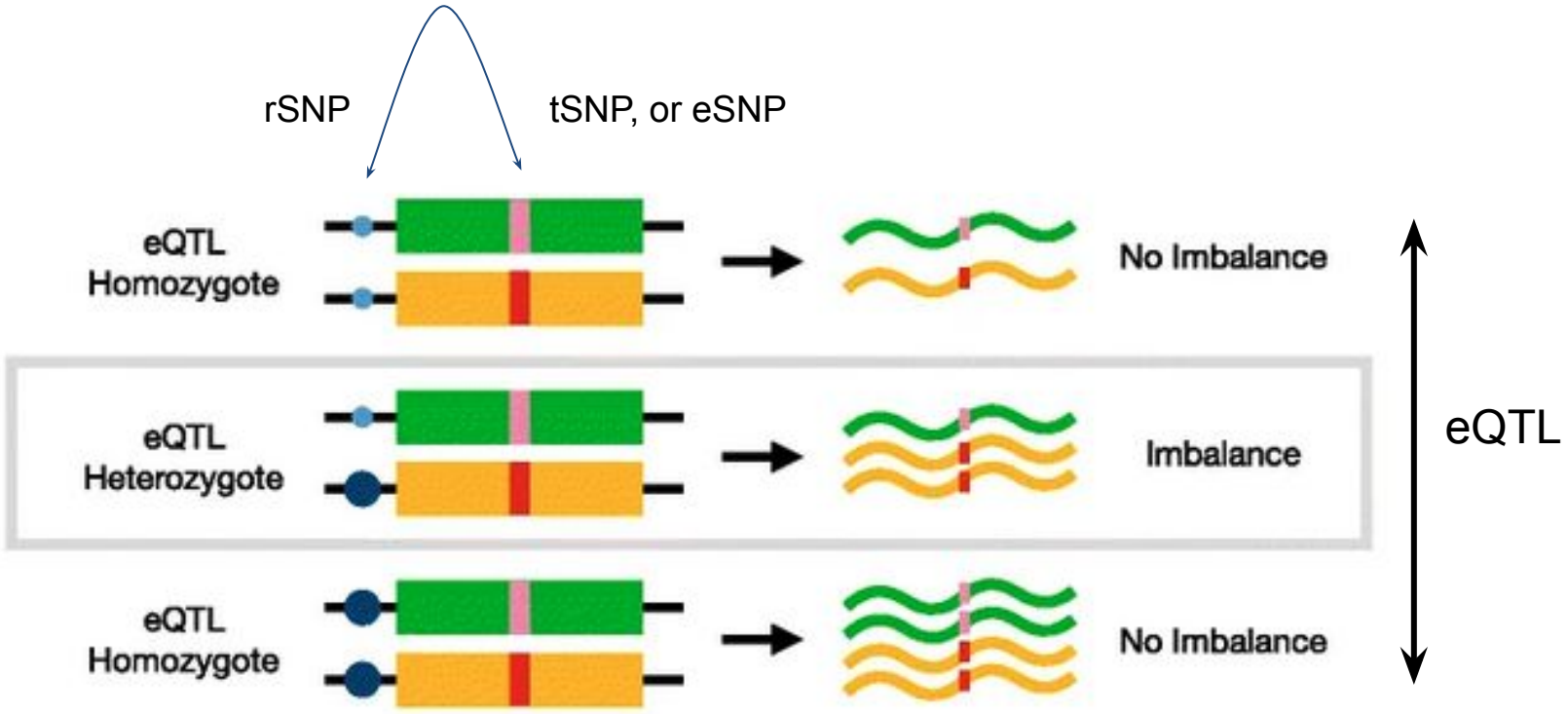


100 simulations, boxplot: 75%, whisker: 95%

Conclusion: Most methods have low FPR in at least 95/100 null comparisons



https://commons.wikimedia.org/wiki/File:Heterozygous.jpg



Castel, S.E., Aguet, F., Mohammadi, P. *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol* **21,** 234 (2020). https://doi.org/10.1186/s13059-020-02122-z