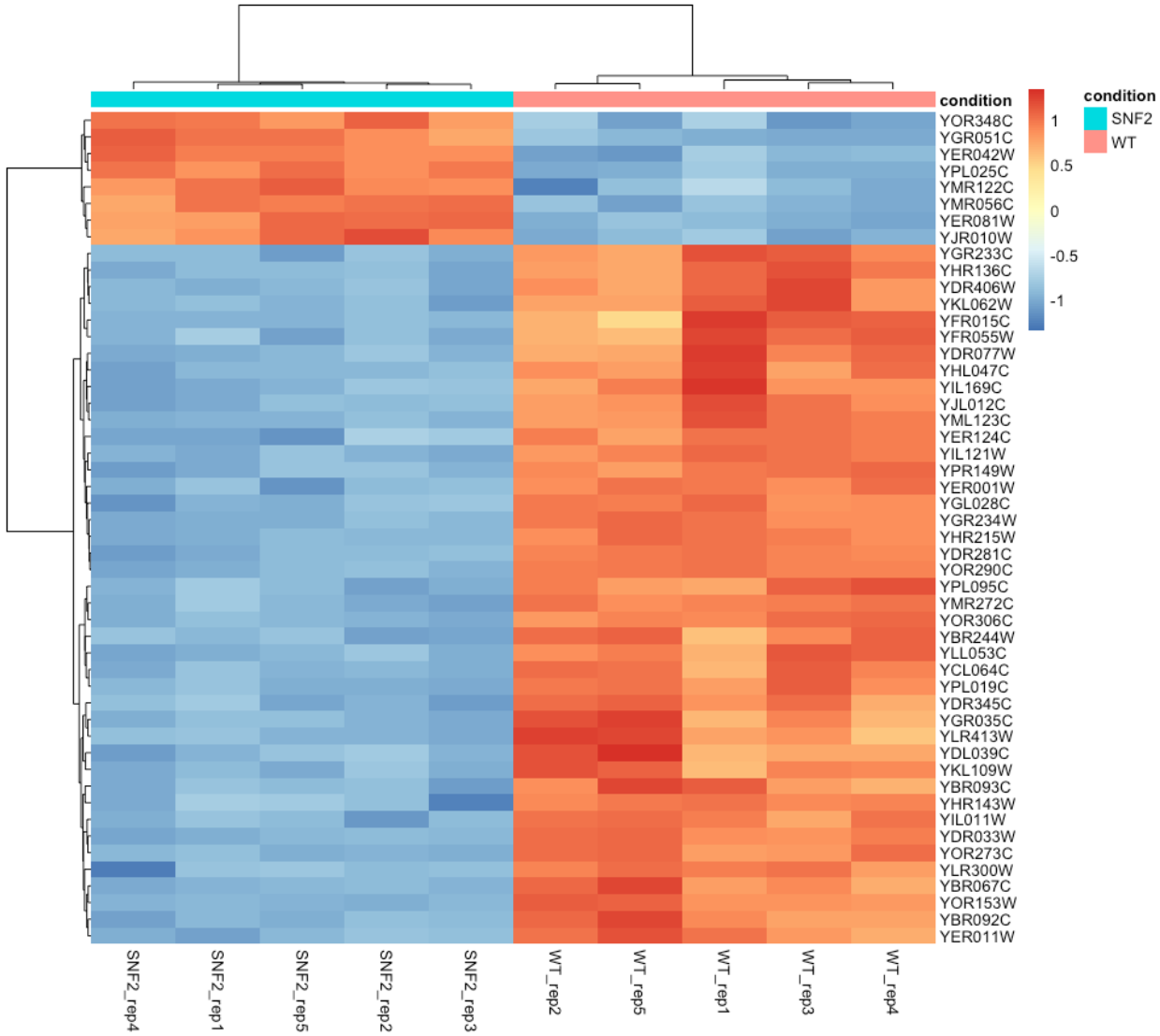
Bioinformatics for RNA-seq

Wenwen Hou
Rebecca Batorsky
Albert Tai

May 2020

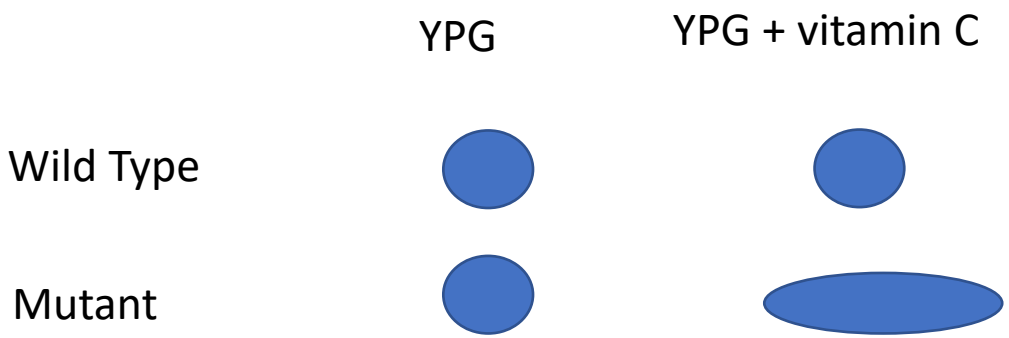
Common RNAseq analysis goals

- Novel transcript discovery
- Transcriptome assembly
- Single cell analysis
- Quantify alternative splicing
- **Differential Expression**



Why is differential expression useful?

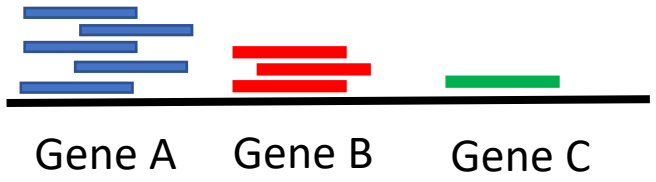
We’re looking for an explanation of observed phenotypes:



What causes difference in phenotype?

Difference in protein activity!

mRNA is easier to measure than protein, so we use it as a proxy



Read Counts

↓

Number of mRNA copies

↓

Concentration of proteins

↓

Difference in protein activity

↓

Phenotype

https://www.slideshare.net/jakonix/part-1-of-rnaseq-for-de-defining-the-goal

Though our assumptions about correlation are often violated

Read Counts

↓

Number of mRNA copies

↓

Concentration of proteins

↓

Difference in protein activity

↓

Phenotype

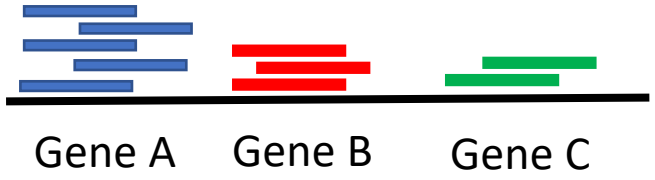
Failure to map to proper location, biases in library prep

mRNA templates have different speeds of protein production, alternative splicing

Protein activity is independently regulated (phosphorylation, ubiquitination)

https://www.slideshare.net/jakonix/part-1-of-rnaseq-for-de-defining-the-goal

As a consequence, we look at comparisons



Read Counts

↓

Number of mRNA copies

↓

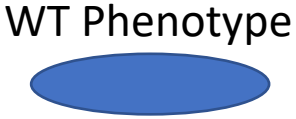
Concentration of proteins

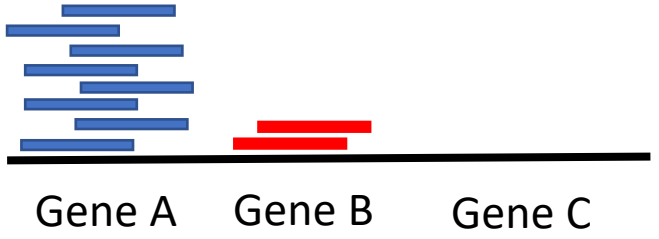
↓

Difference in protein activity

↓

WT Phenotype





Read Counts

↓

Number of mRNA copies

↓

Concentration of proteins

↓

Difference in protein activity

↓

Mutant Phenotype

Our goal

"How can we detect genes for which the counts of reads change between conditions **more systematically** than as expected by chance"

We must design an experiment where this hypothesis can be tested.

Oshlack et al. 2010. From RNA-seq reads to differential expression results. Genome Biology 2010, 11:220
http://genomebiology.com/2010/11/12/220

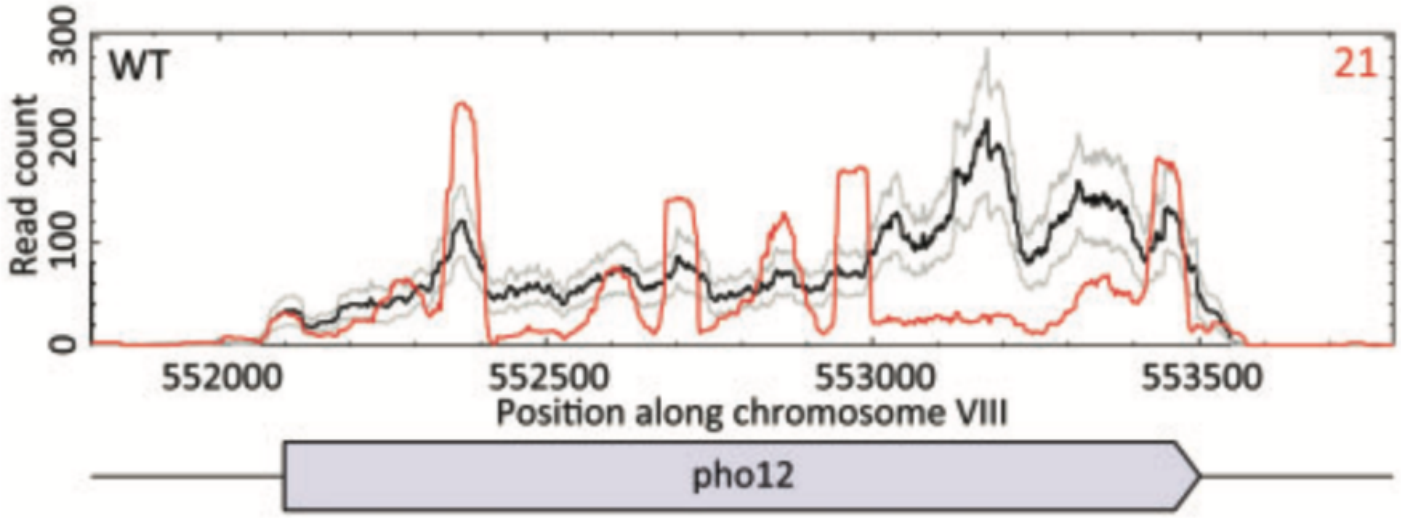
Experiment design

How deep to sequence?
How many biological replicates to choose?

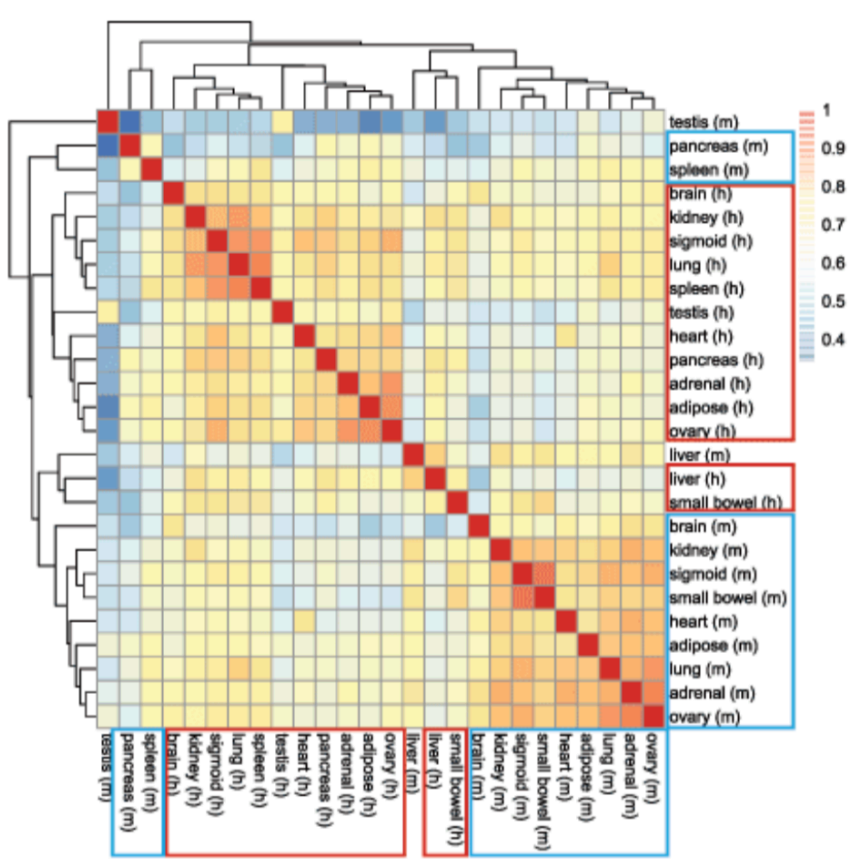
- Difficult to answer in general but certainly >=3 replicates and ~20 M reads/replicate for strongly expressed genes
- Pilot studies are recommended to determine the number of replicates needed to capture the variability (e.g. 2 bio replicates, 10-20 M reads)

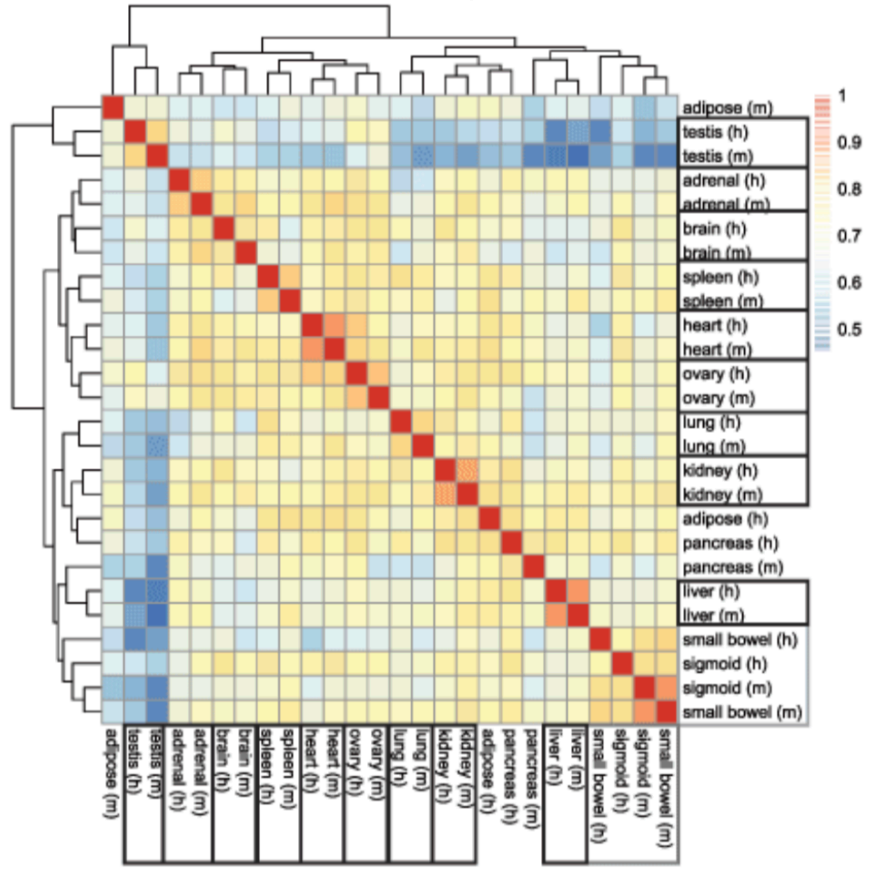
Invest in replicates!

- The most effective way to improve detection of differential expression in low expression genes is to add more replicates, rather than adding more reads
- The following figure from Gierlinski et al shows coverage variation in 4 replicates of a relatively simple yeast transcriptome
- The paper concludes that we should invest in 6 **biological** replicates per condition



Batch effects can happen everywhere





*"Overall, our results indicate that there is **considerable RNA expression diversity between humans and mice**, well beyond what was described previously, likely reflecting the fundamental physiological differences between these two organisms. "*

Lin, Lin, and Snyder (2014). PNAS 111:48

*"**Once we accounted for the batch effect** (…), the comparative gene expression data no longer clustered by species, and instead, we observed **a clear tendency for clustering by tissue**."*

Gilad & Mizrahi-Man (2015). F1000Research 4:121

Credit: http://chagall.med.cornell.edu/RNASEQcourse/

ENCODE's* study design was not optimal

Most human samples were sequenced separately from the mouse samples:

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Many tissues were not sex-matched

Tissue	Human	Mouse
adipose	FEMALE	MALE
adrenal	MALE	FEMALE
brain	FEMALE	MALE
heart	FEMALE	FEMALE
kidney	MALE	FEMALE
liver	MALE	FEMALE
lung	FEMALE	FEMALE
ovary	FEMALE	FEMALE
pancreas	FEMALE	FEMALE
sigmoid colo	MALE	FEMALE
small bowel	FEMALE	FEMALE
spleen	FEMALE	MALE
testis	MALE	MALE

not all variables can be controlled for

human data: deceased organ donors

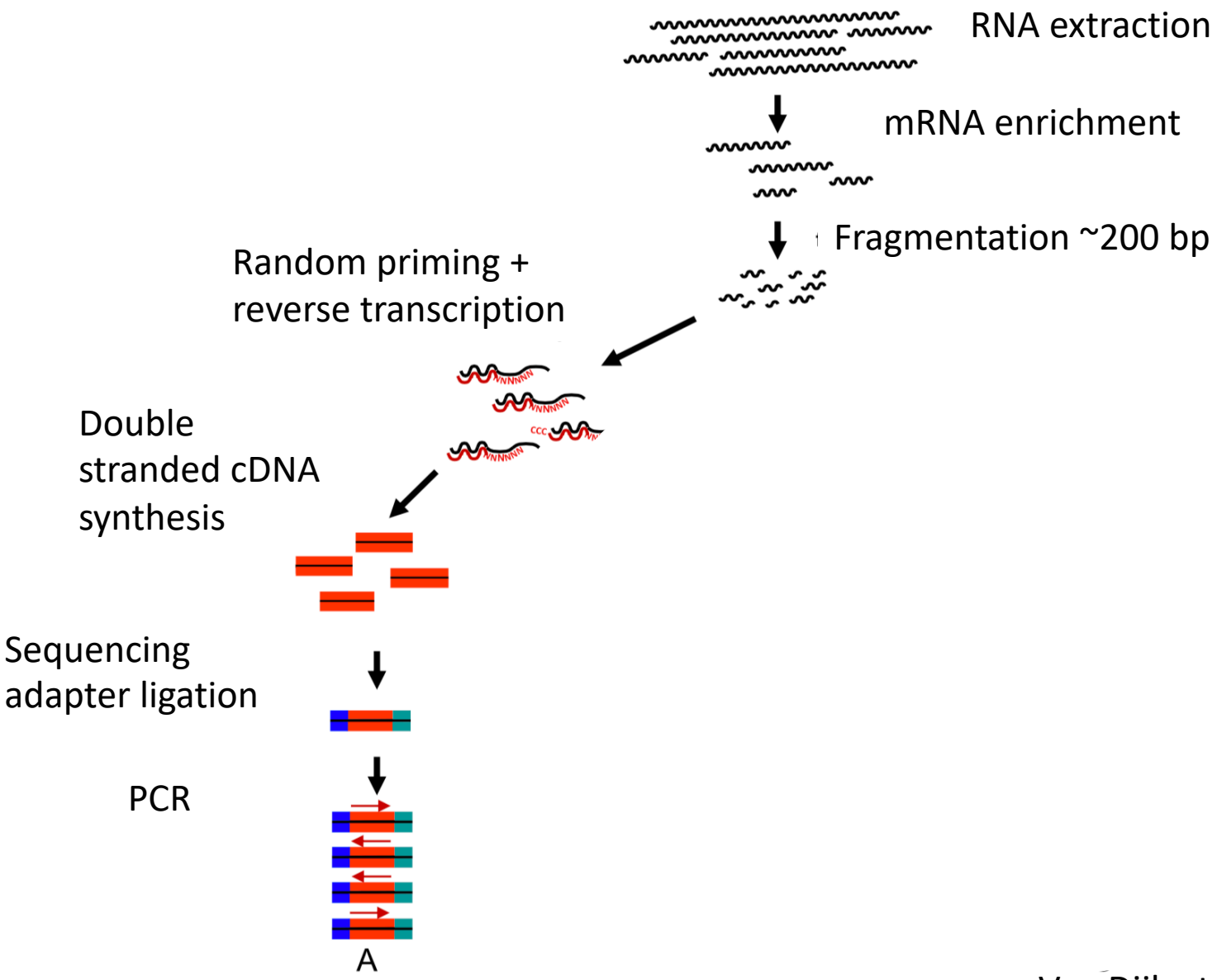
mouse data: 10-week-old littermates

and that's ok, but you got to be mindful of these limitations when making bold claims

A very good read (including the reviews and comments) that discusses many scientific as well as ethical issues: https://f1000research.com/articles/4-121/v1

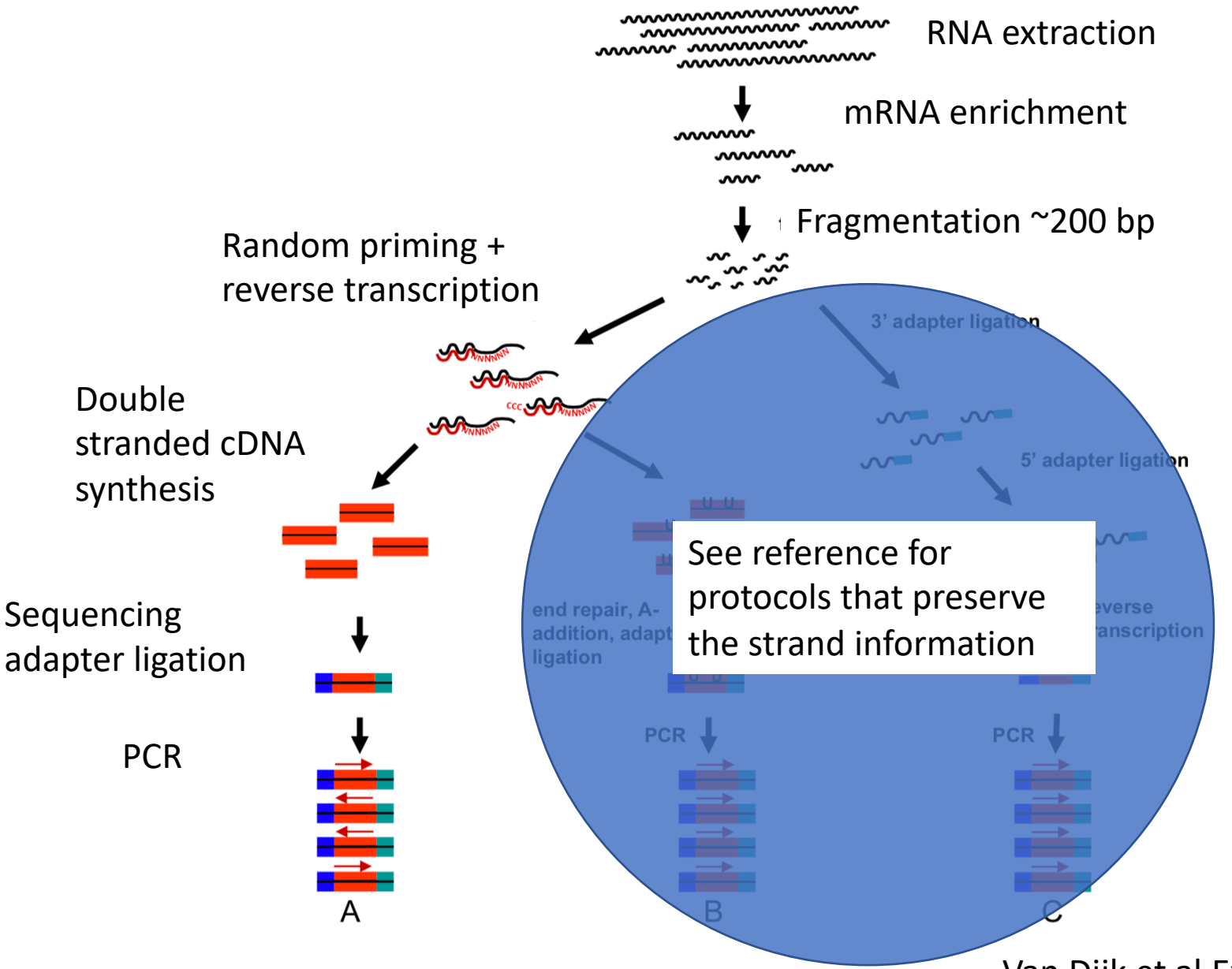
* not just ENCODE: see e.g. Leek et al. (2010) Nat Rev Gen 11(10) 733-739 or Jaffe & Irizarry (2014) Genome Biol 15(R31) 1–9

Classic Illumina RNAseq Library Preparation



Van Dijk et al. Experimental Cell Research 2014

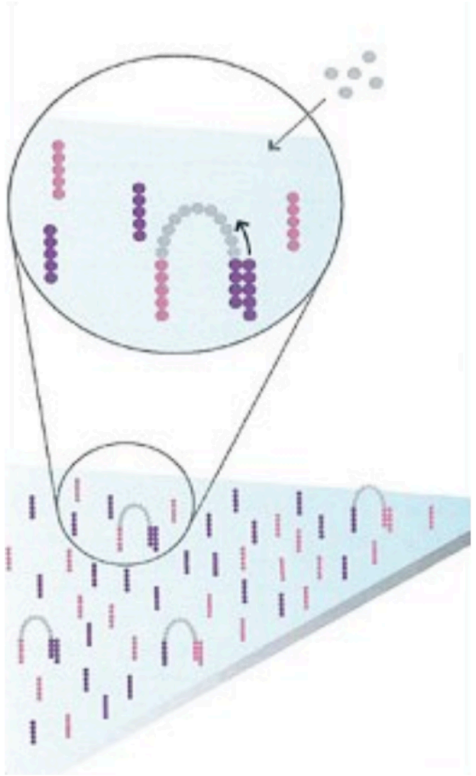
Classic Illumina RNAseq Library Preparation



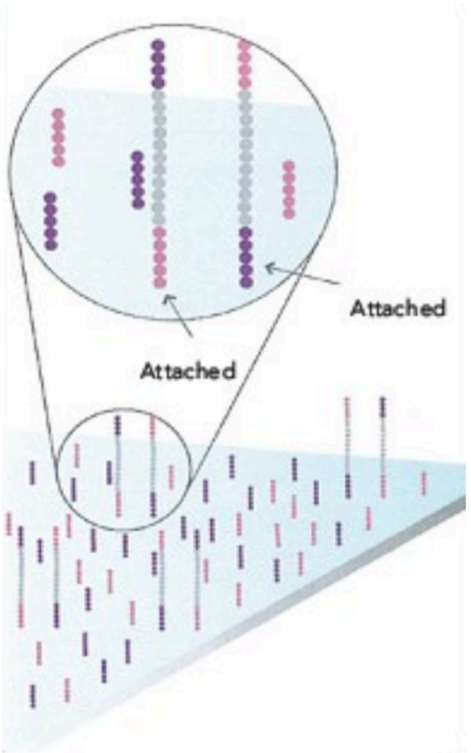
See reference for protocols that preserve the strand information

Van Dijk et al.Experimental Cell Research 2014

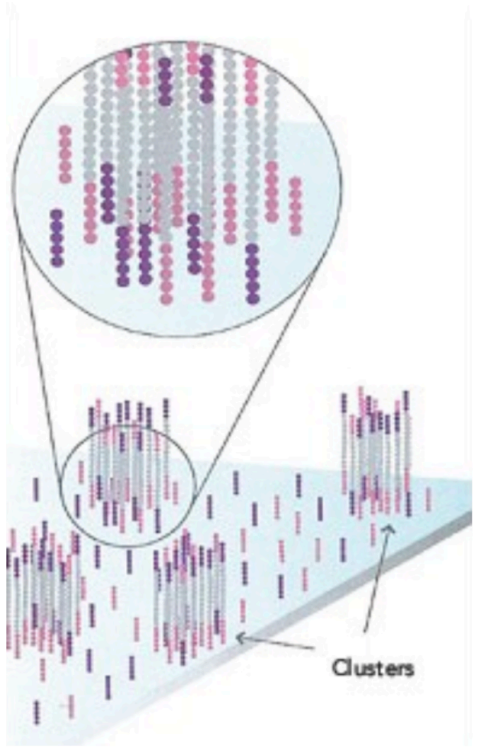
Illumina Sequencing



bridge amplification

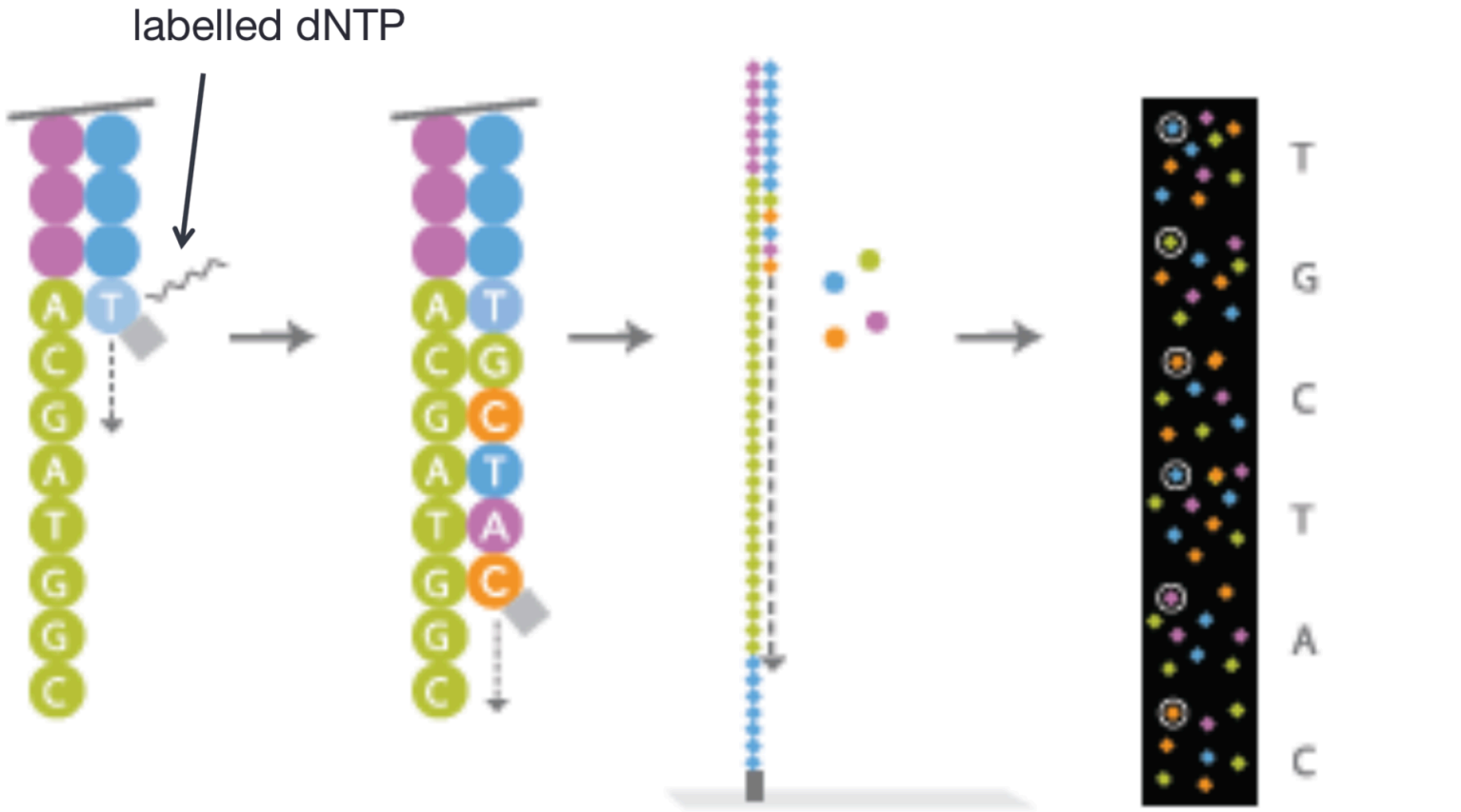


denaturation



cluster generation
removal of complementary strands → identical fragment copies remain

Illumina Sequencing



1. extend 1st base
2. read
3. deblock

repeat for 50 – 100 bp

generate base calls

Avoiding bias by pooling samples