Introduction to functional genomics

Pascal Martin


Pascal.Martin@inra.fr

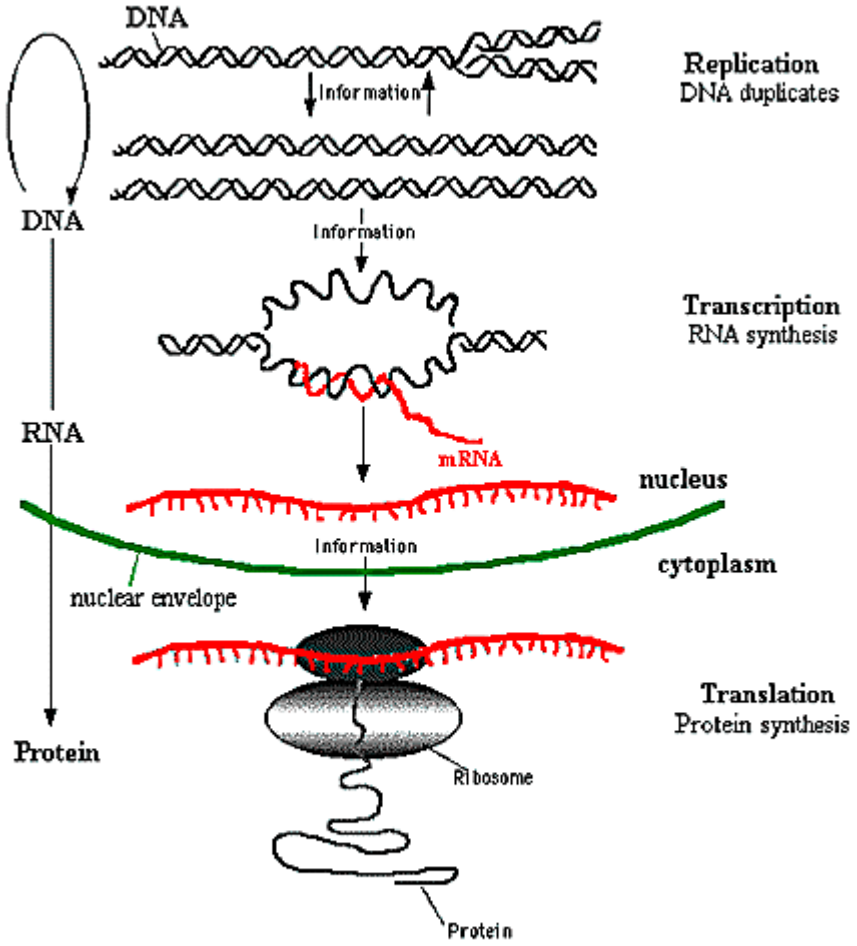

@PgpMartin

- **Introduction**
  - Genes, RNAs & transcriptomes
  - Transcription and the epigenome
  - Why measuring gene expression ?

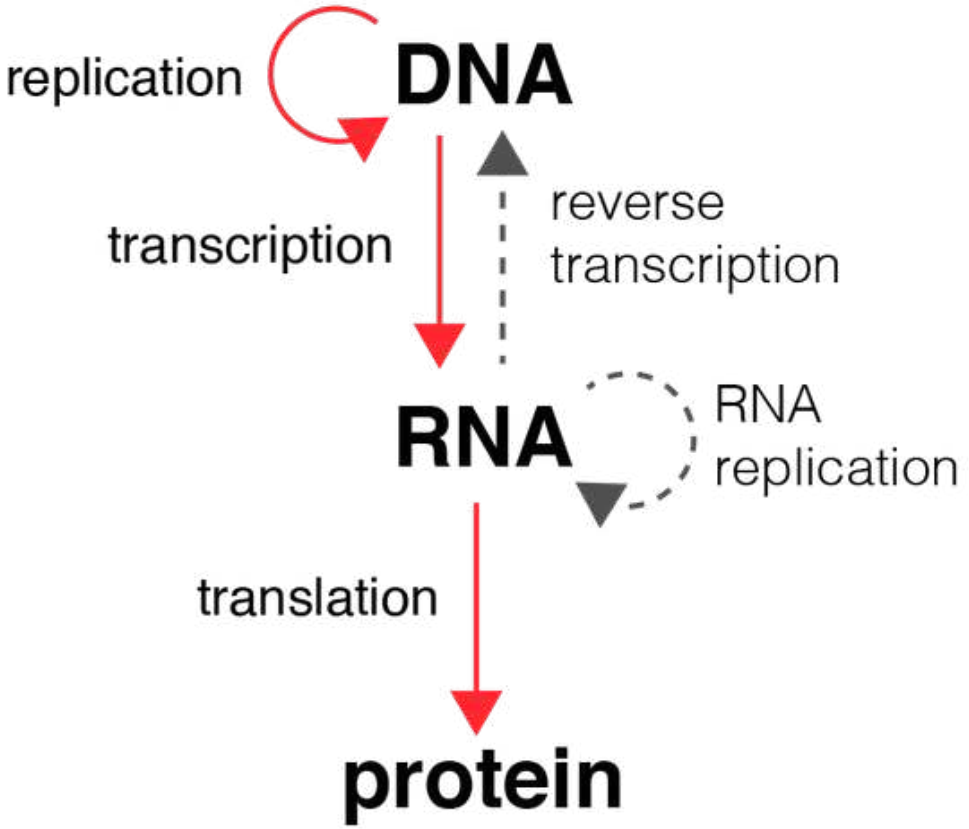
- **Transcriptomics**
  - low throughput methods (Northen blot, RPA, qPCR)
  - microarrays
  - NGS / RNA-seq

- **Epigenomics**
  - ChIP-seq
  - Chromatin accessibility (MNase, DNase, FAIRE, ATAC)
  - Chromatin conformation

The Central Dogma of Molecular Biology



The Central Dogma of Molecular Biology



What is a gene ?

Mendel, 1865
Morgan, 1910
Beadle & Tatum : "One gene, one enzyme", 1941
Avery, McLeod, McCarty, 1944
Watson & Crick, Wilkins & Franklin, 1953
Crick "Central dogma", 1958
Jacob & Monod, lac operon, 1961
Genetic code, 1965
RNA splicing, 1967
GENSCAN gene prediction, 1997
Draft of the human genome, 2001
ENCODE 1st phase, 2003-07

Timeline

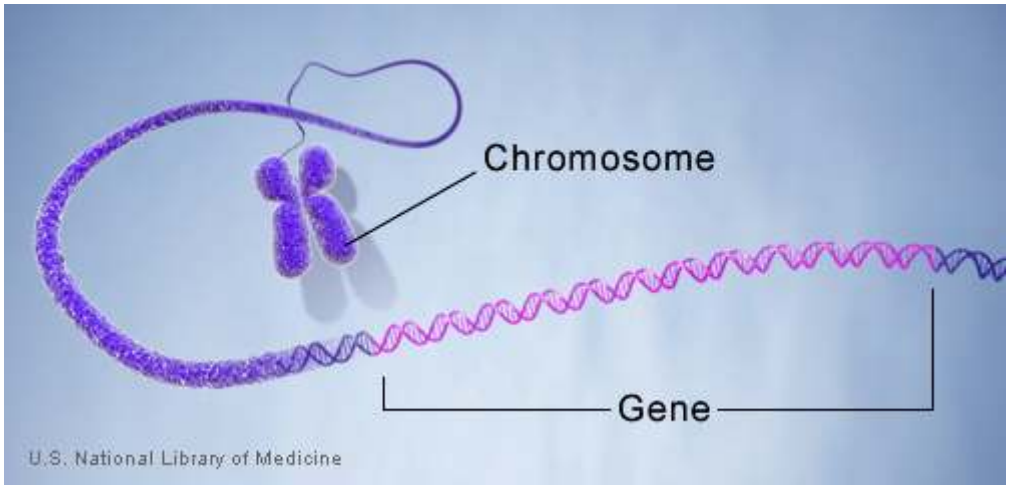
- Discrete Heredity Unit
- Distinct Locus (linear organization, "beads on a string")
- A protein blueprint
- A physical molecule
- A transcribed code
- ORF sequence pattern
- Annotated genomic entity

	U	(RNA only)
C	T	pyrimidines
‖‖	‖	
G	A	purines

Post-ENCODE definition:

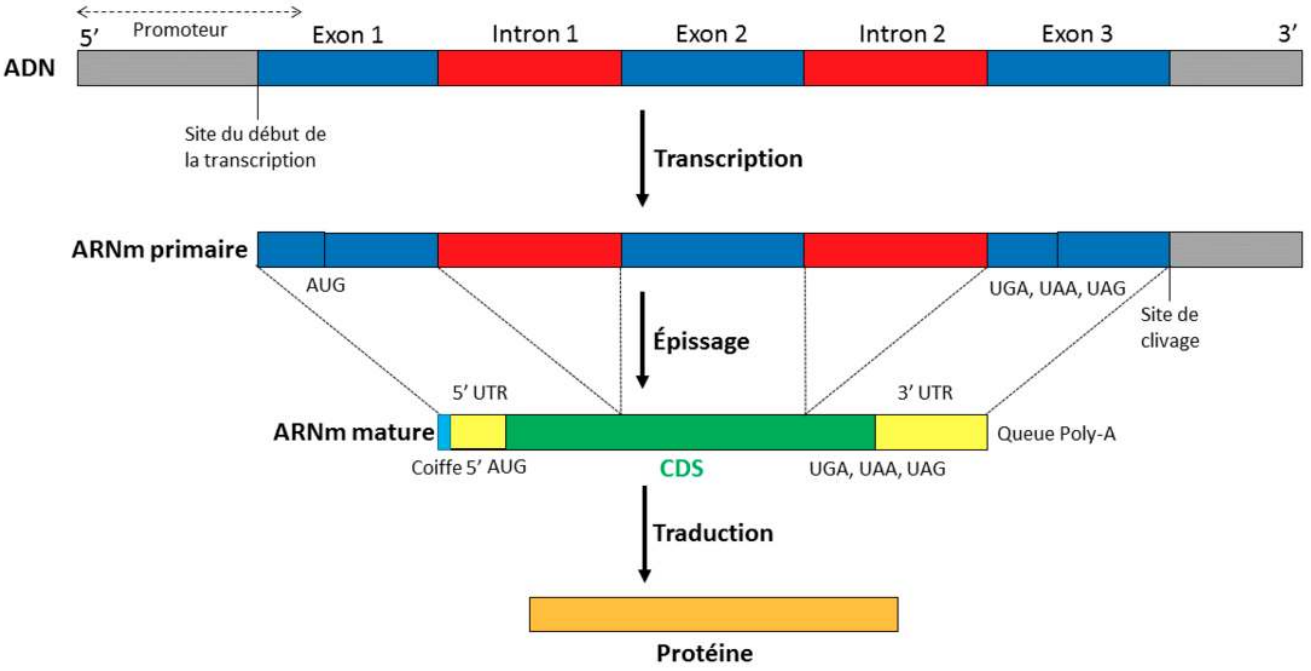
"A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products"

Gerstein et al., Genome Res 2007

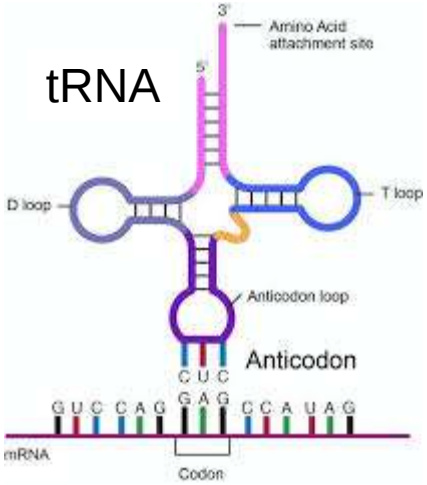


"A basic physical and functional unit of heredity"

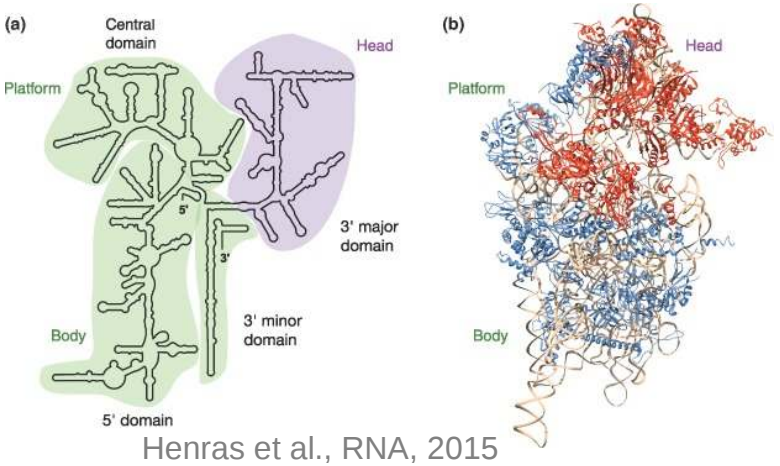
Many genes encode proteins



but many don't

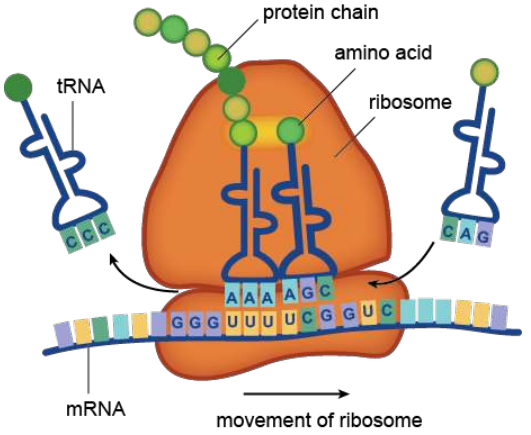


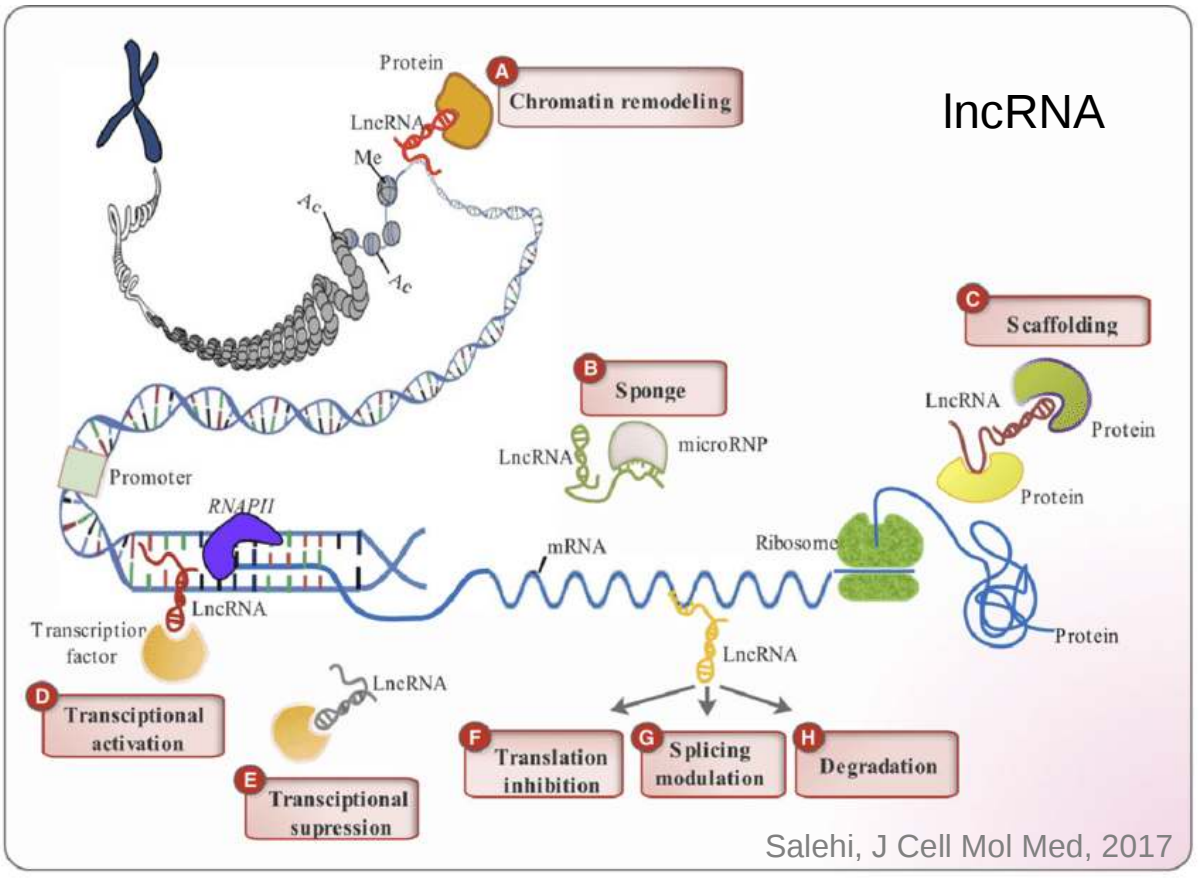
18S rRNA



Henras et al., RNA, 2015

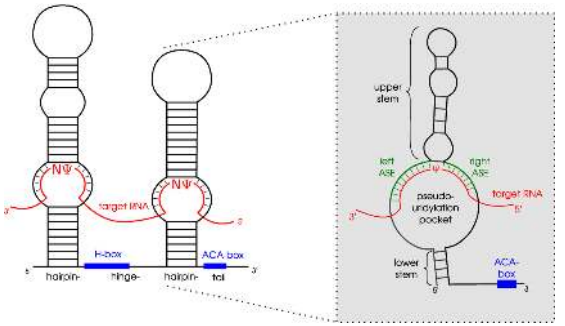
Translation



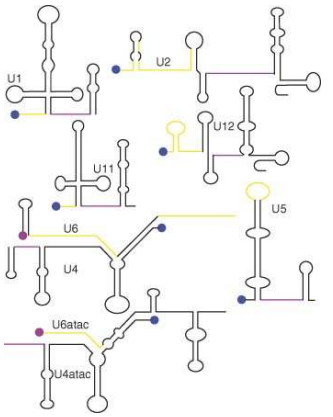


Salehi, J Cell Mol Med, 2017

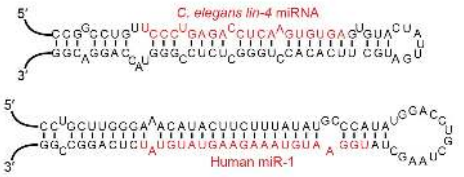
snoRNA/scaRNA (RNA modification)



snRNA (splicing)

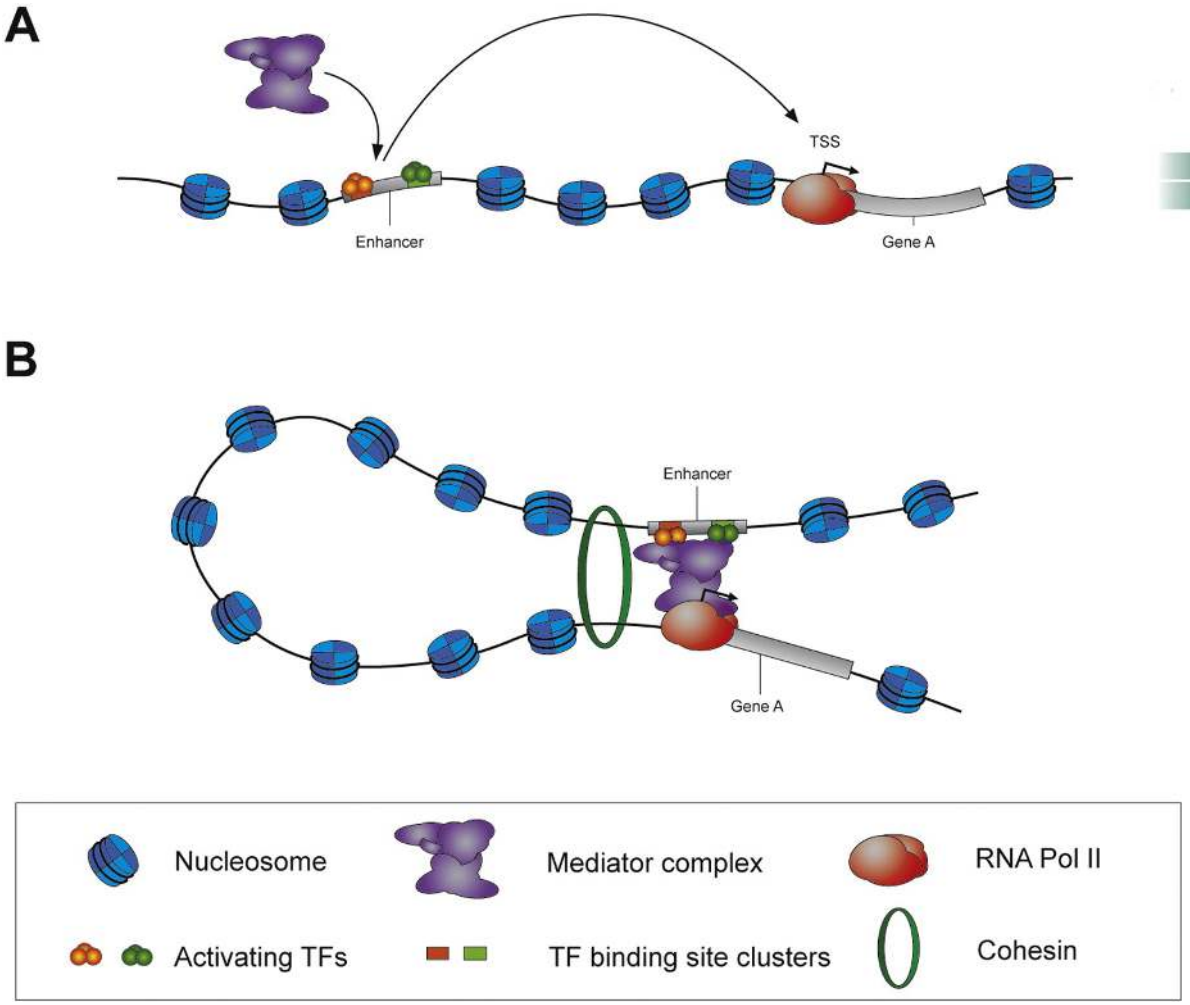


miRNA (regulation)

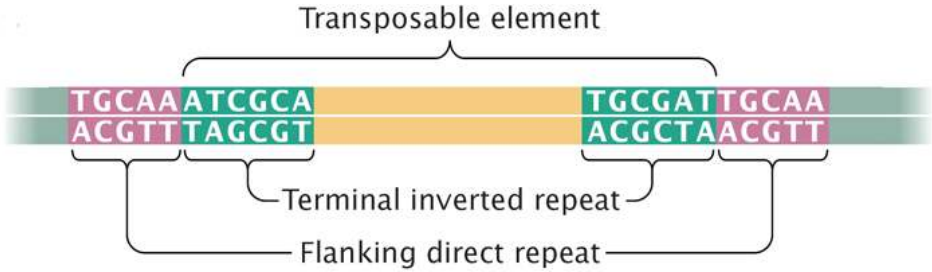


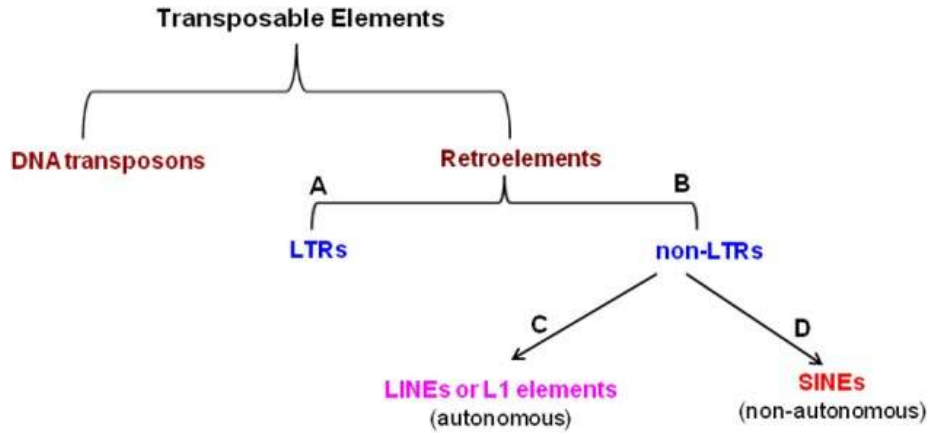
Other transcribed regions

Enhancers



Transposable elements / viruses

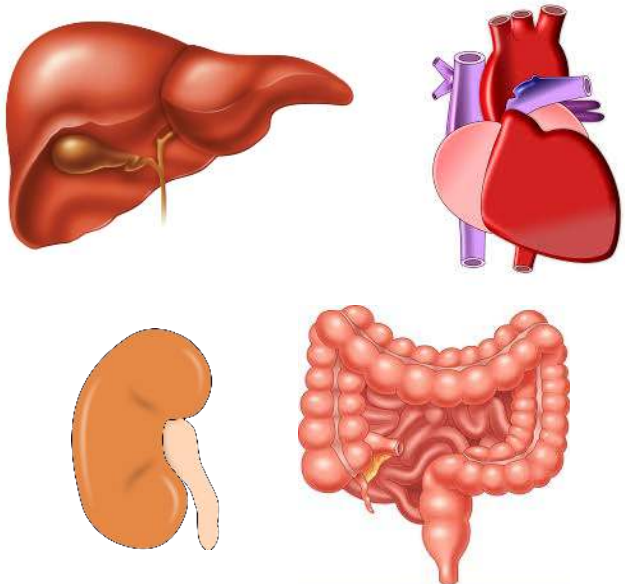




Marand et al., BBA, 2017

Analyzing the transcriptomeS

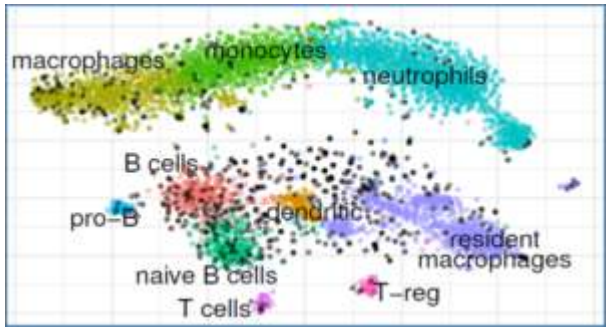
Different tissues / conditions



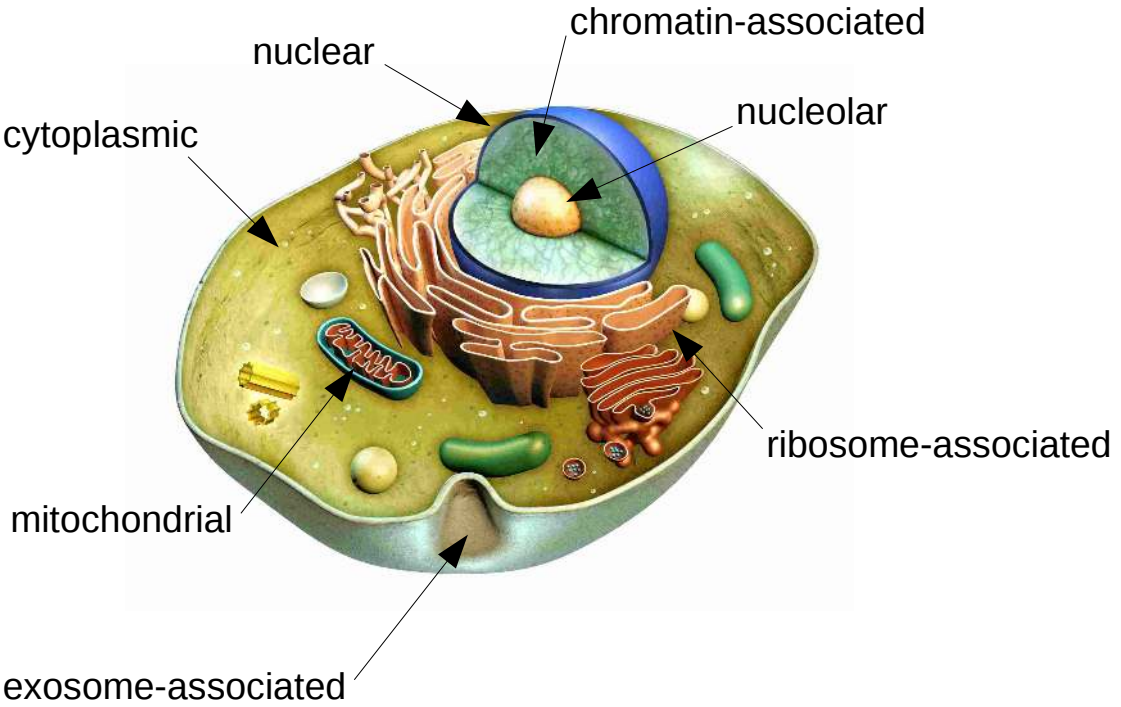
Different types of RNA

- PolyA+/-
- capped / uncapped
- short / long

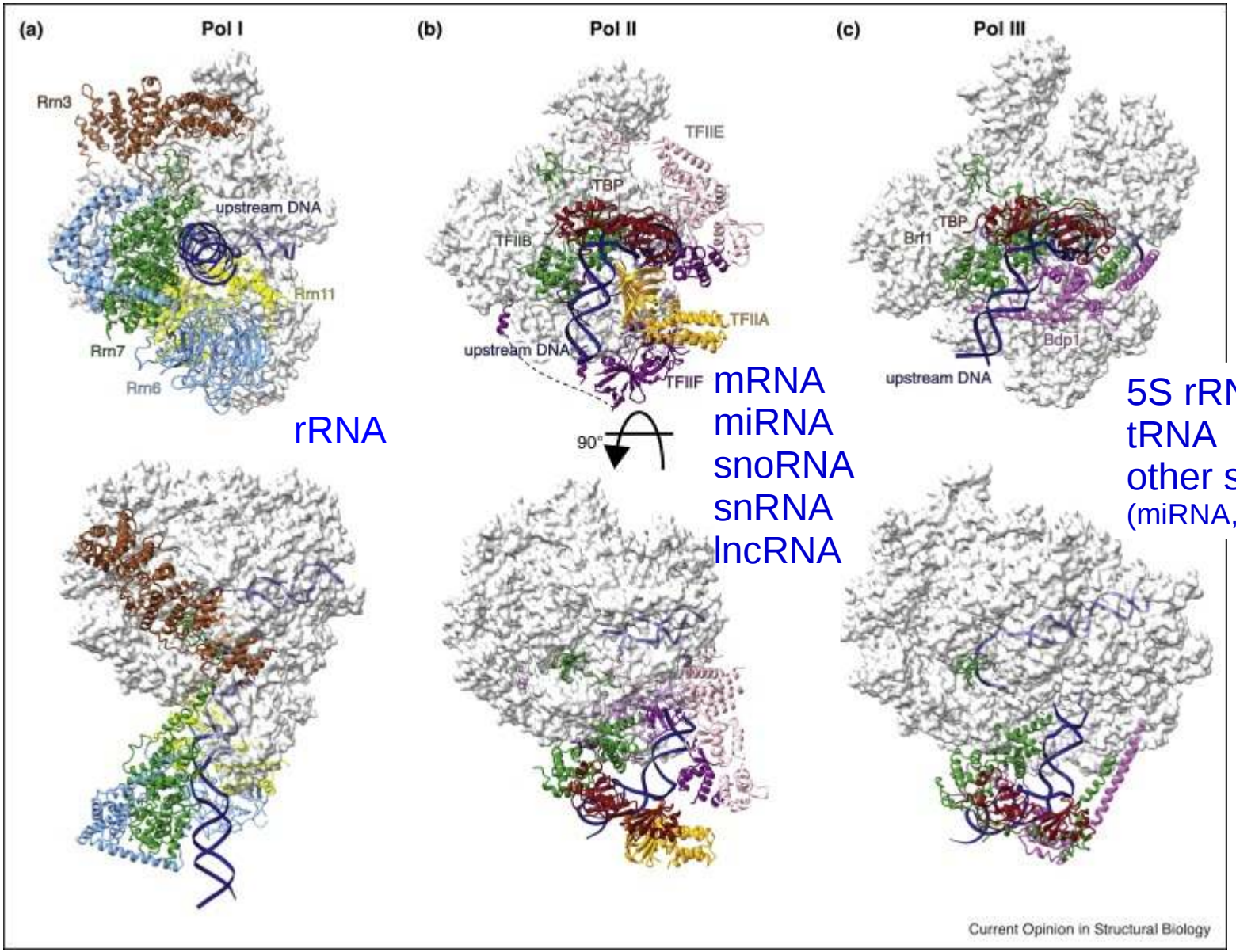
Single-cell transcriptomes



Subcellular fractions



Transcribing DNA



rRNA

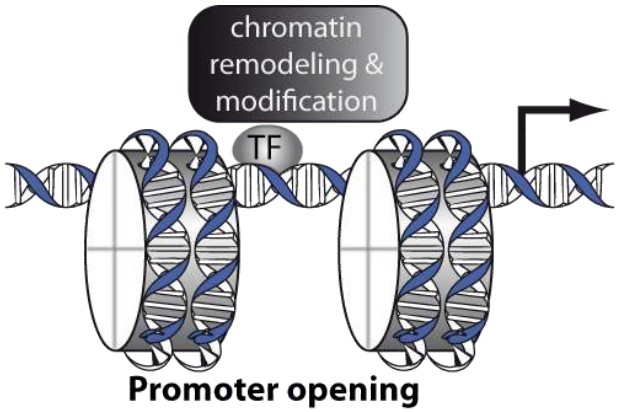
mRNA
miRNA
snoRNA
snRNA
lncRNA

5S rRNA
tRNA
other small RNA
(miRNA, snRNA, 7SK...)

Hanske, Sadian, Müller, Curr Opin Struct Biol, 2018

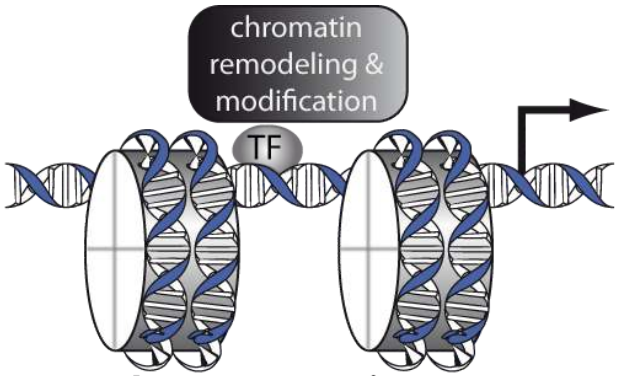
RNA polymerase II transcription

Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008

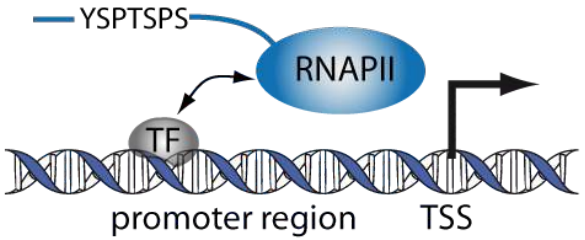


RNA polymerase II transcription

Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008



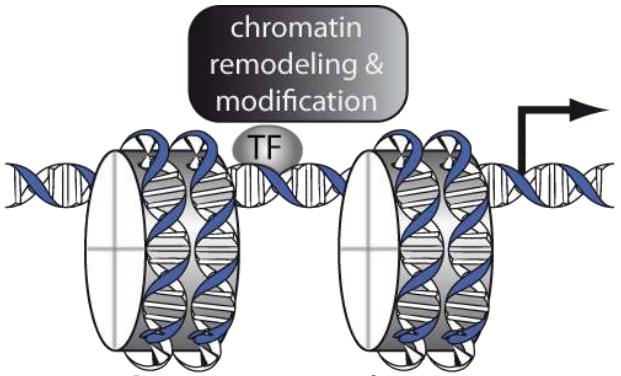
Promoter opening



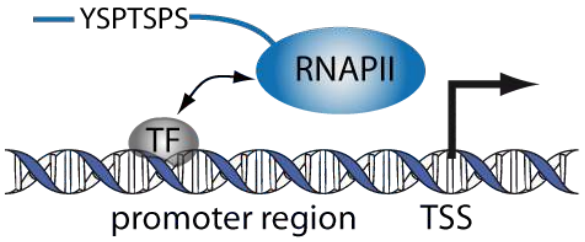
RNA PolII recruitment

RNA polymerase II transcription

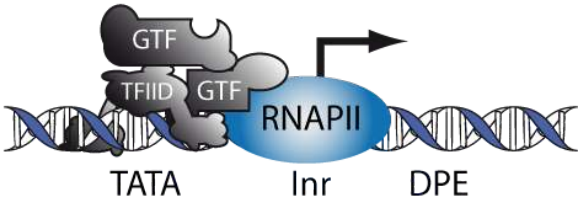
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



Promoter opening



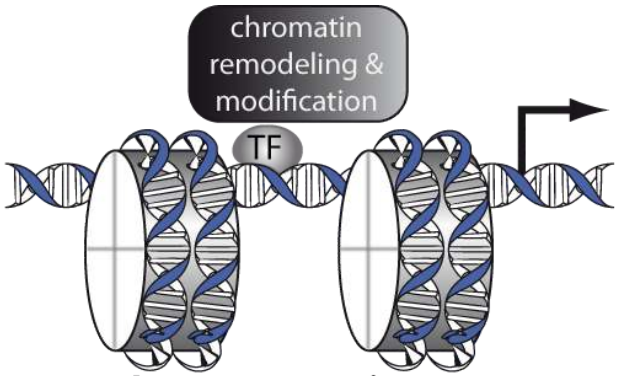
RNA PolII recruitment



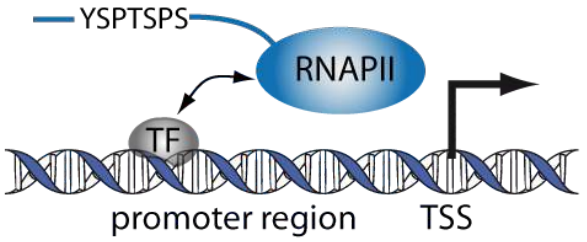
Preinitiation complex formation

RNA polymerase II transcription

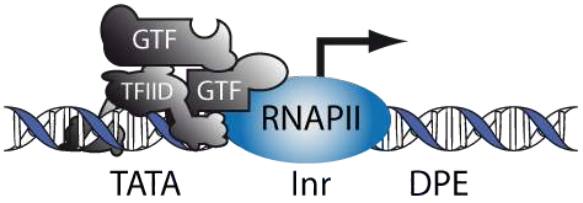
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



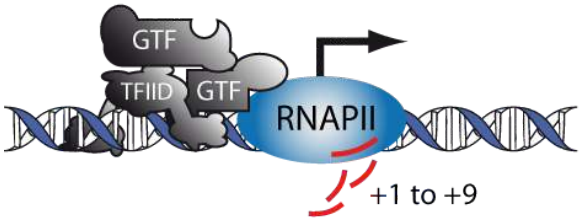
Promoter opening



RNA PolII recruitment



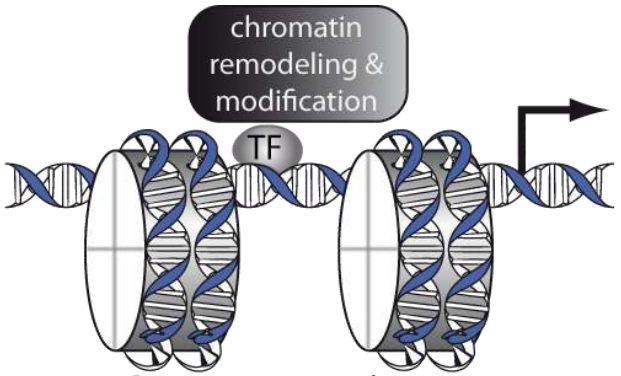
Preinitiation complex formation



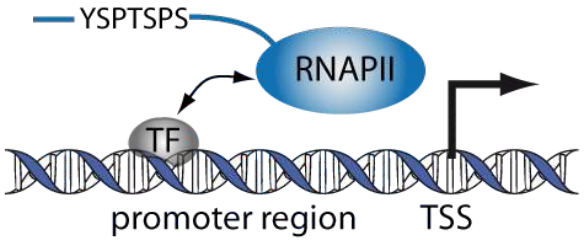
Abortive initiation

RNA polymerase II transcription

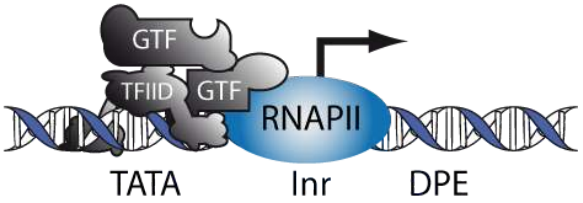
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



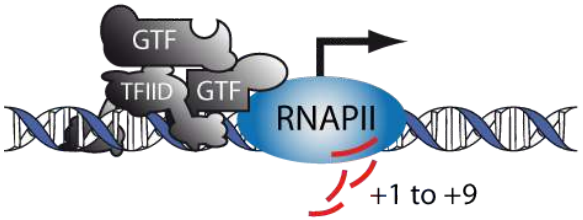
Promoter opening



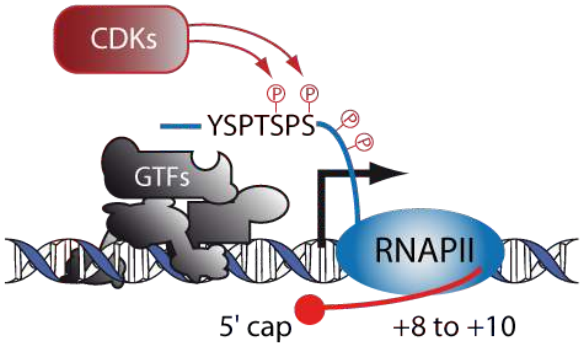
RNA PolII recruitment



Preinitiation complex formation



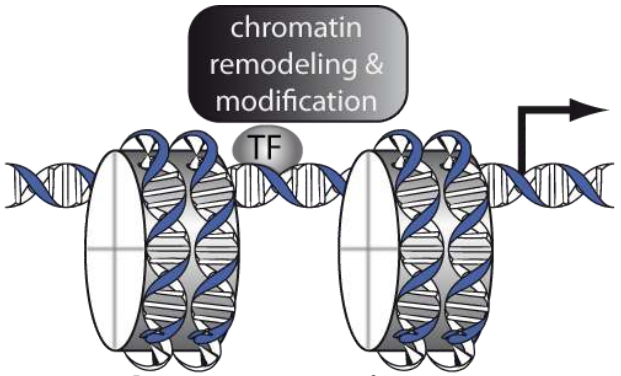
Abortive initiation



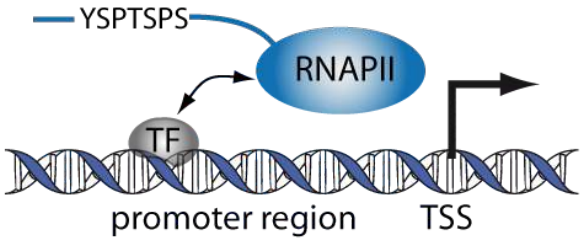
Promoter escape

RNA polymerase II transcription

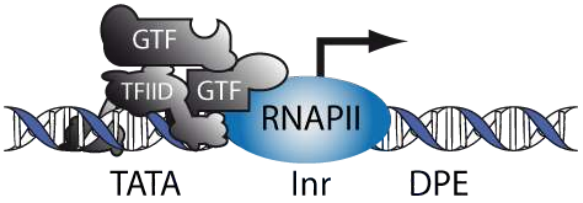
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



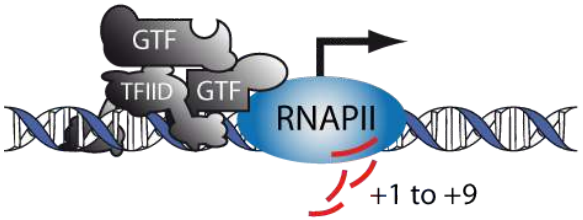
Promoter opening



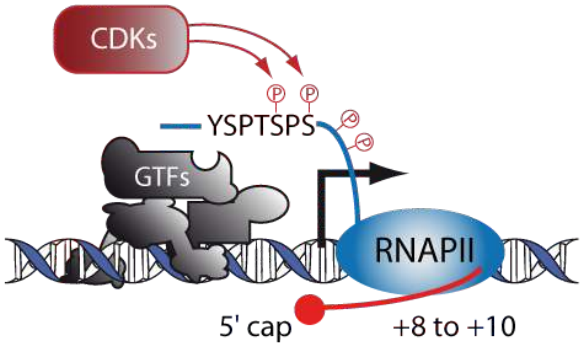
RNA PolII recruitment



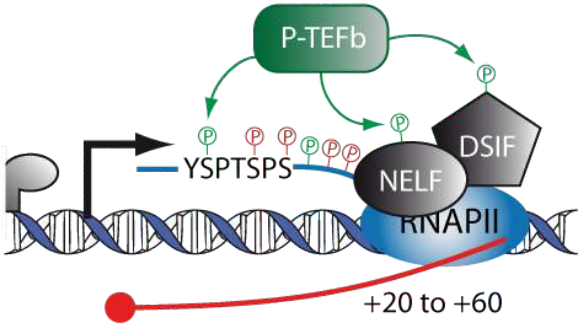
Preinitiation complex formation



Abortive initiation



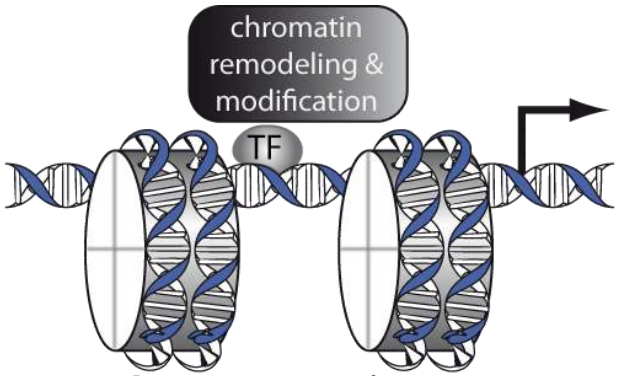
Promoter escape



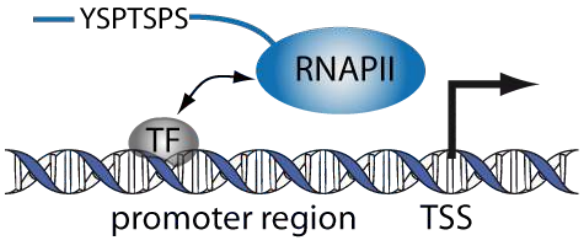
Promoter-proximal pausing

RNA polymerase II transcription

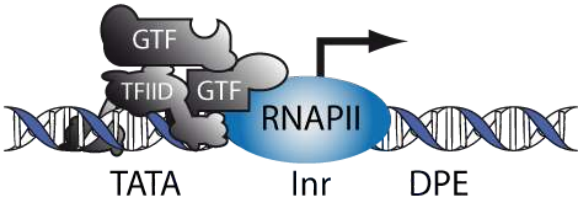
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



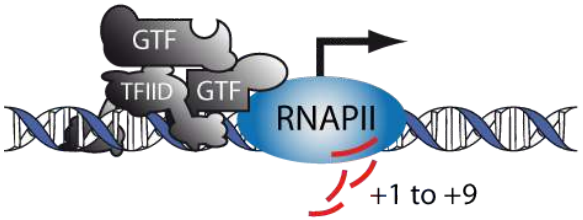
Promoter opening



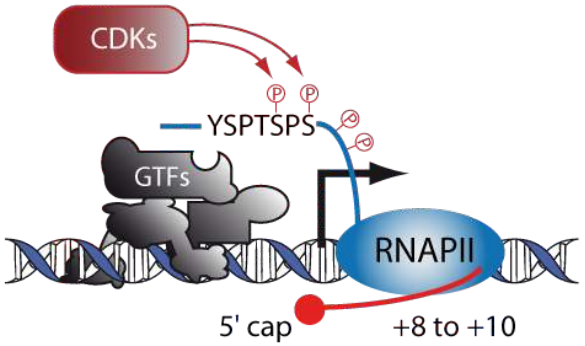
RNA PolII recruitment



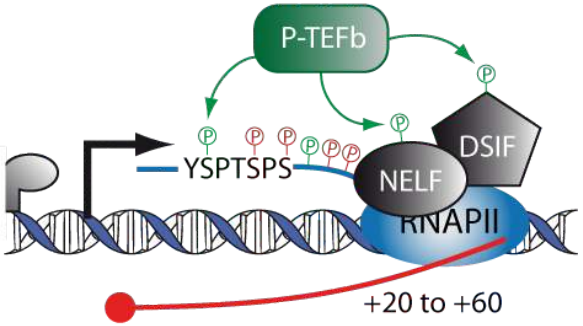
Preinitiation complex formation



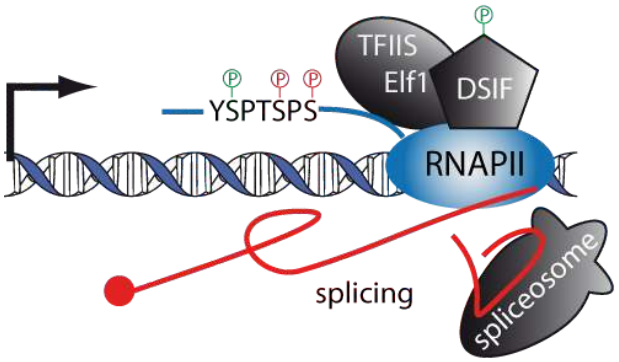
Abortive initiation



Promoter escape



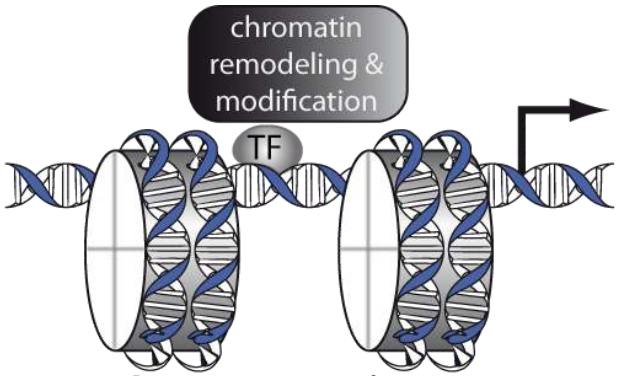
Promoter-proximal pausing



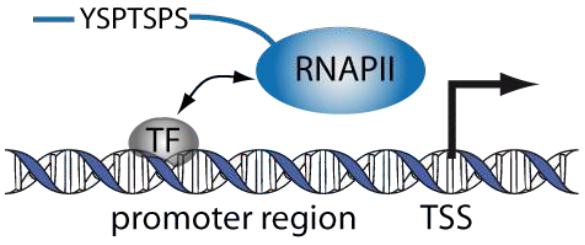
Productive elongation

RNA polymerase II transcription

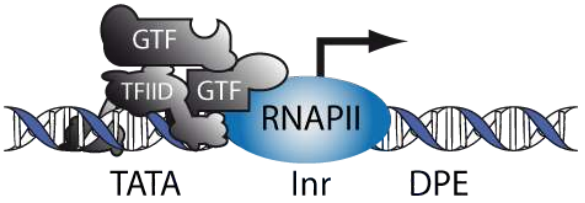
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



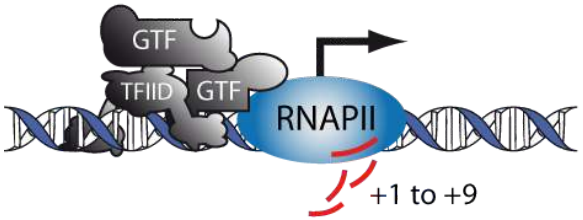
Promoter opening



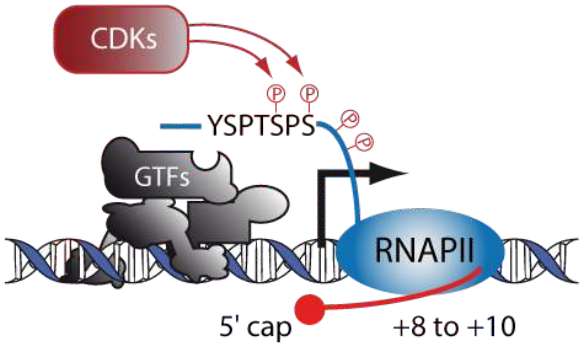
RNA PolII recruitment



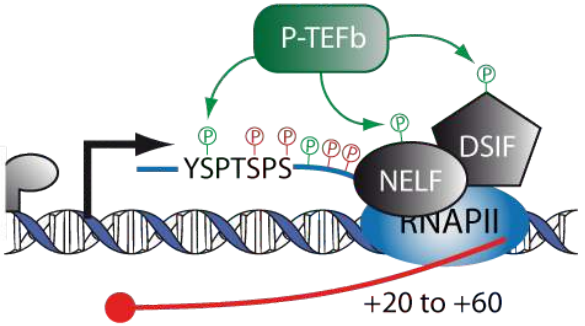
Preinitiation complex formation



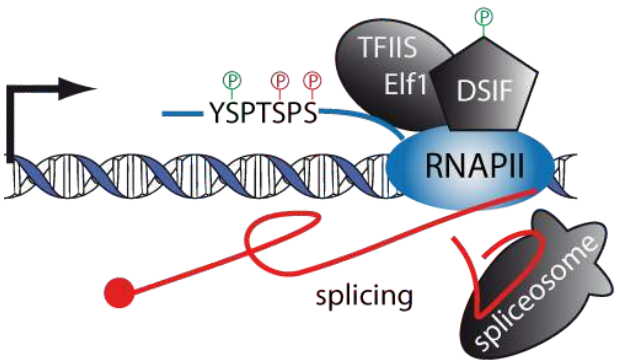
Abortive initiation



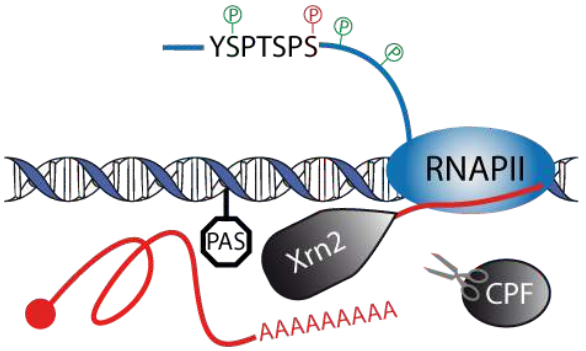
Promoter escape



Promoter-proximal pausing



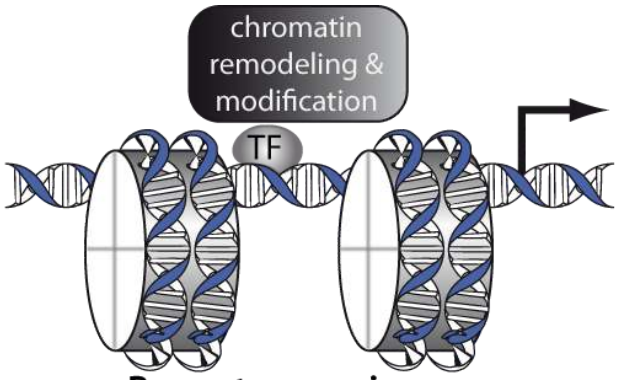
Productive elongation



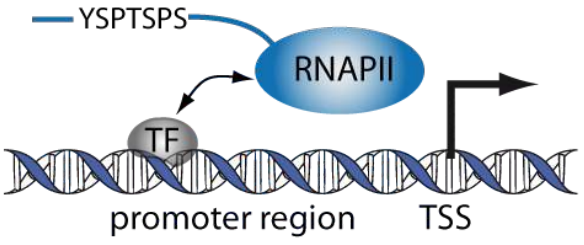
Transcription termination

RNA polymerase II transcription

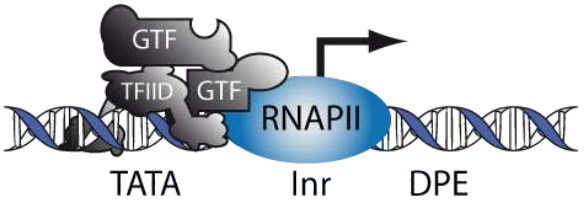
Roeder, TIBS 1996
Margaritis & Holstege, Cell 2008
Zaborowska et al., NSMB 2016
Ngoc et al., Genes Dev 2017



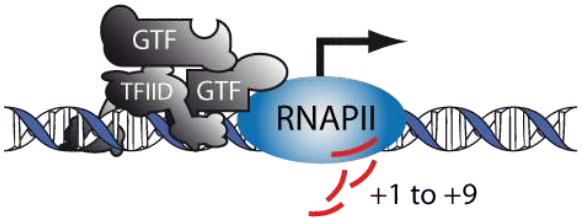
Promoter opening



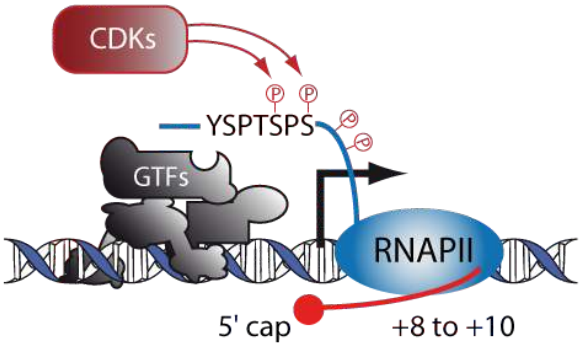
RNA PolII recruitment



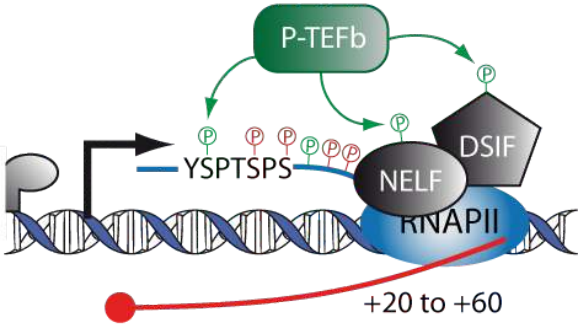
Preinitiation complex formation



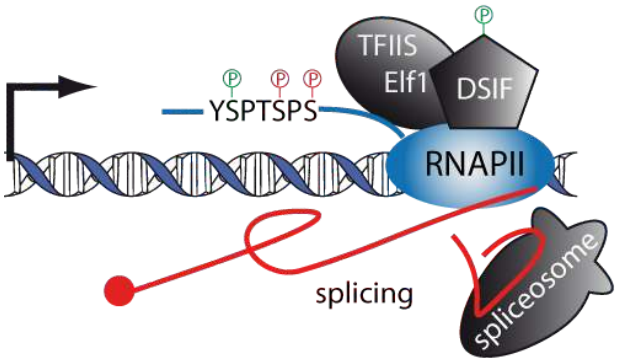
Abortive initiation



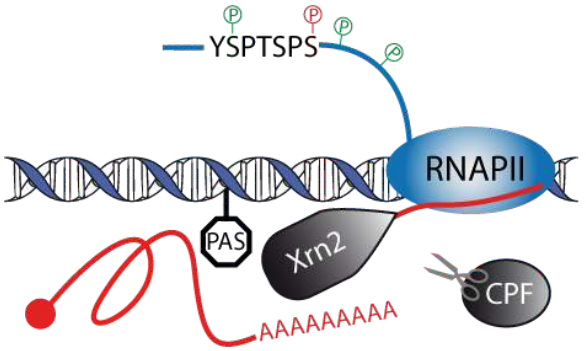
Promoter escape



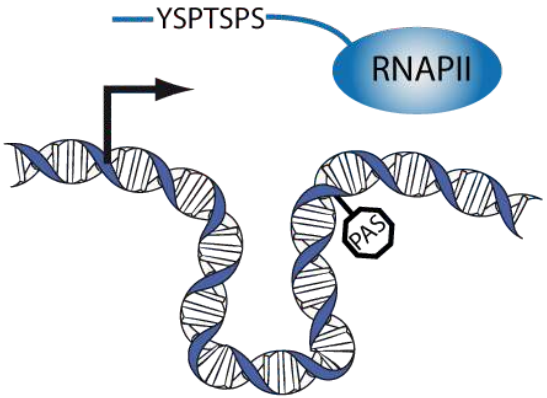
Promoter-proximal pausing



Productive elongation



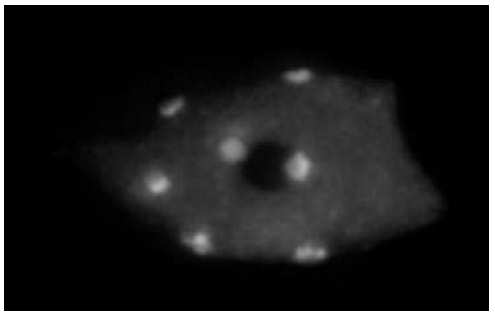
Transcription termination

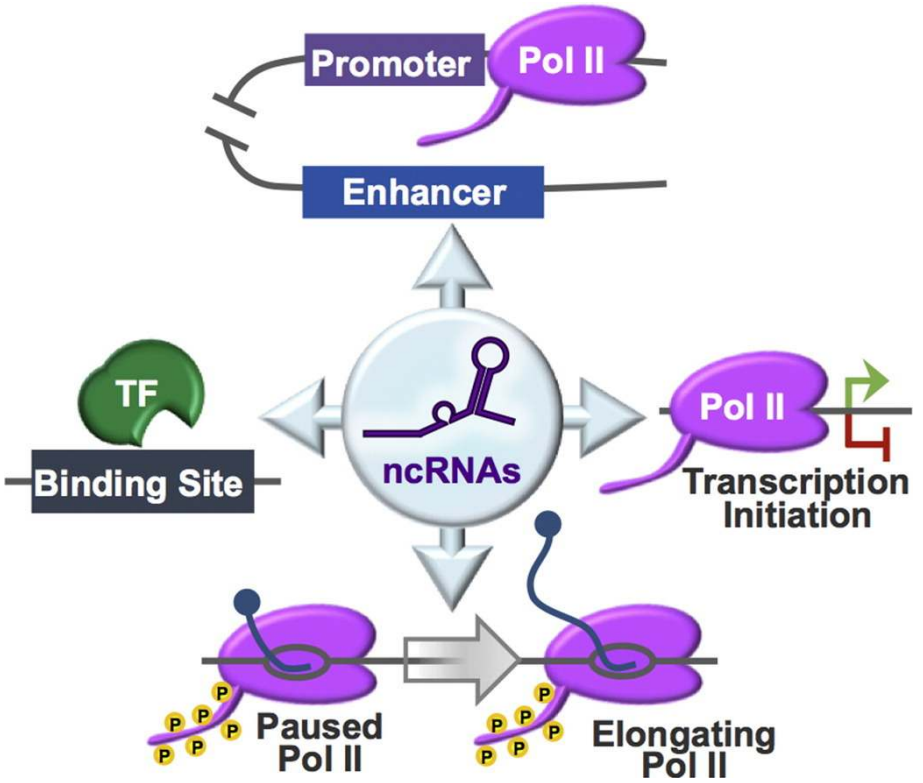


RNAPII recycling

Non coding RNAs & DNA methylation

Nucleus



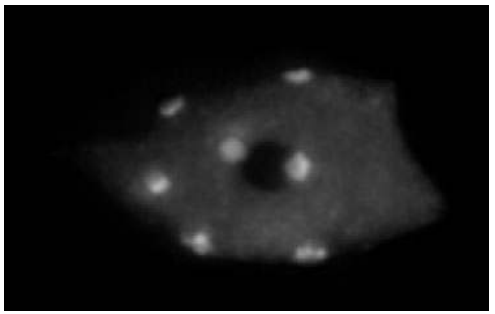


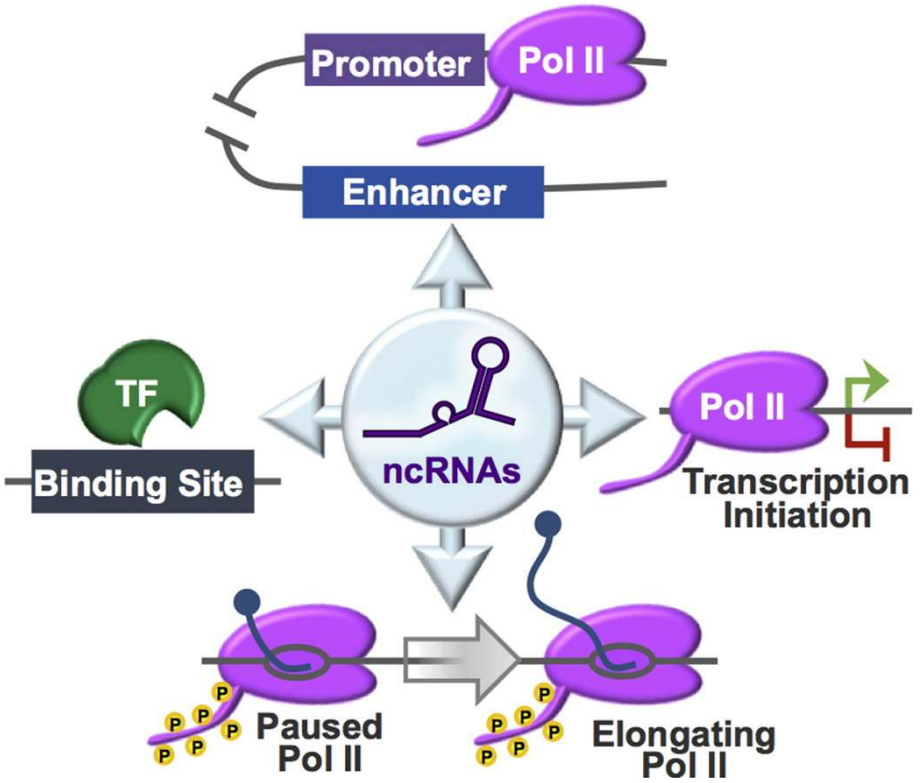
Sabin et al., Mol Cell, 2013
Kornienko et al., BMC Biol, 2013
Mercer & Mattick, NSMB, 2013

Eidem et al., J Mol Biol 2016

Non coding RNAs & DNA methylation

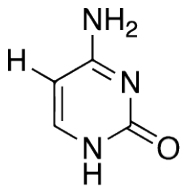
Nucleus



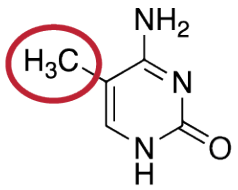


Sabin et al., Mol Cell, 2013
Kornienko et al., BMC Biol, 2013
Mercer & Mattick, NSMB, 2013

Eidem et al., J Mol Biol 2016

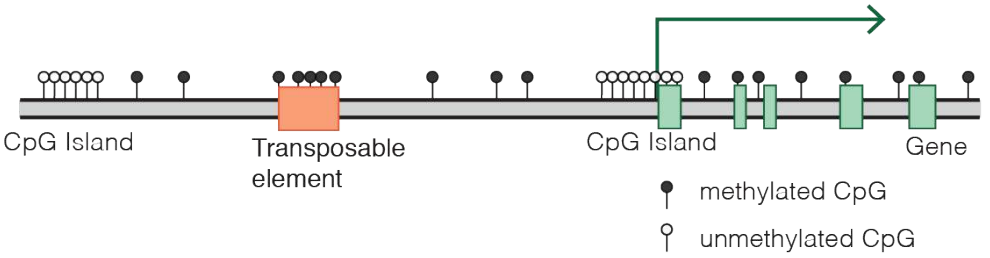


Cytosine



methylated Cytosine

Typical mammalian DNA methylation landscape

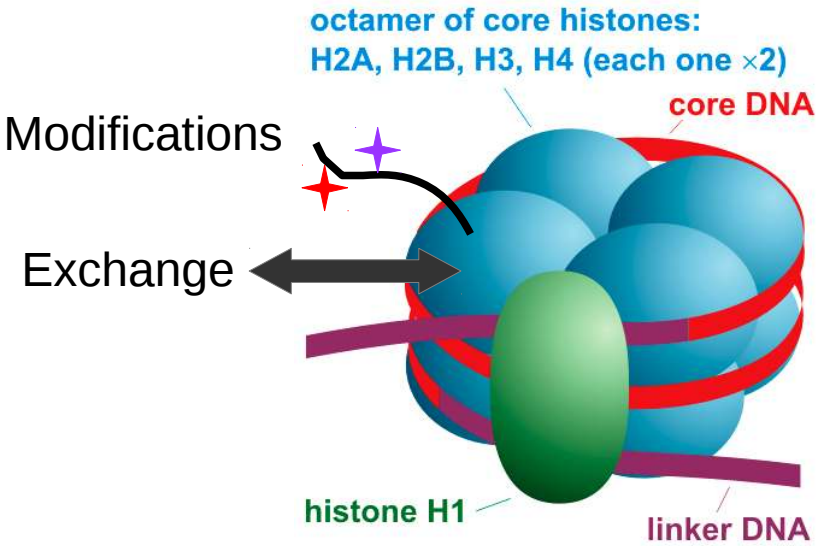


source: Wikipedia

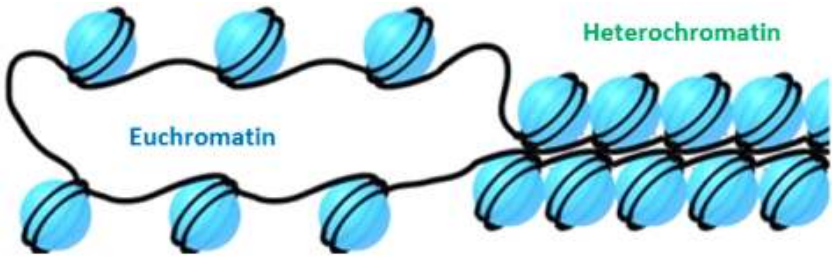
Chromatin

Chromatin = DNA + proteins

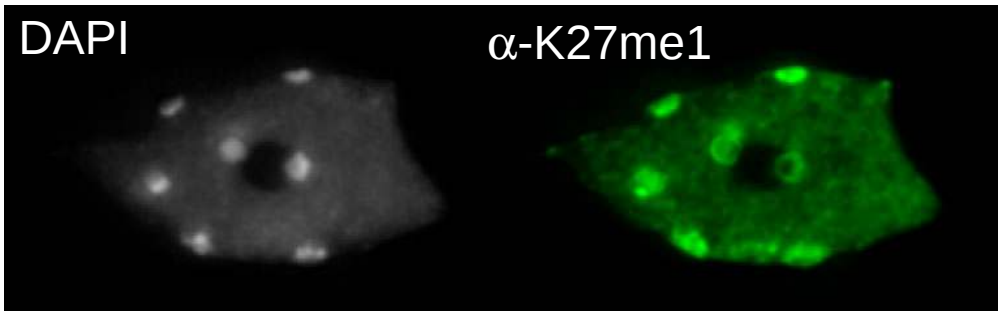
Nucleosome



Lai & Pugh, Nat Rev MCB 2017
Voong et al., Cell 2016
Bednar et al., Mol Cell 2017

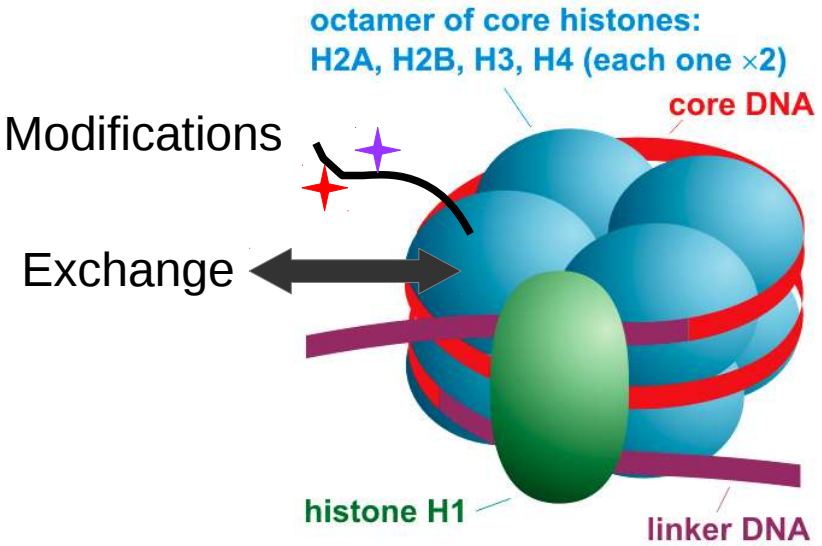


Chromatin

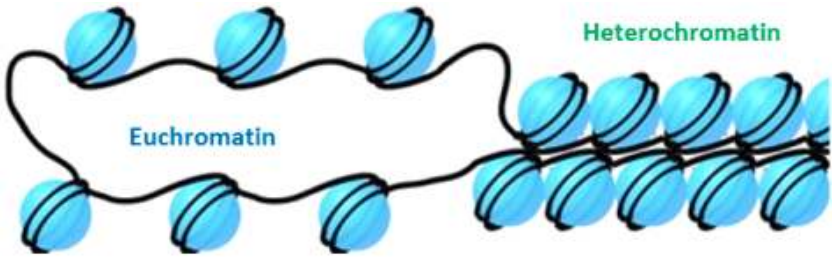


Chromatin = DNA + proteins

Nucleosome



Lai & Pugh, Nat Rev MCB 2017
Voong et al., Cell 2016
Bednar et al., Mol Cell 2017

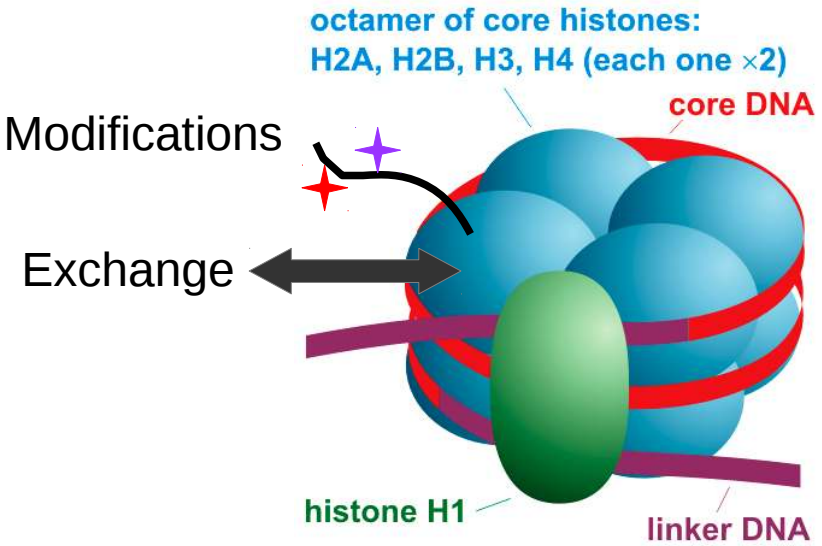


Chromatin

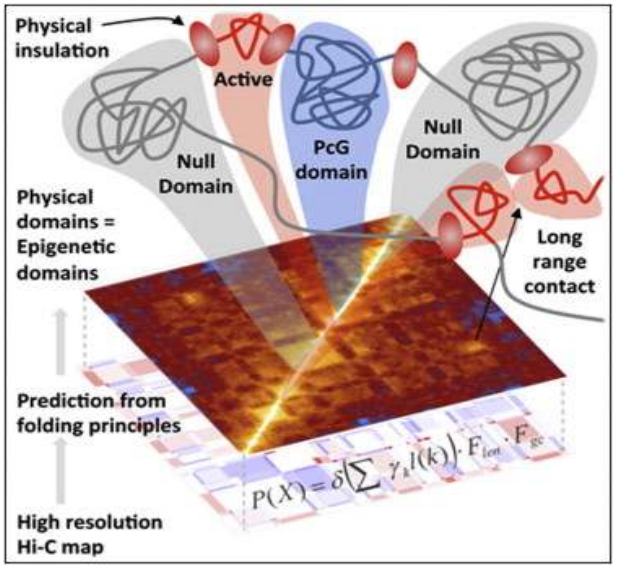
Chromatin = DNA + proteins

3D chromatin organization

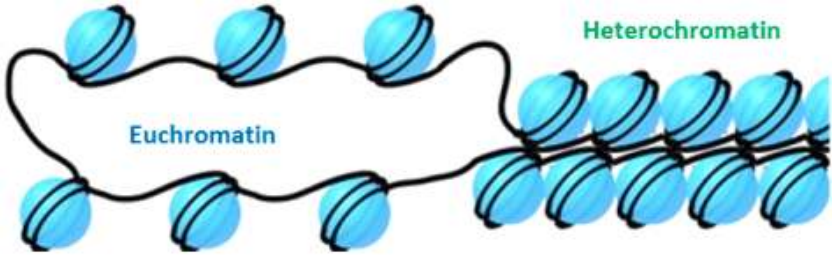
Nucleosome



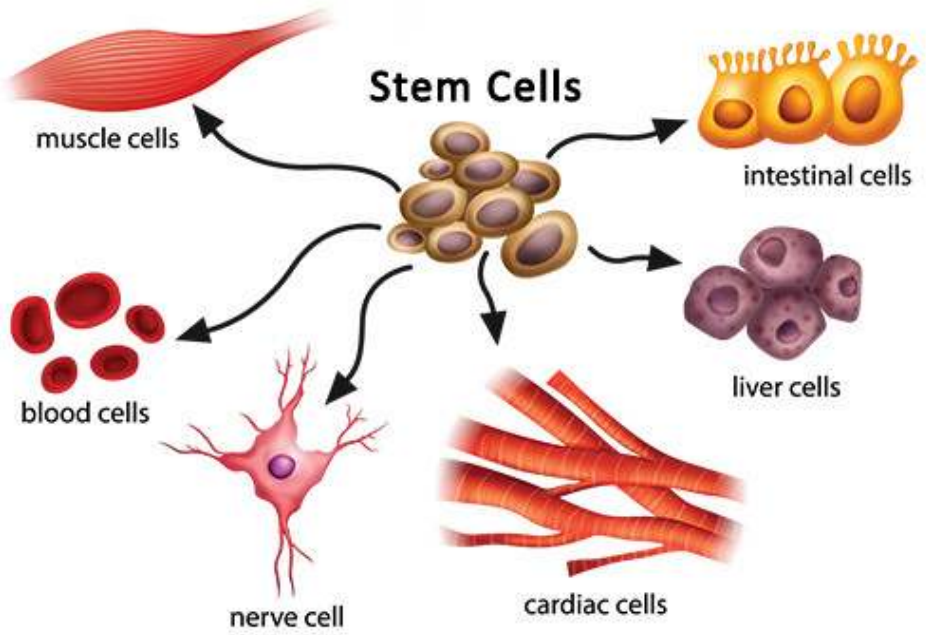
Lai & Pugh, Nat Rev MCB 2017
Voong et al., Cell 2016
Bednar et al., Mol Cell 2017

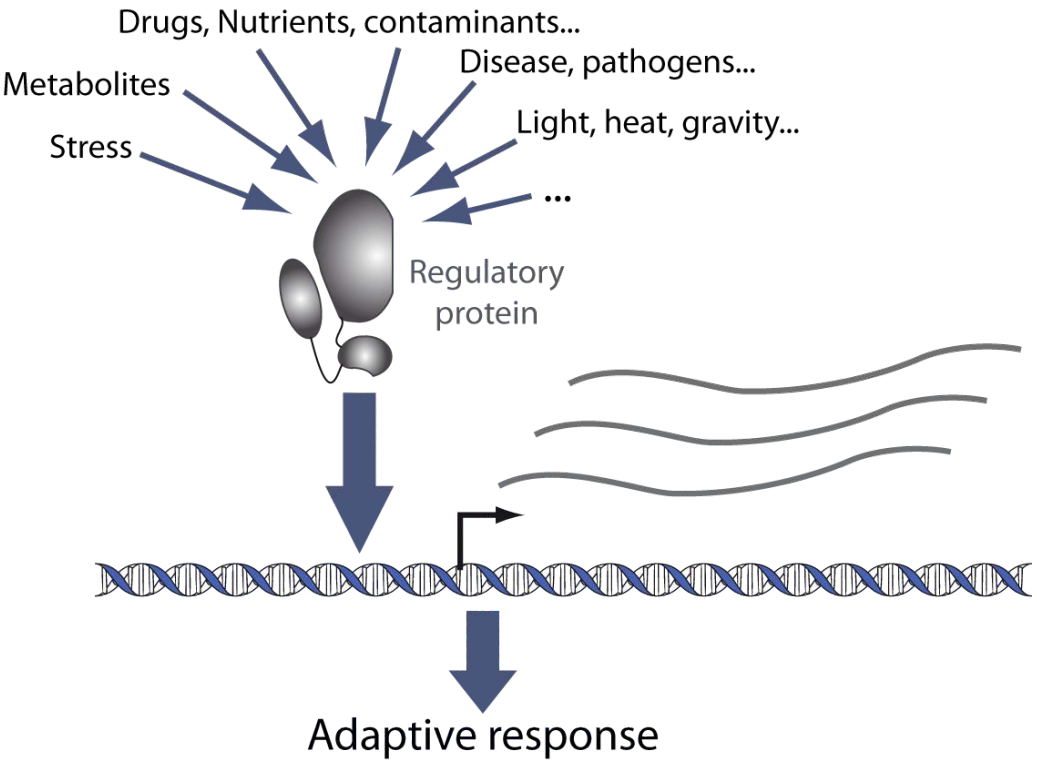


Sexton et al., Cell, 2012
Sanyal et al., Nature 2012
Nora et al., Nature 2012

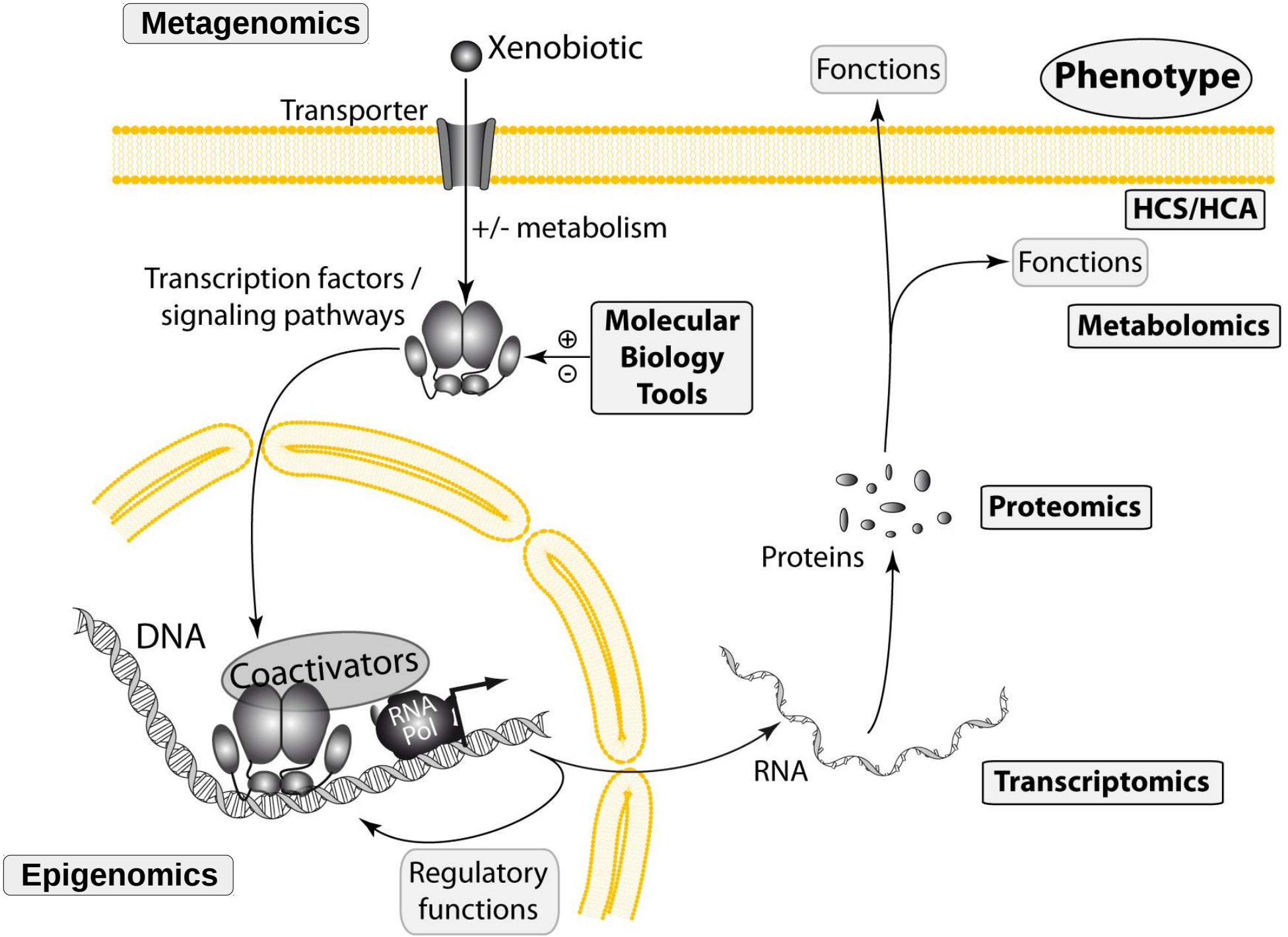


Importance of gene expression regulations





Toxicogenomics



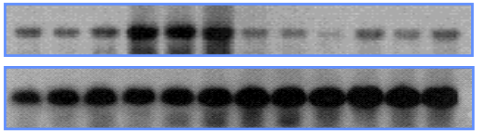
Transcriptomics

Quantify RNA abundance

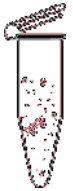
Low throughput / targeted

Few genes

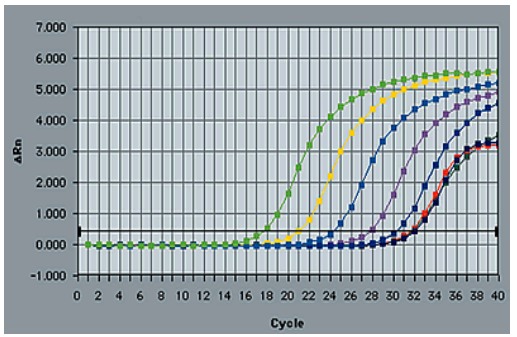
Northern blot / dot blot



Ribonuclease protection assay (RPA)



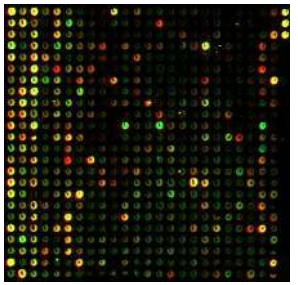
Real-time quantitative PCR (qPCR)



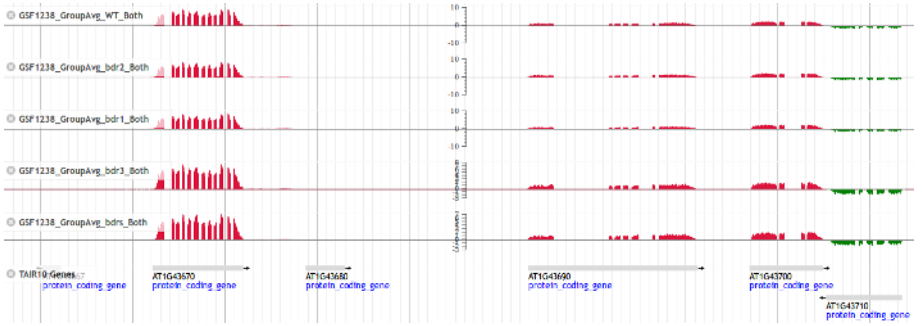
High throughput / genome wide

Many / all genes

Microarrays



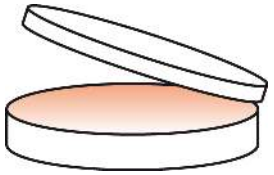
RNA-seq



"Low" throughput methods

Northern blot

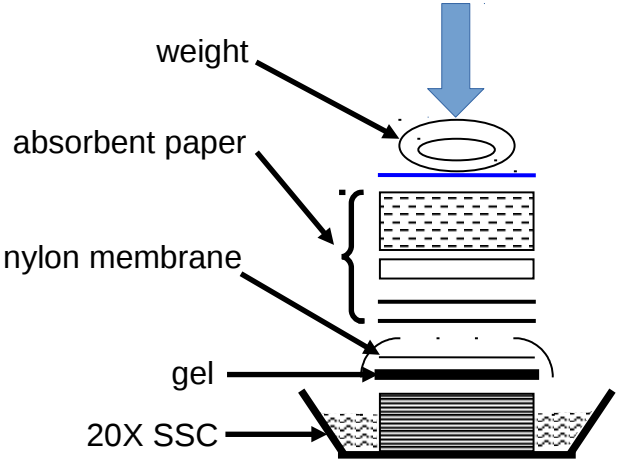
Organ Cells



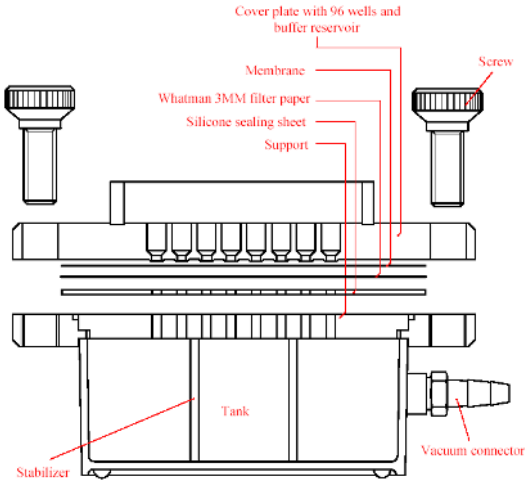
Treated vs untreated
Wild-type vs transgenic

RNA extraction
(total, polyA+, ...)

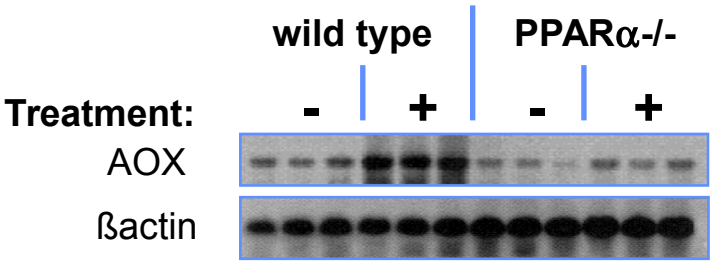
Migration on
agarose gel



Transfer on a
nylon membrane



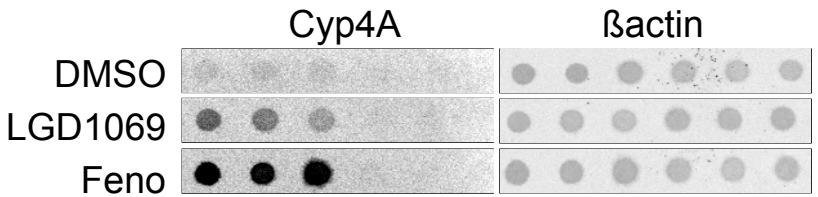
Direct spotting as dots / slots



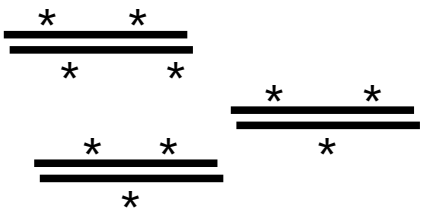
Northern Blot

Densitometric
quantitation

Hybridization and washing



Dot and Slot Blot



Labeling by
random priming

Specific probe

Northern blot

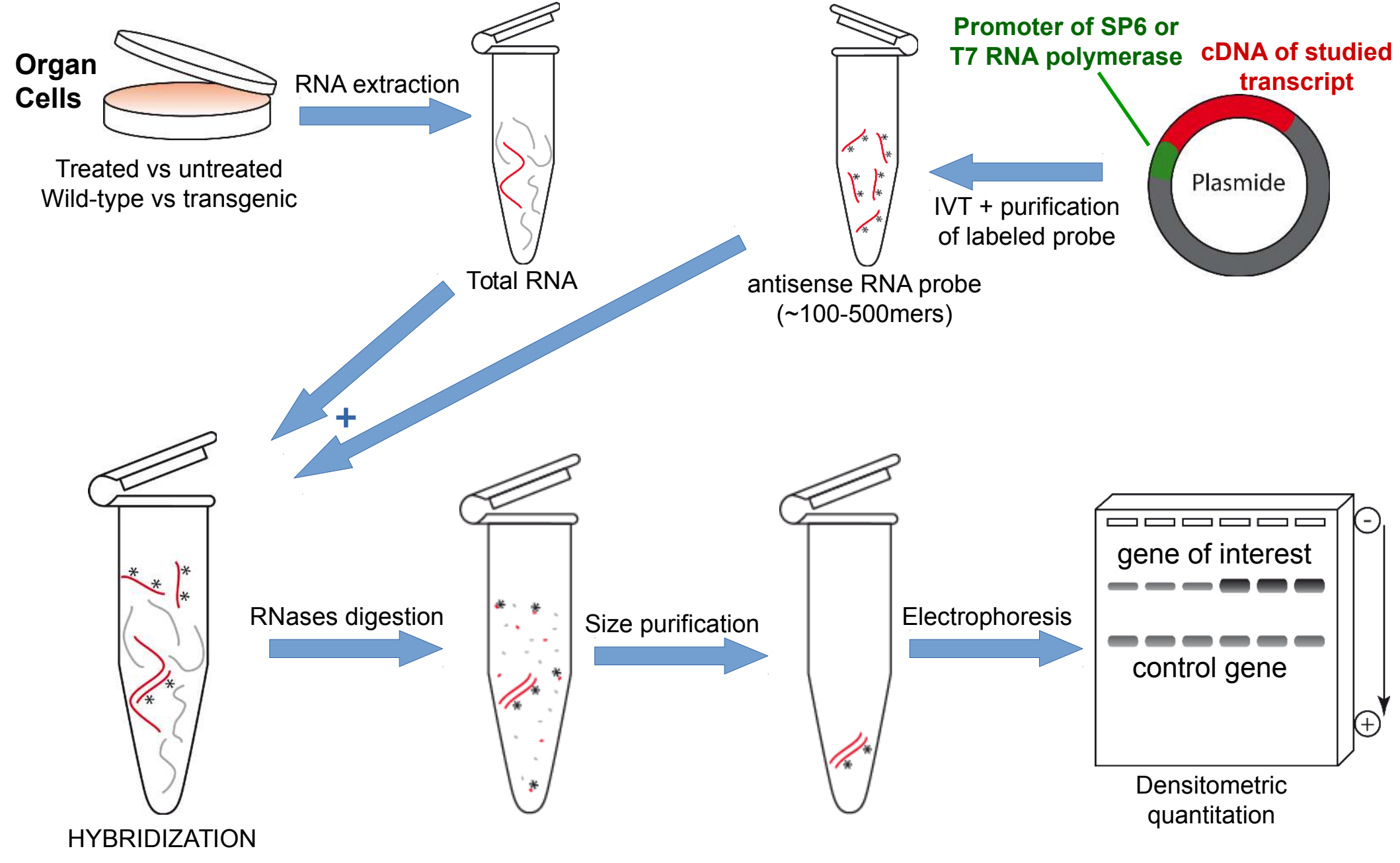
Advantages:

- Detect alternative transcripts (if size are sufficiently different and probe is common to the different transcripts)
- Evaluate transcript(s) size
- Detect regions / exons shared among transcripts
- Low/medium sensitivity to RNA degradation + vizualize / evaluate RNA degradation
- Low cost
- Quantitative and qualitative (see the "bands")

Drawbacks / limits:

- Need to know the sequence (probe ~1Kb)
- Specificity may be hard to obtain (multigenic family)
- Low sensitivity
- Hard to set up (tedious, long, "dirty" : Ethidium Bromide, formaldehyde, formamide, ^{32}P)

RNase protection assay (RPA)



RPA

Advantages:

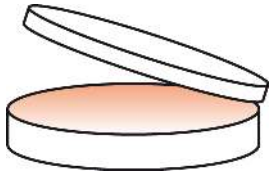
- Liquid phase hybridization => better sensitivity than NB (low abundance transcripts)
- Increase starting RNA (up to 100µg) => detect low abundance transcripts
- No expensive instrument
- Short probe => relatively low sensitivity to moderate RNA degradation
- Can study up to 10 different transcripts simultaneously (probes of different sizes)
- Easier to discriminate alternative transcripts of similar size
- Allows to study transcript ends and exon-exon junctions

Drawbacks / limits:

- No info on transcript size
- Obtaining the probe can be tedious
- Manipulation of material that is sensitive to degradation (single stranded RNA probe)
- Sensitive to small sequence differences between the probe and the target

Real-time quantitative PCR

Organ Cells



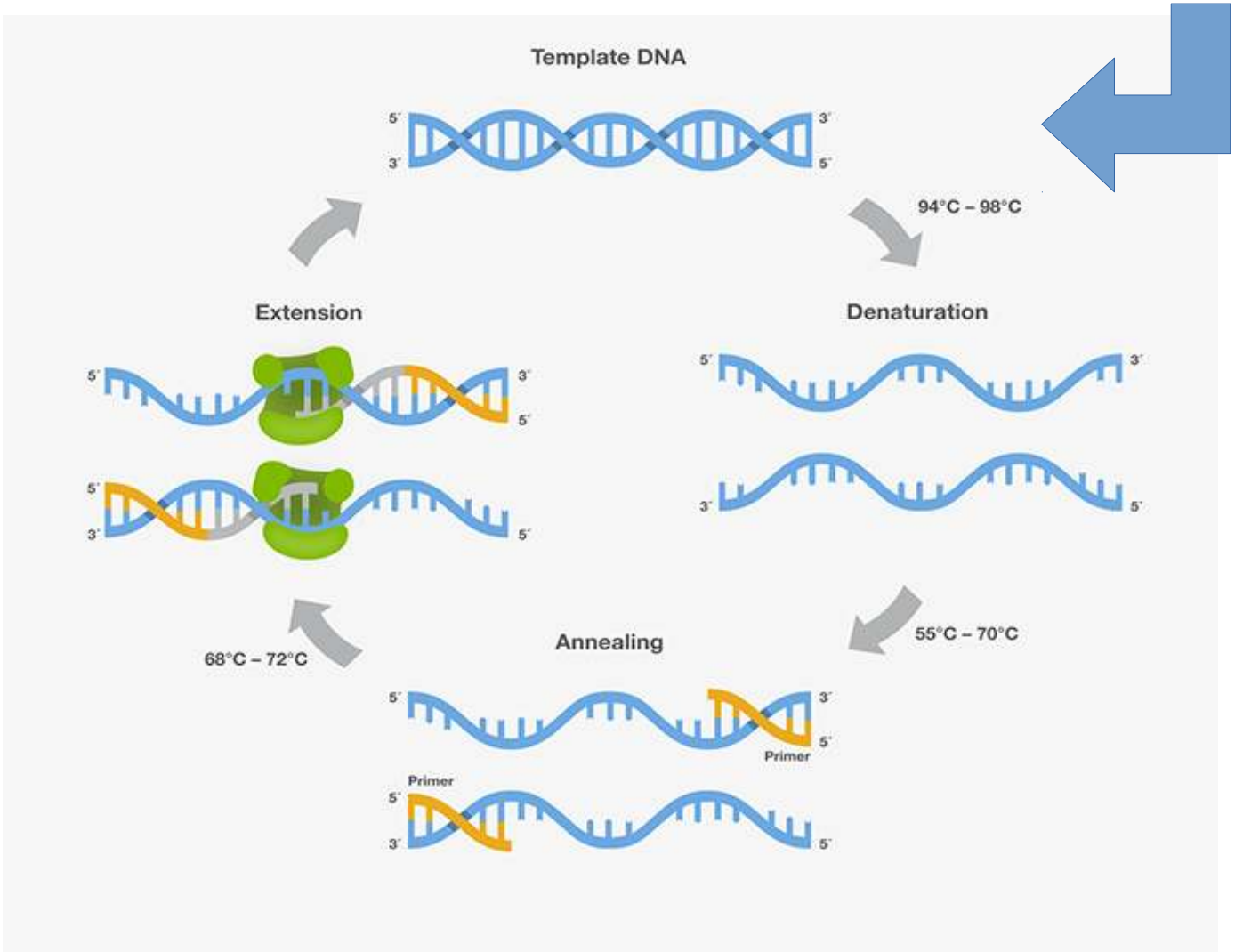
Treated vs untreated
Wild-type vs transgenic

RNA extraction

Total RNA

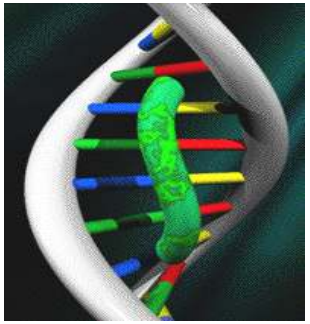
Reverse transcription

- oligodT
- random primers
- specific primers



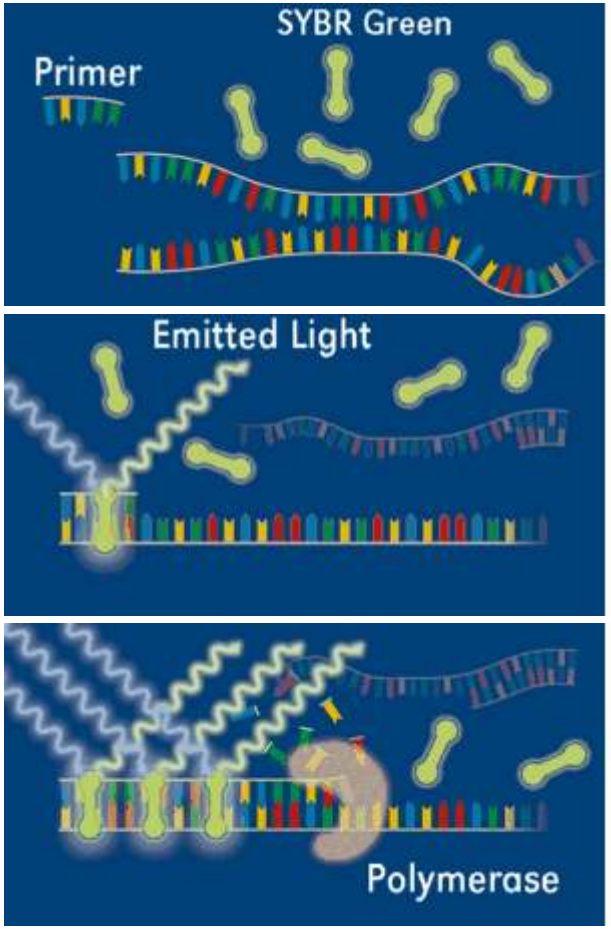
qPCR chemistries

Follow in real-time the production of DNA amplicons



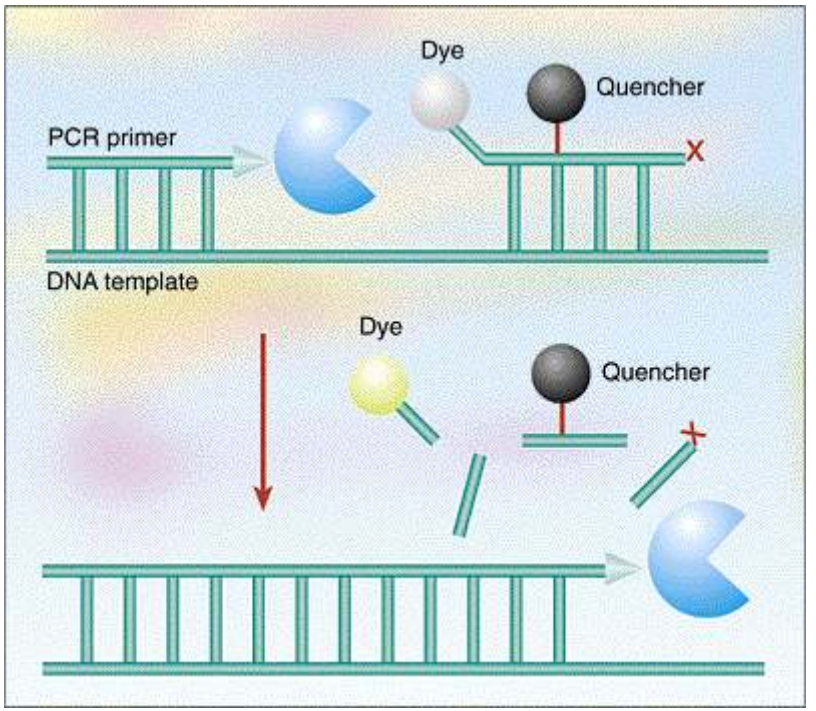
Intercalating agent
(SYBR green, Eva Green...)

non sequence-specific



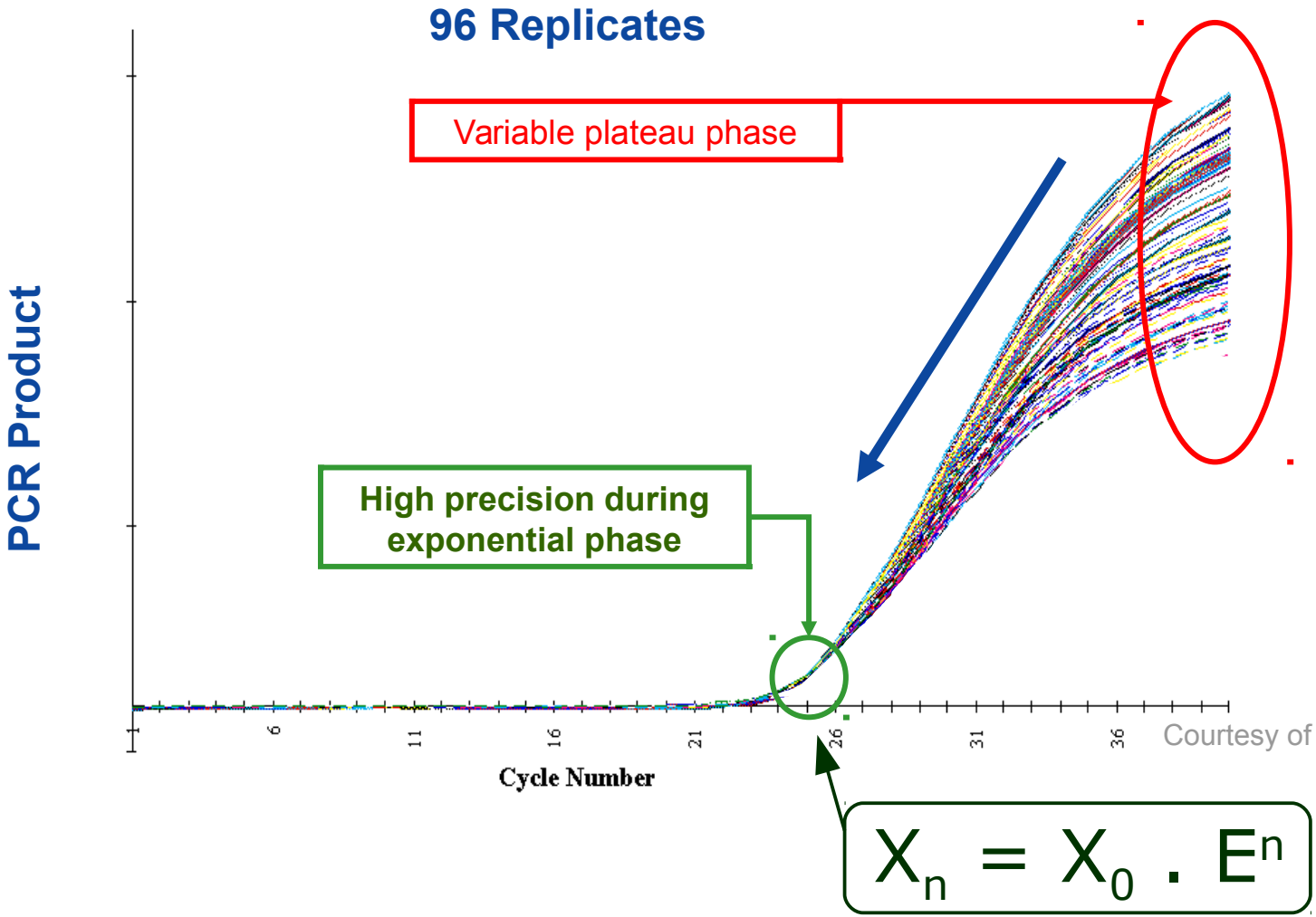
Probe-based chemistries
(TaqMan, Lightcycler probes, ...)

sequence-specific => multiplexing



http://dyes.gene-quantification.info/

qPCR reaction



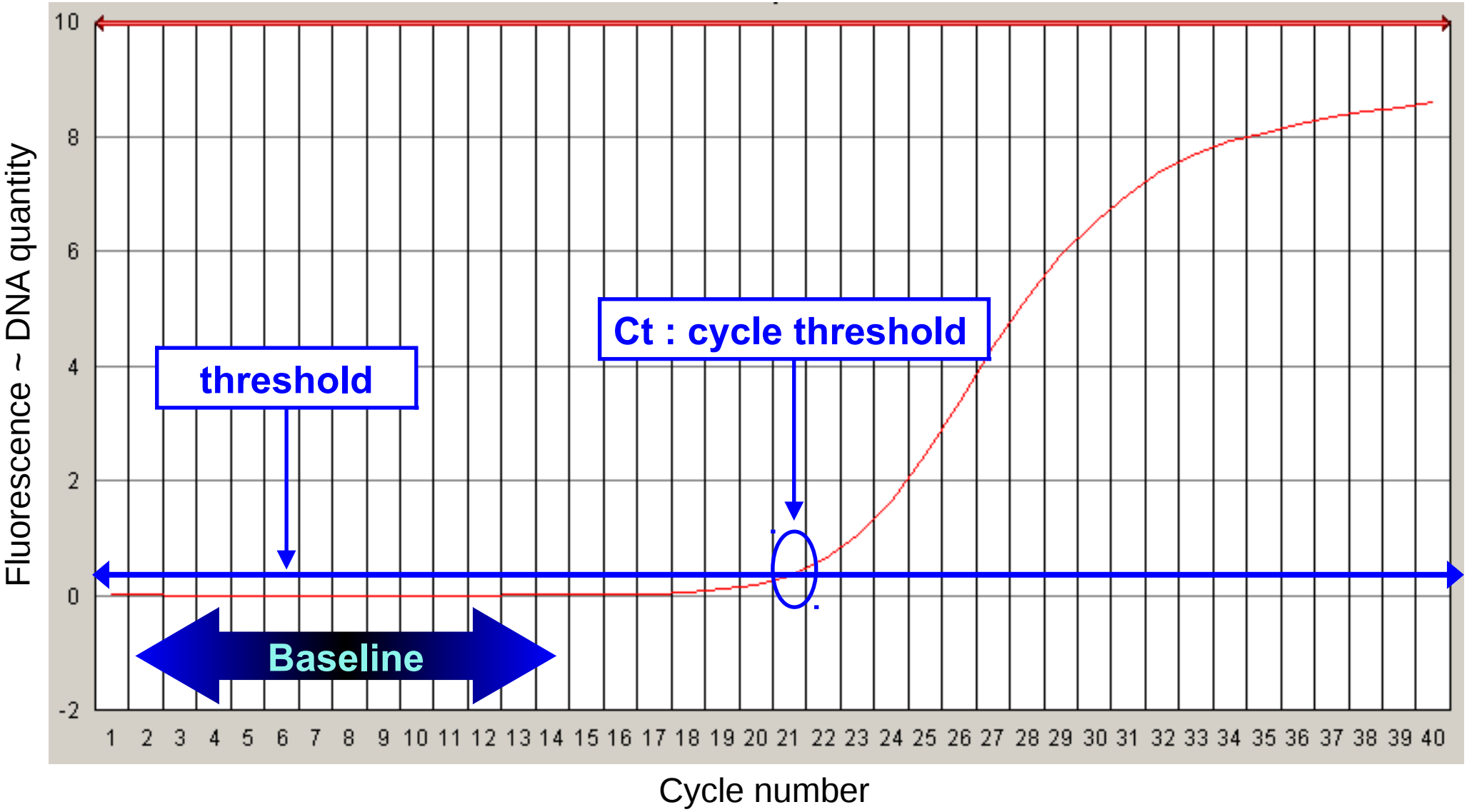
Courtesy of Applied Biosystems, France

X_n : # amplicons at cycle n

X_0 : initial # of transcrits

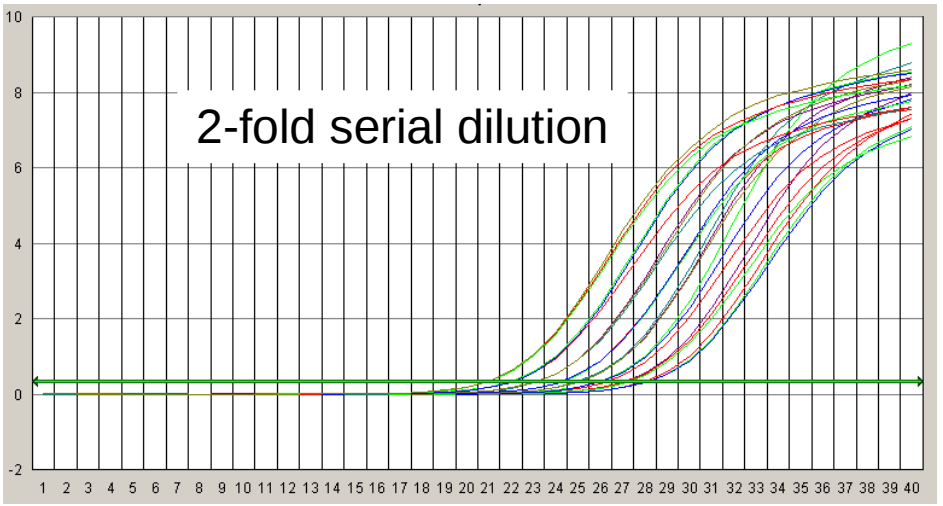
E = PCR efficiency in [1:2]
(E=2 ⇔ 100% efficiency)

qPCR reaction



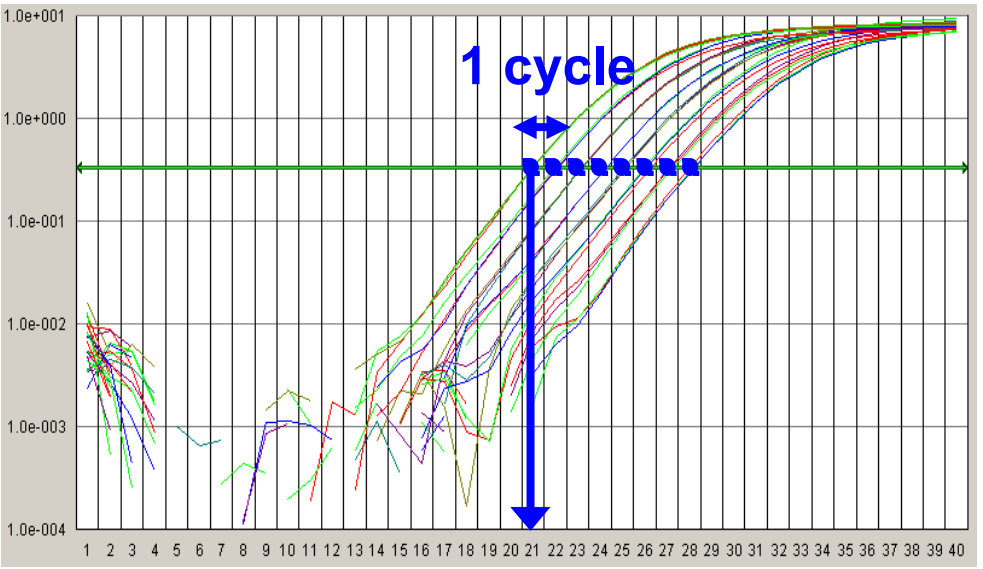
qPCR reaction

Fluo

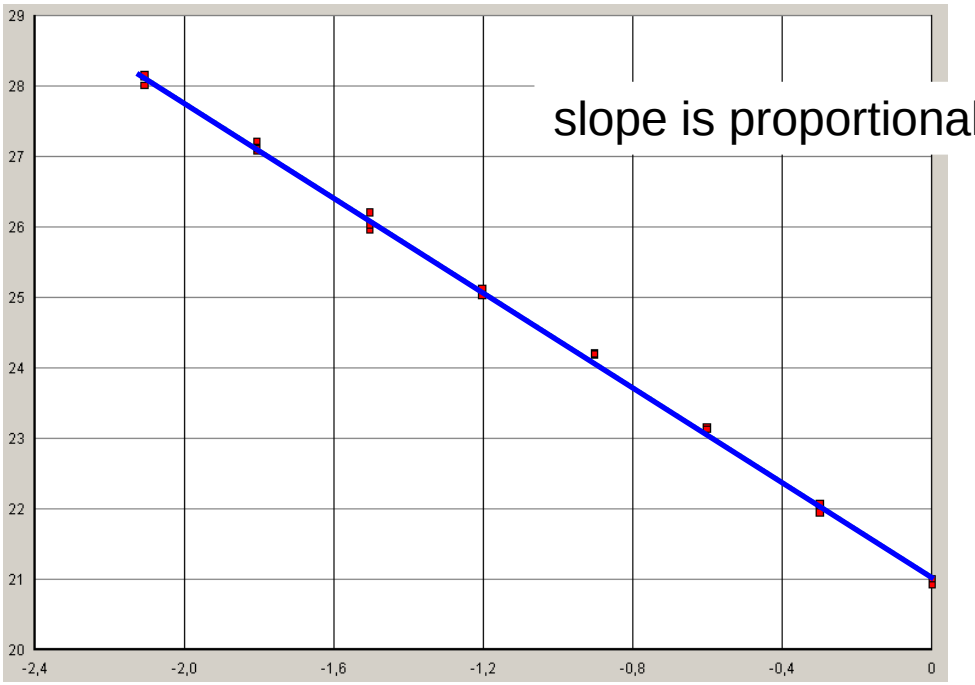


log

Fluo (log scale !!!)



Ct



slope is proportional to PCR efficiency

qPCR reaction

3 fundamental principles:

1) Fluorescence is proportional to amplicon accumulation (specificity, primer/probe selection)

2) PCR efficiency is the same for all samples

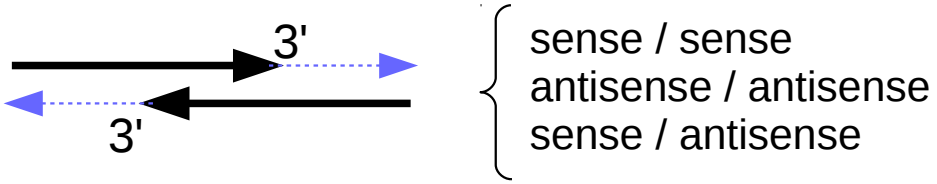
3) Threshold must intersect all PCR curves at the exponential phase

Peirson et al., NAR 2003

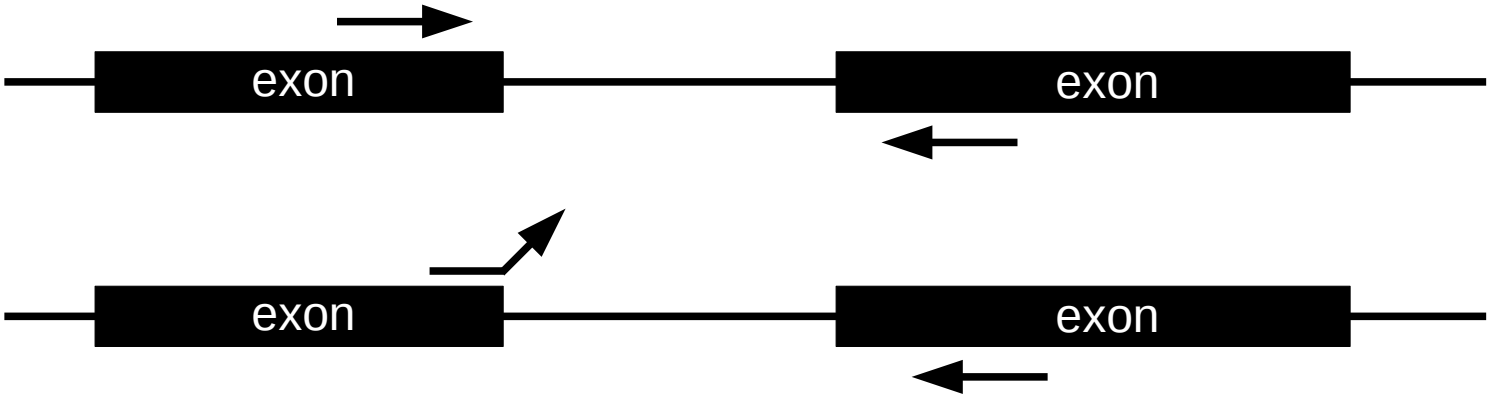
qPCR primer design

1) Fluorescence is proportional to amplicon accumulation (specificity, primer/probe selection)

- Design specific primers ("know your target")
- Avoid 3'-3' primer dimers



- Take advantage of DNA – RNA differences

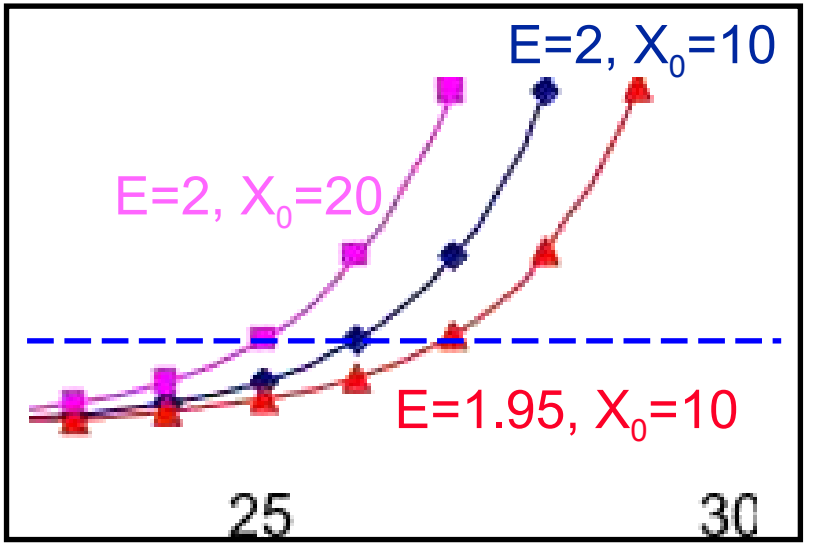


qPCR efficiency

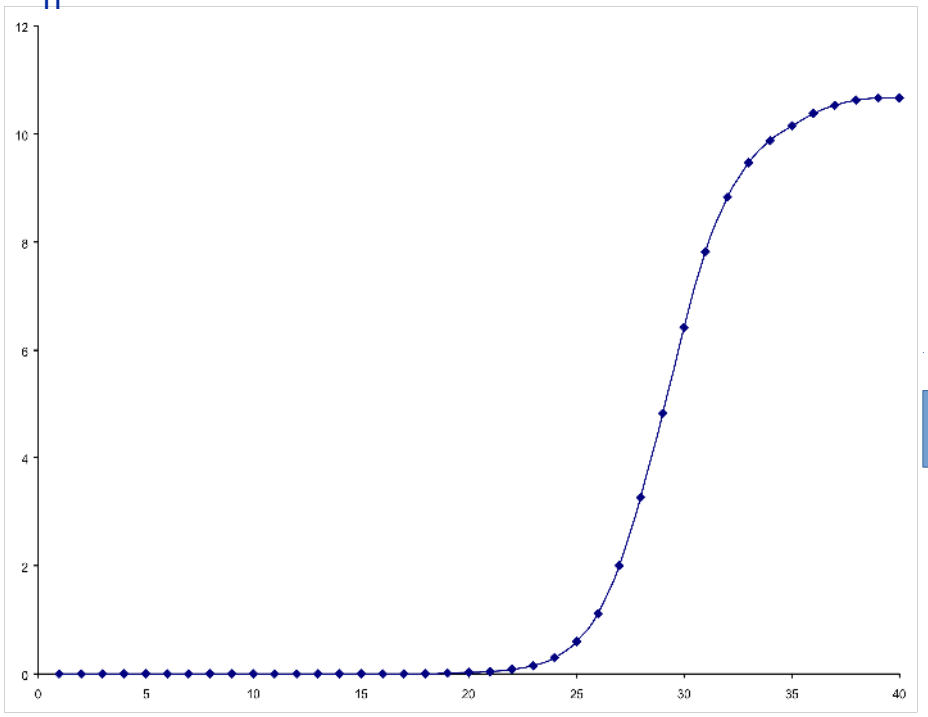
2) PCR efficiency is the same for all samples

$$X_n = X_0 \cdot E^n$$

log $\log(X_n) = \log(X_0) + n \cdot \log(E)$



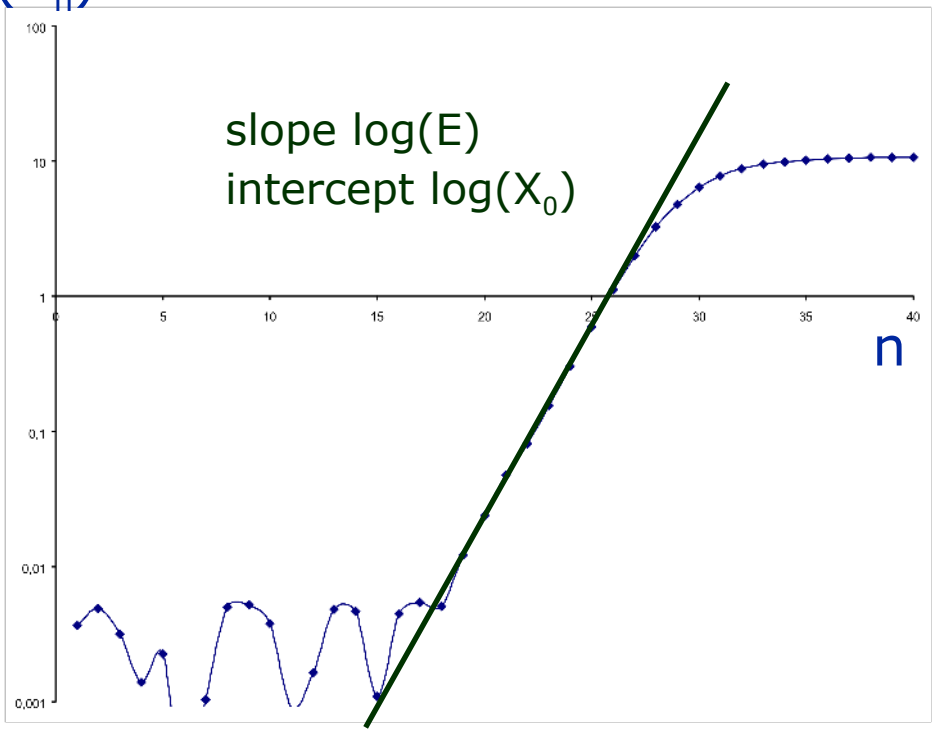
X_n



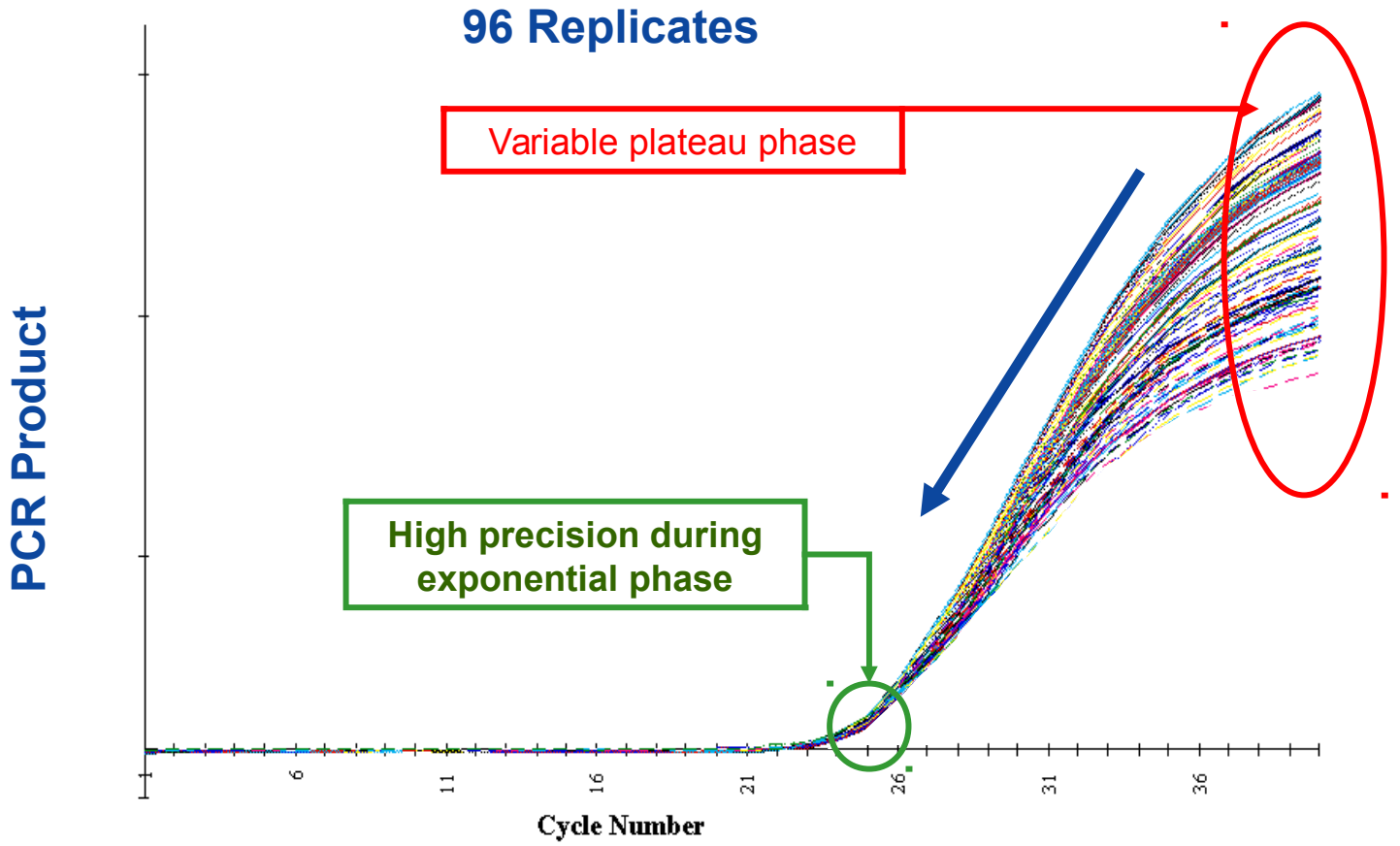
n

log

$\log(X_n)$



3) Threshold must intersect all PCR curves at the exponential phase



Ct is linked to initial number of RNA molecules and not to different PCR kinetics

qPCR

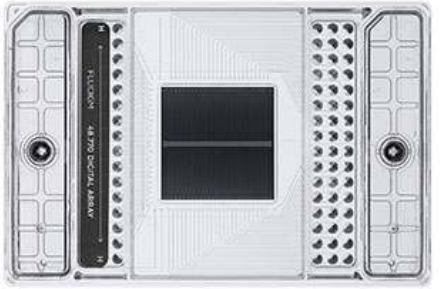
Advantages:

- High sensitivity
- Easy and quick set-up
- Low sensitivity to RNA degradation (short amplicons)
- Highly specific (e.g. single base discrimination at primer 3' end)
- Multiple applications (gene expression, SNP genotyping, CNV, ChIP, etc.)
- Miniaturization

Roche Lightcycler 1536

1536 multiwell plates

Biomark HD (Fluidigm)



96 samples x 96 primers = 9216 reactions
12-48 samples x ~770 primers : up to ~36K reactions

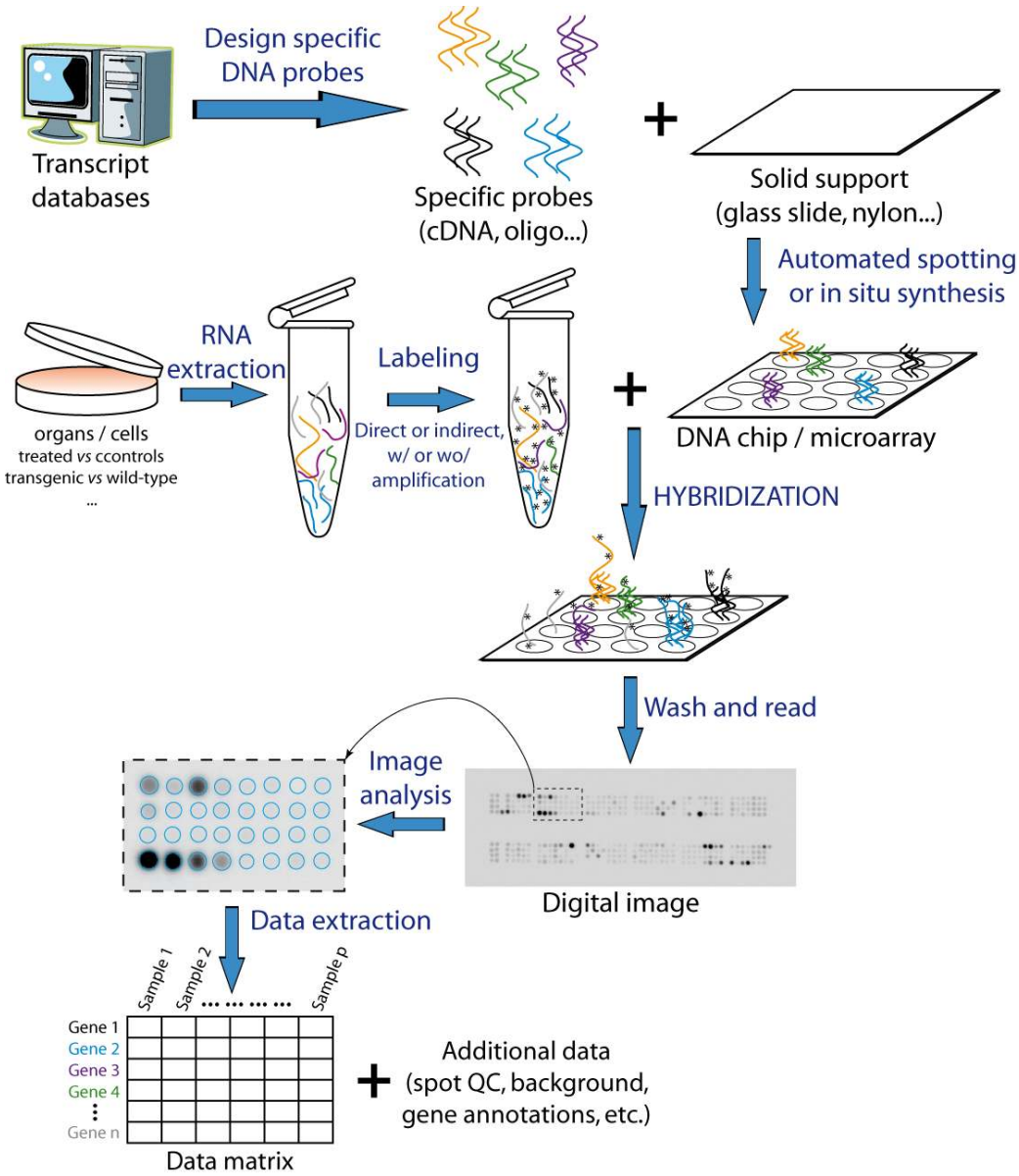
Drawbacks / limits:

- Sensitive to PCR inhibitors (salts, inonic detergents, alcohols, ...)
- Need precise sequence information
- Data analysis a bit more complicated
- No info on transcript size or presence of alternative transcripts

High throughput methods

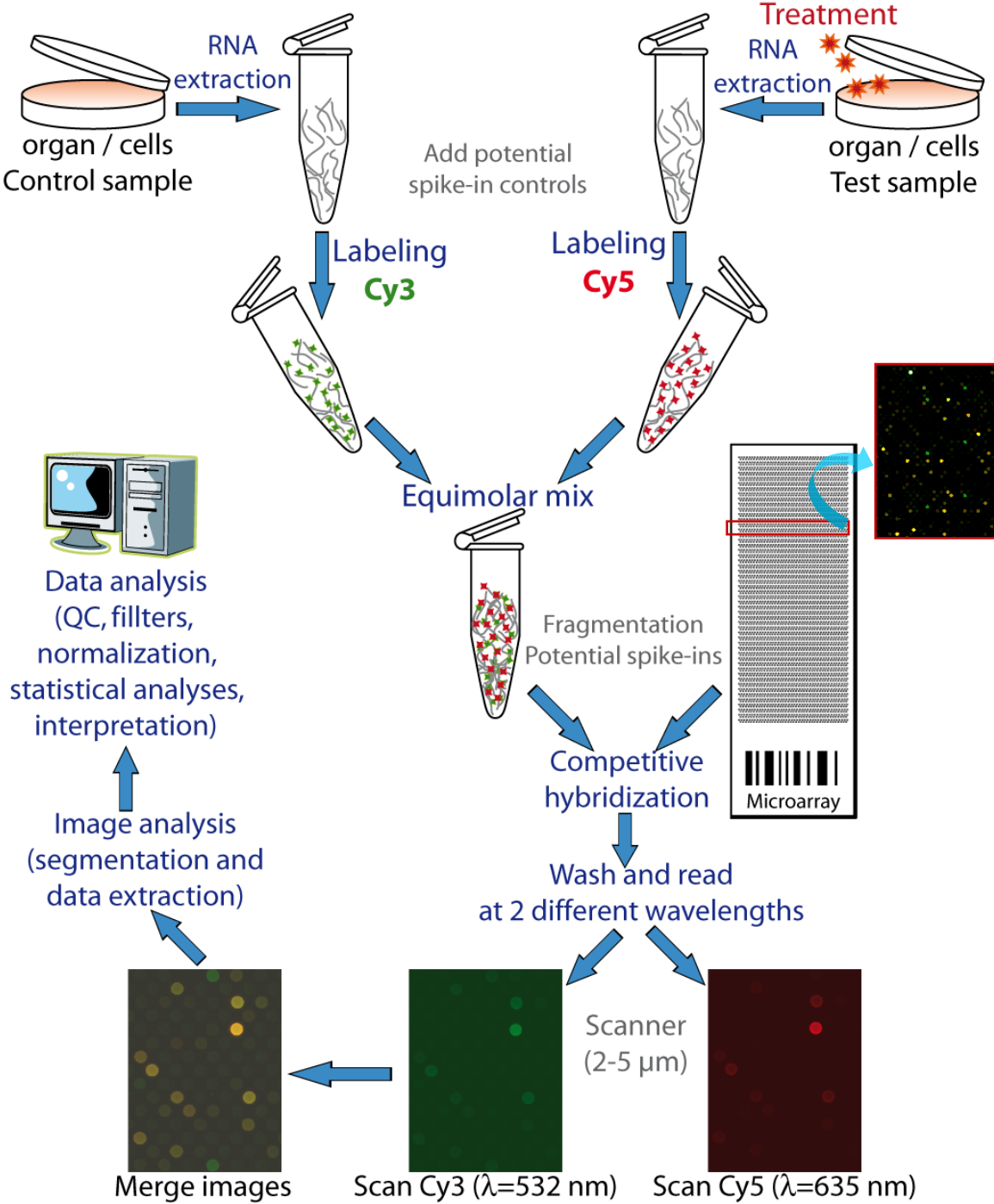
Microarrays

Developed in the mid 90s thanks to improved knowledge of whole genome sequences

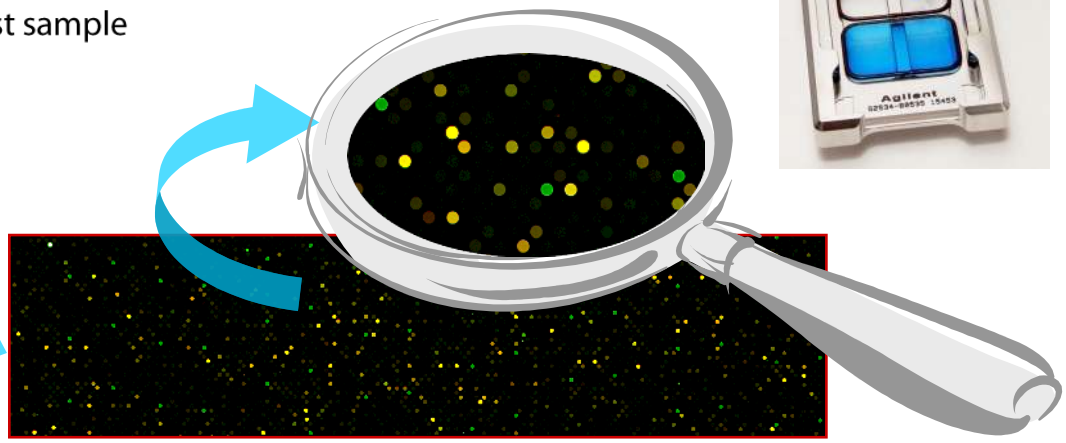


Affymetrix
GeneChip

Dual color microarrays



8 x 60K
4 x 180K
2 x 400K
1M



Agilent microarrays

- Concept applied to :
  - gene expression
  - CNV
  - protein chips
  - tissue array
  - ...

Microarrays

Advantages:

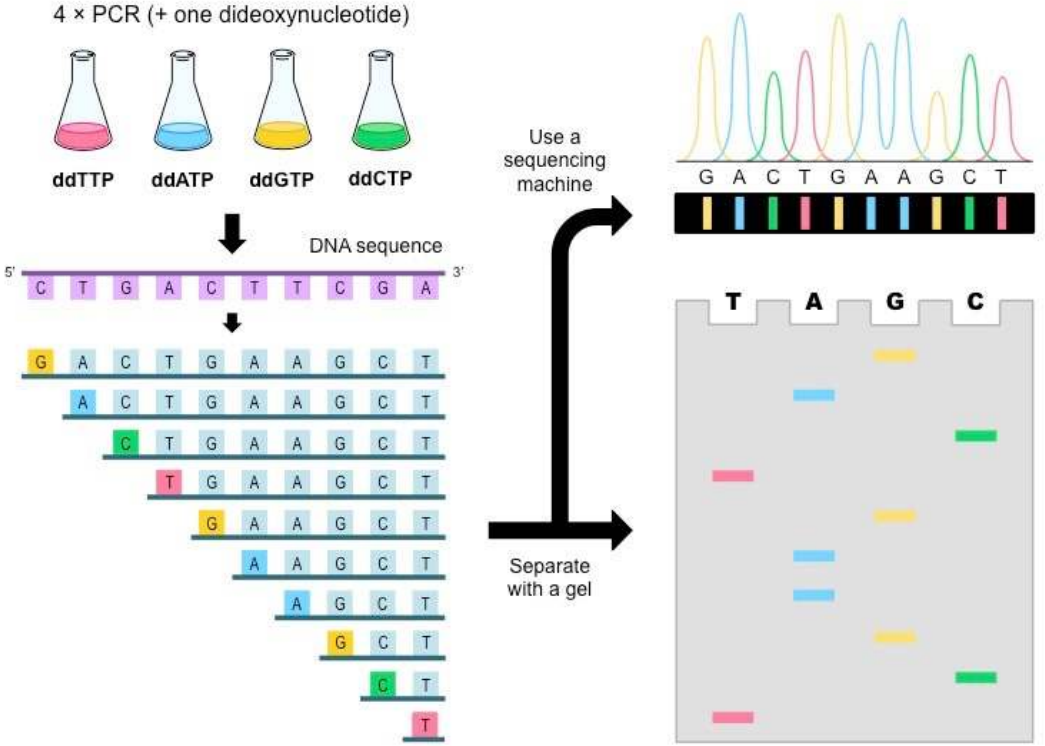
- Fast and easy to use
- Good reproducibility (better intra-platform, MAQC)
- Excellent transcriptome coverage (when sequence is known and annotated)
- Low cost (~150-200 € / sample)
- Measurements for different genes / transcripts are essentially independent from each other

Drawbacks / limits:

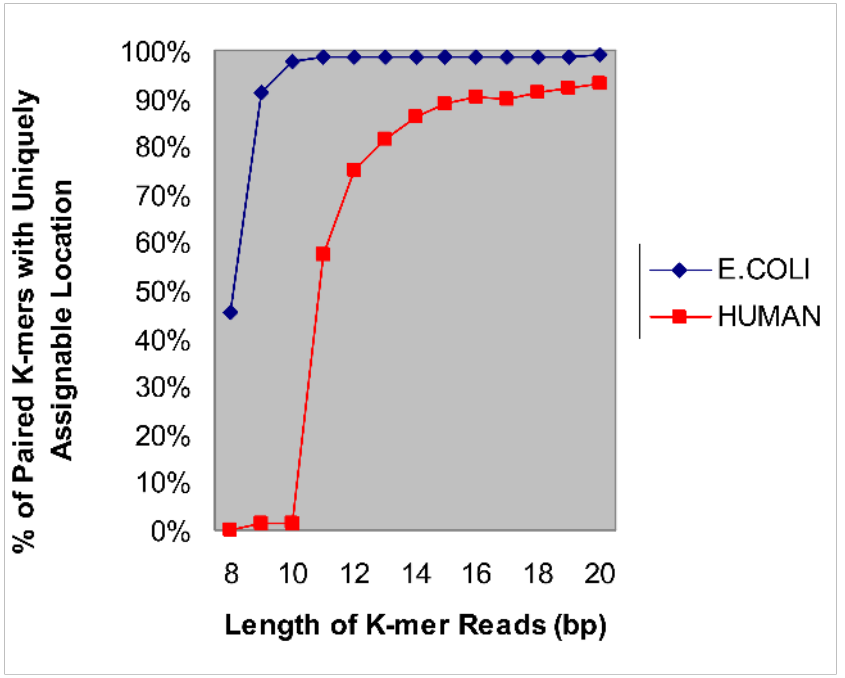
- Need a well annotated genome / transcriptome
- Cross hybridization can occur
- Expensive instruments (scanner) => facilities
- Data analysis is non trivial

NGS development

Sanger sequencing (70s-80s)



For resequencing, short reads work



Jay Shendure

→ Human Genome Project (1990-2003)

→ Development of technologies for massively sequencing short tags (e.g. MPSS)

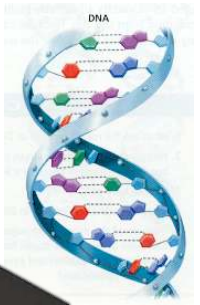
Brenner et al., Nat Biotech 2000

→ Development of currently used technologies (~2005)

What is NGS ?

Adenine
Guanine
Cytosine
Thymine
Uracil

RNA → DNA



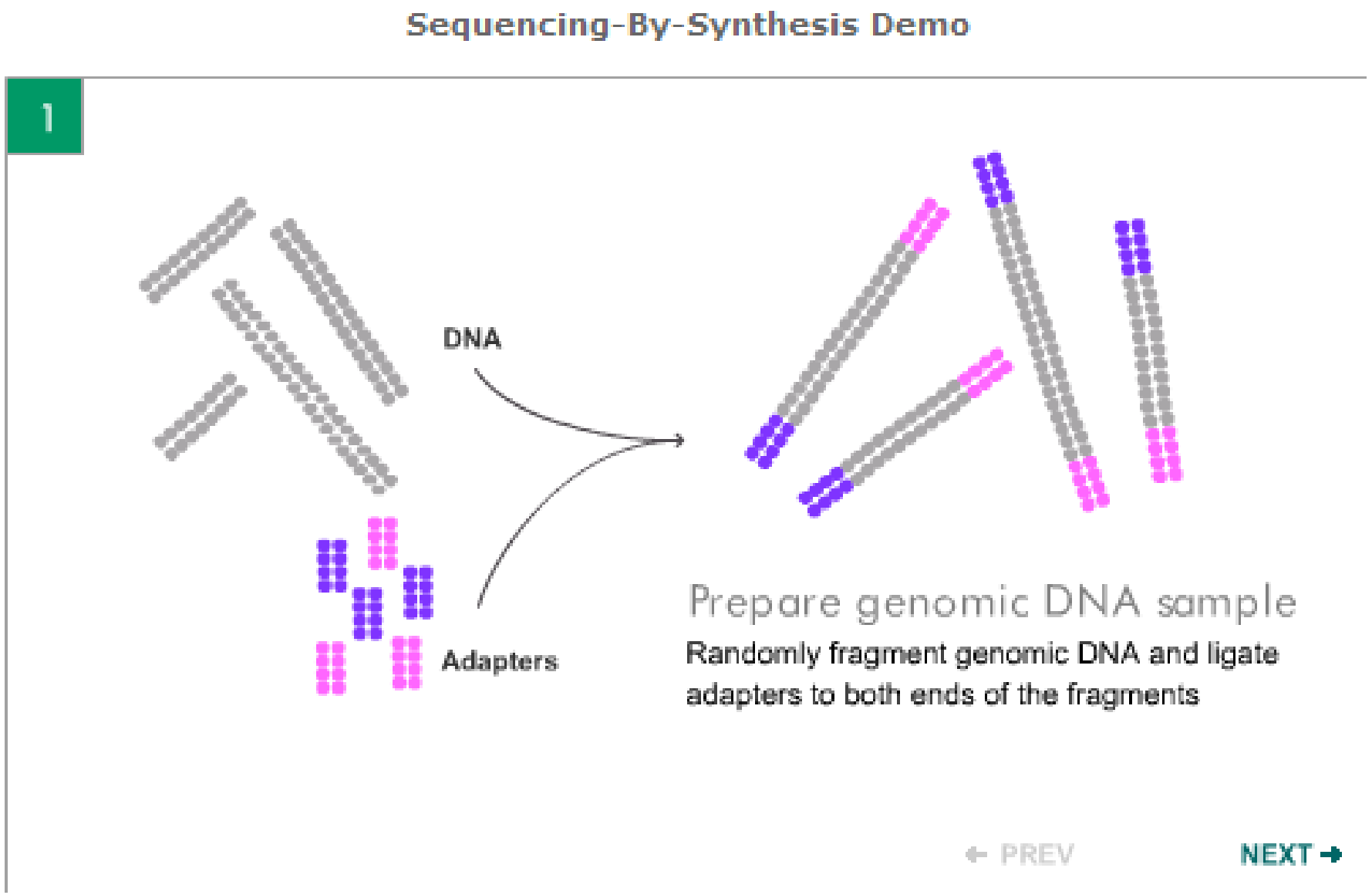
Library prep

Read 1
ATGCGC...
...GATTCA
Read 2

Massively parallel sequencing

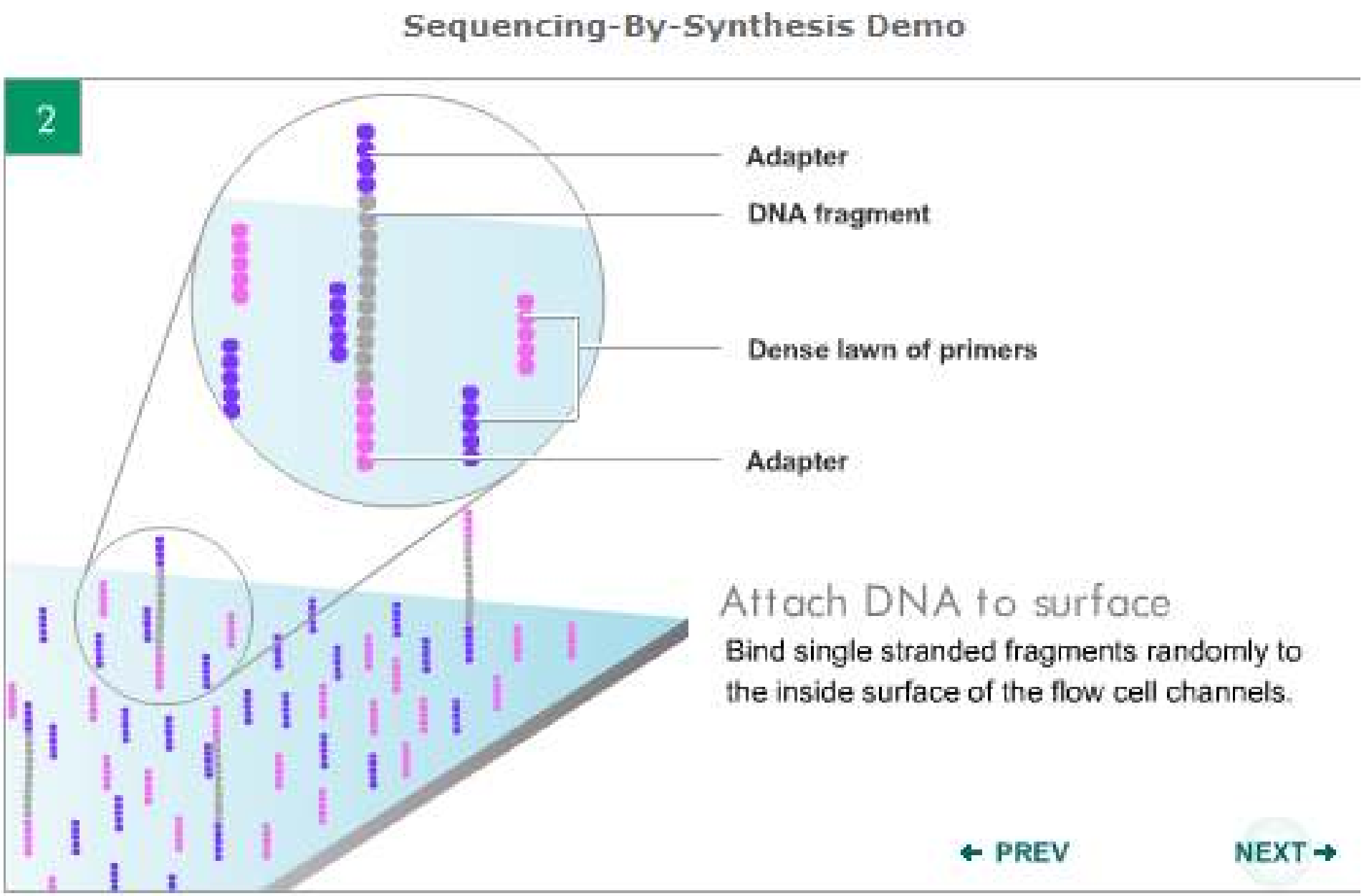
adapters

Illumina sequencing-by-synthesis

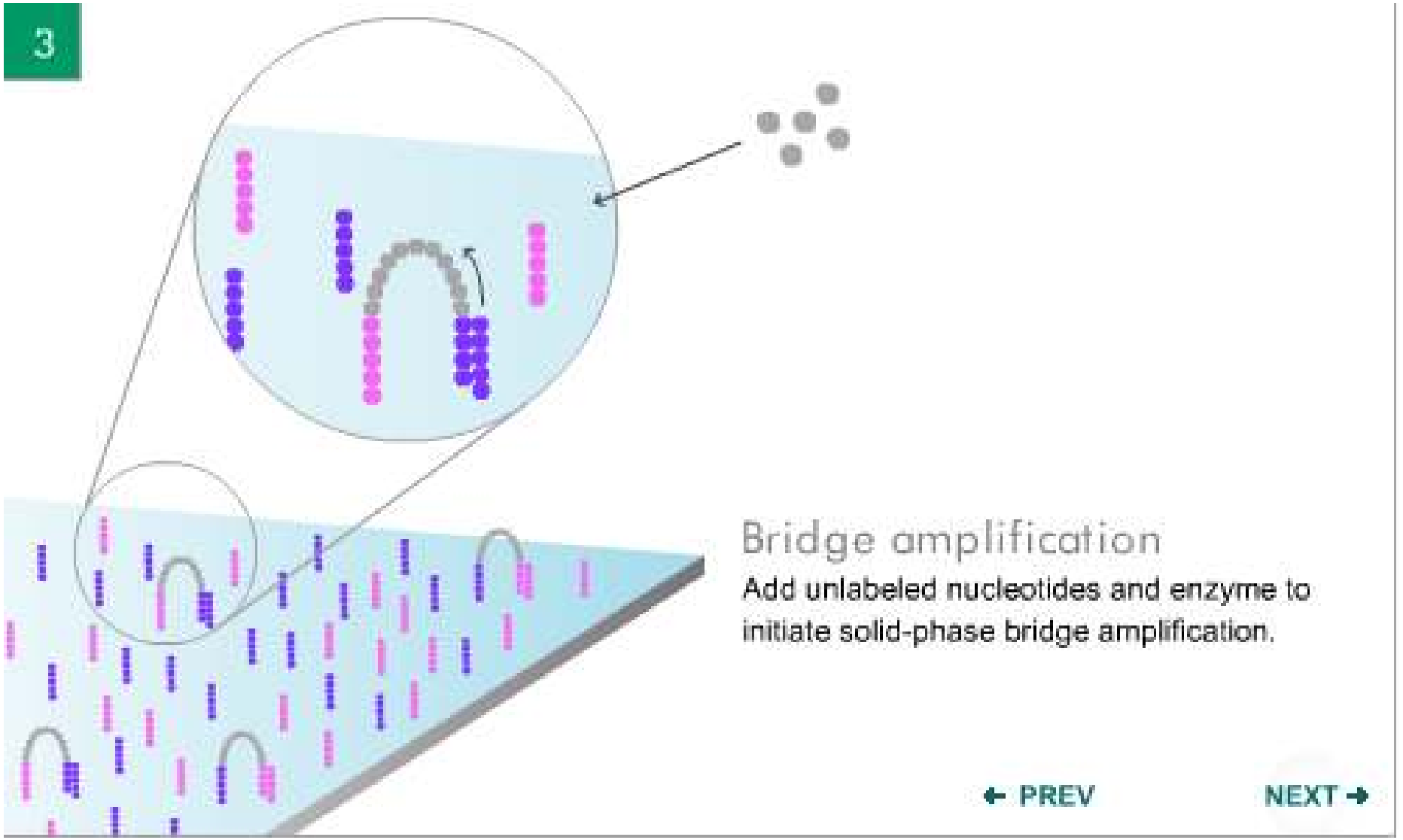


http://www.illumina.com/

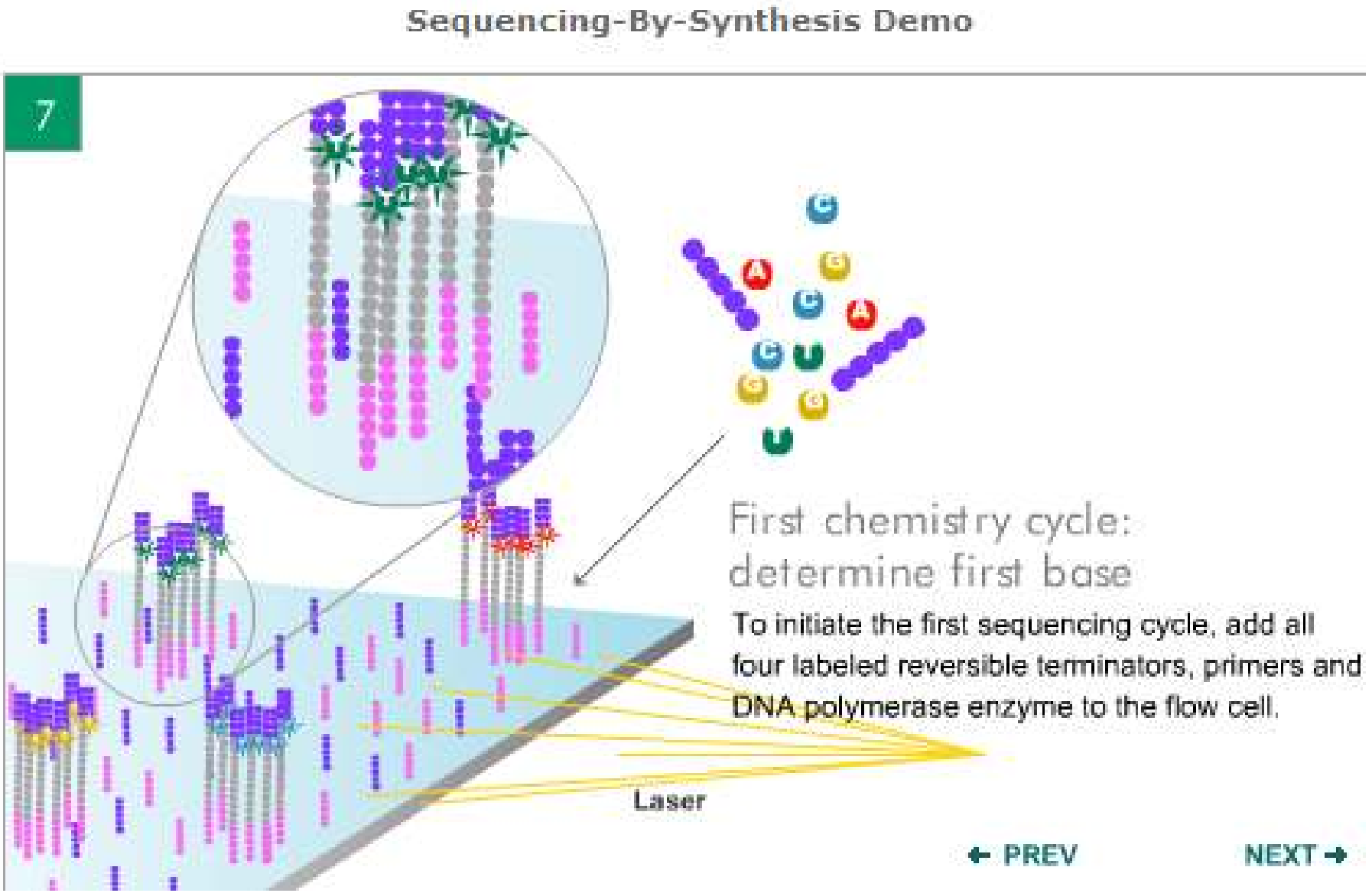
Illumina sequencing-by-synthesis



Illumina sequencing-by-synthesis



Illumina sequencing-by-synthesis



Illumina sequencing-by-synthesis

10

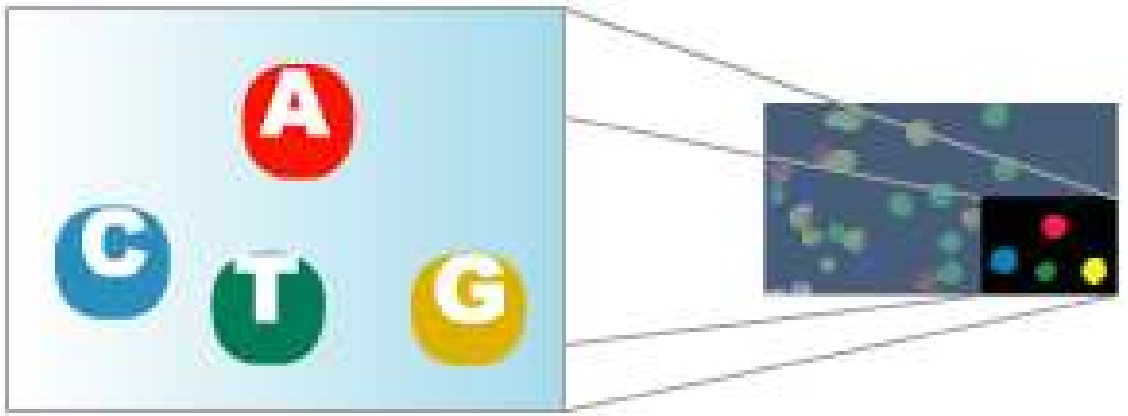
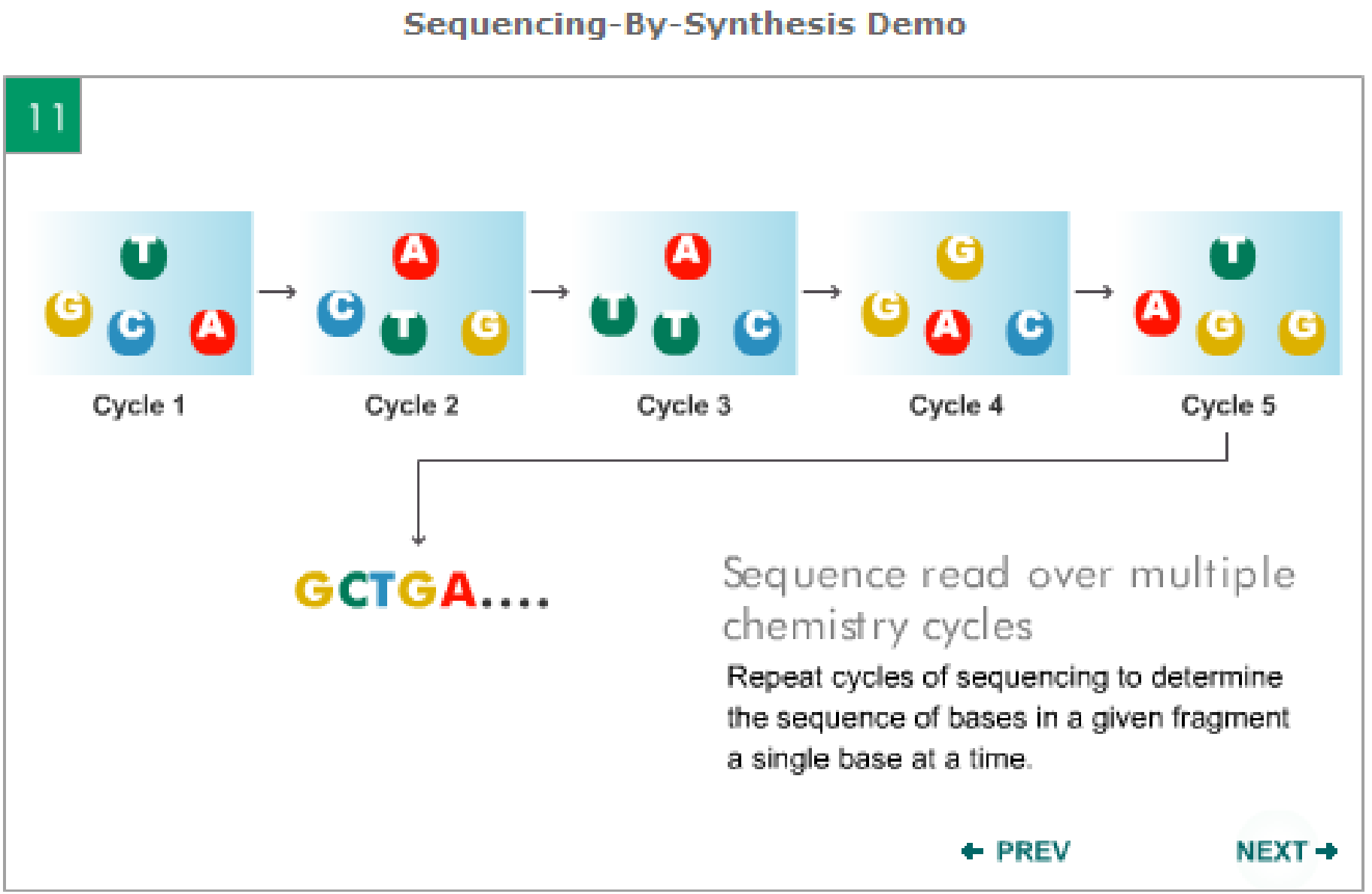


Image of second chemistry cycle is captured by the instrument

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

← PREV NEXT →

Illumina sequencing-by-synthesis



What is NGS ?

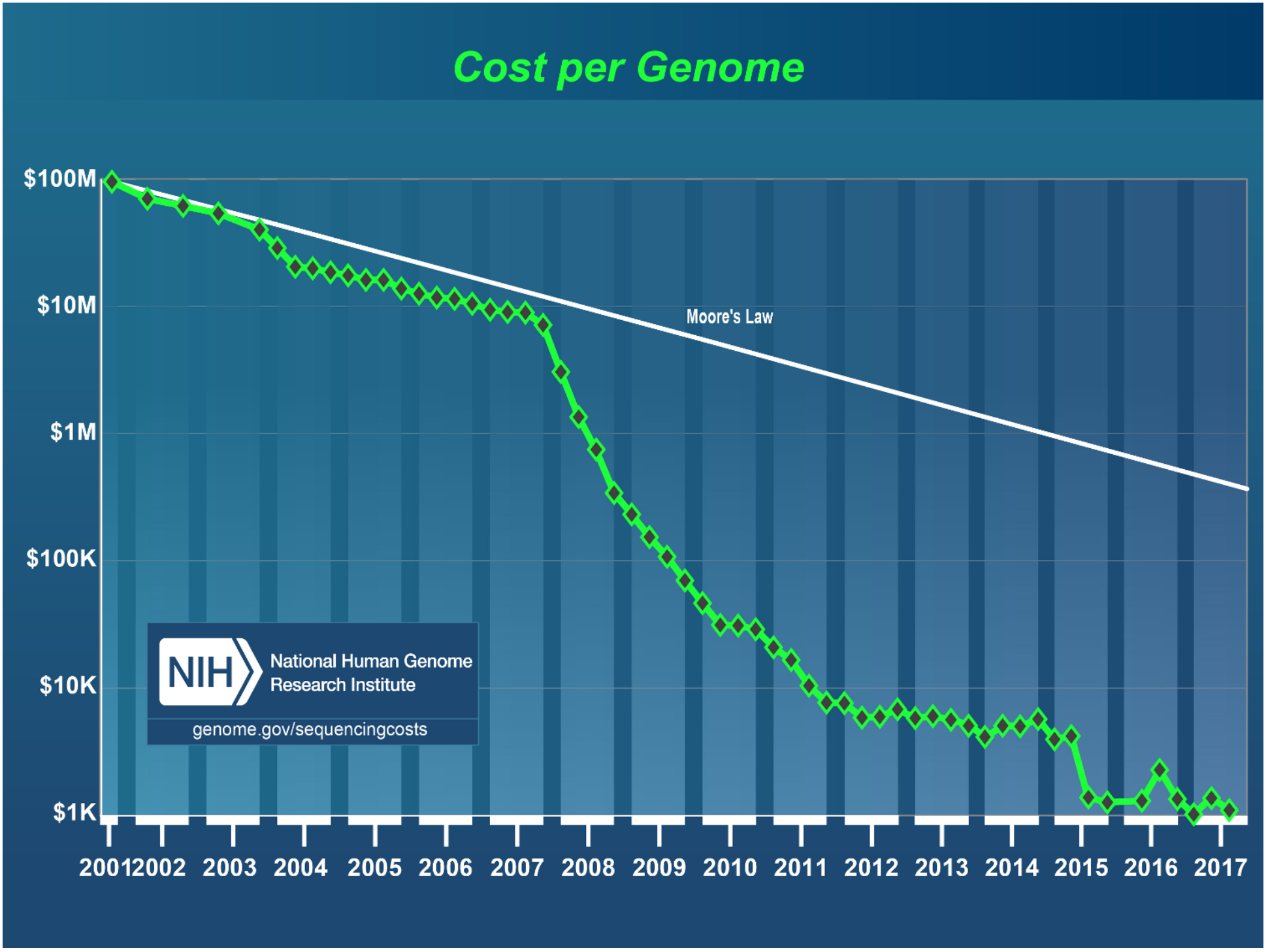
= High-throughput sequencing

Sanger sequencing

NGS

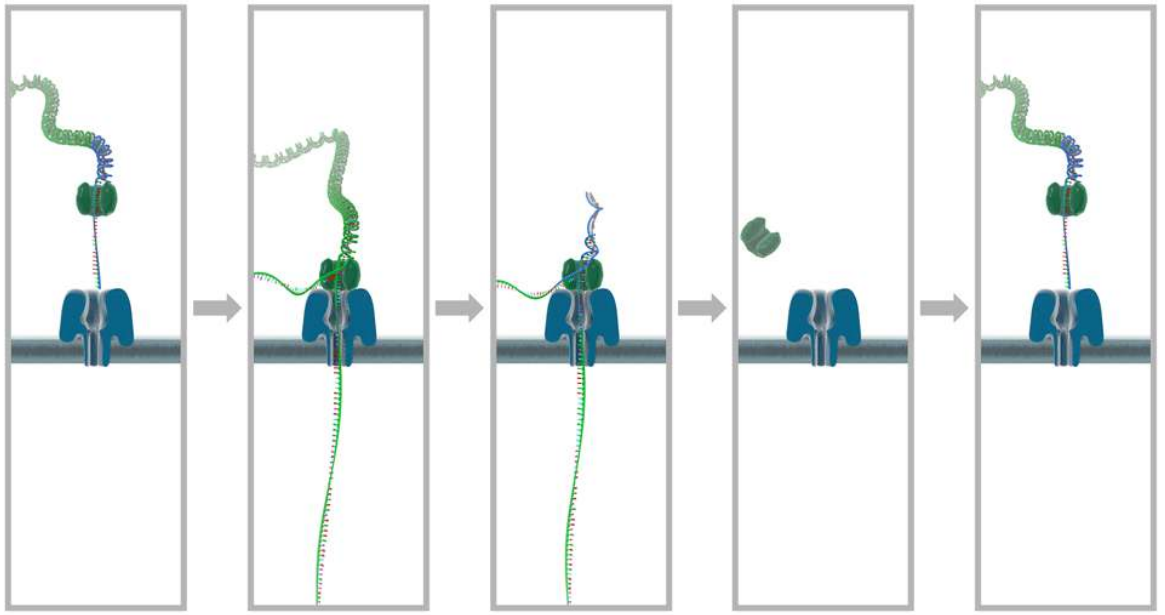
1975-85 2005

Sequencing costs

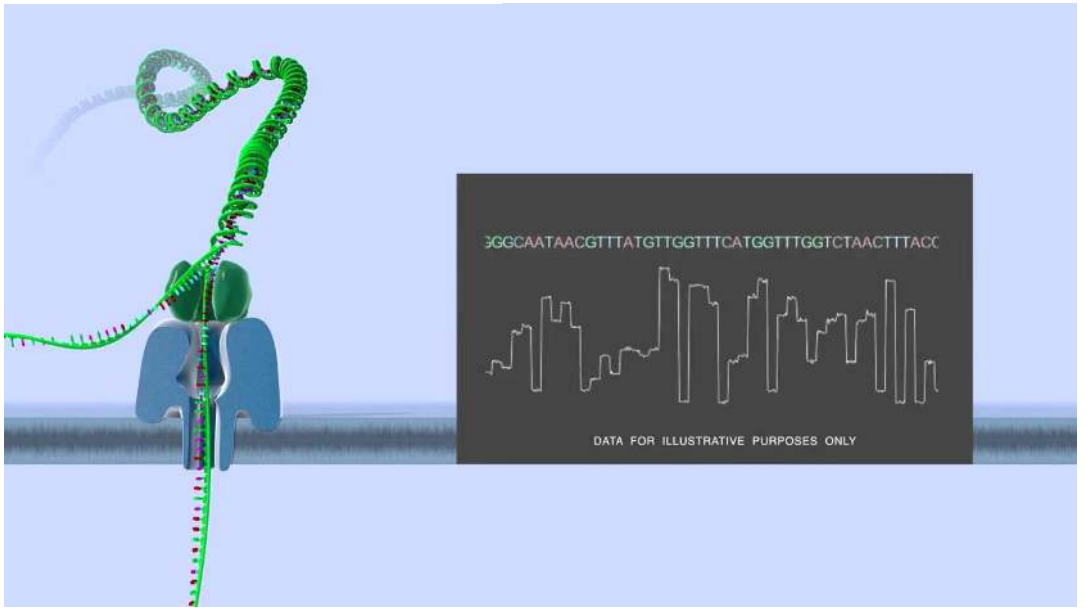


Nanopore sequencing

Long reads are still important (e.g. assembly, phasing, splicing, etc.)



Pollard et al., Hum Mol Genet, 2018



Oxford NANOPORE Technologies

Sequencing technologies

Platform	Sample preparation	Technology / chemistry	Read length	Yield	Quality	Cost
Illumina	Bridge amplification	synthesis + fluorescence	50-250	++ +++++	+++++	++++
SoliD	emulsion PCR	ligation + fluorescence	50-75	++++	+++	+++
Ion Torrent	emulsion PCR	pyrosequencing + pH	200-400	++	+++	+++
Nanopore	Single molecule	synthesis + fluorescence	DNA-dependent up to 500Kb	+	+	++
Pacific Bioscience	Single molecule	synthesis + fluorescence	30Kb->100 Kb	+++	+	+++

Further reading :

http://en.wikipedia.org/wiki/DNA_sequencing
Liu et al., J Biomed Biotechnol, 2012, ID 251364.
McGinn & Gut, N Biotechnol. 2013, 30(4):366-72.
Buermans & den Dunnen, Biochim Biophys Acta. 2014.
Shapiro E. et al. Nat Rev Genet. 2013, 14(9):618-30. Single-cell sequencing
Yang et al., J Nanosci Nanotechnol, 2013, 13(7):4521-38. Nanopore sequencing

NGS applications

Soon et al., Mol Syst Biol, 2013

Population Metagenomics

Protein-DNA / RNA ChIP / RIP / CLIP

RNA RNA / CAGE / GRO

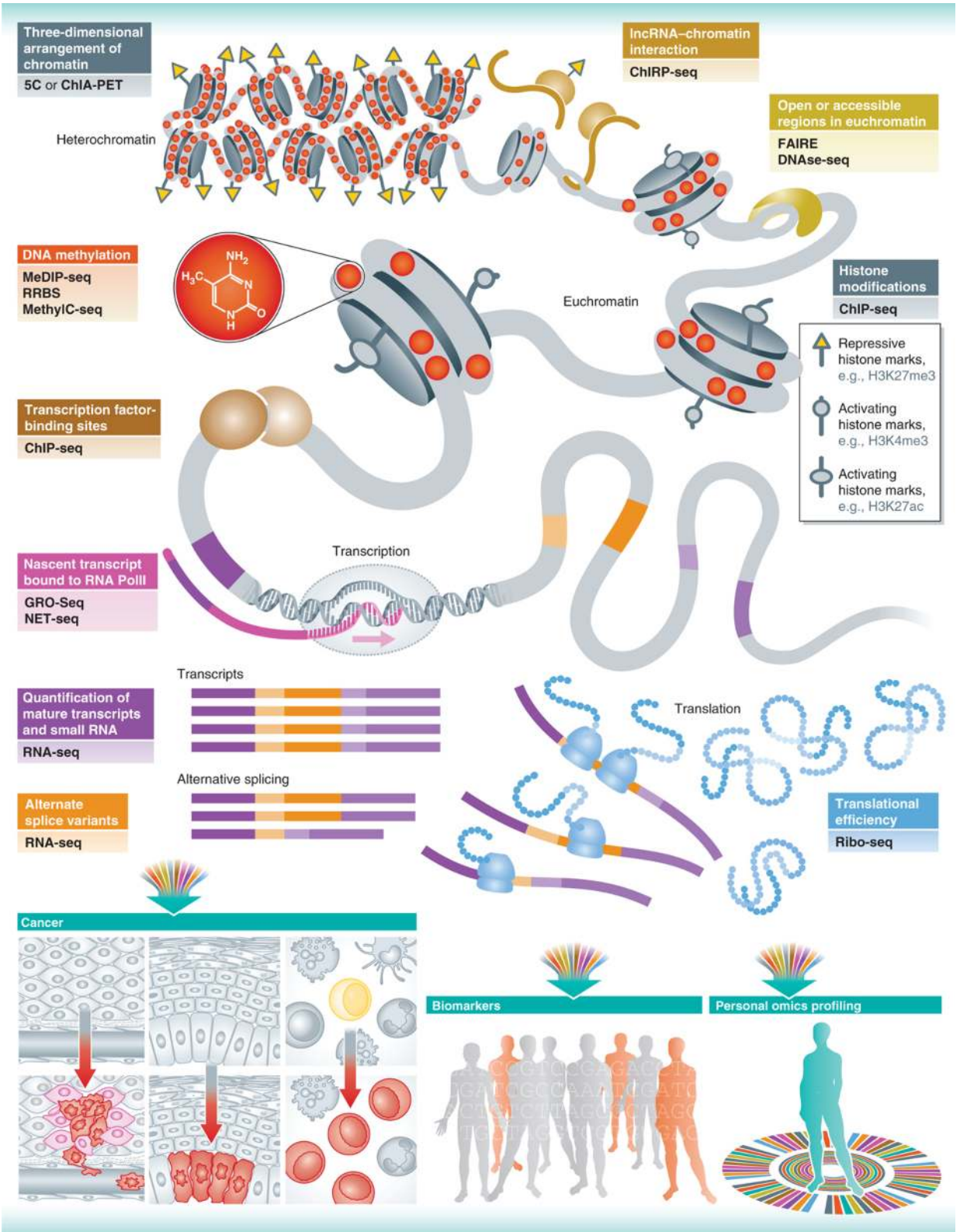
chromatin accessibility DNase / FAIRE / ATAC

chromatin structure MNase / ChIP / Hi-C

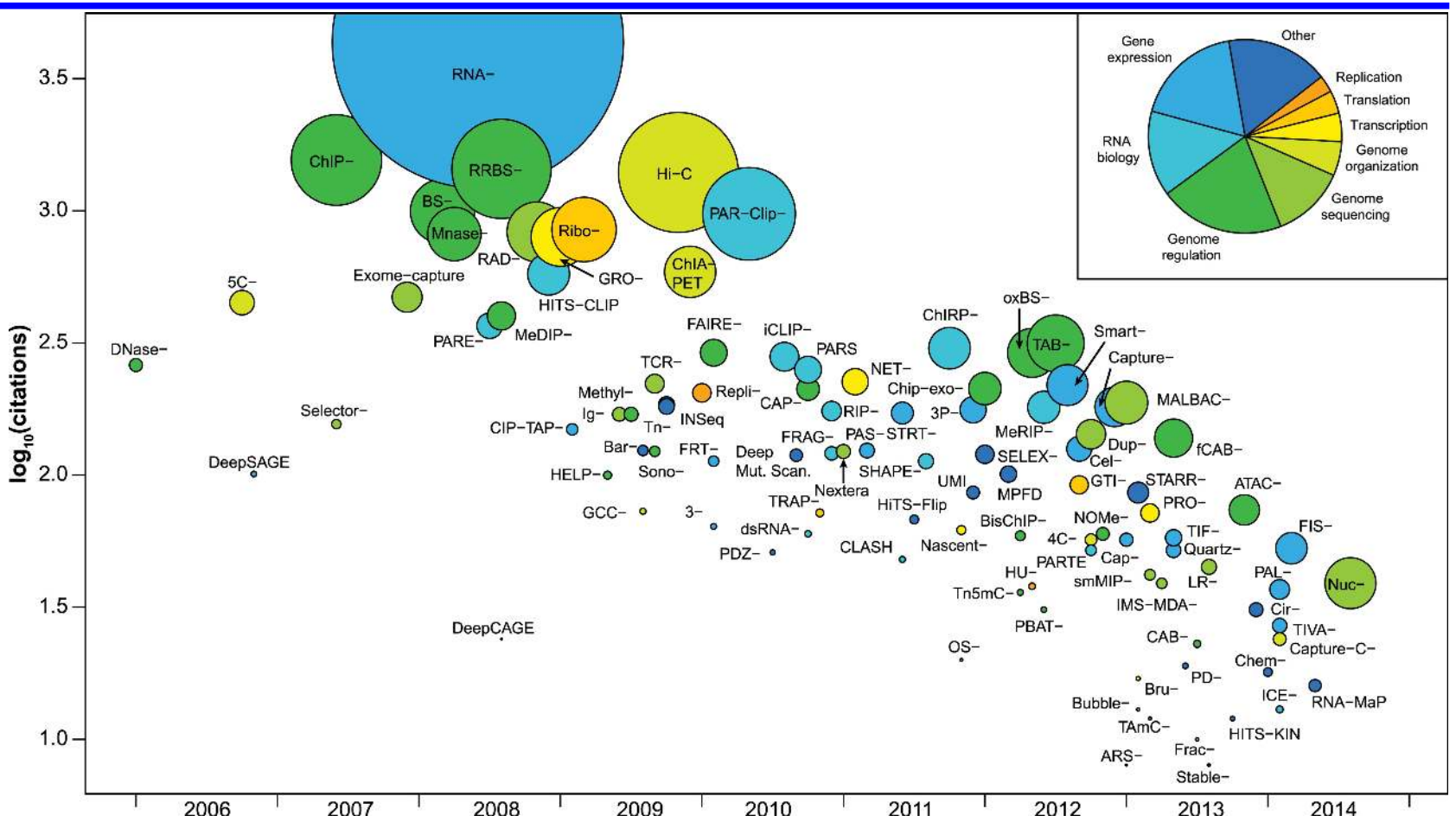
DNA methylation BS / MeDIP / RRBS

DNA DNA / exomes

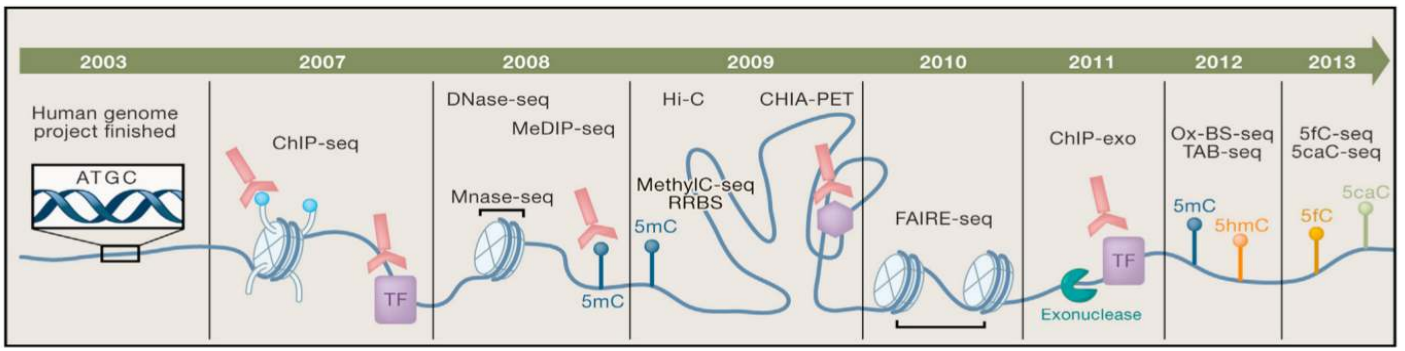
Many others ...



NGS applications



Reuter et al., Mol Cell, 2015



Rivera & Ren, Cell, 2013

See also : Shendure & Lieberman Aiden, Nat Biotechnol, 2012

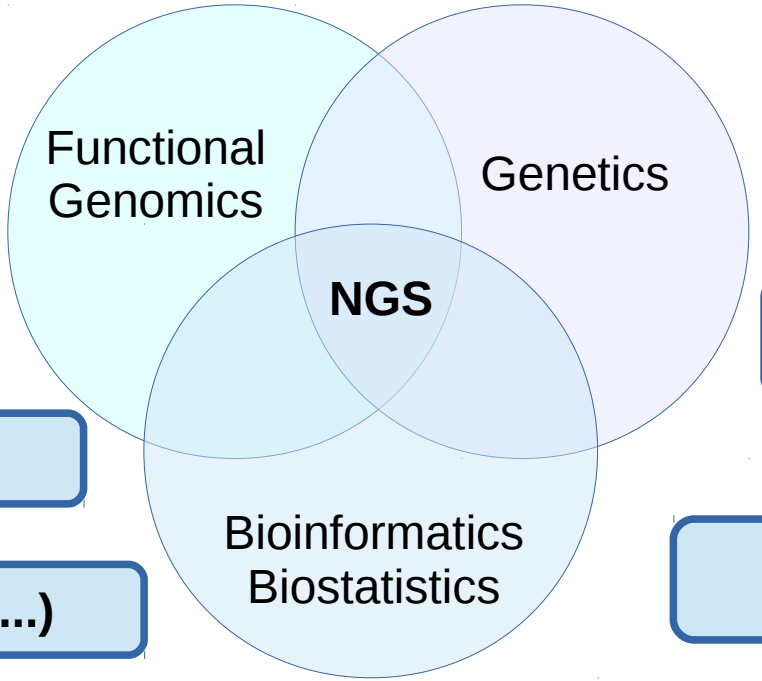
Multiple objectives, often complementary

Analyze structures (interactions, secondary structures,...)

Identify variations (SNPs, indels, CNV...)

Identify peaks

Annotate



Measure abundances

Search for motifs (TFs, ...)

Count events / screen

Find associations (QTL, GWAS, ...)

Compare groups (KO/wt...)

Cluster / order (states, sequences, individuals,...)

Predict / integrate (clinical data, other omics, ...

Model (systems biology)

etc...

What do the data look like ?

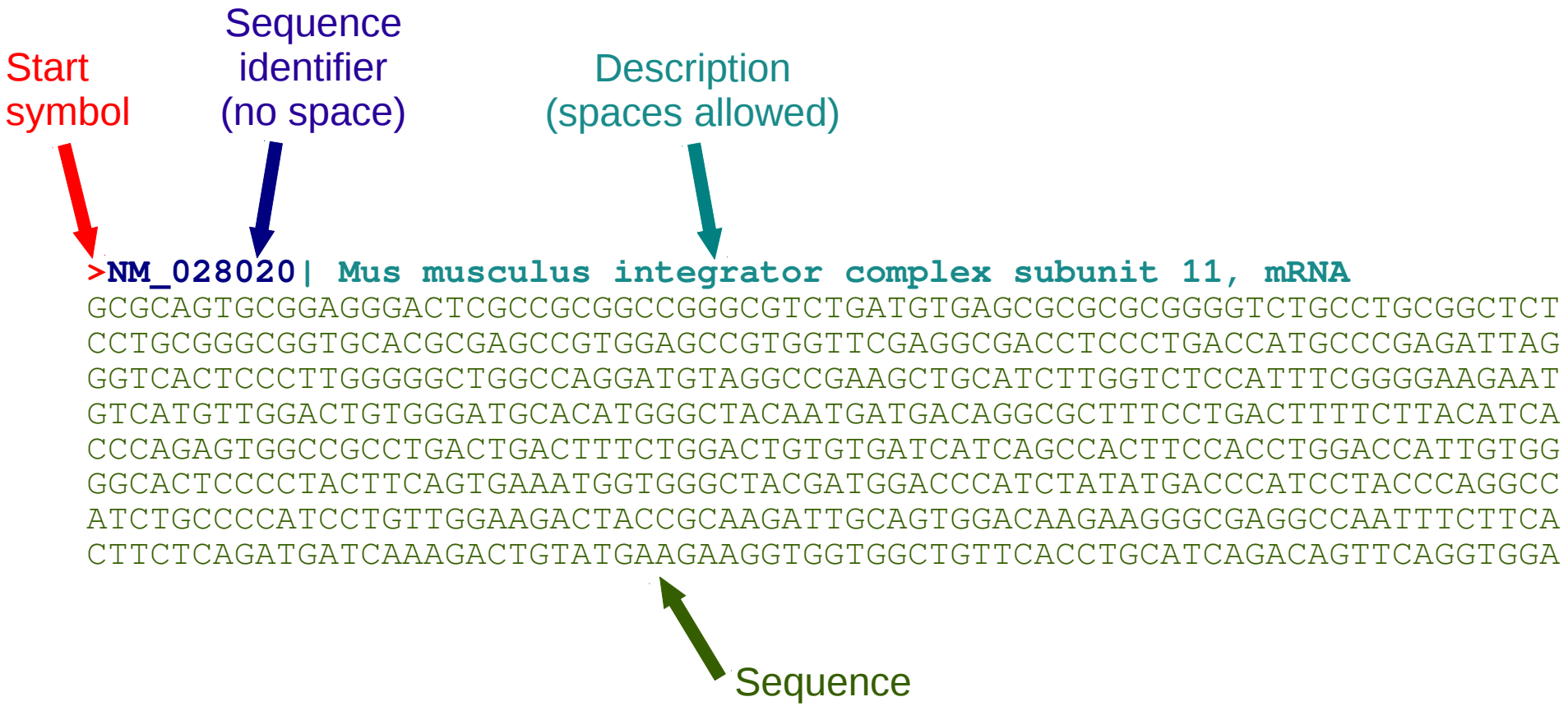
~50-200 bp

TCAGGTGCGCACTCTATGGC.....TAACGGGCTAAAATCAAGTCT
GACTAAACATGGCGCGTCAG.....ATGCGAACTACGTGCGAGGC
AGGACGACGGGTCGGGAAC.....ATAGCGGCGGGCCCATTGATT

Several million reads (large files!!)

TCAGGTGCGCACTCTATGGC.....GCCATGCGTAACATTAAGTCT
TTAGACCGTAACGGCTATTTG.....ATGCGAACTACGTGCGAGGC
AGGACGACGGGTCGGGAAC.....ATAGCGGCGGGCCCATTGATT

Fasta format



Extension	Alphabet	Utilisation
.fasta, .fa	any	Generic
.fna	DNA	Large genomic regions (contigs, chromosomes, ...)
.ffn	DNA	Coding sequence
.faa	Protein	Amino acid sequence
.frn	DNA	Non coding RNAs (e.g. tRNA, rRNA...)

http://en.wikipedia.org/wiki/FASTA_format

Multi-fasta

- Concatenation of fasta files
- The > sign separates sequences

>**Sequence1 tRNA A**
GCGCAGTGCGGAGGGACTCGCCGCGGCCGGGCGTCTGATGTGAGCGCGCGCGGGGTCTGCCTGCGGCTCT
CCTGCGGGCGGTGCACGCGAGCCGTGGAGC
>**Sequence2 protein coding RNA B**
GGTCACTCCCTTGGGGGCTATGCAGGATGTAGGCCGAAGCTGCATCTTGGTCTCCATTTCGGGGAAGAAT
GTCATGTTGGACTGTGGGATGCACATGGGCTACAATGATGACAGGCGCTTTCCTGACTTTTCTTACATCA
CCCAGAGTGGCCGCCTGACTGACTTTCTGGACTGTGTGATCATCAGCCACTTCCACCTGGACCATTGTGG
GGCACTCCCCTACTTCAGTGAAATGGTGGGCTACGATGGACCCATCTATATGACCCATCCTACCCAGGCC
ATCTGCCCCATCCTGTTGGAAGACTACCGCAAGATTGCAGTGGACAAGAAGGGCGAGGCCAATTTCTTCA
>**Sequence3 non coding RNA C**
CTTCTCAGATGATCAAAGACTGTATGAAGAAGGTGGTGGCTGTTCACCTGCATCAGACAGTTCAGGTGGA
TGATGAGCTGGAAATCAAGGCATACTATGCAGGCCACGTGCTGGGGGCAGCCATGTTCCAGATTAAAGTG
GGCTCGGAGTCTGTGGTCTACACGGGTGACTATAACATGACCCCAGACCGGCATTTGGGGGCTGCGTGGA
TTGACAAGTGTCGCCCCAACTTGCTCATCACAGAATCCACATATGCTACAACCATCCGAGATTCCAAACG
CTGCAGAGAGCGAGATTTCCTAAAGAAAGTTCATGAAACTGTGGAGCGTGGAGGAAAGGTGCTGATTCCT
GTGTTTGCACTGGGCCGAGCACAAGAACTCTGCATCCTGCTGGAGACCTTCTGGGAGCGCATGAACCTGA
AGGTGCCCATATACTTCTCTACGGGCCTGACAGAGAAAGCCAACCACTACTACAAGCTCTTCATCACCTG
GACCAACCAGAAGATCCGGAAGACATTTGTCCAGAGGAACATGTTTGAGTTTAAGCACATCAAAGCCTTT

Phred quality score

- A quality score for each nucleotide
- Measures the reliability of the base call (physical process: chromatogram, image, pH...)
- Formalized in the Phred software developed for the Human Genome Project

Ewing et al., Genome Res. 1998. 8(3):175-85 & 186-94

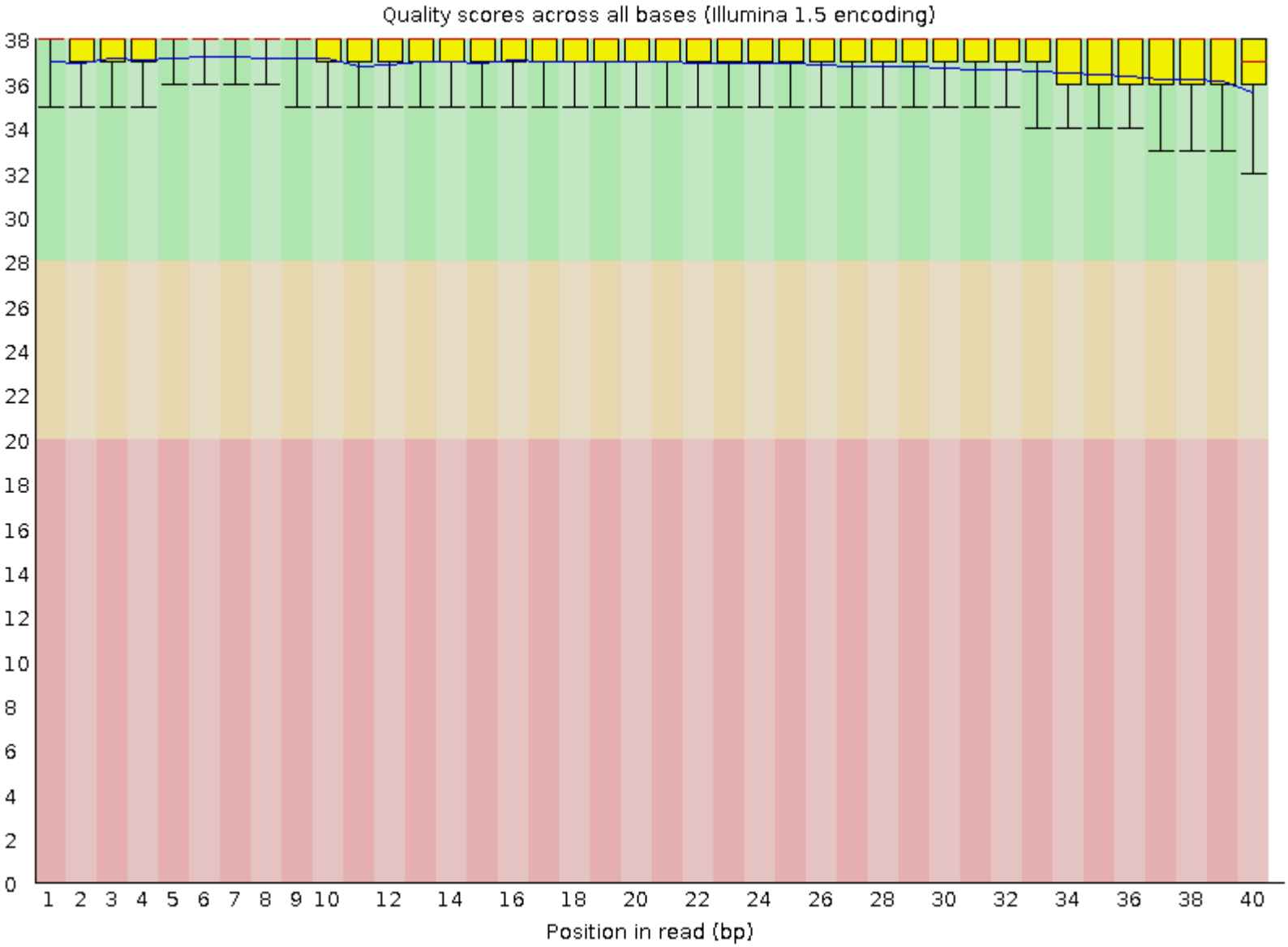
Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1 000	99,9 %
40	1 in 10 000	99,99 %
50	1 in 100 000	99,999 %

$$Q = -10 \times \log_{10}(P) \iff P = 10^{-Q/10}$$

- Q : Phred quality score
- P : Probability of base calling error

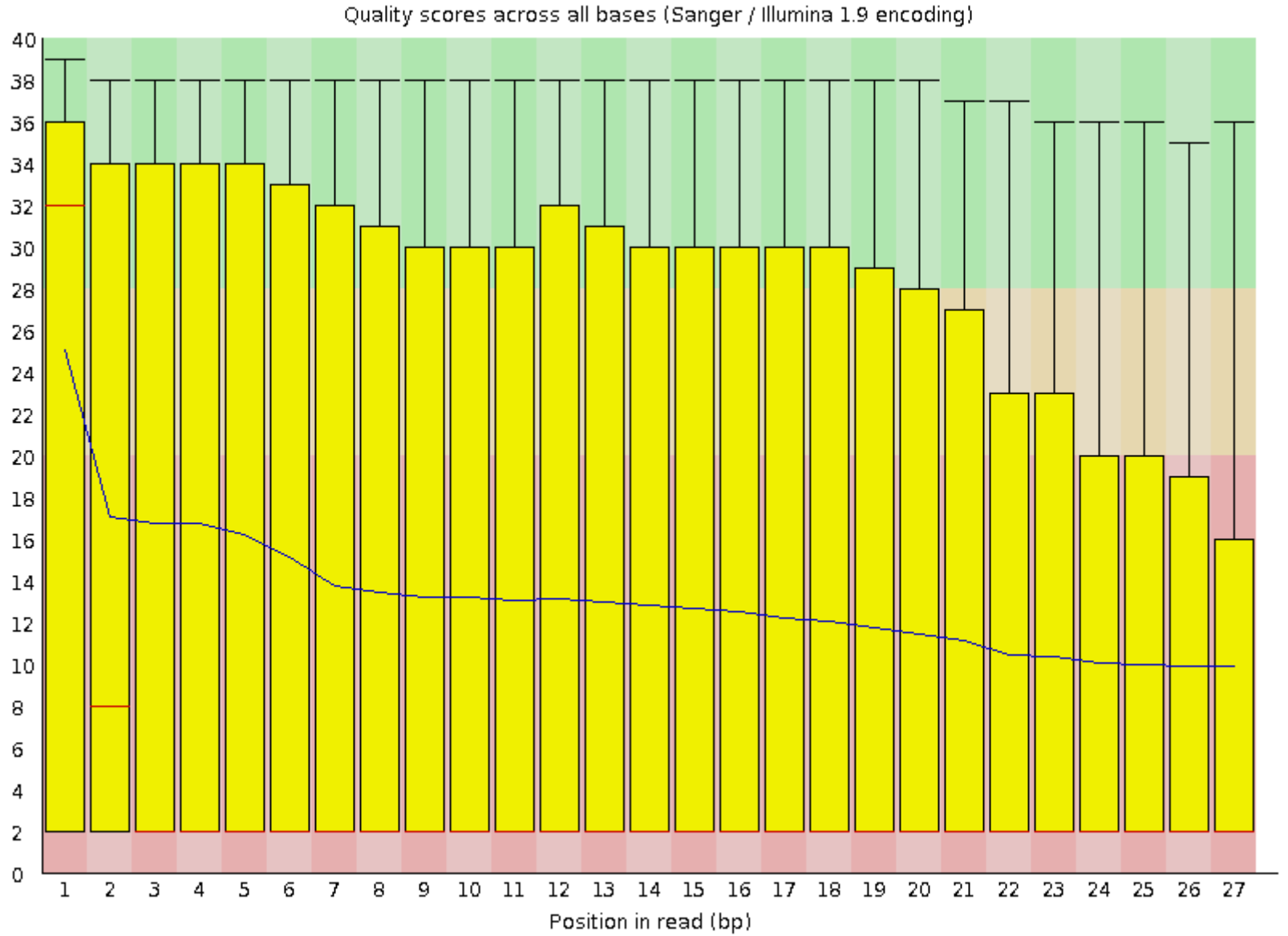
Quality

- Exemple of high quality reads (Illumina)



Quality

- Low quality reads (CAGE-seq, Illumina GAIIX)

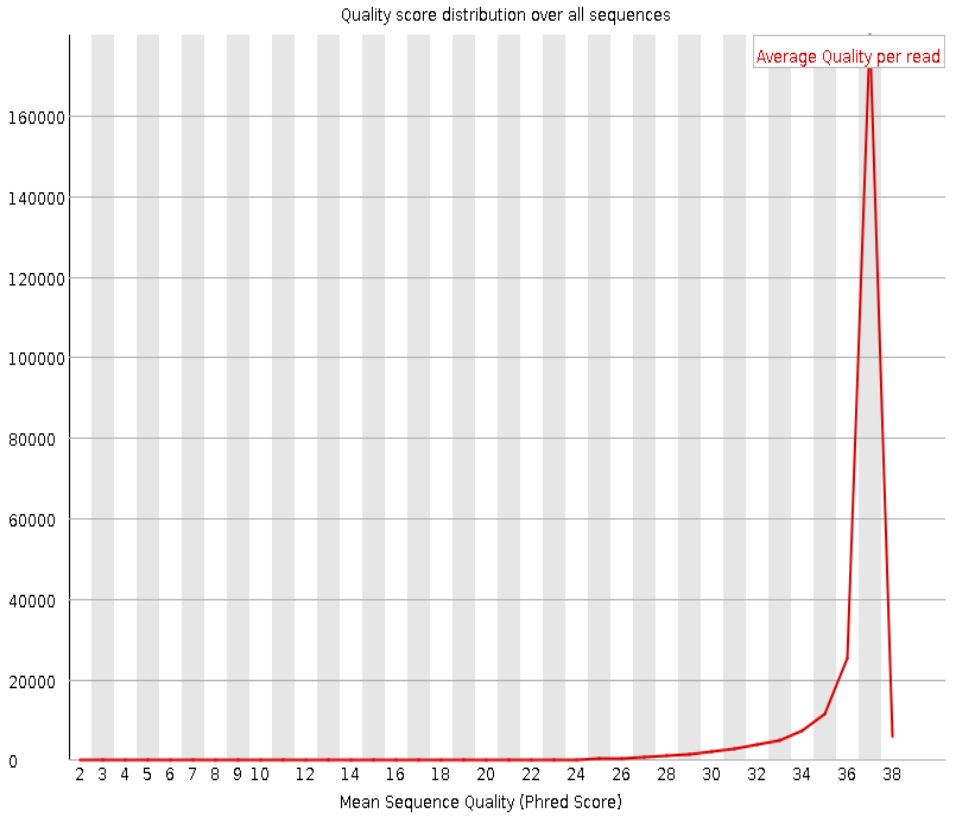


Read quality

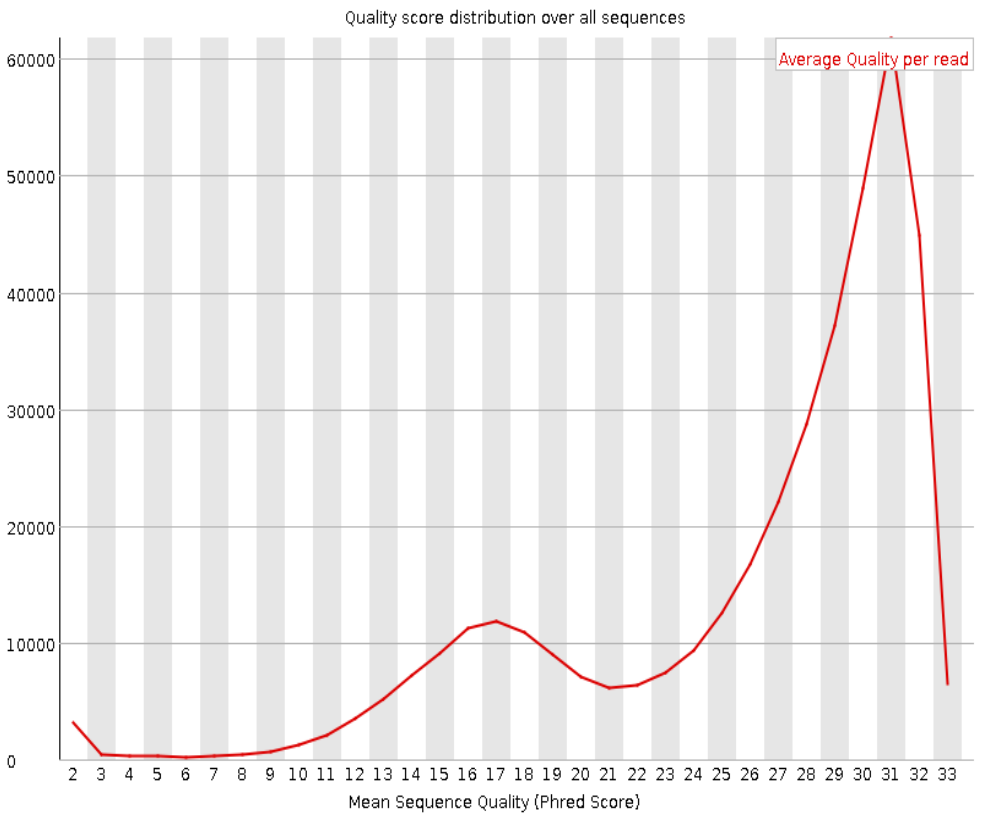
- Average phred score used to evaluate quality of entire read
- Different from alignment quality !!

Li et al., Genome Res. 2008. 18(11):1851-8. MAQ

Excellent quality (Illumina)



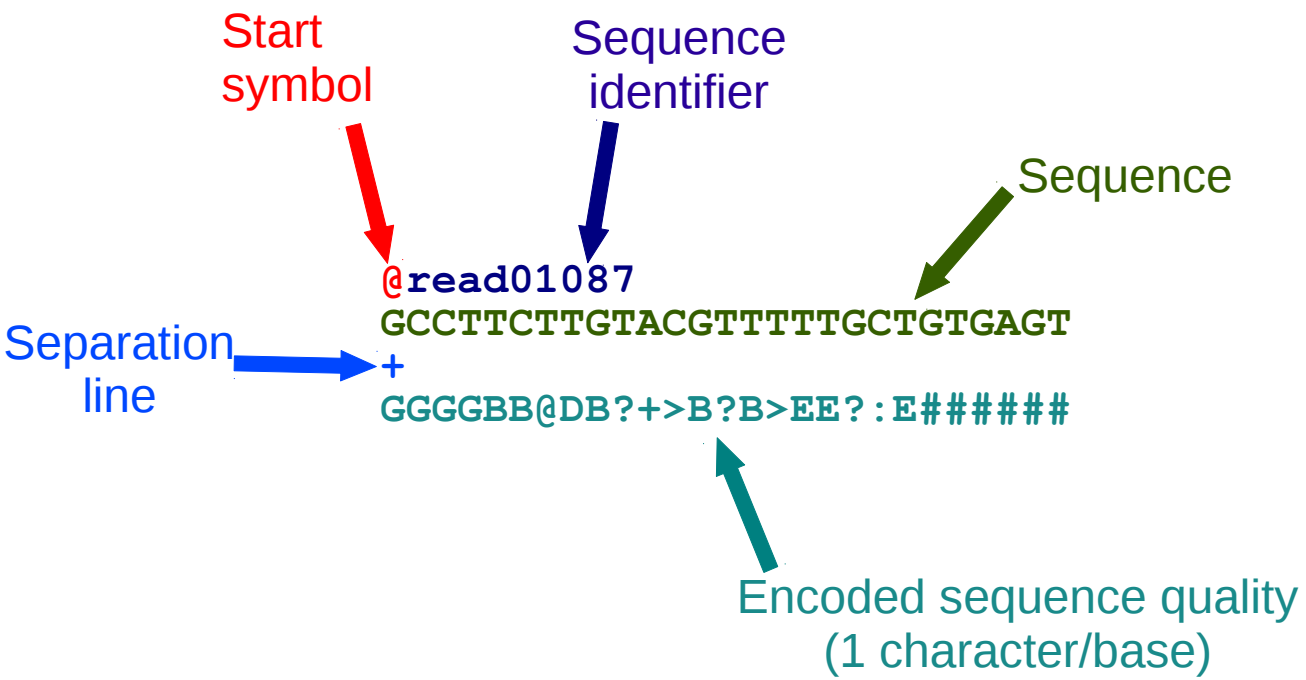
Low quality (Illumina)



Fastq format

http://en.wikipedia.org/wiki/FASTQ_format

File extension: .fq or .fastq



Sequence quality encoding (ASCII, Phred+33):

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
|         |         |         |         |
Q0        Q10       Q20       Q30       Q40
bad       average   ok        good      excellent
```

Multi-fastq

- Simple concatenation

```
@SRR488286.8911073
GCCTTCTTGTACGTTTTTGCTGTGAGT
+
GGGGBB@DB?+>B?B>EE?:E######
@SRR488286.10295868
GCTTTCGTTCTCAAACCGTCGNCCAGA
+
G>GGGG:>?A;GGGGGB?;BB(BB??2
@SRR488286.2166439
GAGTTGGGCTTCCACCTCGACCGGGAA
+
C@<@B0:<'87;B0;B;;52;;06=:C
@SRR488286.8412402
GAGTAAAACATACAACTTAAACGAGAA
+
A??;*A:>@.<@(@############
```

- Often very large files (>10Go)
- Contain redundancy
- Compression (.gz, .bz2) => size often reduced to <20 % of orginal size
- Saves disk space and accelerate transfers but reading/writing is slower

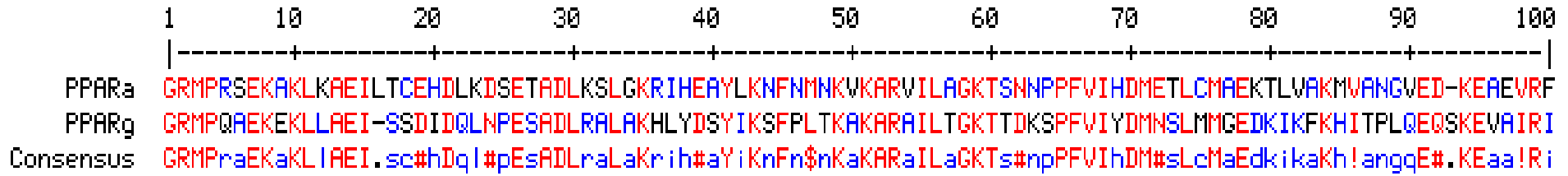
Alignment

- We have :
  - Millions of reads in fastq file
  - A reference sequence in a fasta file

Where do the reads come from?
Alignment problem

What is an alignment ?

Sequence alignment on Wikipedia



- "Optimal" alignment / adjustment of 2 sequences.
- Match/mismatch ? gaps/extensions ? What criterion defines an optimal alignment ?

Alignment : edit distance

The edit distance (Levensthein, 1965) is the minimum number of characters that need to be replaced (mismatch), inserted or suppressed (gap/extension) to go from one character string to the other.

- Insertion (i)
- Deletion (d)
- Substitution (s)

```
TCGCGTA-CGTACG
-CGCCTACCGGA-G
d   s   i   s d
```

Edit distance between query (q) and subject (s) : **D(q,s)=5**

- Calculating the edit distance is not trivial.
  → Dynamic programming
- The algorithm provides the optimal alignment between 2 sequences

Pairwise and multiple alignement

➢ Pairwise alignment (2 séquences) : Blast (n, p,…), blast2seq, Needle, Water...

```
query    1   GRMPRSEKAKLKAEILTCEHDLKDSETADLKSLGKRIHEAYLKNFNMNKVKARVILAGKTSNNPPFVIHDMETLCMAEKTLVAKMVANGVED-KEAEVRF  99
subject  1   GRMPQAEKEKLLAEI-SSDIDQLNPESADLRALAKHLYDSYIKSFPLTKAKARAILTGKTTDKSPFVIYDMNSLMMGEDKIKFKHITPLQEQSKEVAIRI  99
Consensus    GRMP++EK KL AEI + + D  + E+ADL++L K ++++Y+K+F + K KAR IL GKT++  PFVI+DM +L M E  +  K +    E  KE  +R
```

- A **query** sequence vs a **subject** sequence
- Blast aligns 1 query vs thousands of subjects (database)
- provides an e-value (or E score) the number of hits one can "expect" to see by chance when searching a database of a particular size (background noise). See BLAST tutorial.
- NGS aligner align millions of queries to a few references

➢ Multiple alignment (>2 sequences) : Multalin, ClustalW, Muscle, T-coffee ...

```
           1       10        20        30        40        50        60        70        80        90       100
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
PPARa      GRMPRSEKAKLKAEILTCEHDLKDSETADLKSLGKRIHEAYLKNFNMNKVKARVILAGKTSNNPPFVIHDMETLCMAEKTLVAKMVANGVED-KEAEVRF
PPARb      GRMPEAEKRKLVAGLTASEGCQHNPQLADLKAFSKHIYNAYLKNFNMTKKKARSILTGKSSHNAPFVIHDIETLWQAEKGLVWKQLVNGLPPYNEISVHV
PPARg      GRMPQAEKEKLLAEI-SSDIDQLNPESADLRALAKHLYDSYIKSFPLTKAKARAILTGKTTDKSPFVIYDMNSLMMGEDKIKFKHITPLQEQSKEVAIRI
Consensus  GRMP.aEK.KL.Aei..s#.dq.#p#.ADLkal.Khiy#aYlKnFn$tK.KAR.ILtGKts.n.PFVIhDm#tL.maEk.lv.K...ng.e..kE..!r.
```

Global vs Local alignment

- Global alignement:

FTFTALILLAVAV
F--TAL-LLA-AV

  - Alignment on the whole length of the sequences
  - Used for sequences of similar length
  - Needlemann-Wunsh algorithm

- Local alignment:

FT**FTALILL**-AVAV
FTAL-LL

  - Only keeps the parts of the 2 sequences that match
  - Well suited to search conserved protein domains
  - Smith-Watermann algorithm

- Glocal or semi-local alignment:

FT**FTALILL-AV**AV subject
FTAL-LLAAV query

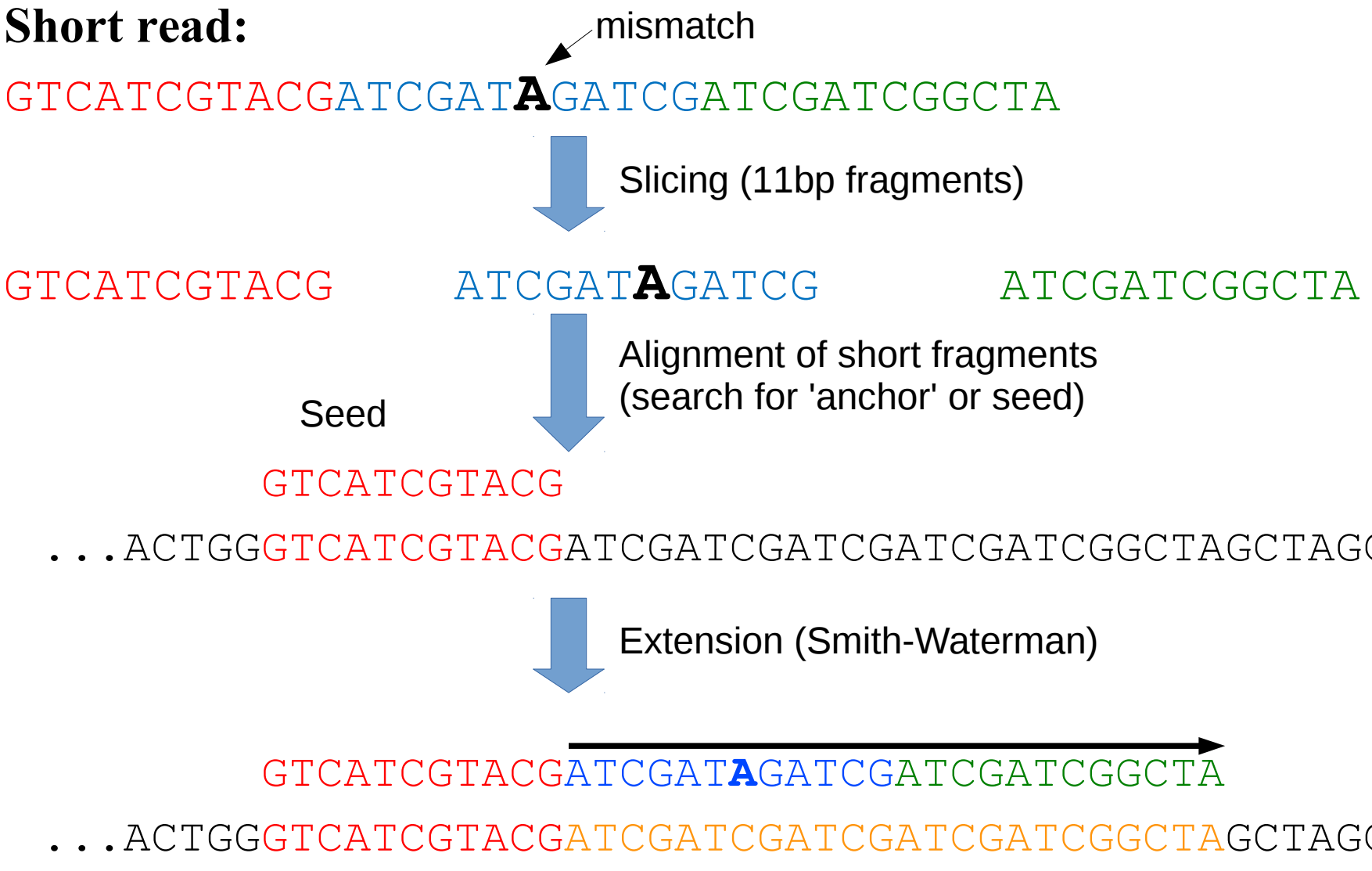
  - global for the **query** and local for the **subject**

Alignment : seed-and-extend

Reference sequence :

...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...

Short read:



NGS alignment

- Data volume imposes high-performance tools

  (computation, storage, algorithms, code)

- The reads have subtle differences compared to the reference genome

  (reference is assembled from multiple sequences, SNP, evolution, sequencing errors...)

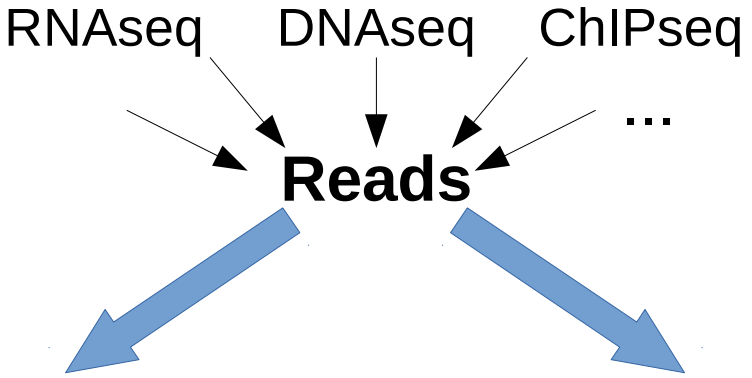
  ➡ Can't use exact string matching

- The genome contains a lot of sequence repeats

  ➡ Some reads can align at multiple positions (up to >50 % sometimes...)

- Some reads do not align to the reference genome (primers, contaminations...)

- Often complementary strategies can be used :



Alignement on a reference ⟺ de novo **assembly**

NGS alignement - tools

Li & Homer, Brief Bioinform, 2010 ; Fonseca et al., Bioinformatics, 2012

➢ Algorithms based hash tables

- ELAND
- BFAST
- MAQ (Li et al., Genome Res, 2008)
- SHRIMP, ...

➡ An « index », obtained on reads and/or the reference

➡ Mapping quality scores.
(ex : Ruffalo et al., Bioinformatics, 2012)

Phred scores
Mismatchs/indels
Repeats
Paired reads

➡ Use 'Seed-and-extend' concept

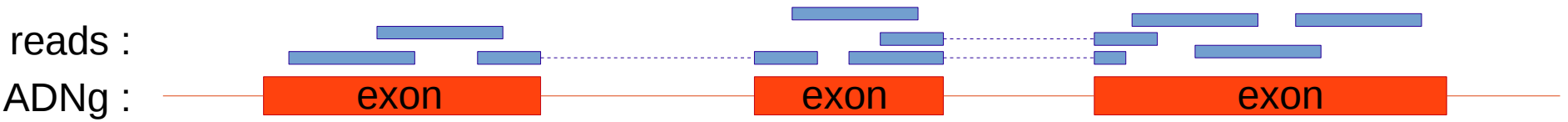
➢ Algorithms based on Burrows-Wheeler transform

➡ reversible recoding of sequences that optimises their storage

- Bowtie (Langmead et al., Genome Biol., 2009) / Bowtie2 (Langmead & Slazberg, Nat Methods, 2012)
- BWA (Li & Durbin, Bioinformatics, 2009 ; Li & Durbin, Bioinformatics, 2010)

➢ Spliced reads aligners (RNA-seq)

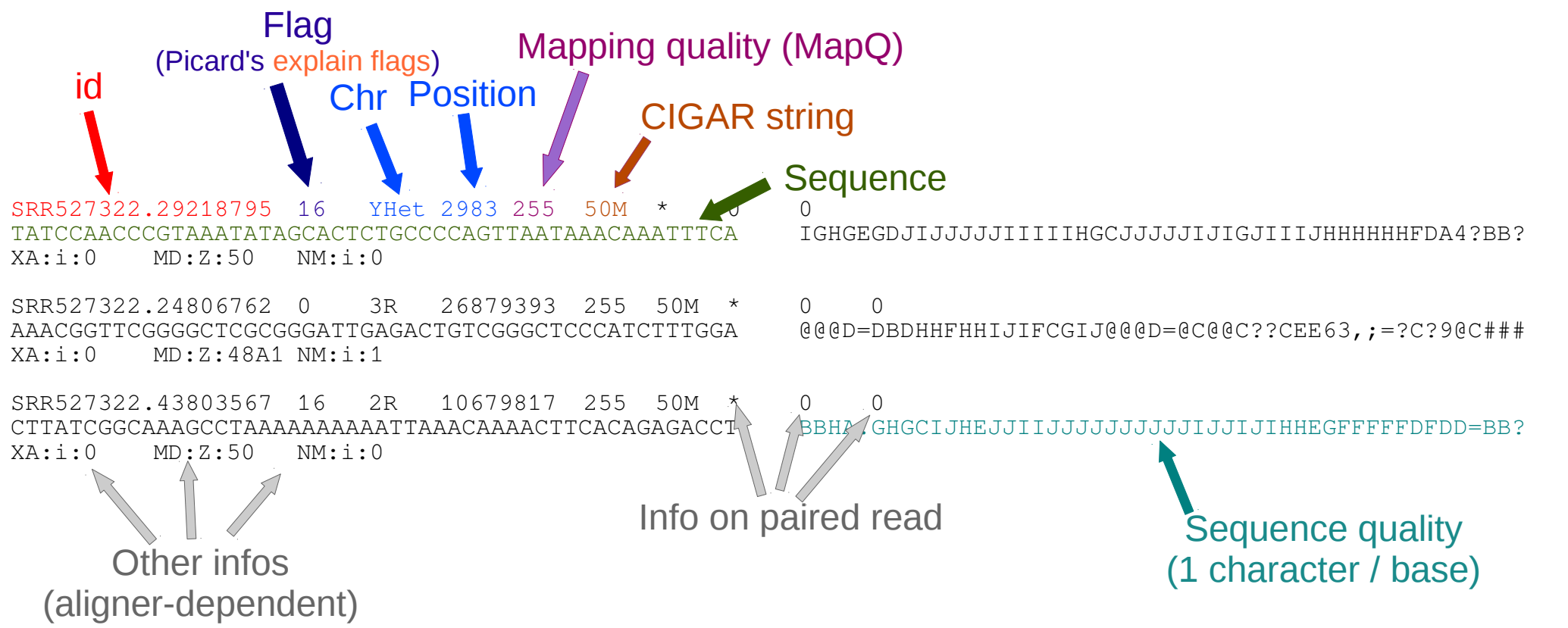
- HISAT2 (replaces TopHat2 Kim et al., Genome Biol, 2013, Pertea et al. Nat Protoc 2016)
- STAR (Dobin et al., 2013)



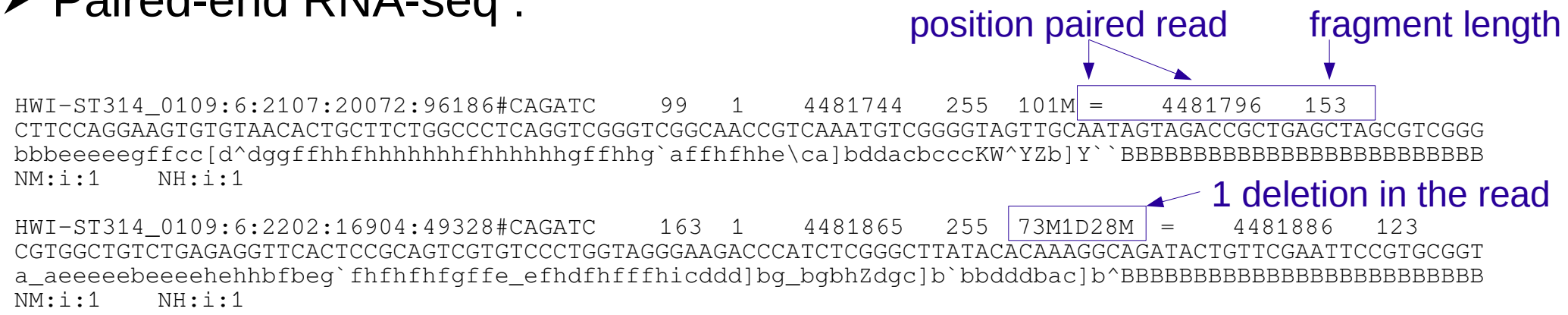
See also other alignment softwares on Wikipedia

SAM / BAM file format

http://samtools.github.io/
Li et al., Bioinformatics, 2009
See also: SAMtools, Picard, genome.sph.umich.edu/wiki/SAM



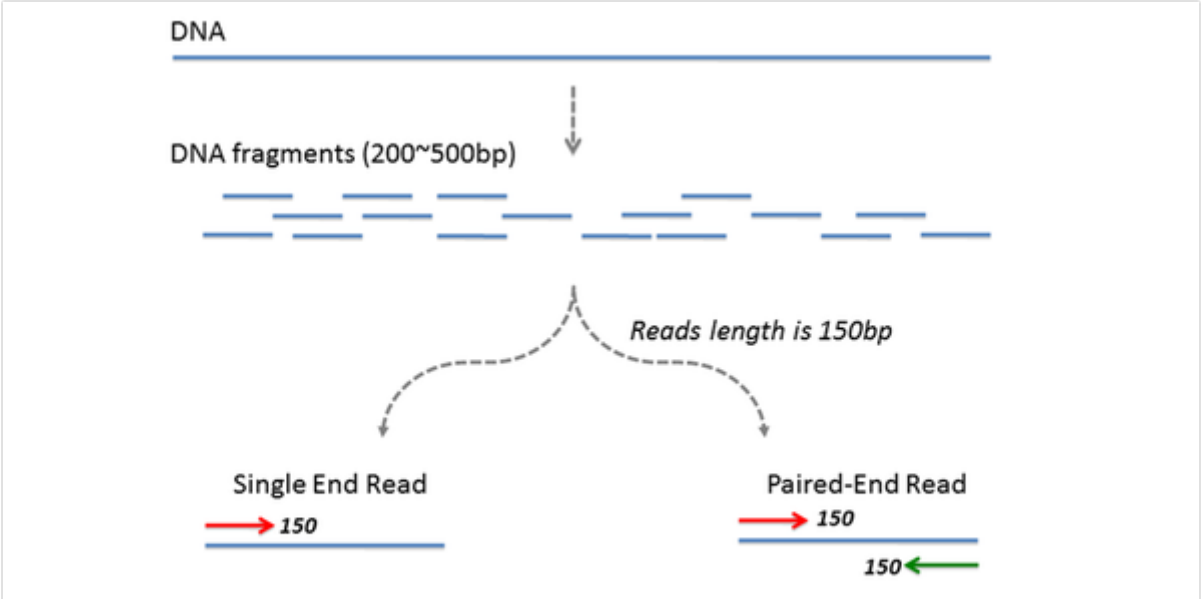
- Paired-end RNA-seq :



SAM / BAM file format

http://samtools.github.io/
Li et al., Bioinformatics, 2009
See also: SAMtools, Picard, genome.sph.umich.edu/wiki/SAM

- Paired-end reads (RNA-seq) :



id
Flag
Chr
Position
MapQ
position paired read
fragment length

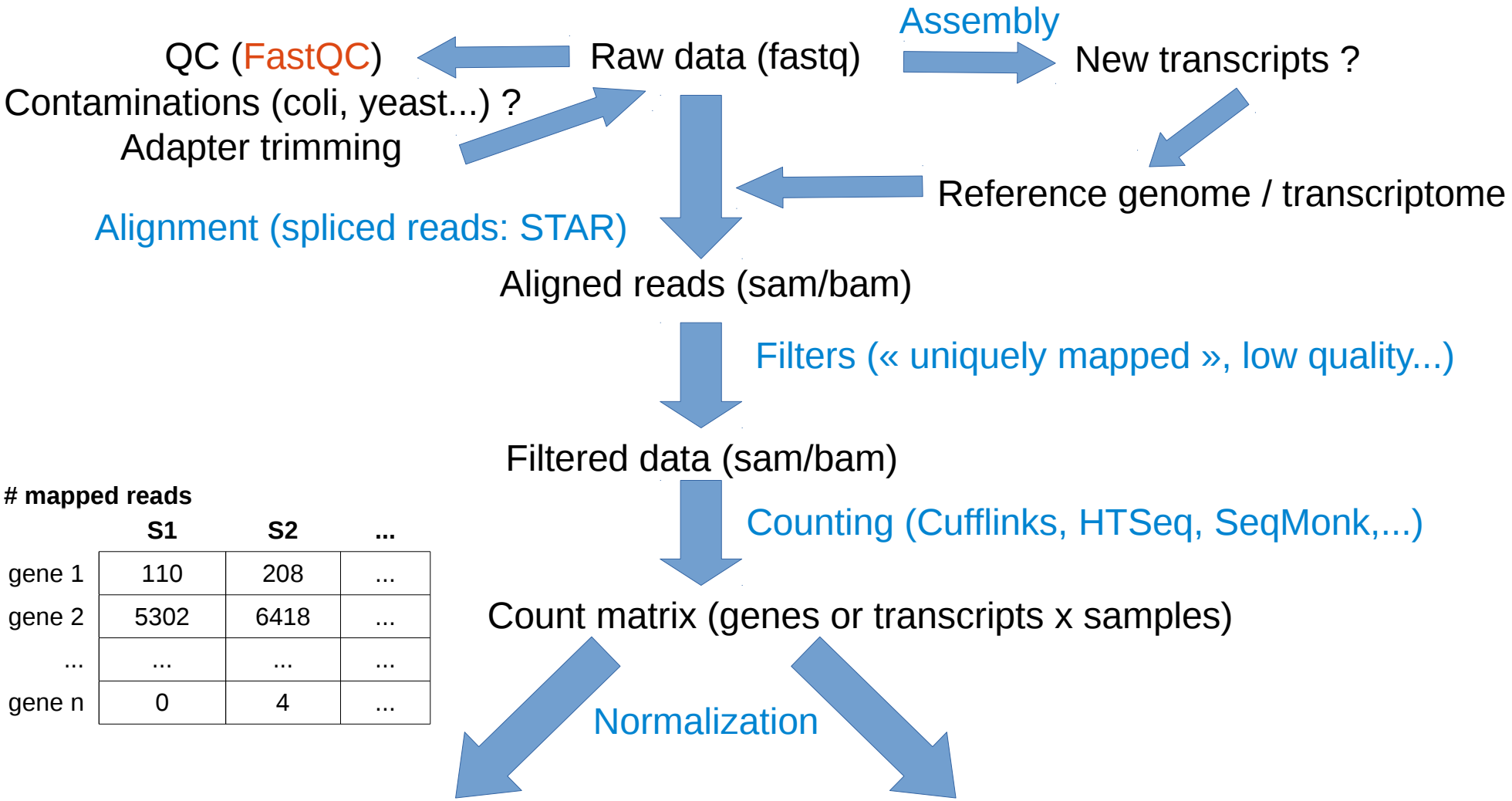
```
HWI-ST314_0109:6:2107:20072:96186#CAGATC     99   1    4481744   255  101M  =    4481796   153
CTTCCAGGAAGTGTGTAACACTGCTTCTGGCCCTCAGGTCGGGTCGGCAACCGTCAAATGTCGGGGTAGTTGCAATAGTAGACCGCTGAGCTAGCGTCGGG
bbbeeeeegffcc[d^dggffhhfhhhhhhhfhhhhhhgffhhg`affhfhhe\ca]bddacbcccKW^YZb]Y``BBBBBBBBBBBBBBBBBBBBBBB
NM:i:1    NH:i:1

HWI-ST314_0109:6:2202:16904:49328#CAGATC     163  1    4481865   255  73M1D28M  =    4481886   123
CGTGGCTGTCTGAGAGGTTCACTCCGCAGTCGTGTCCCTGGTAGGGAAGACCCATCTCGGGCTTATACACAAAGGCAGATACTGTTCGAATTCCGTGCGGT
a_aeeeeebeeeehehhbfbeg`fhfhfhfgffe_efhdfhfffhicddd]bg_bgbhZdgc]b`bbdddbac]^^BBBBBBBBBBBBBBBBBBBBBBB
NM:i:1    NH:i:1
```

1 deletion in the read

CIGAR string

Typical RNA-seq pipeline



 # mapped reads

	S1	**S2**	**...**
gene 1	110	208	...
gene 2	5302	6418	...
...	...	...	...
gene n	0	4	...

RNA-seq

Advantages:

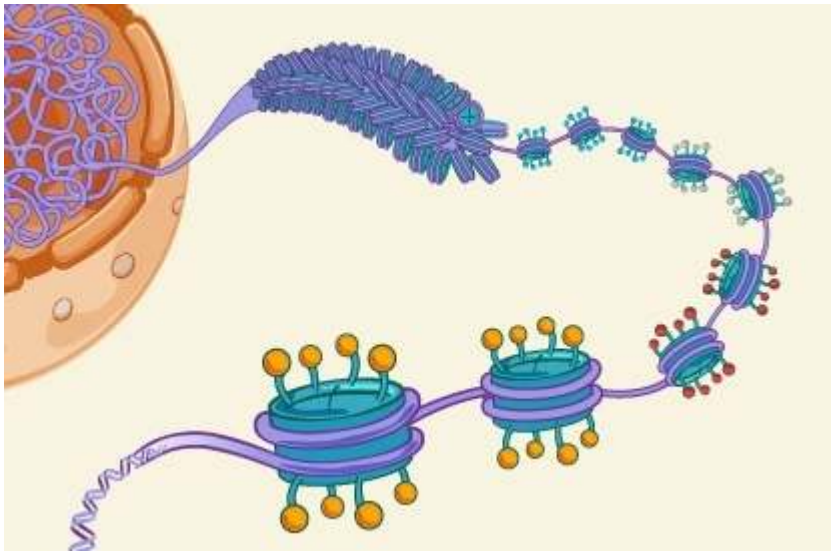
- Improved sensitivity and dynamic range
- Better reproducibility (SEQC)
- Not limited to prior knowledge of sequence and annotations
- Single base resolution of transcribed features (novel transcripts, splicing, allele-specific expression)
- Can detect quantitative as well as qualitative/structural changes to the transcriptome

Drawbacks / limits:

- Some GC content bias
- Mapping ambiguity for paralogous sequences
- Some protocols are highly sensitive to RNA degradation.
- Measurements for different genes are not independent from each other
- Higher cost per sample
- Data analysis is non trivial

Epigenomics

Epigenome



- DNA methylation
- Histone modifications
- Chromatin accessibility
- Chromatin conformation
- ncRNA

International intiatives / consortium

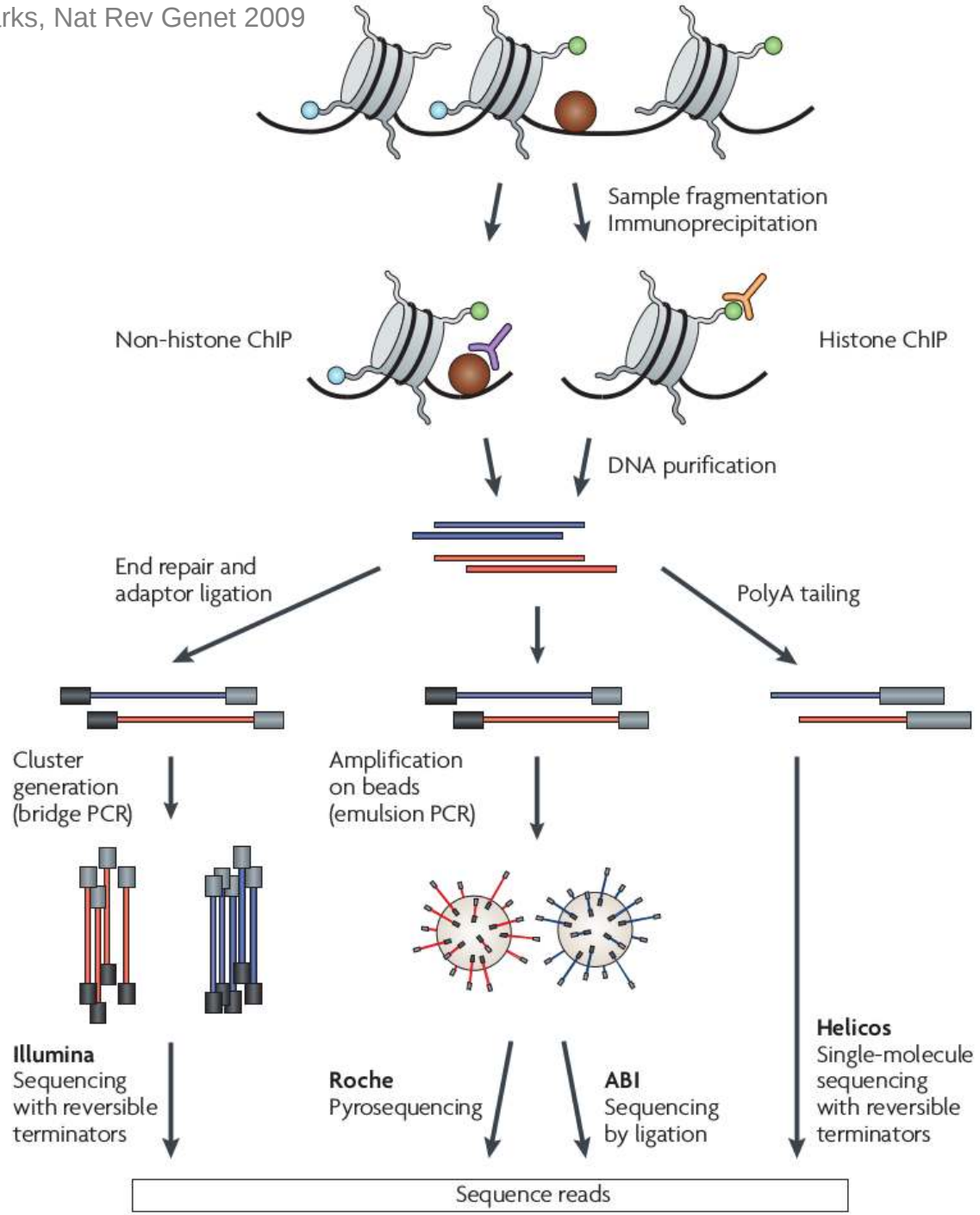
International Human Epigenome Consortium

- ENCODE and modENCODE
- NIH Roadmap Epigenomics
- Blueprint epigenome (Haematopoietic epigenomes)
- 4D Nucleome

...

ChIP-seq

Parks, Nat Rev Genet 2009



- Mapping protein-DNA interactions genome-wide
- Transcription factors, RNA polymerase, histones, insulators, ...
- Numerous protocol variants
- Generally requires lots of starting material (>10^6-10^7 cells)

Typical ChIP-seq pipeline

Voir aussi : Bailey et al., Plos Comp. Biol, 2013

QC (FastQC) ⇄ Raw data (fastq)

Adapter trimming

Reference genome → Alignment (unspliced reads: bowtie2, bwa,...)

Aligned reads (sam/bam)

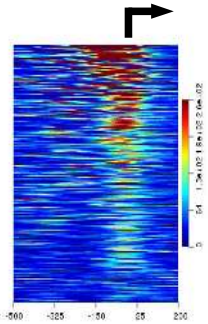
Filters (« uniquely mapped », unique, low quality...)

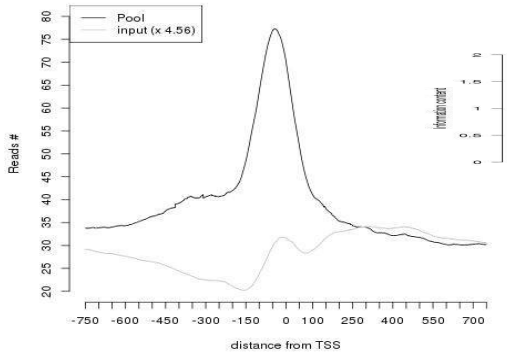
Filtered reads (sam/bam)

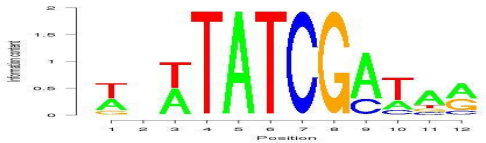
coverage / profiles / heatmaps

R / Bioconductor

Python / Perl / ...







Peak calling → BED file

MACS, MACS2, SPP

Sicer (histone mod),...

Counting and differential analysis

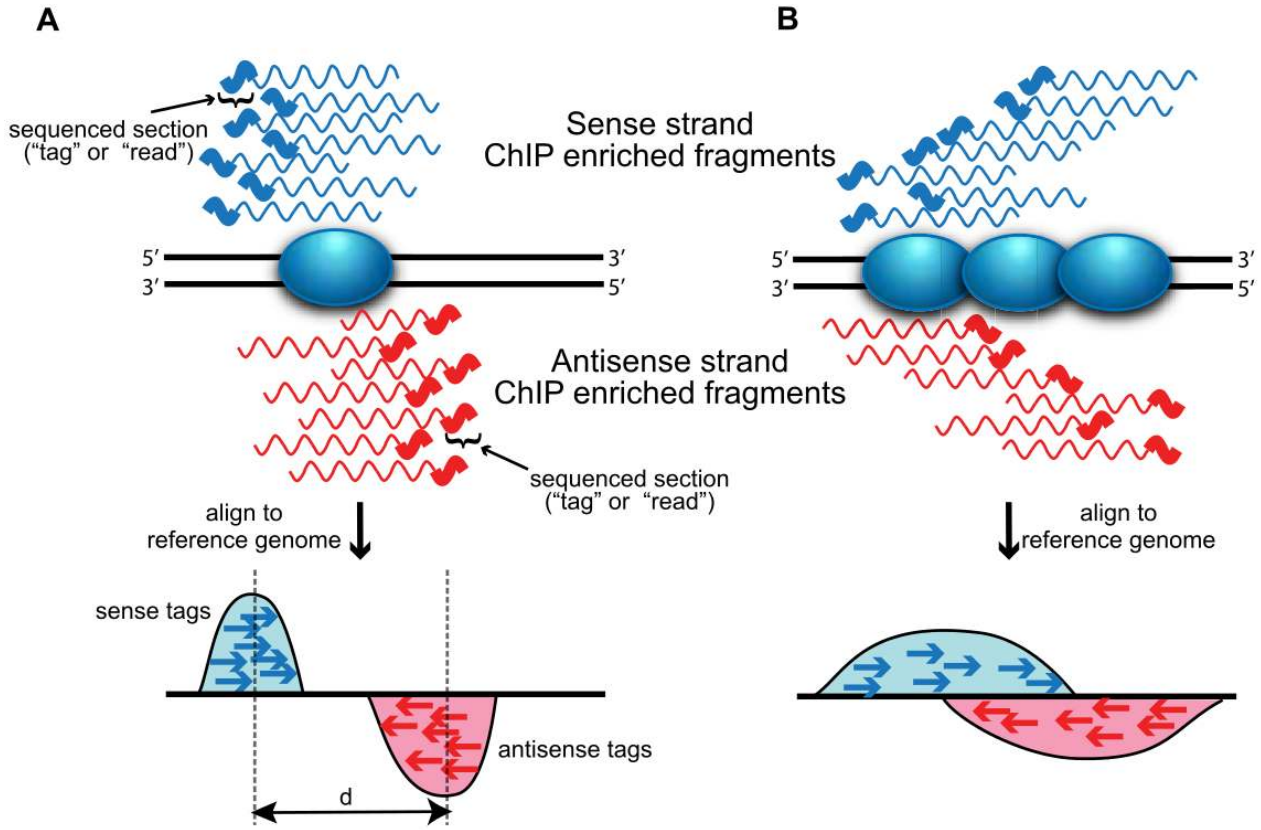
R / Bioconductor...

Motif search and analysis

RSAT, XXmotif, MEME, R, ...

Peak calling

The orientation of sequencing (5' to 3') and short read length create a bias in read distribution on +/- strands



A. Sequence-spécific binding (TF...)
B. Broad binding (histone, RNA Pol,...)

Wilbanks & Facciotti, PloS One, 2010

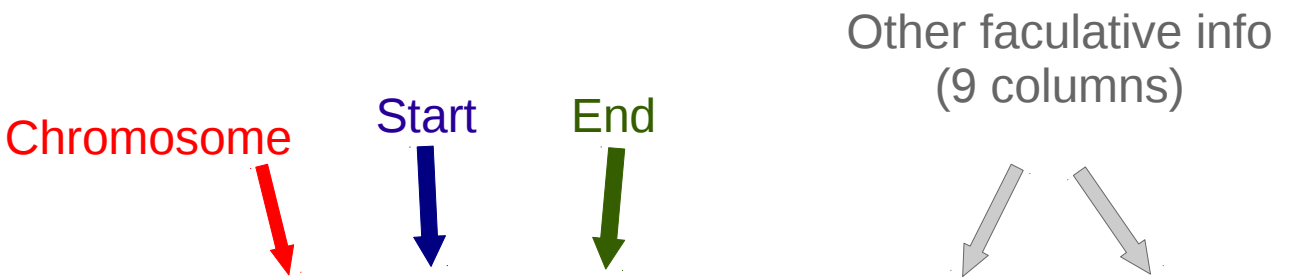
General principles for peak calling :

- Estimate fragment size (d) then shift (or extend) the reads by d/2
- Search for regions of enriched signal (# de reads) compared to background (local ideally)

→ Peak list : BED file

BED format

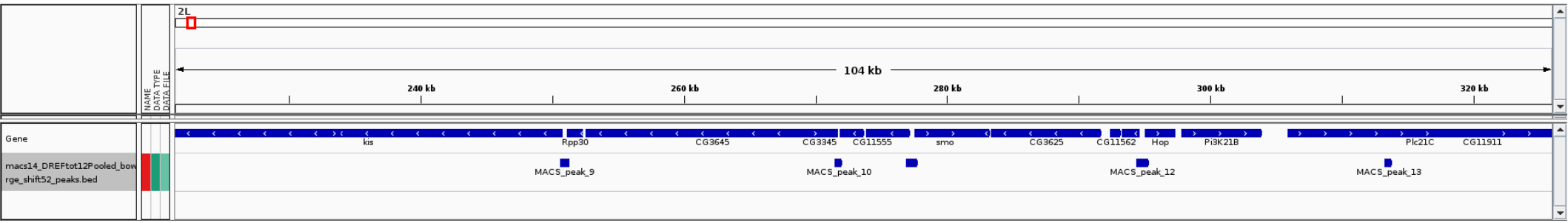
http://fr.wikipedia.org/wiki/BED_(format_de_fichier)
http://genome.ucsc.edu/FAQ/FAQformat.html




```
2L  47348  52731  MACS_peak_1   342.66
2L  71901  73894  MACS_peak_2   175.37
2L  347522 355682 MACS_peak_3   588.17
2L  403198 404266 MACS_peak_4   75.21
2L  489732 491282 MACS_peak_5   228.43
```

Colonne	Contenu
1	chrom
2	chromStart
3	chromEnd
4	name
5	score
6	strand
7	thickStart
8	thickEnd
9	itemRgb
10	blockCount
11	blockSizes
12	blockStarts

Storage of genomic intervals
Frequently used in genome browser to represent e.g. annotations or detected signal peaks

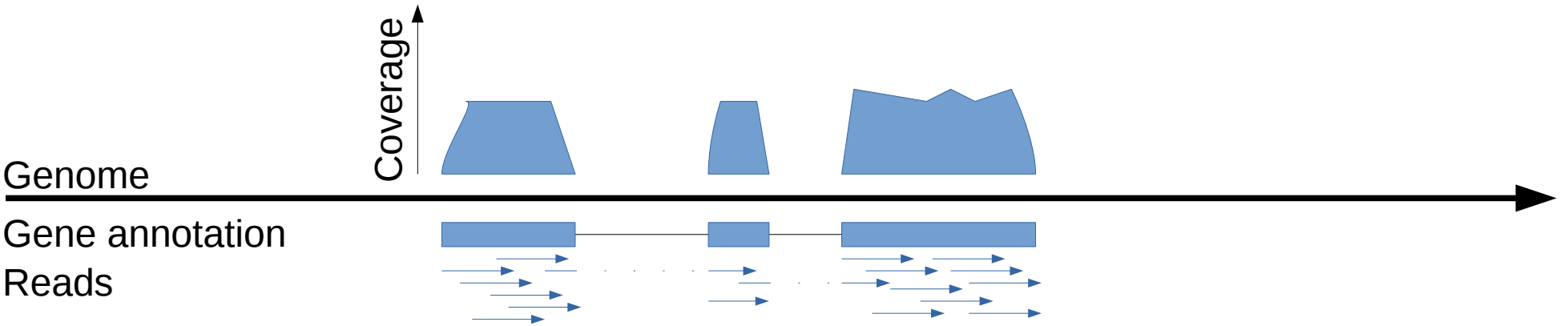


Other format for annotations : GFF, GTF (cf UCSC)

Formats for data values along the genome

Wig/bigwig
Bedgraph

Store a value (coverage) along the genome



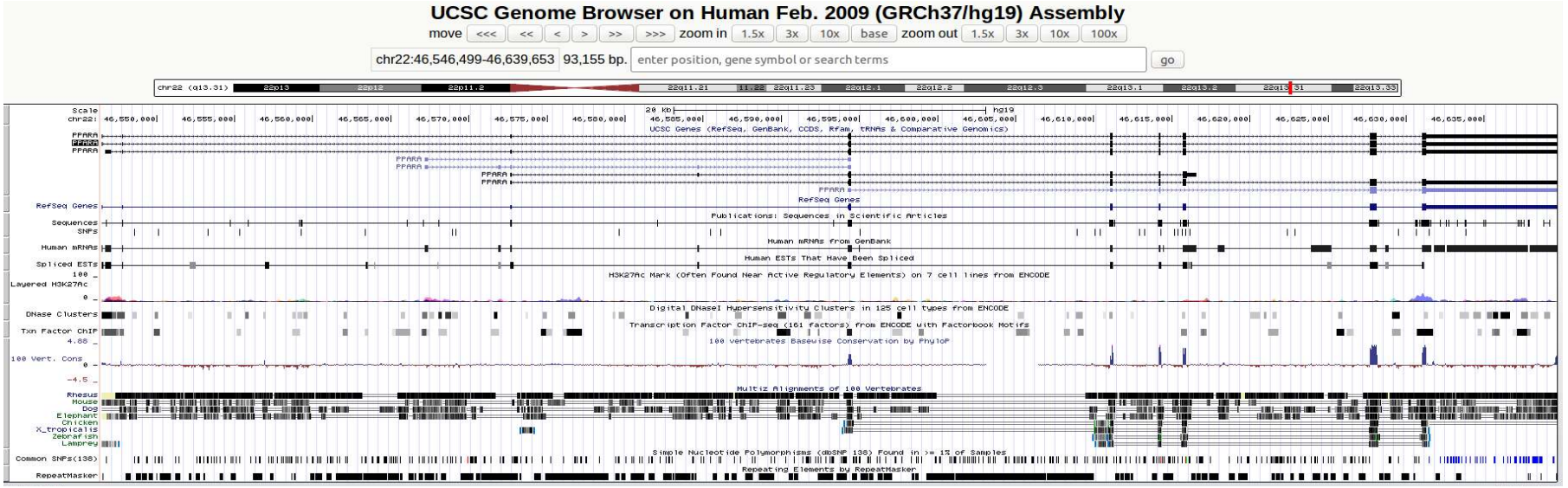
...0000000000123432200000022300023443433320000000000...

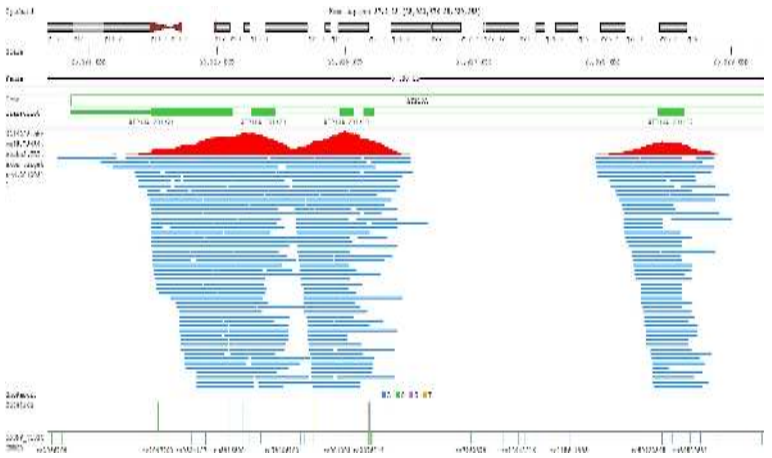
Chr	start	end	value
Chr1	1	125	0
Chr1	126	132	1
Chr1	133	134	2
...			

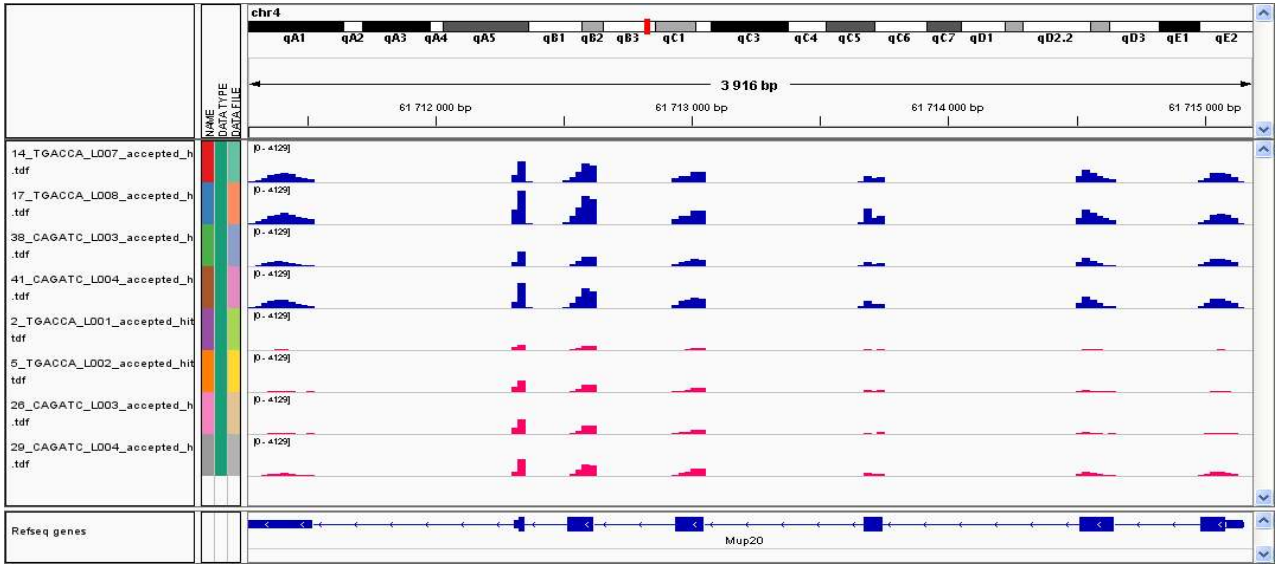
Genome browser

A genome browser allows to navigate your data

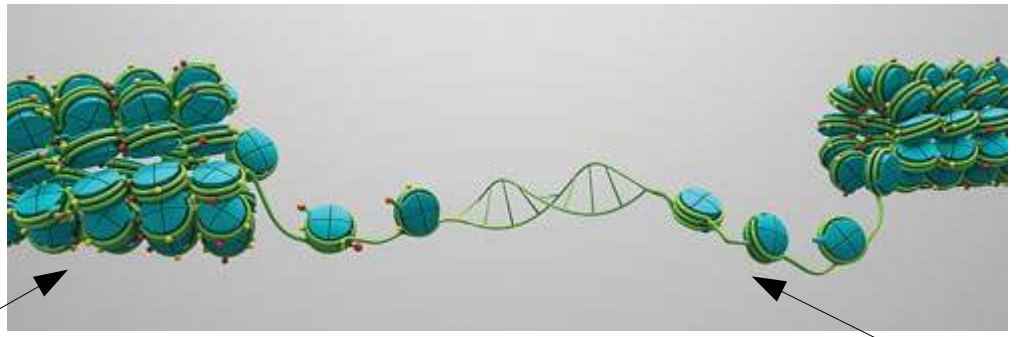
Ex : UCSC GB, Integrative Genomics Viewer, Ensembl GB, jbrowse, ...





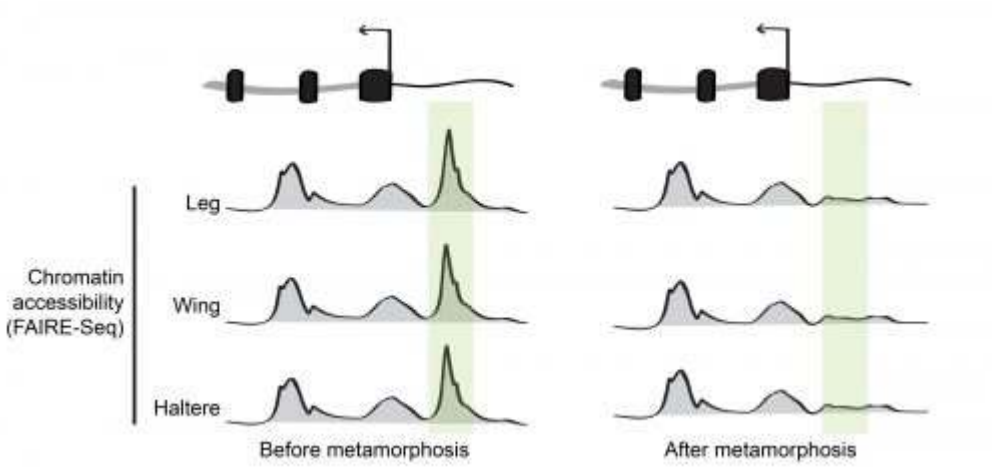


Why mapping chromatin accessibility ?

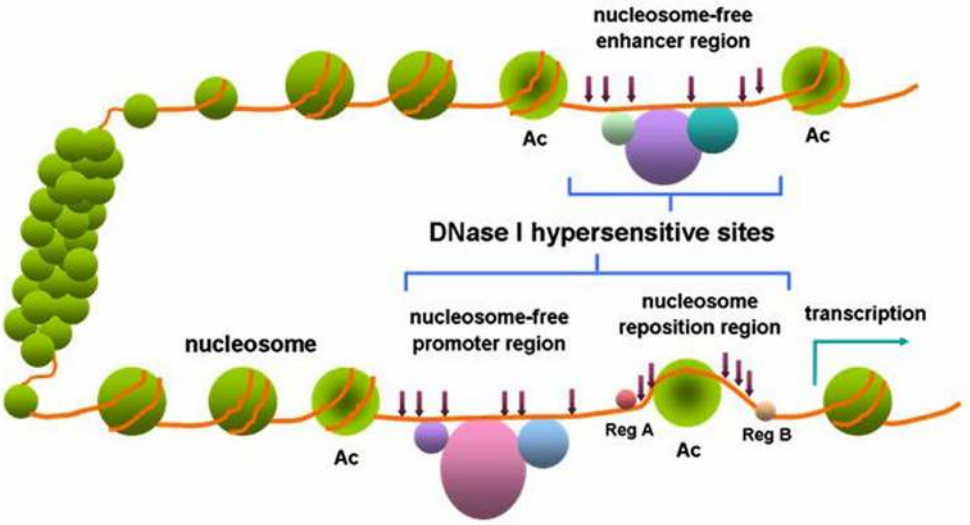


heterochromatin => inaccessible

euchromatin => more accessible



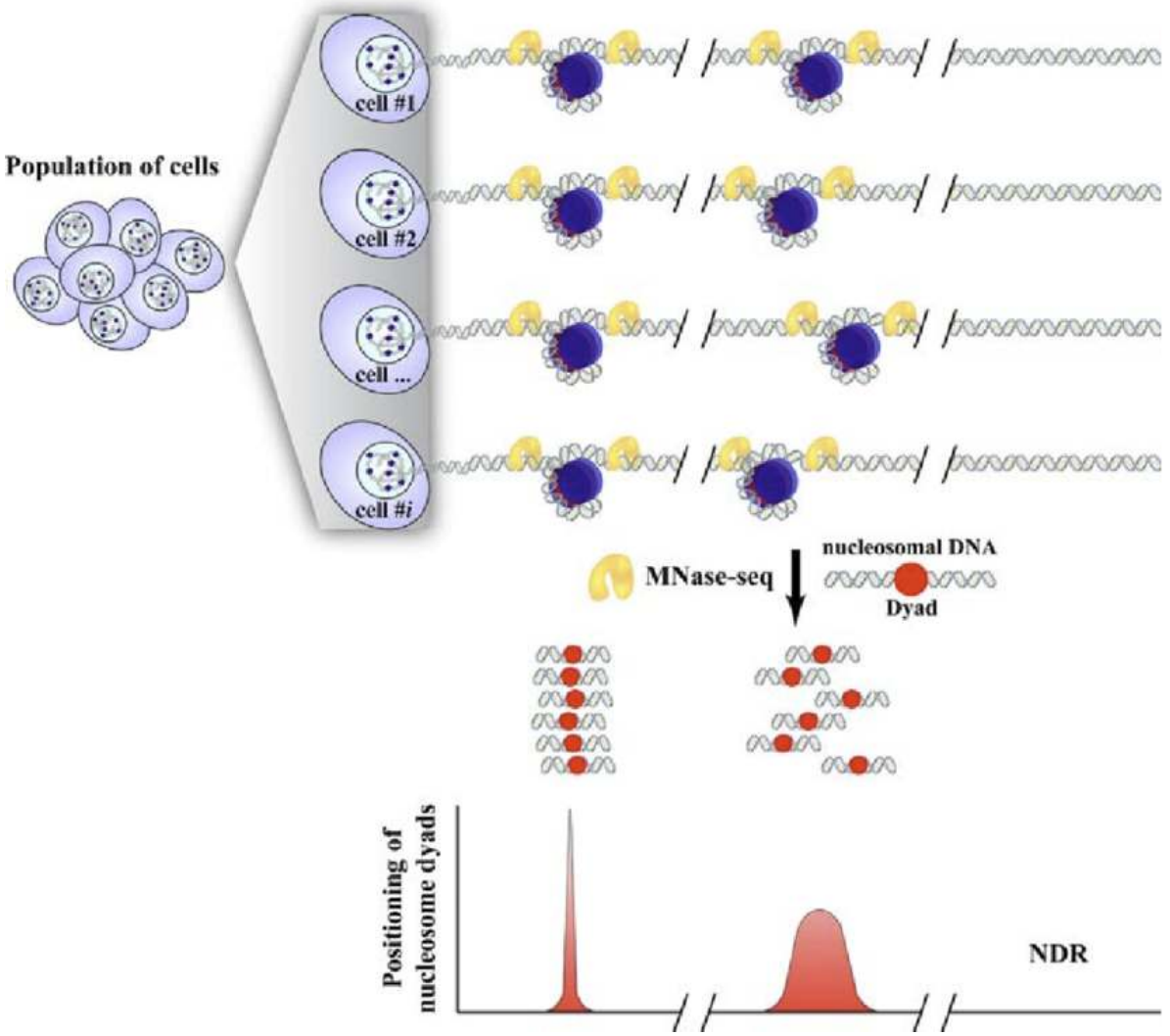
accessible <=> active



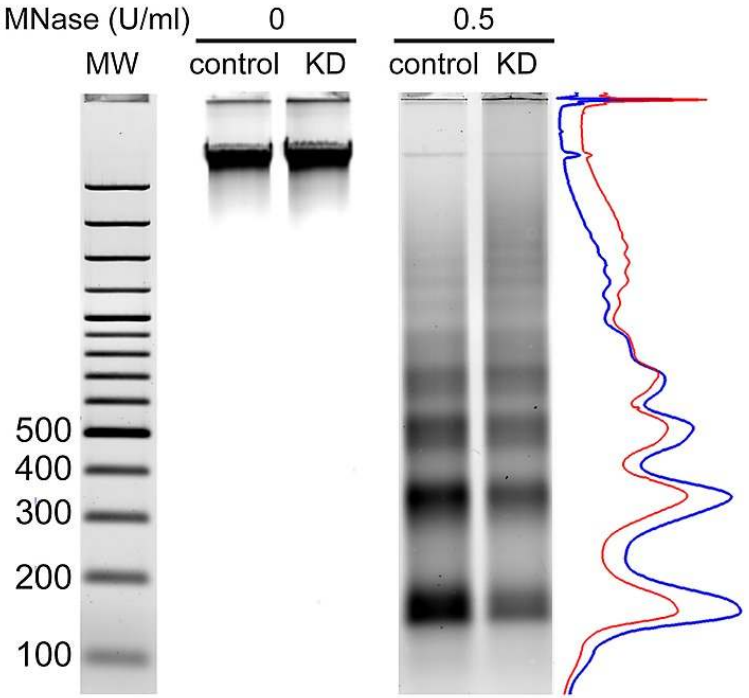
active promoters and enhancers are highly accessible regions

MNase-seq

- Mapping nucleosome positions (MNase)



Moshkin, AIMS Biophysics, 2015

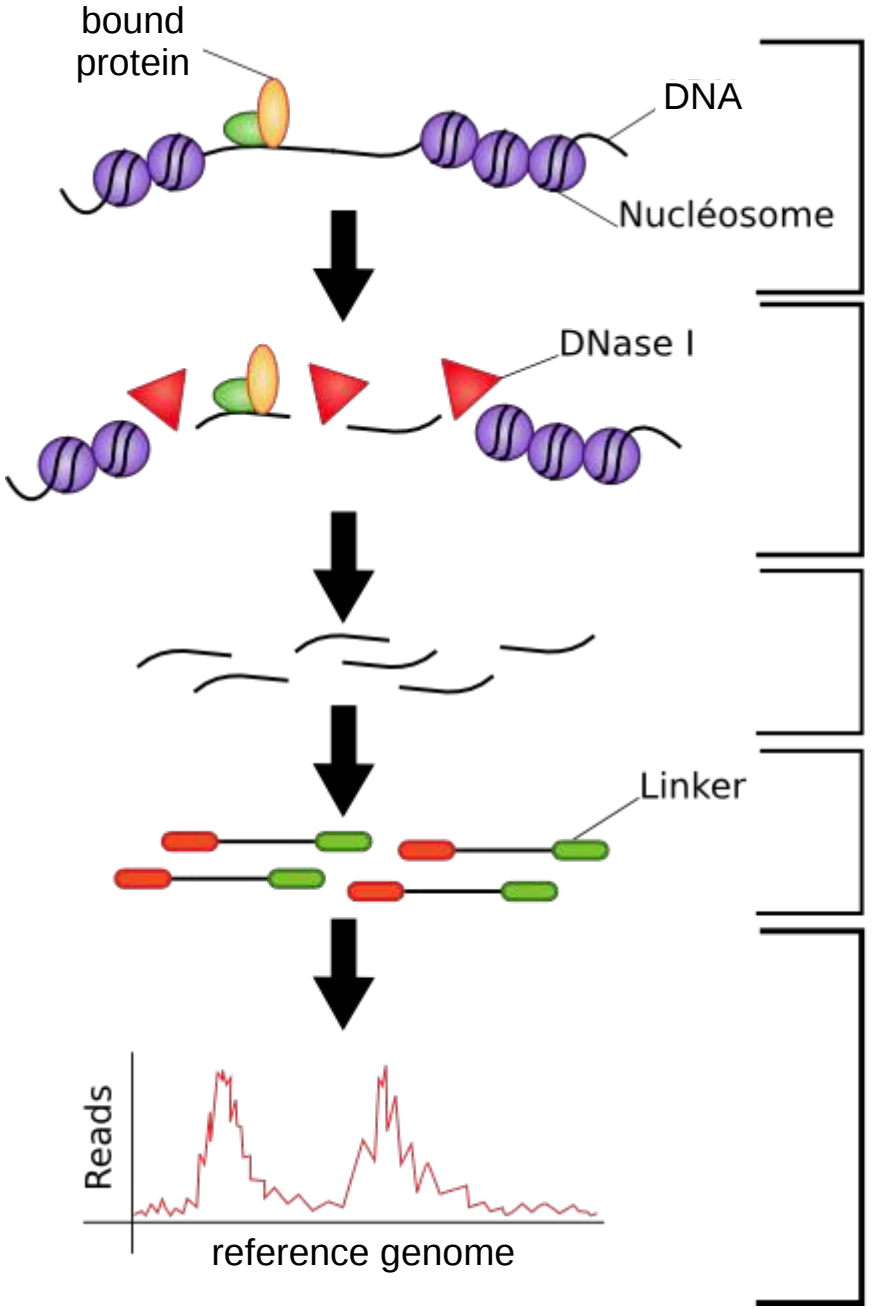


Celona et al., PloS Biol, 2011

- Can be combined with ChIP => Native ChIP

DNase-seq

- Mapping accessible / regulatory regions (DNase)



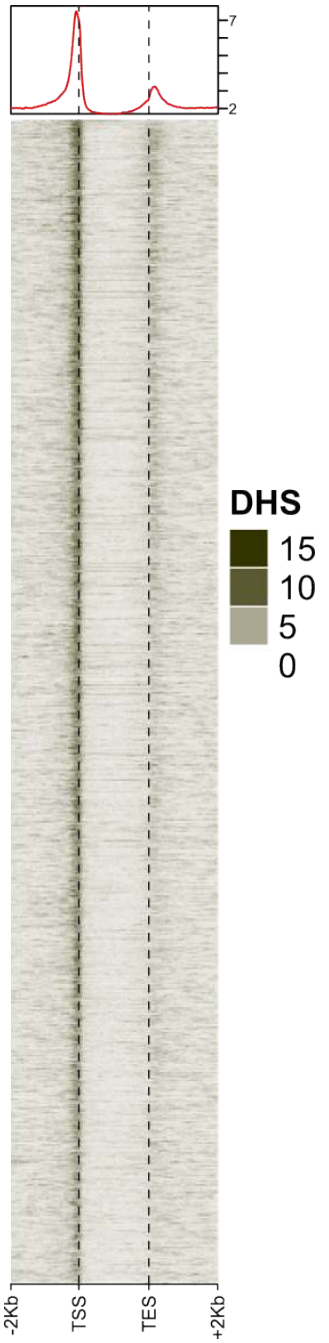
DNA extraction

DNase I digestion

DNA purification

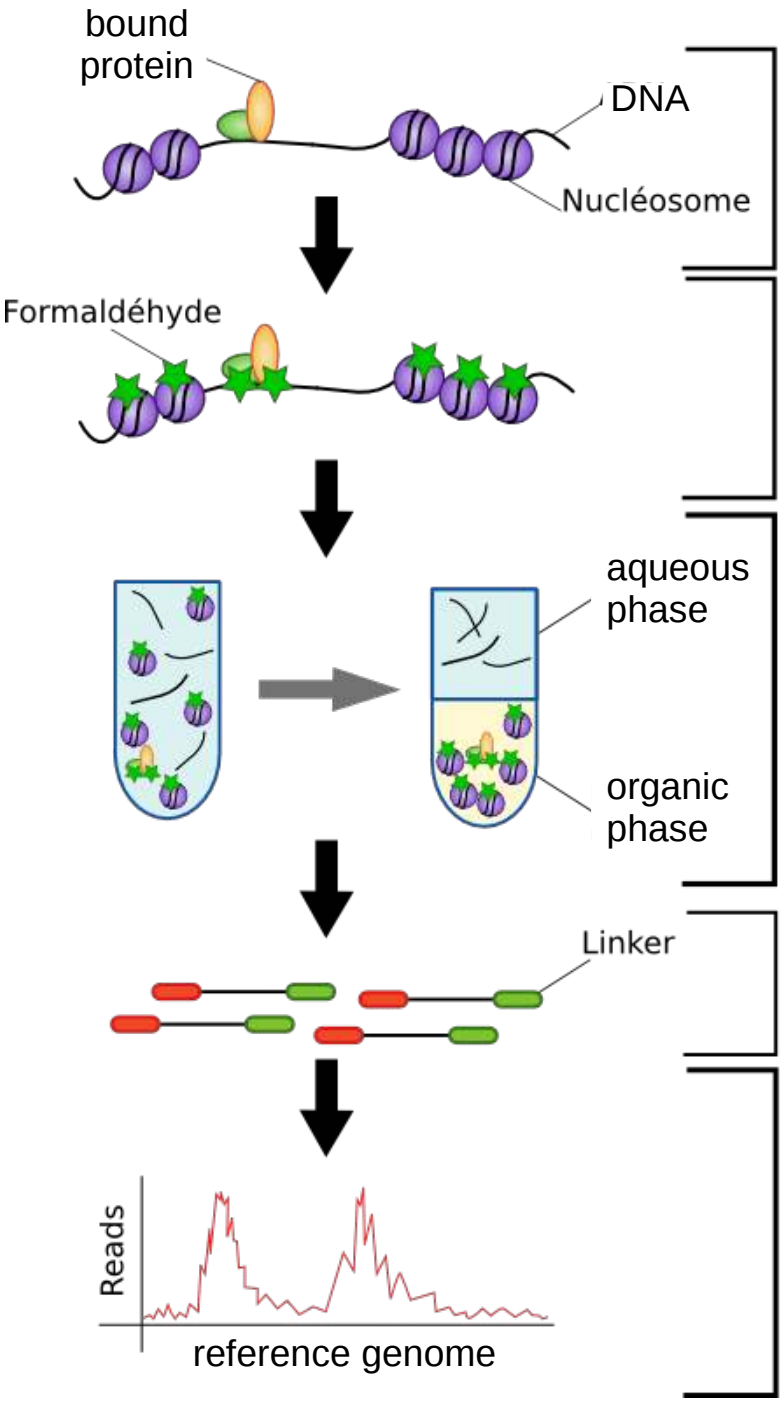
Library prep & amplification

Sequencing & bioinfo



FAIRE-seq

- Mapping accessible / regulatory regions



DNA extraction

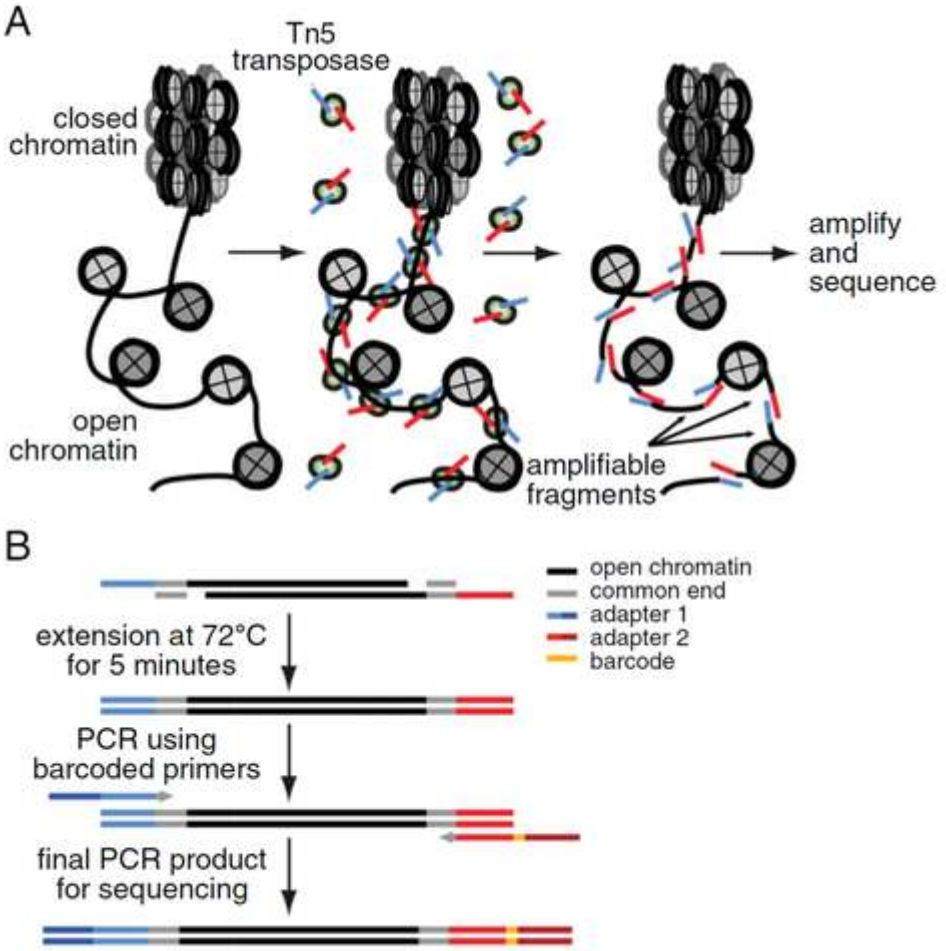
Crosslinking

Sonication and DNA purification

Library prep & amplification

Sequencing & bioinfo

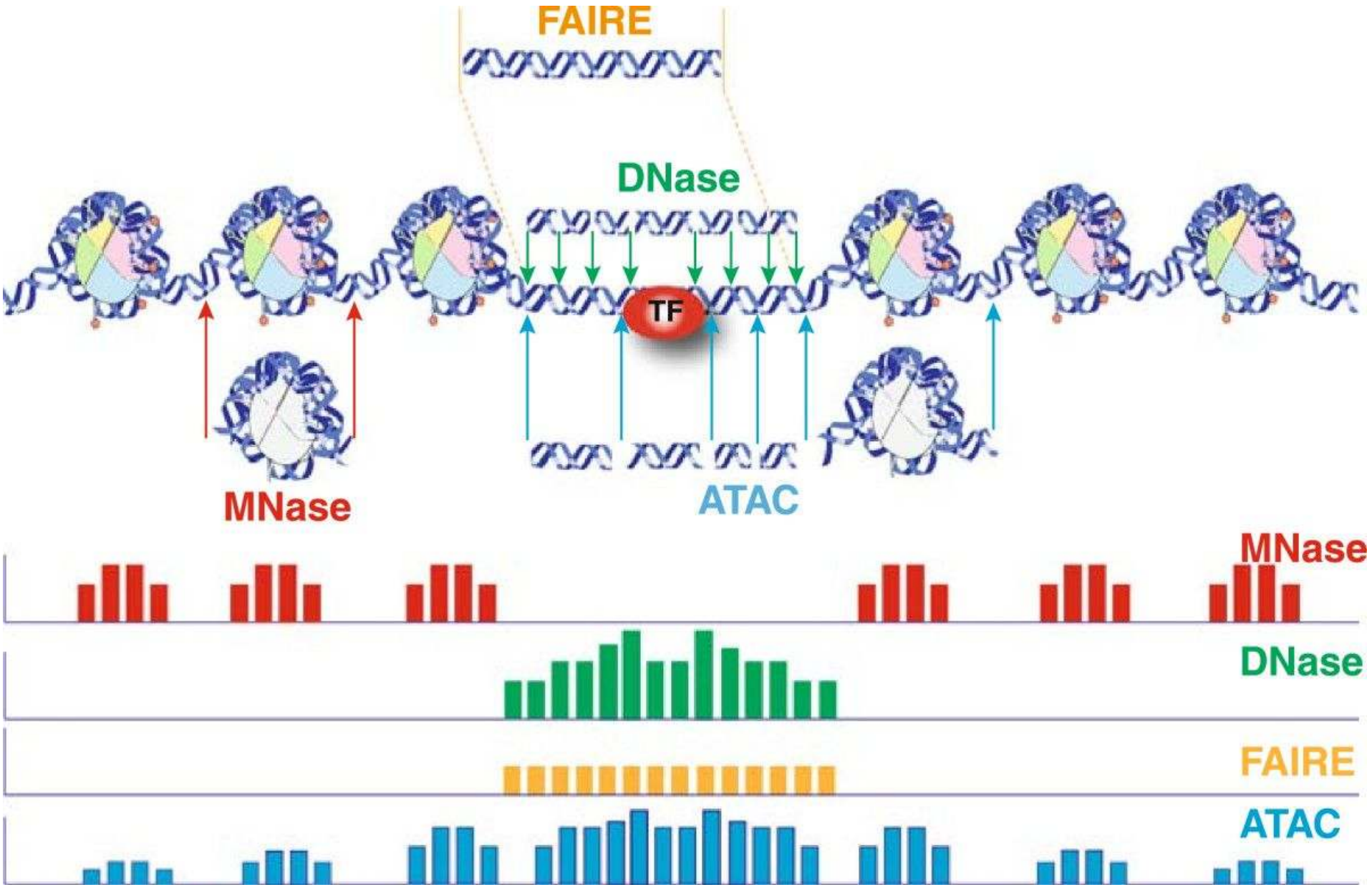
ATAC-seq



- Mapping of chromatin accessibility
- Fragment sizes carry important information (TF vs nucleosomal)
- Sequencing depth determines usable information
- Low input / single cell
- Often high mDNA contamination

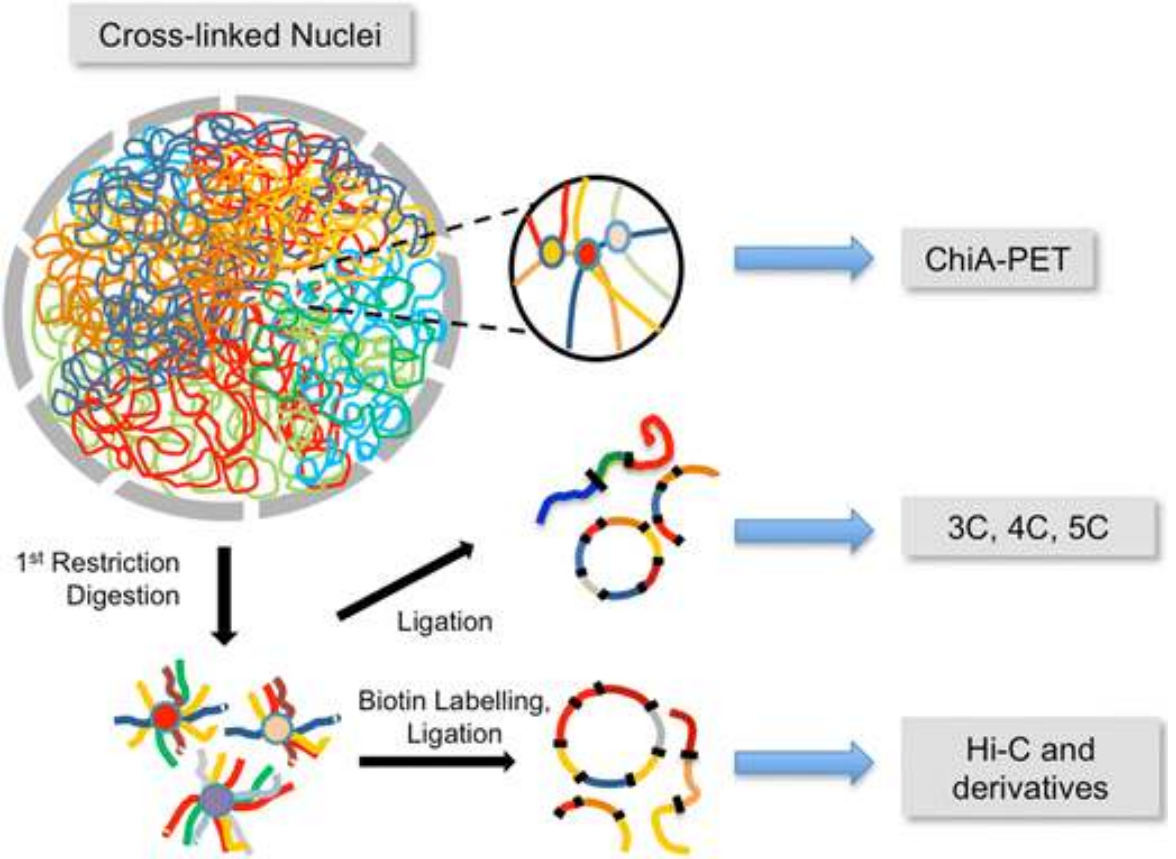
Buenrostro et al., Curr Protoc Mol Biol, 2015

Chromatin accessibility methods



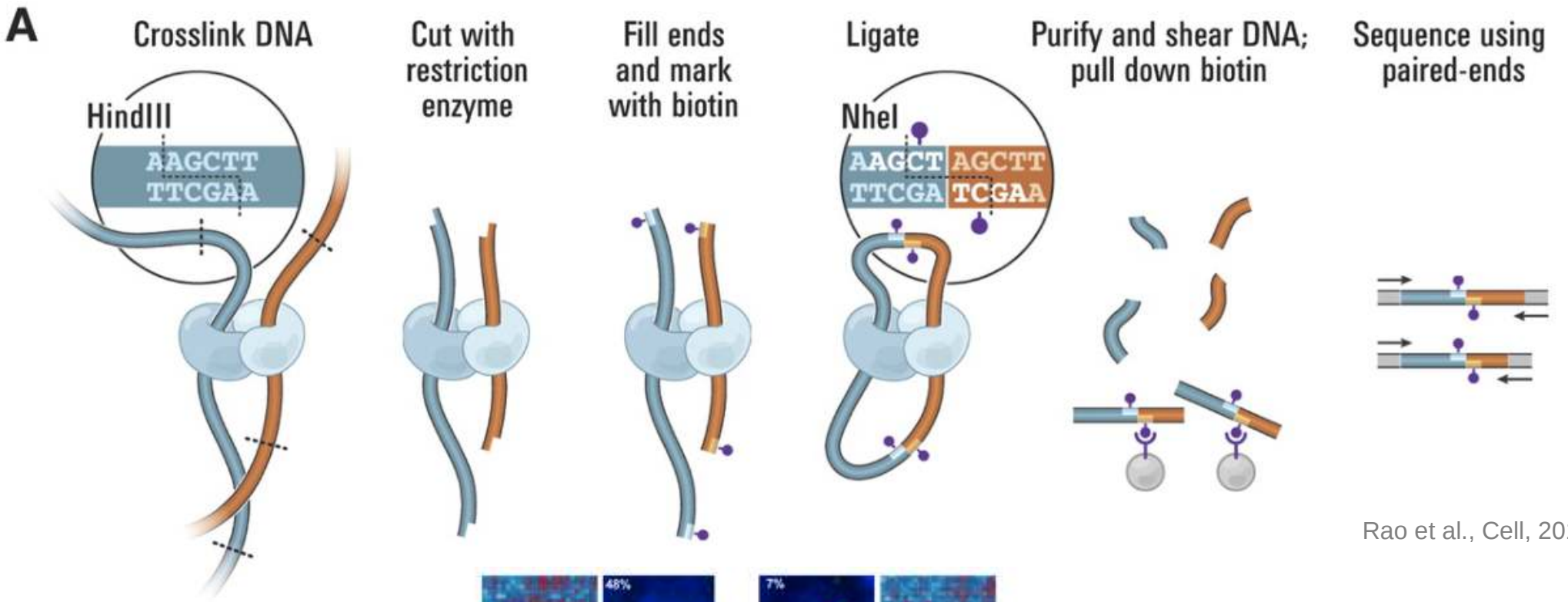
Tsompana & Buck, Epigenet Chrom, 2014

Chromatin Conformation Capture (3C) and related methods

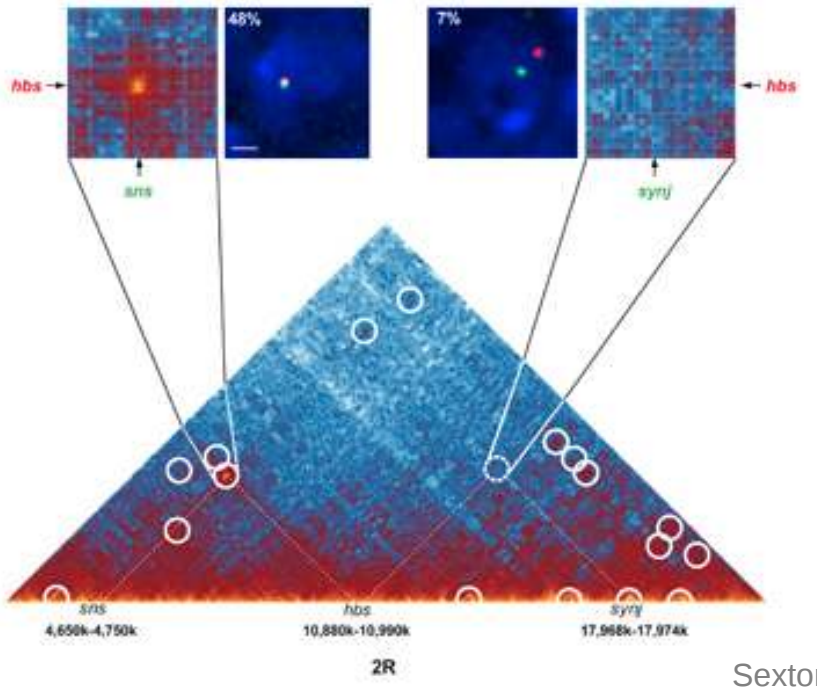


Sati & Cavalli, Chromosoma, 2017

Chromatin Conformation Capture (Hi-C)

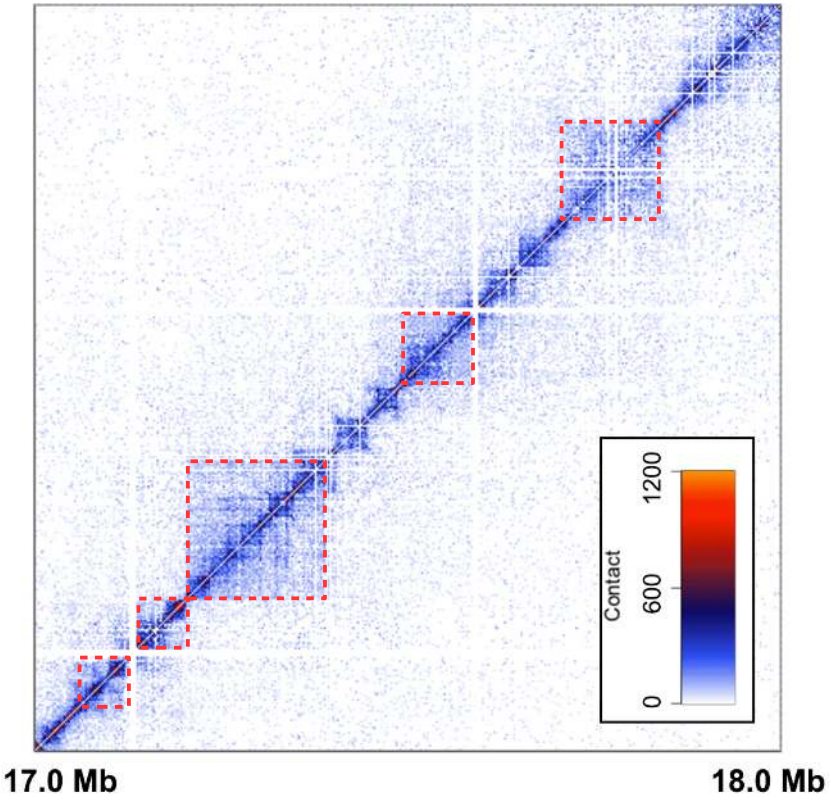


Rao et al., Cell, 2014

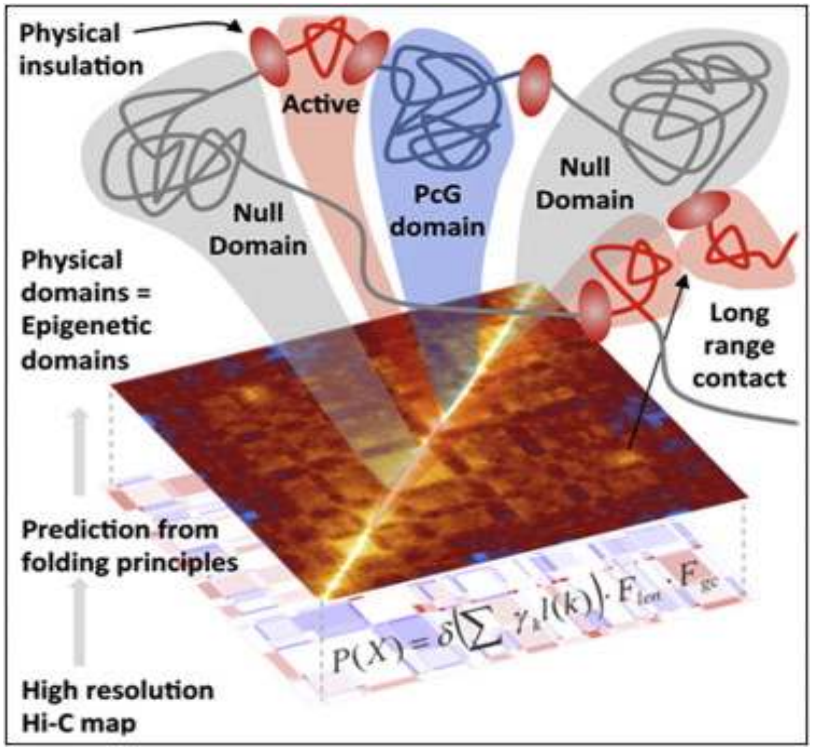


Sexton et al., Cell, 2012

Topologically associating domains

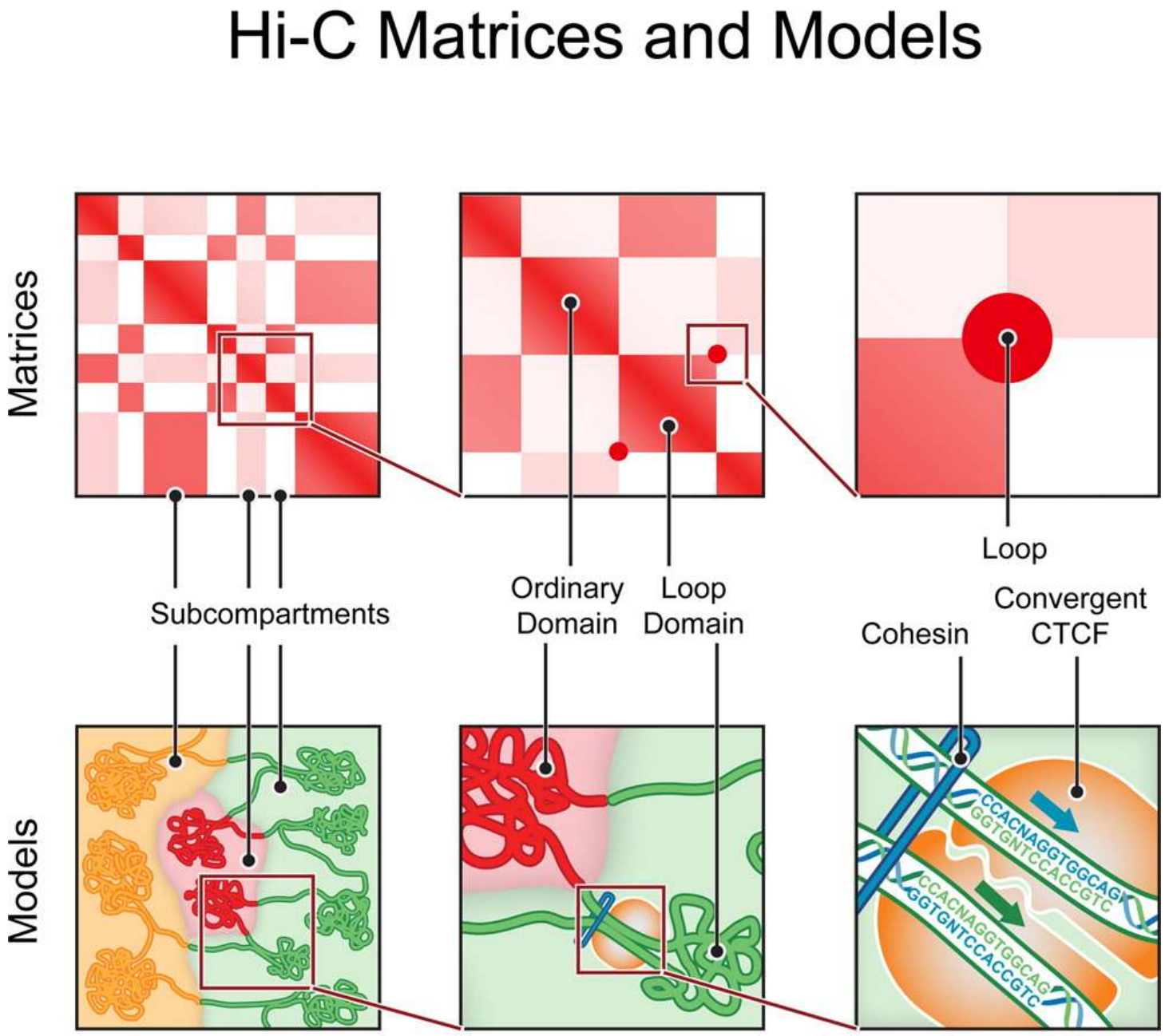


Hi-C data from Drosophila



Sexton et al., Cell, 2012

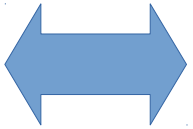
Compartments and loops



Rao et al., Cell, 2014

Conclusion

- NGS-based techniques provide an unprecedented view of genome function
- Improved scale, sensitivity and specificity
- New methods constantly developed (e.g. single cell)
- Need for standardization
- Data integration still a challenge



github.com/pgpmartin

Thank you for your attention

Pascal.Martin@inra.fr

@PgpMartin