Bioinformatics for RNA-seq

Wenwen Hou
Rebecca Batorsky
Albert Tai

May 2020

Requirements

- HPC Cluster Account available to Tufts affiliates
- VPN if working off campus
- Basic knowledge of Linux and HPC:
  - Intro to Linux
  - HPC Quick Start guide or Intro to HPC
  - Introduction to R

We'll test out access together during this session.
Depending on the number/type of questions, we may choose to follow up after the session.

Course Format

1-hour Zoom Introduction

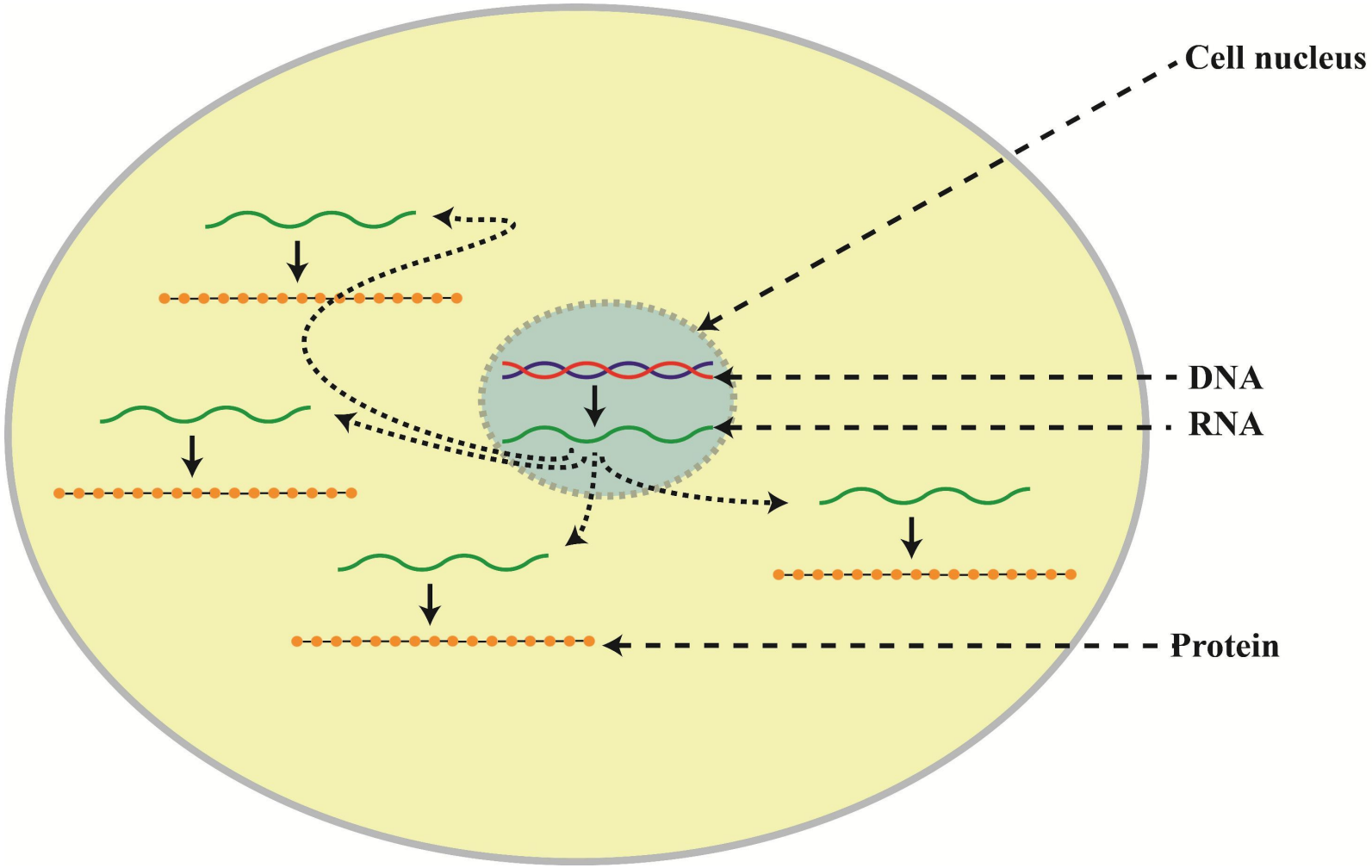
~3 hours of self-guided material on github, suggested to be completed over the **next week:** https://huoww07.github.io/Bioinformatics-for-RNA-Seq/

(working with a partner is encouraged)

Piazza

- Please ask and answer questions liberally on Piazza
- Steps to enroll in class if you are not auto-enrolled:
  - https://piazza.com/tufts
  - Bioinformatics 2: Intro to RNA sequencing Bioinformatics
  - Join as student
- If you can't access Piazza for some reason please let us know Wenwen.Huo@tufts.edu or Rebecca.Batorsky@tufts.edu

DNA and RNA in a cell

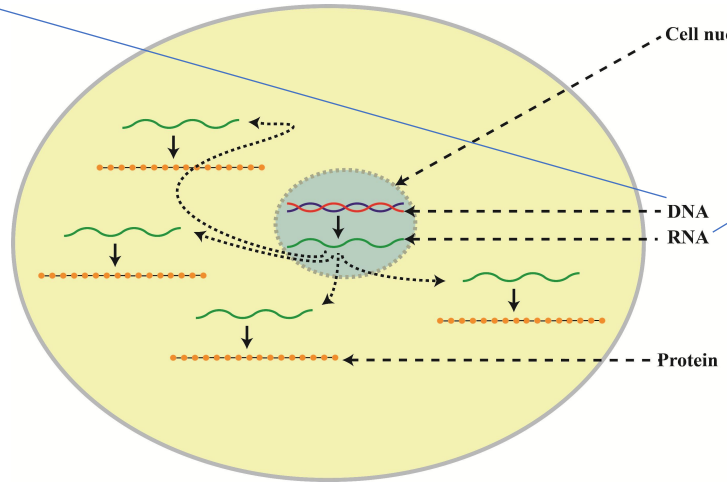


https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Two common analysis goals

DNA Sequencing

- Fixed copy of a gene per cell
- Analysis goal:
  Variant calling and interpretation



RNA Sequencing

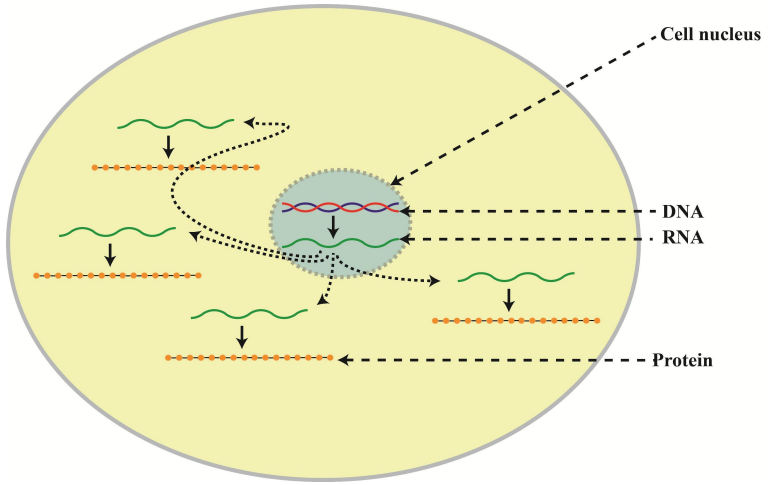
- Copy of a transcript per cell depends on gene expression
- Analysis goal: Differential expression and interpretation

https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

This workshop will cover RNA sequencing

Sequen

- Fixed co[illegible]ene per cell
- Analy[illegible]
  Vari[illegible]ling a[illegible]terpretation



RNA Sequencing

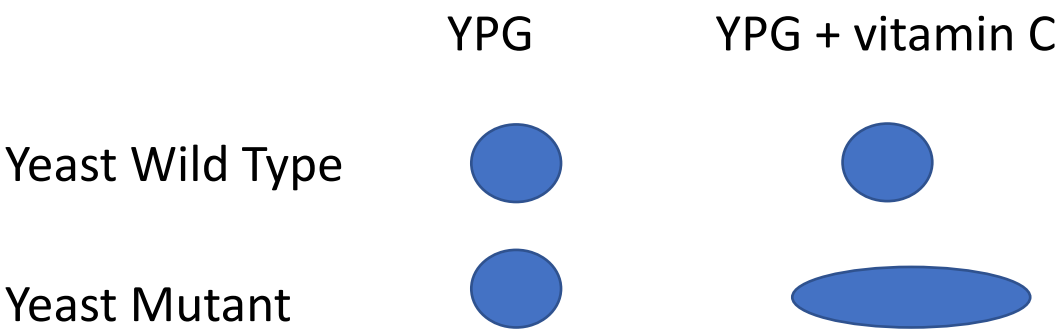
- Copy of a gene per cell depends on gene expression
- Analysis goal: Differential expression and interpretation

Not today!
Check out our “Intro to NGS” workshop:
https://rbatorsky.github.io/intro-to-ngs-bioinformatics/

https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

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

↓

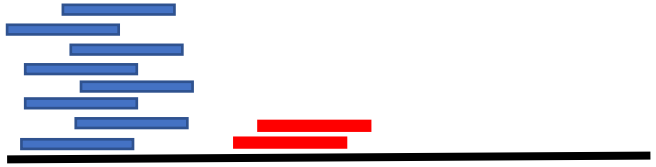
Concentration of proteins

↓

Difference in protein activity

↓

Yeast Wild Type Phenotype



Read Counts

↓

Number of mRNA copies

↓

Concentration of proteins

↓

Difference in protein activity

↓

Yeast 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)

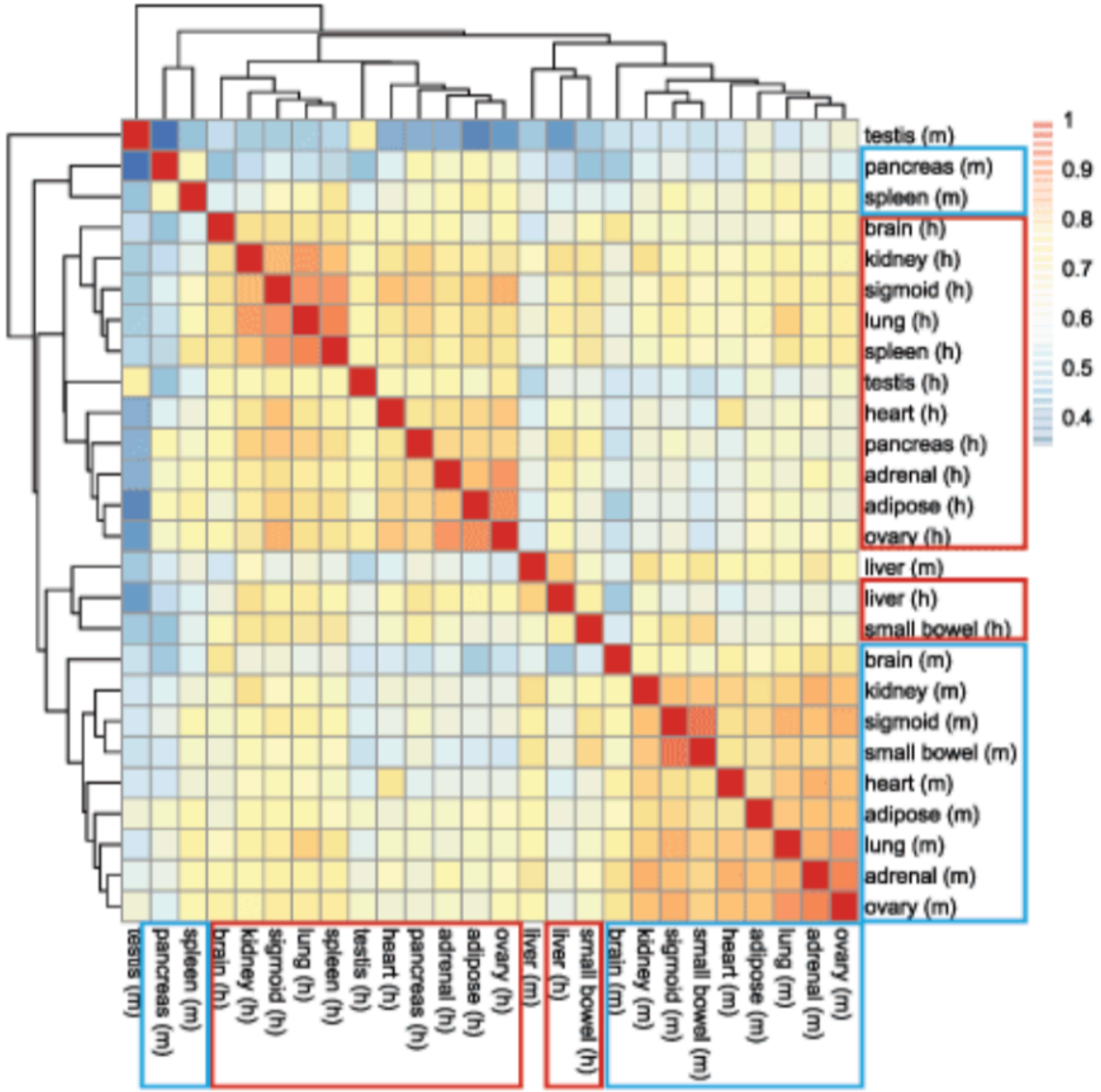
Lessons from the ENCODE study (2014)

ENCODE was designed to test “the common notion that major developmental pathways are highly conserved across a wide range of species, in particular across mammals.”

How close are mouse and human in terms of gene expression across multiple tissues?

https://f1000research.com/articles/4-121

Initial publication showed mouse and human cluster separately

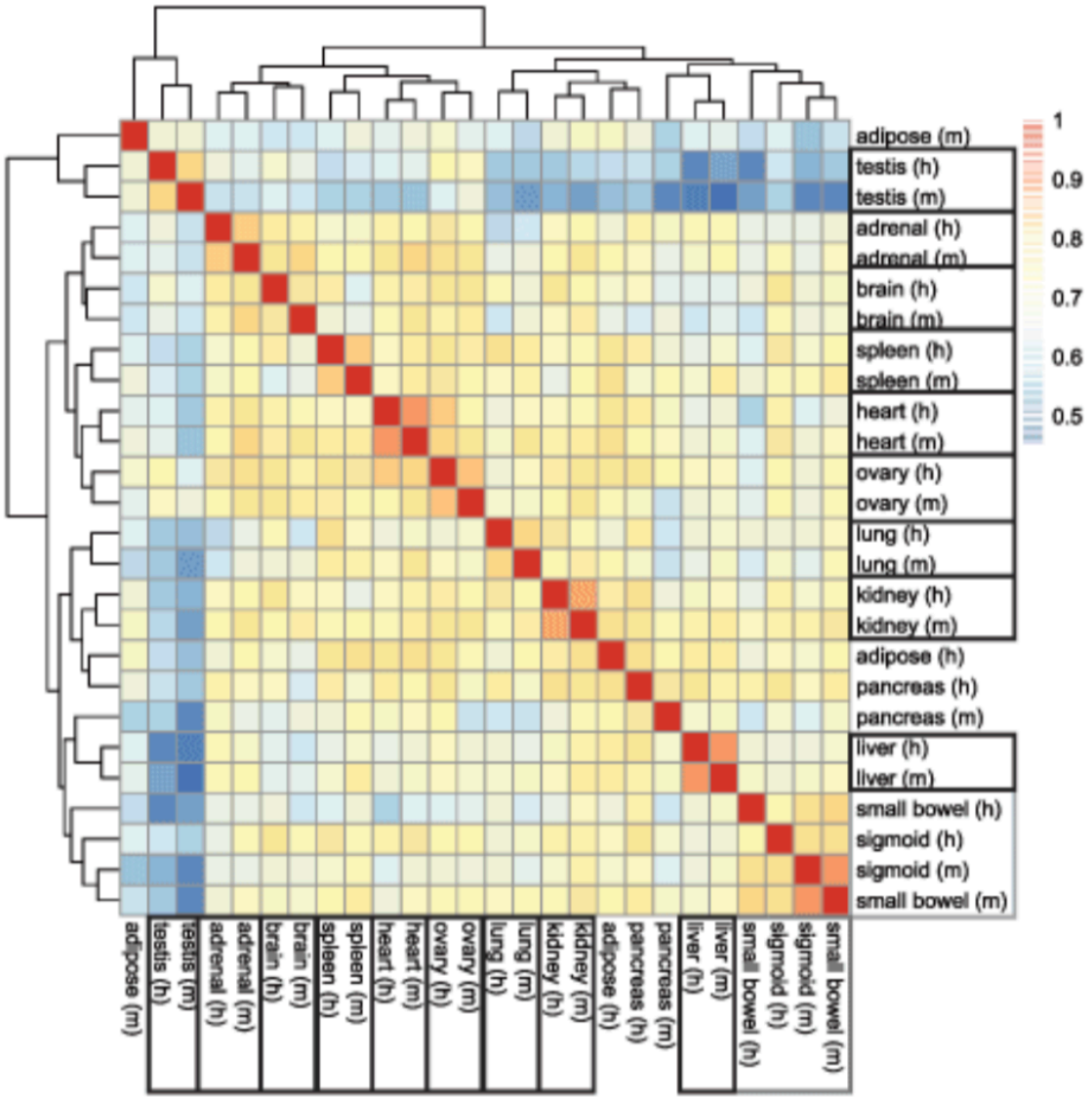


"Overall, our results indicate that there is ***considerable RNA expression diversity between humans and mice****, well beyond what was described previously, likely reflecting the fundamental physiological differences between these two organisms. "*

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

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

Once batch effects were accounted for: clustering by tissue



> *"**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* 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

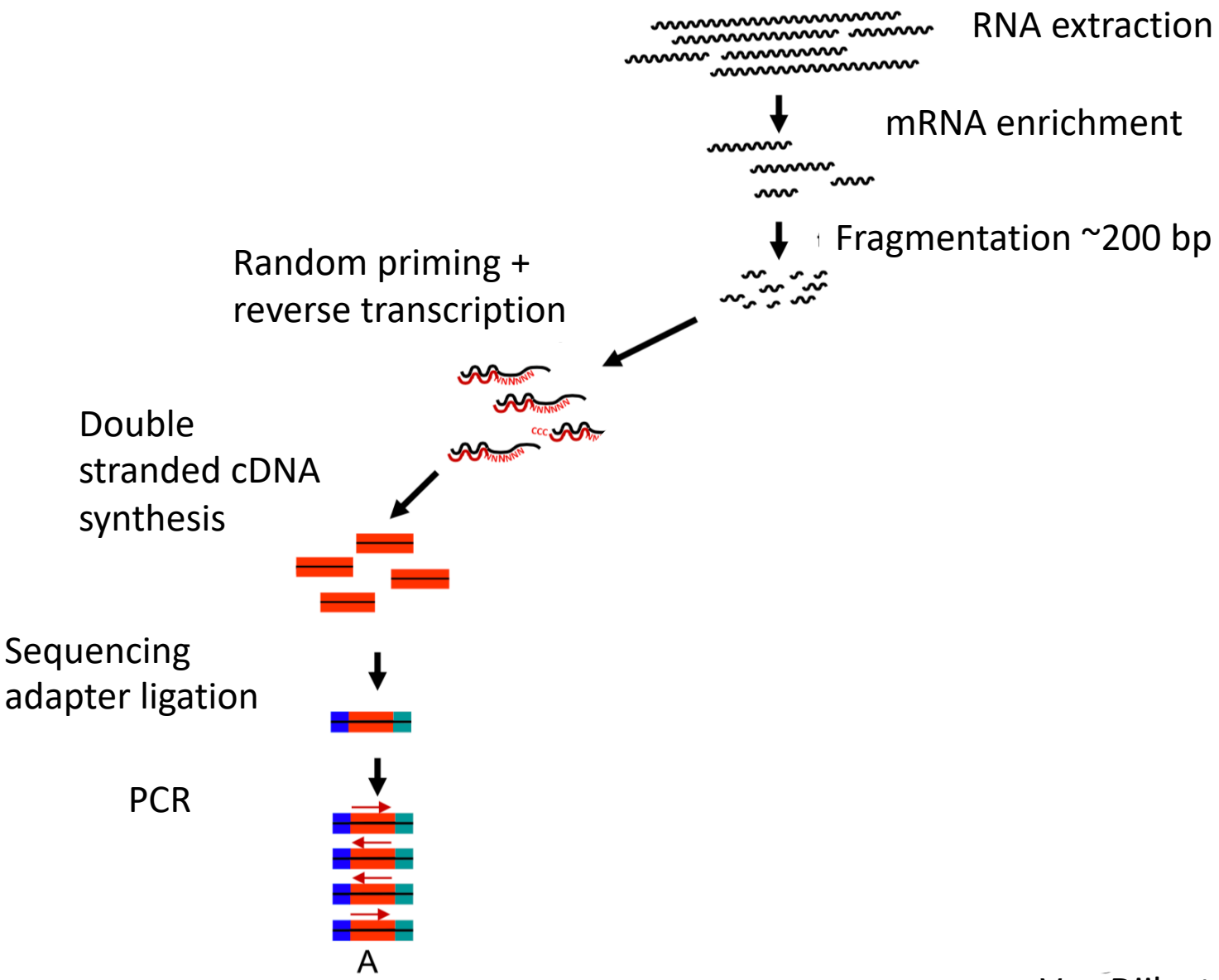
Many tissues were not sex-matched

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 just ENCODE! Good review! https://f1000research.com/articles/4-12

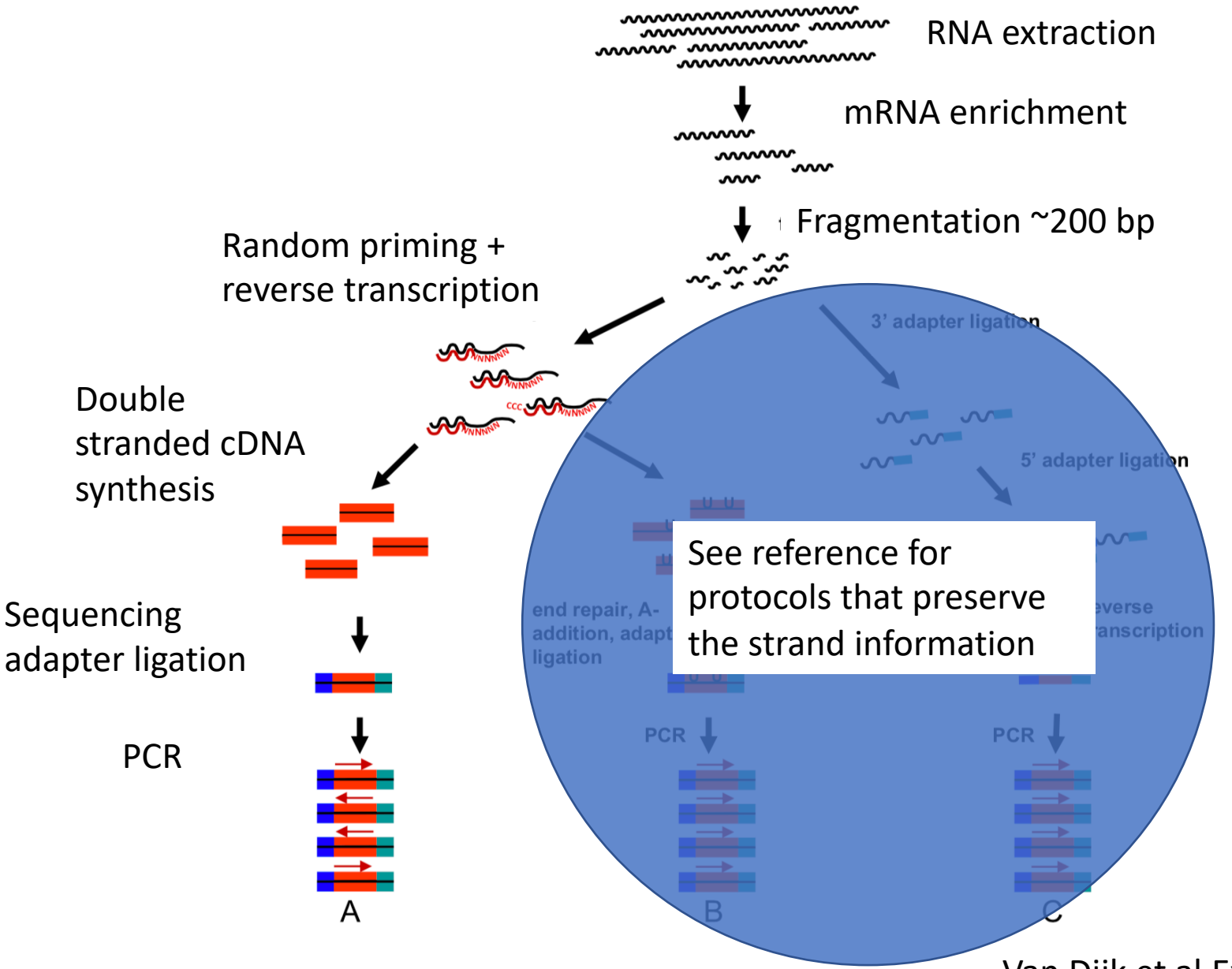
Credit: http://chagall.med.cornell.edu/RNASEQcourse/

Classic Illumina RNAseq Library Preparation



Van Dijk et al. Experimental Cell Research 2014

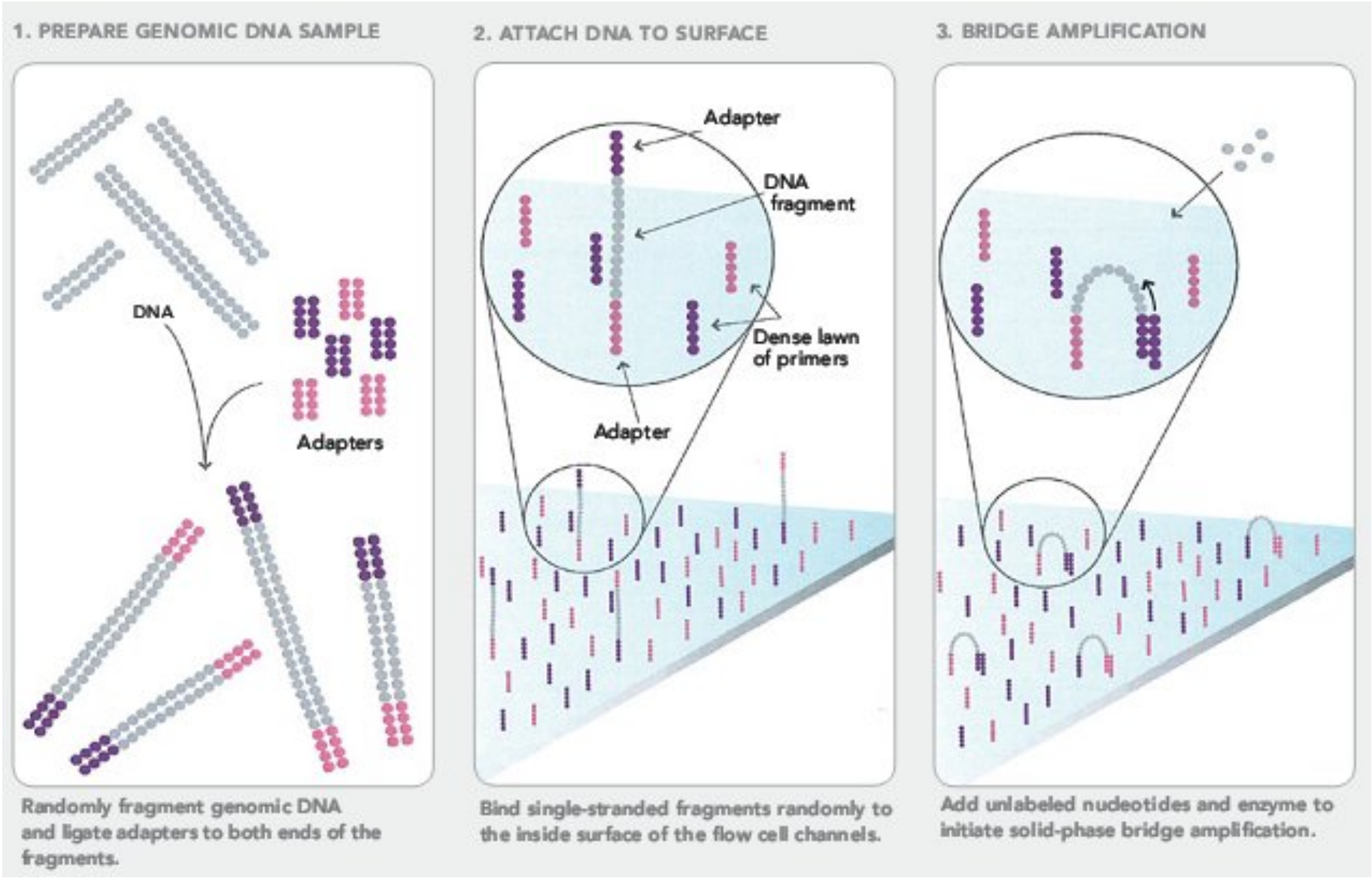
Classic Illumina RNAseq Library Preparation



See reference for protocols that preserve the strand information

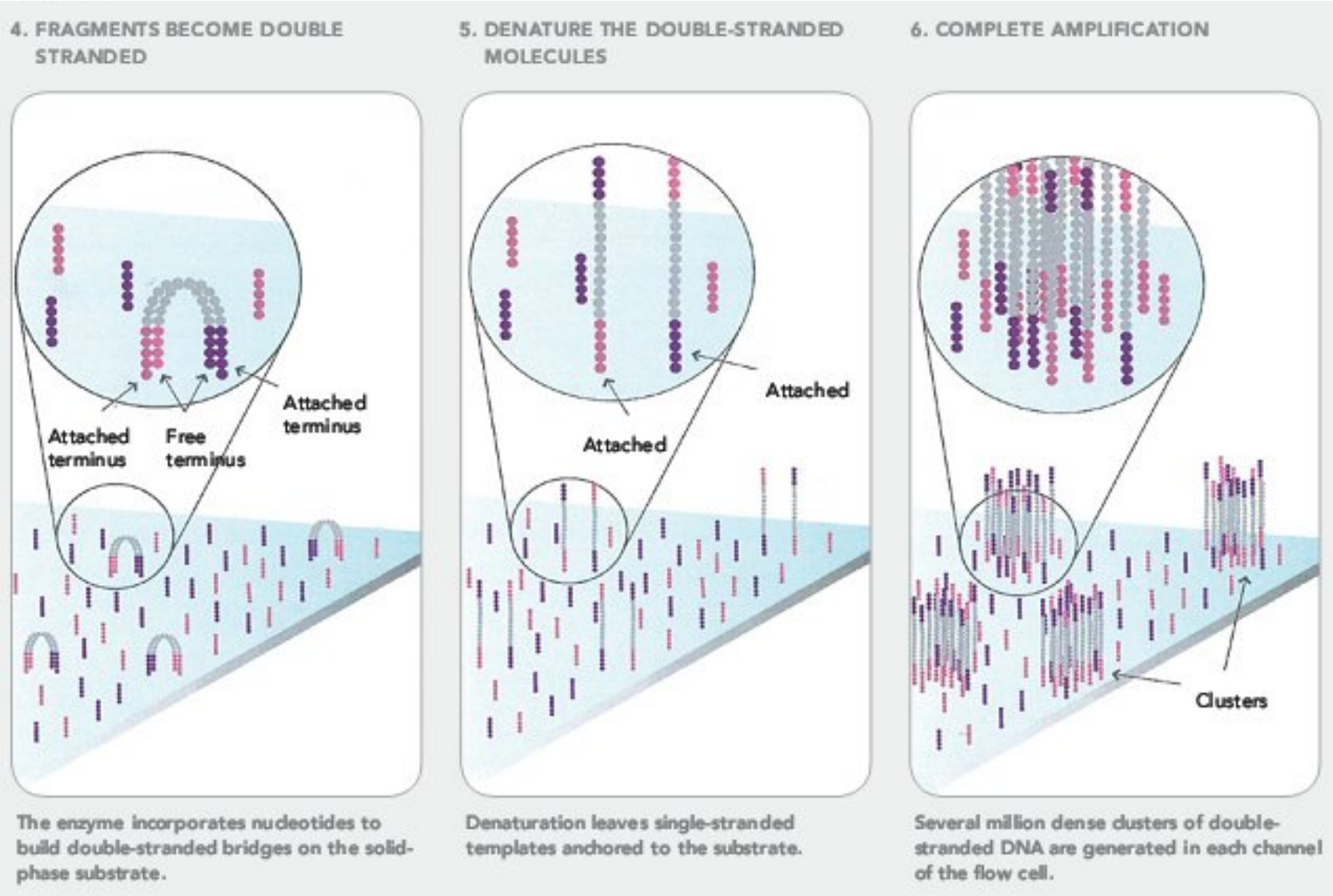
Van Dijk et al.Experimental Cell Research 2014

Next Generation Sequencing (NGS)



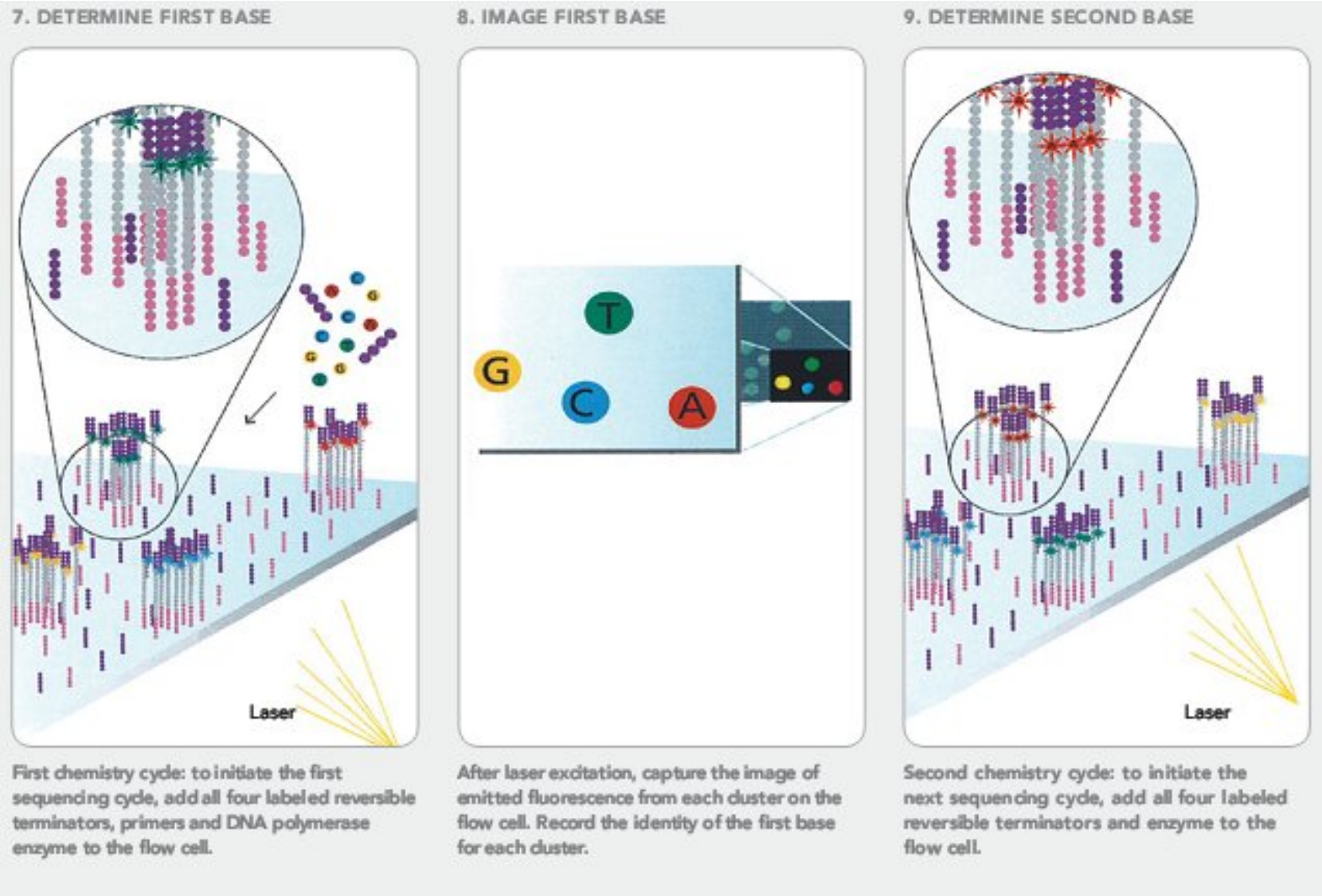
https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



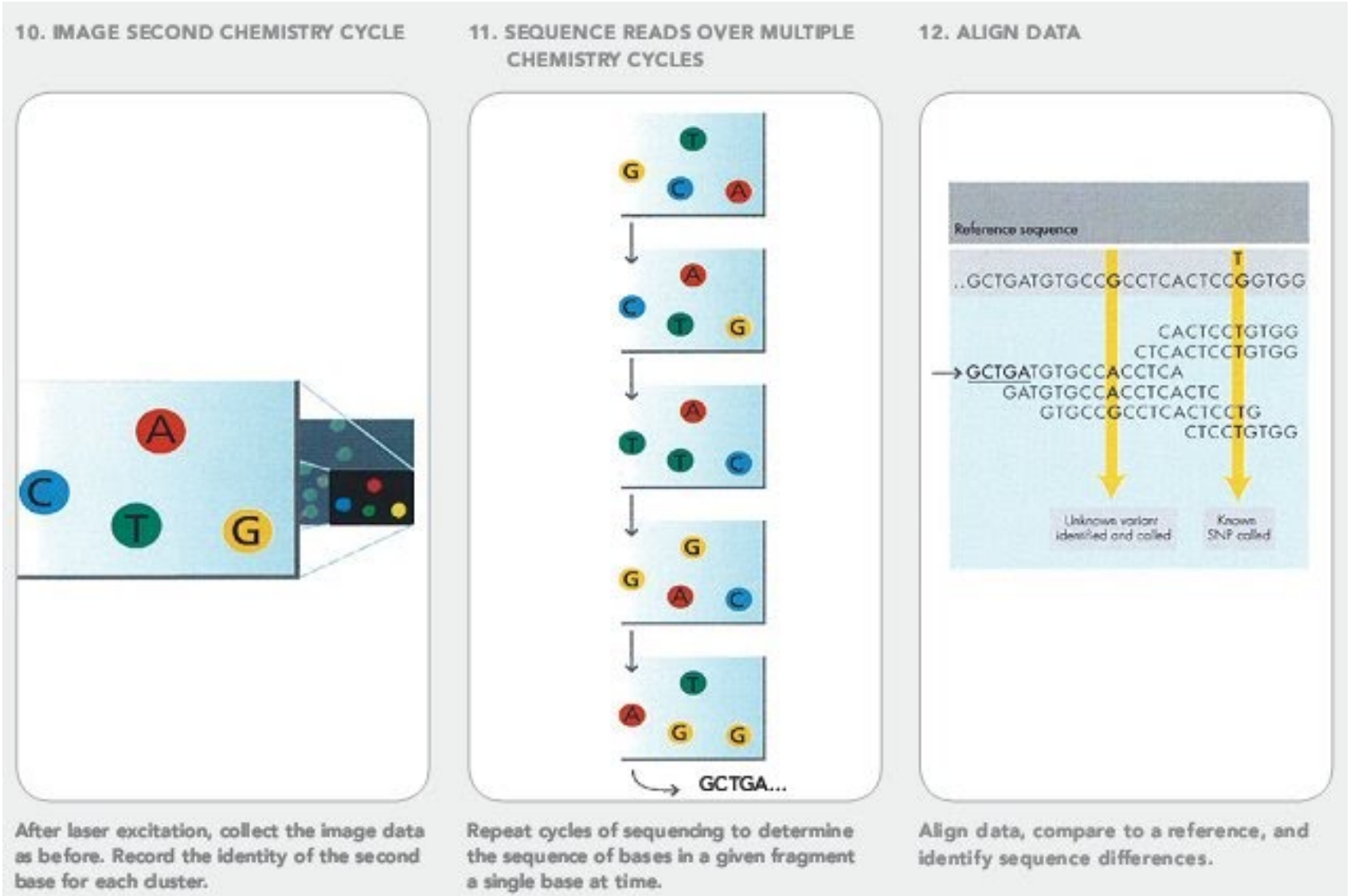
https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



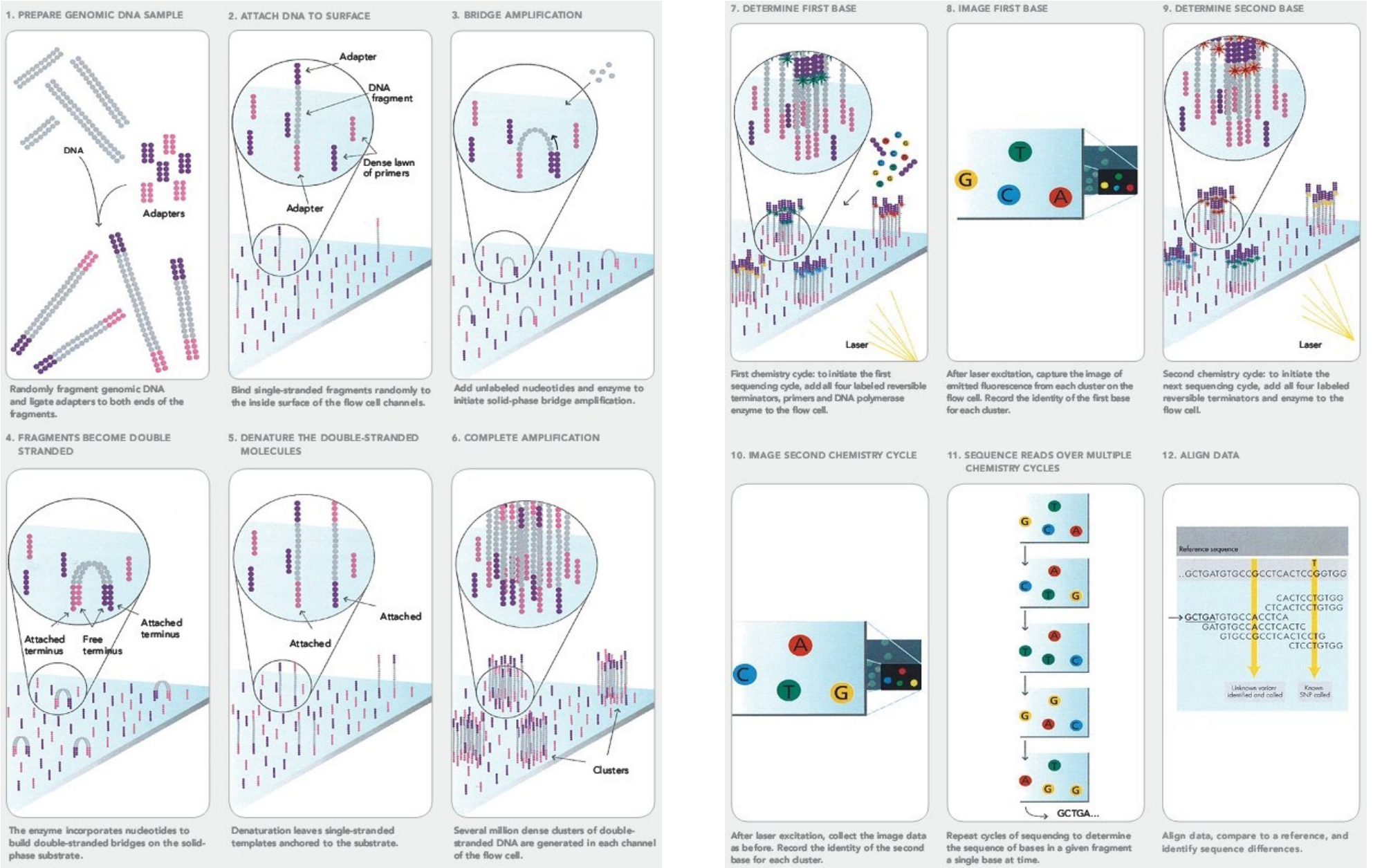
https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)

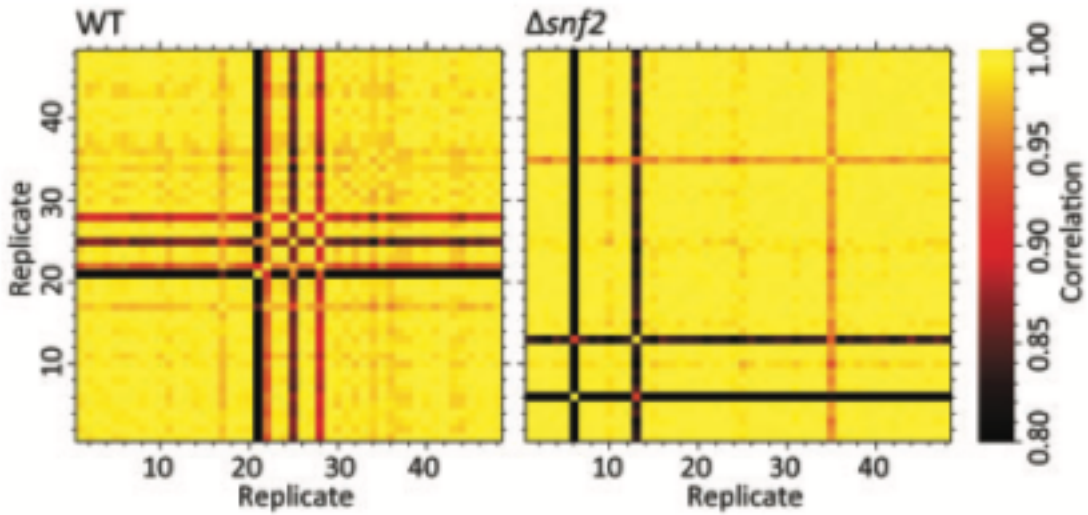


This Illumina Video is helpful for visualization!

Dataset for this course

"Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment"
Gierlinski et al Bioinformatics 2015

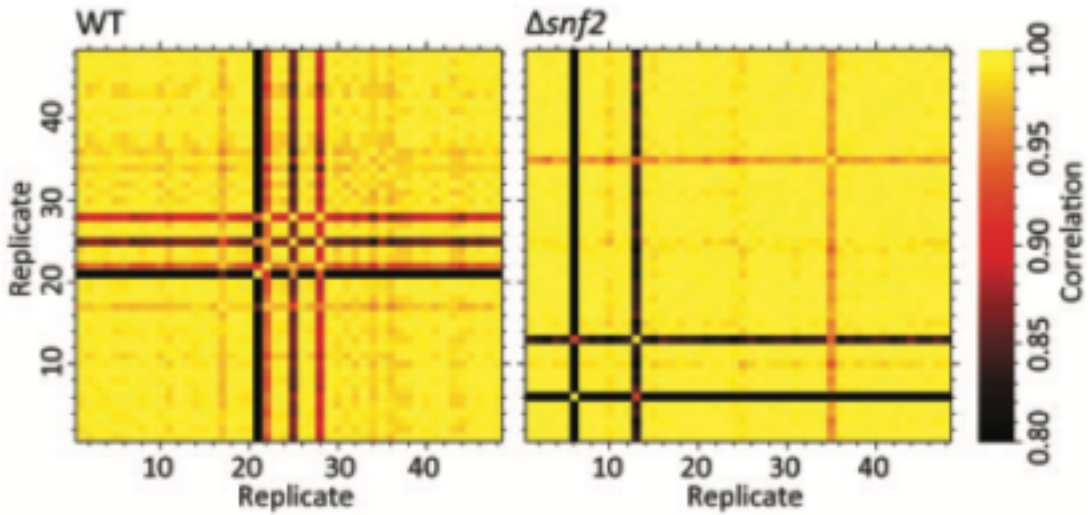
- mRNA data from 48 biological replicates of two Saccromyces cerevisiae populations
- Wildtype (WT) and SNF2 knock-out (Δsnf2)
- Unusually comprehensive analysis of variability in sequencing replicates



Dataset for this course

"Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment"
Gierlinski et al Bioinformatics 2015

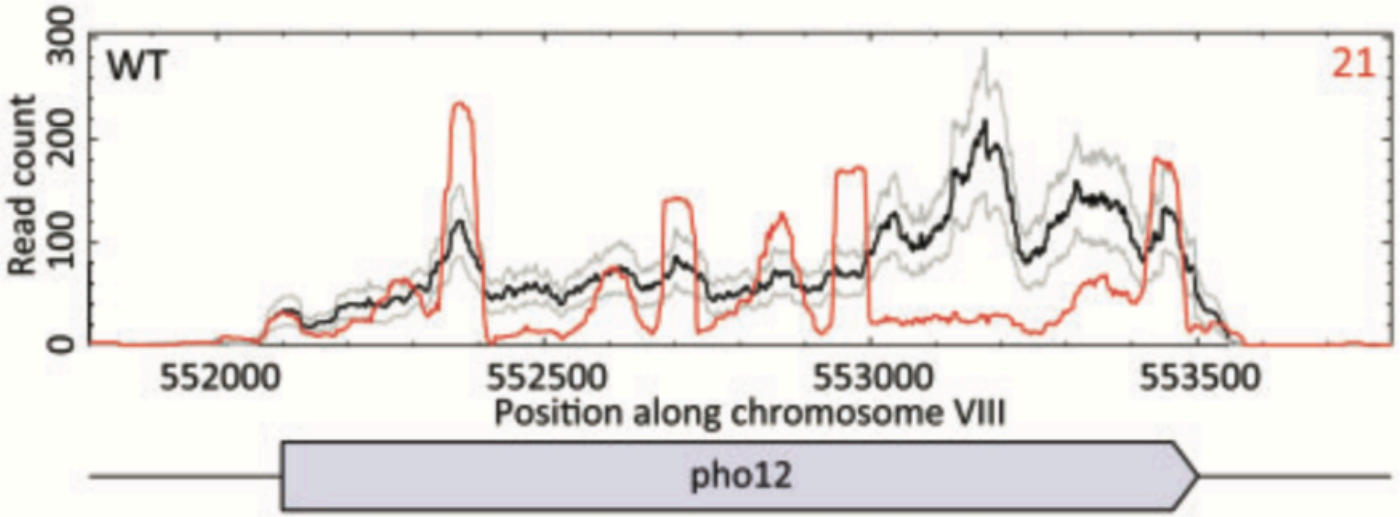
- mRNA data from 48 biological replicates of two Saccromyces cerevisiae populations
- Wildtype (WT) and SNF2 knock-out (Δsnf2)
- Unusually comprehensive analysis of variability in sequencing replicates



**Course dataset will consider 7 subsamples of one WT replicate and one SNF2 mutant, to demonstrate differences between populations and details of processing batches from different conditions

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



Gierlinski et al Bioinformatics 2015
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4754627