Introduction

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

January 8, 2019

Class details

Instructors: Friederike Dündar (frd2007@med.cornell.edu)
and
Luce Skrabanek (las2017@med.cornell.edu)

Supported by: Akanksha Verma (akv3001@med.cornell.edu)

This is a hands-on class: **Please always bring your laptop!**

We will provide the slides and code before or during class here:
https://bit.ly/2CUdS9z[1]

The **final grade** will be made up by homework assignments (30%; starting next week) and a bioinformatics project (70%; we will give you more details during the third class).

Questions?

[1] http://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2018/

ANGSD and Bioinformatics

NGS data

Next-generation sequencing	
High-throughput	millions of nucleotides can be sequenced at once
Relatively cheap	experiments involving NGS have become abundant

Human Genome Project (1990-2003)

ANGSD relies on bioinformatics

Next-generation sequencing	
High-throughput	millions of nucleotides can be sequenced at once
Relatively cheap*	experiments involving NGS have become abundant

relatively large data files are being generated on a regular basis

Bioinformatics

Processing

formatting, data wrangling

Alignment

Statistical analyses

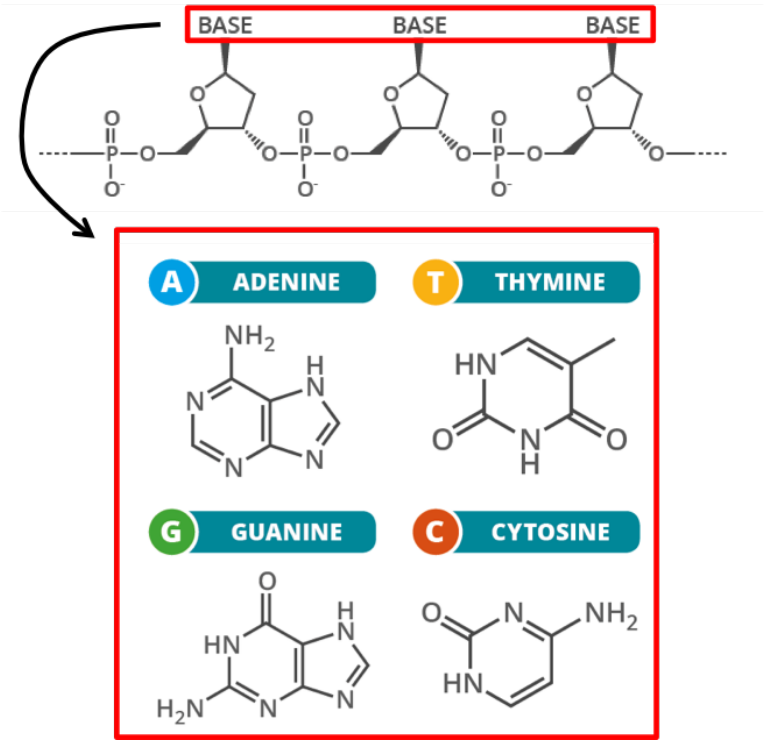
Interpretation

* The cost of analysis has remained high and is difficult to estimate!

What do we sequence?

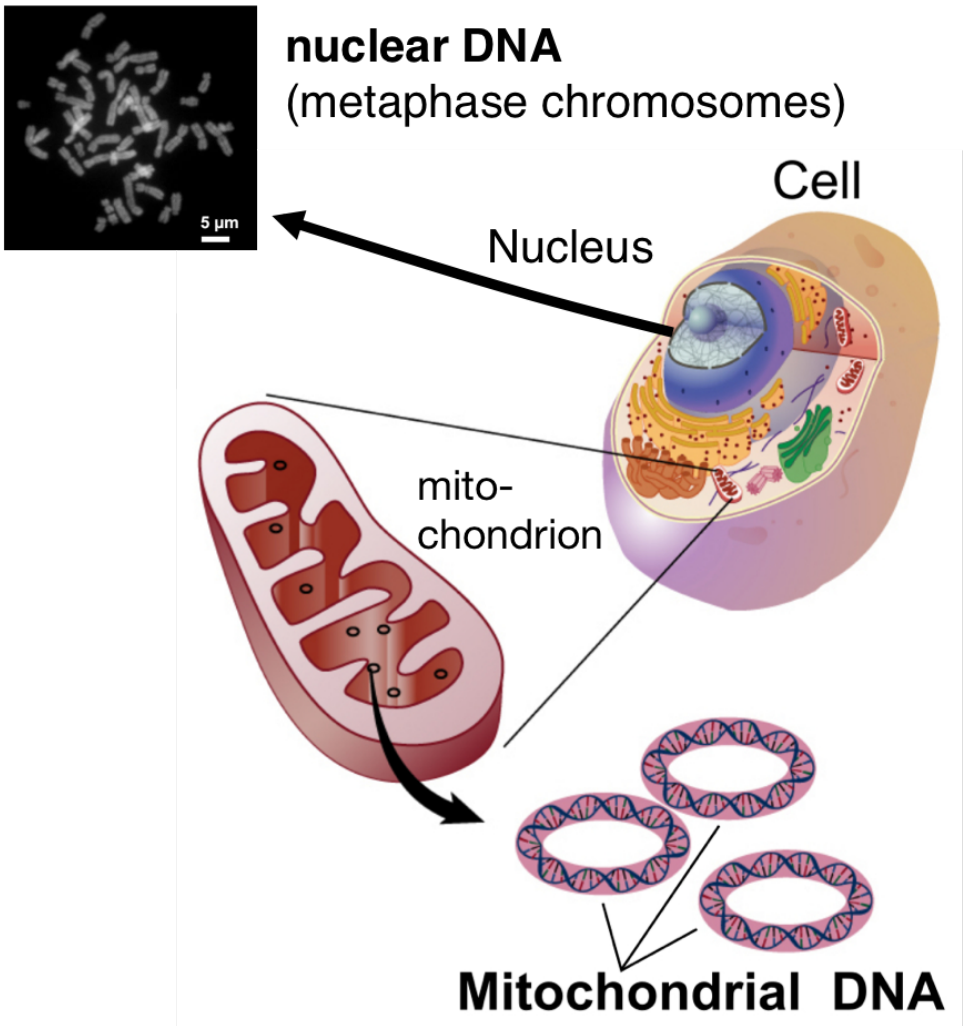
DNA: Deoxyribonucleic acid

sugar-phosphate backbone +
nitrogenous base



- a single nucleotide consists of 3 components:
  1. **sugar**: 2'-deoxyribose (5 carbon atoms = pentose)
  2. **phosphate**: 1-3 linked phosphate units attached to the 5'-carbon of the sugar
  3. **nitrogeneous base**: either a single-ring pyrimidine (cytosine, thymine) or a double-ring purine (adenine, guanine)

The molecular basis of inheritance: DNA



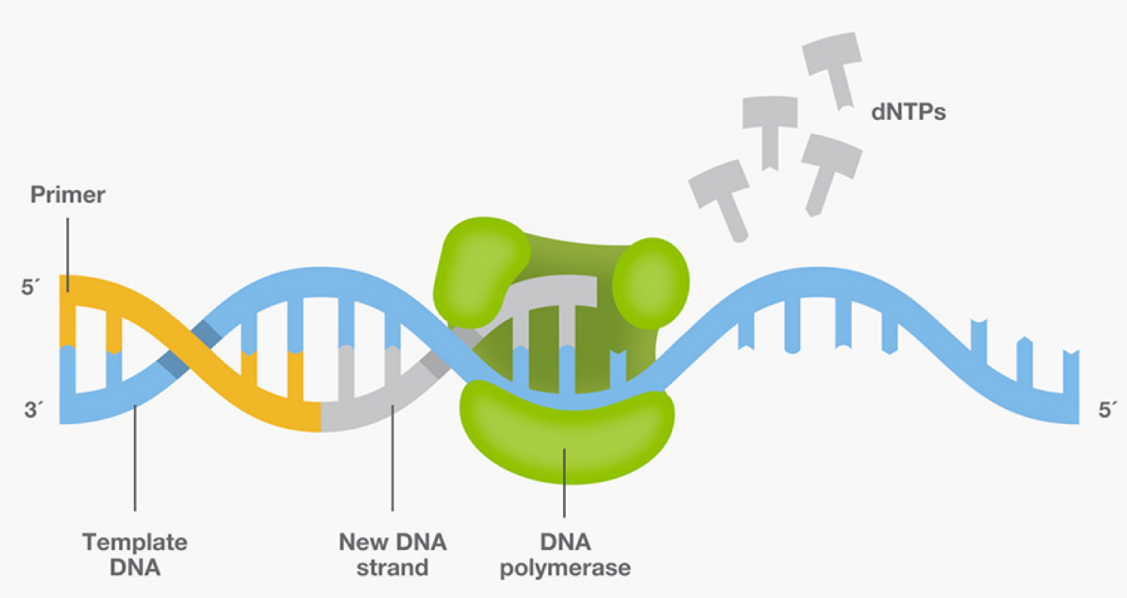
- genomic (nuclear) DNA
  - contained and replicated within the **nucleus**
  - "linear"
  - multiple chromosomes, which are inherited from both parents
- mitochondrial DNA
  - contained and replicated within **mitochondria**
  - circular
  - represents 1 chromosome
  - inherited (only!) from the mother

How do we sequence?

Decoding the DNA

Typically involves the enzymes that have naturally evolved to "read" the DNA, most notably **DNA Polymerases**.

- cannot start DNA synthesis from scratch, always need **primers**
- rely on the presence of a **template** strand, which they **complement**



https://www.thermofisher.com/uk/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/pcr-education/pcr-reagents-enzymes/pcr-basics/_jcr_content/

Next-generation sequencing (= 2nd generation)

- refers to **highly parallelized sequencing** of millions of DNA fragments at the same time
- typically encompasses 2 basic types of sequencing:
  - sequencing by **ligation** (SOLiD, Complete Genomics)
  - sequencing by **synthesis** (DNA-Pol-dependent)
    - fluororescent nt label (Solexa/Illumina)
    - proton release; ion semiconductor sequencing (Ion Torrent)

See Goodwin et al. [2016] for detailed descriptions of NGS platforms.

Next-generation sequencing

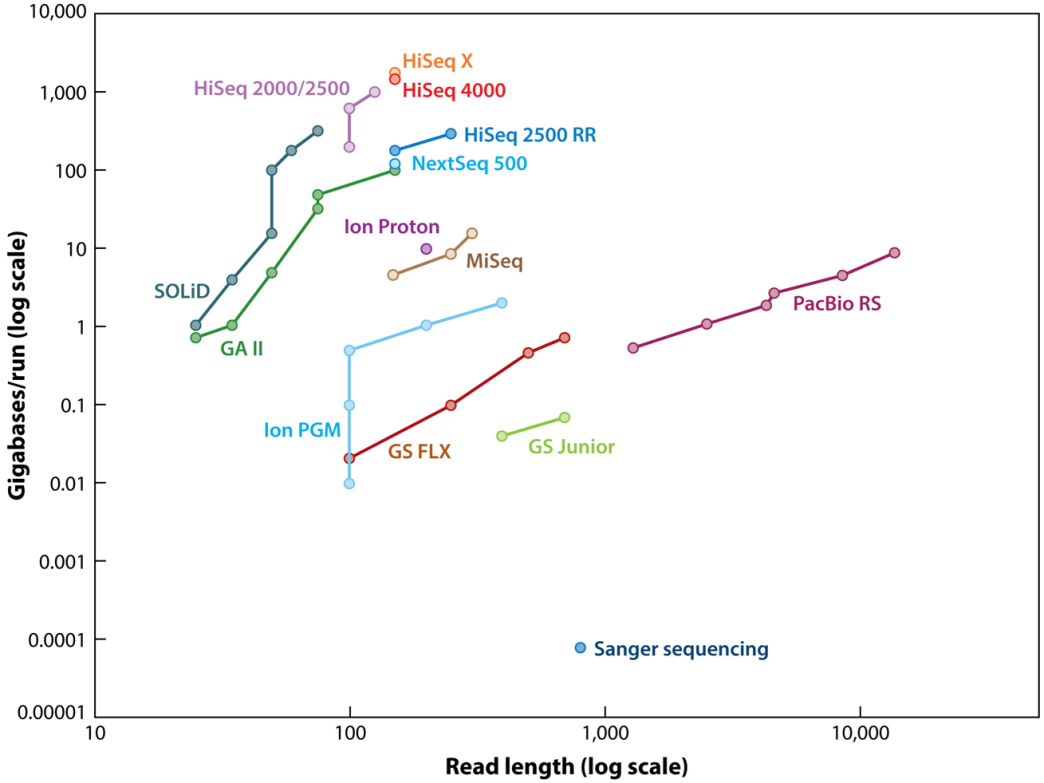
- unifying characteristics of the different NGS platforms:
- **short reads (50-250 bp)** and **short fragments (250-1000 bp)**
- require **clonal amplification** of every single DNA fragment
- markedly **higher error rates** than Sanger sequencing (0.1–15%)

NGS = Illumina-based sequencing

In practice, **Illumina's** sequencing platform is by far the most dominant one thanks to its high throughput, constant improvements, and library preparation support (kits).

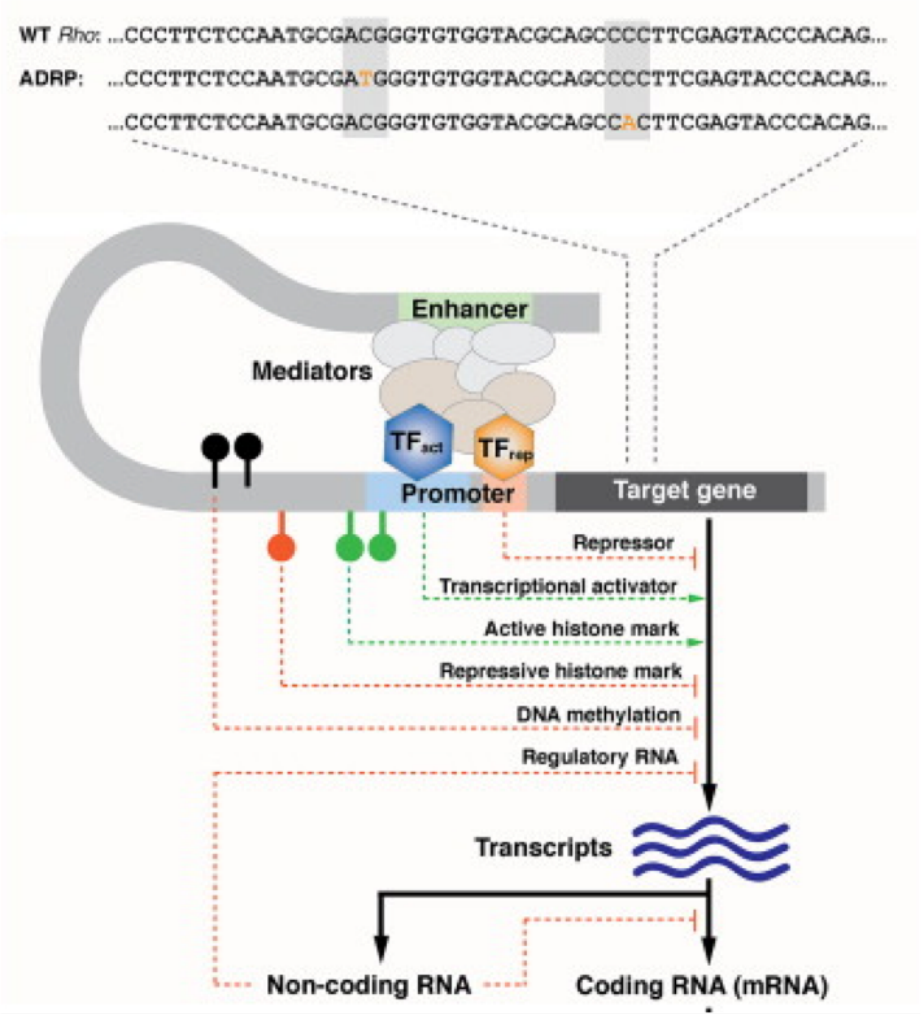
Since acquiring Solexa in 2006, Illumina has been setting the pace in terms of optimizing yield and costs (e.g. Reuter et al. [2015]).

By mid-2019, PacBio is expected to belong to Illumina, too.



Levy SE, Myers RM. 2016.
Annu. Rev. Genom. Hum. Genet. 17:95–115

Applications of NGS: not just decoding the genome

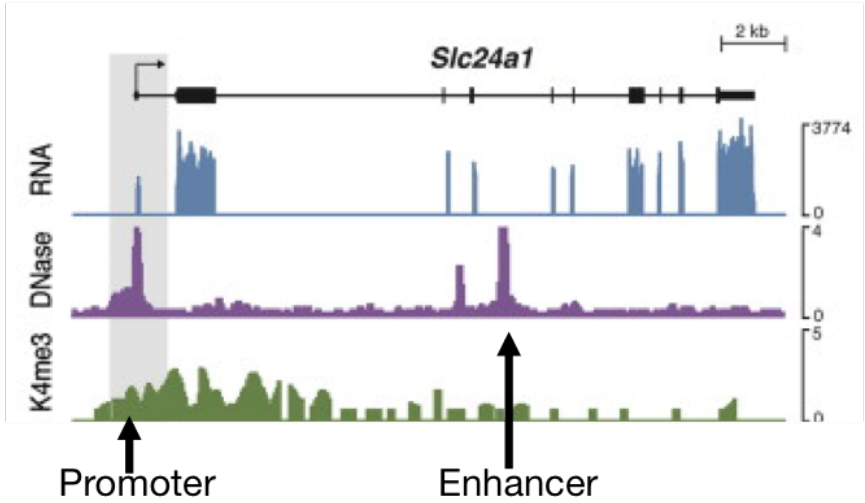


Figures from Yang et al. [2015]

(A) "Reading" the actual sequence

(B) Characterizing ('mapping') regions with certain properties

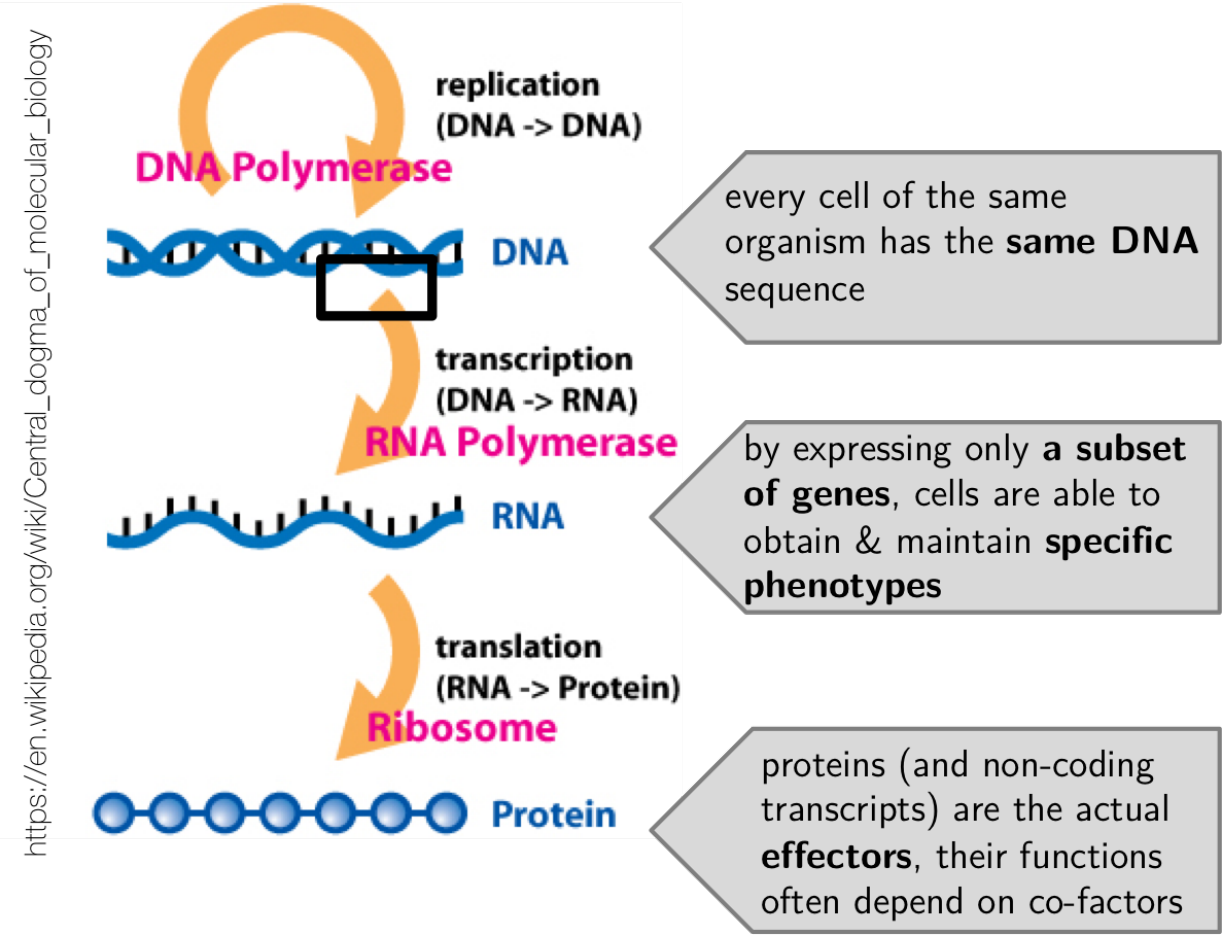
by enriching them biochemically and inferring their genome location based on statistically higher numbers of reads – the actual sequence is only needed to identify the locus of origin, the information of interest is based on abundance



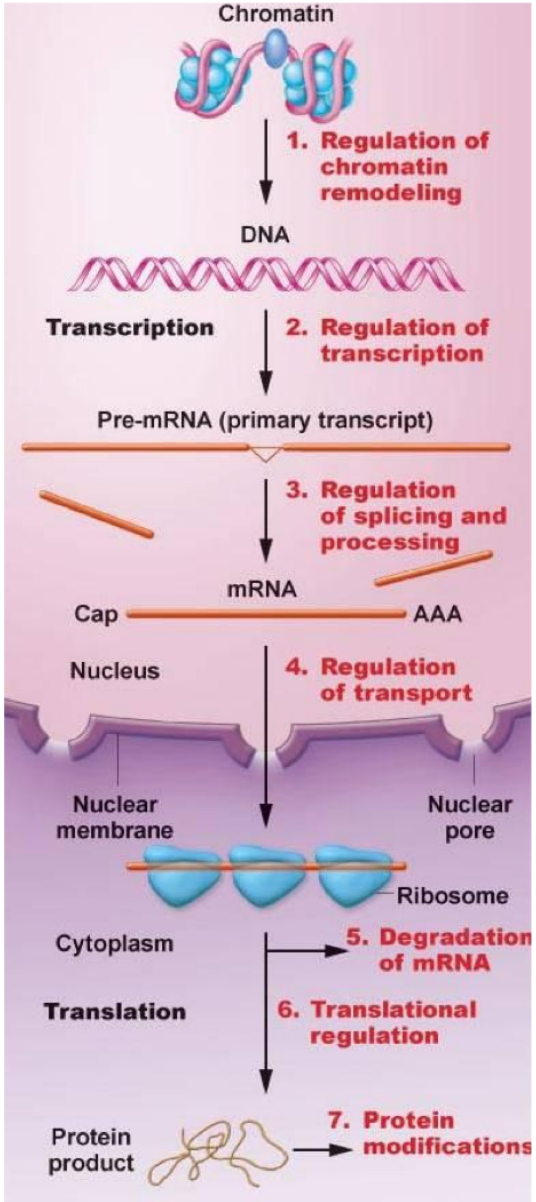
Why do we sequence?

Why do we sequence?

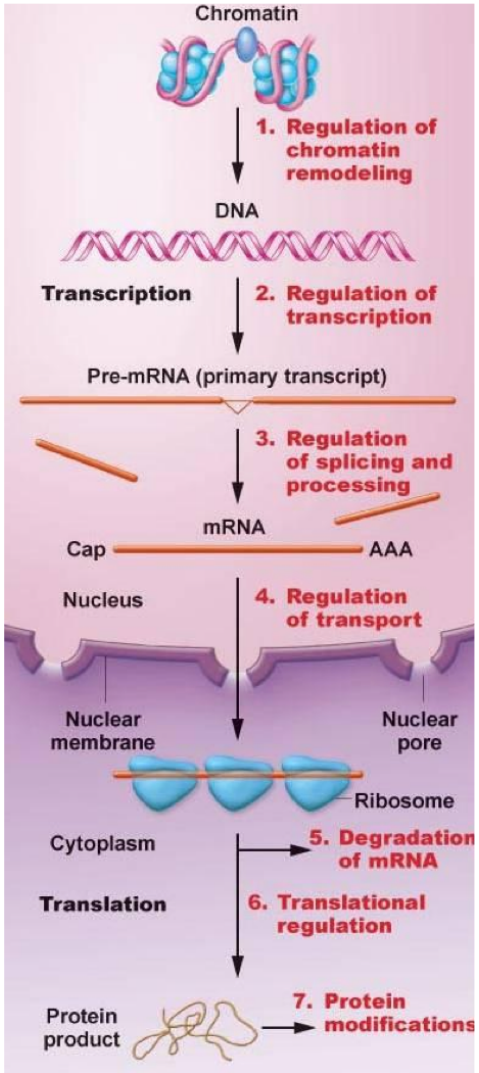
. . . to understand the molecular basis of different **phenotypes** (organisms, cells).



Understanding the genetic code and its interpretation



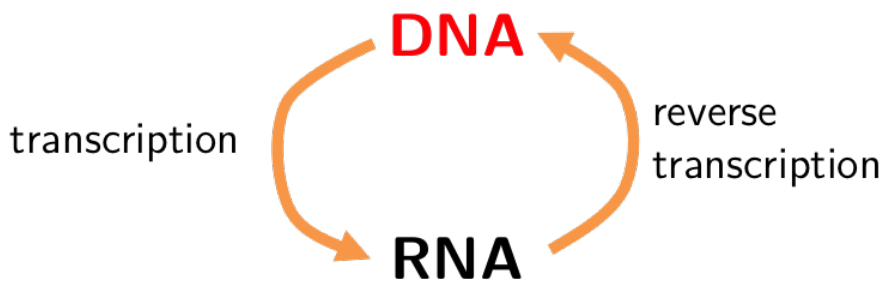
Understanding the genetic code and its interpretation



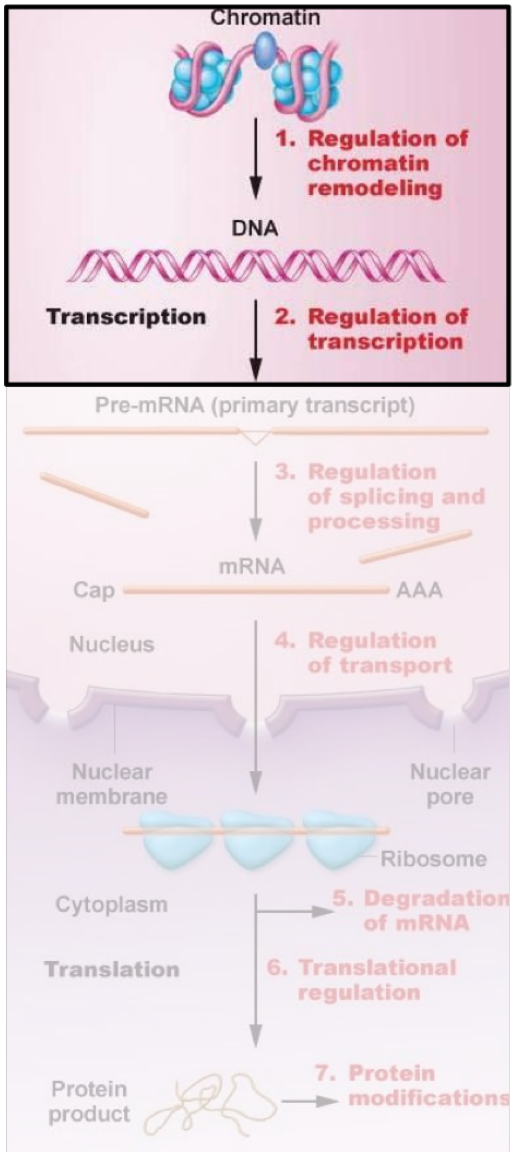
Both RNA and DNA molecules can be sequenced fairly easily in a high-throughput manner.

(A) "Reading" the actual sequence

(B) Characterizing ('mapping') regions with certain properties



Understanding DNA: it's not just about the letters



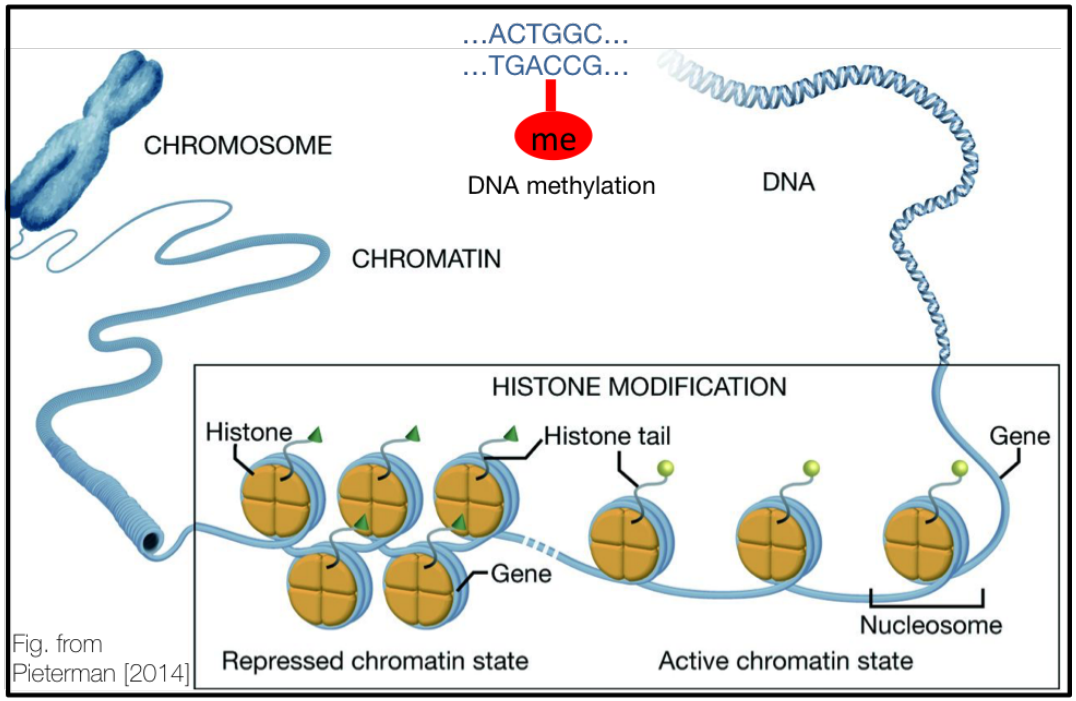
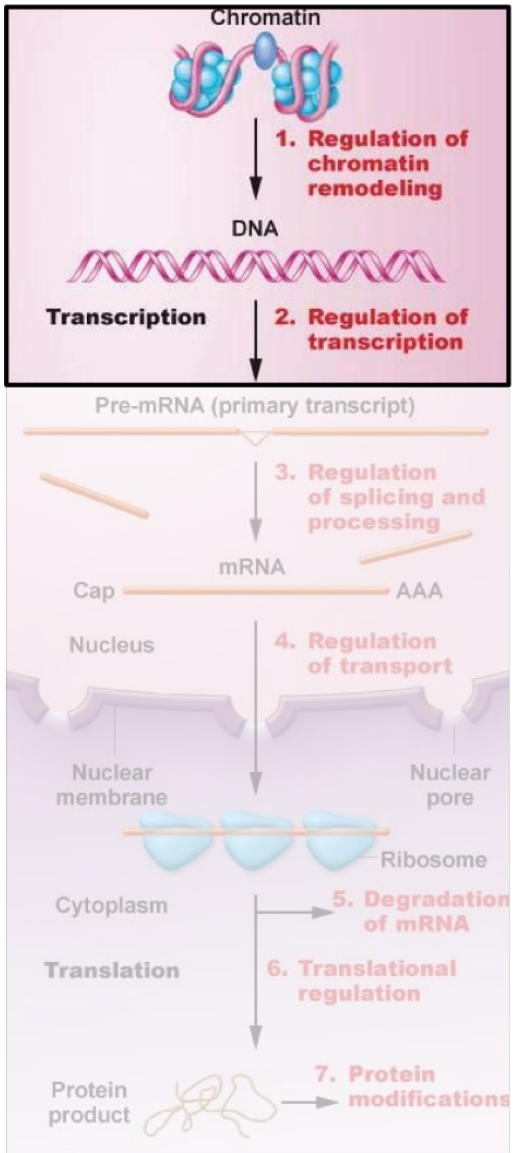


Fig. from Pieterman [2014]

Understanding DNA



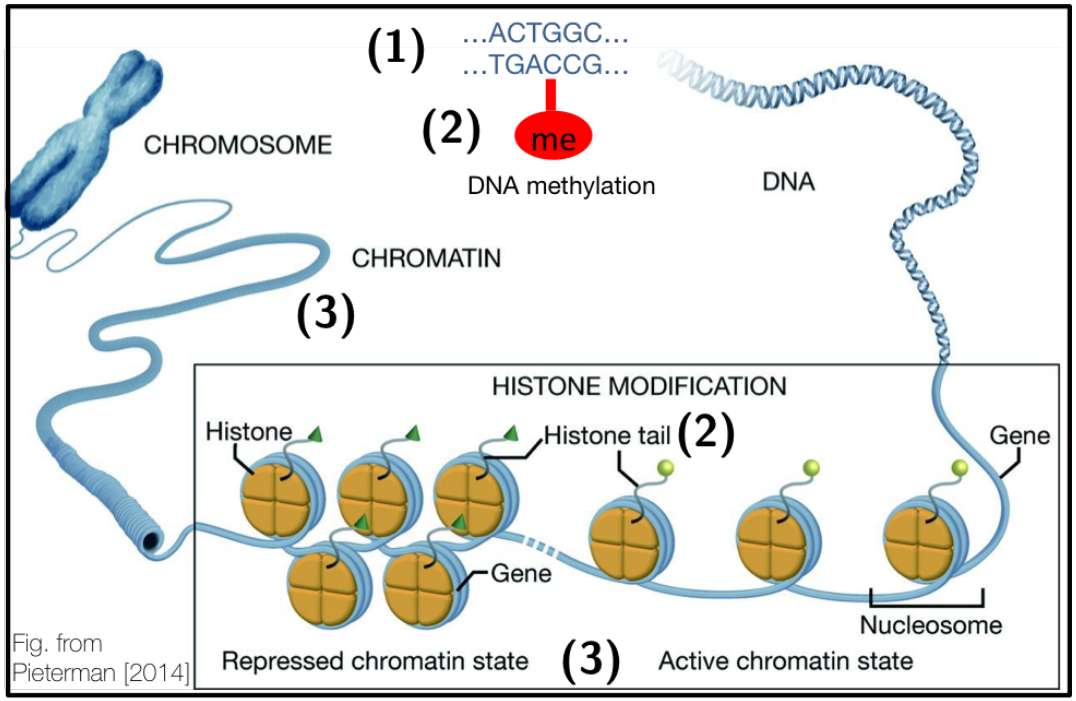


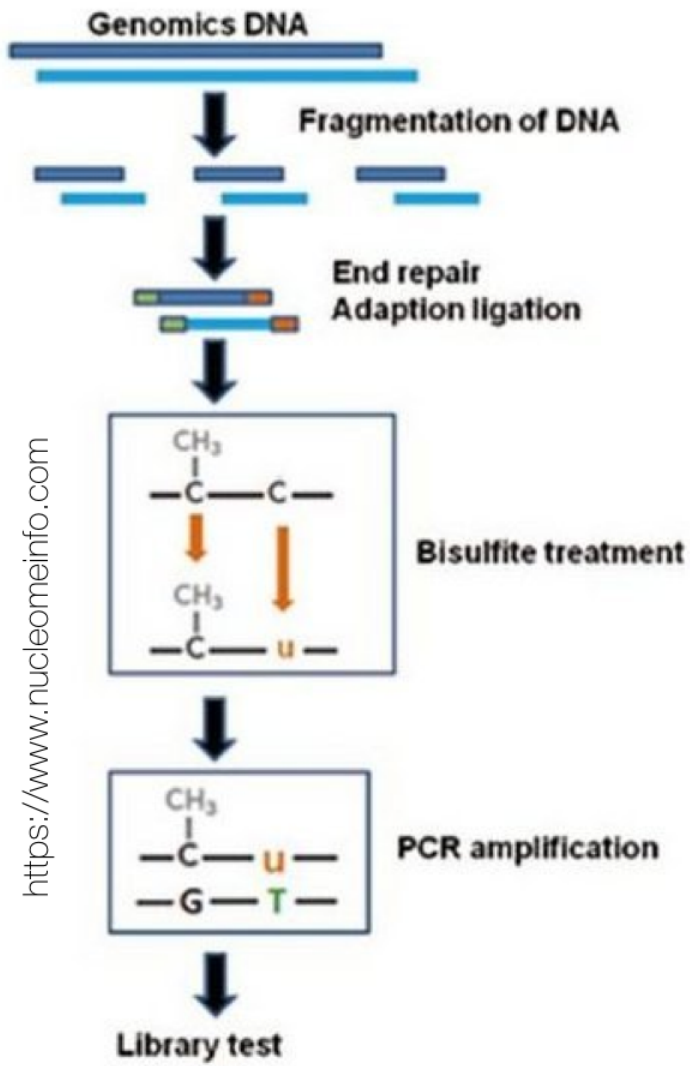
Fig. from Pieterman [2014]

(1) DNA sequence: genome assembly, variant detection
- whole genome (WGS), whole exome (WES), amplicons

(2) Interrogating **epigenetic marks**
- histone marks, TF binding: ChIP-seq
- DNA methylation: bisulfite sequencing

(3) Understanding the **chromatin structure**
- active/repressed: ATAC-, DNase, MNase-seq, ...
- 3D interactions: Hi-C, ChIA-PET

Understanding DNA: assessing genome-wide DNA methylation



DNA methylation is a true epigenetic mark as it has been shown to be inheritable.

Regions with high levels of methylated cytosines are generally considered to be transcriptionally repressed.

Understanding DNA: identifying protein-DNA interaction sites

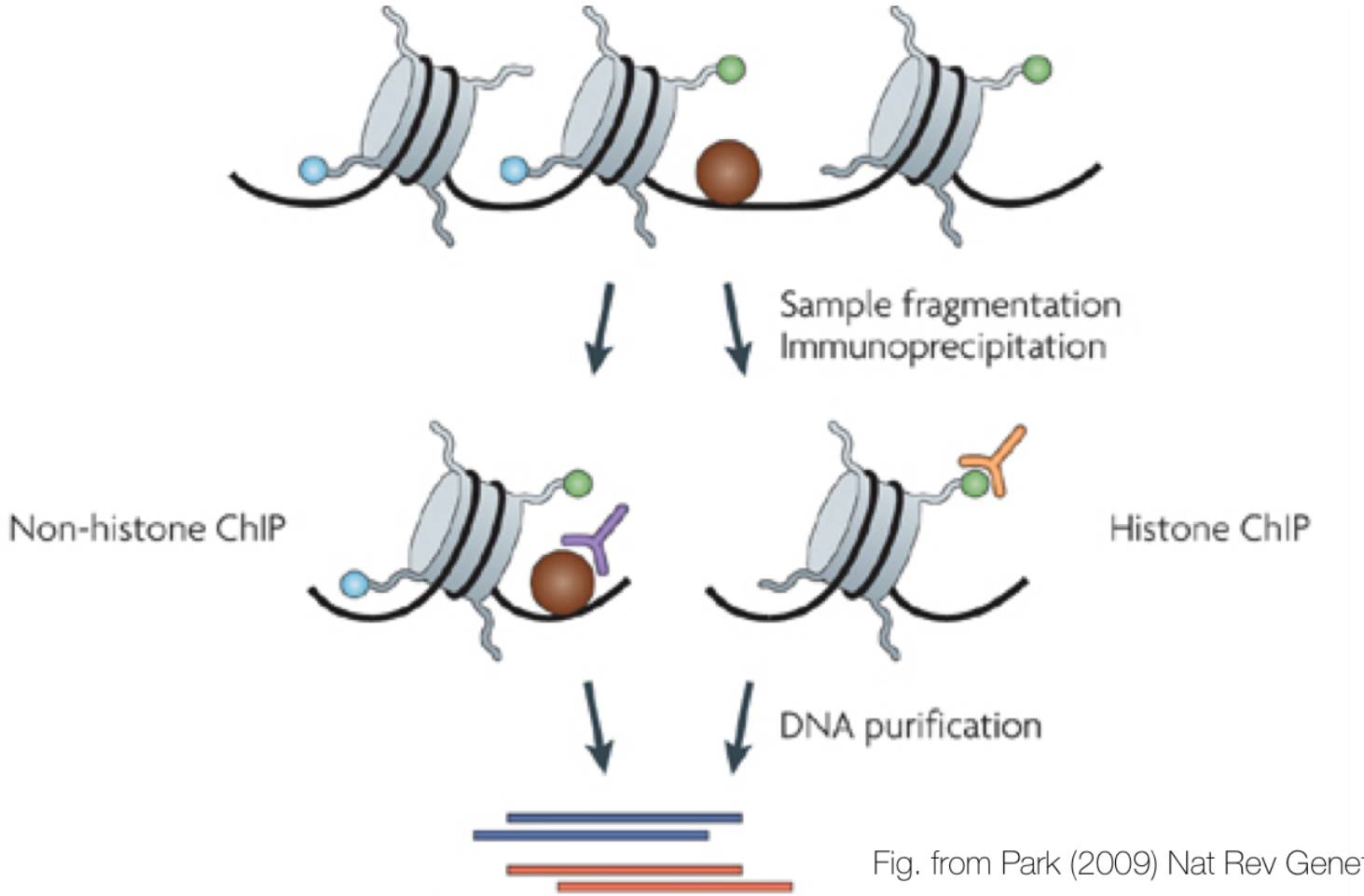
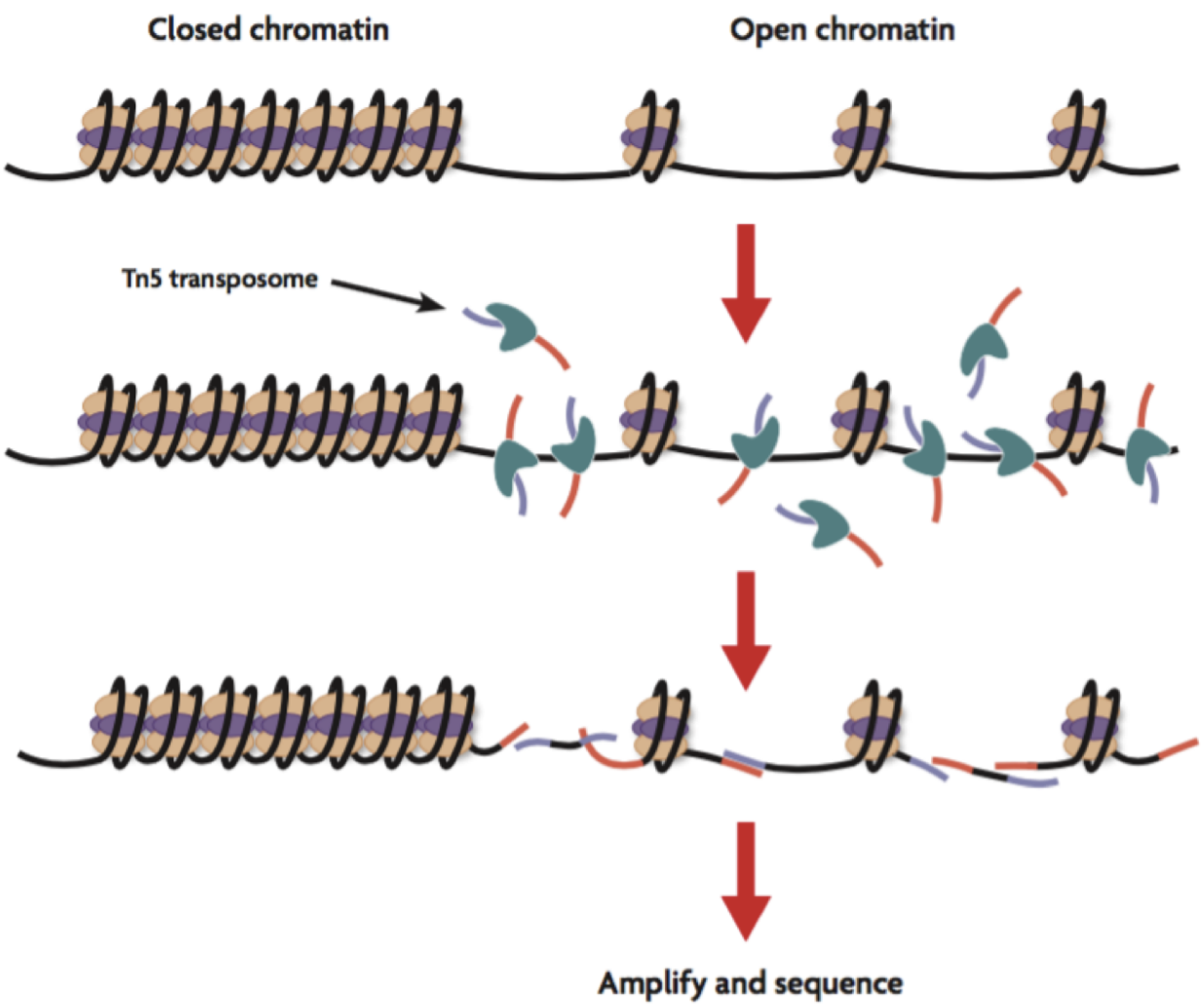


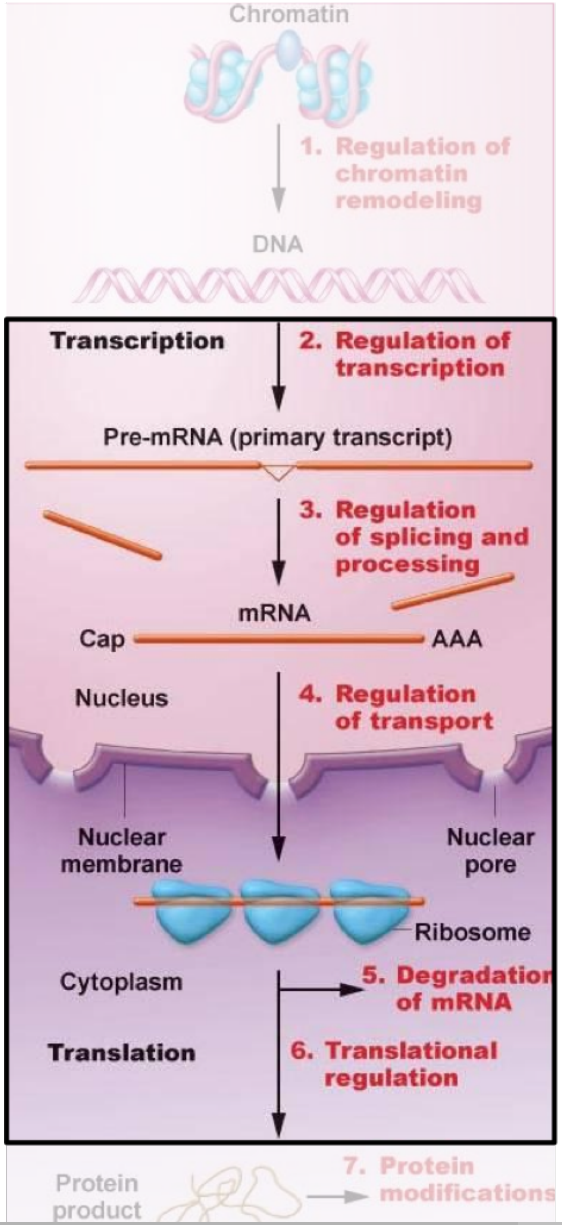
Fig. from Park (2009) Nat Rev Genetics

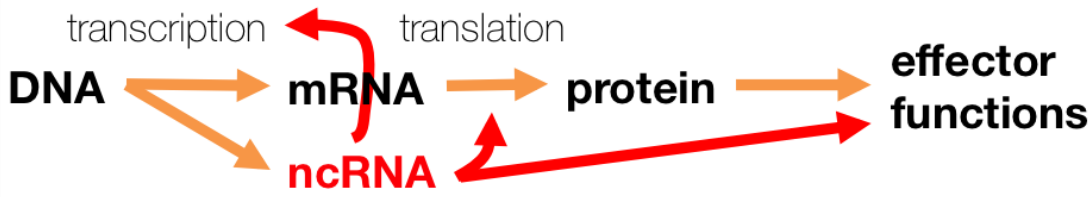
Understanding DNA: identifying regions with open chromatin (ATAC-seq)



https://www.activemotif.com/uploads/images/web_site/services-hp-dec2015/open-chromatin.png

Understanding RNA





(1) Gene expression: sequencing transcripts
- transcript identification & quantification (including non-coding transcripts): RNA-seq
- nascent transcripts: PRO-, GRO-seq

(2) Identifying **RNA-binding proteins**
- RIP-seq, CLIP-seq

(3) Determining **RNA structures**
- PARS, Structure-seq

(4) Assessing **RNA modifications**
- meRIP-seq, ICE (adenosine-inosine editing)

(5) Understanding **translation**
- Ribo-seq

Applications of NGS: RNA-seq is the most common one

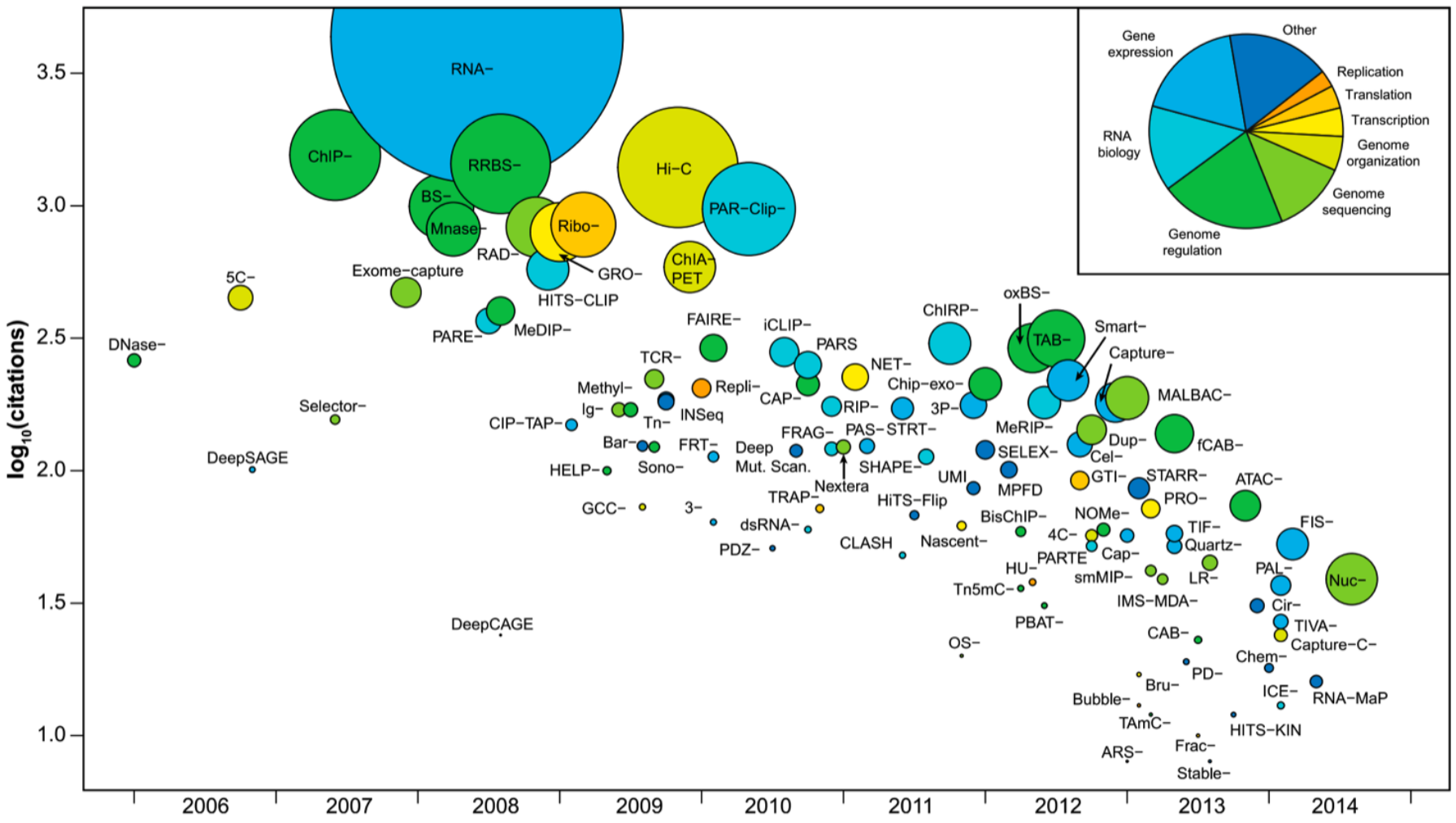


Fig. from Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). Molecular Cell, 58(4), 586–597.

Main steps of typical NGS experiments

TEMPLATE PREP

Obtaining the molecules of interest:
DNA, RNA, nucleotide-protein complexes

Library preparation:
fragmentation and ligation of sequencing adapters

Amplification

SEQUENCING

Sequencing by **Synthesis**
Sequencing by **Ligation**

short reads vs. long reads

BIOINFORMATICS

Base calling

Alignment
Identifying loci of the sequenced fragments

Additional processing

Interpretation

Experimental design

Where to sequence at WCM?

Genomics and Epigenomics Sequencing Services

- highly experienced staff
- nevertheless: know the issues you need to discuss with them!

Jenny Xiang, M.D.
Director of Genomics Services
WCM CLC Genomics and Epigenomics Core Facility
(212) 746-4258
jzx2002@med.cornell.edu

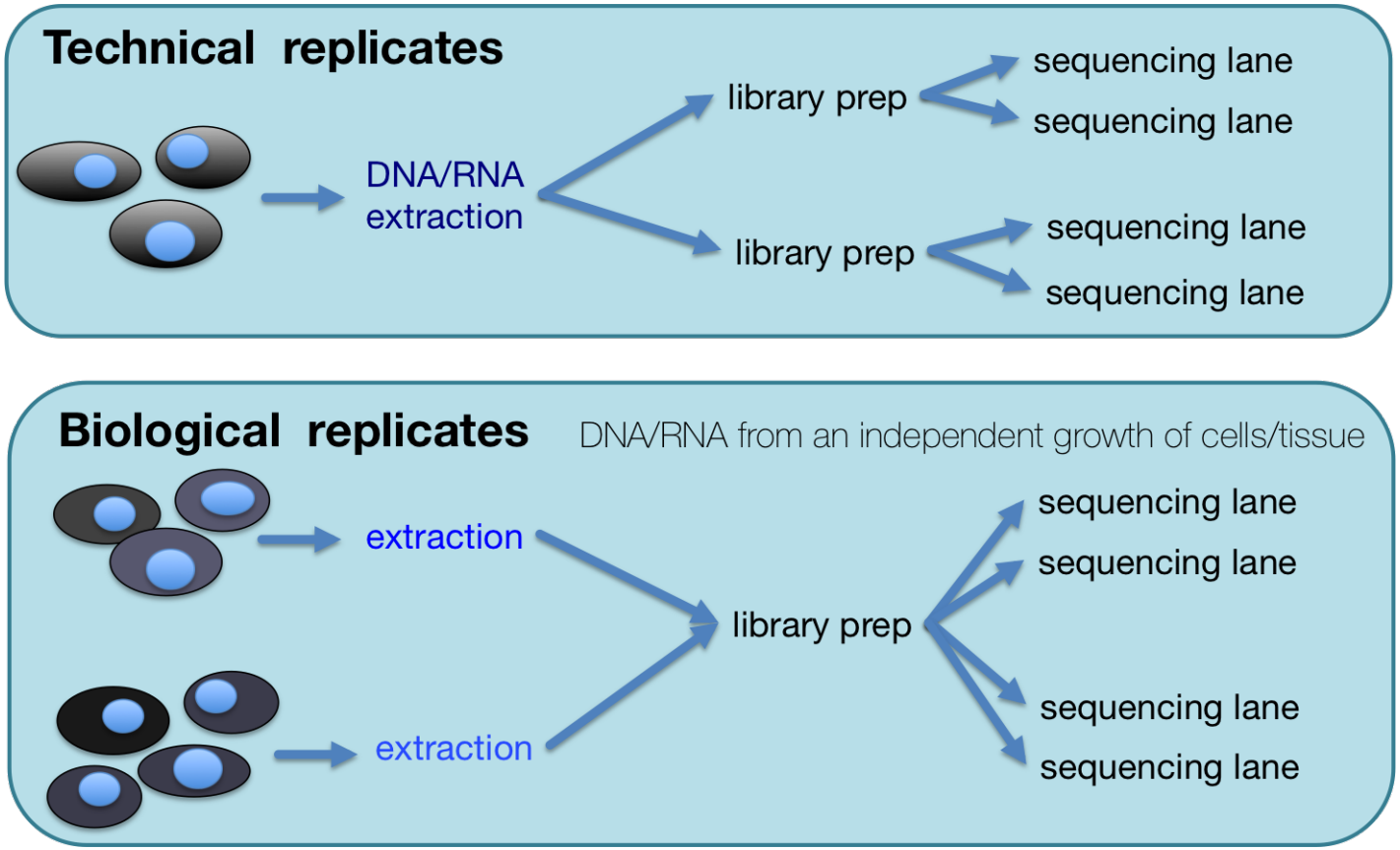
Alicia Alonso, Ph.D.
Director of Epigenomics Services
WCM CLC Genomics and Epigenomics Core Facility
(212) 746-3260
ala2035@med.cornell.edu

Experimental design considerations

- **How many replicates?**
- How to avoid batch effects?
- How many reads?

Why do we need replicates?

- replicates are needed to understand the **level of noise**



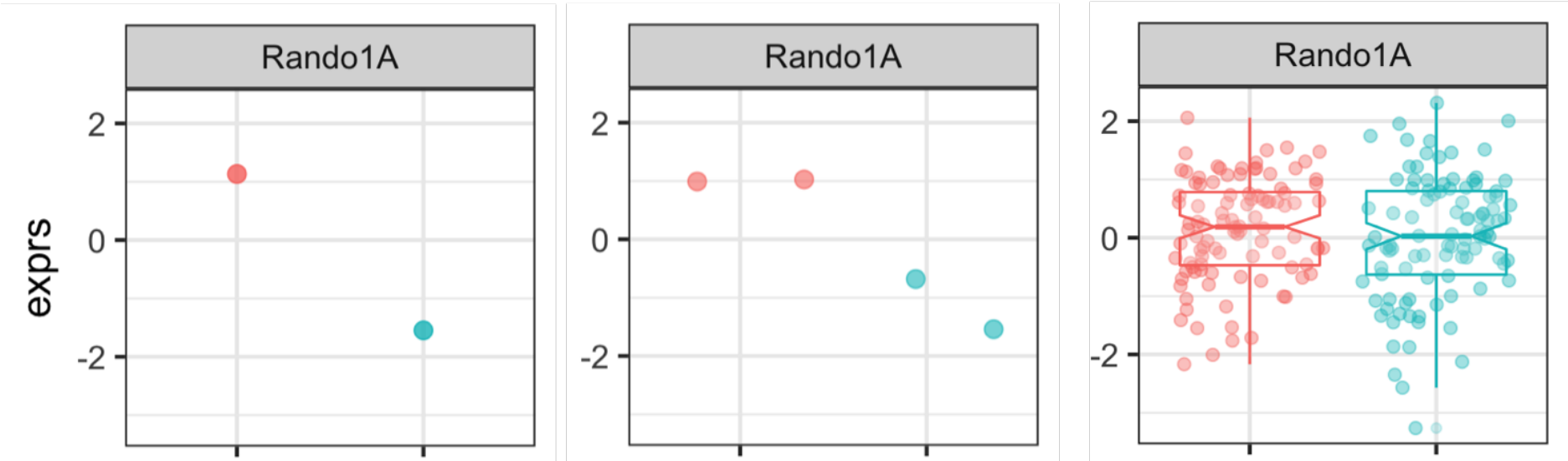
Cross-platform replicates sometimes may make sense, too.

Why do we need replicates?

- definitely needed for quantitative assessments, e.g. RNA-seq for determining expression level differences, but qualitative approaches such as variant calling also benefit from technical replicates [Robasky et al., 2014], [Derryberry et al., 2016]

> *"Samples are our windows to the population, and their statistics are used to estimate those of the population."*
> Martin Krzywinski & Naomi Altman

```
testdat <- data.frame(exprs = rnorm(200),
                      condition = c("WT","MUT"),
                      gene_name = "Rando1A")
```



Experimental design considerations

- How many replicates?
- **How to avoid batch effects?**
  - Understanding typical sources of noise and artifacts
- How many reads?

General problems for NGS

Problems = sources of technical noise

Sample preparation

- DNA/RNA extraction with varying degrees of degradation
- contaminations
- mislabelling, mishandling

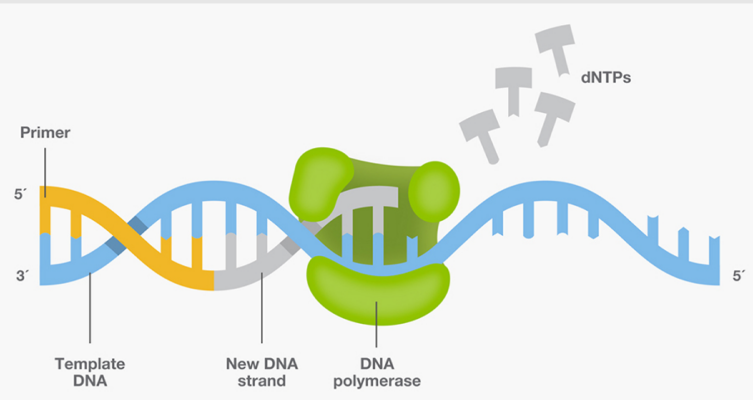
Biases of Illumina-based DNA sequencing

Somewhat **sequencing-machine**-specific problems

- sequencing errors
- miscalled bases

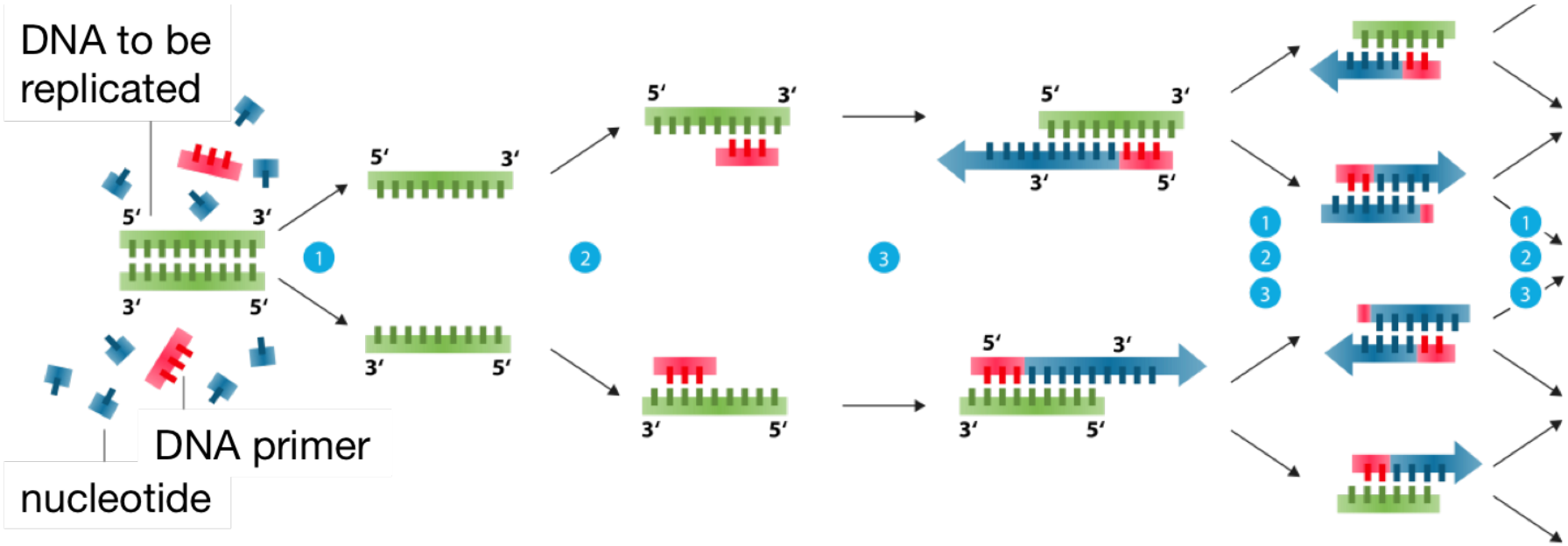
Sample-specific problems: **PCR artifacts**

- duplicated fragments (low library complexity)
- GC bias: fragments with moderate GC content are preferably amplified
- length bias: fragments between 250-700bp are strongly favored



https://www.thermofisher.com/uk/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/pcr-education/pcr-reagents-enzymes/pcr-basics/_jcr_content/

The most important biochemical assay for NGS: PCR



(1) Denaturation at 94-96C
(2) Annealing at ca. 68C
(3) Elongation at ca. 72C

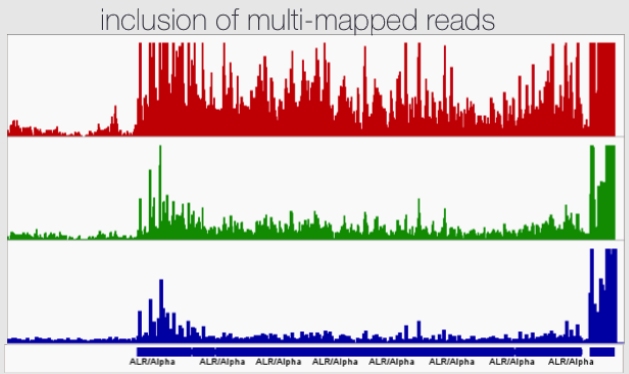
https://laboratoryinfo.com/wp-content/uploads/2015/07/Polymerase_chain_reaction.svg_.png

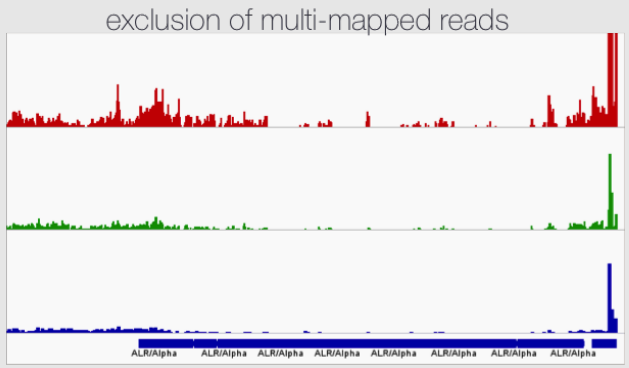
For NGS applications, template DNA fragments vary in size and GC content!
Due to the exponential nature of the amplification process, small differences in the starting population can lead to strongly skewed final populations.
Always keep the number of PCR cycles to an absolute minimum!

Biases of Illumina-based DNA sequencing

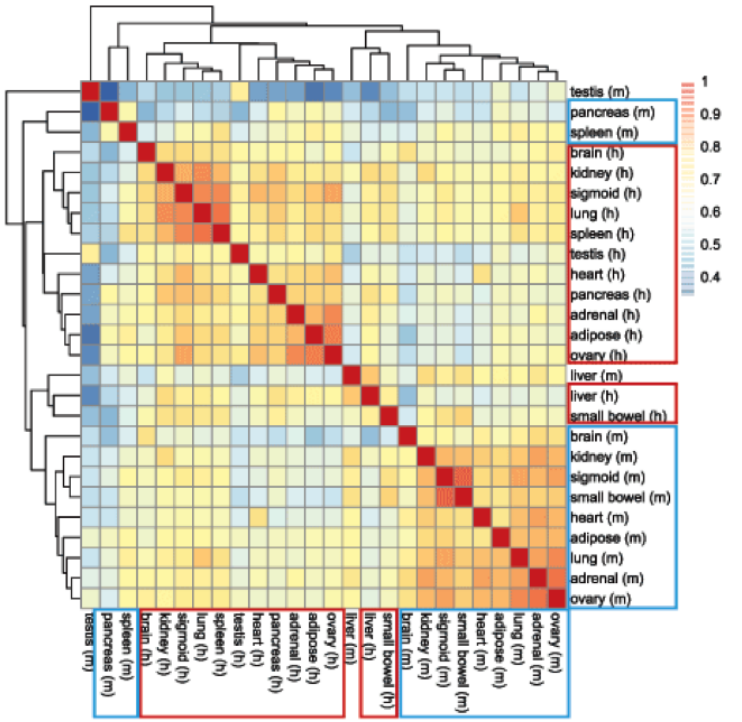
Bioinformatics problems

- **DNA**: long, repetitive elements are difficult to align to with short reads ("mappability" issue)
  - abundance of (structural) variants may complicate alignments
- **RNA**: great dynamic range (lowly expressed to extremely abundant)
  - saturation point is hardly reached: number of distinct transcripts depends on the overall make-up of the library
  - strongly affected by contaminations (DNA, rRNA, ...)
- inappropriate **data processing**, e.g. wrong parameter choices





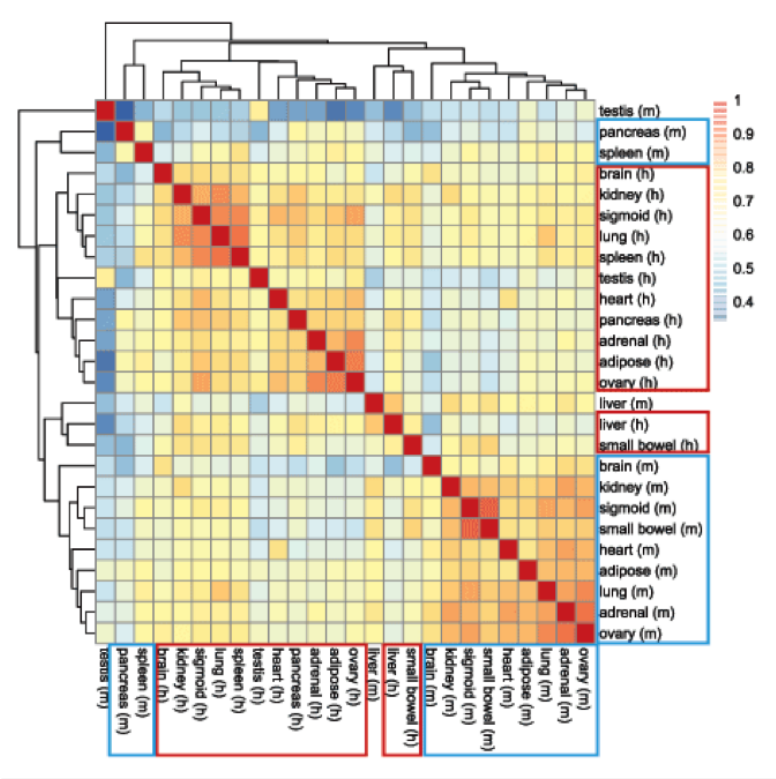
Case study: ENCODE's comparison of mouse and human tissues



"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. "*

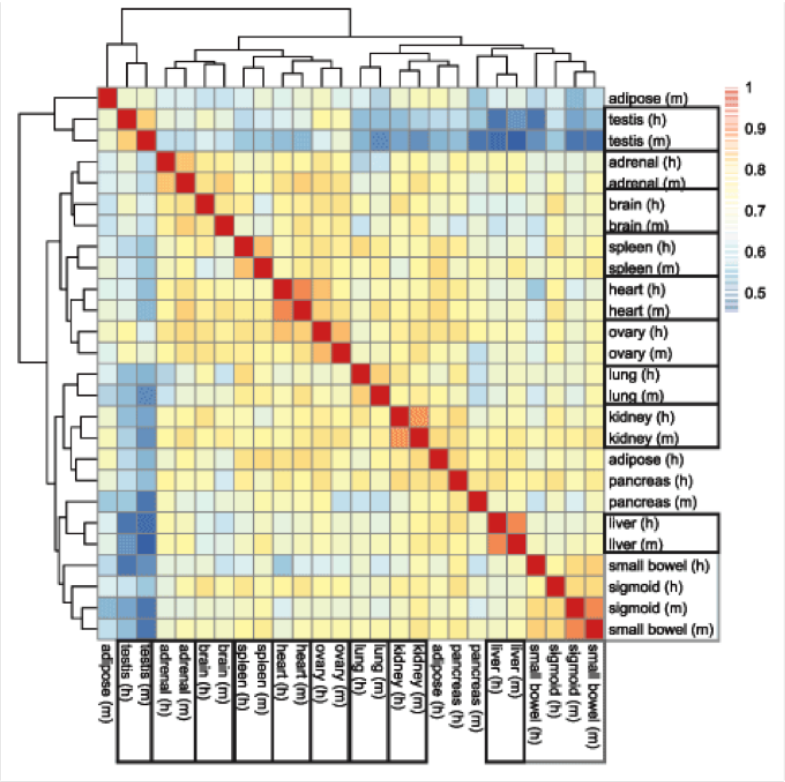
Lin, Lin, and Snyder (2014). PNAS 111:48

Case study: ENCODE's comparison of mouse and human tissues



"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

Suboptimal study design

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

- human data: deceased organ donors
- mouse data: 10-week-old littermates

Not all variables can be controlled for! Know the limitations of your study before making bold claims! Recommended reading: `https://f1000research.com/articles/4-121/v1`

Avoiding bias by relying on randomization

Completely randomized design

STRESS	A	B	A	A	B	A	B	A	A	B	B	B
DIET	1	2	1	2	2	1	1	2	2	1	2	1

Restricted randomized design

GENOTYPE	A	A	A	A	A	A	B	B	B	B	B	B
DIET	1	2	1	2	2	1	1	2	1	1	2	2

Blocked & randomized design

GENOTYPE	A	A	B	B	A	A	B	B	A	A	B	B
DIET	1	2	1	2	1	2	1	2	1	2	1	2
WEIGHT	•	•	•	•	●	●	●	●	⬤	⬤	⬤	⬤

Block what you can,
randomize what you cannot.

*What factors are of **interest**? Which ones might introduce noise? Which nuisance factors do you absolutely need to account for?*

Experimental design considerations

- How many replicates?
- How to avoid batch effects?
- **How many reads?**

How deep is deep enough?

lower limit should usually be whatever ENCODE says:
https://www.encodeproject.org/about/experiment-guidelines/

Application	Recommended seq. depth
differential gene expression	20 - 50 mio SR, 75 bp
variant calling	30-200x coverage
whole-genome bisulfite sequencing	30x coverage
ChIA-PET	200 mio PE

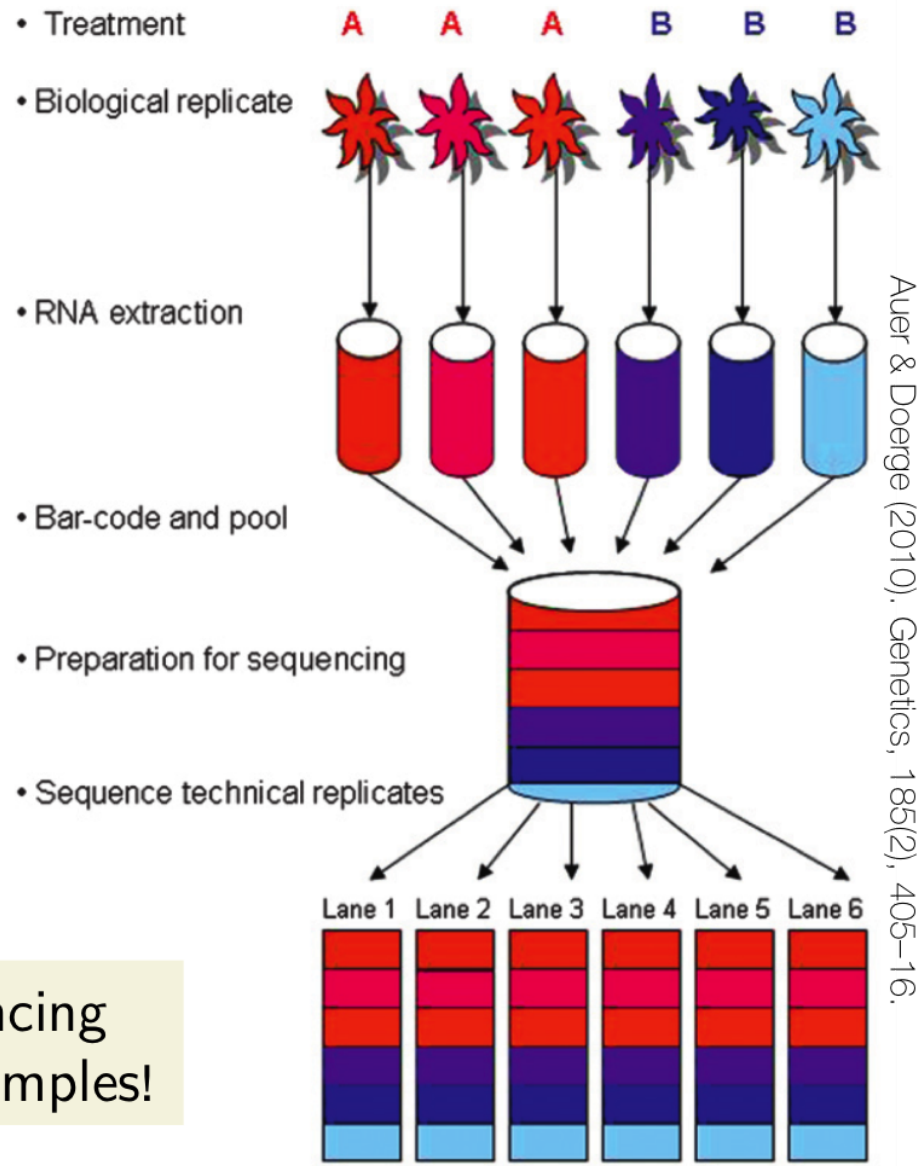
- you may need more, longer, and possibly paired-end reads
  - novel transcript identification
  - alternative splicing
  - ChIP-seq for broad histone marks
  - 3D chromatin structure assessment assays

Sometimes the addition of replicates is more meaningful than increased sequencing depth!

Typical experimental setup

- keep the **technical nuisance** factors (harvest date, RNA extraction kit, sequencing date...) to a minimum
- cover only as much of the **biological variation** as needed (but keep possible limitations for the final conclusions in mind)

Make sure the sequencing core **multiplexes** all samples!



Auer & Doerge (2010). Genetics, 185(2), 405–16.

References

[Auer and Doerge, 2010, Krzywinski and Altman, 2013, Gilad and Mizrahi-Man, 2015, Lin et al., 2014, Park, 2009, Pieterman et al., 2014, Reuter et al., 2015, Stirzaker et al., 2014]

References

Paul L Auer and R W Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–16, Jun 2010. ISSN 1943-2631. doi: 10.1534/genetics.110.114983.

Dakota Z. Derryberry, Matthew C. Cowperthwaite, and Claus O. Wilke. Reproducibility of SNV-calling in multiple sequencing runs from single tumors. *PeerJ*, 2016. doi: 10.7717/peerj.1508.

Yoav Gilad and Orna Mizrahi-Man. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research*, 2015. doi: 10.12688/f1000research.6536.1.

Sara Goodwin, John D Mcpherson, and W Richard Mccombie. Coming of age : ten years of next- generation sequencing technologies. *Nature Genetics*, 17(6):333–351, 2016. doi: 10.1038/nrg.2016.49.

Martin Krzywinski and Naomi Altman. Points of significance: Importance of being uncertain. *Nature Methods*, 10(9):809–810, 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2613. URL http://www.nature.com/doifinder/10.1038/nmeth.2613.

Shin Lin, Yiing Lin, Joseph R. Nery, Mark A. Urich, Alessandra Breschi, Carrie A. Davis, Alexander Dobin, Christopher Zaleski, Michael A. Beer, William C. Chapman, Thomas R. Gingeras, Joseph R. Ecker, and Michael P. Snyder. Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 2014. doi: 10.1073/pnas.1413624111.

Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–80, Oct 2009. doi: 10.1038/nrg2641.

C. R.C. Pieterman, E. B. Conemans, K. M.A. Dreijerink, J. M. De Laat, H. Th M. Timmers, M. R. Vriens, and G. D. Valk. Thoracic and duodenopancreatic neuroendocrine tumors in multiple endocrine neoplasia type 1: Natural history and function of menin in tumorigenesis. *Endocrine-Related Cancer*, 2014. doi: 10.1530/ERC-13-0482.

Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4): 586–597, May 2015. doi: 10.1016/j.molcel.2015.05.004.

Kimberly Robasky, Nathan E. Lewis, and George M. Church. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 2014. doi: 10.1038/nrg3655.

Clare Stirzaker, Phillippa C. Taberlay, Aaron L. Statham, and Susan J. Clark. Mining cancer methylomes: Prospects and challenges. 2014. doi: 10.1016/j.tig.2013.11.004.