Announcement

Seminar room change:

Frank Forward 317

Statistical Methods for High-dimensional Biology

Genomics technology

Yongjin Park, UBC Path&Lab, STAT, BC Cancer

Today’s lecture: Genomics technology

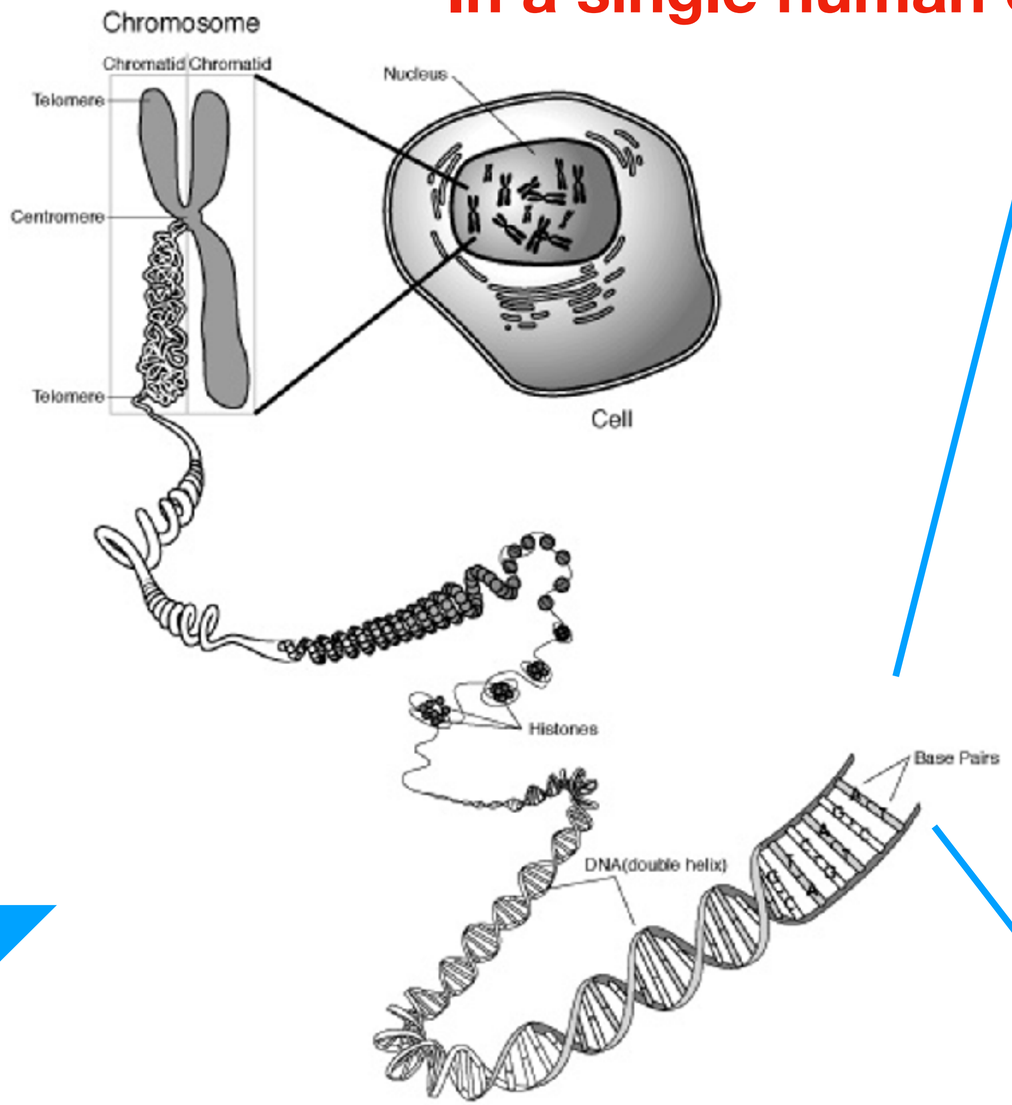
- **What is genomics technology?**
  - Measuring every step of the information flow: DNA, mRNA, splicing, protein, etc.
  - Why do we (statisticians) need to be aware of it?
- **Obtaining the book of life: a rough history of genomics methods**
  - Sequencing-based methods
  - Array-based methods
- **Understanding the book of life: omics technology**
  - Focusing on variations: mutations and expressions
  - Efforts to build epigenetic, transcriptomic, cell type references

Statistical methods for High-dimensional Biology: An overview

What is the high-dimensional aspect of biology?

In a single human cell (n=1)

zoom-in down to the DNA base pair level



- **3.2B** base pairs (ATGC characters)
- **30k** mRNA/genes (protein coding) ± 5k to 100k non-coding
- **100+** Epigenetic modifications per gene regulatory element
- **4+** isoforms due to post-transcriptional modifications
- Post-translational modifications

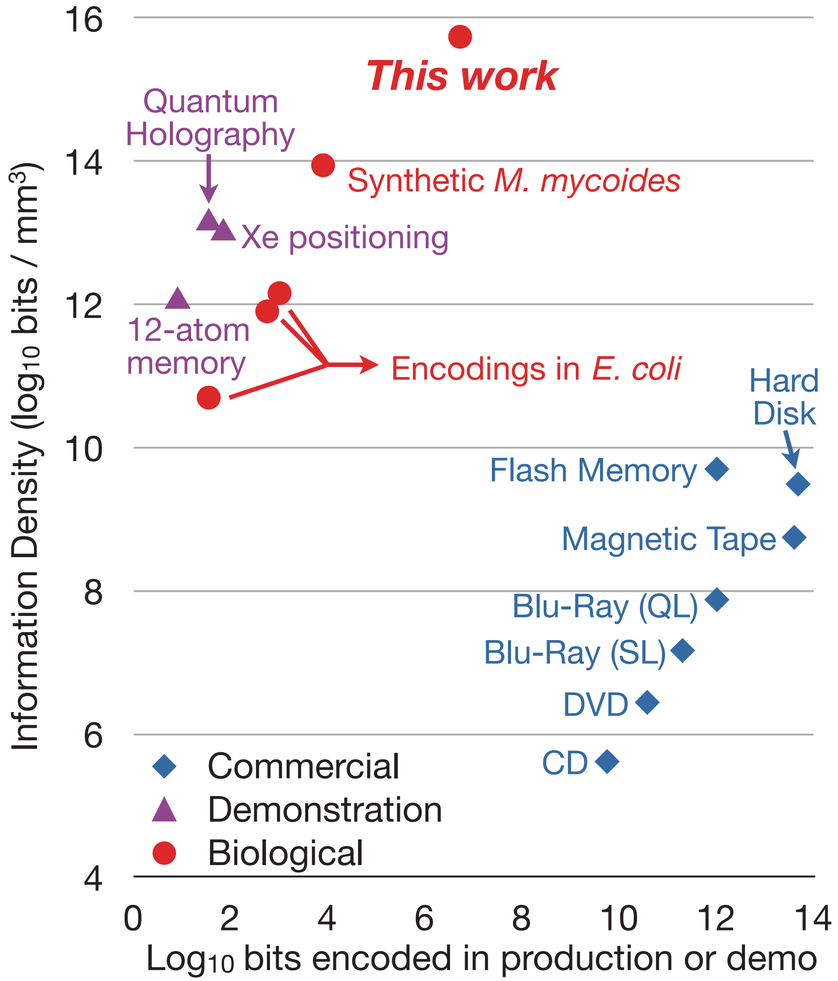
DNA can store a lot of information

BREVIA

Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2]*

- Hard to write but easy to read for a cell
- An efficient way to copy and edit
- Highly parallelized operations (many, many base pairs match simultaneously)
- A long history of evolution



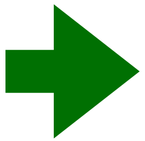
What if all the 3.2B variables are independent?

How can "statistics" help navigate high-dimensional space?

What can help us in the high-dimensional setting? **Essentially, our only hope is that the data is endowed with some form of low-dimensional structure, one which makes it simpler than the high-dimensional view might suggest.**

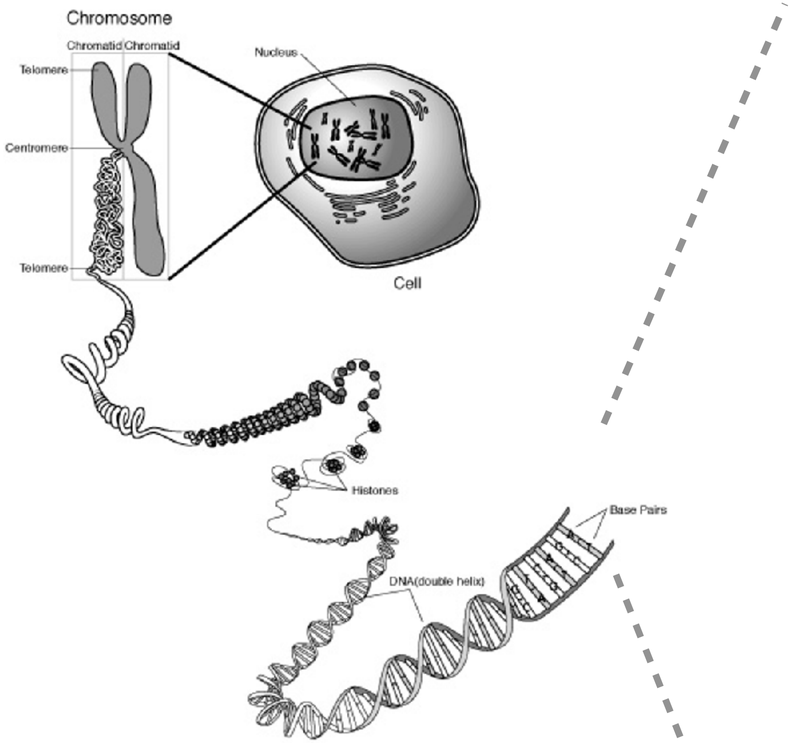
Wainwright, High-dimensional Statistics

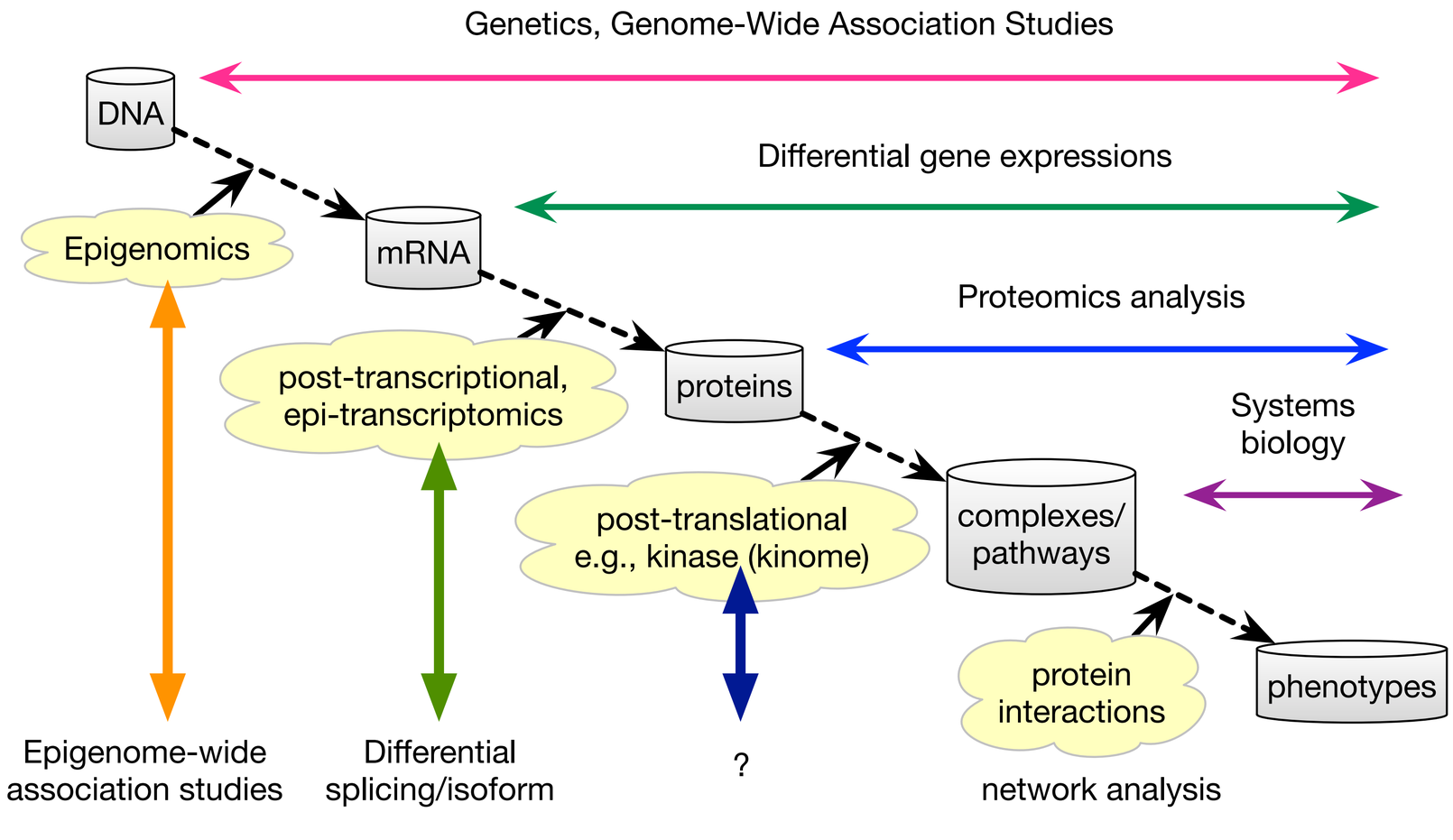
- Biological knowledge/constraints
- Multi-omics data integration
- Dependency structures between variables
- Common cell types and pathways
- Evolutionary history
- (...)



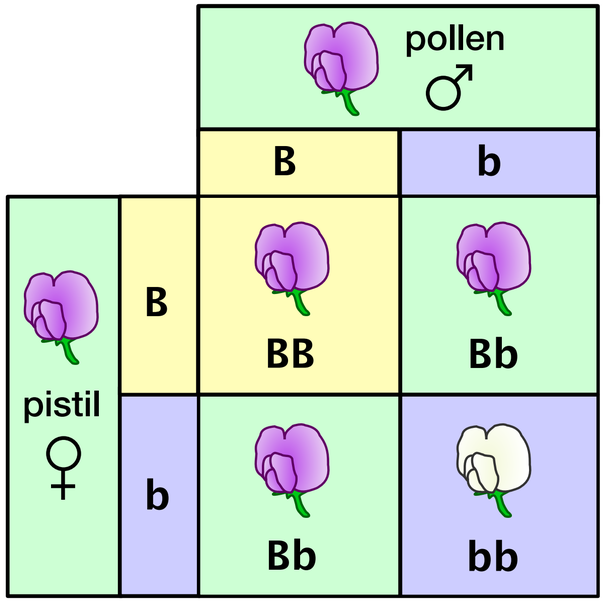
Understand unique aspects of biological data-generating processes

Understand the data-generating process





Mendel's discovery by comparing DNA variants with phenotypes



Gregor Mendel
(1822-1884)

Key concepts emerged:

- Gene = a unit of heredity that transfers from a parent to offspring
- Allele = a different form of a gene [from a Greek word, αλληλο, αλλος, "allos", other]

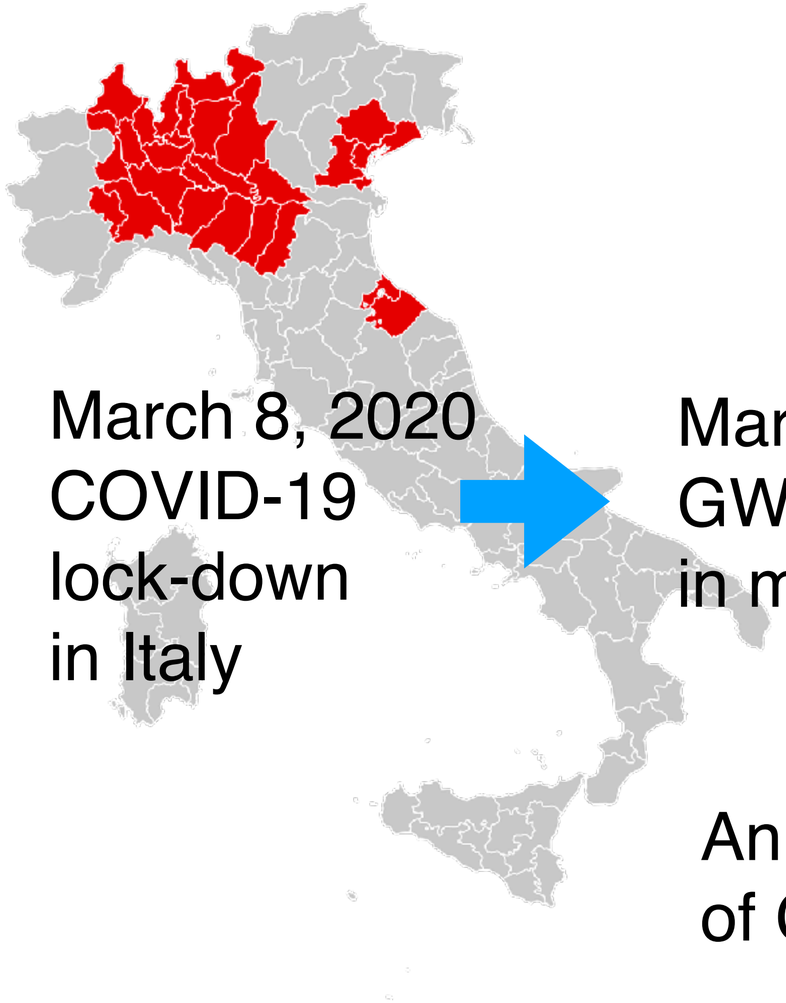
Human genetics revolution

To Solve 3 Cold Cases, This Small County Got a DNA Crash Course

Forensic genealogy helped nab the Golden State Killer in 2018. Now investigators across the country are using it to revisit hundreds of unsolved crimes.

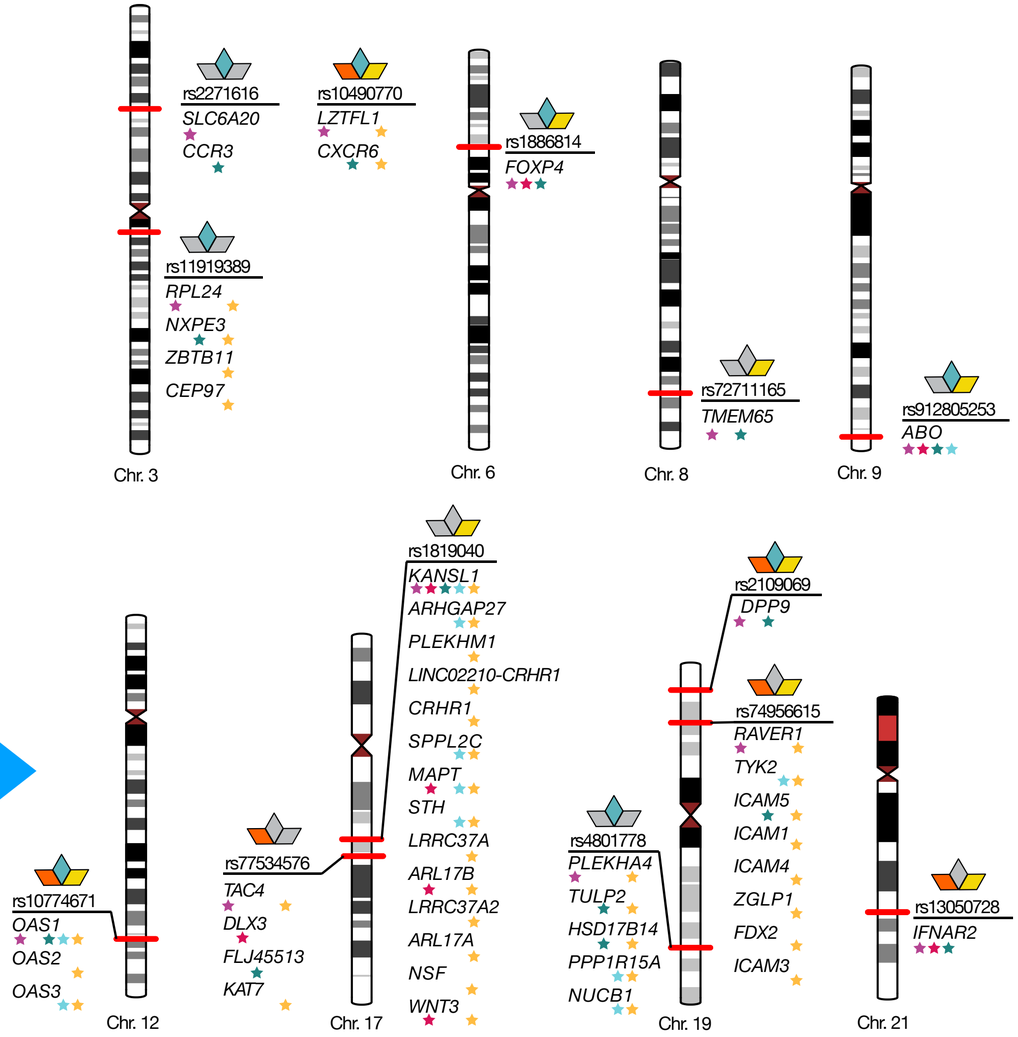
https://www.nytimes.com/2021/05/03/science/cold-cases-genetic-genealogy.html

COVID-19 Host Genetics



March 8, 2020
COVID-19
lock-down
in Italy

March 12, 2021,
GWAS paper
in medRxiv



An unprecedented pace
of GWAS profiling

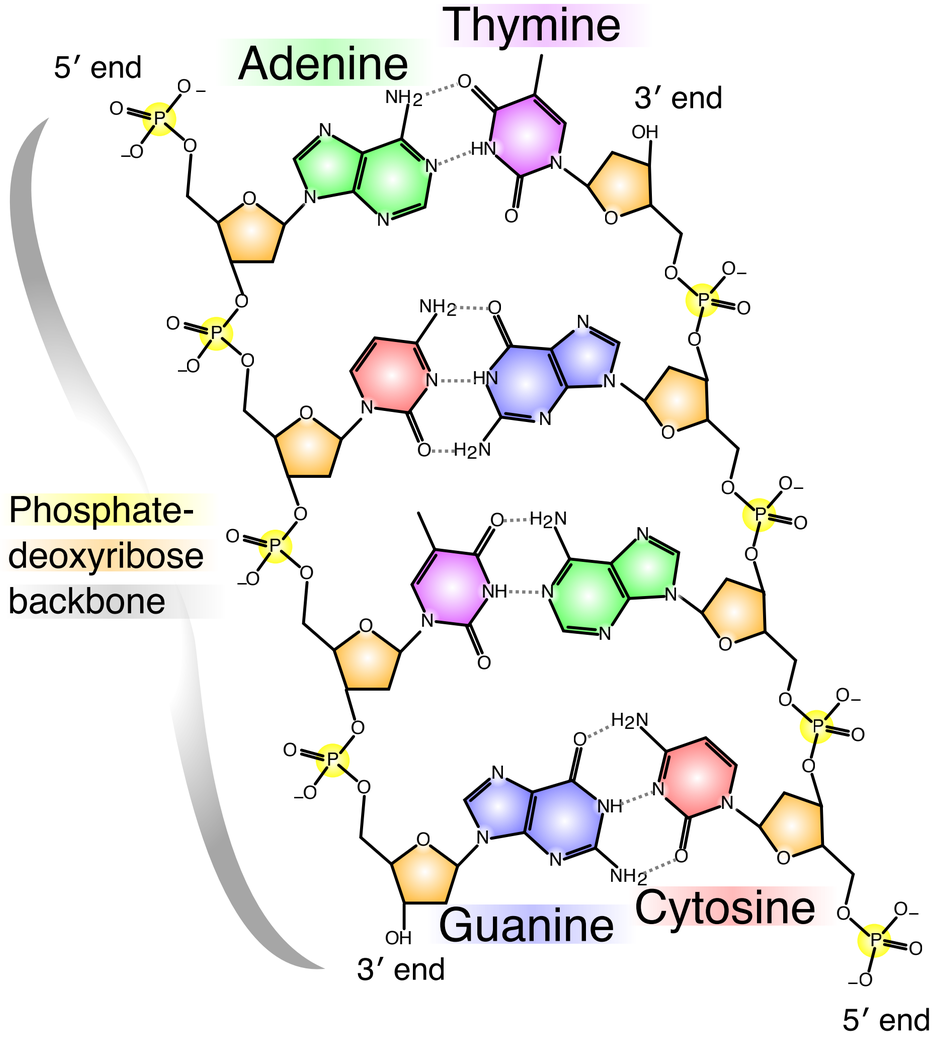
https://www.medrxiv.org/content/10.1101/2021.03.10.21252820v1.article-info

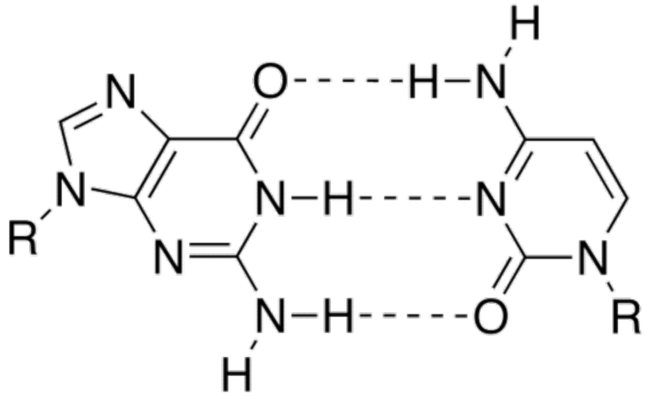
COVID-19 Host Genetics Initiative, *Nature* (2021)

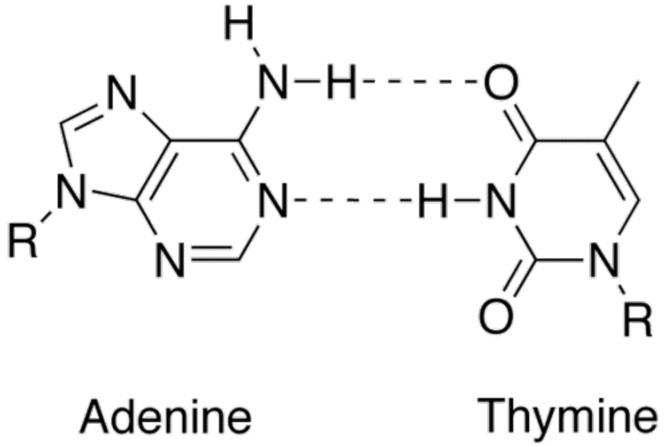
Today’s lecture: Genomics technology

- **What is genomics technology?**
  - Measuring every step of the information flow: DNA, mRNA, splicing, protein, etc.
  - Why do we (statisticians) need to be aware of it?
- **Obtaining the book of life: a rough history of genomics methods**
  - Sequencing-based methods
  - Array-based methods
- **Understanding the book of life: omics technology**
  - Focusing on variations: mutations and expressions
  - Efforts to build epigenetic, transcriptomic, cell type references

Key forces: hydrogen bond between complementary “characters”

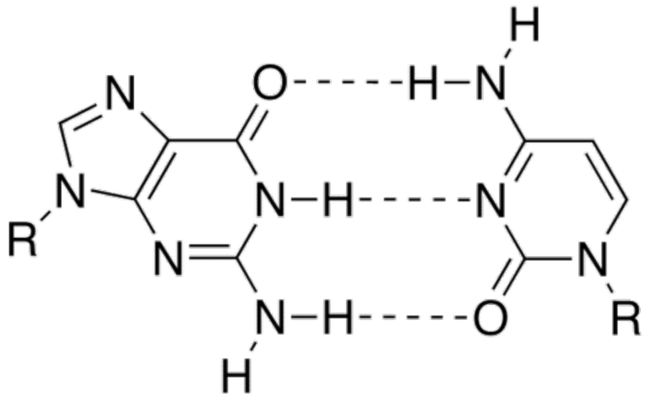


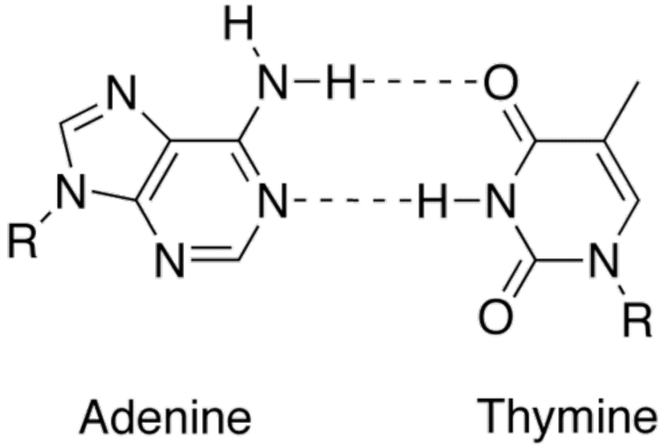


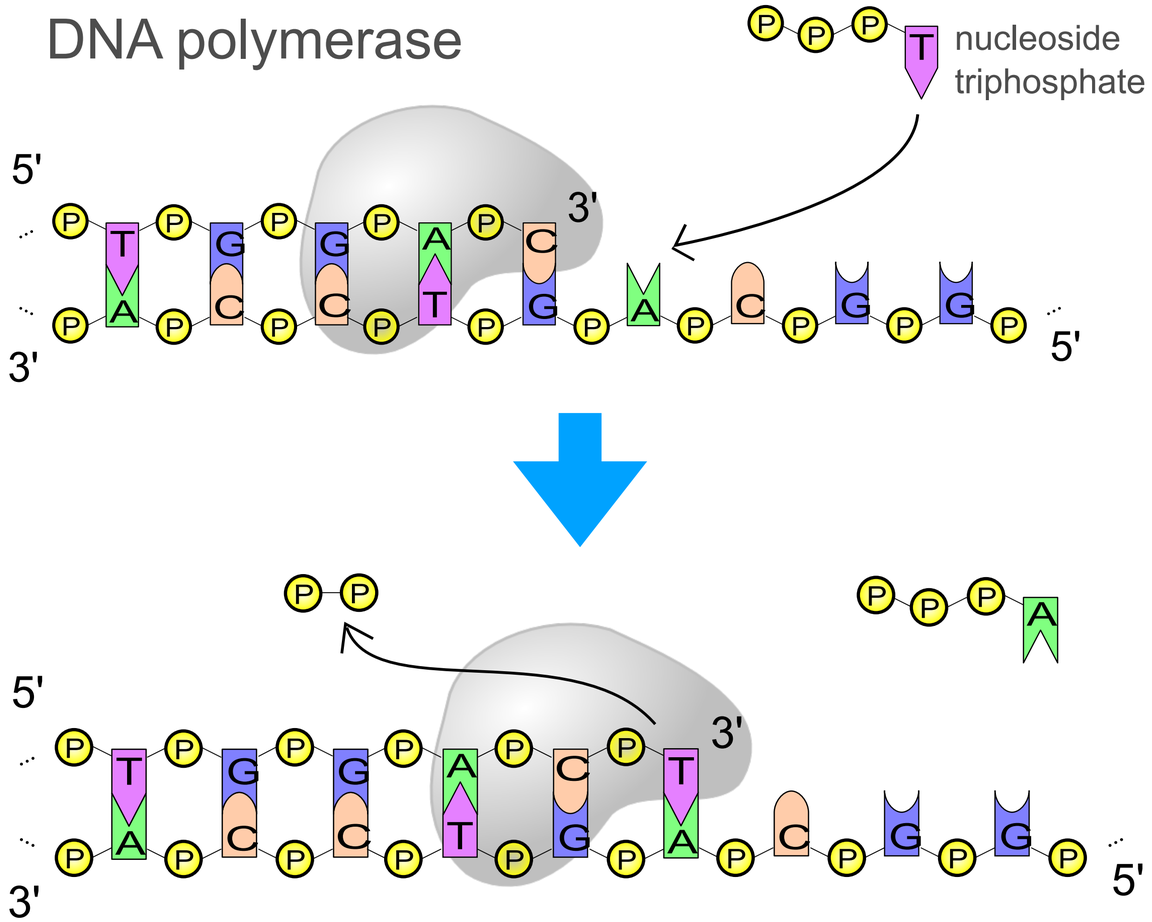


https://en.wikipedia.org/wiki/Complementarity_(molecular_biology)

How DNA sequences were replicated

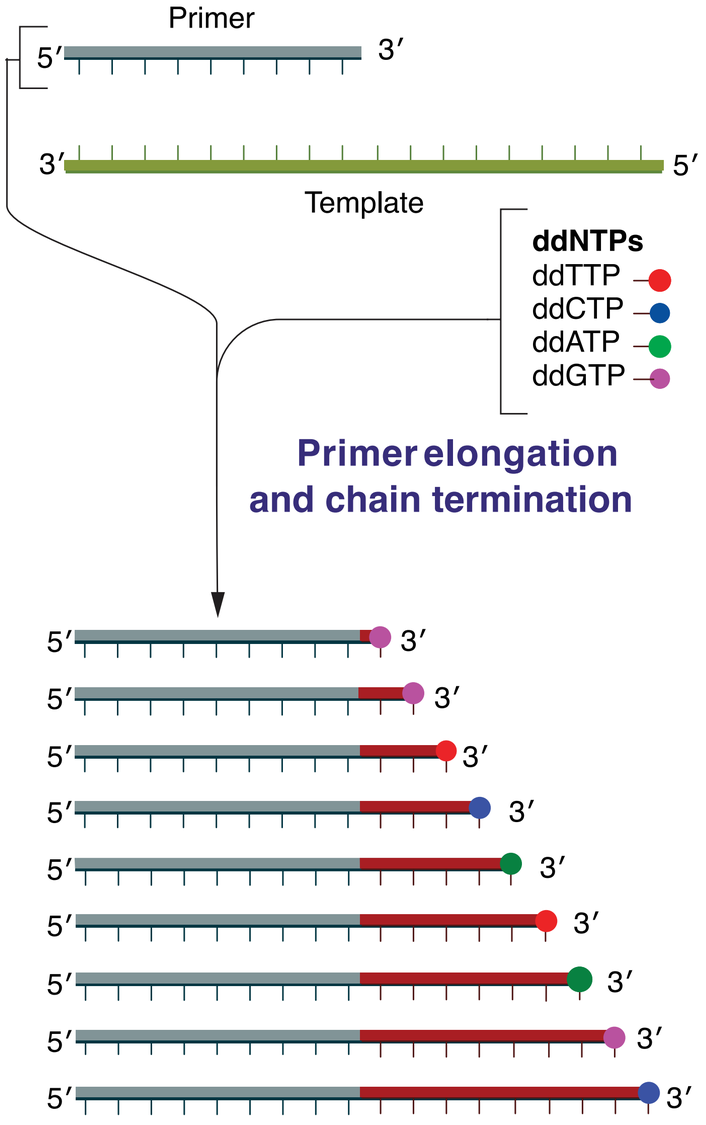


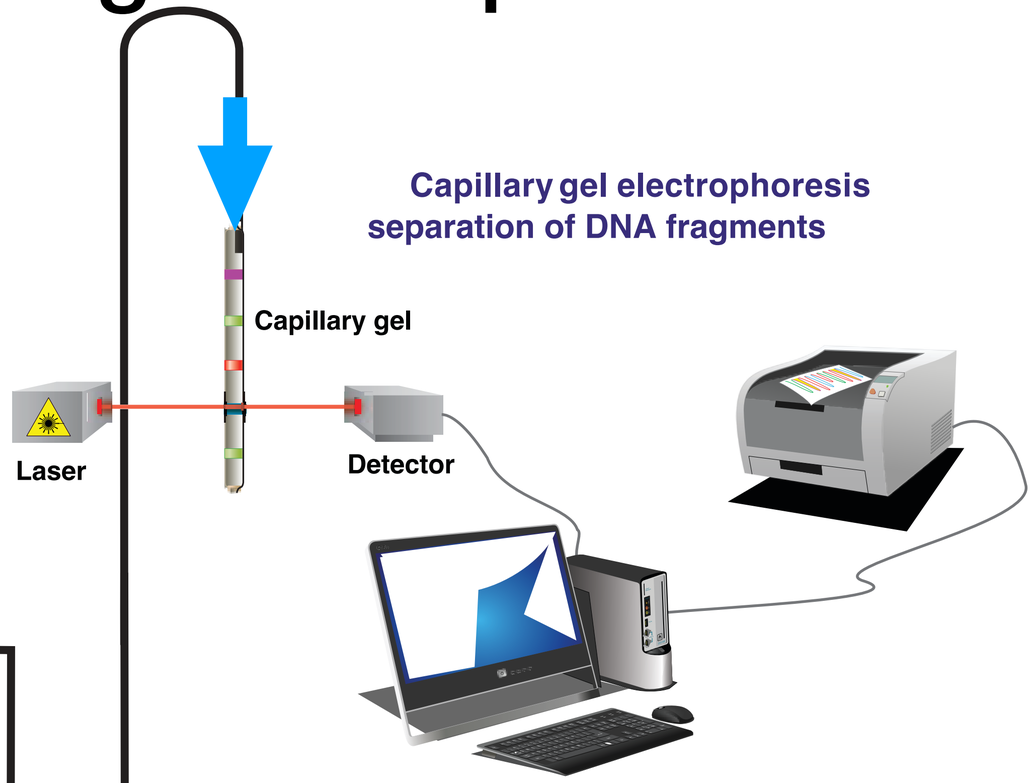




https://en.wikipedia.org/wiki/DNA_replication

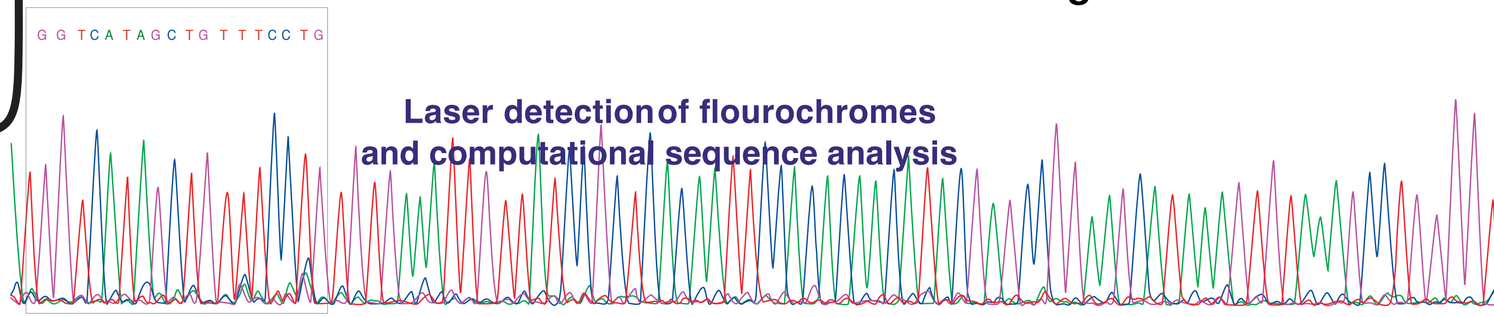
How do we read a DNA sequence as if we were copying the sequence?





Frederick Sanger

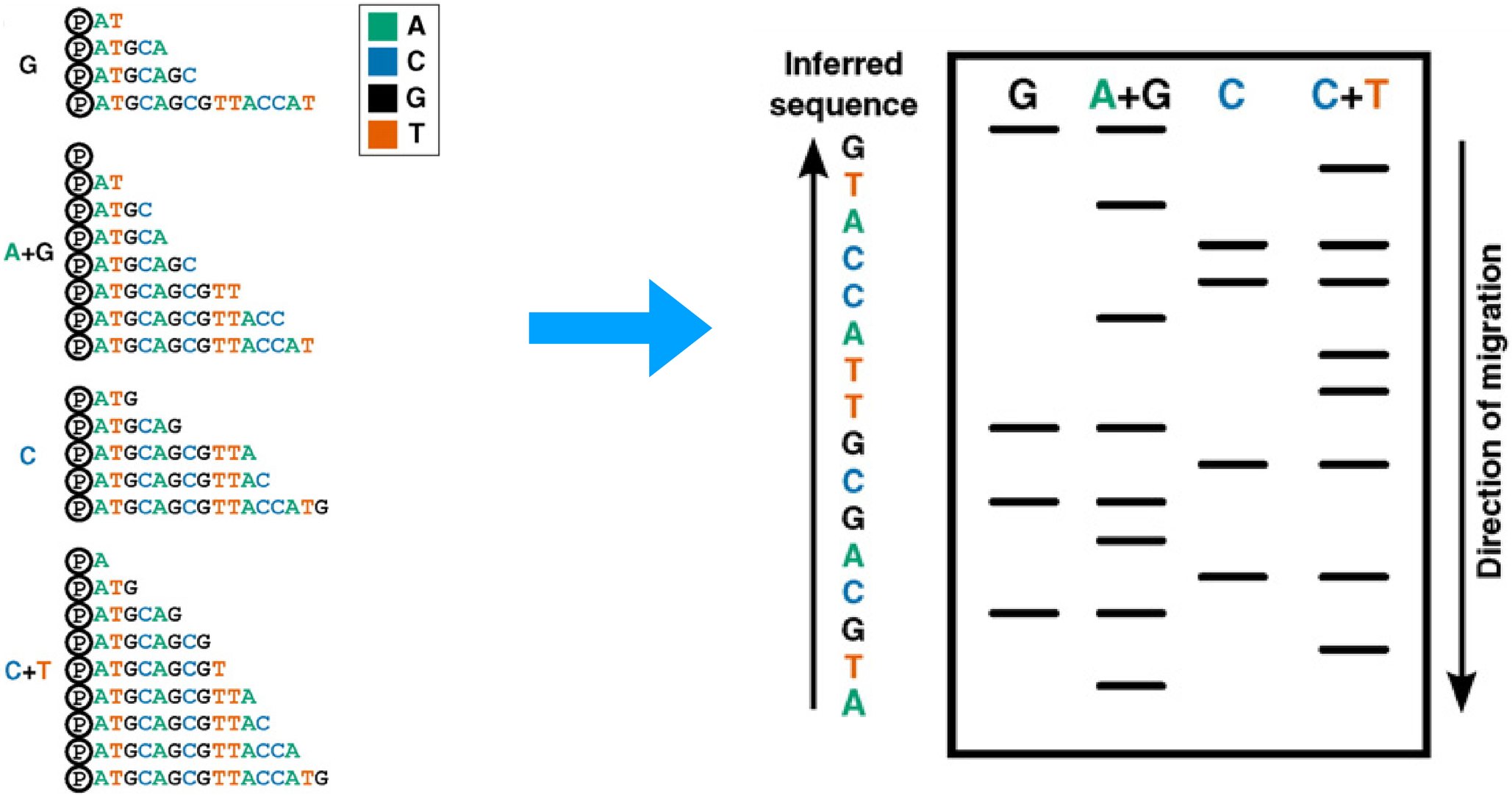
Leroy Hood



Chromatograph

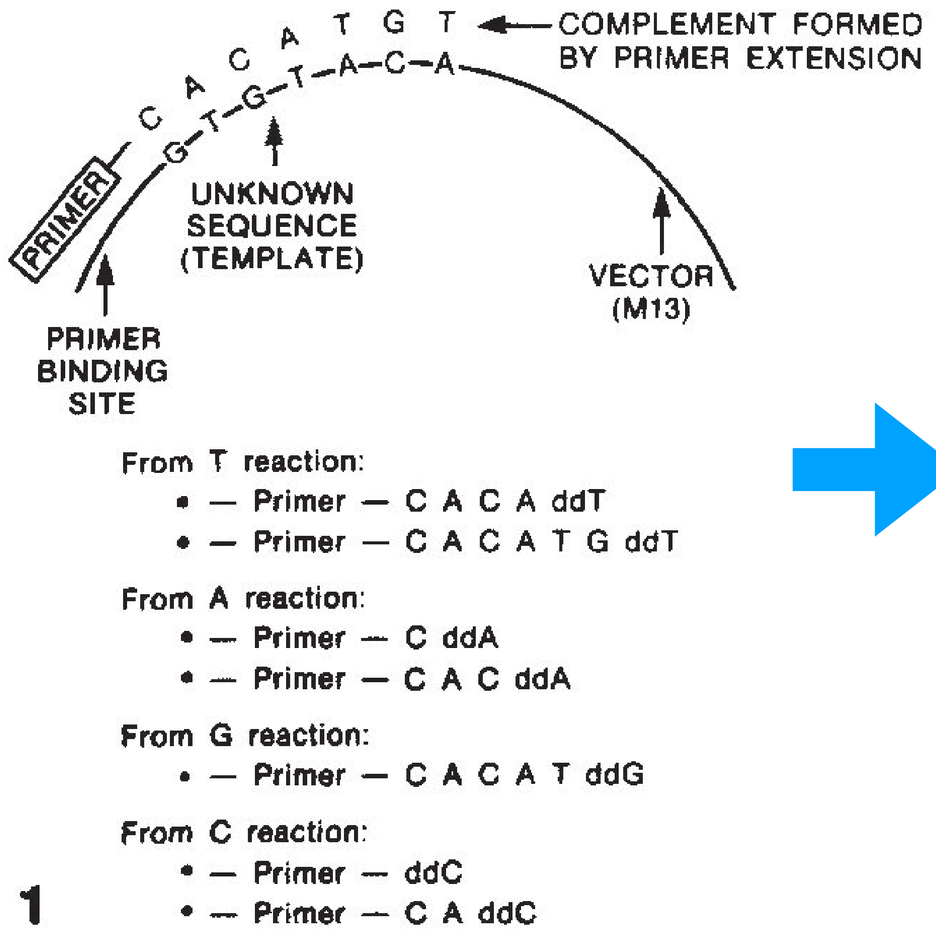
https://en.wikipedia.org/wiki/Sanger_sequencing

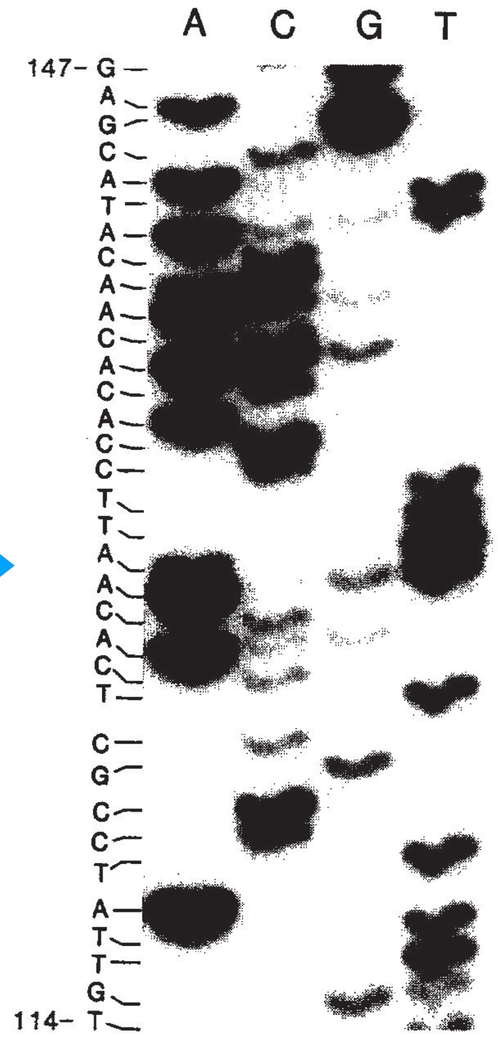
A long time ago before an automated sequencer

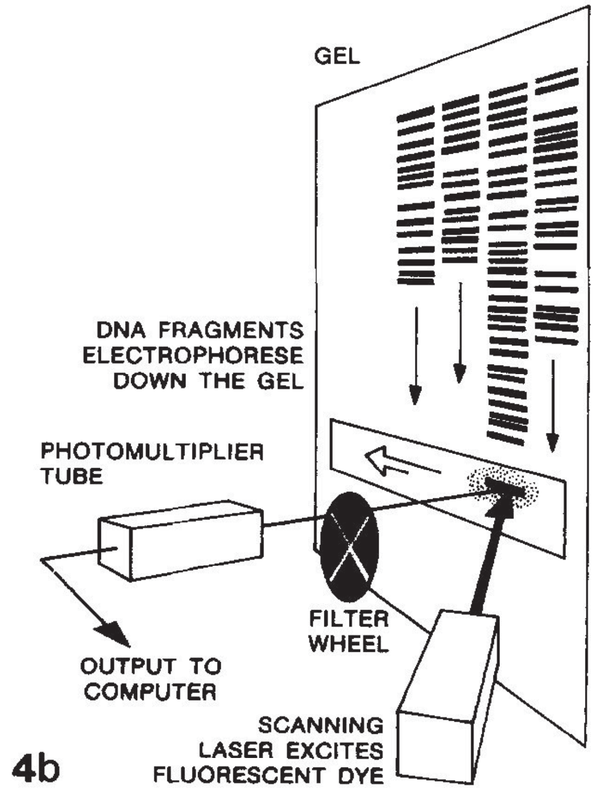


J.M. Heather, B. Chain /Genomics 107 (2016) 1–8

Can we automate DNA sequencing?

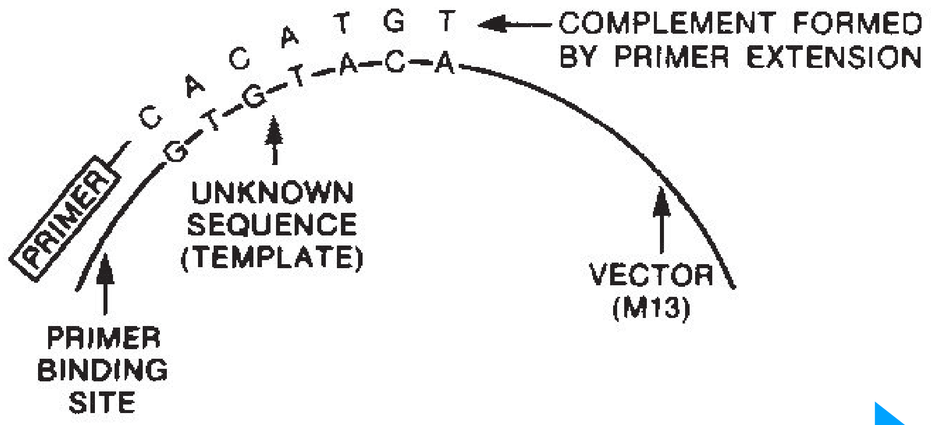






Hood *et al.* (1987)

Can we read them via a faster flow?



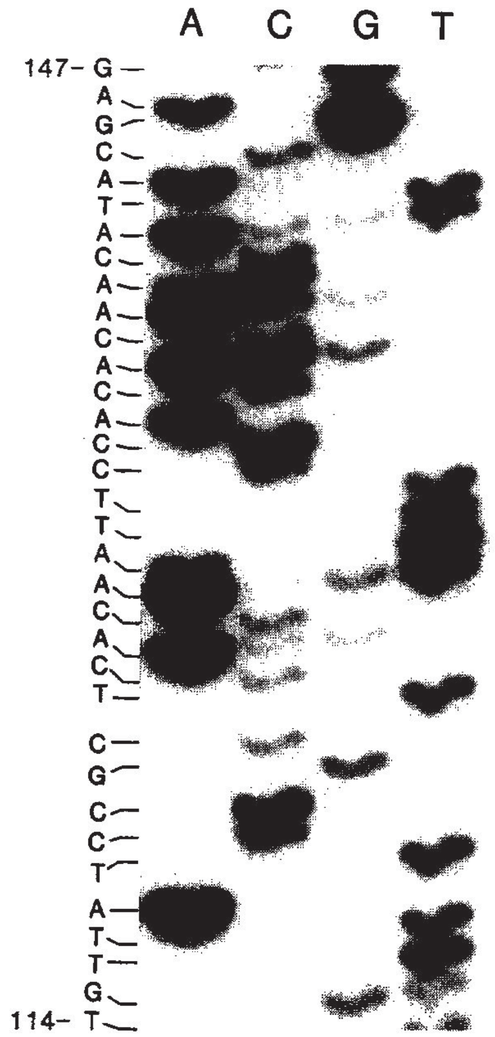
From T reaction:
- — Primer — C A C A ddT
- — Primer — C A C A T G ddT

From A reaction:
- — Primer — C ddA
- — Primer — C A C ddA

From G reaction:
- — Primer — C A C A T ddG

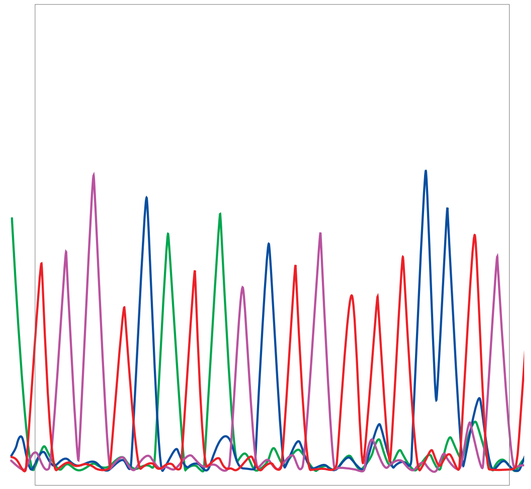
From C reaction:
- — Primer — ddC
- — Primer — C A ddC

1



Running four-coloured shot reads (sorted by the size of fragments) read them one by one fast by laser

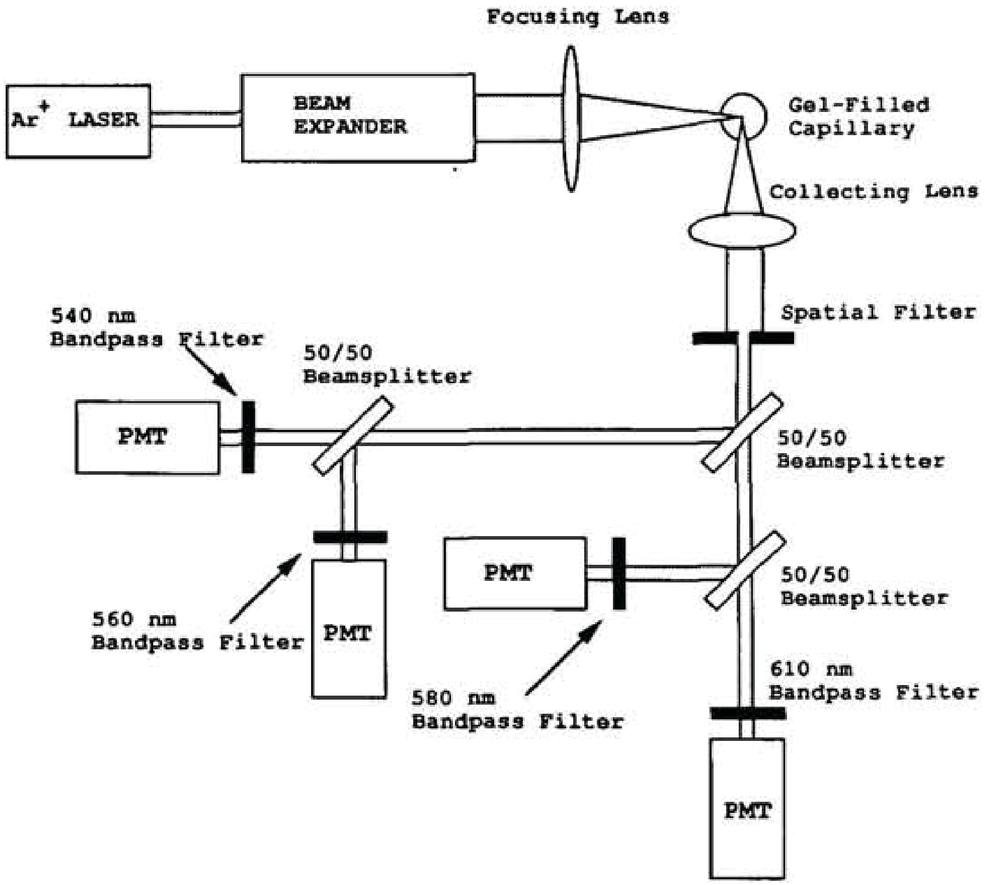
vs.



Chromatograph

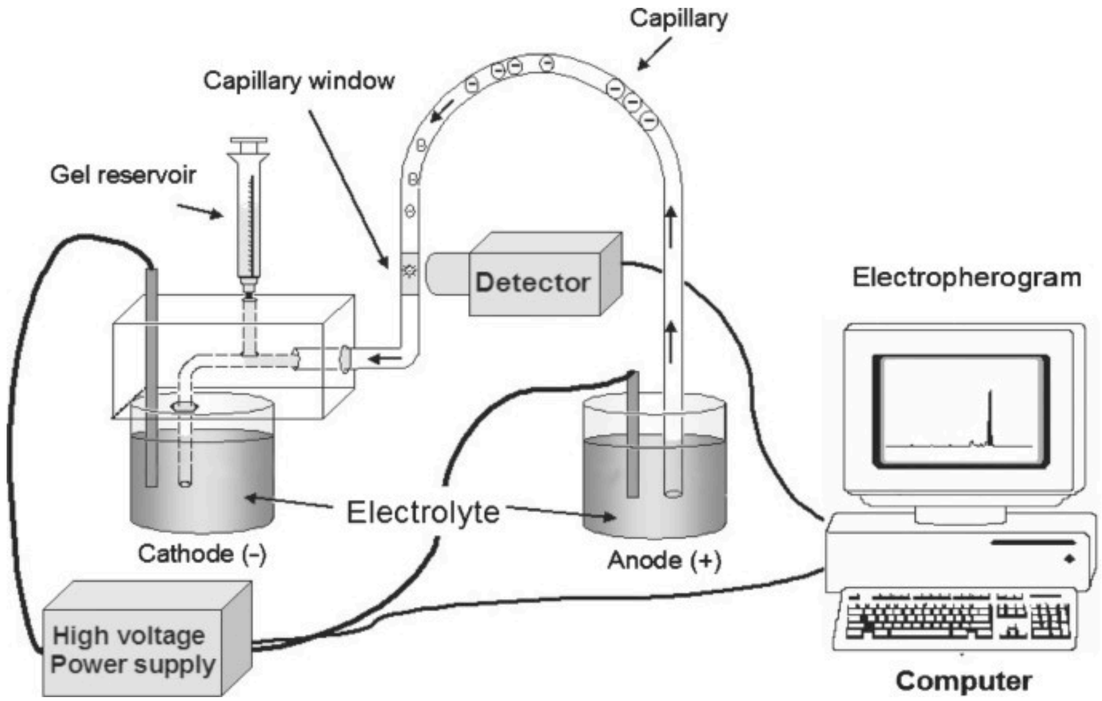
Hood *et al.* (1987)

Capillary electrophoresis to speed up the reading process



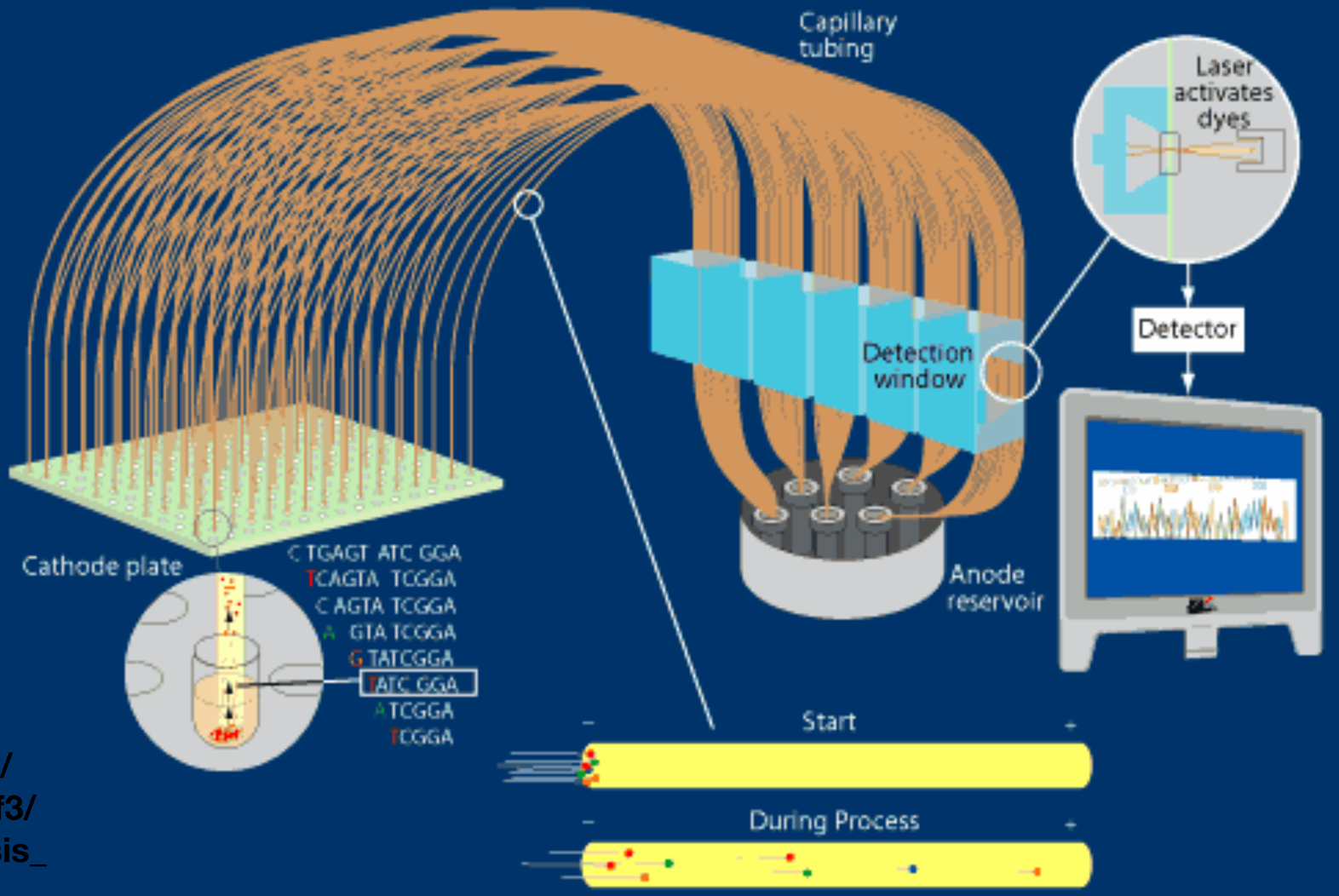
Luckey *et al*. NAR (1990)

Why DNA will move through capillary?



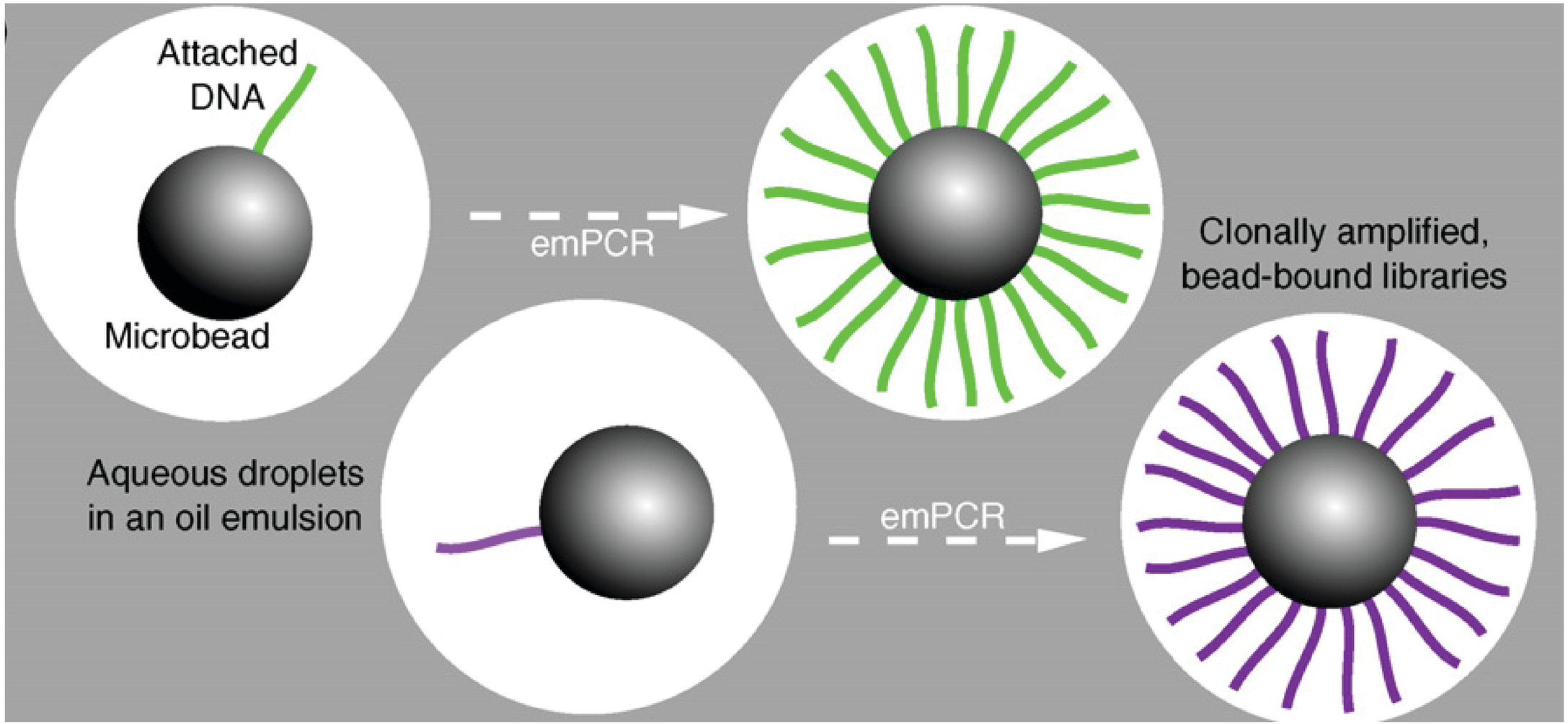
Beatriz Sanchez-Vega

Parallelizing many capillaries to further speed up



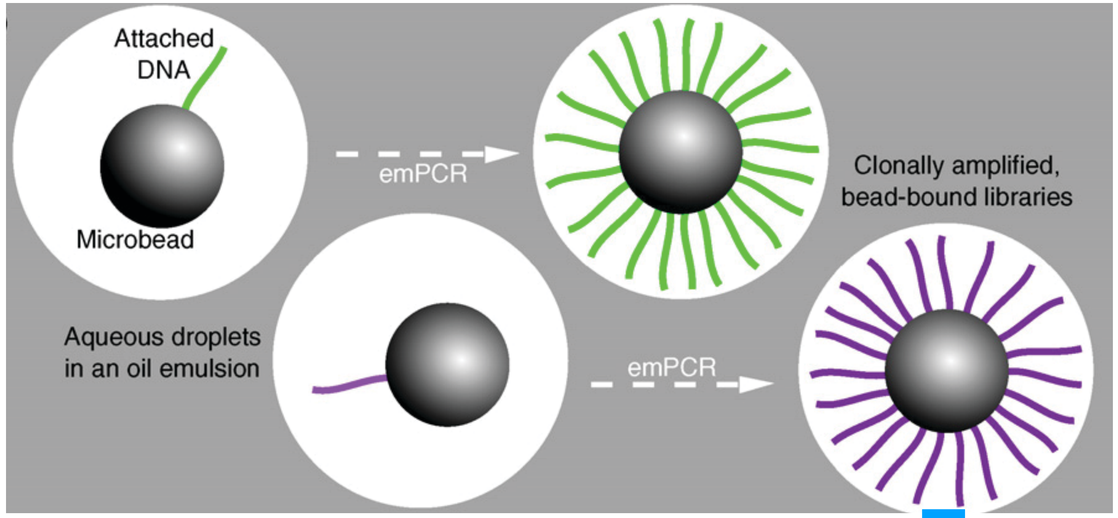
https://upload.wikimedia.org/wikipedia/commons/f/f3/Capillary_electrophoresis_sequencing.png

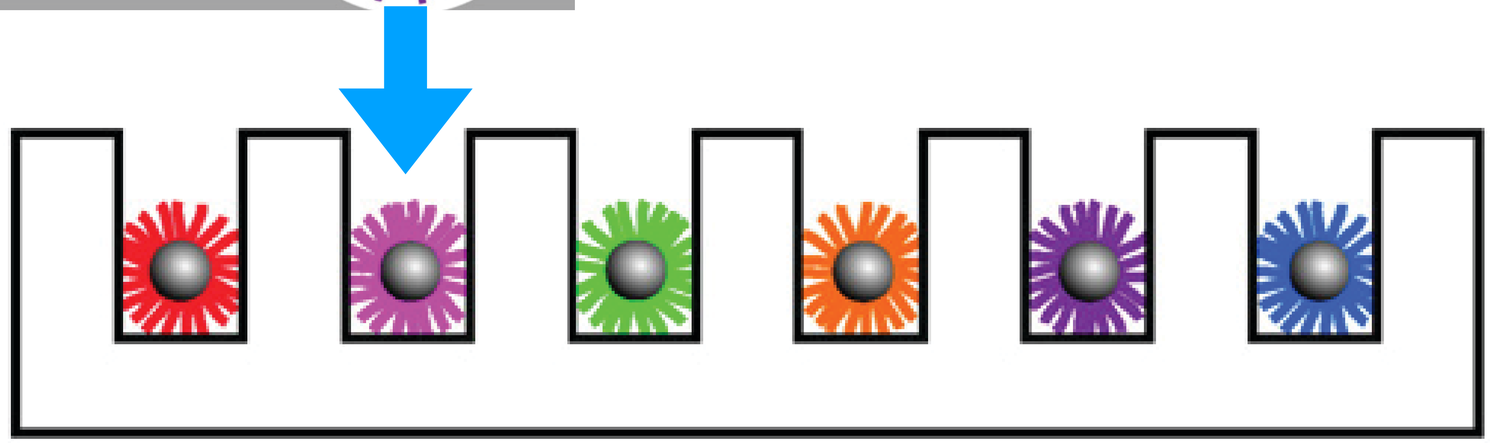
Micro bead-based applications followed by massively parallel short read sequencing



J.M. Heather, B. Chain /Genomics 107 (2016) 1–8

Micro bead-based applications followed by massively parallel short read sequencing

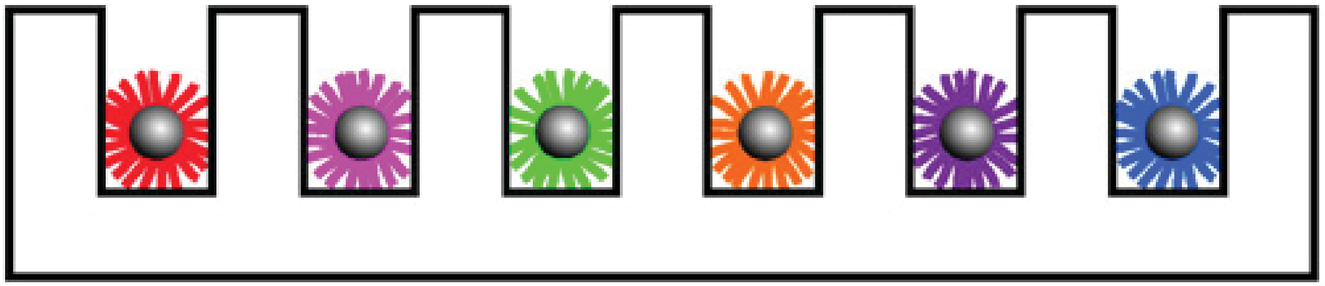




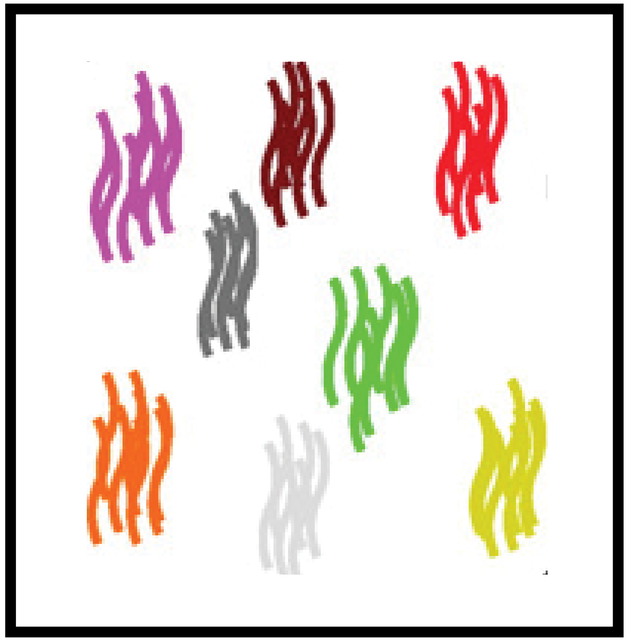
Picotitre wells containing DNA-bearing microbreads

J.M. Heather, B. Chain /Genomics 107 (2016) 1–8

Micro bead-based applications followed by massively parallel short read sequencing



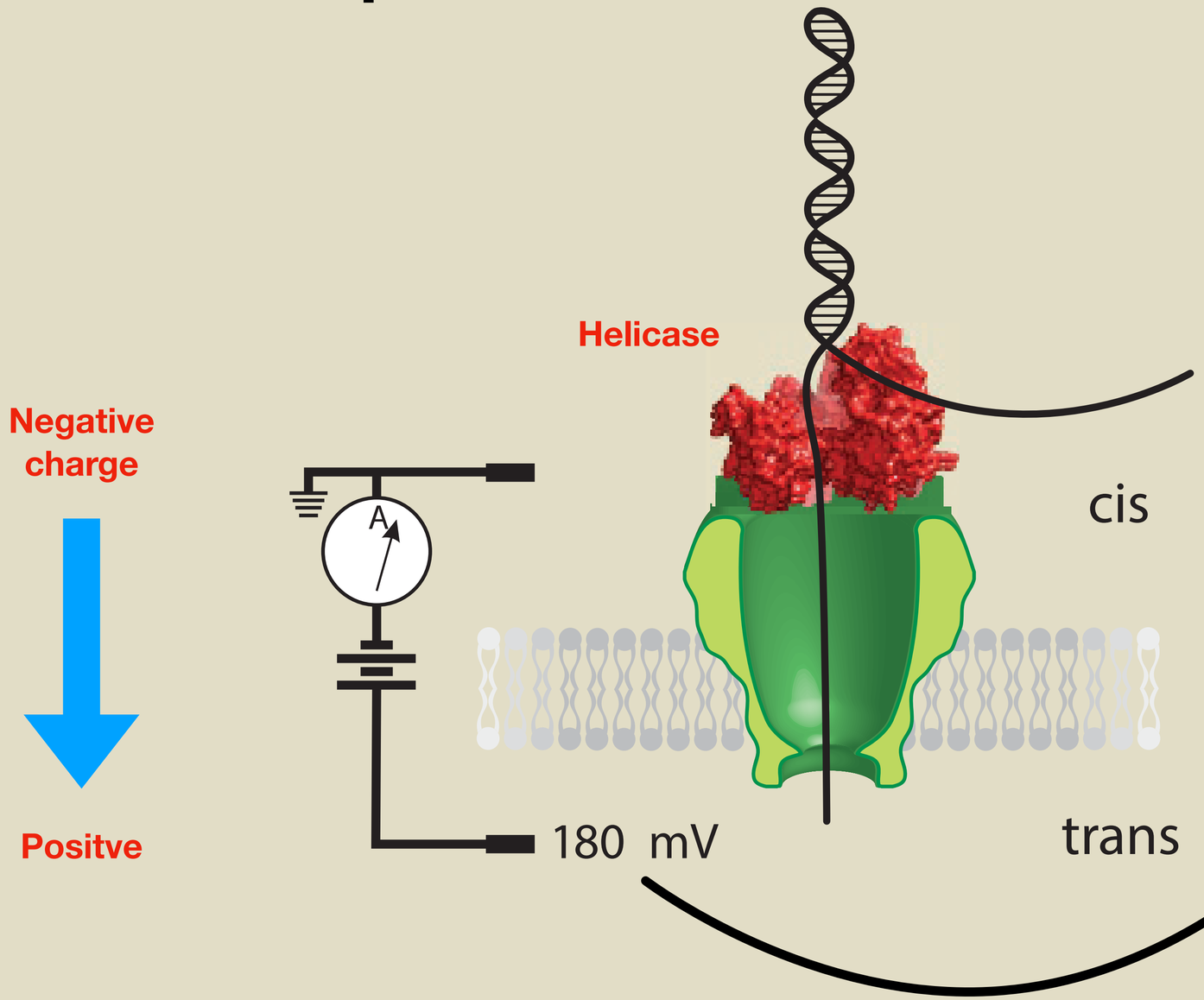
Picotitre wells containing DNA-bearing microbreads

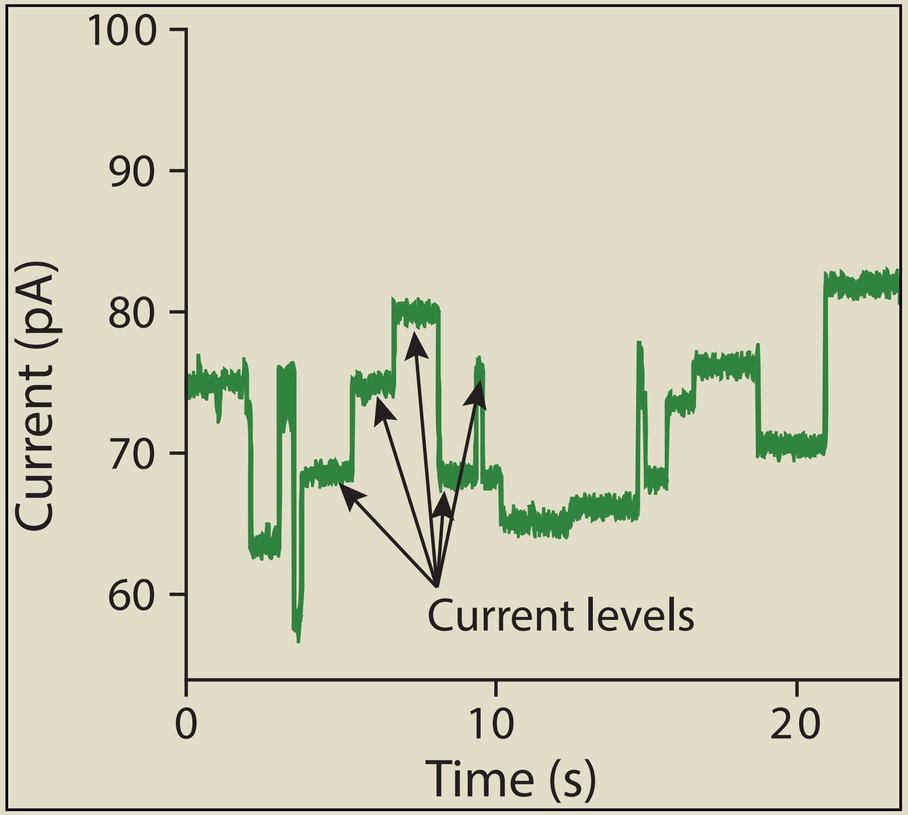


Flow cell-bound clusters

J.M. Heather, B. Chain /Genomics 107 (2016) 1–8

Nanopore: Yet another method for sequencing

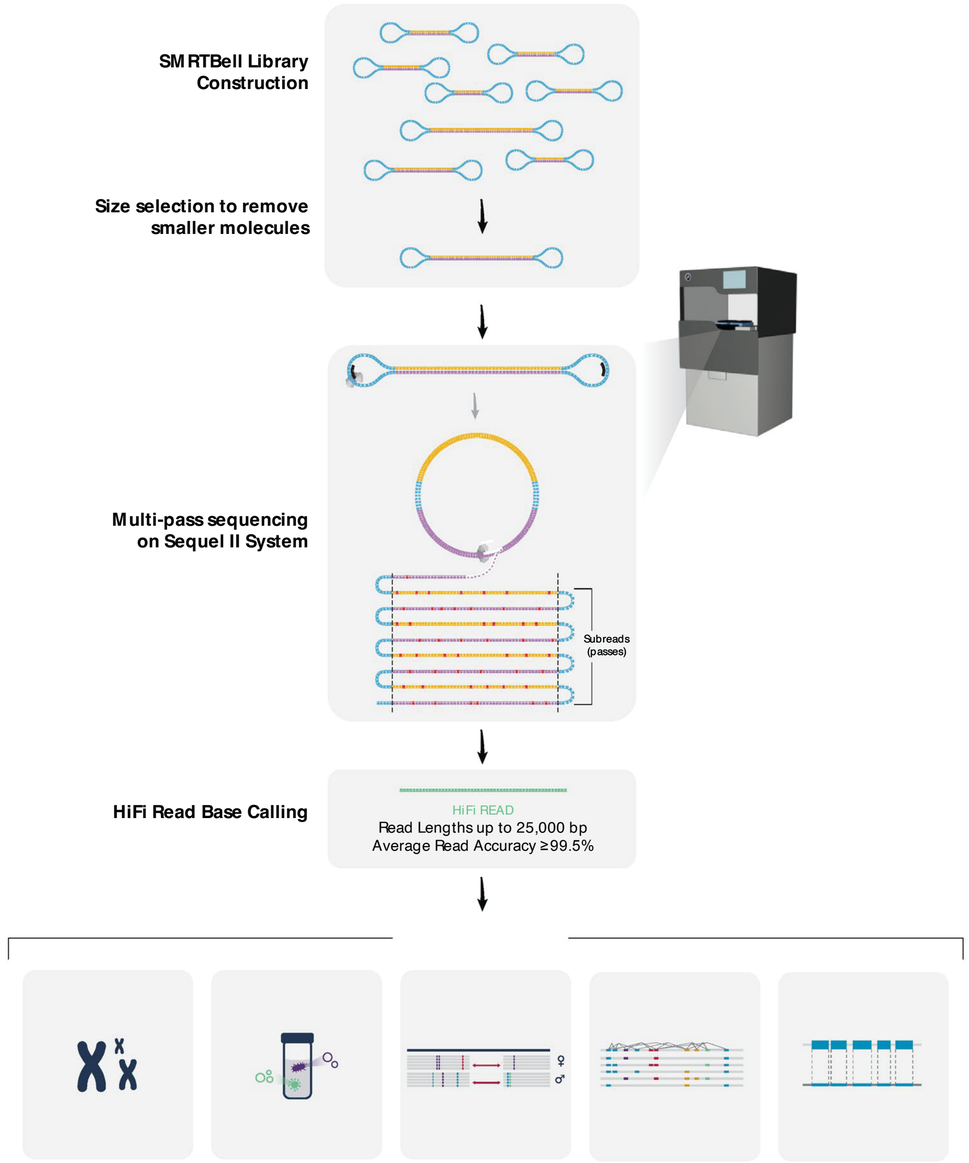




- The ionic conductivity is sensitive to the presence of the nucleotide.
- Charatersitic "squiggle"

Deamer, Akeson, Branton, Nature Biotech. (2016)

Why not read the full length of a genome?



Highly accurate long-read HiFi sequencing data for five complex genomes

Ting Hon[1], Kristin Mars[1], Greg Young[1], Yu-Chih Tsai[1], Joseph W. Karalius[1], Jane M. Landolin[2], Nicholas Maurer[3], David Kudrna[4], Michael A. Hardigan[5], Cynthia C. Steiner[6], Steven J. Knapp[5], Doreen Ware[7,8], Beth Shapiro[3,9], Paul Peluso[1] & David R. Rank[1]✉

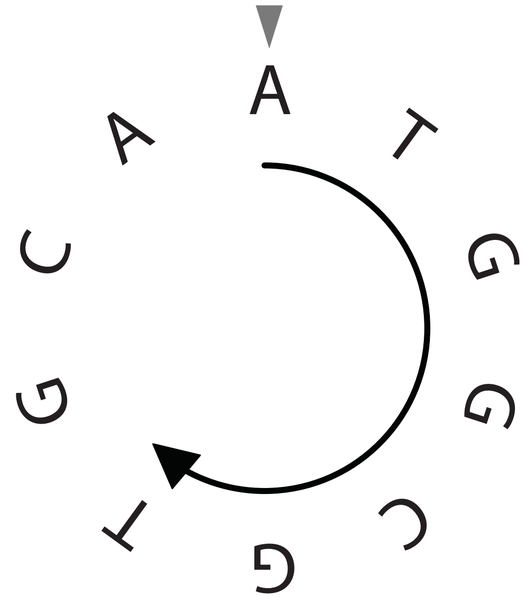
The PacBio® HiFi sequencing method yields highly accurate long-read sequencing datasets with read lengths averaging 10–25 kb and accuracies greater than 99.5%. These accurate long reads can be used to improve results for complex applications such as single nucleotide and structural variant detection,

State of the art

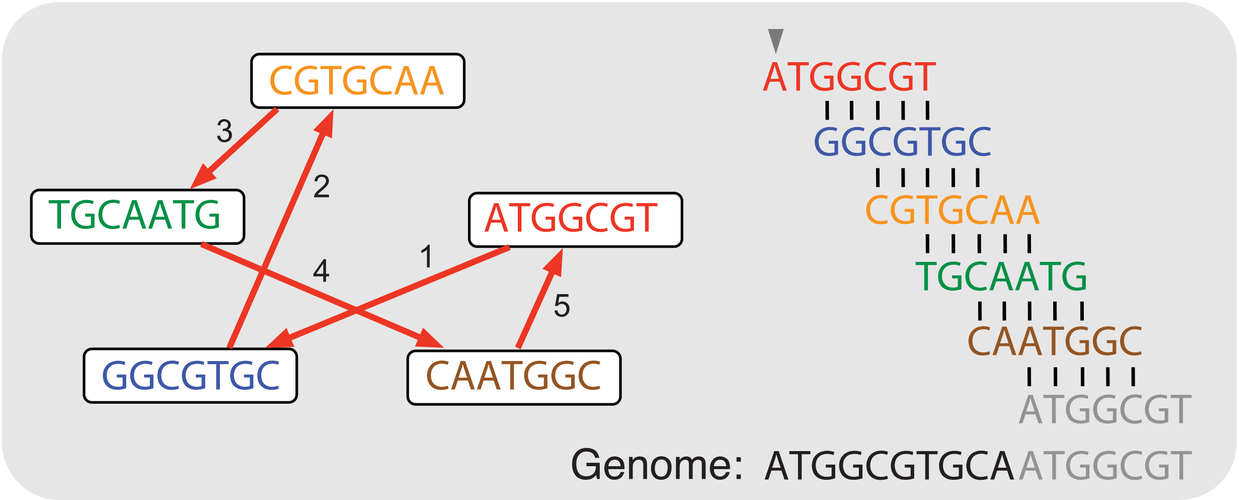
10-25kb length ≪ 3.2B

Hon *et al.* (2020)

Genome assembly: putting short reads together in one sequence

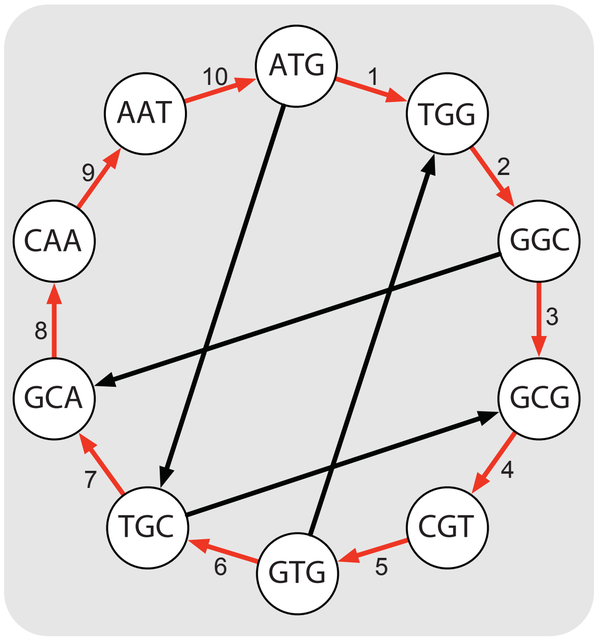


Short-read sequencing

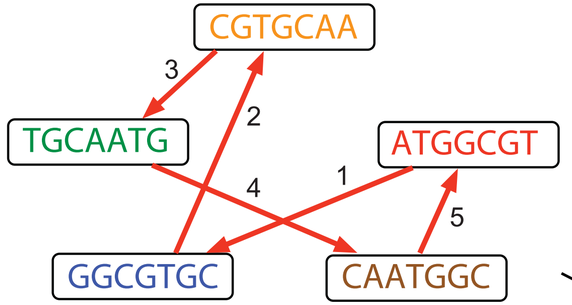


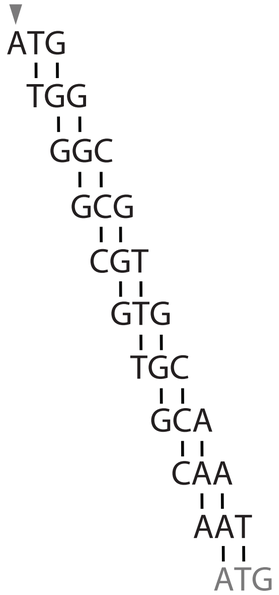
Compeau *et al.* Nature Biotech. (2011)

Genome assembly is a computational problem

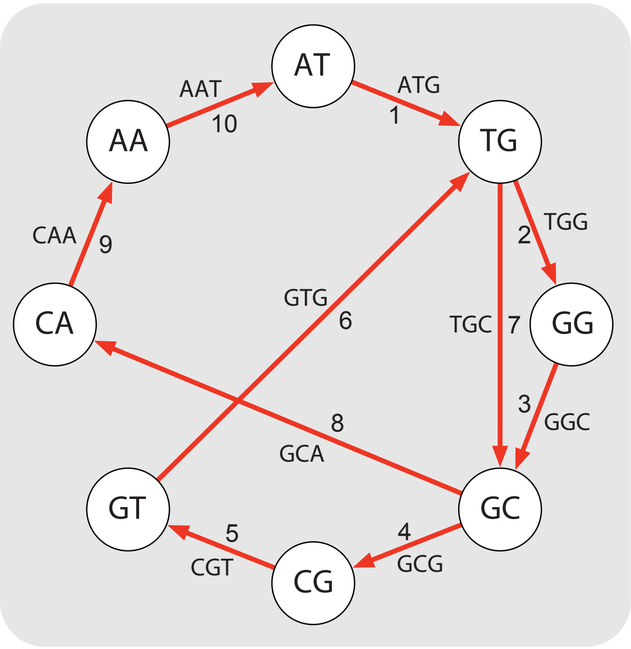


Hamiltonian cycle
Visit each vertex once
(harder to solve)



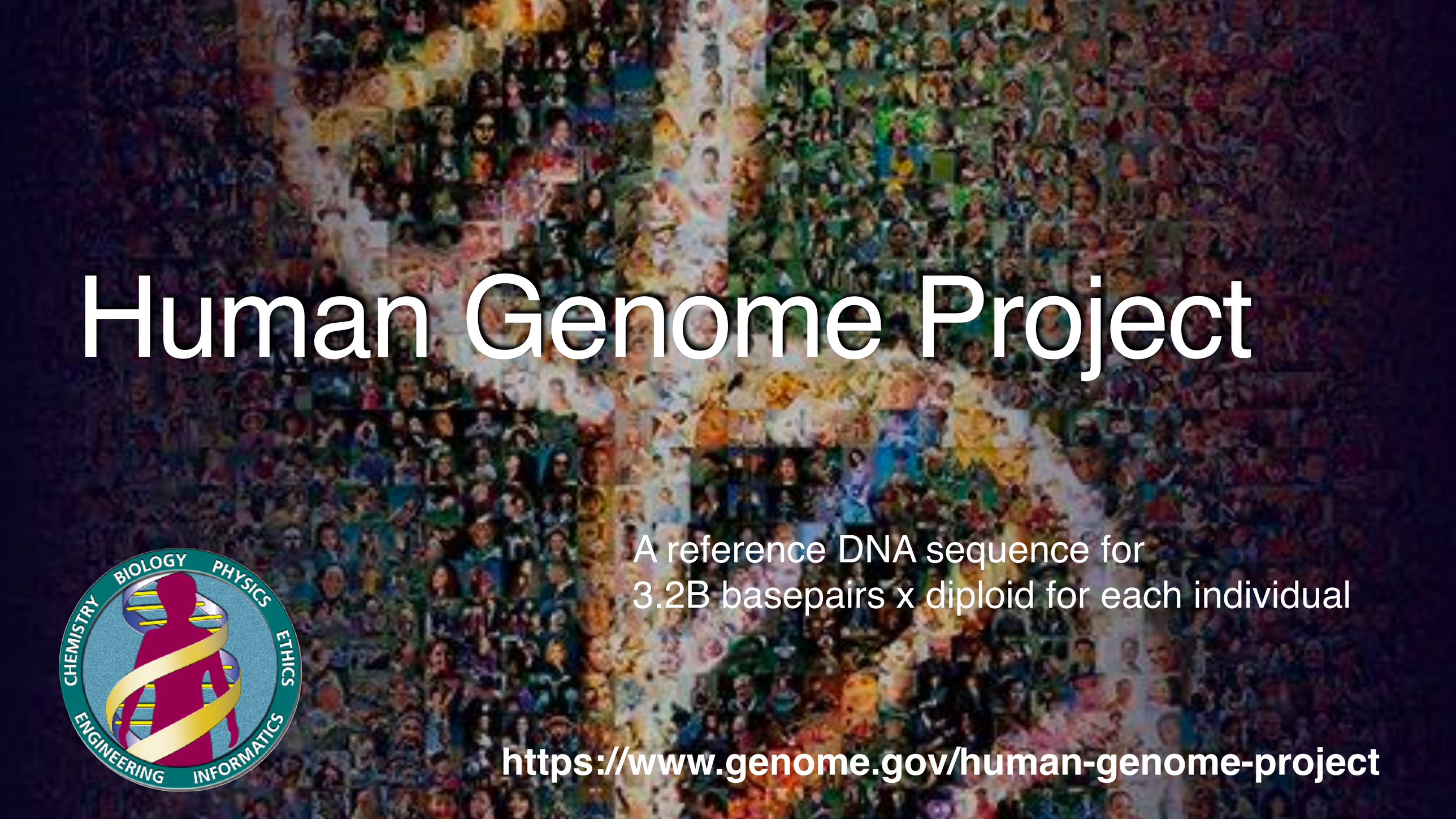


Genome: ATGGCGTGCAATG



Eulerian cycle
Visit each edge once
(easier to solve)

Compeau *et al.* Nature Biotech. (2011)



Human Genome Project

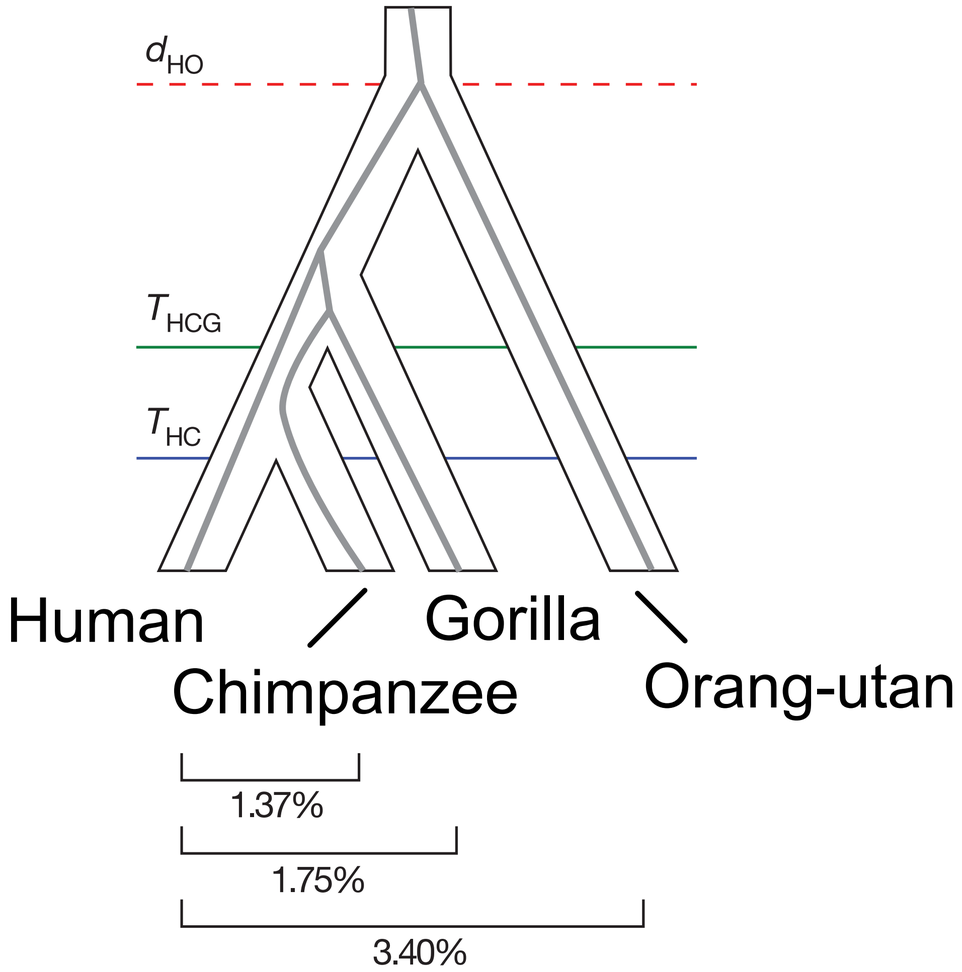
A reference DNA sequence for
3.2B basepairs x diploid for each individual

https://www.genome.gov/human-genome-project

How did Human Genome Project start the revolution?

We have a reference panel of genomic sequence information...

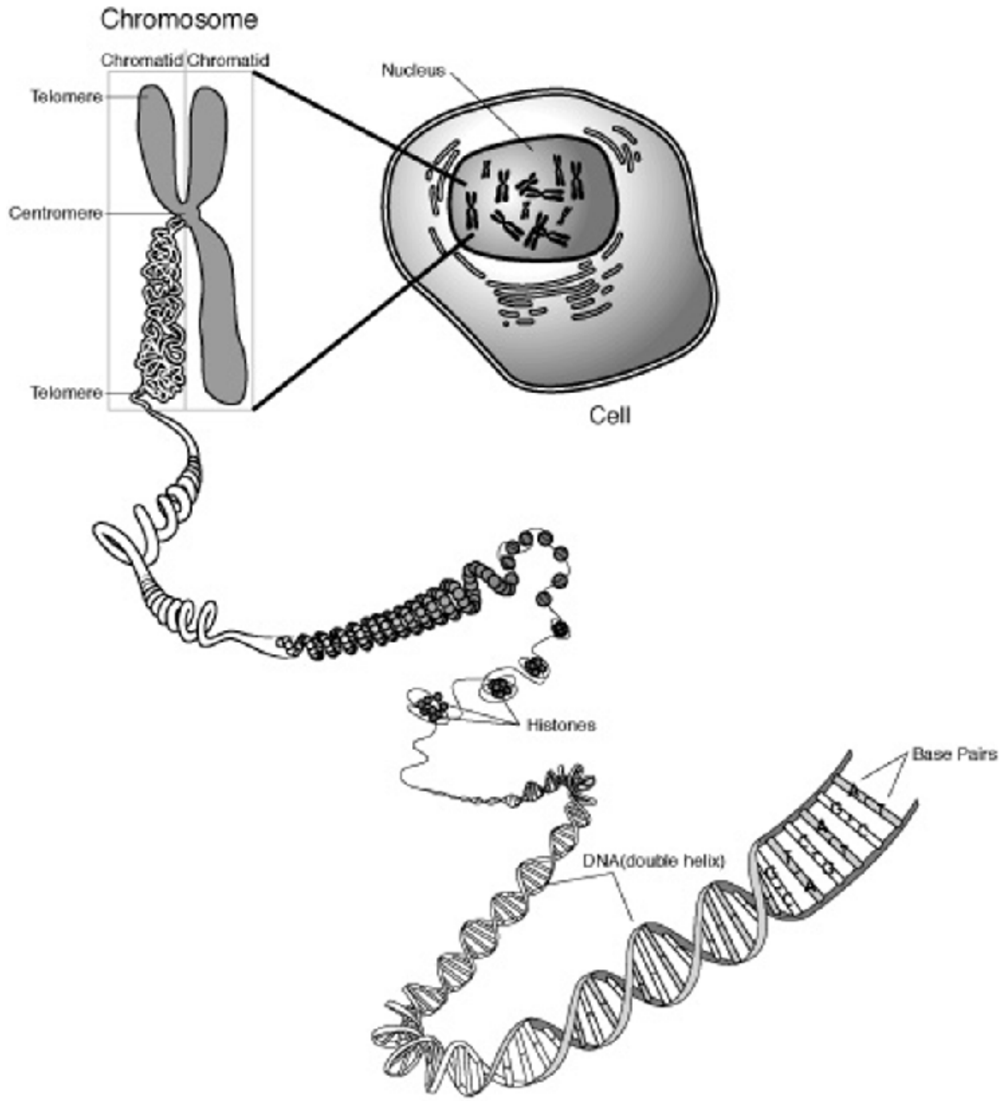
Why is it important to me?

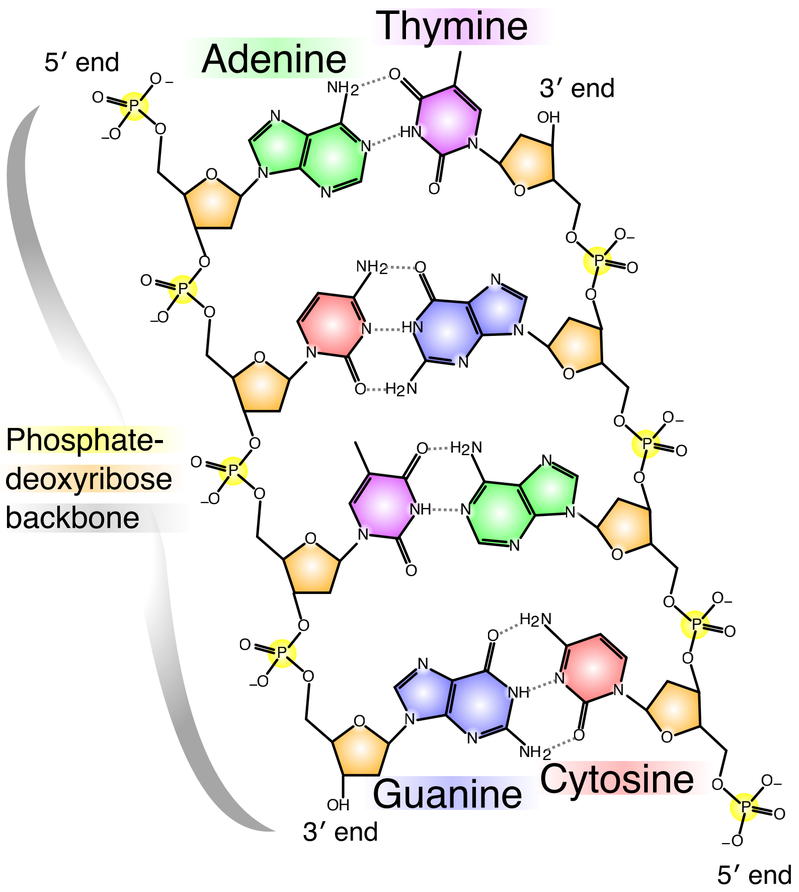


99% genetic information shared across humans

Scally .. Durbin, Nature (2012)

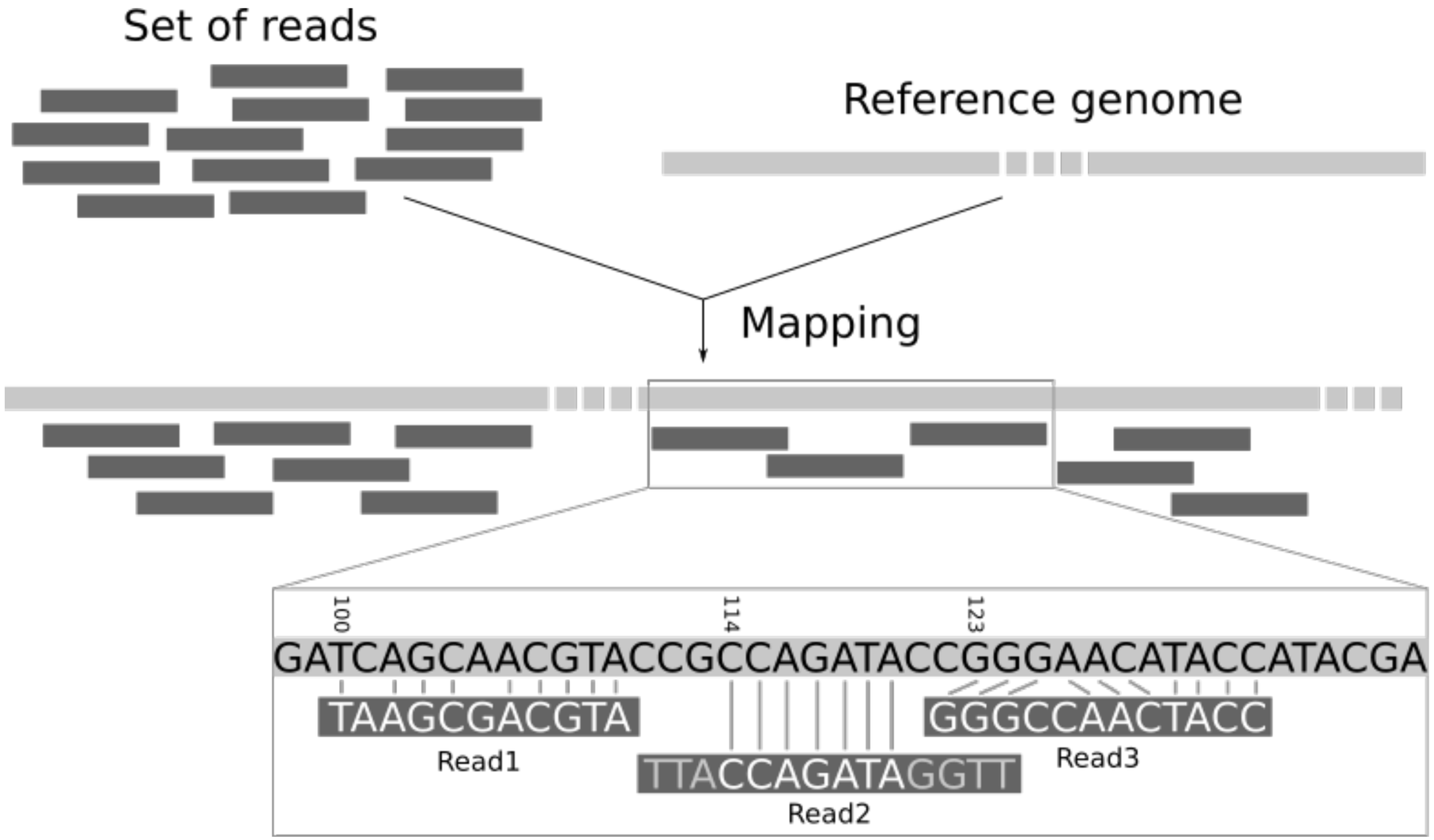
A reference genome can serve as an efficient computing tool (a template for matching)





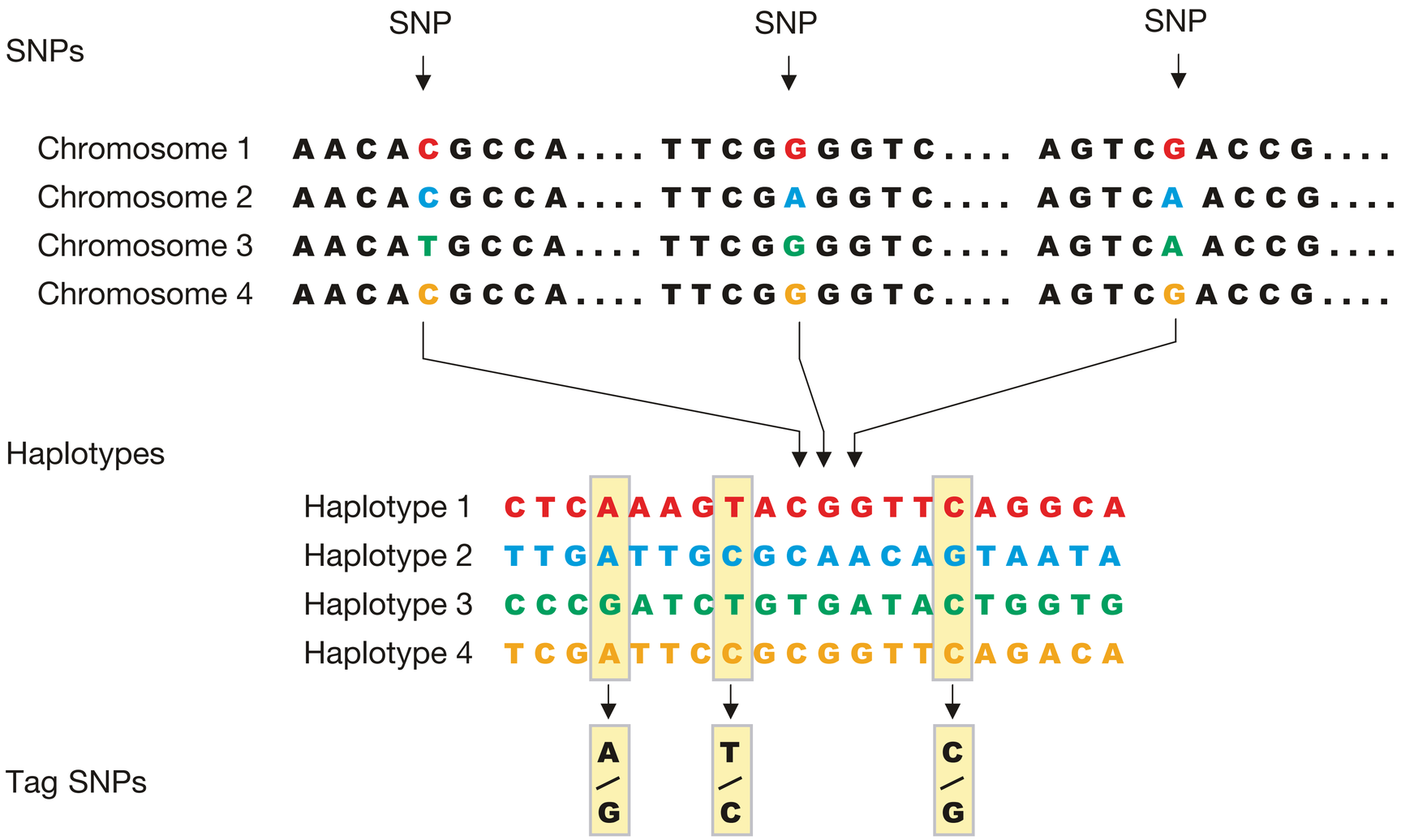
Given reference, we can look up locations

Given reference, we can look up locations

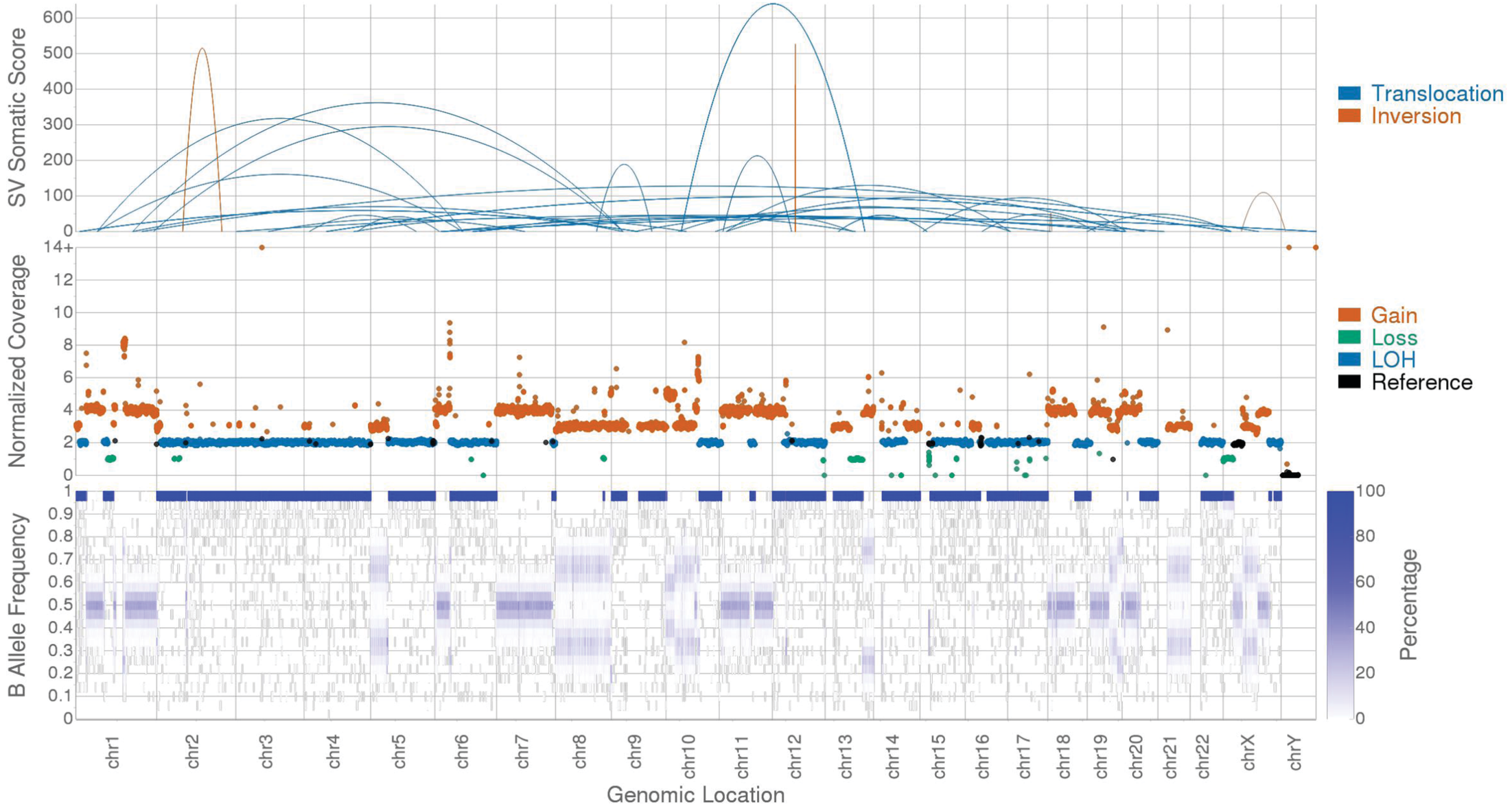


https://training.galaxyproject.org/

DNA variation in genetics

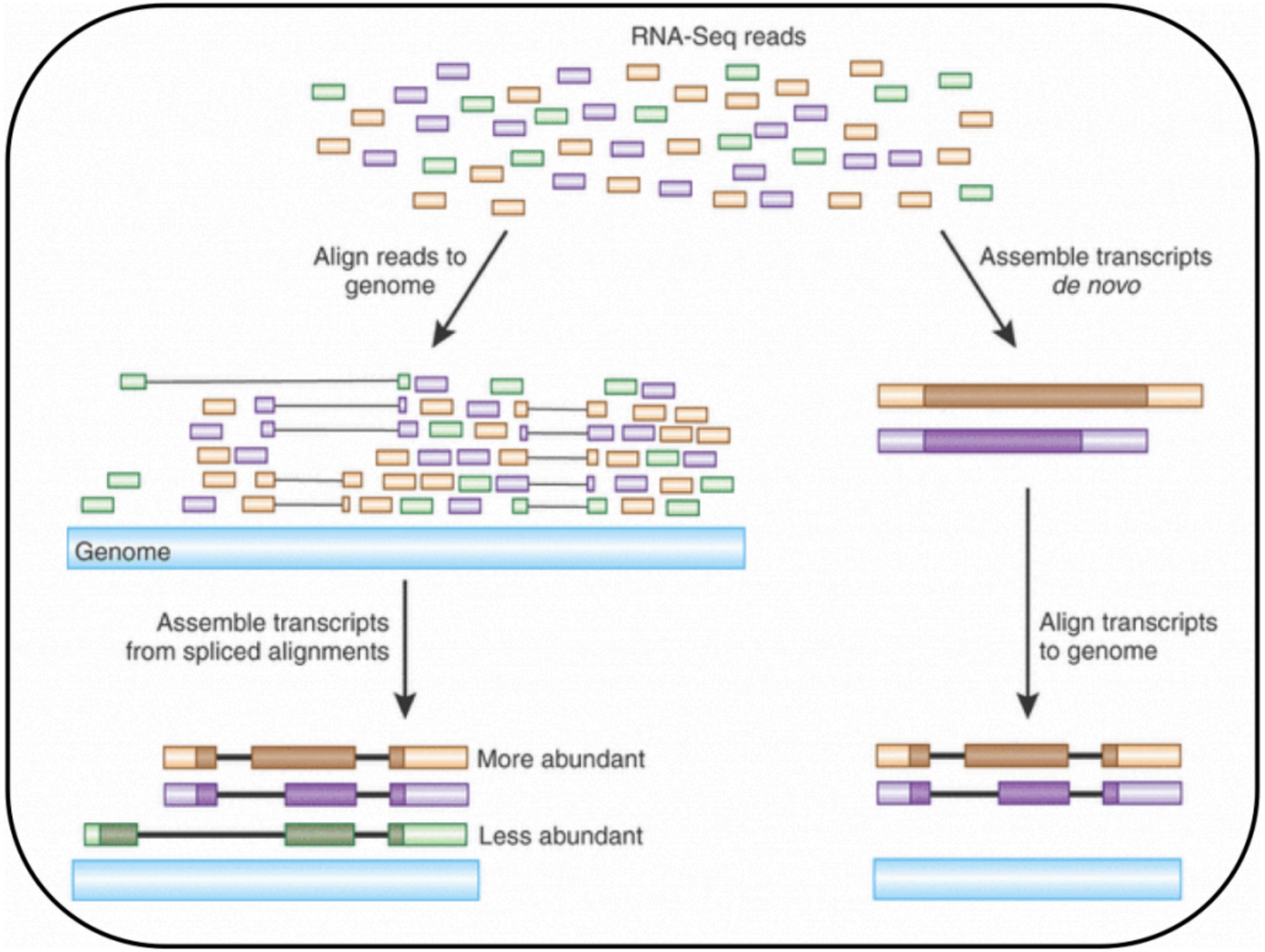


Hapmap Project



Illumina Cancer Analysis Service

Quantification of protein-coding genes: (1) convert mRNA to cDNA (2) match/map

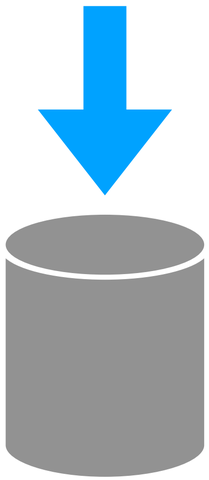


Haas BJ and Zody MC. Advancing RNA-seq analysis. 28:421-423

How do we align/map/match a short read to known genomic DNAs?

Computational Method

```
TTATATTGAATTTTCAAAAATTCTTACTTTTT
TTTTGGATGGACGCAAAGAAGTTTAATAAT
CATATTACATGGCATTACCACCATATACATA
TCCATATCTAATCTTACTTATATGTTGTGGA
AATGTAAAGAGCCCCATTATCTTAGCCTAA
AAAAACCTTCTCTTTGGAACTTTCAGTAAT
ACGCTTAACTGCTCATTGCTATATTGAAGT
```



Index reference genome for a quick lookup (e.g., BWT, Bloom filter)

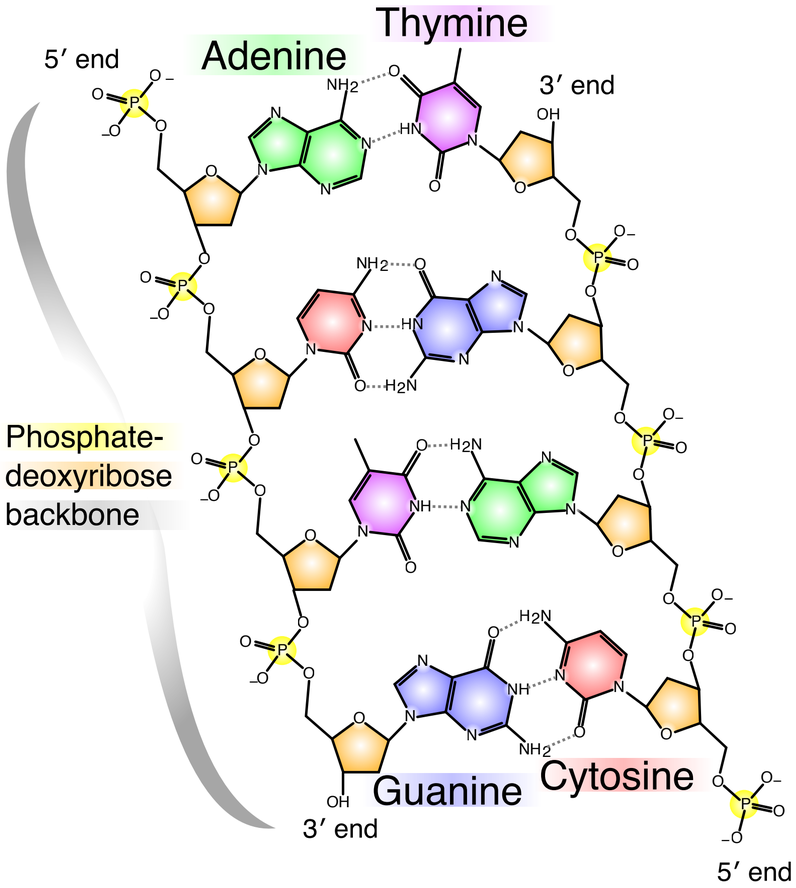
GCCCCATTATCTT?

vs.

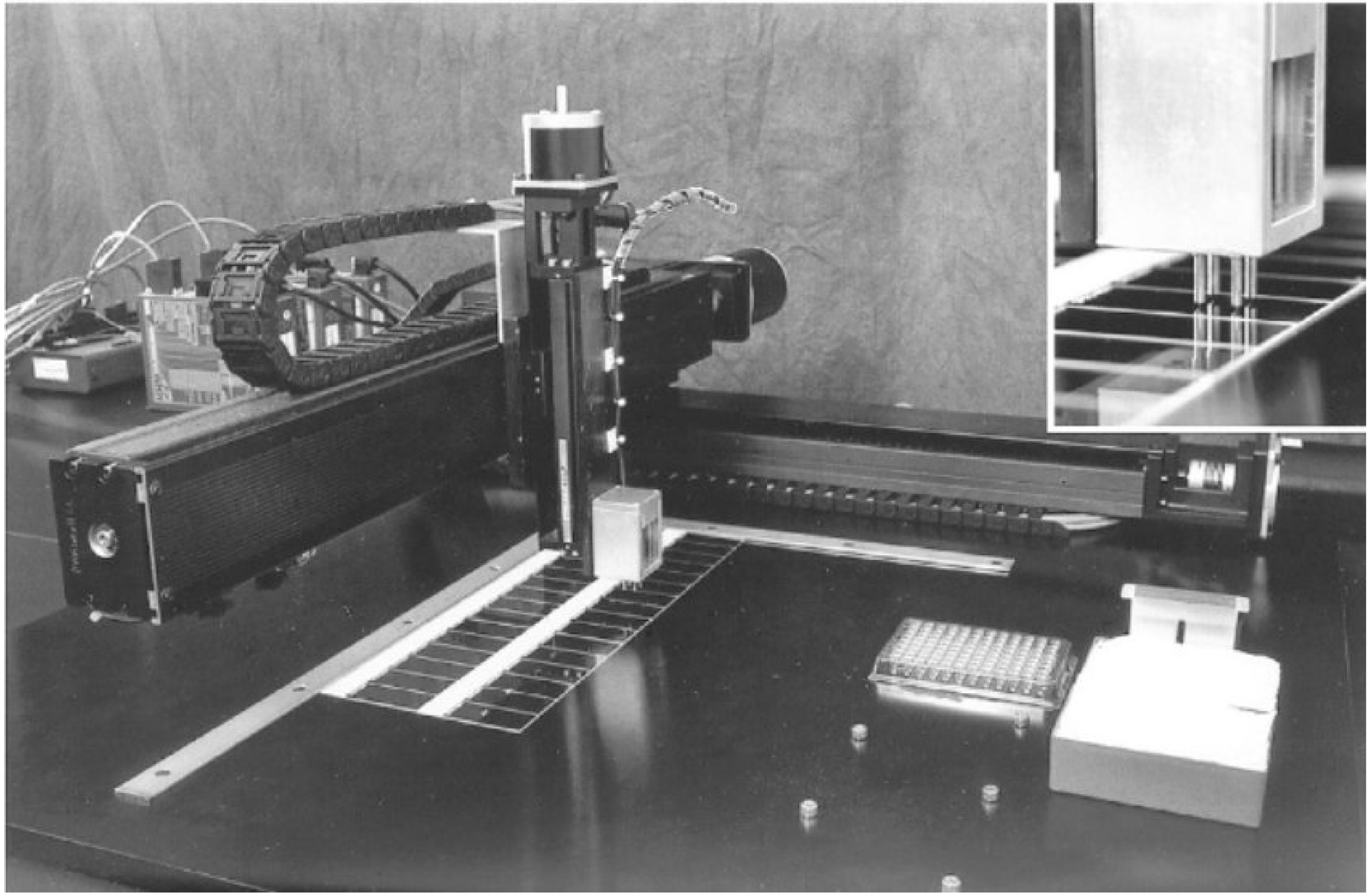
GCCTTTTTATCTT?

Can we fine this short read? If so, where?

Biochemical Method

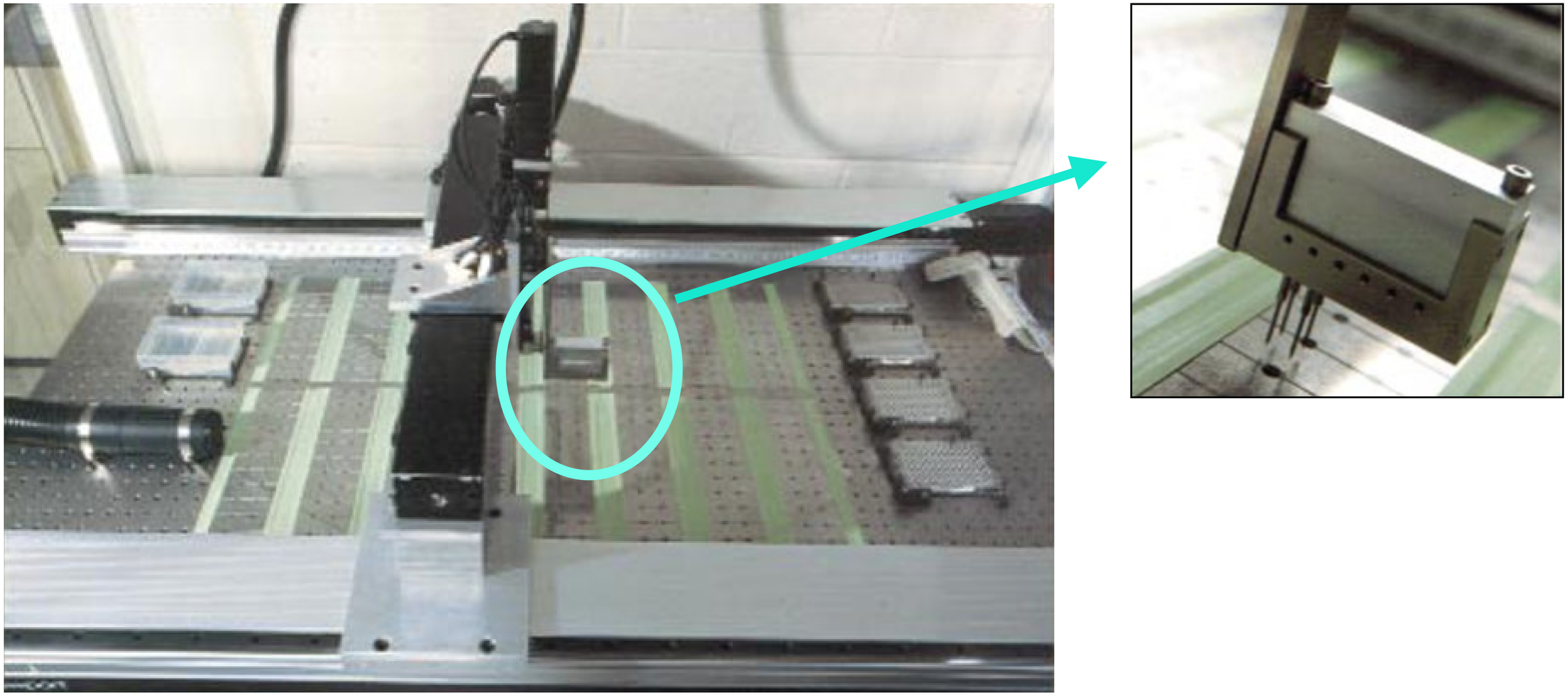


A long time ago…
Array technology (Pat Brown's pipetting machine)



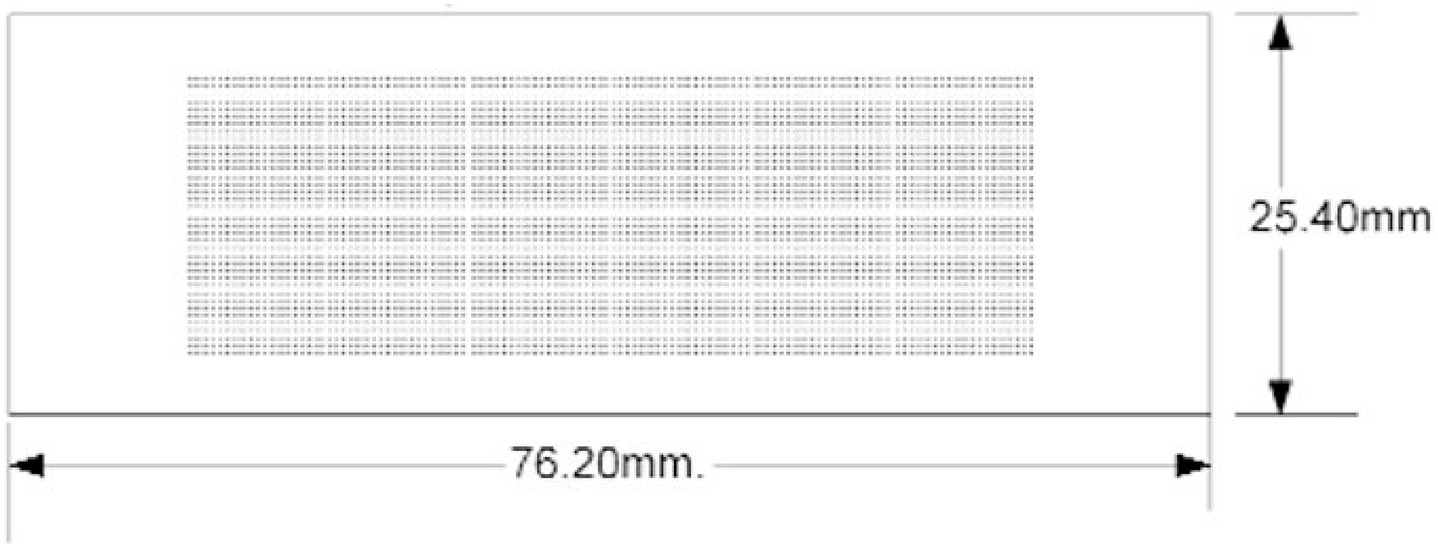
Cheung *et al*. (1999)

Pat Brown’s pipetting machine

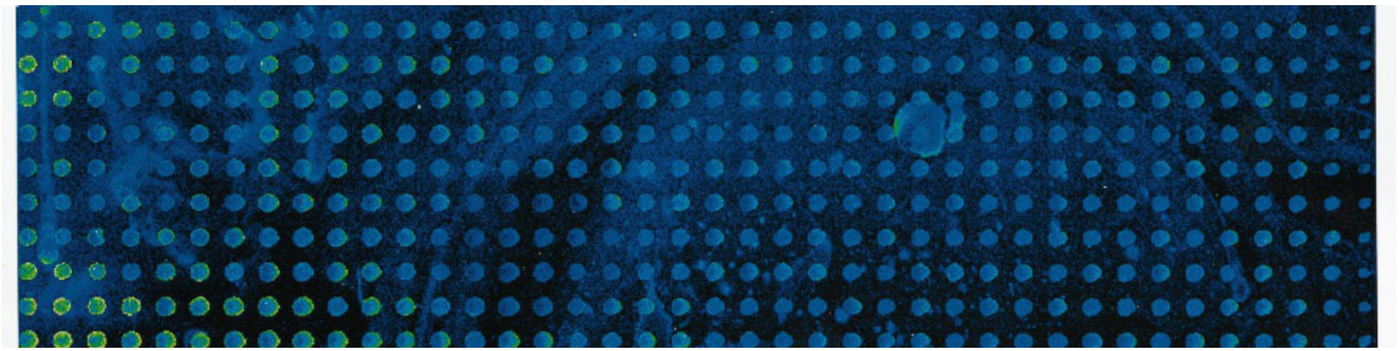


Cheung *et al*. (1999)

cDNA microarray

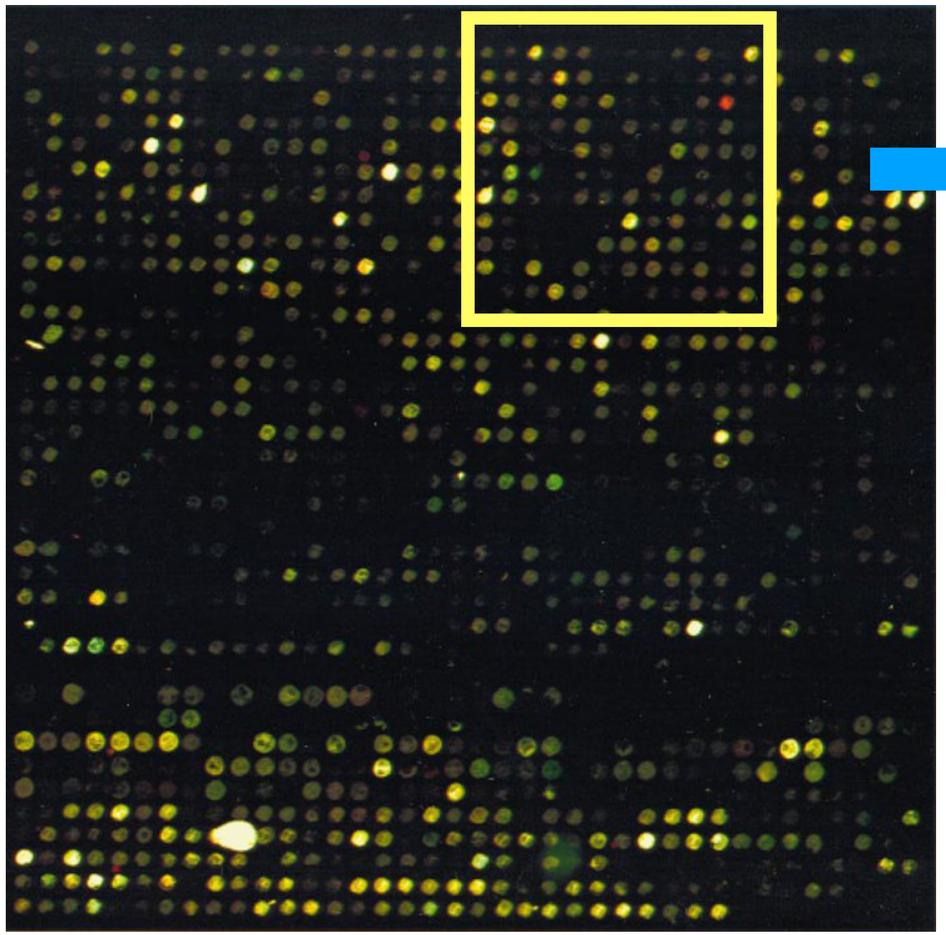


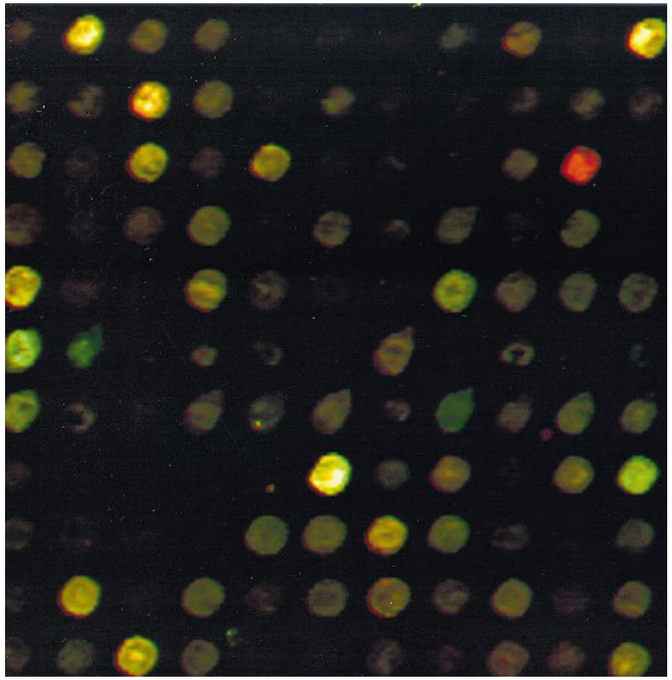
Micro-pipetting to print each gene's sequence in each tiny spot



Cheung *et al*. (1999)

cDNA array for 5k mouse genes





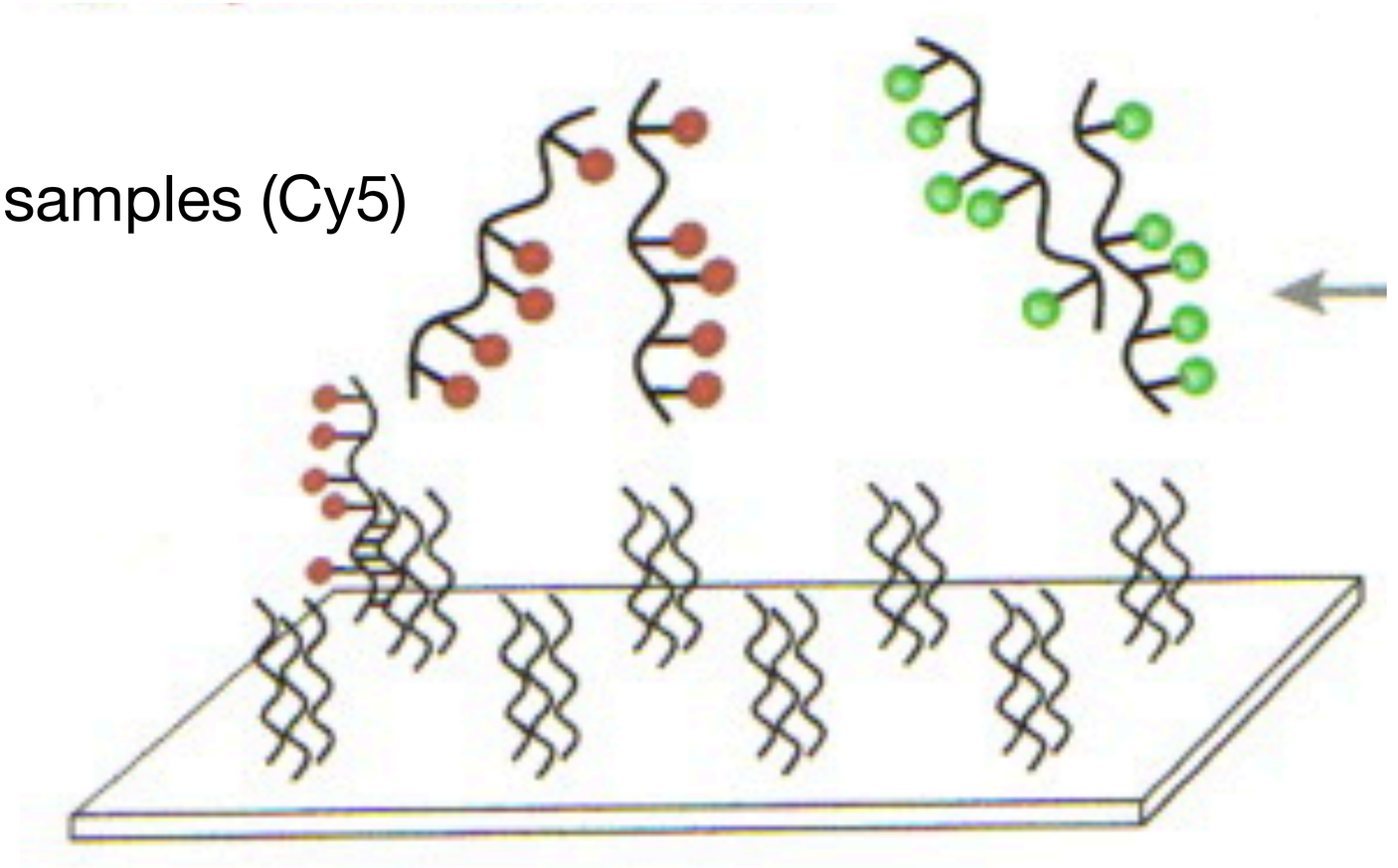
Compare Cyanine5-dUTP (red) vs. Cyanine5 (green)

E.g., wild type vs. mutant

Cheung *et al*. (1999)

Quantifying the amount of transcripts based on hybridization

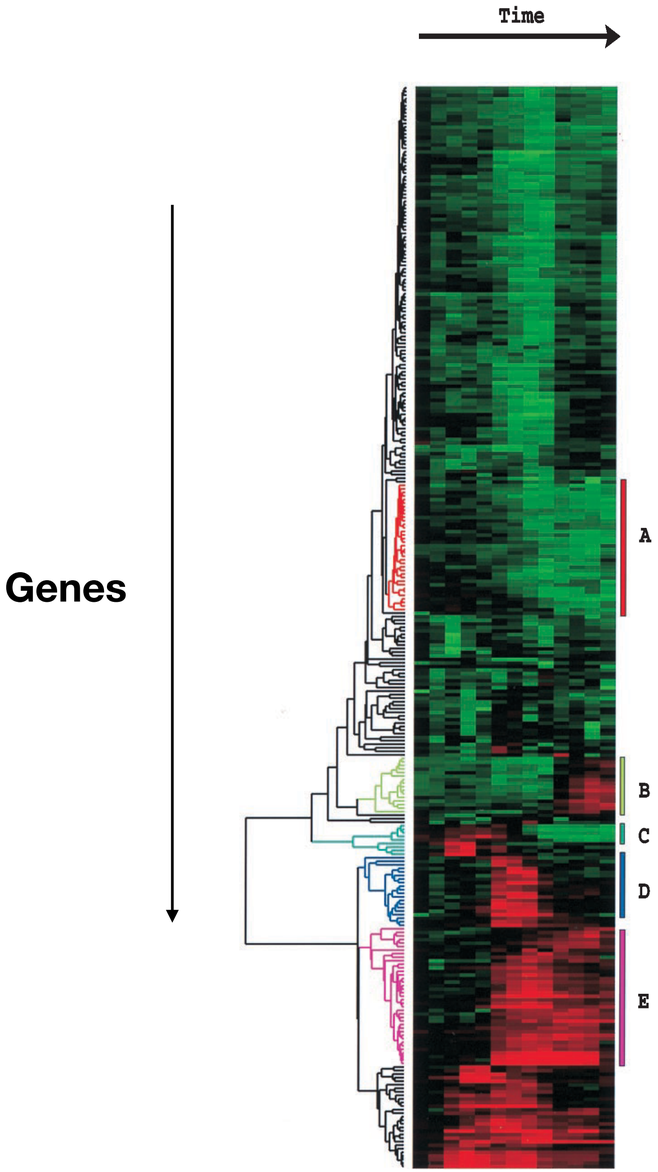
Wild type samples (Cy5)



Mutant samples (Cy3)

Image: Boonham *et al*. (2007)

Gene expression microarrays



Cluster analysis and display of genome-wide expression patterns

MB Eisen, PT Spellman, PO Brown... - Proceedings of the ..., 1998 - National Acad Sciences Paperpile

A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes ...

☆ Save 99 Cite Cited by 20059 Related articles Web of Science: 12261 Import into BibTeX »

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Contributed by David Botstein, October 13, 1998

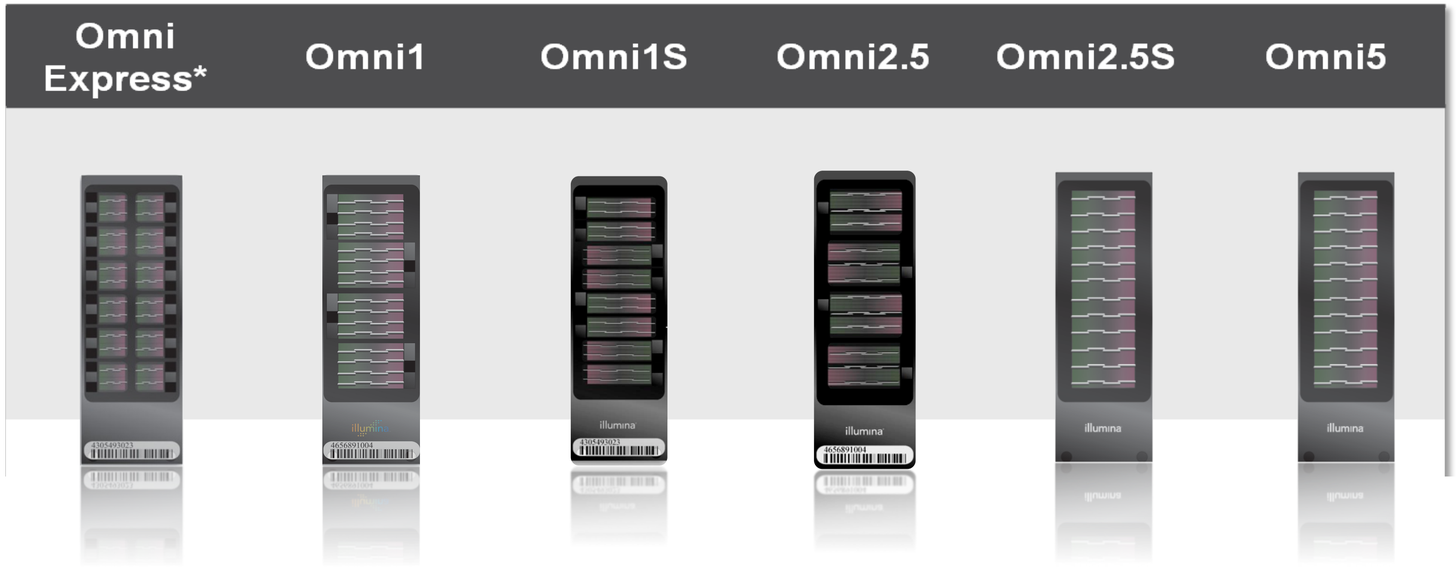
ABSTRACT **A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, conveying the clustering and the underlying expression data simultaneously in a form intuitive for biologists. We have found in the budding yeast *Saccharomyces cerevisiae* that clustering gene expression data groups together efficiently genes of known similar function, and we find a similar tendency in human data. Thus patterns seen in genome-wide expression experiments can be interpreted as indications of the status of cellular processes. Also, coexpression of genes of known function with poorly characterized or novel genes may provide a simple means of gaining leads to the functions of many genes for which information is not available currently.**

be used, such as the Euclidean distance, angle, or dot products of the two n-dimensional vectors representing a series of n measurements. We have found that the standard correlation coefficient (i.e., the dot product of two normalized vectors) conforms well to the intuitive biological notion of what it means for two genes to be "coexpressed;" this may be because this statistic captures similarity in "shape" but places no emphasis on the magnitude of the two series of measurements.

It is not the purpose of this paper to survey the various methods available to cluster genes on the basis of their expression patterns, but rather to illustrate how such methods can be useful to biologists in the analysis of gene expression data. We aim to use these methods to organize, but not to alter, tables containing primary data; we have thus used methods that can be reduced, in the end, to a reordering of lists of genes. Clustering methods can be divided into two general classes, designated supervised and unsupervised clustering (4). In

Eisen, Spellman, Brown, Botstein (1998)

DNA microarrays for genotyping



	Omni Express*	Omni1	Omni1S	Omni2.5	Omni2.5S	Omni5
SNPs:	700,000	1,100,000	1,250,000	2,500,000	2,500,000	5,000,000
CNV:	No	Yes	Yes	Yes	Yes	Yes
ETA:	Now	Now	Now	Now	Early 2011	Early 2011

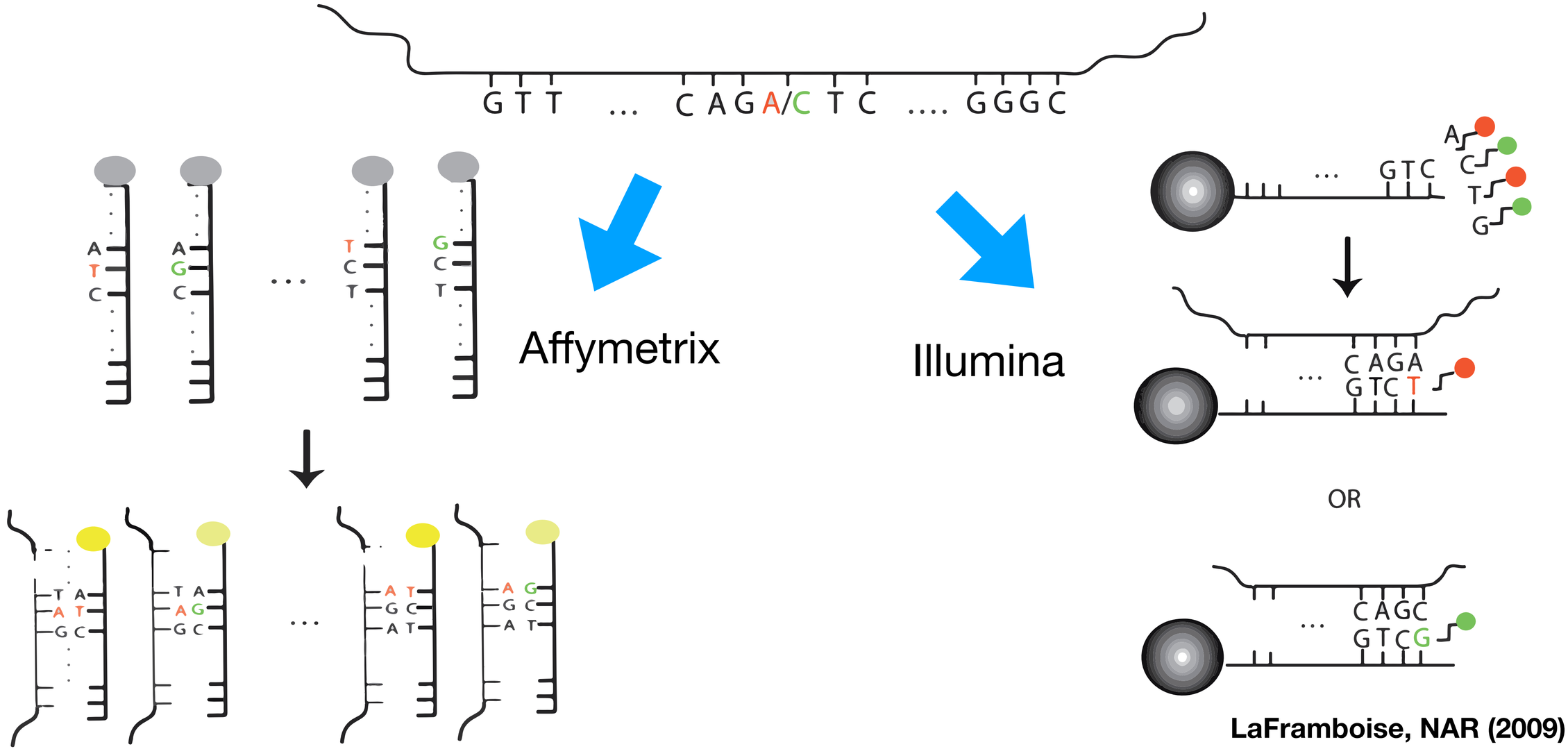
MAF > 5%

MAF > 2.5%

MAF > 1%

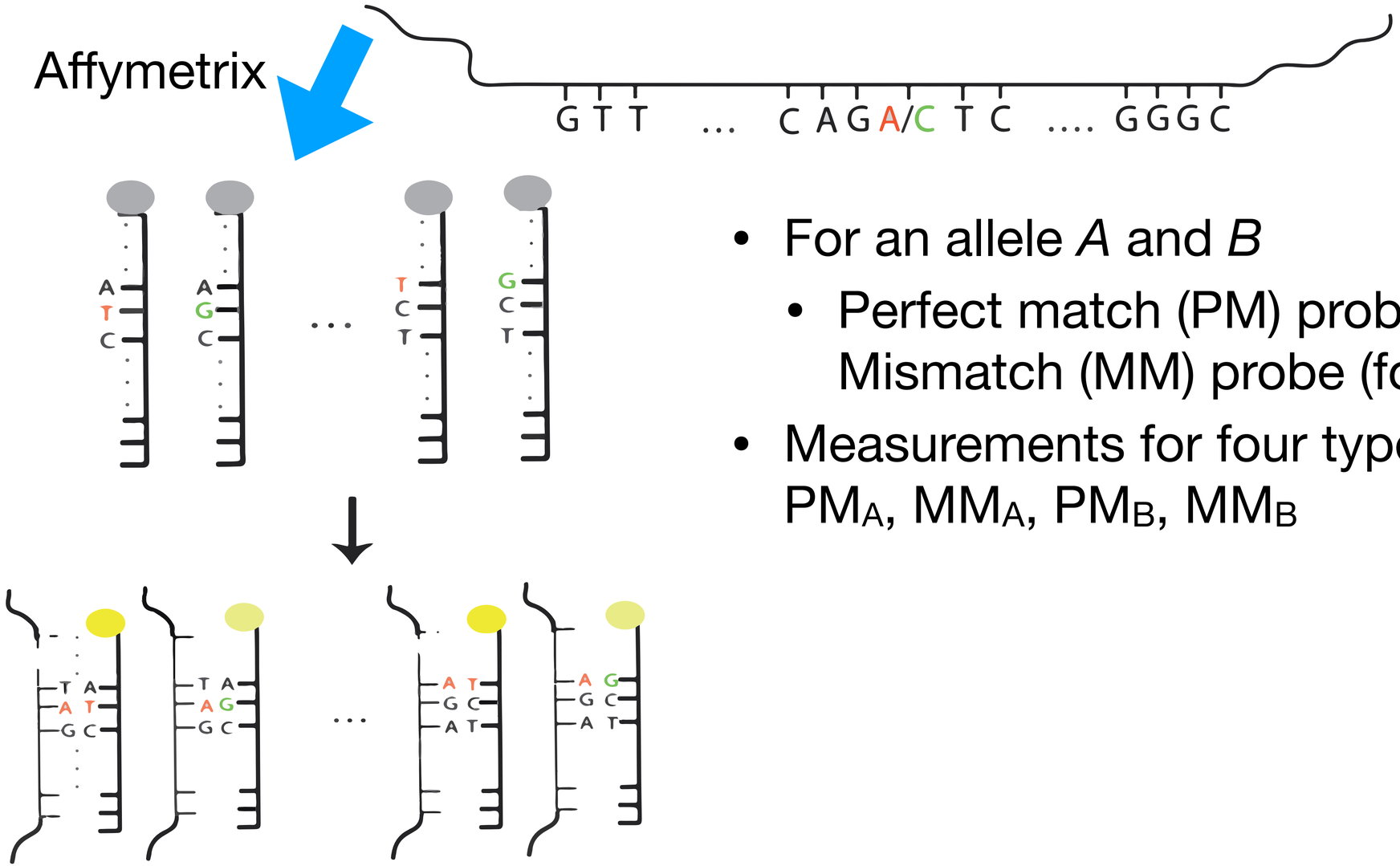
Broad Institute, Winckler, MPG (2011)

Genotype calling by arrays



LaFramboise, NAR (2009)

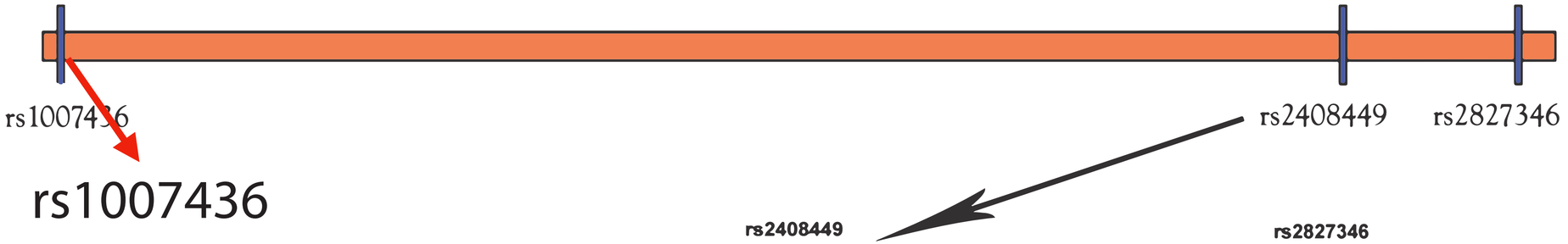
Genotype calling by arrays

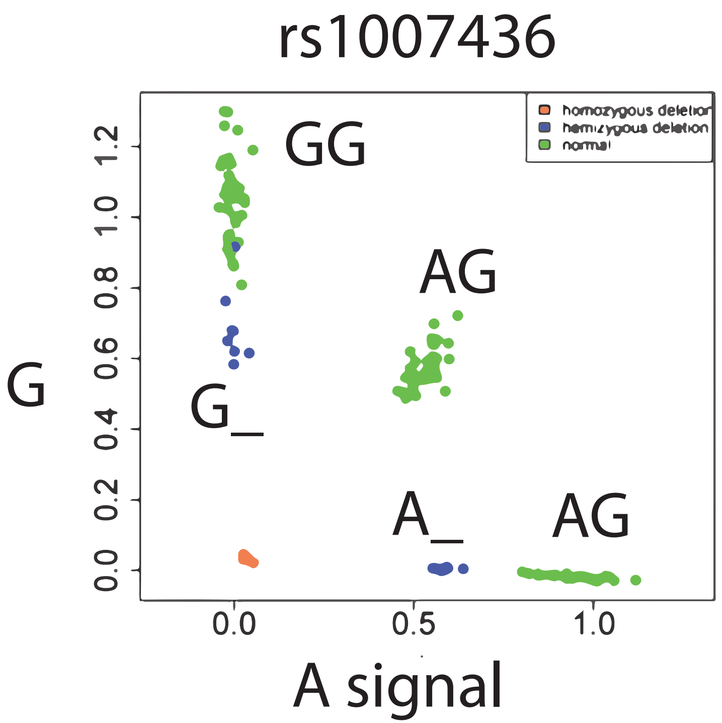


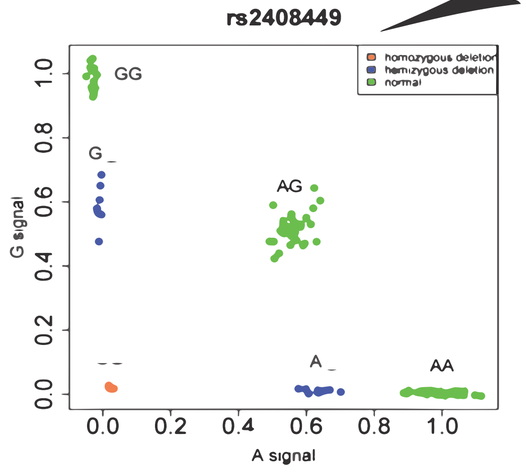
- For an allele *A* and *B*
  - Perfect match (PM) probe
    Mismatch (MM) probe (for background)
- Measurements for four types of probes:
  PM_A, MM_A, PM_B, MM_B

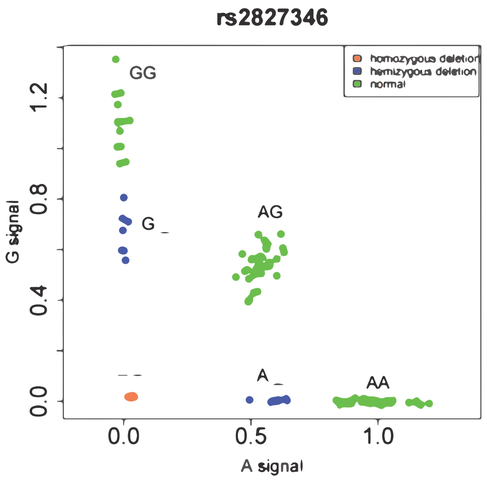
LaFramboise, NAR (2009)

Genotype-calling ≈ a mixture of Gaussians







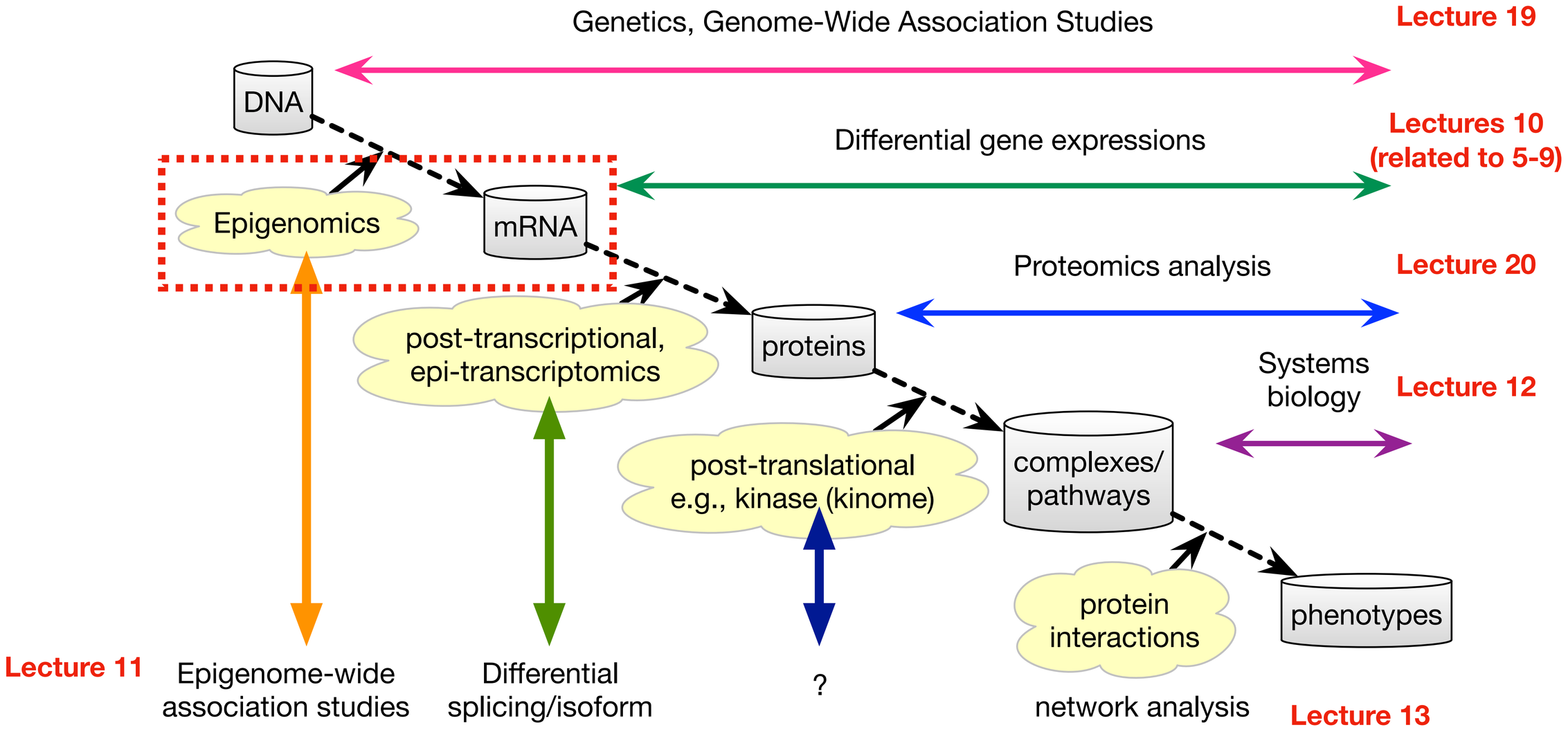


LaFramboise, NAR (2009)

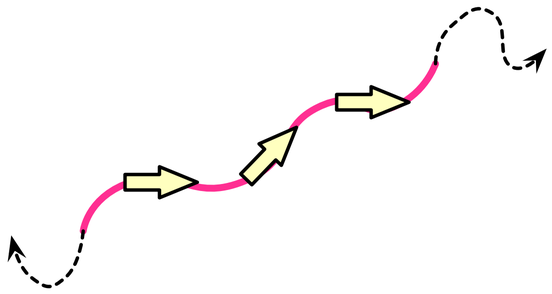
Today’s lecture: Genomics technology

- **What is genomics technology?**
  - Measuring every step of the information flow: DNA, mRNA, splicing, protein, etc.
  - Why do we (statisticians) need to be aware of it?
- **Obtaining the book of life: a rough history of genomics methods**
  - Sequencing-based methods
  - Array-based methods
- **Understanding the book of life: omics technology**
  - Focusing on variations: mutations and expressions
  - Efforts to build epigenetic, transcriptomic, cell type references

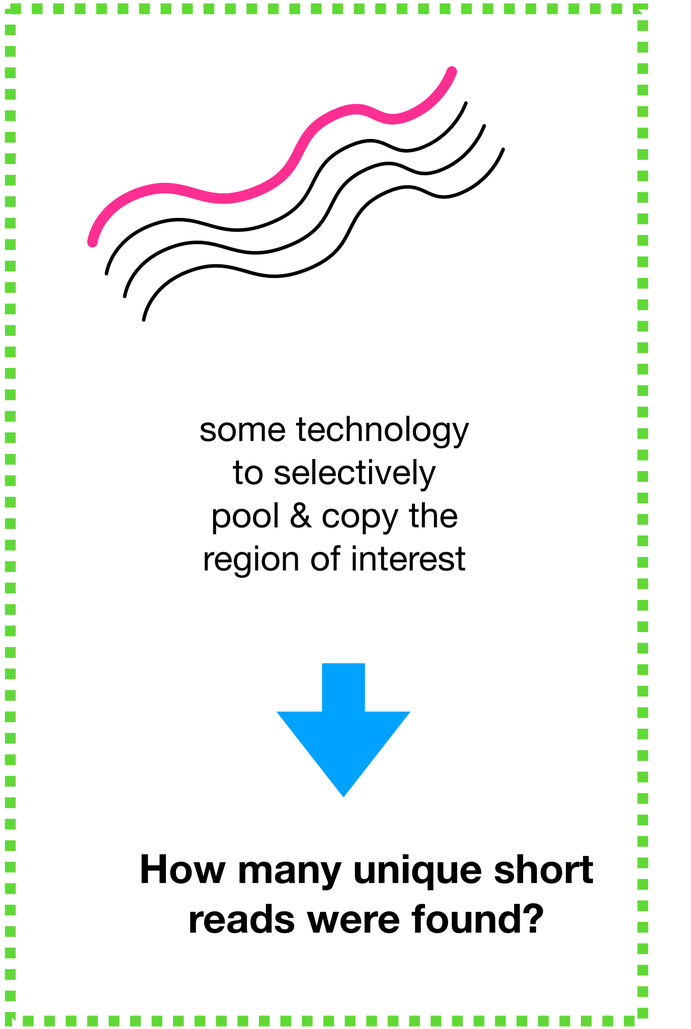
Biology: Measuring intermediate layers by multiple types of high-throughput assays



The basic idea of High-throughput methods: selection followed by sequencing

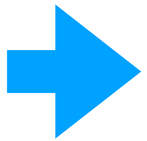


some region
of interest
(gene, regulatory
elements, etc.)

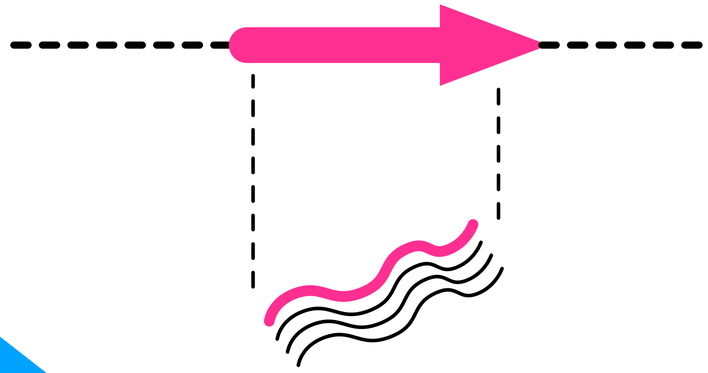


some technology
to selectively
pool & copy the
region of interest

How many unique short reads were found?



Which regions and how do we select the region?

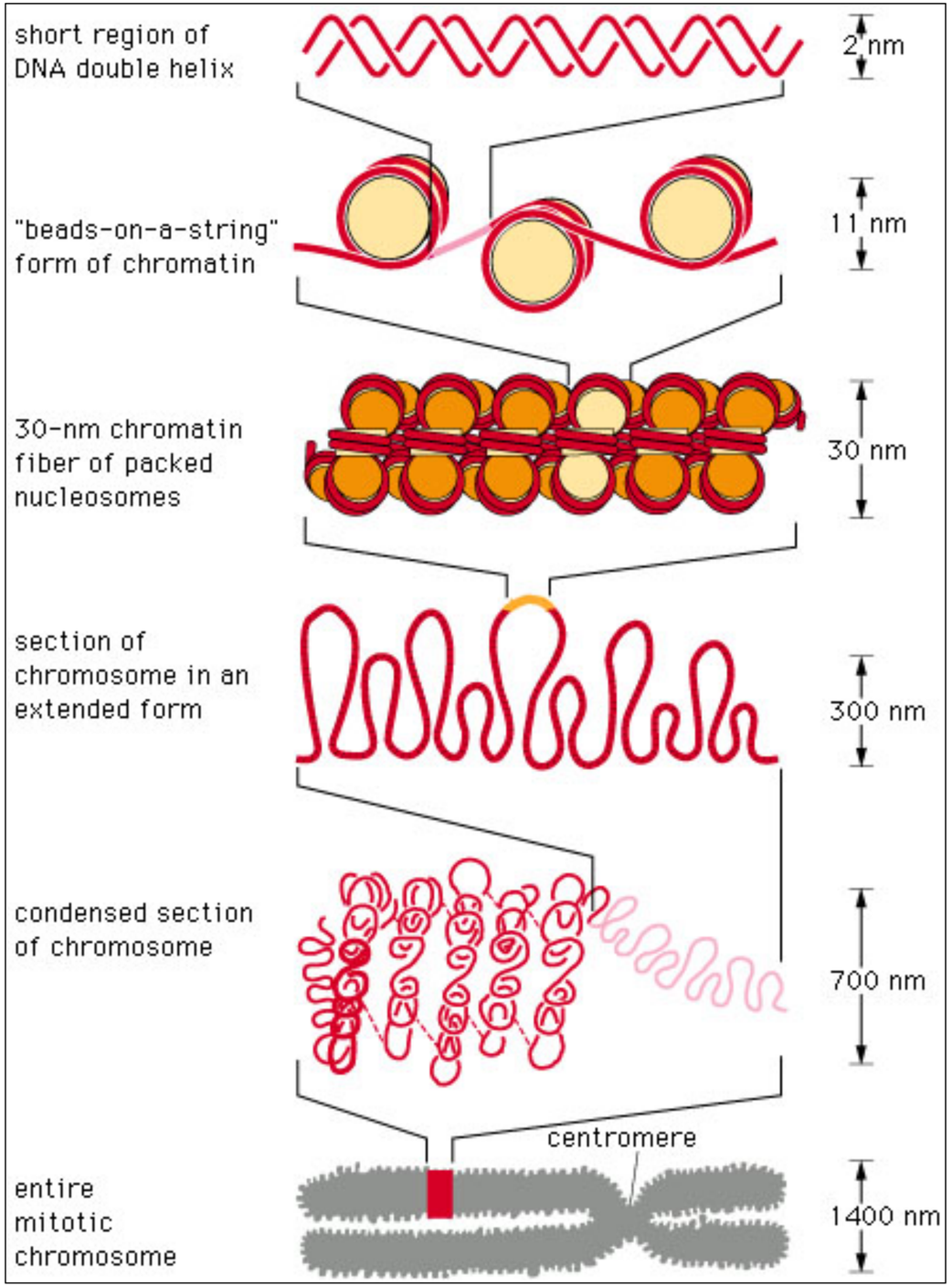


Align/match
to locate this region
in the genome

Genome ≈ an anchor point

DNA packaging

- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally
- Now emerging:
  - Role of accessibility
  - State in chromatin itself
  - Role of 3D interactions



Slide: Manolis Kellis

Epi-genetic modifications

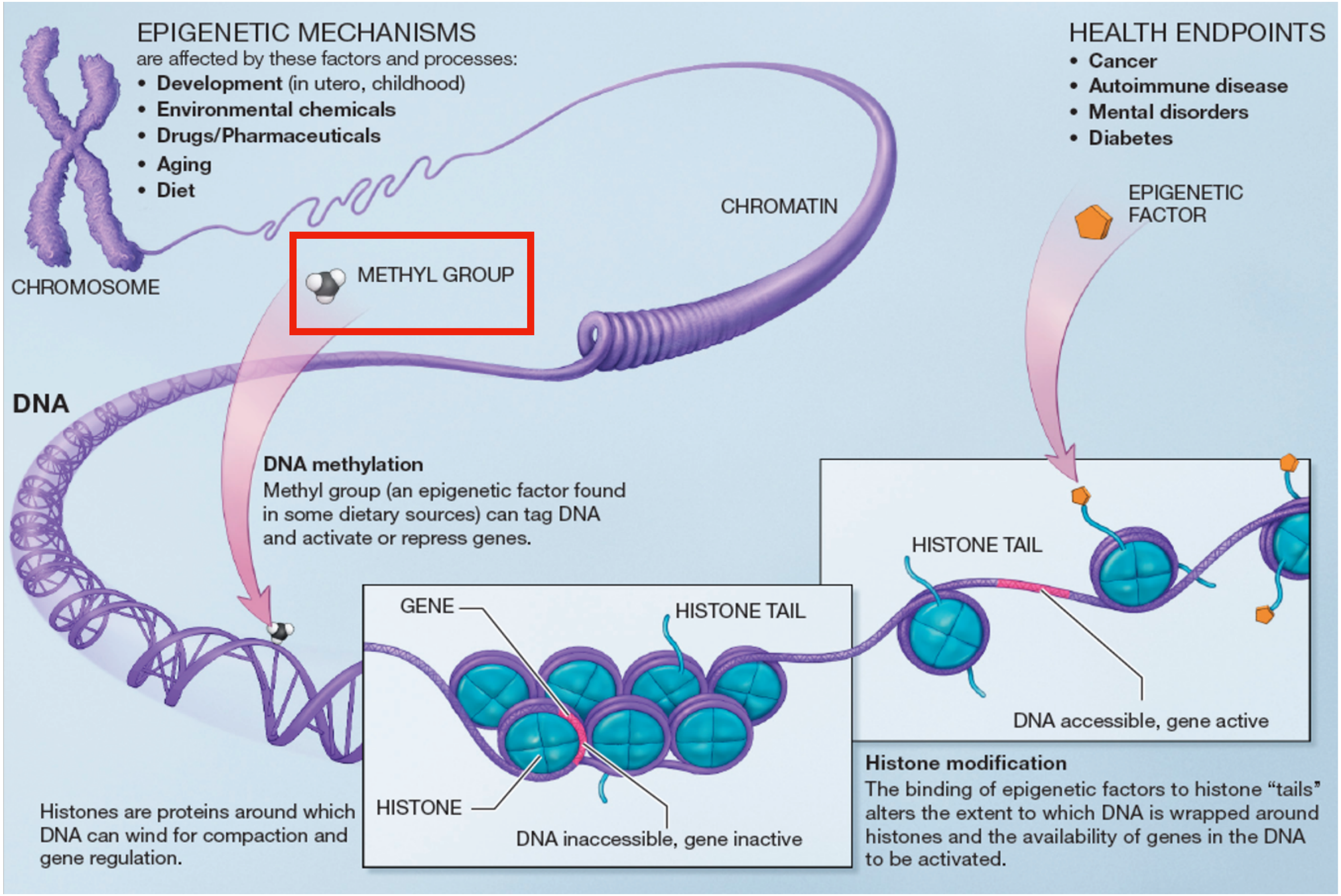
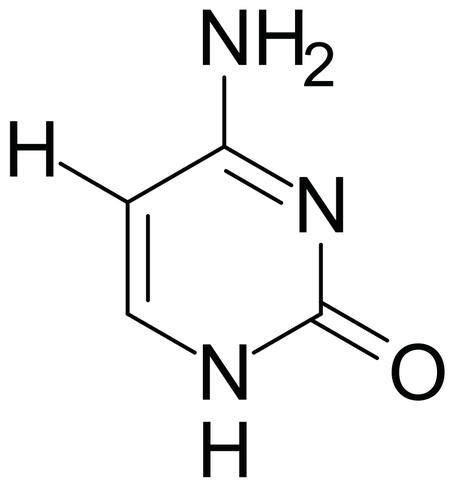
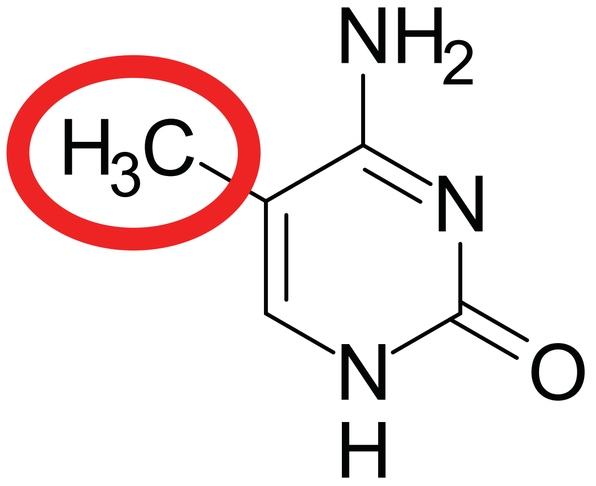


Image source: http://nihroadmap.nih.gov/epigenomics/

DNA methylation modifies the biochemical properties of a particular base



cytosine

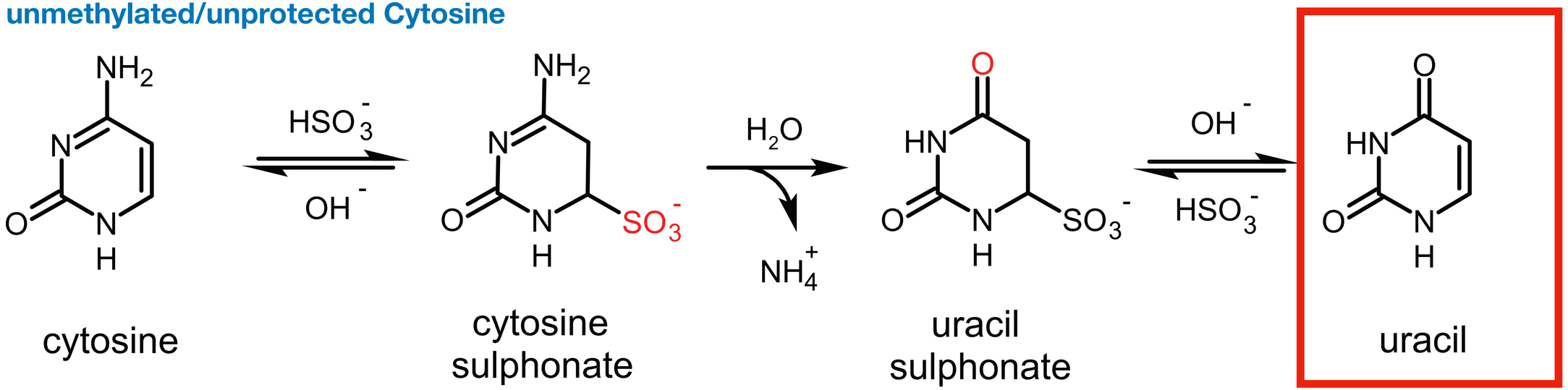


methylated cytosine

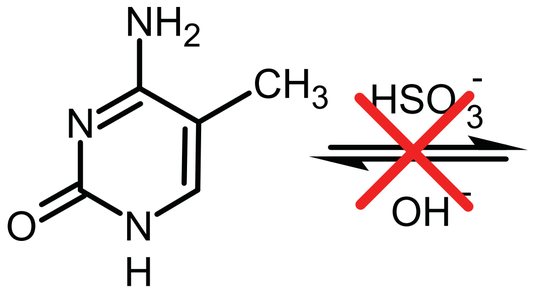
https://en.wikipedia.org/wiki/DNA_methylation

Bisulfite conversion: "unmethylated" C ☞ U

For unmethylated/unprotected Cytosine

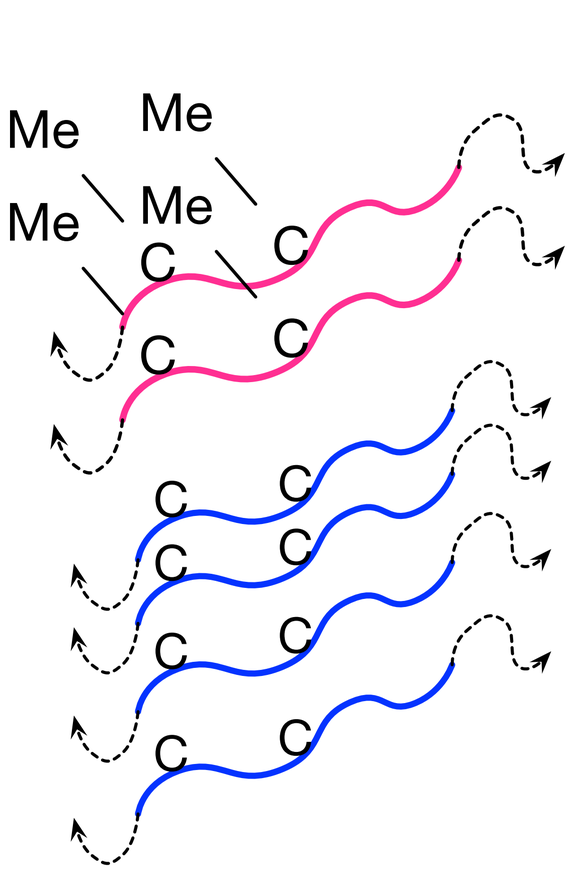


For methylated Cytosine

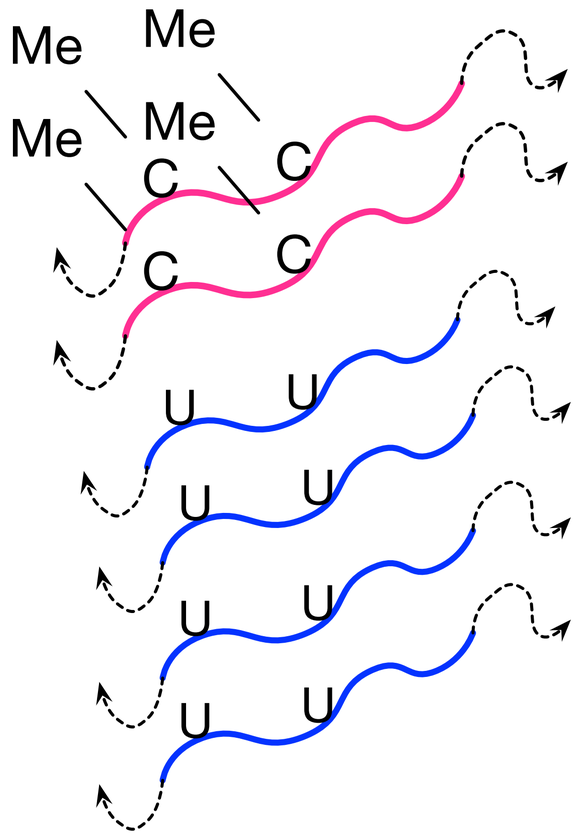


https://en.wikipedia.org/wiki/Bisulfite_sequencing

Bisulfite sequencing/array: Distinguish DNAs methylated vs. un-methylated



bisulfite treatment



sequencing & alignment

Where in the genome?

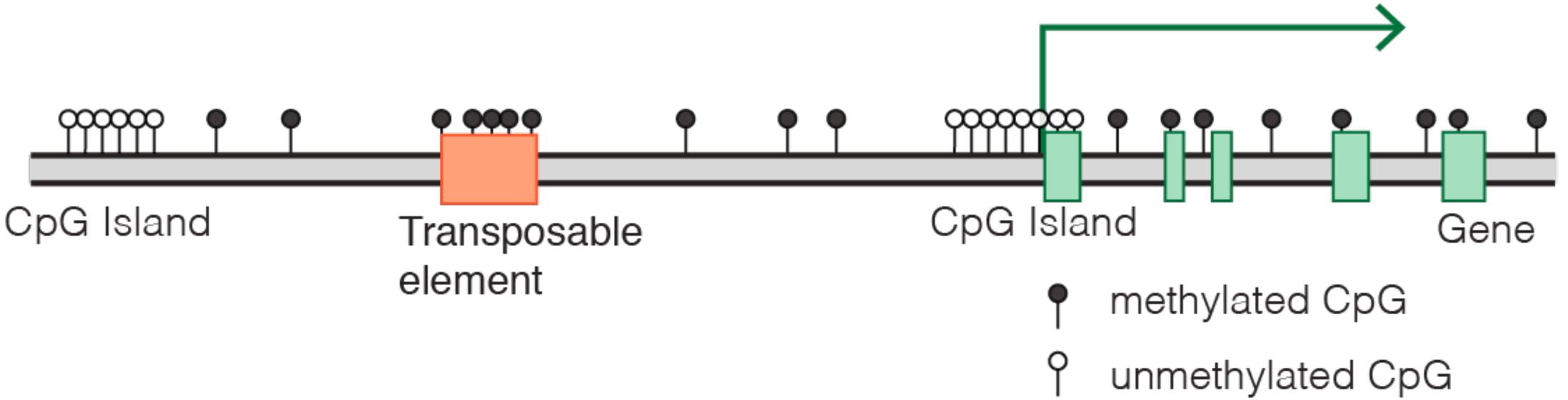
How many with C?

How many with T/U?

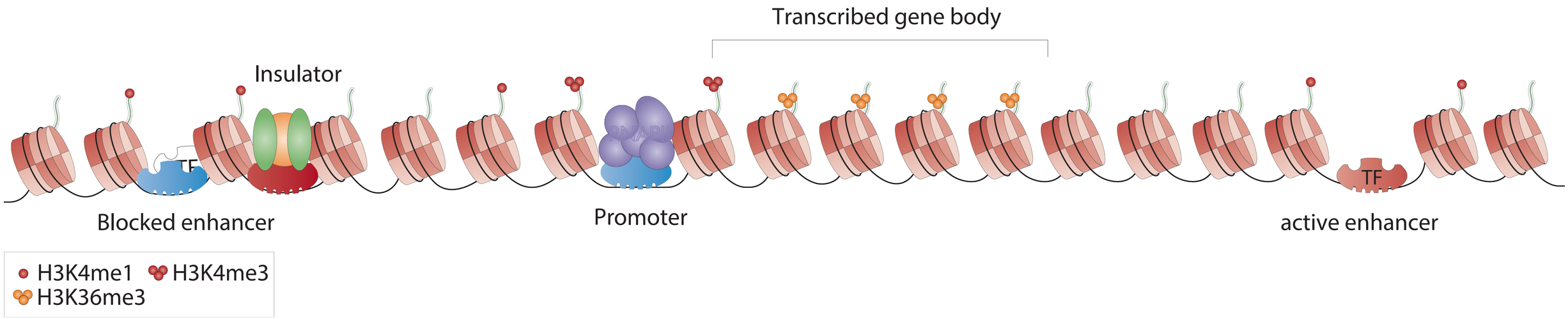
Quantify β-value:

$$\beta = \frac{\#\mathbf{meth}}{\#\mathbf{meth} + \#\mathbf{unmeth}}$$

Typical mammalian DNA methylation landscape

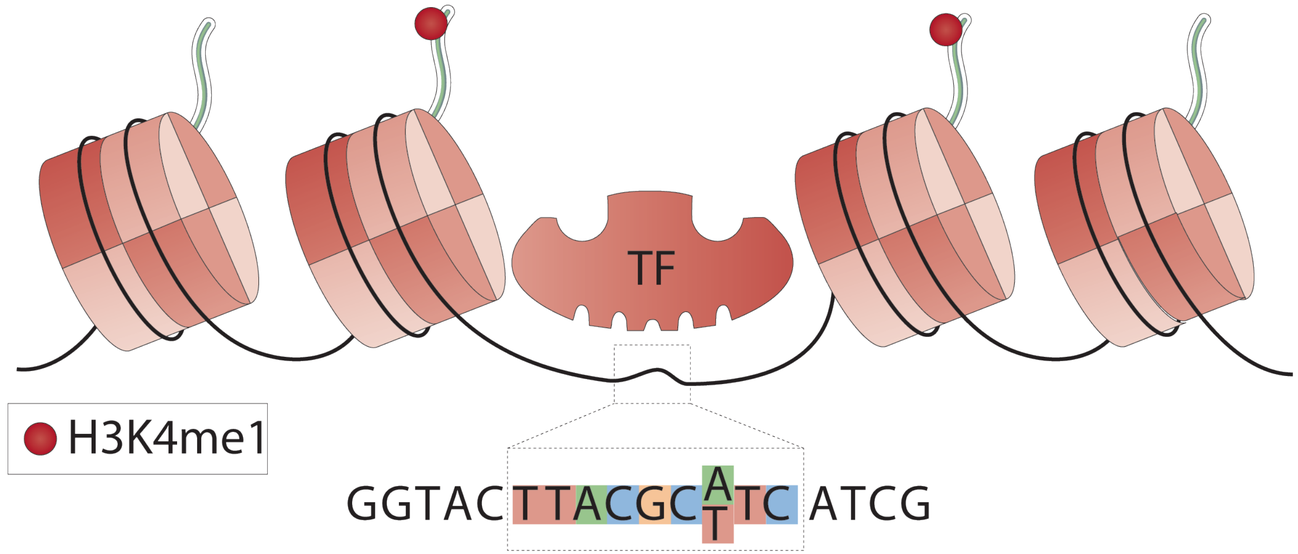


Different histone code marks different types of regulatory elements



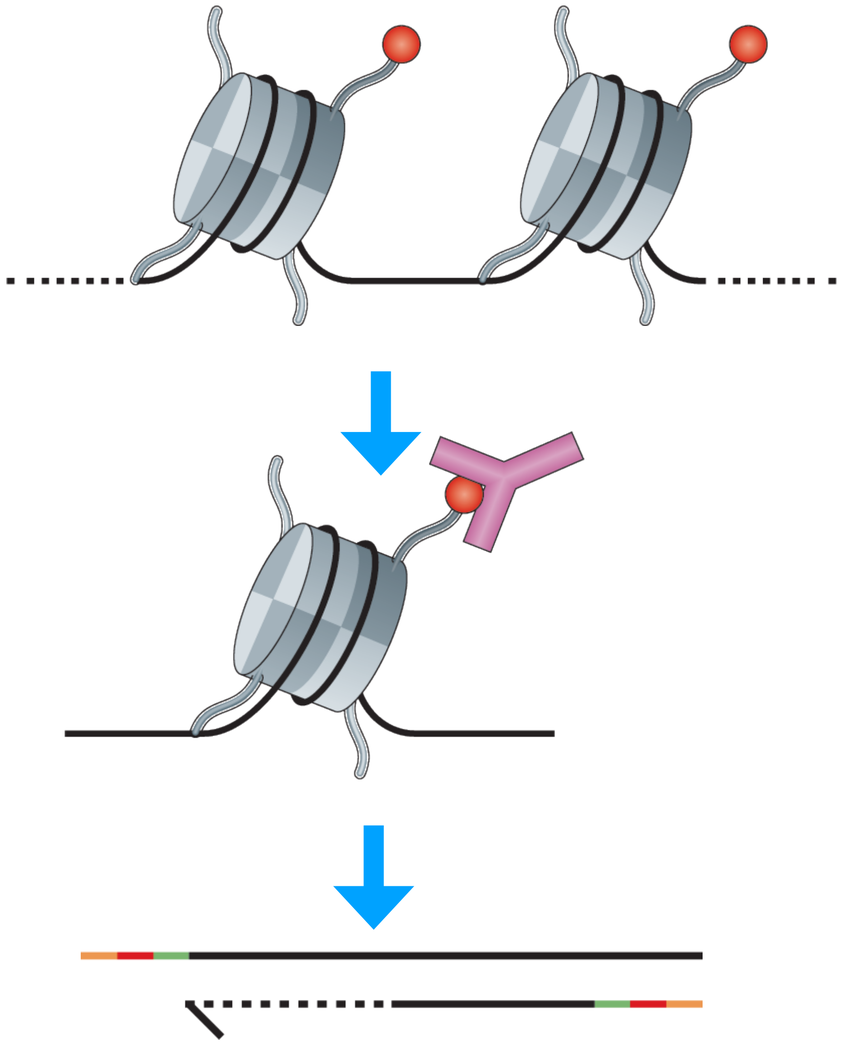
Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

One of the goals is to understand the logic of non-coding DNA sequences



Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

ChIP-seq: Quantifying histone modifications



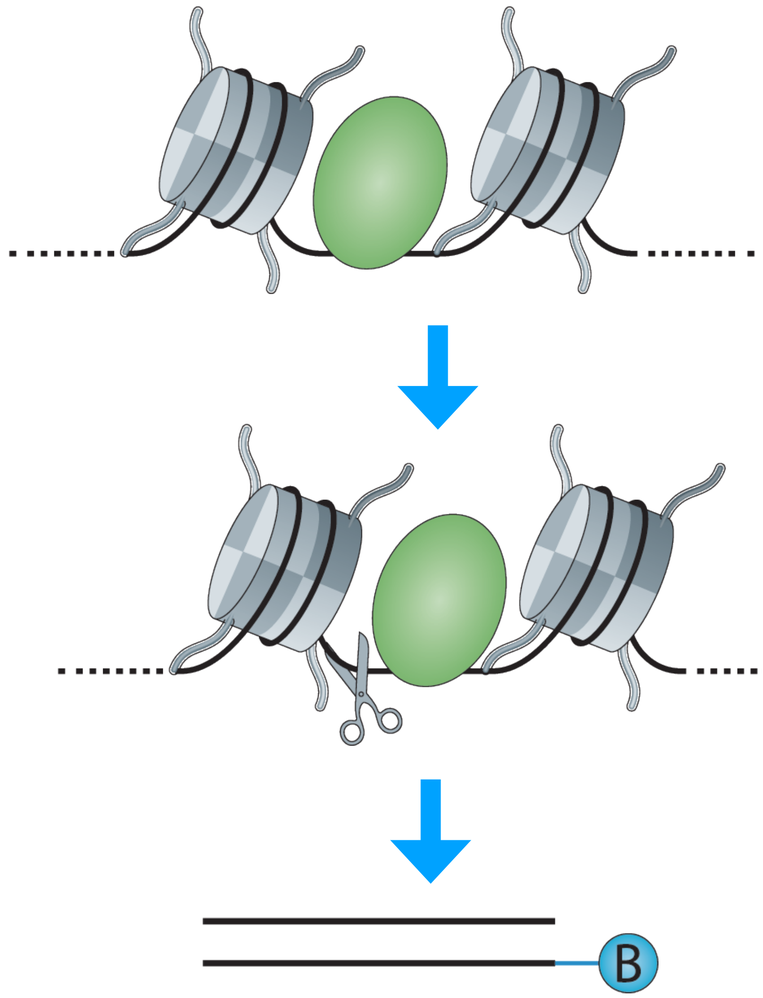
Crosslink proteins & DNA

Immunoprecipitate &
purify DNA bound

Amplify/duplicate
the extracted
DNA segments

Furey, *Nature Reviews Genetics* (2012)

DNase-seq: Quantifying DNA accessibility



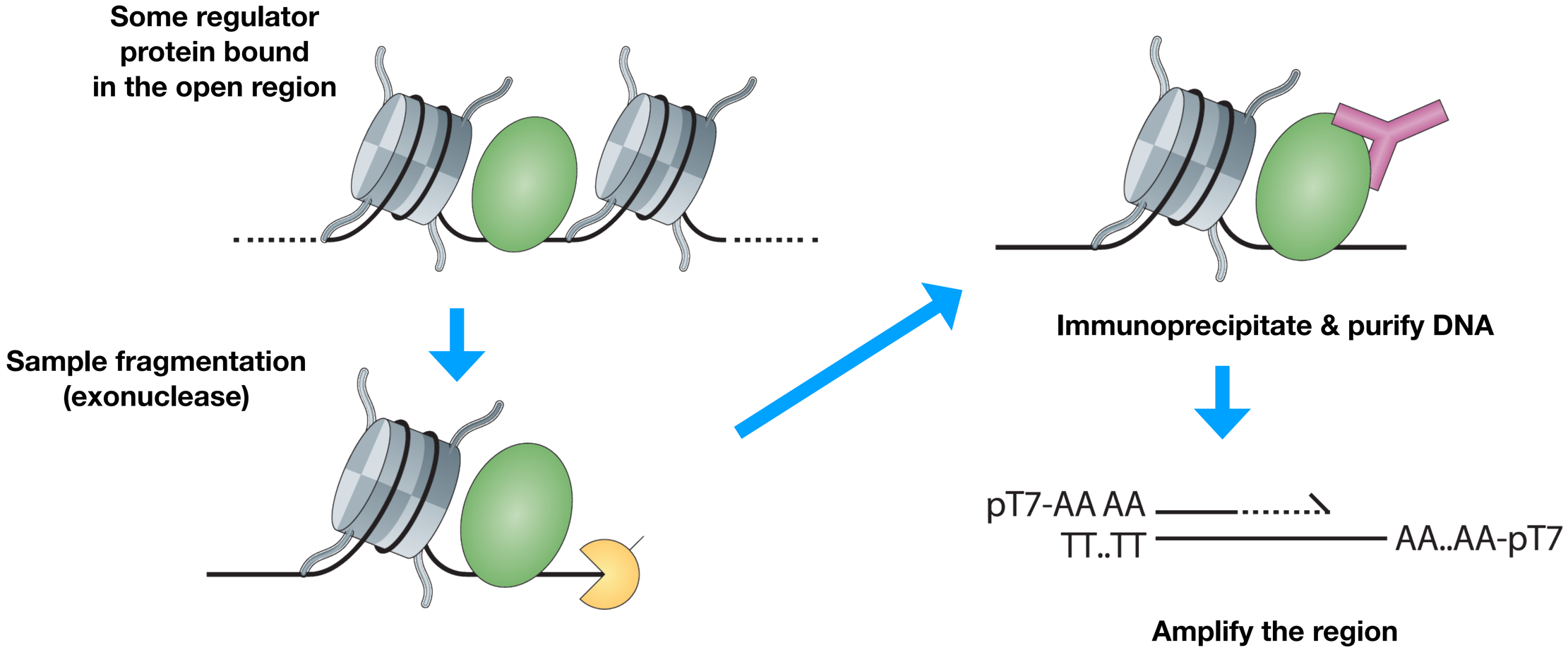
Some regulator protein bound
in the open region

Sample fragmentation
(DNaseI digestion)

Add biotynylated linkers
and extract

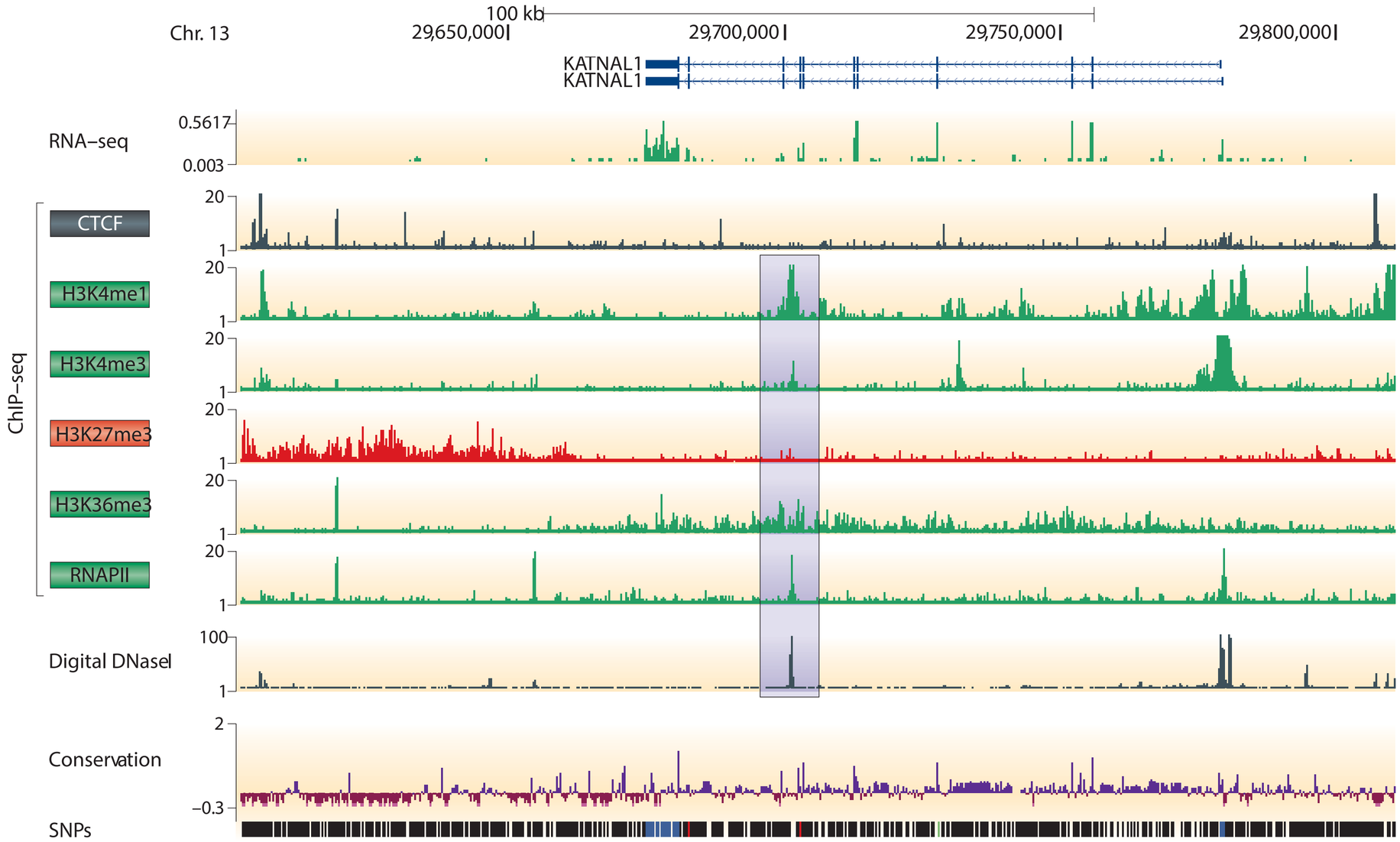
Furey, *Nature Reviews Genetics* (2012)

ChIP-seq: Quantifying regions bound by a particular protein



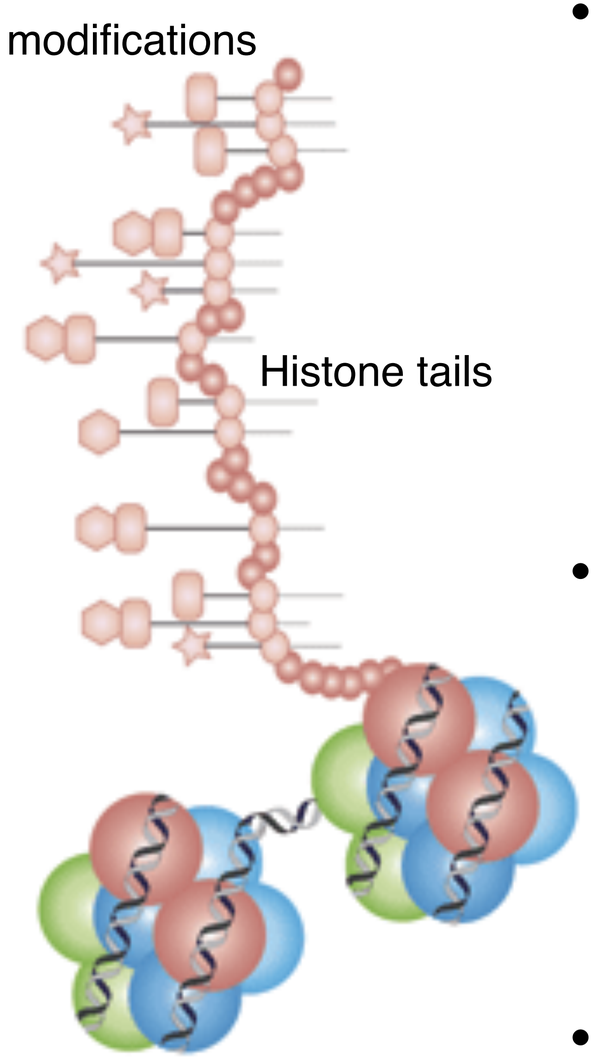
Furey, *Nature Reviews Genetics* (2012)

An example of multiple ChIP-seq tracks



Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

Diversity of epigenetic modifications

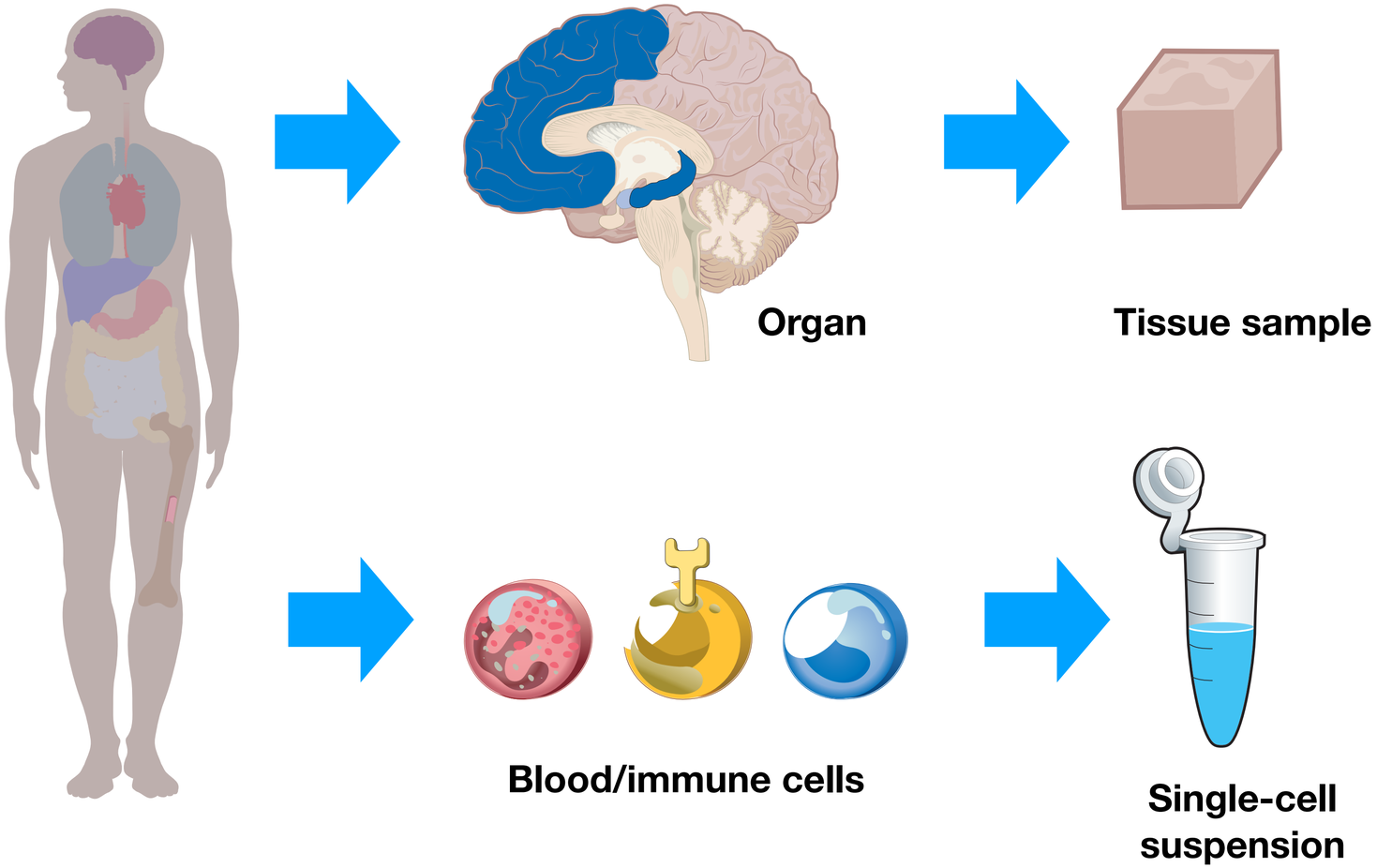


DNA wrapped around
histone proteins

- 100+ different histone modifications
  - Histone protein → H3/H4/H2A/H2B
  - AA residue → Lysine4(K4)/K36…
  - Chemical modification → Met/Pho/Ubi
  - Number → Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

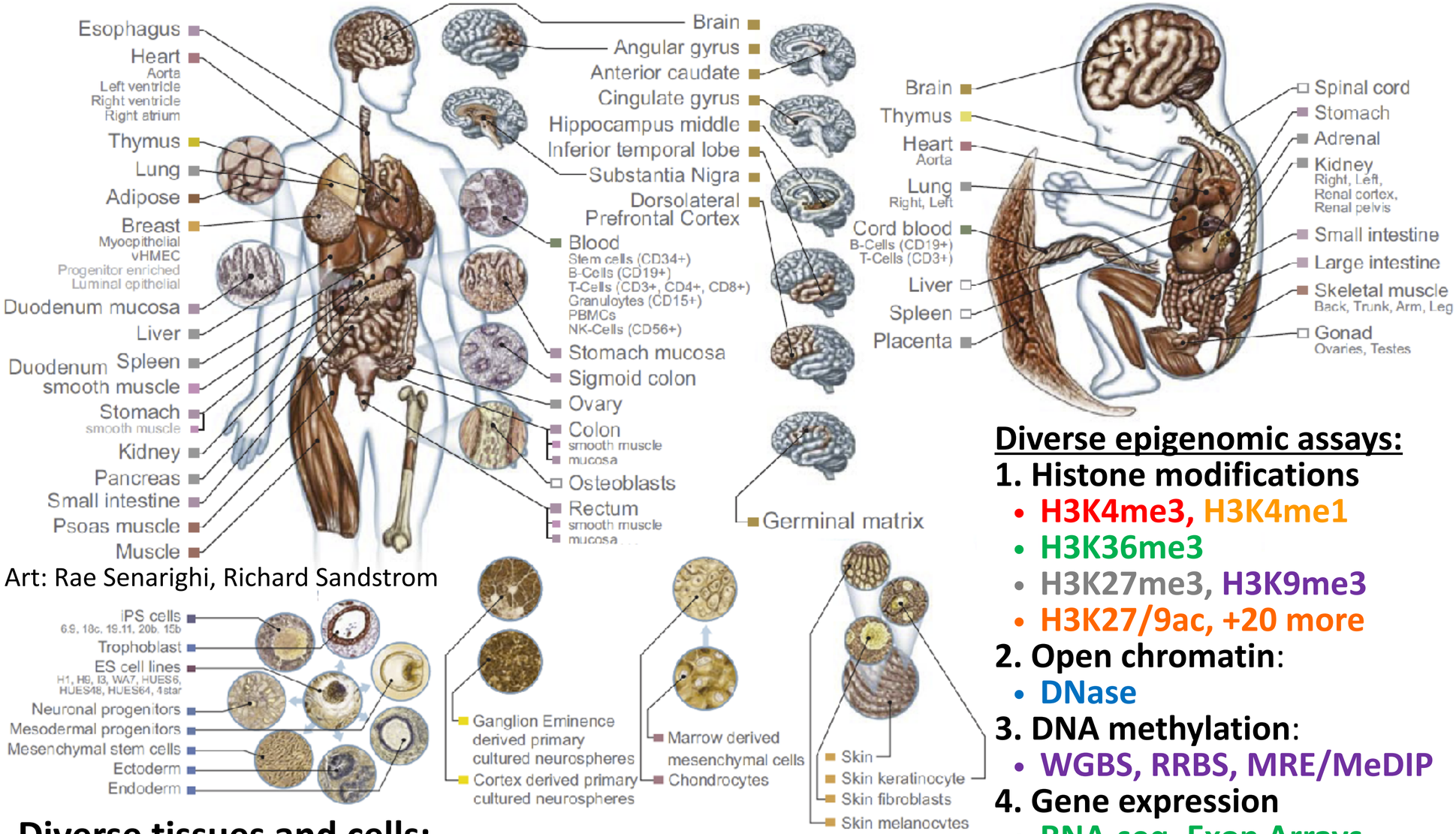
Slide: Manolis Kellis

Full-scale investigation: Mapping epigenomics/ transcriptomics across human body



Romanoski& Glass, Nature (2015)

Epigenomics Roadmap across 100+ tissues/cell types



Art: Rae Senarighi, Richard Sandstrom

Diverse epigenomic assays:

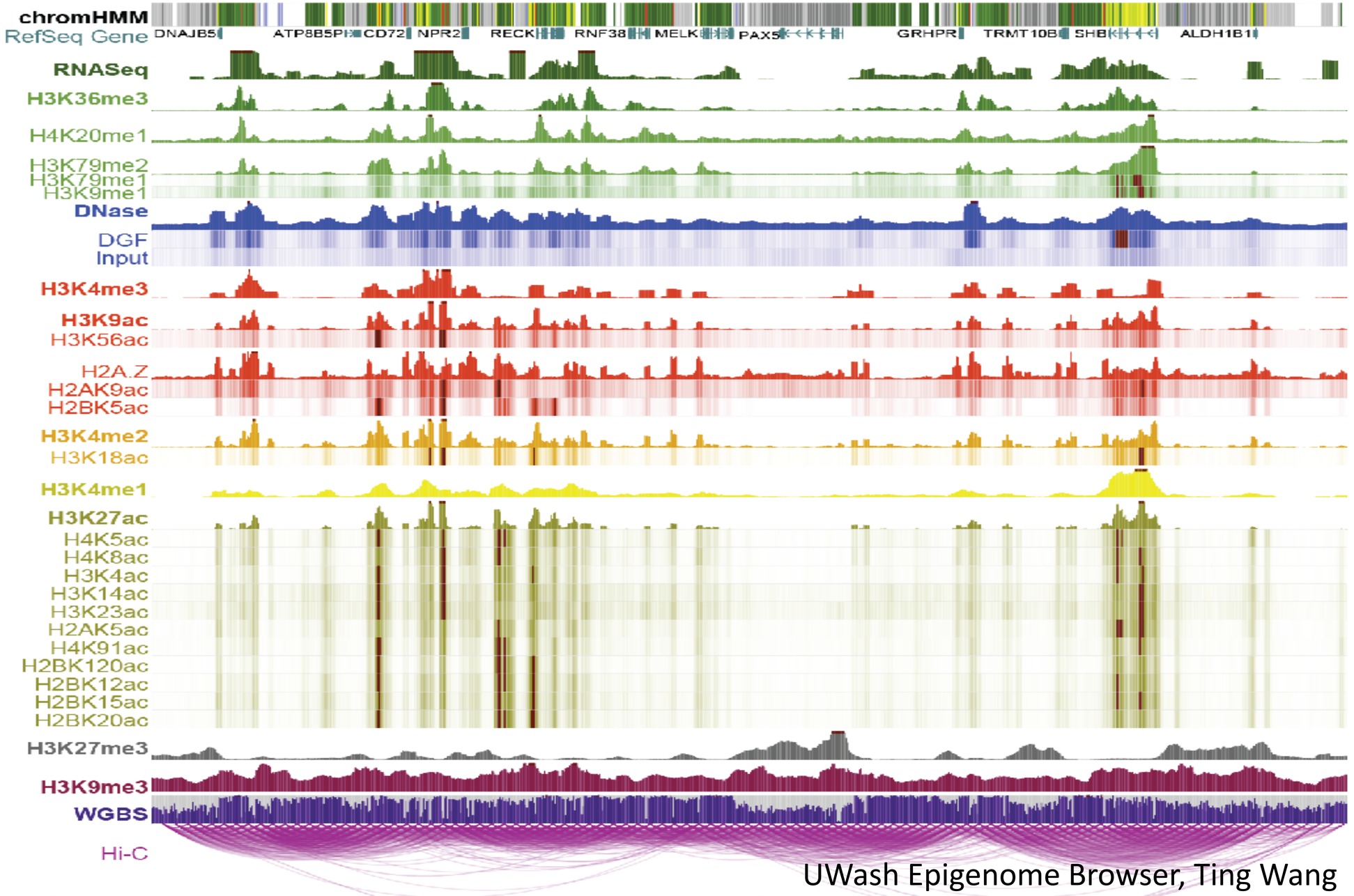
1. **Histone modifications**
   - **H3K4me3, H3K4me1**
   - **H3K36me3**
   - **H3K27me3, H3K9me3**
   - **H3K27/9ac, +20 more**
2. **Open chromatin**:
   - **DNase**
3. **DNA methylation**:
   - **WGBS, RRBS, MRE/MeDIP**
4. **Gene expression**
   - **RNA-seq, Exon Arrays**

Slide: Manolis Kellis

Diverse tissues and cells:

1. **Adult tissues and cells** (brain, muscle, heart, digestive, skin, adipose, lung, blood…)
2. **Fetal tissues** (brain, skeletal muscle, heart, digestive, lung, cord blood…)
3. **ES cells, iPS, differentiated cells** (meso/endo/ectoderm, neural, mesench, trophobl)

Deep sampling of 9 reference epigenomes (e.g. IMR90)

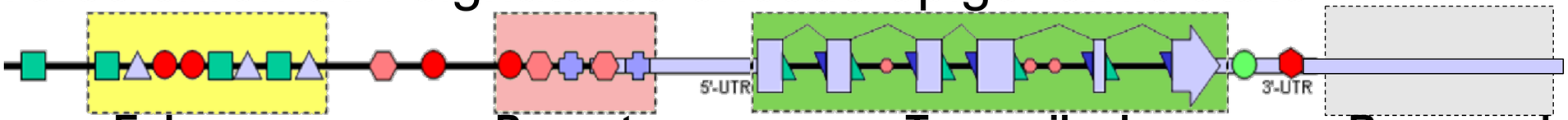


UWash Epigenome Browser, Ting Wang

Slide: Manolis Kellis

Chromatin state+RNA+DNAse+28 histone marks+WGBS+Hi-C

Diverse chromatin signatures encode epigenomic state

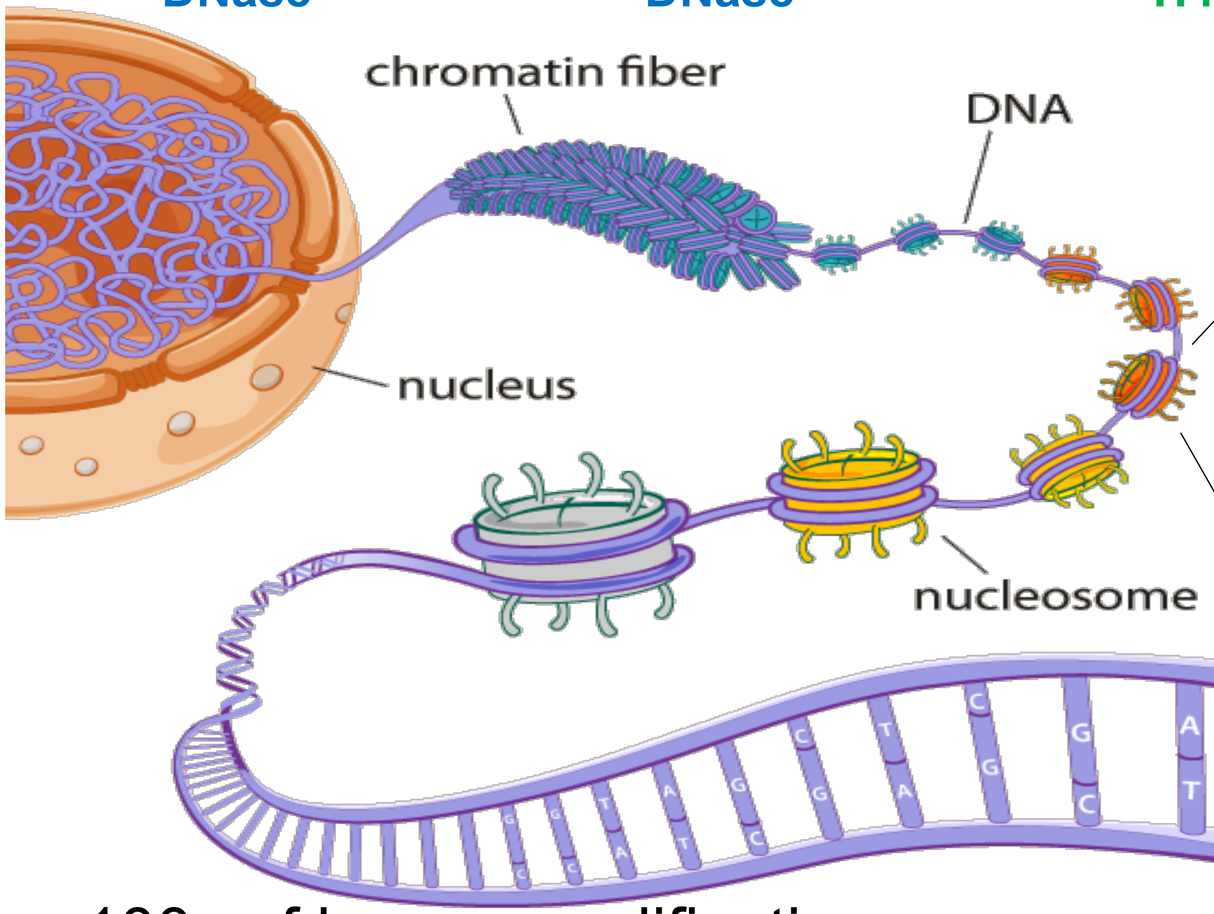


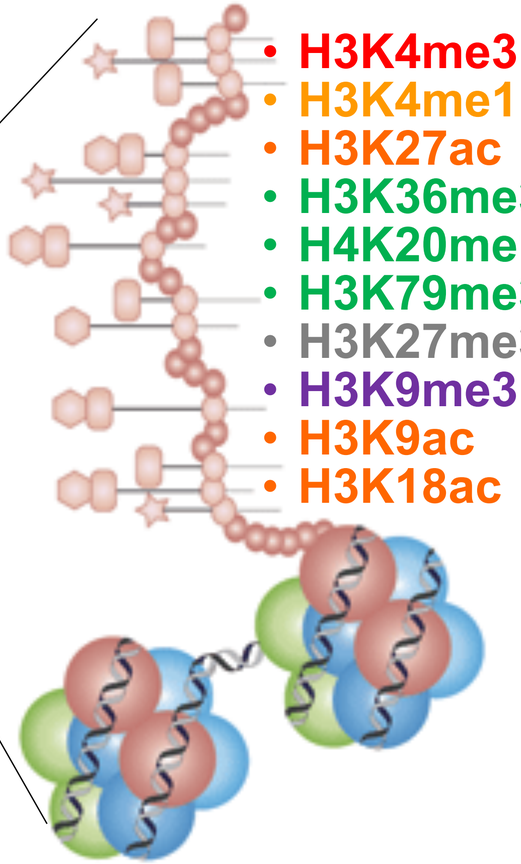
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl



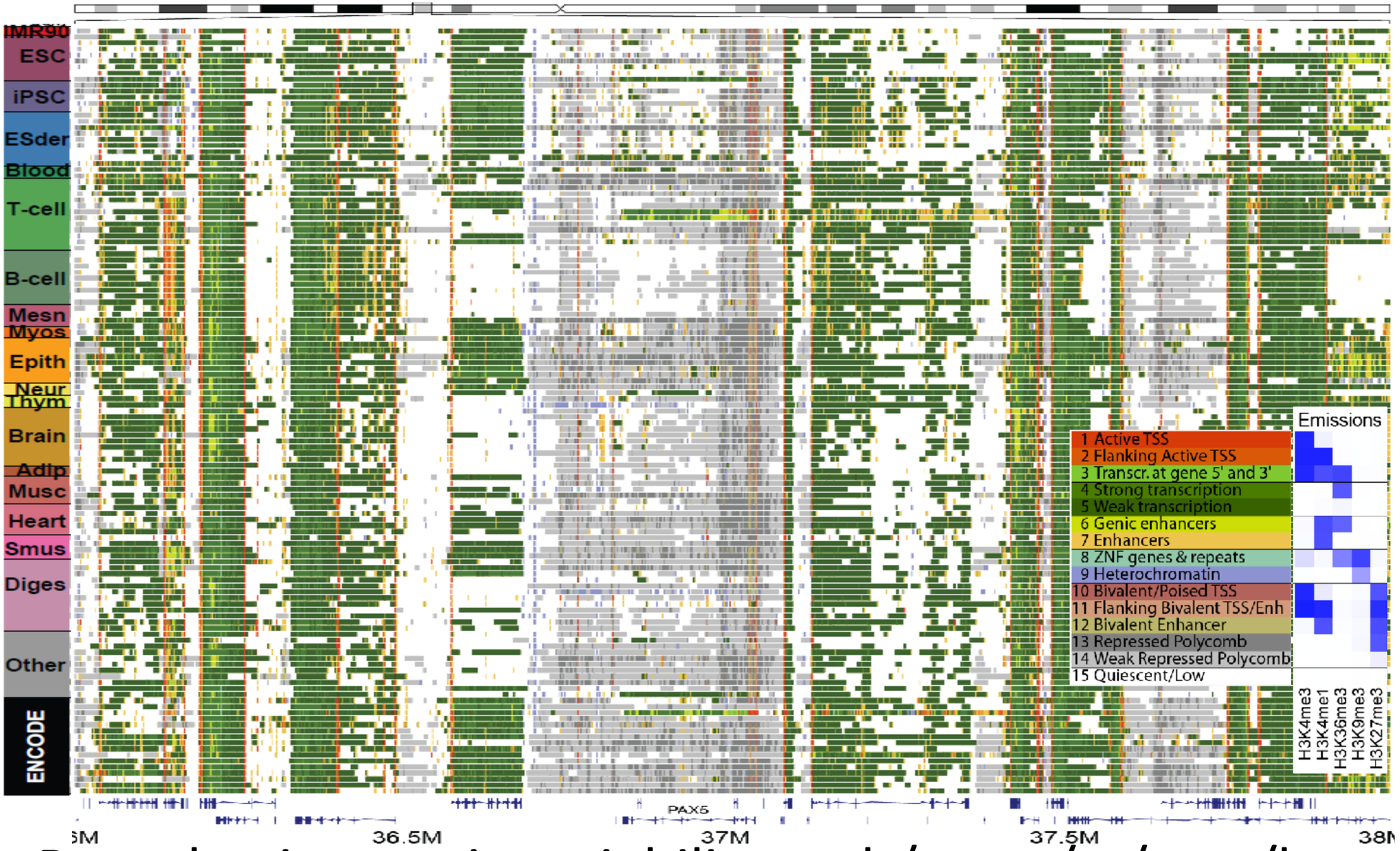


- H3K4me3
- H3K4me1
- H3K27ac
- H3K36me3
- H4K20me1
- H3K79me3
- H3K27me3
- H3K9me3
- H3K9ac
- H3K18ac

Slide: Manolis Kellis

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Chromatin state annotations across 127 epigenomes

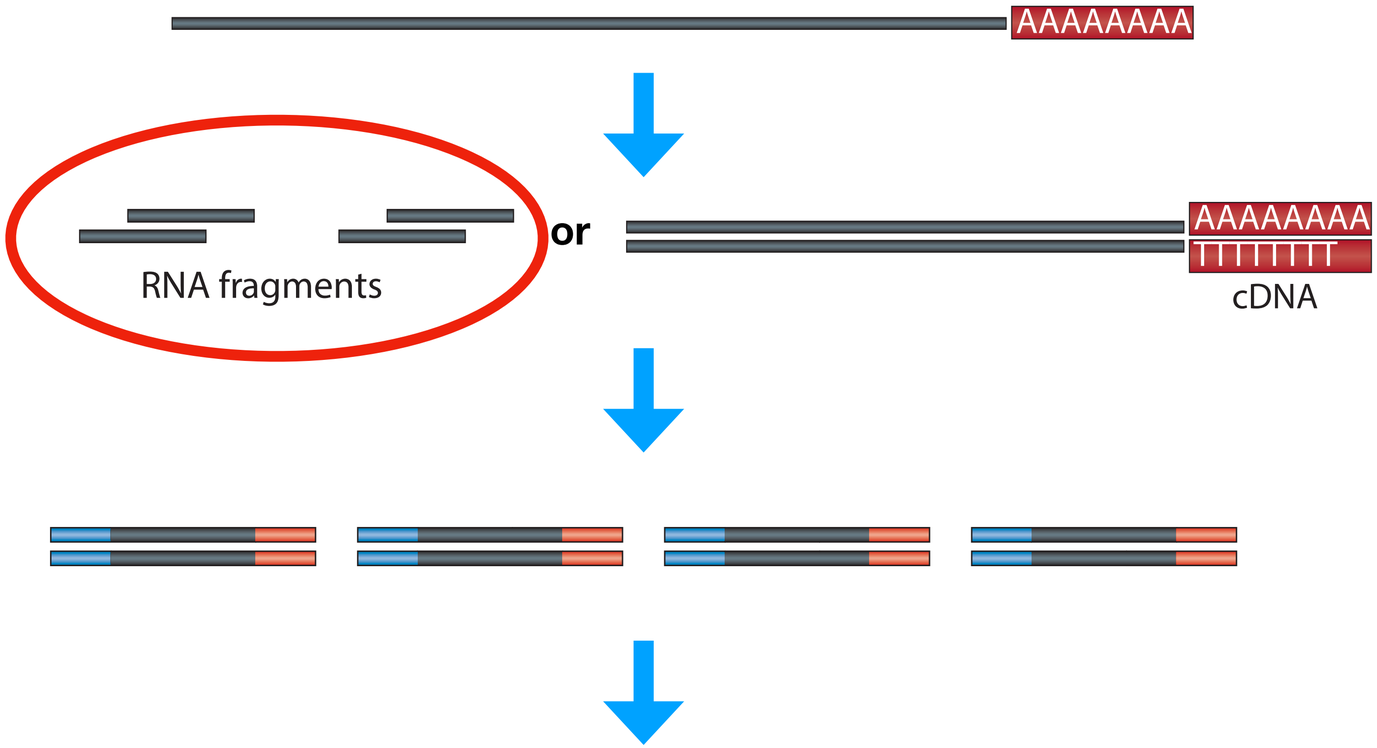


Slide: Manolis Kellis

Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje

Quantifying gene expressions



Pool poly-A tail-containing mRNA sequences

Construct cDNA or fragmentation

Construct library with 5', 3' adaptors

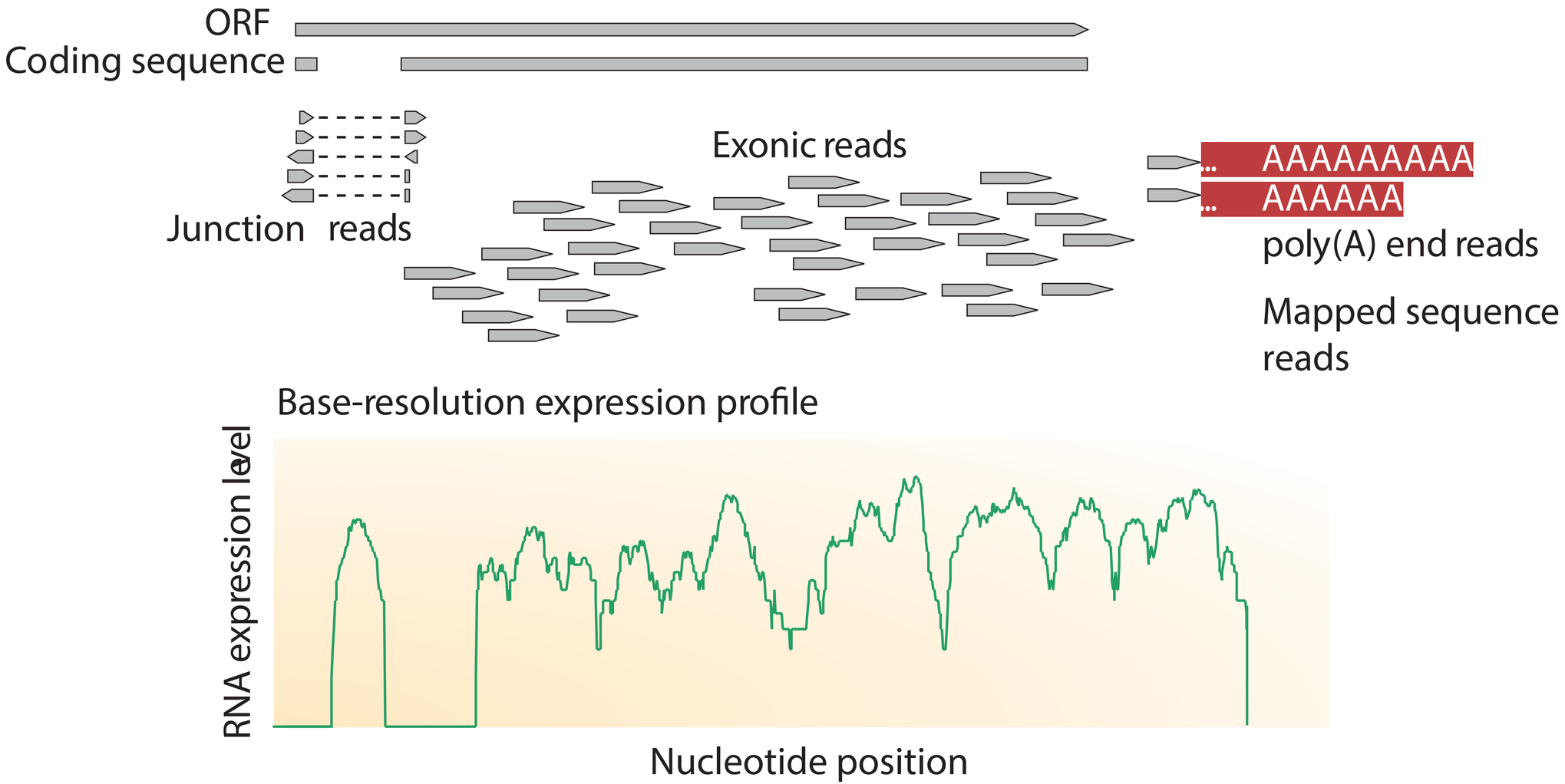
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGG
ATGAAACATTAAAGTCAAACAATATGA
.....

Hybridization to known sequences (array-based technology)

Short read sequencing (FASTQ files)

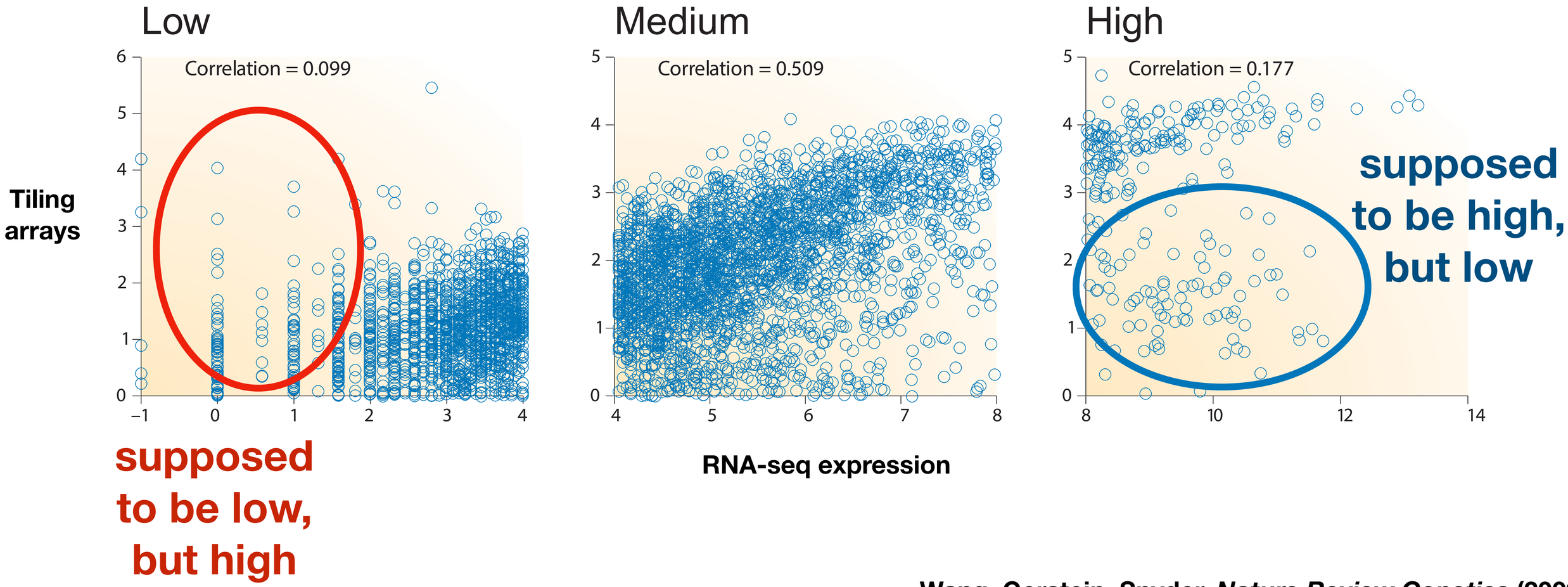
Wang, Gerstein, Snyder, *Nature Review Genetics* (2009)

Quantifying gene expressions



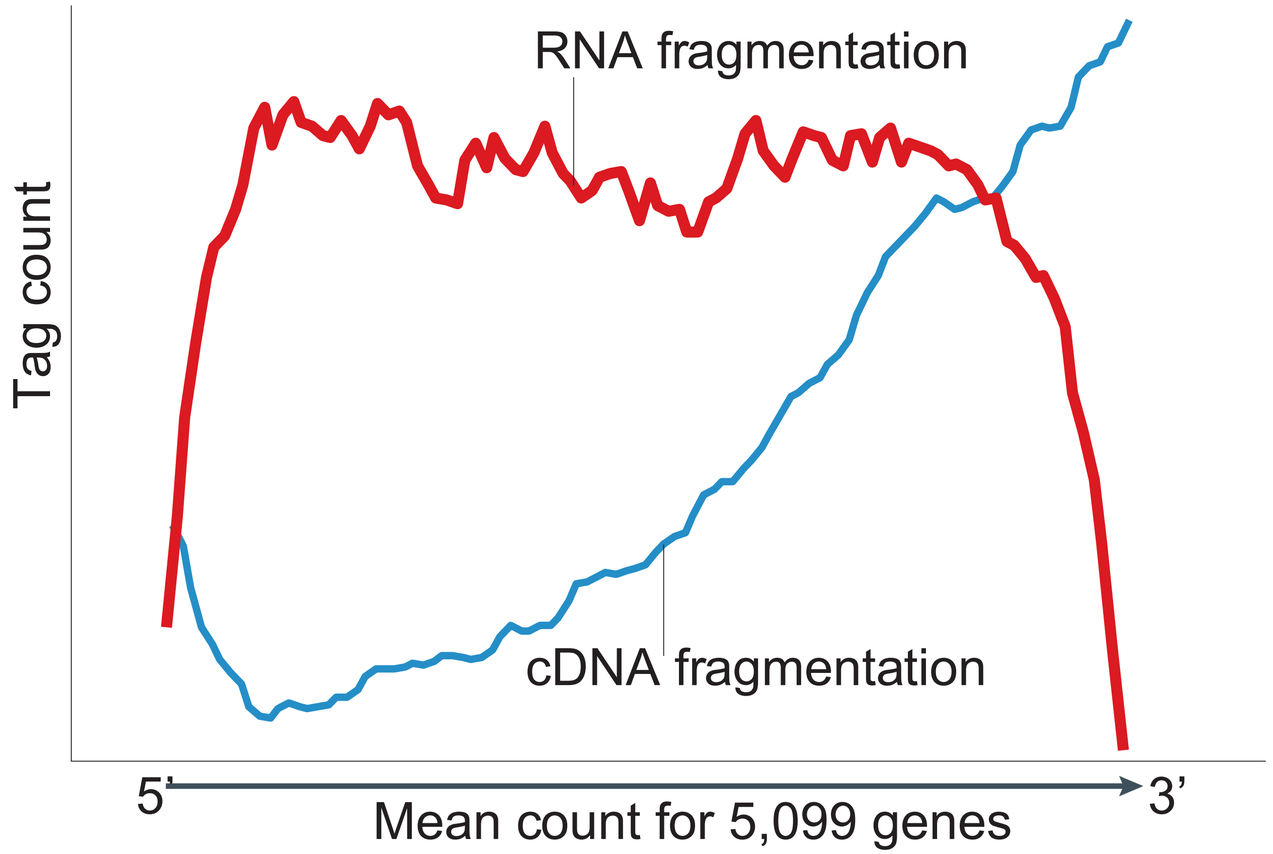
Wang, Gerstein, Snyder, *Nature Review Genetics* (2009)

RNA-seq is generally more accurate

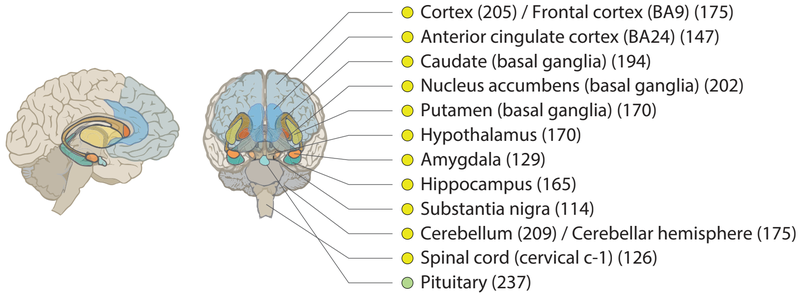


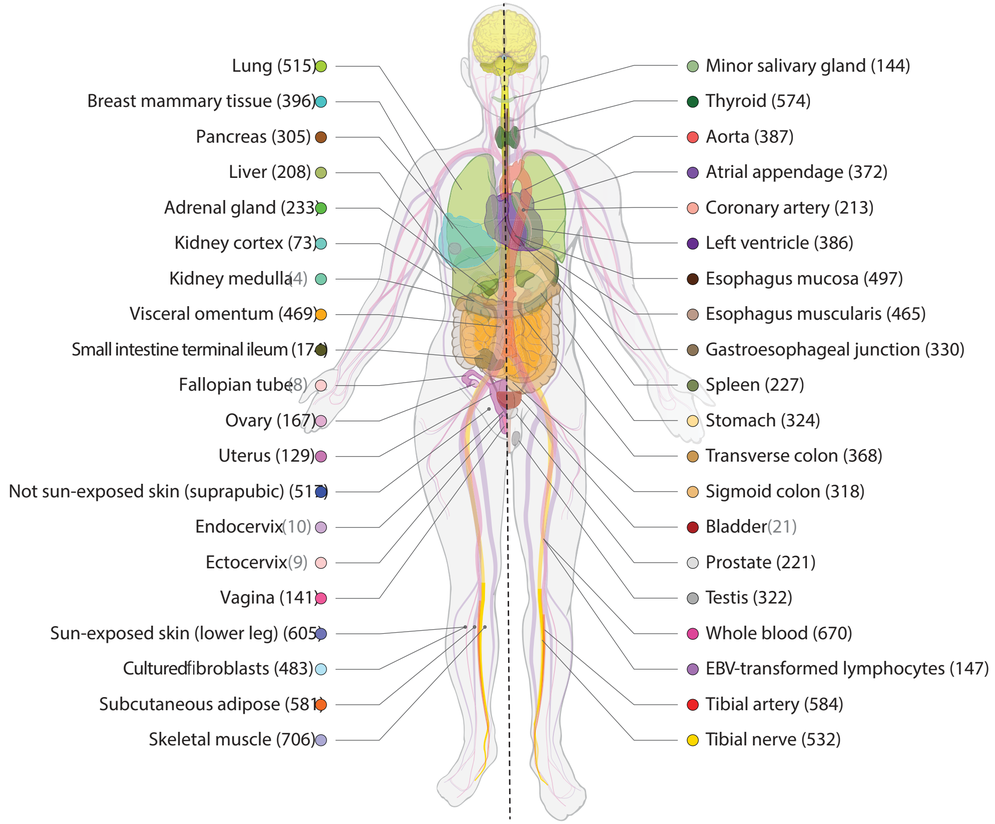
Wang, Gerstein, Snyder, *Nature Review Genetics* (2009)

RNA-seq generally provides more uniform coverage of a gene body than cDNA



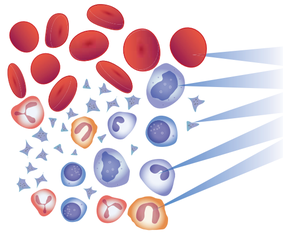
Wang, Gerstein, Snyder, *Nature Review Genetics* (2009)



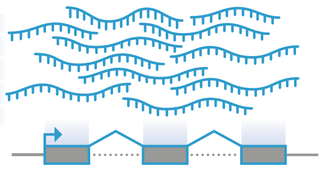


GTEx Portal

Cell type composition in tissues



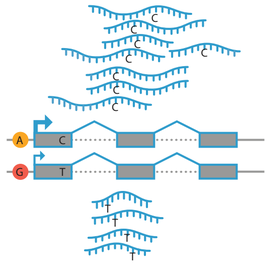
Gene expression and splicing



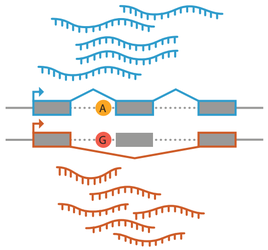
Expression quantitative trait loci (eQTLs)

Splicing quantitative trait loci (sQTLs)

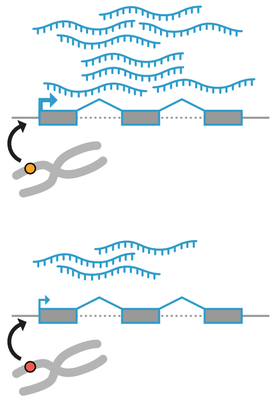
cis-eQTLs



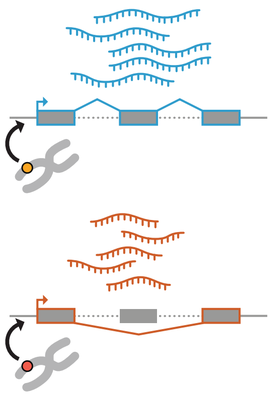
cis-sQTLs



trans-eQTLs



trans-sQTLs



GTEx, *Science* (2020)

Today’s lecture: Genomics technology

- **What is genomics technology?**
  - Measuring every step of the information flow: DNA, mRNA, splicing, protein, etc.
  - Why do we (statisticians) need to be aware of it?
- **Obtaining the book of life: a rough history of genomics methods**
  - Sequencing-based methods
  - Array-based methods
- **Understanding the book of life: omics technology**
  - Focusing on variations: mutations and expressions
  - Efforts to build epigenetic, transcriptomic, cell type references