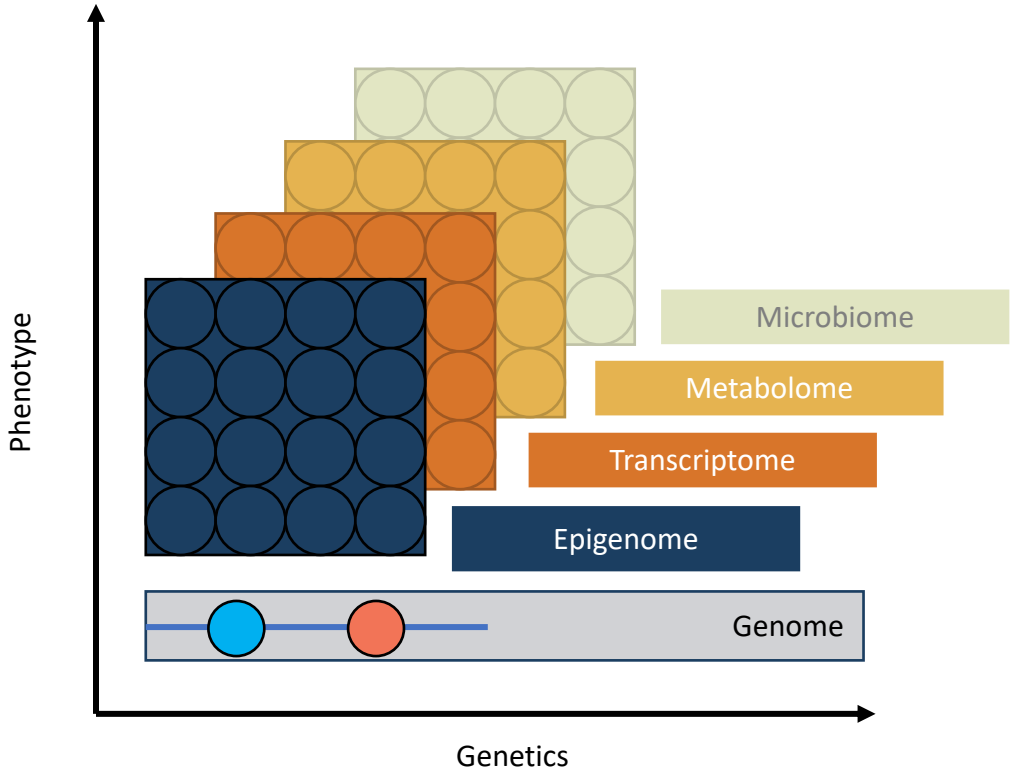
Applied Computational Genomics

(95555-PHC 6934-002, Fall 2020)

Time: 5:00 pm- 7:45 pm, Wednesday, Aug 24 – Dec 10, 2020



Instructor:

Chengqi(Charley) Wang, PhD
Jenna Oberstaller, PhD
Justin Gibbons, PhD
Thomas Keller, PhD
Anujit Sarkar, PhD
Min Zhang, MD

Teaching Assistant:

Jan Dahrendorff
Samira Jahangiri
Janelle Donglasan

UNIVERSITY OF SOUTH FLORIDA
A PREEMINENT RESEARCH UNIVERSITY

GENOMICS PROGRAM

Program Structure and mission overview

USF Genomics Program

❖ Establish a collaborative and open community to support all areas of Omics research at USF

Omics Hub

❖Build awareness for in-house capabilities

❖promote departmental and community growth

❖Computational and laboratory consulting in partnership with the Core

❖Encourage grants submission as collaborators

❖Provide training (bench, data analysis)

USF Genomics Core Facility

❖Core equipment to which certified users have **SUPERVISED** access after training and project-consultation with our Core experts

Omics expertise, consultation, and community-building

Career biologists, bench scientists, bioinformaticians, programmers, and scientific writing experts here to help you design experiments, generate preliminary data, and collaborate with you to facilitate your omics research and strengthen the USF Genomics Program

Computational expertise

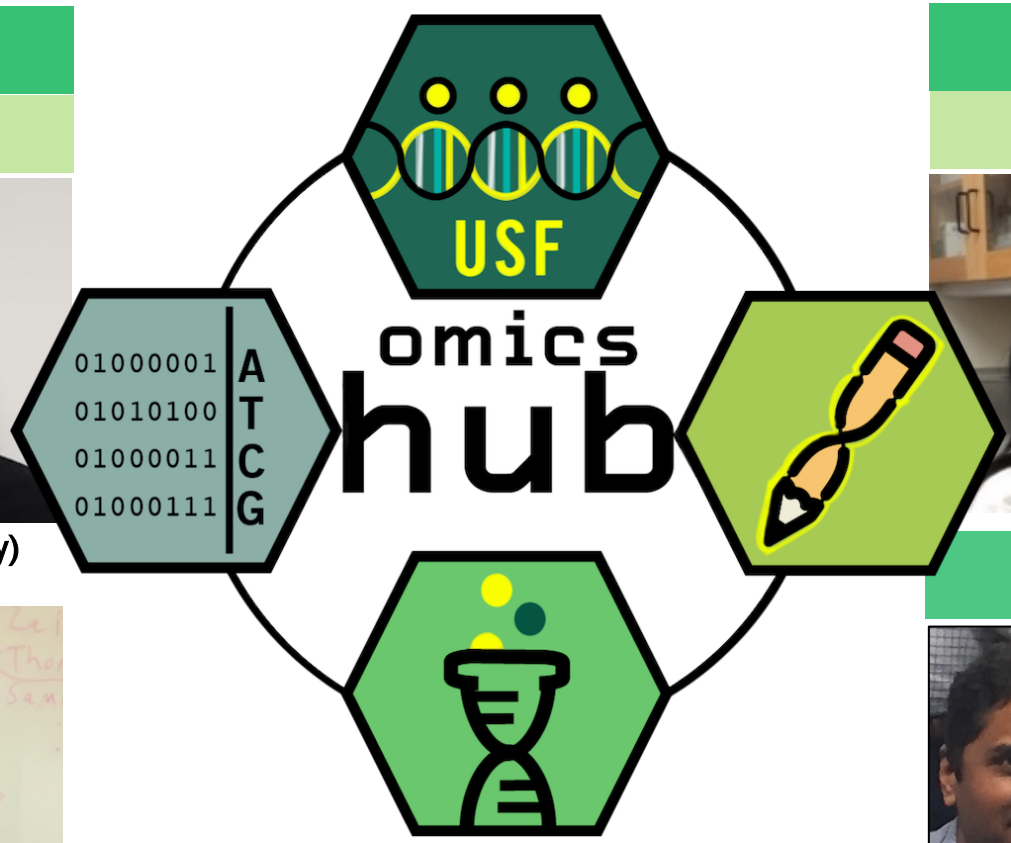
genomics@usf.edu

Jenna Oberstaller, PhD

Chengqi(Charley) Wang, PhD

Thomas Keller, PhD

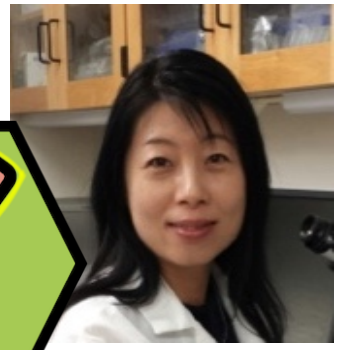
Justin Gibbons, PhD



genomics@usf.edu

Laboratory expertise

genomics@usf.edu



Min Zhang, MD

Core instrumentation expert

Microbiome specialist

Anujit Sarkar, PhD

All Things Microbiome

Swamy Rakesh Adapa, MS

A founding Hub contributor

Key Hub Course Concepts

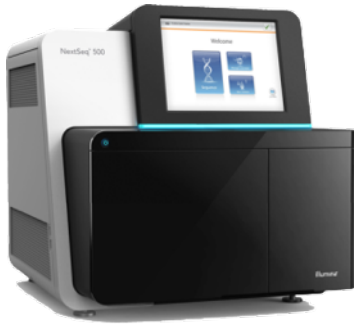
- EXPERIMENTAL DESIGN is paramount
- BEST PRACTICES
  - at the bench
  - at the computer
- Data-analyses
  - QUALITY CONTROL
  - APPROPRIATE analyses to apply to answer YOUR biological question
    - many different methods, stats to be considered
  - REPRODUCIBILITY
- Training for biologists
  - we can't remove ALL the code
  - but focus = straightest path from data to biology/interpretation

USF GENOMICS CORE FACILITIES AND RESOURCES

- Supervised access to sequencing core
- Grant-application support
- Consultation for sequencing and data analysis

FREE 2 hours per USF PI!

Illumina - MiSeq System



NextSeq 550 System

10x Chromium SC

TapeStation / Bioanalyzer

M220 Focused-Ultrasonicator

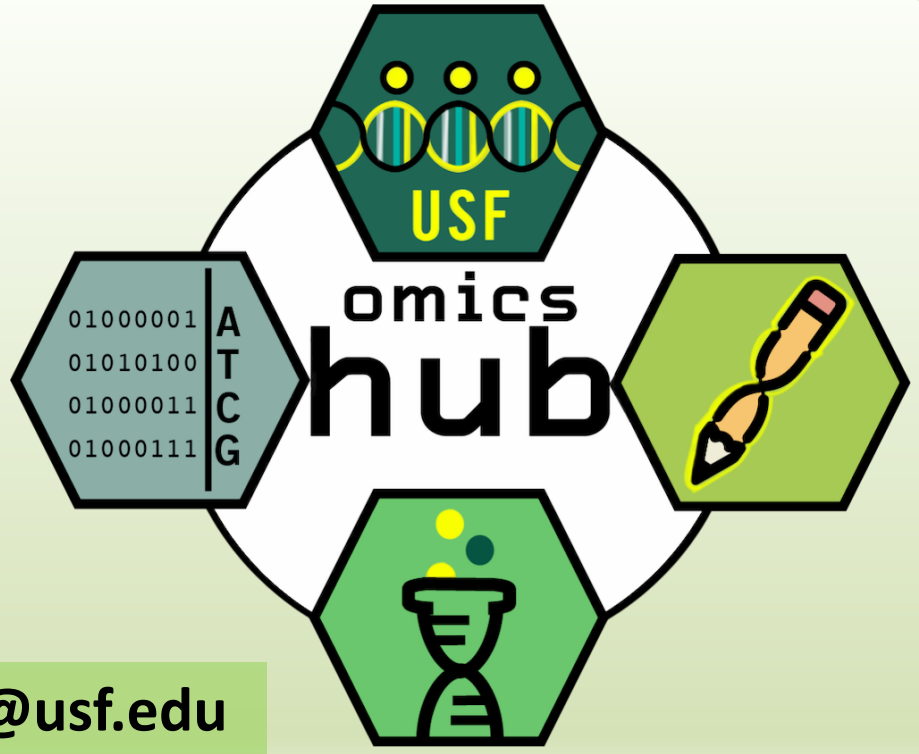
Qubit® 2.0 Fluorometer

Roche LightCycler 96 qPCR

- Whole-Genome Sequencing
- RNA Sequencing
- Targeted Gene Sequencing (amplicon, gene panel)
- Gene Expression Profiling
- Targeted Gene Expression Profiling
- miRNA & Small RNA Analysis
- Methylation Sequencing
- 16S Methylation Sequencing
- Single-Cell Sequencing
- DNA, RNA Quantification
- Sequencing Consultation

USF Genomics Core Computational Facilities and Resources

- Omics Consultation
- Transcriptomics
- Epigenomics
- Phylogenomics
- Microbiome and Metagenomics
- Genome Assembly and Annotation
- Custom Scripts and Analysis Pipelines
- Data Visualization
- Genomics Training



genomics@usf.edu

Core Computational Resources:

- USF's High-Performance Computing (HPC) Cluster, managed by the USF Department of Research Computing
- Sequence data are securely stored on HIPAA-compliant architecture of USF's HPC cluster

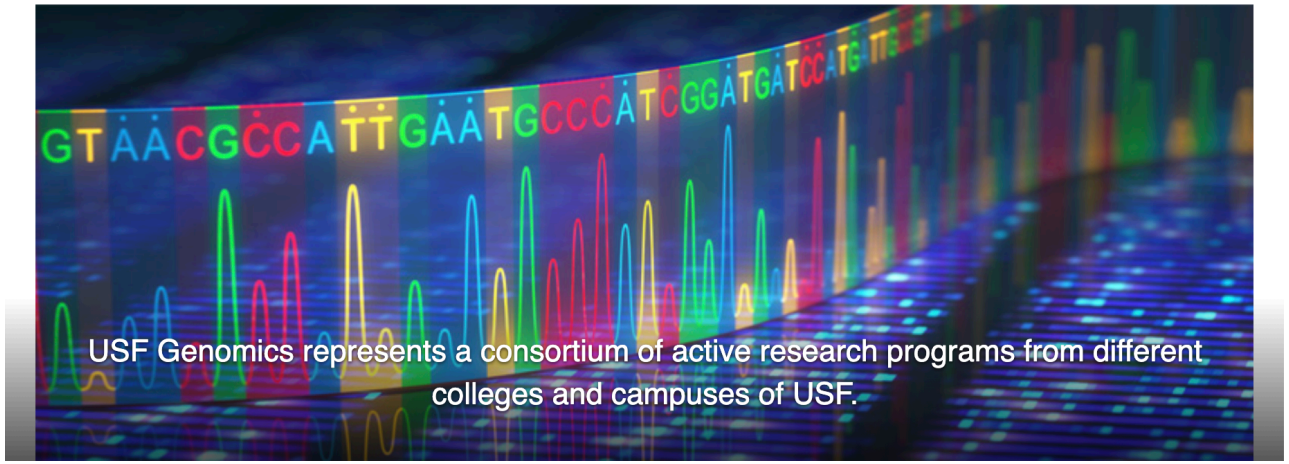
USF Genomics home page

USF HEALTH

College of Public Health

USF Genomics

- Home
- Researchers
- Grants
- Degree Programs
- Undergraduate Opportunities
- Annual Symposium
- College of Arts and Sciences News
- Genomics Training
- Request Services



About USF Genomics

Our projects integrate modern genomic approaches to understand and help develop solutions for some of the major challenges affecting global health and especially our Florida community. USF Genomics research projects are naturally interdisciplinary with studies that cross many boundaries of traditional scientific fields, integrating bioscience approaches in experimental and environmental research with computation biology. Opportunities exist for new graduate students to join well-funded research programs to study disease causes from underlying drug resistance and virulence genes in pathogenic organisms to changes of microbial communities within our bodies to the marine environment.

Register and Submit Abstracts for the USF Genomics Annual Symposium on November 15, 2019

Upcoming Events

USF Genomics Seminars: Tuesdays once a month, 4:00-5:00 pm - IDRB 302

USF Genomics Seminar - Dr. Maureen Groer - Microbiome State of the Science - Tuesday, September 17, 4:00-5:00 pm - IDRB 302

USF Genomics Forum Discussion Series - Tuesdays once a Month, 5:00-6:00 pm - IDRB 302

USF Genomics Annual Symposium: Personal Genomics - November 15, 2019

Genomics Laboratory and Data-Analysis Training Courses: February, May, September 2019

Full COPH Calendar

Introduction

Instructor:

-- Chengqi(Charley) Wang, PhD

-- My own research

Functional Genomics, Epigenetics, Machine Learning and Deep Learning algorithm

Instructor:

-- Jenna Oberstaller, PhD

-- My own research

Functional genomics, evolutionary genomics, systems biology, omics tools-development for biologists

Instructor:

-- Justin Gibbons, PhD

-- My own research

He is currently working on several projects ranging from drug resistance in malaria to how the microbiome effects the health of preterm infants

Introduction

Instructor:

-- Thomas Keller, PhD

-- My own research

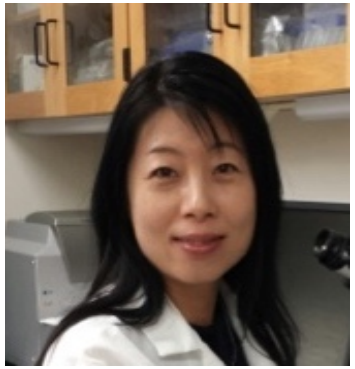
Epigenomics, microbiome, clinical applications to machine learning & deep learning

Invited lecturer:

-- Anujit Sarkar, PhD

-- My own research

Interrelationship between cardiovascular disease, T1D and oral microbiome; Role of gut microbiome in the long term growth and development of preterm infants



Invited Lecturer:

-- Min Zhang, MD

-- My own research

Molecular Genetics, Next Generation Sequencing, Single-cell Sequencing,
Malaria parasite biology, Genetic analysis of critical metabolic process

Introduction

TA:

-- Jan Dahrendorff

-- My own research

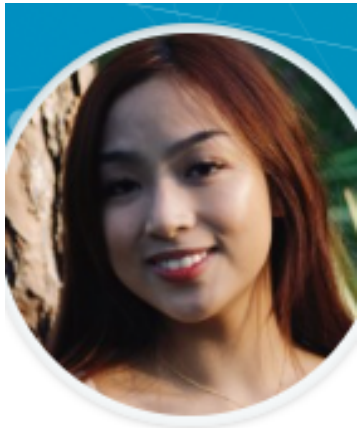
Differential gene expression in longitudinal cohort of post traumatic stress disorder

TA:

-- Samira Jahangiri

-- My own research

Variant analysis in Genome-wide associated studies, differential gene expression analysis and the functional study of disease causing mutations using high-throughput data.



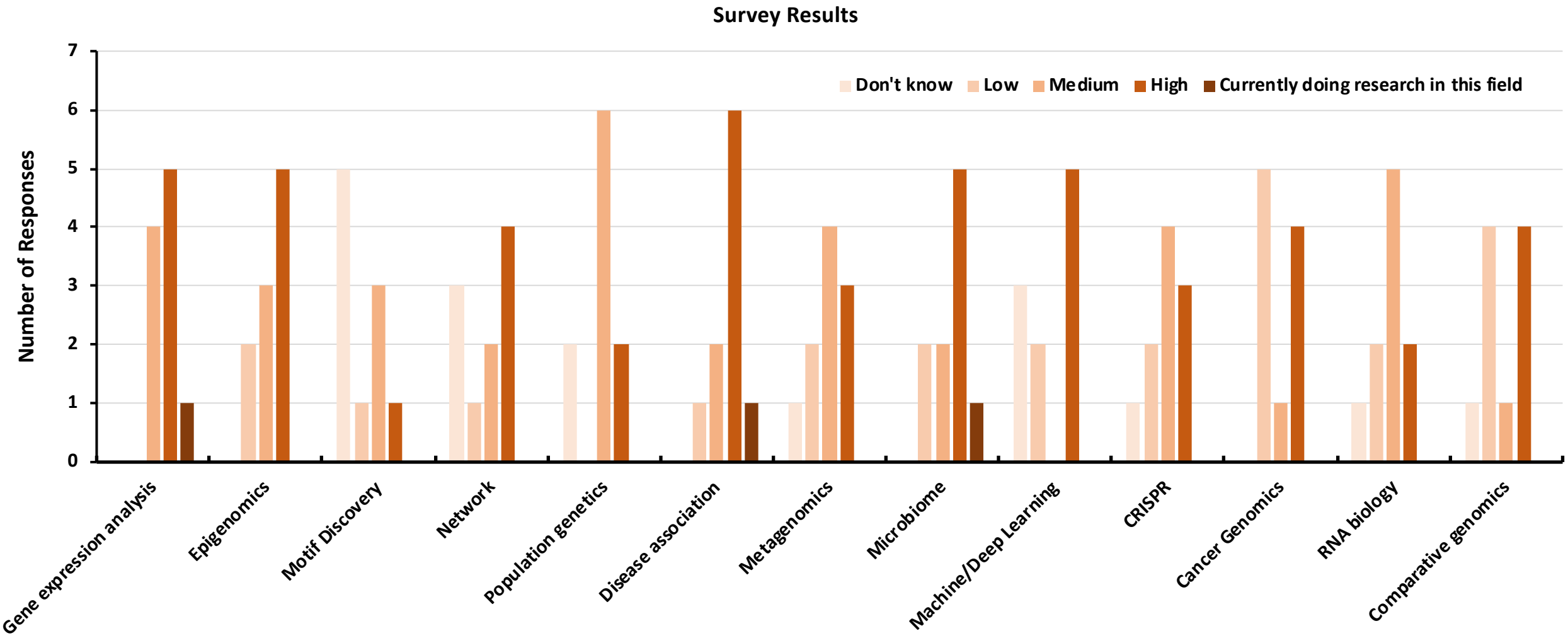
TA:

-- Janelle Donglasan

-- My own research

Identify genetic and epigenetic predictors of stress-related mental disorders, with a particular focus on depression and post-traumatic stress disorder

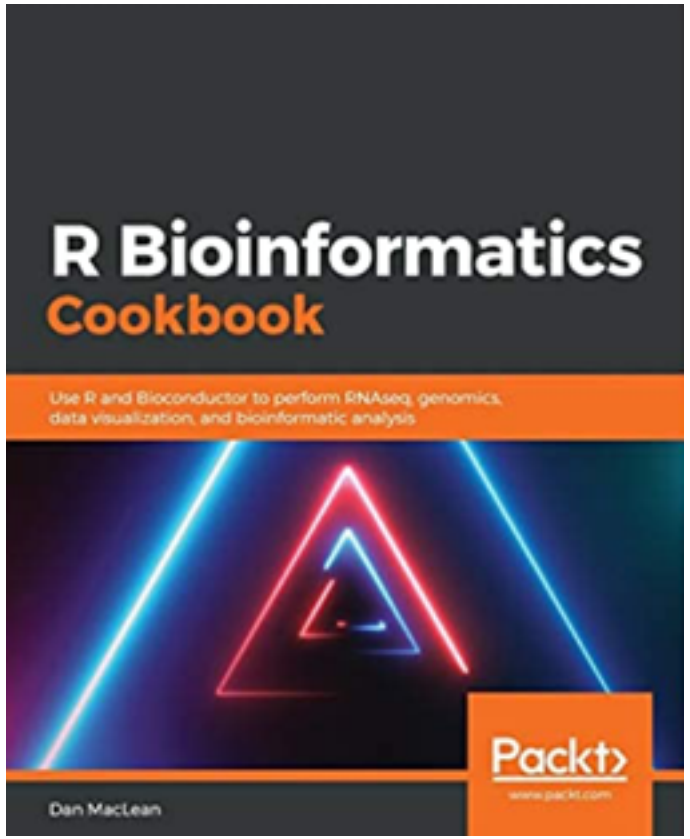
Students (From first-day survey)



Interest in specific topic

	Don't know	Low	Medium	High	Currently doing research in this field	Total
Gene expression analysis			3	5	1	9
Epigenomics		2	2	5		9
Motif Discovery	4	1	3	1		9
Network	2	1	2	4		9
Population genetics	2		5	2		9
Disease association		1	1	6	1	9
Metagenomics	1	1	4	3		9
Microbiome		2	1	5	1	9
Machine/Deep Learning	2	2		5		9
CRISPR	1	2	3	3		9
Cancer Genomics		5		4		9
RNA biology	1	2	4	2		9
Comparative genomics	1	3	1	4		9
Total	14	22	29	49	3	117

(Optional) Book for the Course



R Bioinformatics Cookbook: Use R and Bioconductor to perform RNAseq, genomics, data visualization, and bioinformatic analysis

Maclean Dan

Grading Criteria

Assessment	Percent of Final Grade
Project 1	30%
Project 2 (Final)	40%
In-Class Assignment	20%
Participation and Attendance	10%
Total	**100%**

In-lass Assignment

Each class will have an In-Class Assignment to complete the code discussed in class (shell or R-script).

Each assignment must be submitted to the canvas within 72 hours.

Lesson 2 [9/2/20]	Chenggi (Charley) Wang **TA:** Samira Jahangiri	**Part 1:** Introduction to the Linux directory system and how to get around it **Part 2:** Text Editor (Vim) in Linux **Part 3:** Piping and Redirection in Linux **Part 4:** Slurm work load manager	**Pre-class reading:** Basic knowledge of Shell script (Learning the Shell) http://linuxcommand.org/lc3_learning_the_shell.php In-class assignment	7:45 pm [9/5/20]

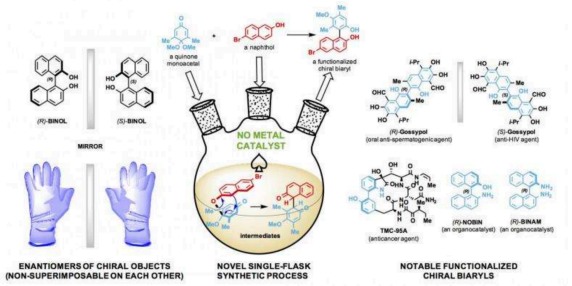
Q&A

Q&A (Every Thursday evening from 8:00pm -8:40 pm)

How can genetics research help meet outstanding medical challenges

Informing therapeutic development

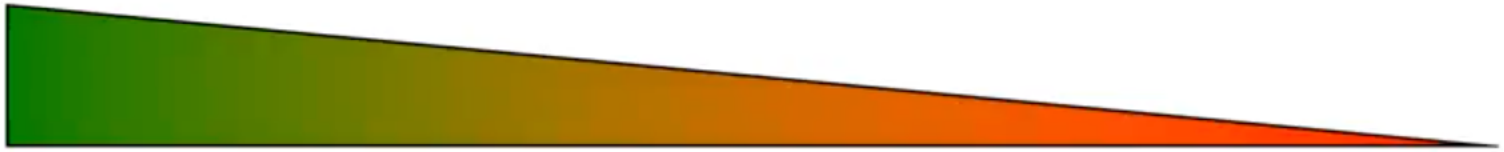
Identify target



Chemical screen

Animal models

Clinical trails

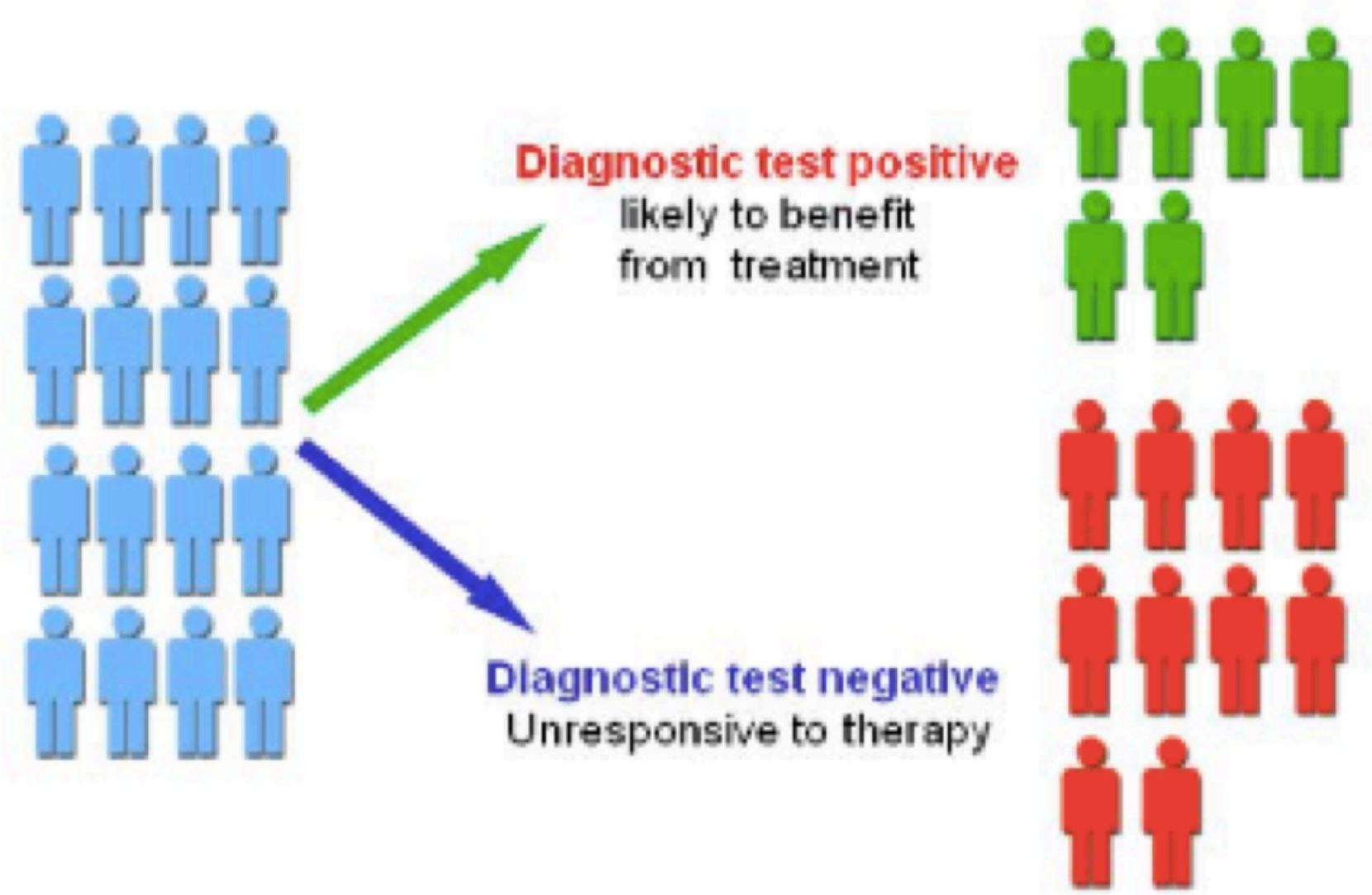


Most therapies developed through traditional approaches fail to have efficacy

Therapeutic development desperately needs into human biology

Genetics provides a validated means to these insights

Informing therapeutic development



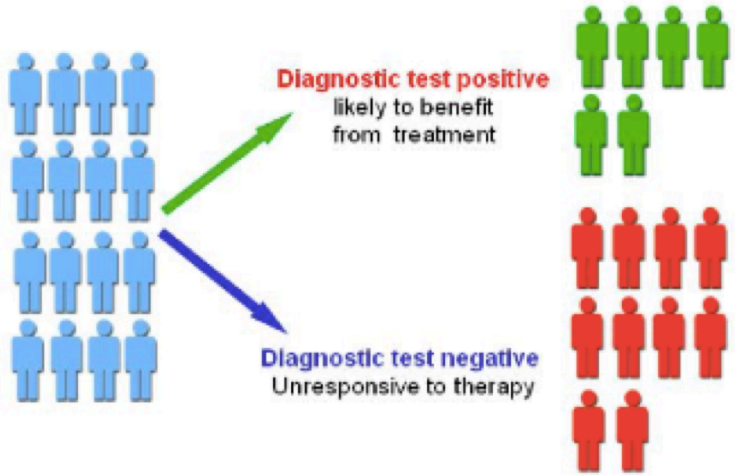
Personalized genomic medicine

Genetics can in some cases provide individualized medical insights

- diagnostics in cases of severe genetic disorders

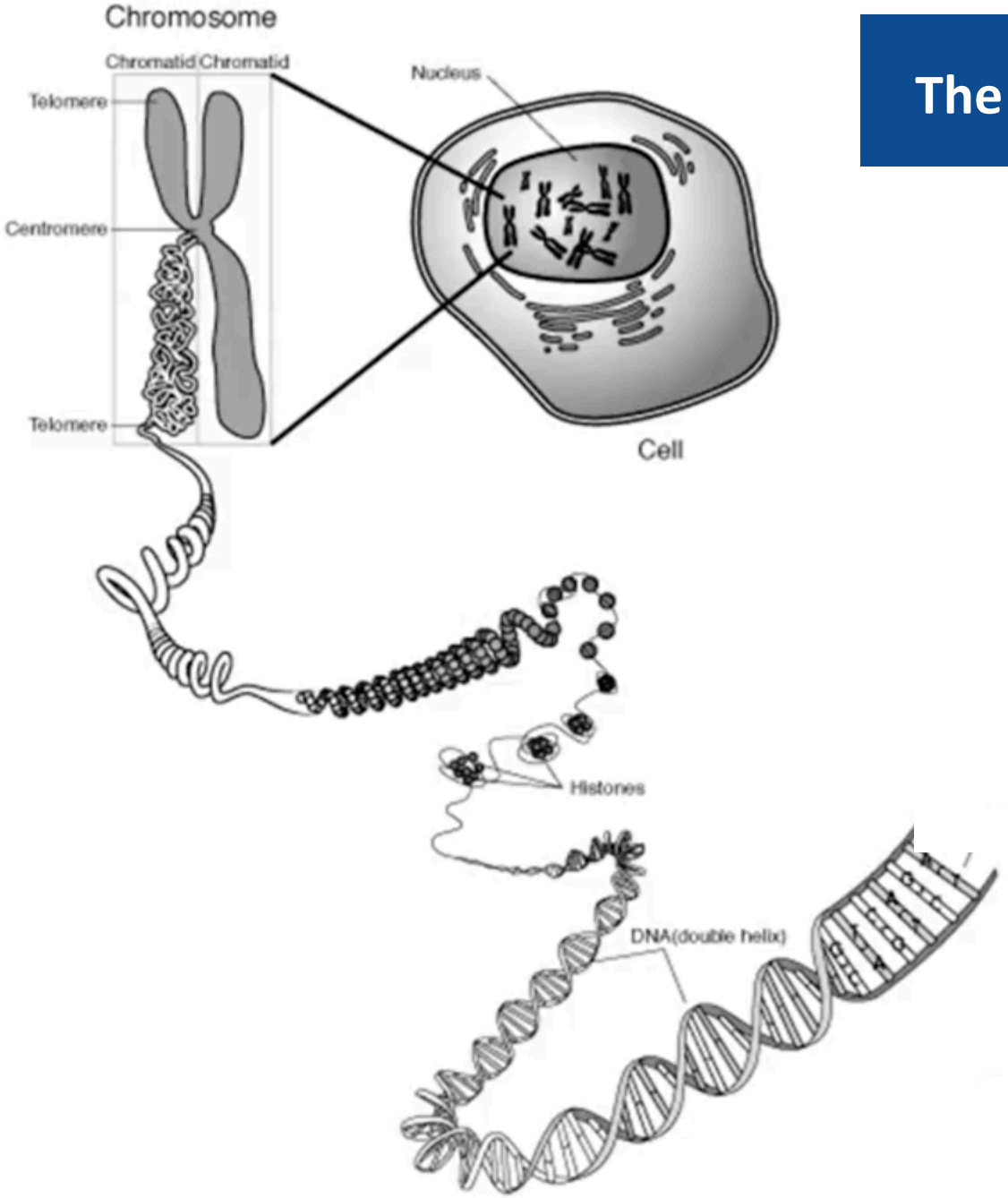
- identification of individuals more or less likely to benefit from specific therapeutic interventions

- prediction of individuals at risk for severe idiosyncratic adverse drug responses



Vastly oversold for years - technological advances are bringing what has long been a futuristic vision much closer to reality

The scope of the challenge



Within each cell:

2 copies of the genome

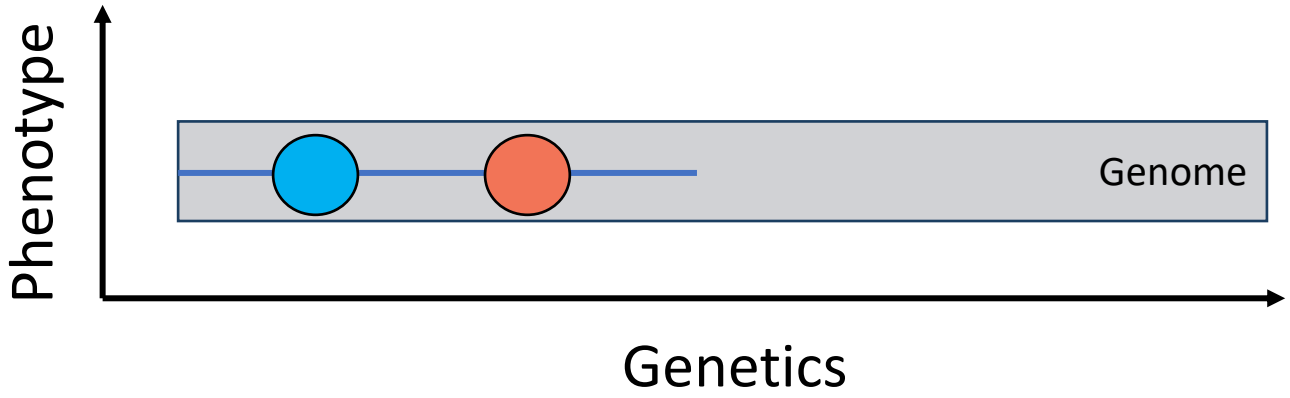
23 chromosomes

~20,000 genes

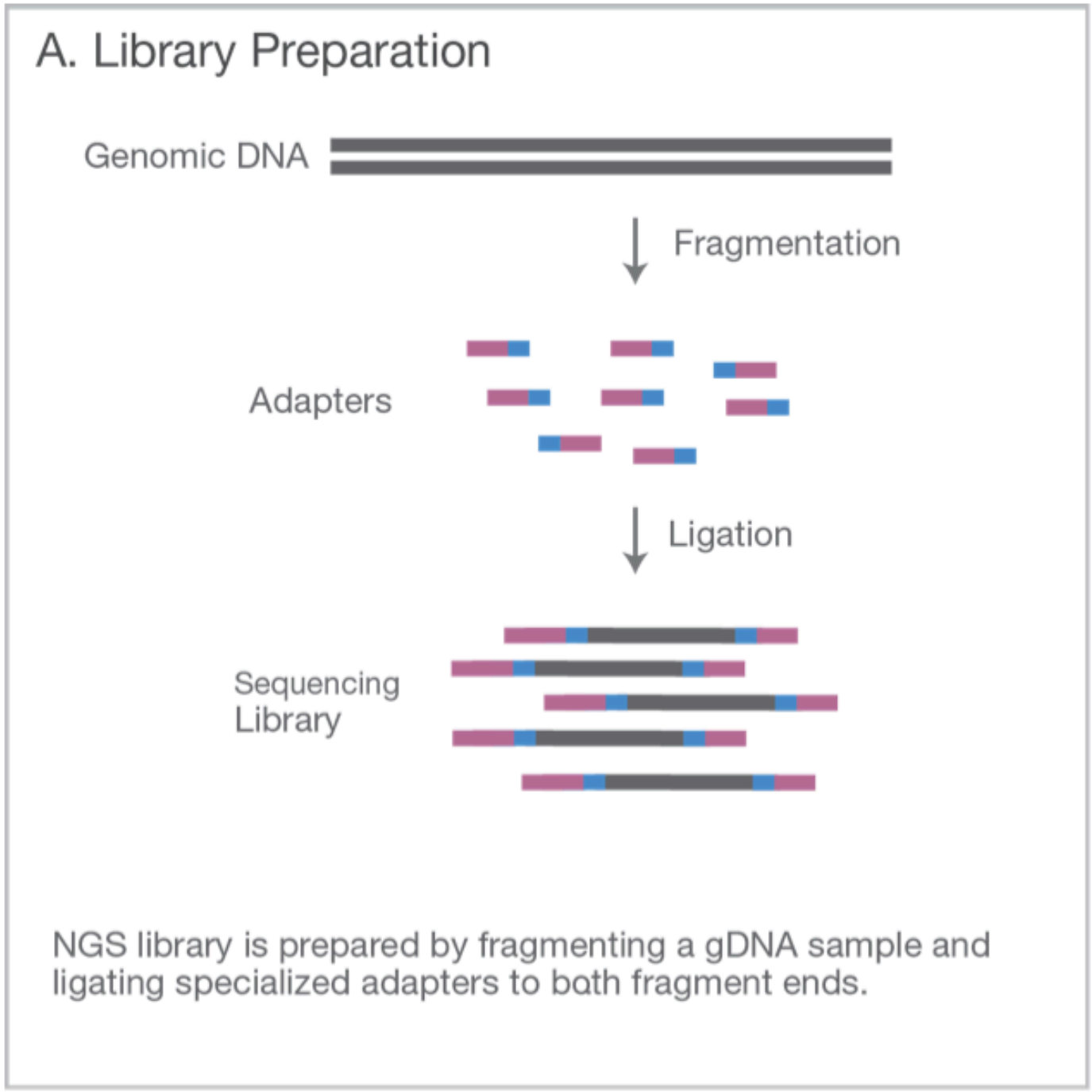
3.2B letters of DNA

Millions of polymorphic sites

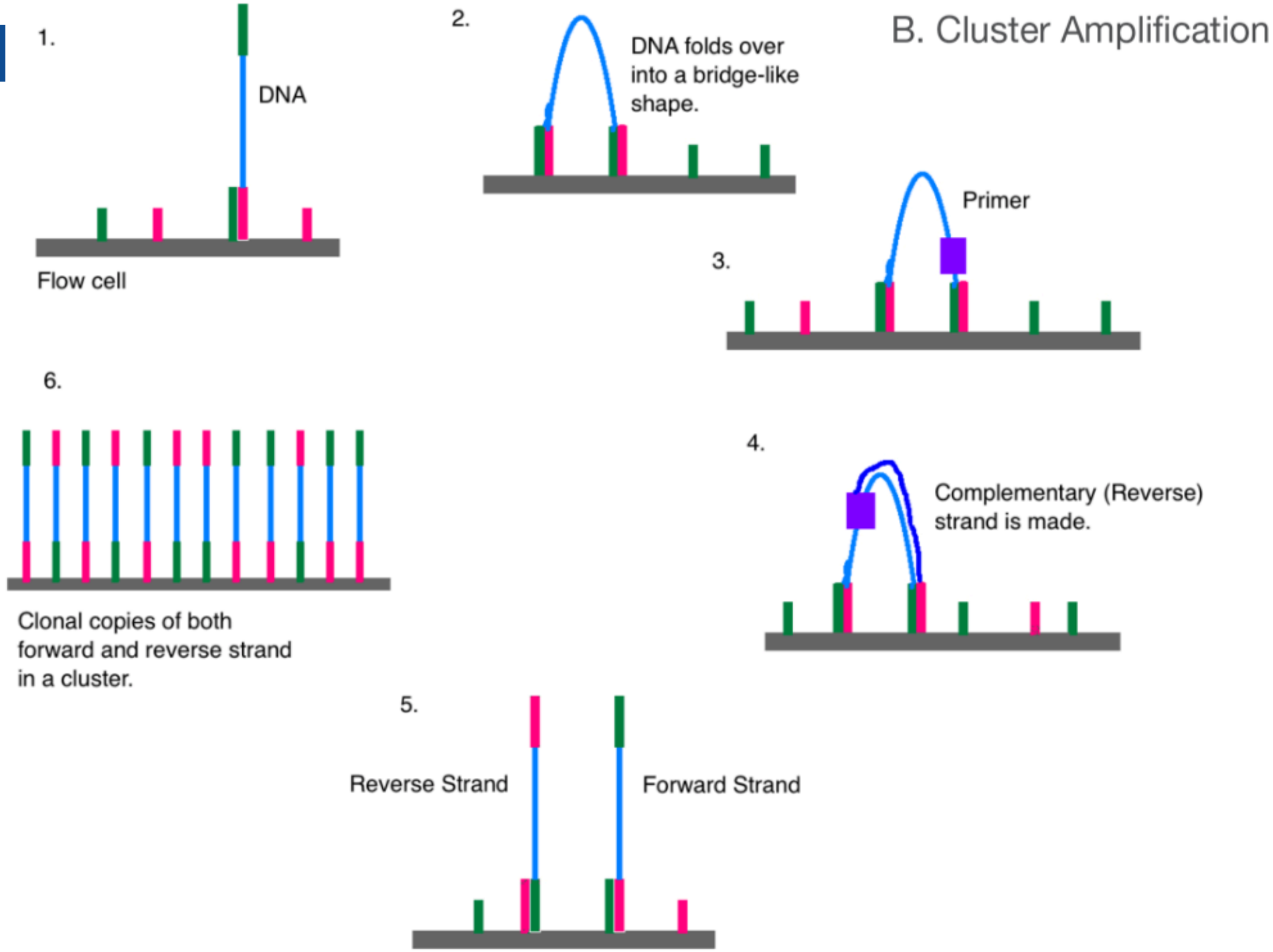
Genomics



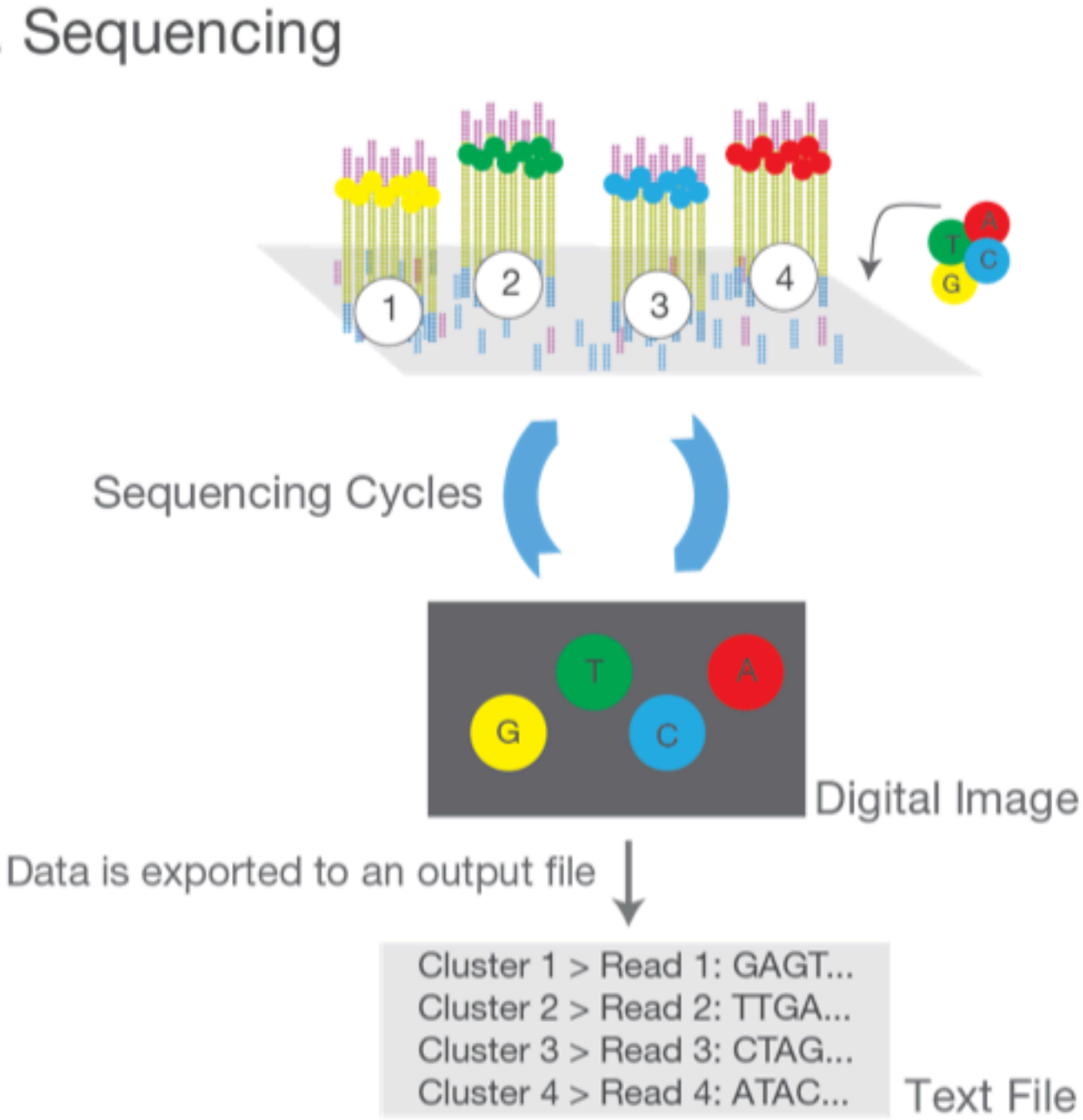
Illumina sequencing



Illumina sequencing



Illumina sequencing



Discovering genetic variation: Sequencing

Some mapping procedure ignores mismatches, and matched reads that were highly similar.

But they may represent true sequence variants

Statistical methods needed to distinguish true variants (Variant calling)

Whole genome variant callline: GATK HaplotypeCaller

A haplotype is a group of genes or DNA sequences that are inherited together on the same chromosome

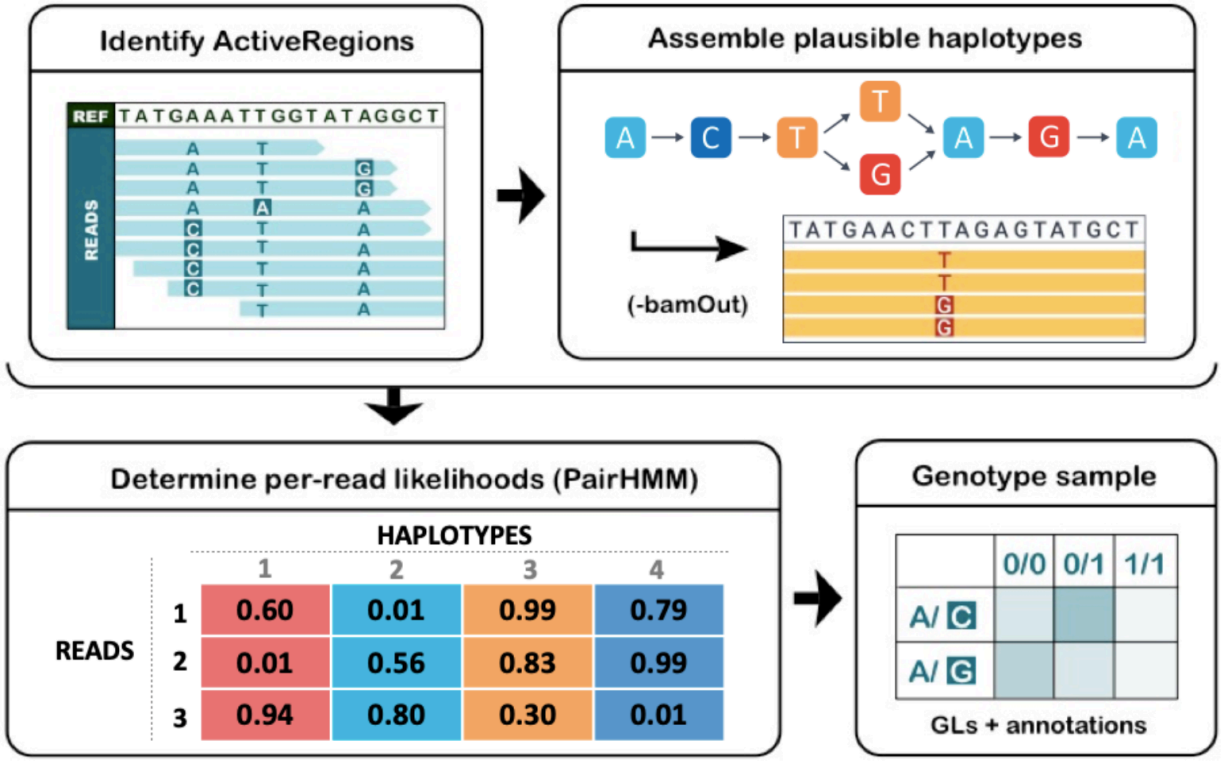
1. Use heuristic to find mismatches not explained by noise
2. Use assembly graph to identify possible haplotypes
3. For each haplotype, estimate

P(read | haplotype)

Using probabilistic sequence alignment

- Hidden Markov Model
- States: insertion, deletion, substitution
- Emission: pairs of aligned nucleotides/gaps
- Transitions: equivalent to insertion/deletion/gap penalties from Smith-Waterman algorithm
- Get **P(read | haplotype)** using forward-backward algorithm

4. Use Bayes rule to get P(Haplotype | read)
5. Assign genotypes to each sample based on the max a posteriori haplotypes



https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

Type of genetic variation

- **99% of DNA is shared** between two individuals
- Variation in the remainder explains all our **predisposition** differences
- **Remaining** phenotypic variation: environmental/stochastic differences

Name	**Example**	**Frequency in one genome**
Single nucleotide polymorphisms (**SNPs**)	GAGGAGAACG[C/G]AACTCCGCCG	1 per 1,000 bp
Insertions/deletions (**indels**)	CACTATTC[C/CTATGG]TGTCTAA	1 per 10,000 bp
Short tandem repeats (**STRs**)	ACGGCAGTCGTCGTCGTCACCGTAT	1 per 10,000 bp
Structural variants (**SVs**) / Copy Number Variants (**CNVs**)	Large (median 5,000 bp) deletions, duplications, inversions	1 per 1,000,000 bp

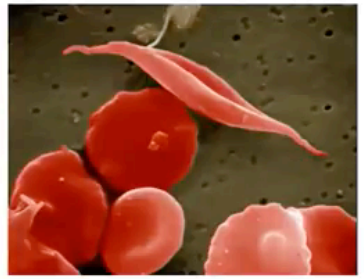
Type of genetic variation

CATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTG

CATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTG

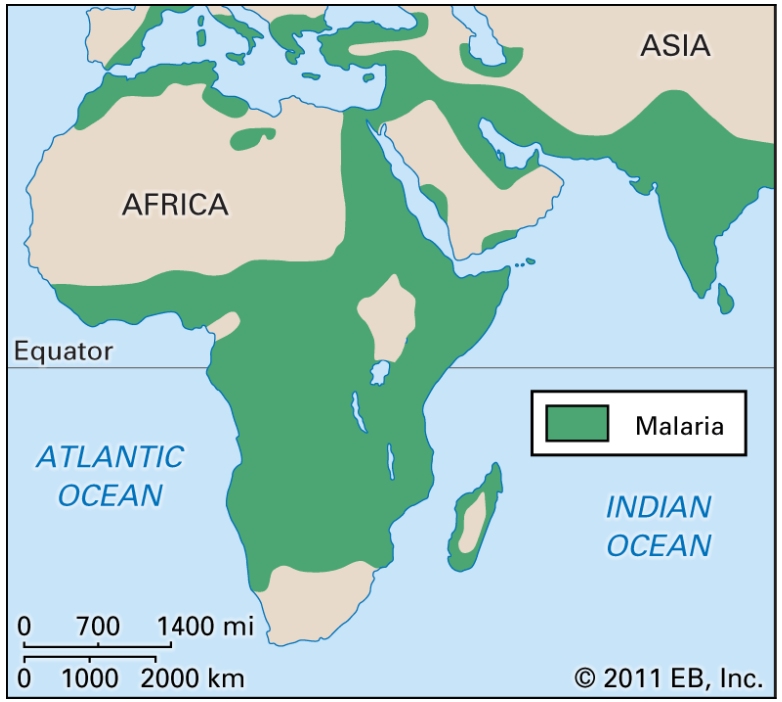
glutamic acid > valine

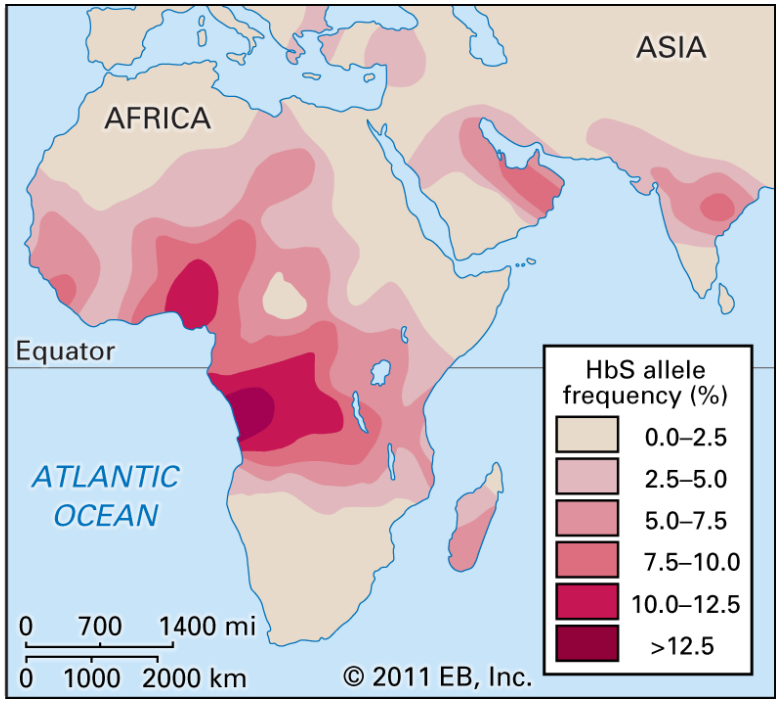
Sickle Cell Anemia



The disease is characterized by many of the symptoms of chronic anemia (fatigue, pale skin, and shortness of breath) as well as susceptibility to infection, jaundice and other eye problems, delayed growth, and episodic crises of severe pain in the abdomen, bones, or muscles.

The distribution of malaria and the distribution of sickle cell anemia overlap in areas of Africa, southern Asia, and the Mediterranean. The persistence of the *HbS* gene, which causes sickle cell anemia, has been explained by the fact that heterozygous persons are resistant to malaria.





https://www.cdc.gov/malaria/about/biology/index.html

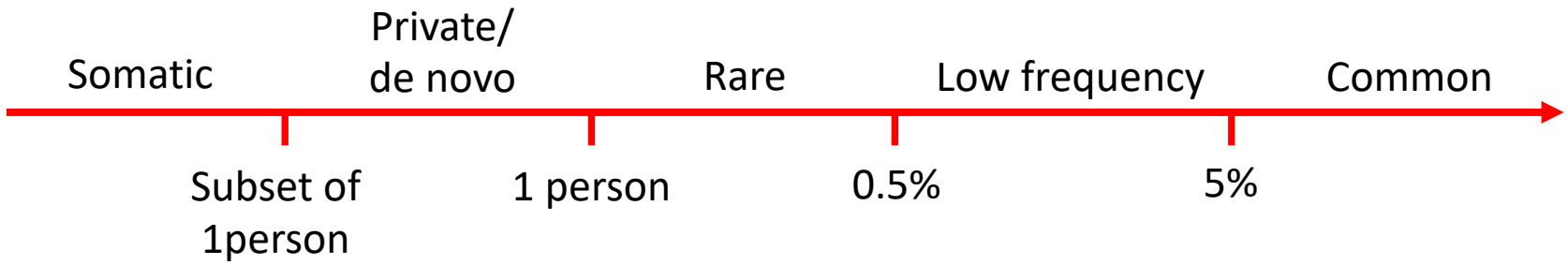
- Many modern analyses (GWAS, eQTL) focus on SNPs/indels
- Often have only two **alleles** (states)
- Identified as reference SNP clusters (**rsid**)
- Submitted sequences containing a variant are clustered to build a database (**dbSNP**)
- To date, >100 M known variants in dbSNP

rs189107123

GAGGAGAACG[C/G]AACTCCGCCG

Classifications of SNP alleles

Classifying variants by minor allele frequency:



Distinguishing the two alleles

- Matching the human reference sequence (Reference/Alternate)
- Being more frequent in the population (Major/Minor)
- Matching the most recent common ancestor between human and chimpanzee (ancestral/derived)
- Based on their disease association (risk/non-risk)

Example: rs189107123

GAGGAGAACG[C/G]AACTCCGCCG

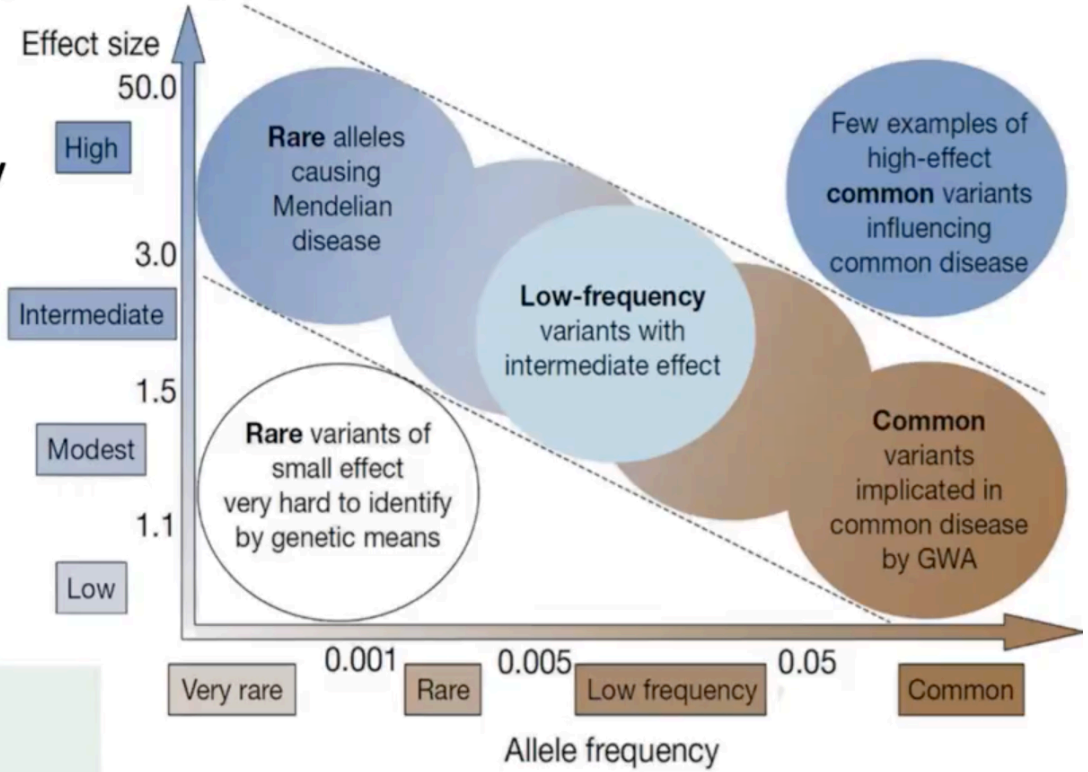
Reference allele: C

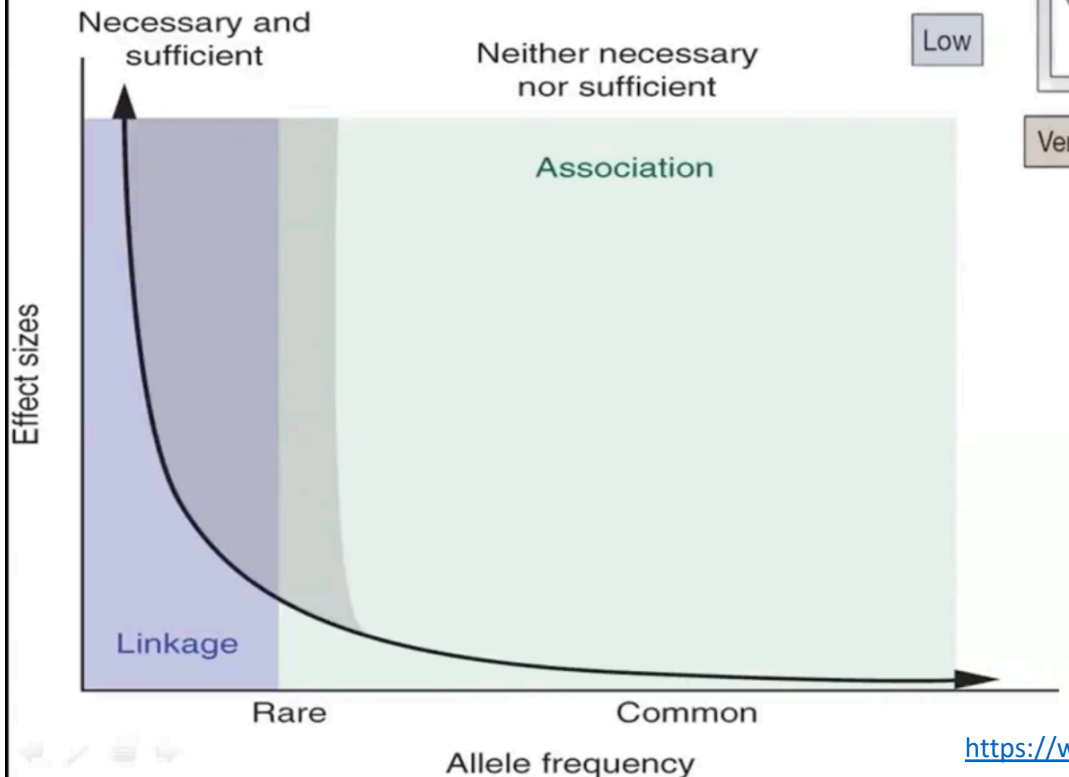
Minor allele: G (frequency 0.03 in Europeans)

Ancestral allele: unknown (**why?**)

Common alleles typically have small effects

Discovery method tuned to variant effect size/frequency



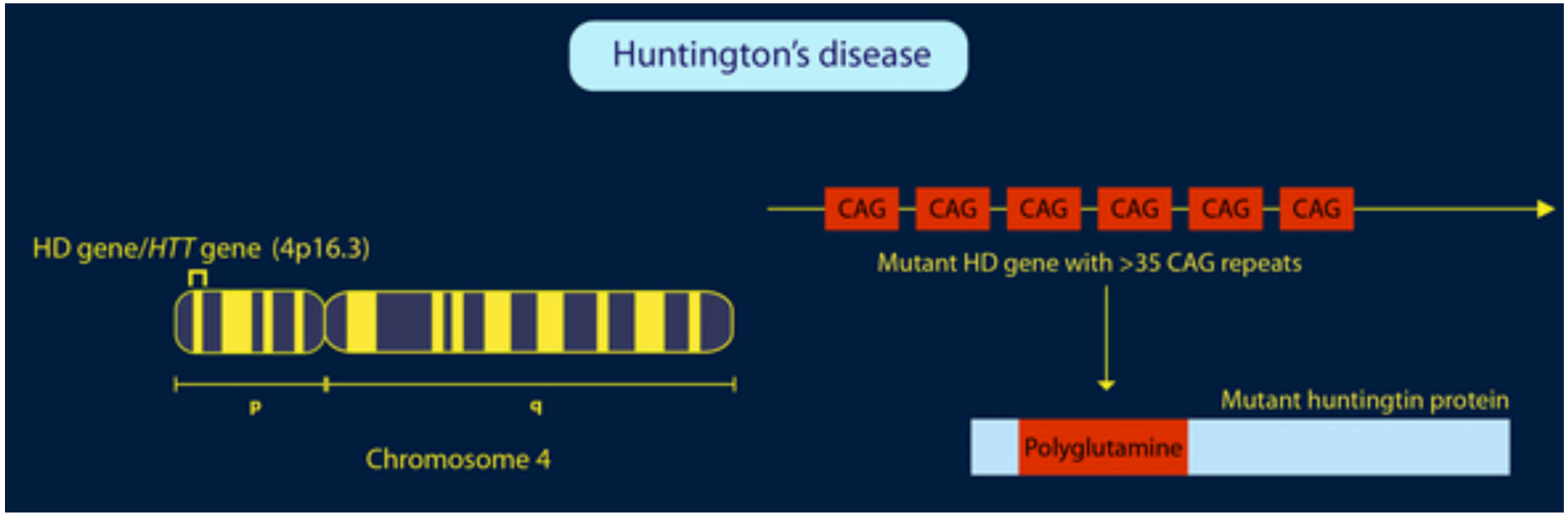


Discovery method tuned to variant effect size/frequency

https://www.dna-ghost.com/single-post/2017/09/04/Variant-frequency-effect-size-and-GWAS-limitation

Beyond SNPs: Tandem repeats

https://ghr.nlm.nih.gov/condition/huntington-disease



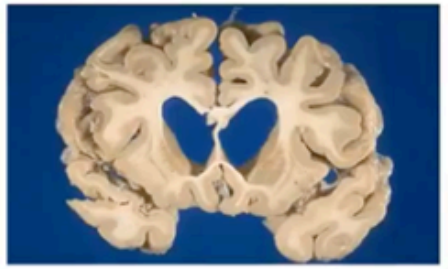
9 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT

10 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT

12 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT

> 30 Huntington's Disease

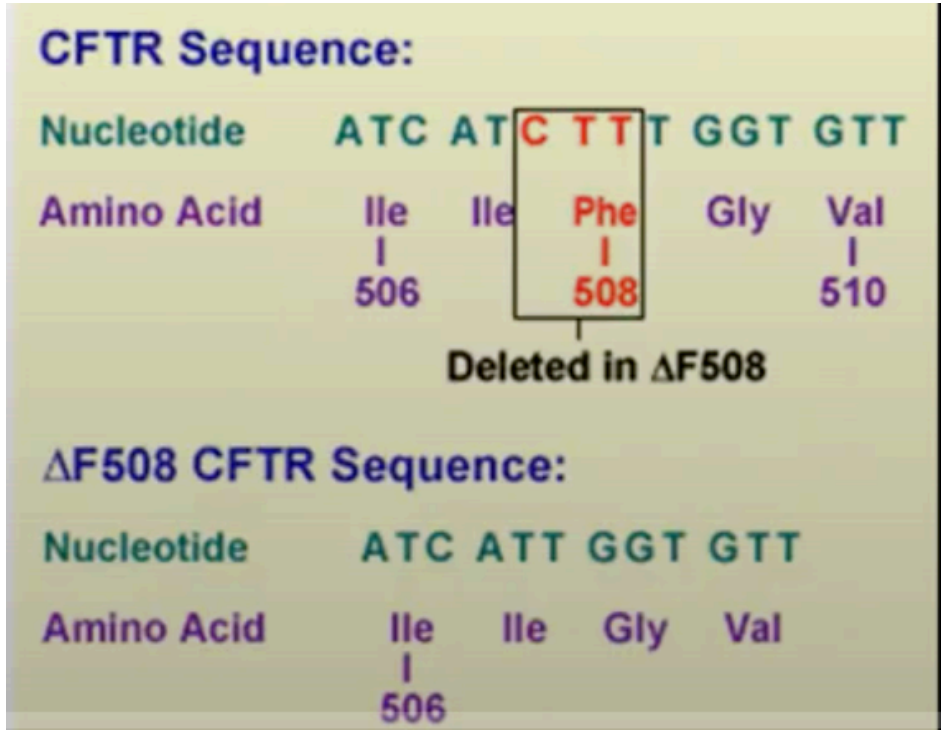
Abnormal protein, damages neurons, brain cell death, mood, coordination, speaking, dementia, etc



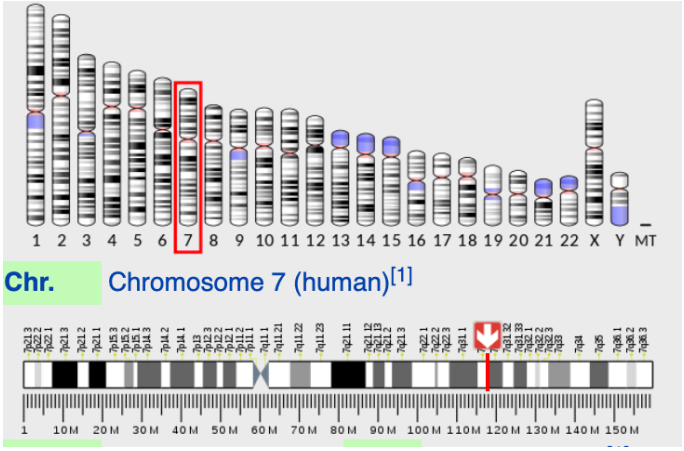
Beyond SNPs: Insertion/Deletions

Cystic fibrosis transmembrane conductance regulator (CTFR) → Lung infections, Cystic fibrosis

```
CATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAAT
CATTAAAGAAAATATCATTGGTGTTTCCTATGATGAATA
```



DeltaF508 (ΔF508) results from a deletion (Δ) of three nucleotides which results in a loss of the amino acid phenylalanine (F) at the 508th position on the protein. As a result, the protein does not fold normally and is more quickly degraded. This protein does not escape the endoplasmic reticulum for further processing. Having two copies of this mutation (one inherited from each parent) is by far the most common cause of cystic fibrosis (CF), responsible for nearly two-thirds of cases worldwide



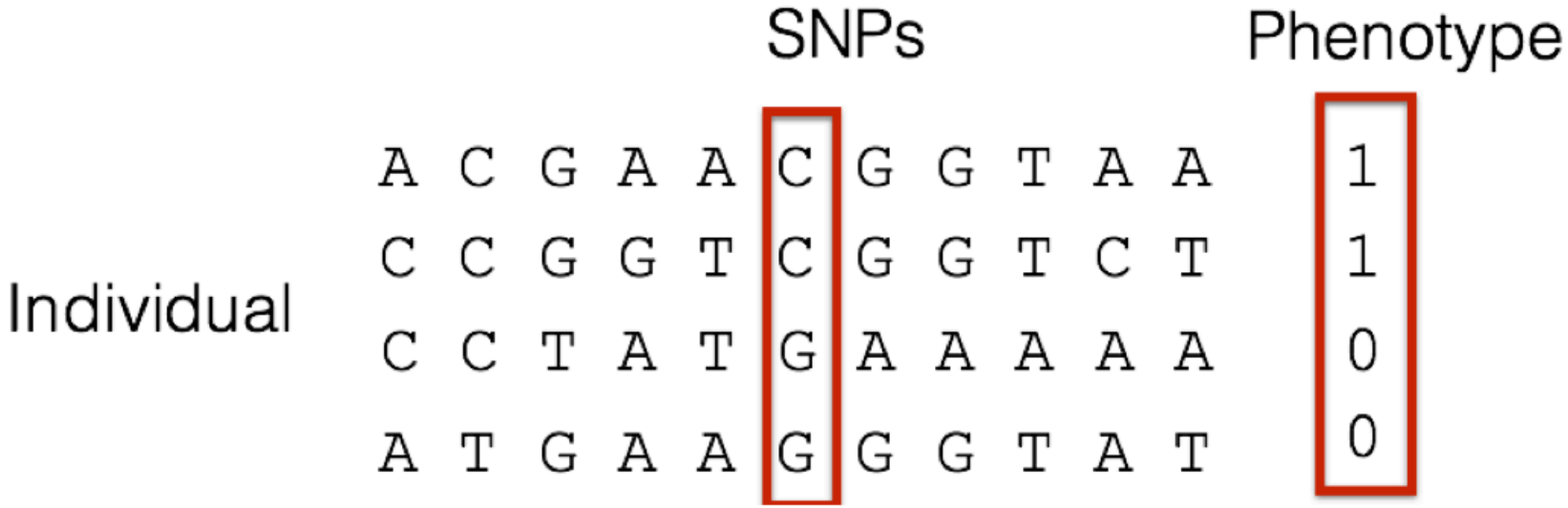
https://en.wikipedia.org/wiki/Cystic_fibrosis_transmembrane_conductance_regulator

"Unrelated" population of individuals measured for a phenotype

Individual	SNPs	Phenotype
	A C G A A C G G T A A	1
	C C G G T C G G T C T	1
	C C T A T G A A A A A	0
	A T G A A G G G T A T	0

Population Association studies

- Find an association between a SNP and phenotype
- Really want to find a SNP that is causal.



Testing for association

Individual	SNPs										Phenotype
	A	C	G	A	A	C	G	G	T	A	A	1
	C	C	G	G	T	C	G	G	T	C	T	1
	C	C	T	A	T	G	A	A	A	A	A	0
	A	T	G	A	A	G	G	G	T	A	T	0

R11209026	Allele A	Allele G
Case	22	976
Controls	68	932

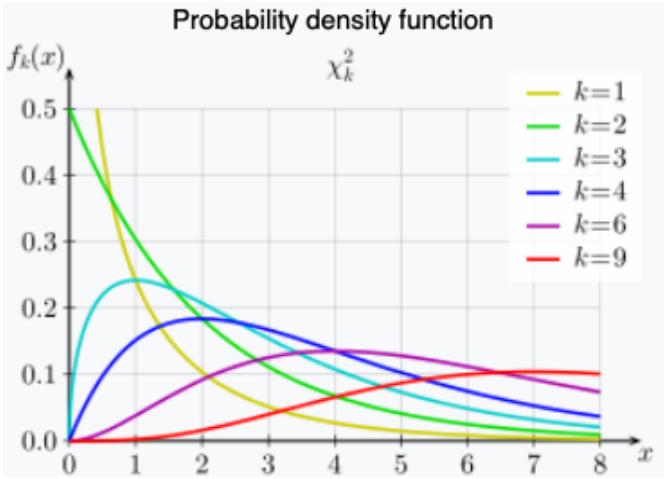
Expected

R11209026	Allele A	Allele G
Case	45	953
Controls	45	955

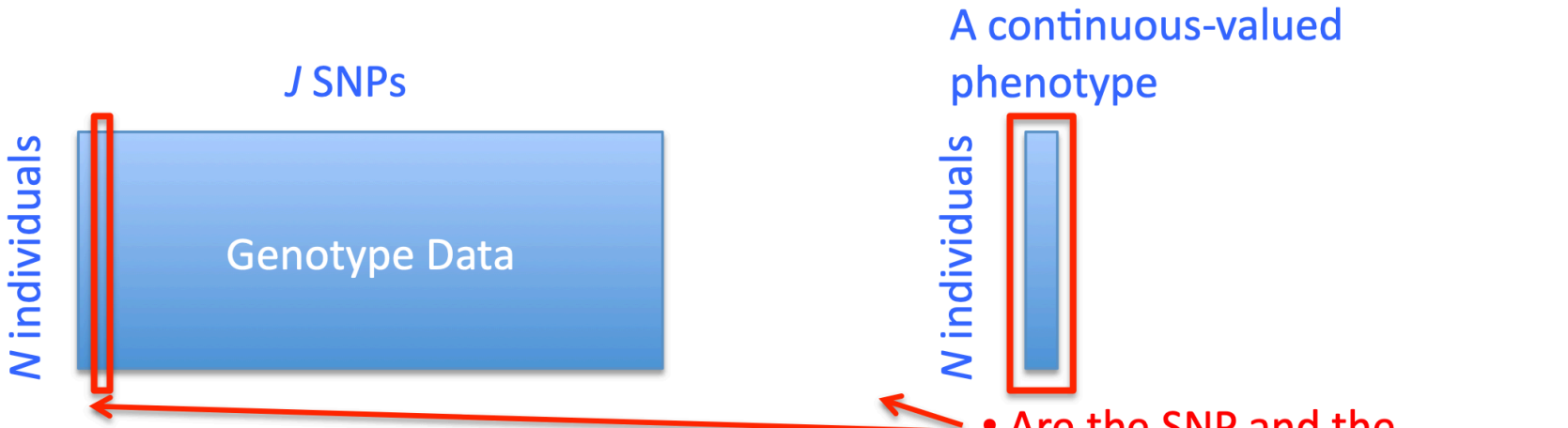
$$\chi^2 = \sum(O - E)^2/E$$

R11209026	Allele A	Allele G
Case	11.72	0.55
Controls	11.70	0.55

- Simplest tests (single marker regression, X^2) rule the data
- Complex multi-marker models are often less reliable



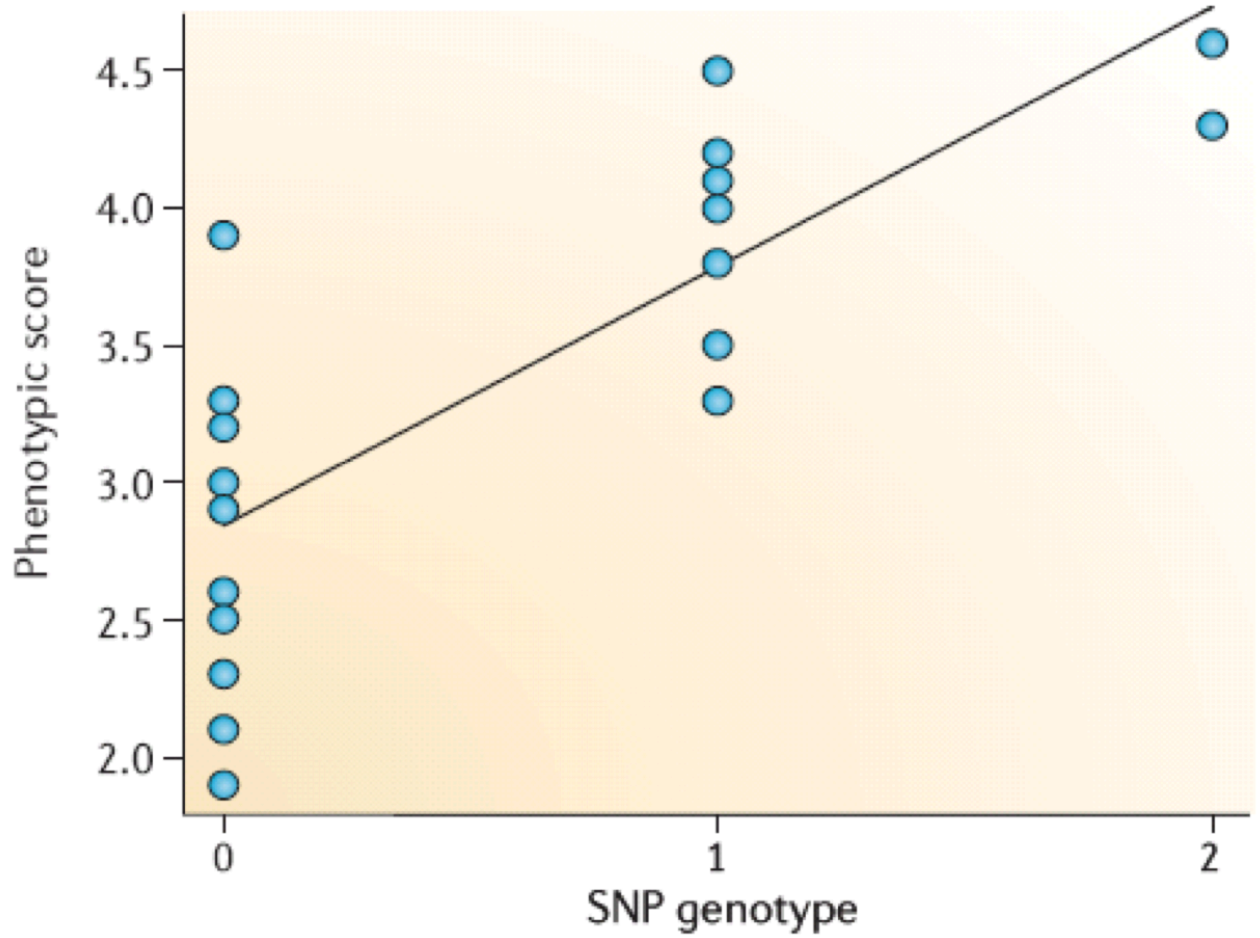
Using linear regression for SNP association testing



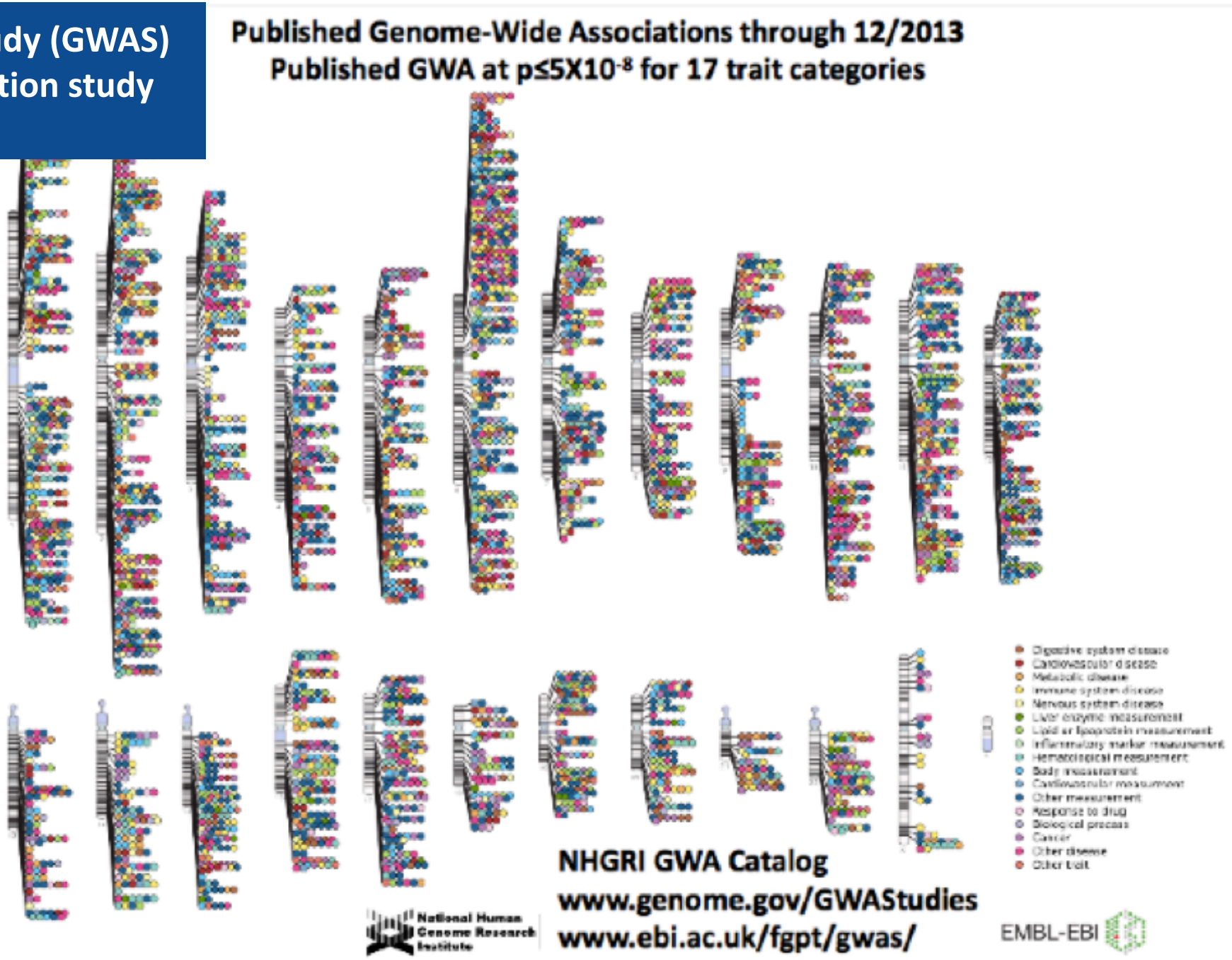
- Are the SNP and the phenotype significantly associated?
- Fit a linear regression model for each (SNP, phenotype) pair

Using linear regression for SNP association testing

- Continuous-valued traits
  - Also called quantitative traits
  - Cholesterol level, blood pressure etc.
- One cannot create a contingency matrix as in case/ control studies
- For each locus, fit a linear regression using the number of minor alleles at the given locus of the individual as covariate



Genome-wide association study (GWAS)
Perform a population association study across the genome

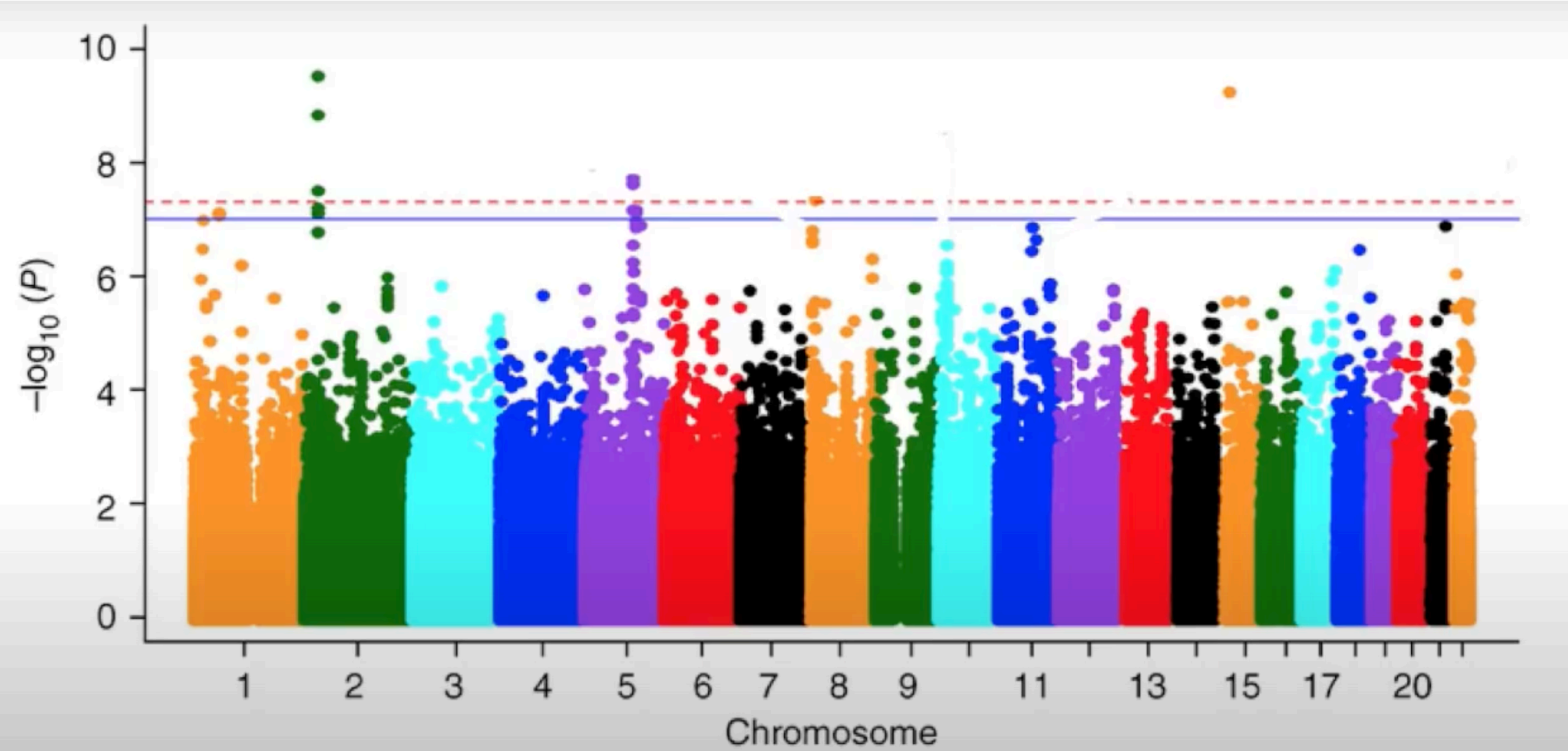


Multiple testing

- In GWAS, we're performing $O(10^6)$ tests that are largely independent
  - Each study has hundreds of $p<.001$ purely by statistical chance (no real relationship to disease)
  - "Genome-wide significance" often set at $p=5x10^{-8}$ (= .05 / 1 million tests)

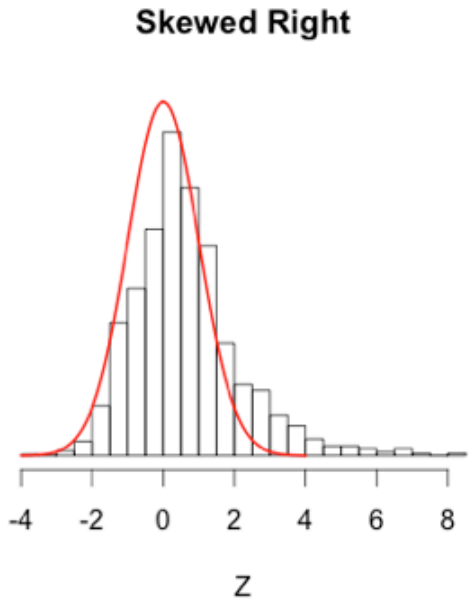
1996: Risch and Merikangas propose that a **p**-value of **5×10^{-8}** (equivalent to a **p**-value of 0.05 after a Bonferroni correction for 1 million independent tests) is a conservative threshold for declaring significant association in a genome-wide study.

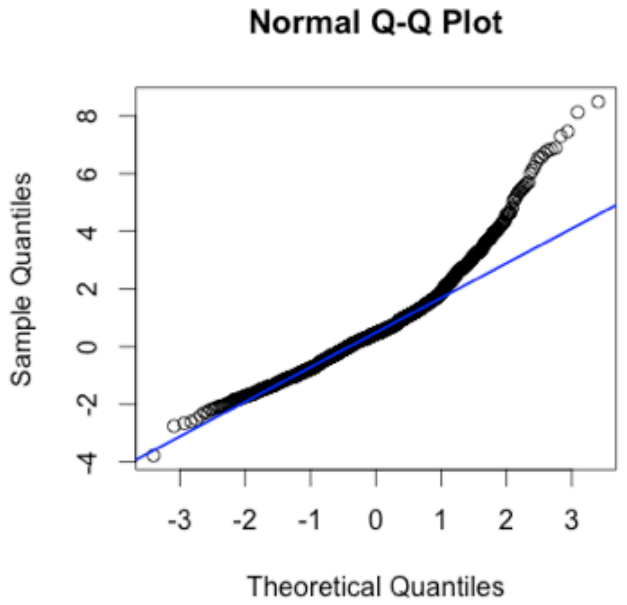
Understanding Manhattan Plots

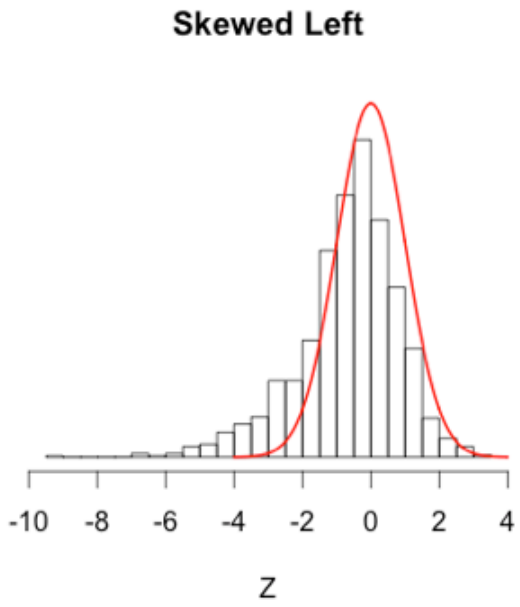


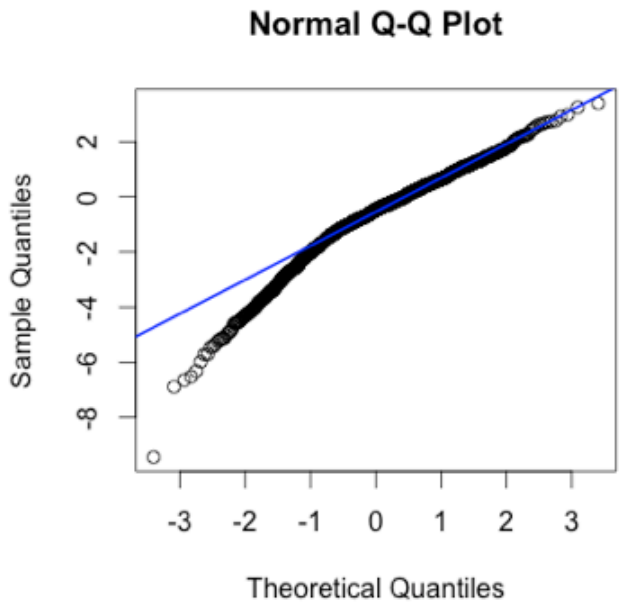
Kottyan et al. 2014 Nature Genetics 46: 895

Understanding Q-Q Plots



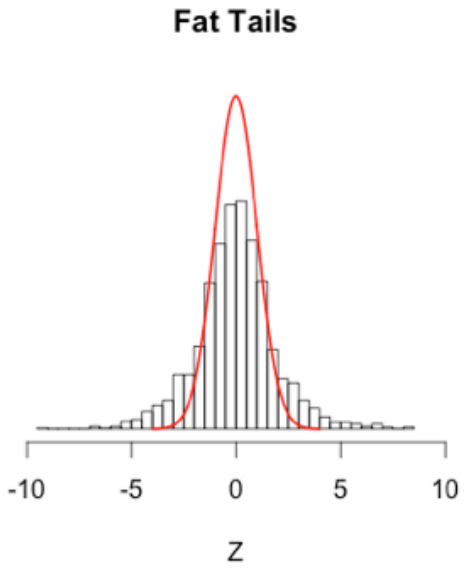


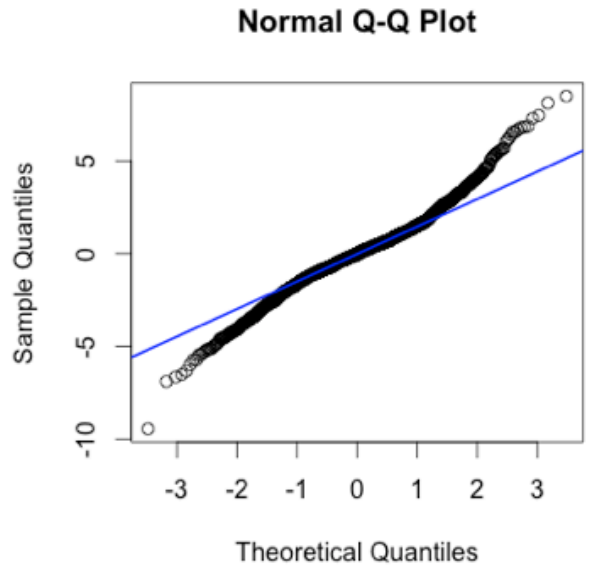


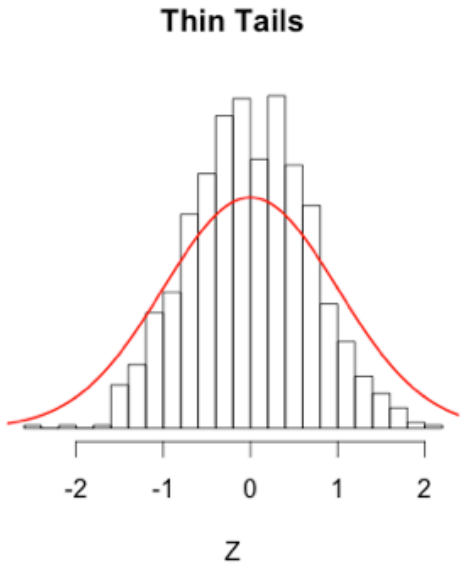


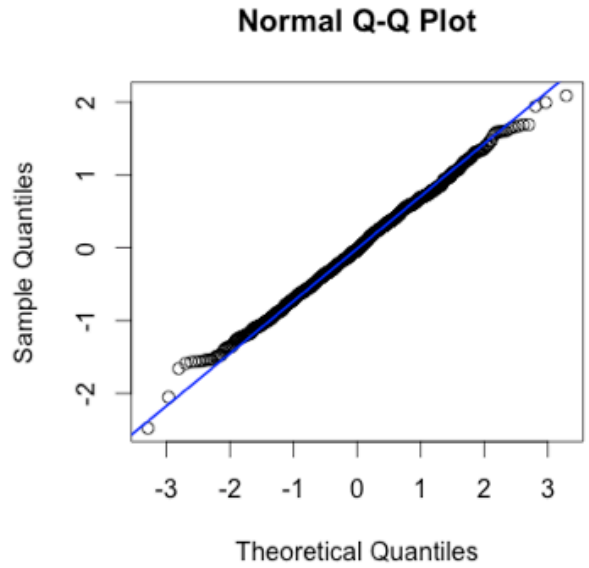
https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

Understanding Q-Q Plots





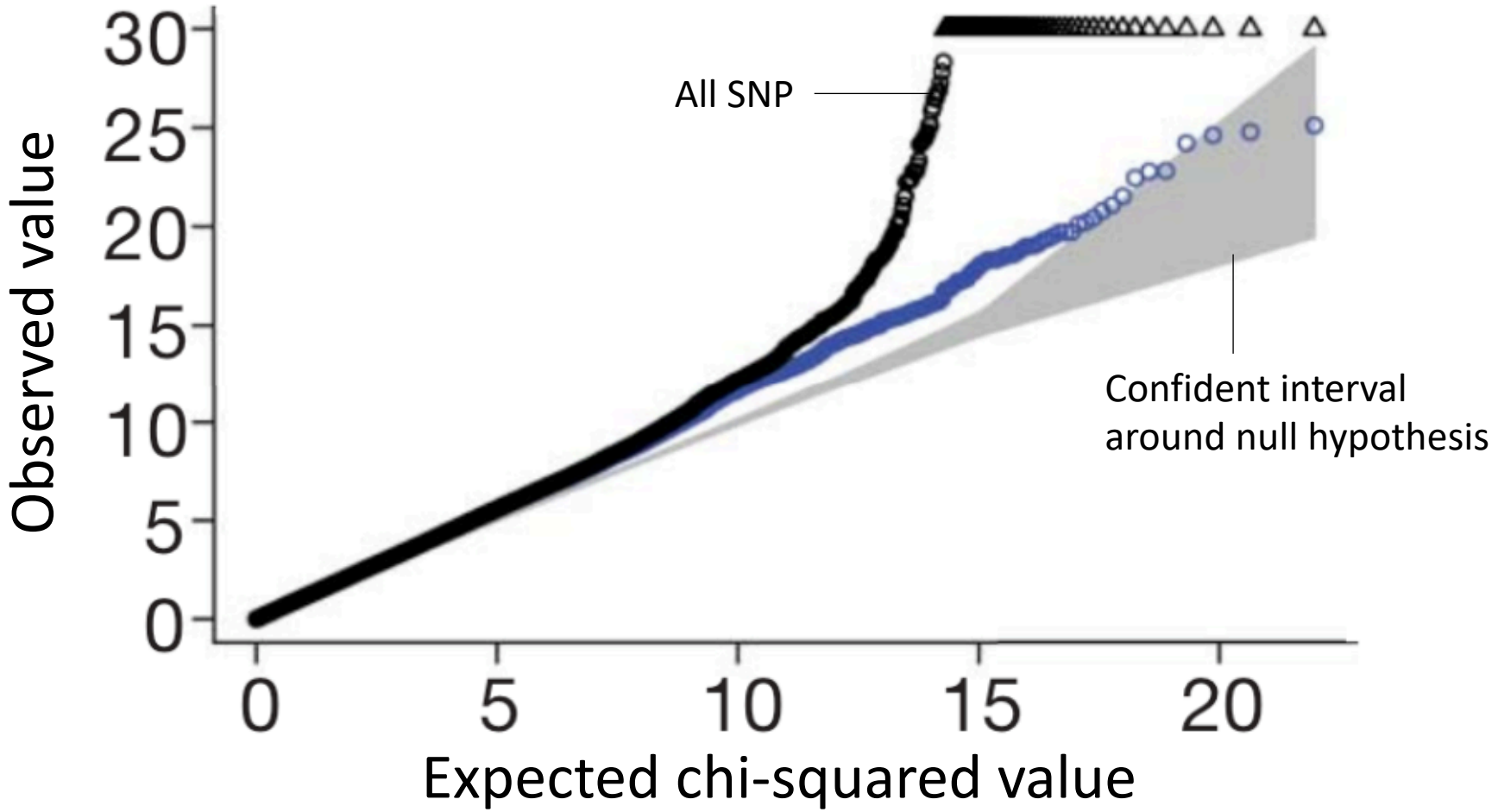




https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

Understanding Q-Q Plots

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.



Genomics

Genomics is the most mature of the omics fields. In the realm of medical research, genomics focuses on identifying genetic variants associated with disease, response to treatment, or future patient prognosis.

GWAS is a successful approach that has been used to identify thousands of genetic variants associated with complex diseases (GWAS catalog https://www.ebi.ac.uk/gwas/home) in multiple human populations. In such studies, thousands of individuals are genotyped for more than a million genetic markers, and statistically significant differences in minor allele frequencies between cases and controls are considered evidence of association.

GWAS studies provide an invaluable contribution to our understanding of complex phenotypes. Associated technologies include genotype arrays, NGS for whole-genome sequencing, and exome sequencing

Rationale for the development of systems biology technologies that involve the integration of different omics data types to identify molecular patterns associated with disease

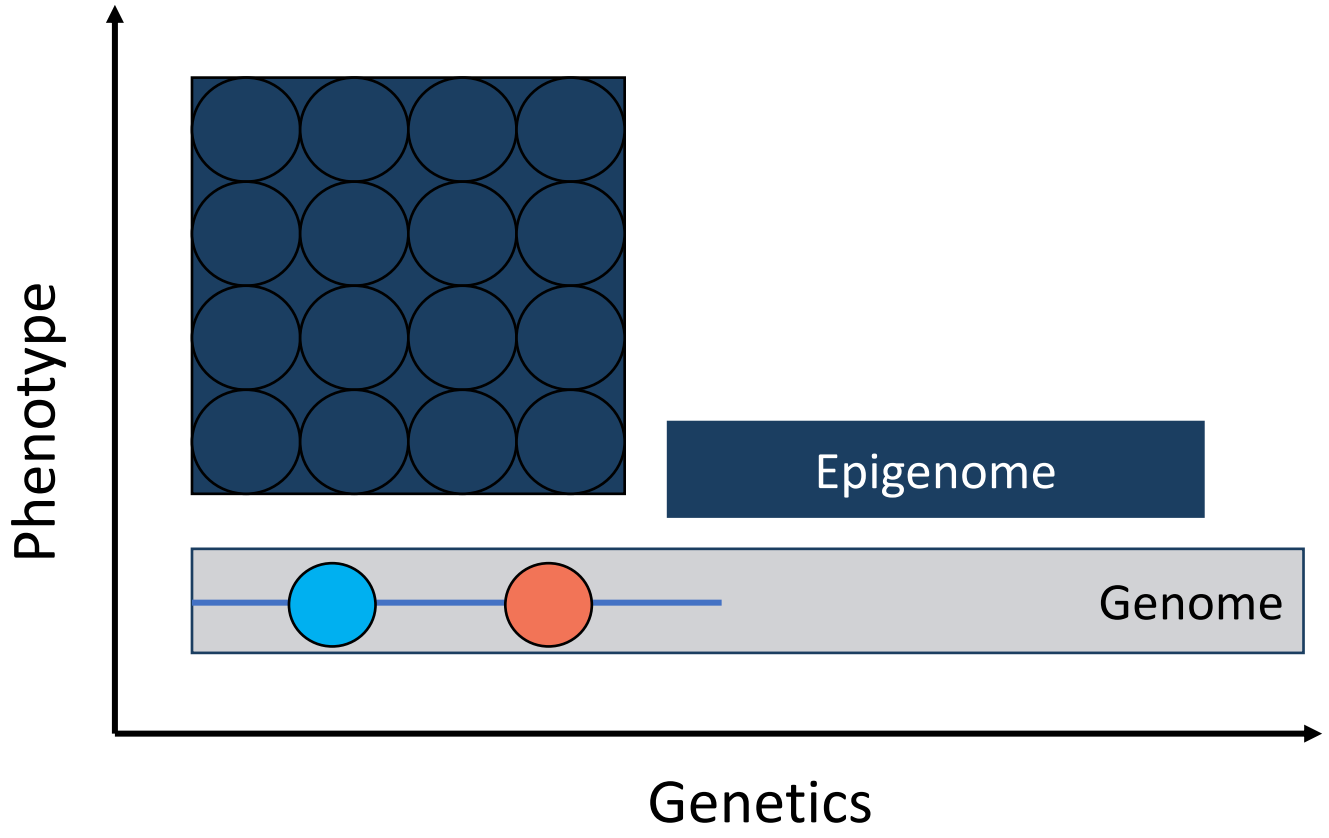
However, as our power to identify genetic variants associated with complex disease increased several realizations were reached that have shaped subsequent approaches to elucidating the causes of disease.

First, the loci that have been identified so far generally explain only a fraction of the heritable component for specific diseases.

Second, while Mendelian diseases generally result from changes in coding regions of genes, common diseases usually result from changes in gene regulation.

Third, the same genetic variants often contribute to different final outcomes, depending on the environment and genetic background.

Epigenomics



ORF and CDS

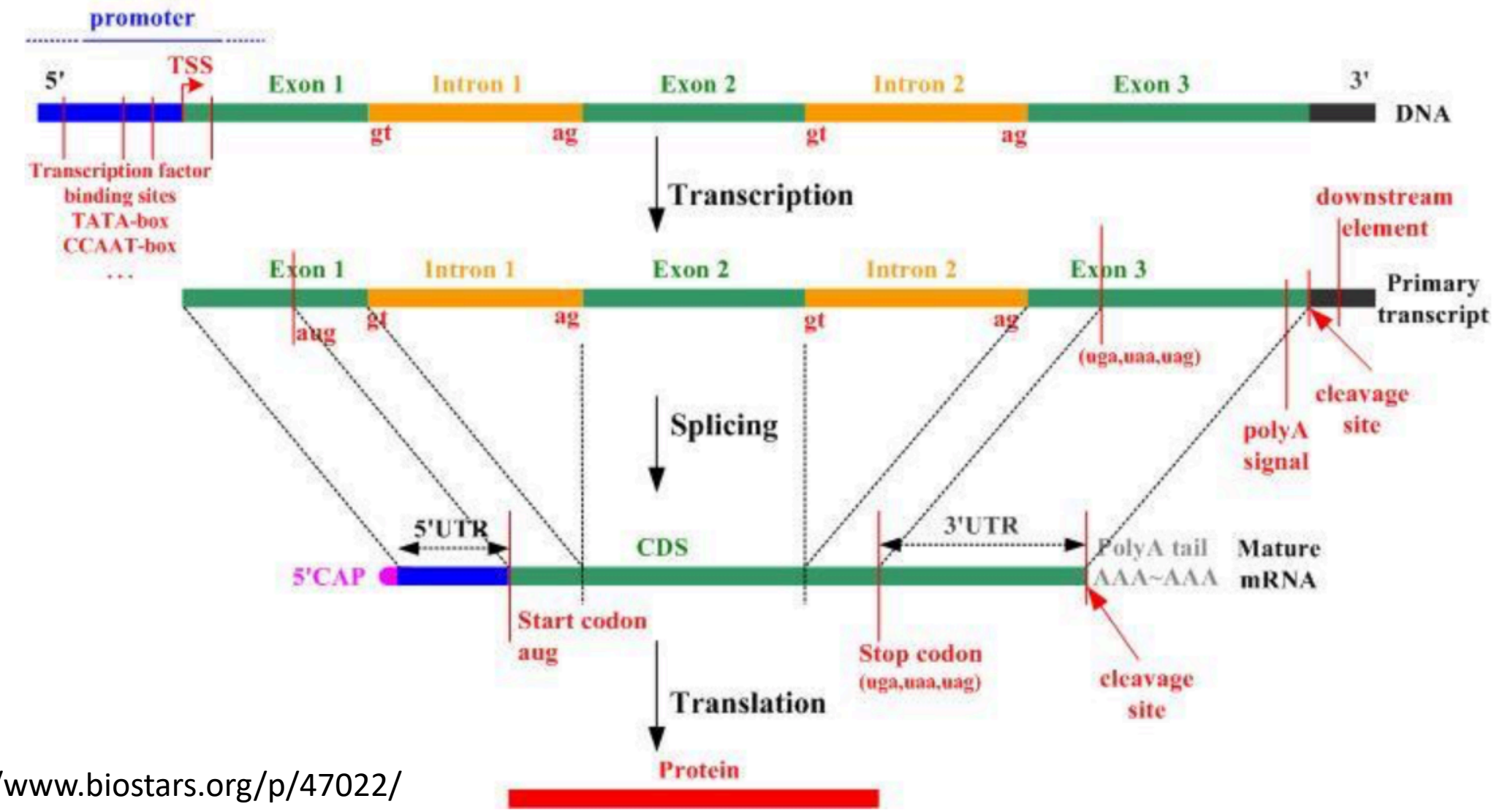
The region of the nucleotide sequences from the start codon (ATG) to the stop codon is called the **Open Reading frame**.

The **Coding Sequence (CDS)** is the actual region of DNA that is translated to form proteins. While the ORF may contain introns as well, the CDS refers to those nucleotides(concatenated exons) that can be divided into codons which are actually translated into amino acids by the ribosomal translation machinery.

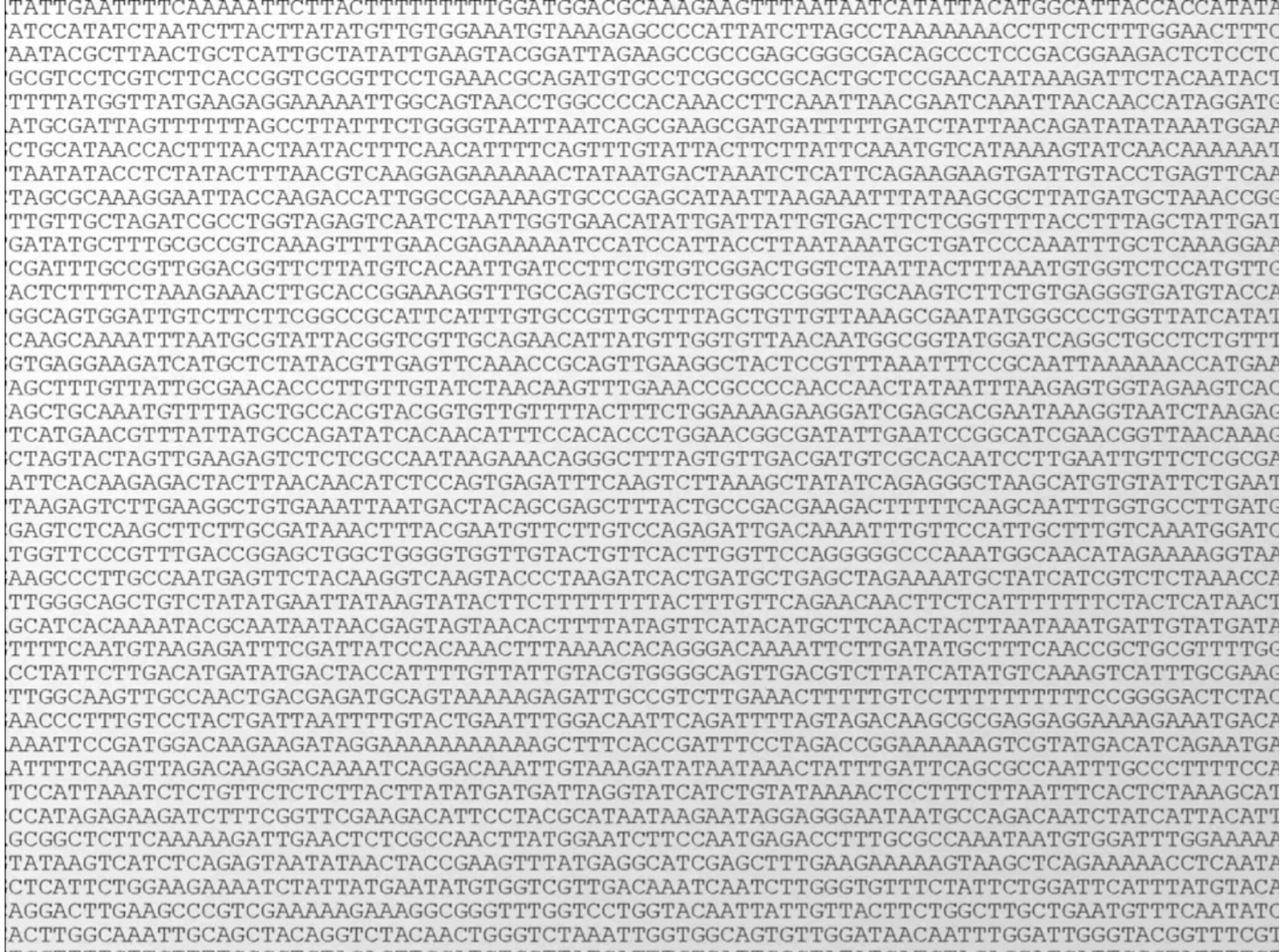
Mainly: CDS means only that the sequence is known to be transcribed and, therefore, it is coding for something -- neither gene nor protein has to be known. Any full mRNA sequence (obtained from cDNA sequencing) will have a full coding sequence. ORF is usually predicted based on DNA sequence and not proven to be transcribed.

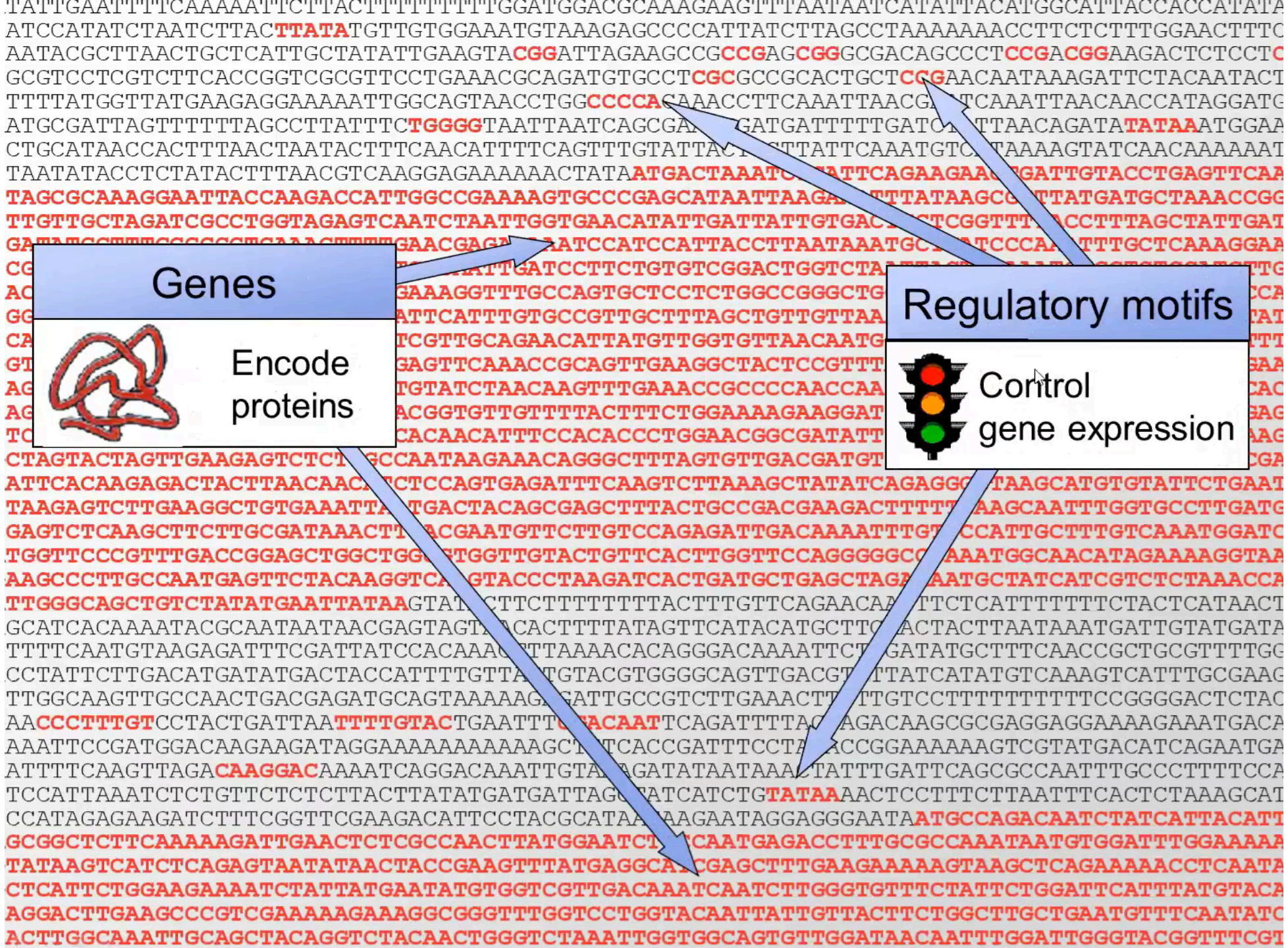
https://www.biostars.org/p/47022/

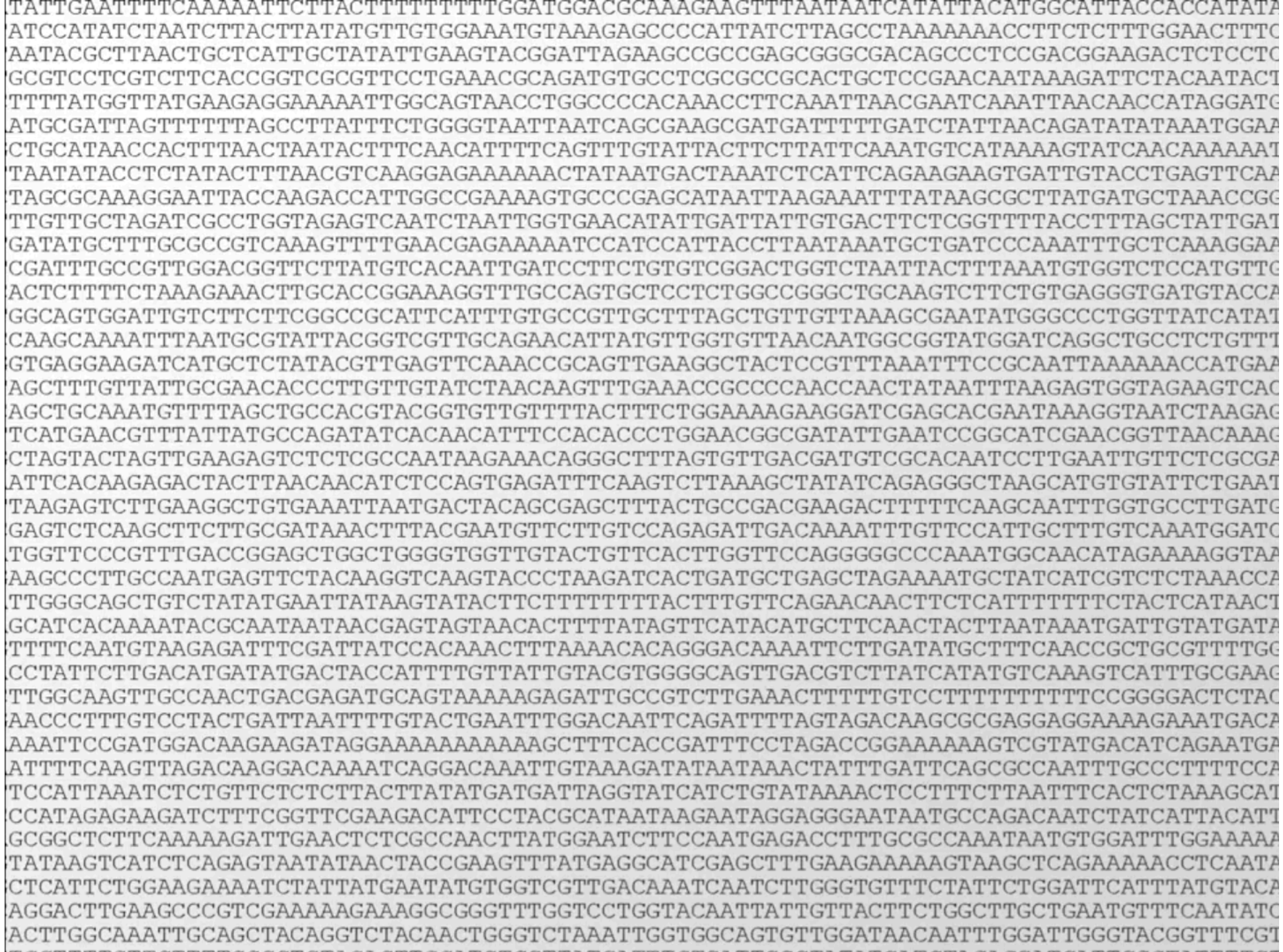
ORF and CDS

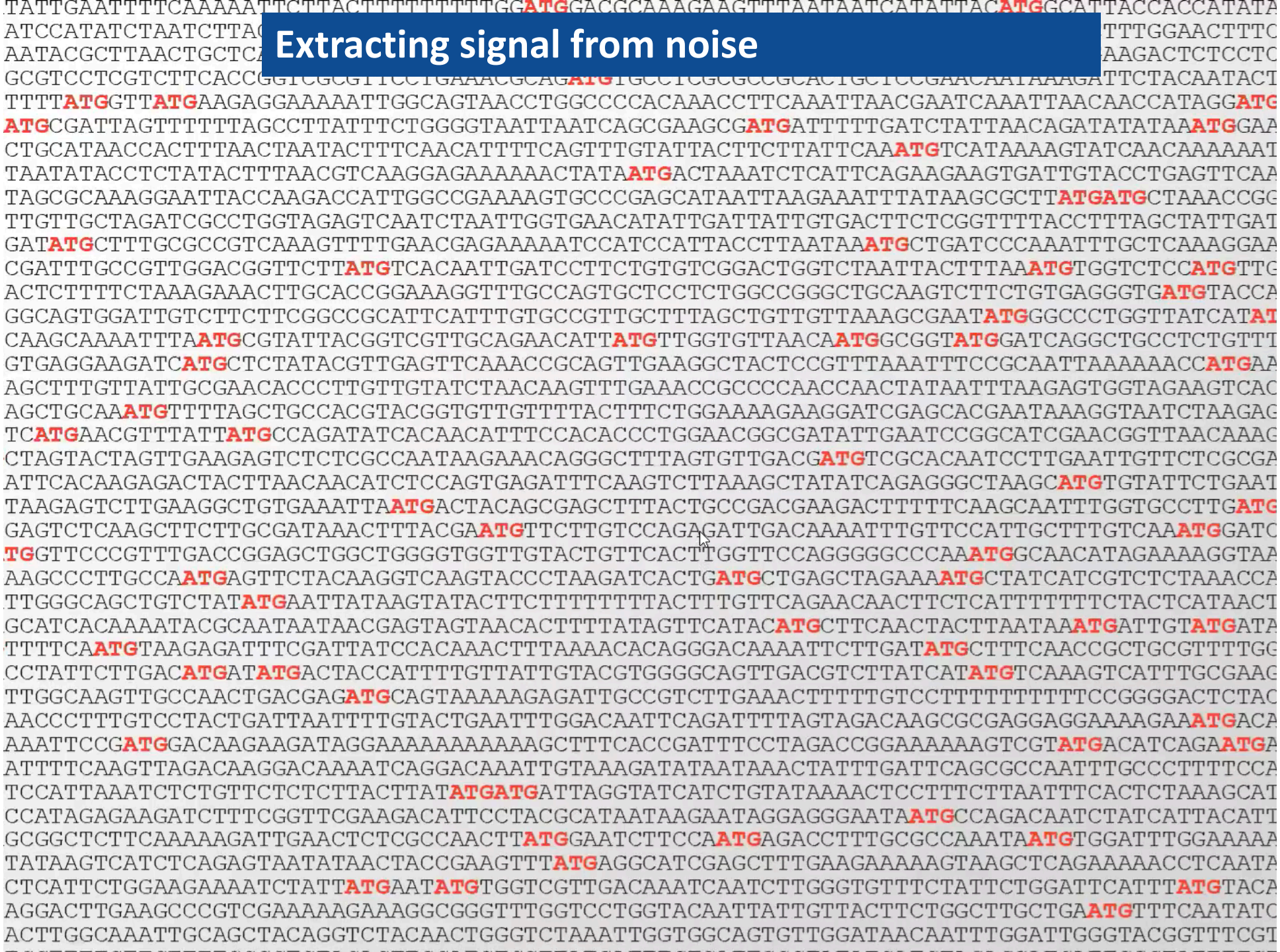


https://www.biostars.org/p/47022/

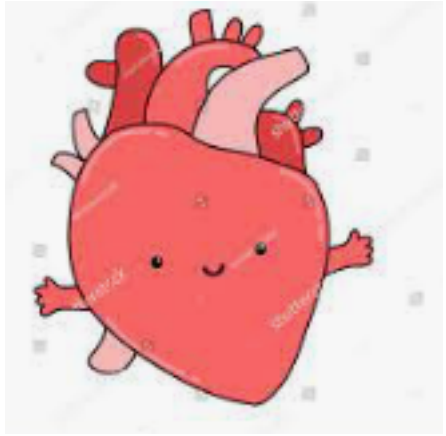


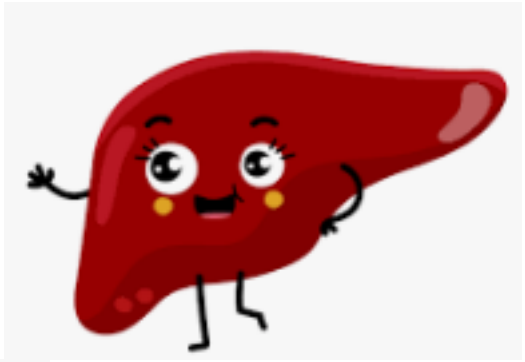

Genes
Encode proteins
Regulatory motifs
Control gene expression

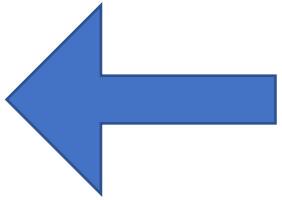


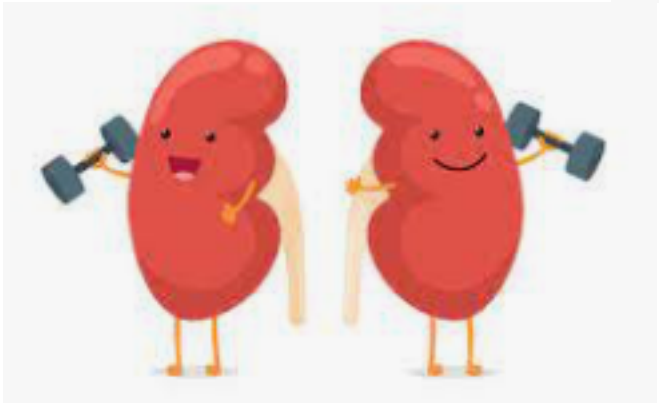

Extracting signal from noise

One genome – Many cells



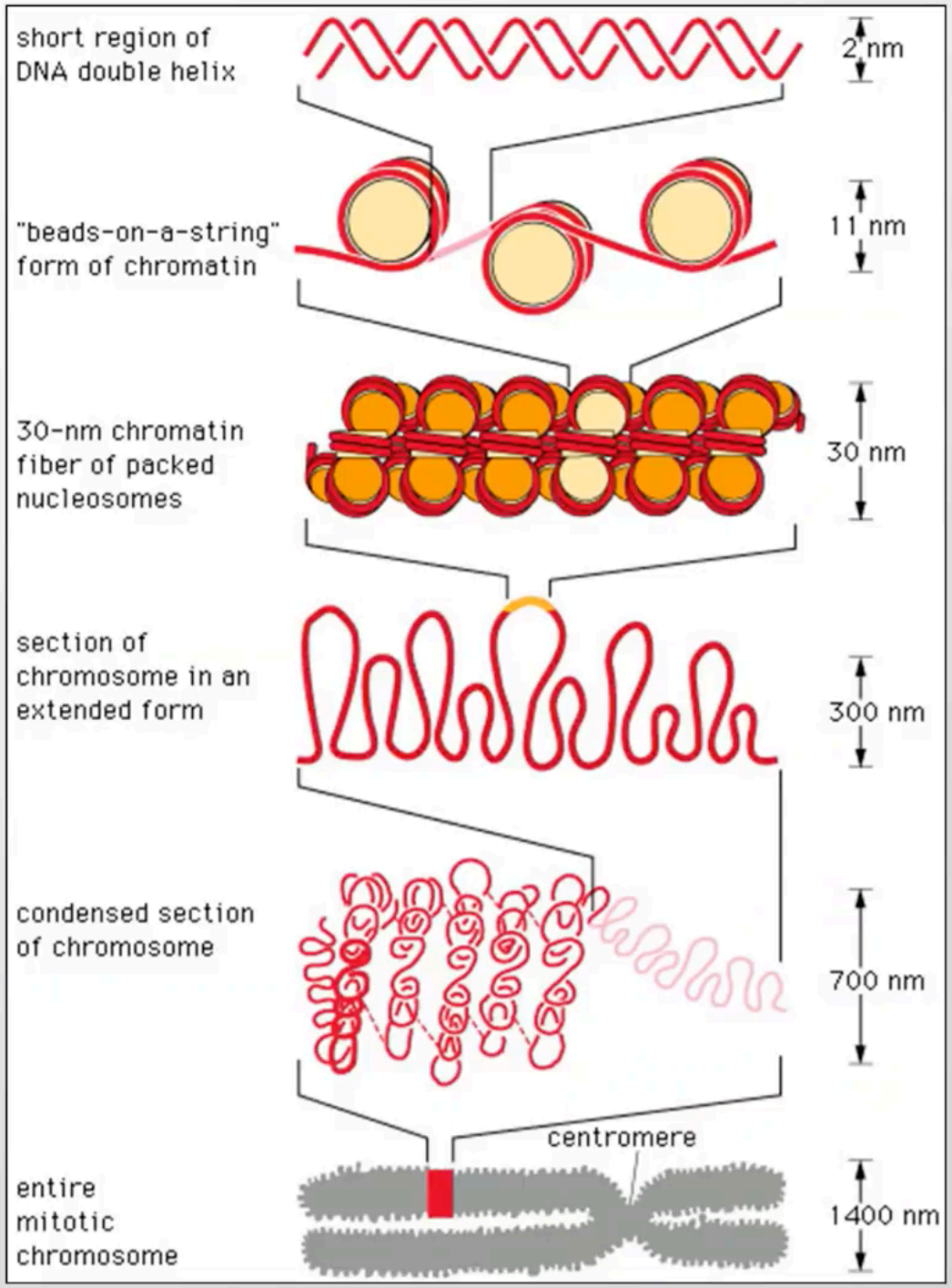






ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA

DNA-packing



Why packing

--DNA is very long

stretch out the entire human genome: 300 x distance between earth and sun

--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 compart structure must open locally

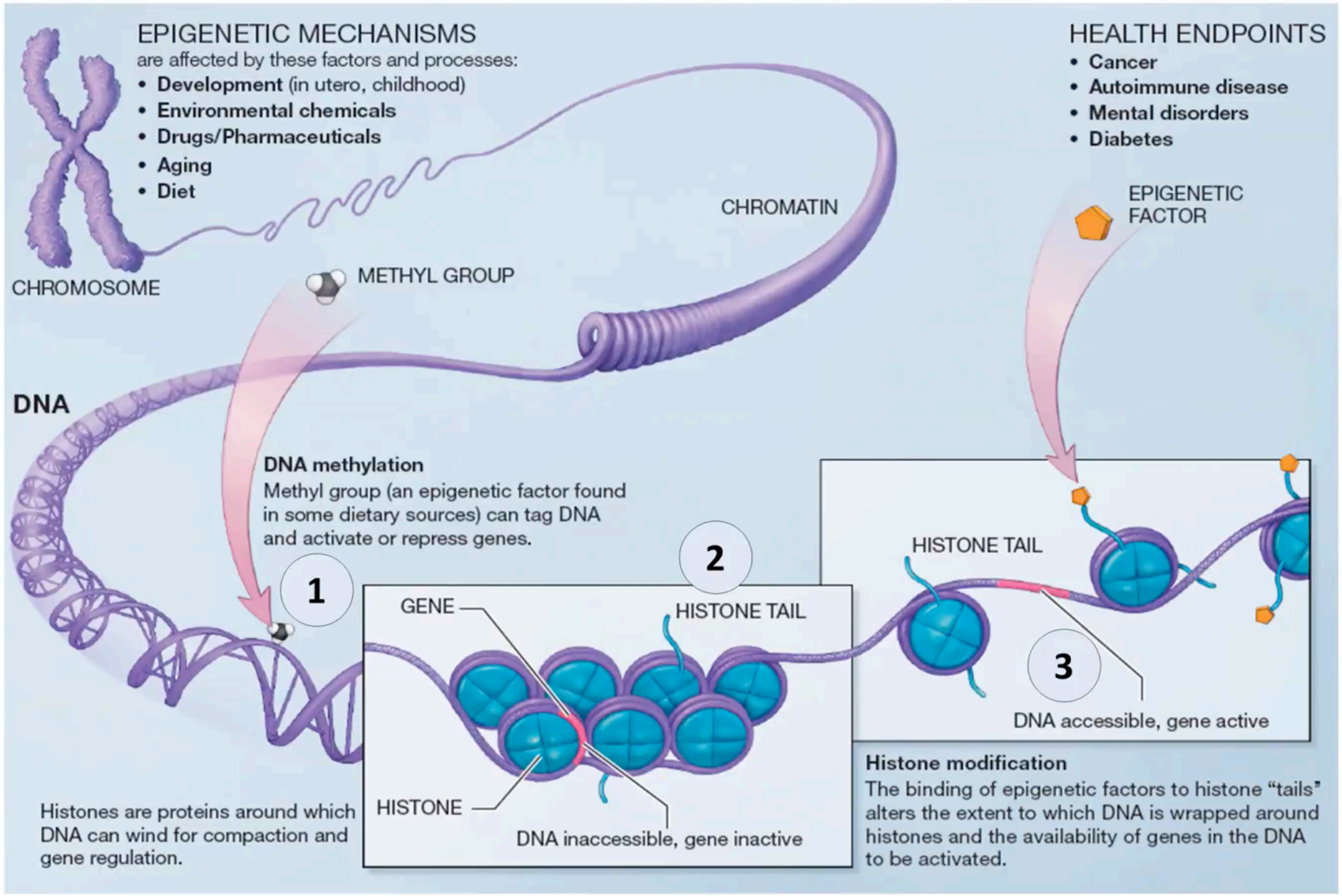
Now emerging:

--Role of accessibility

--State in chromatin itself

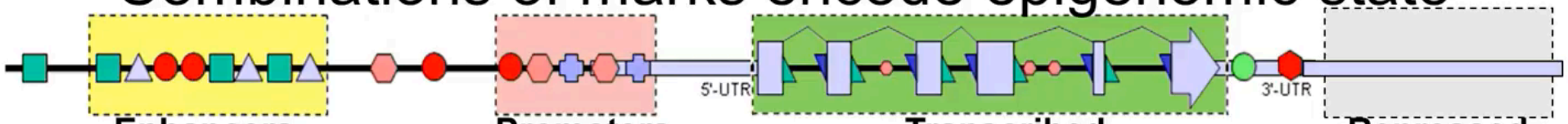
--Role of 3D interaction

Three types of epigenetic modifications

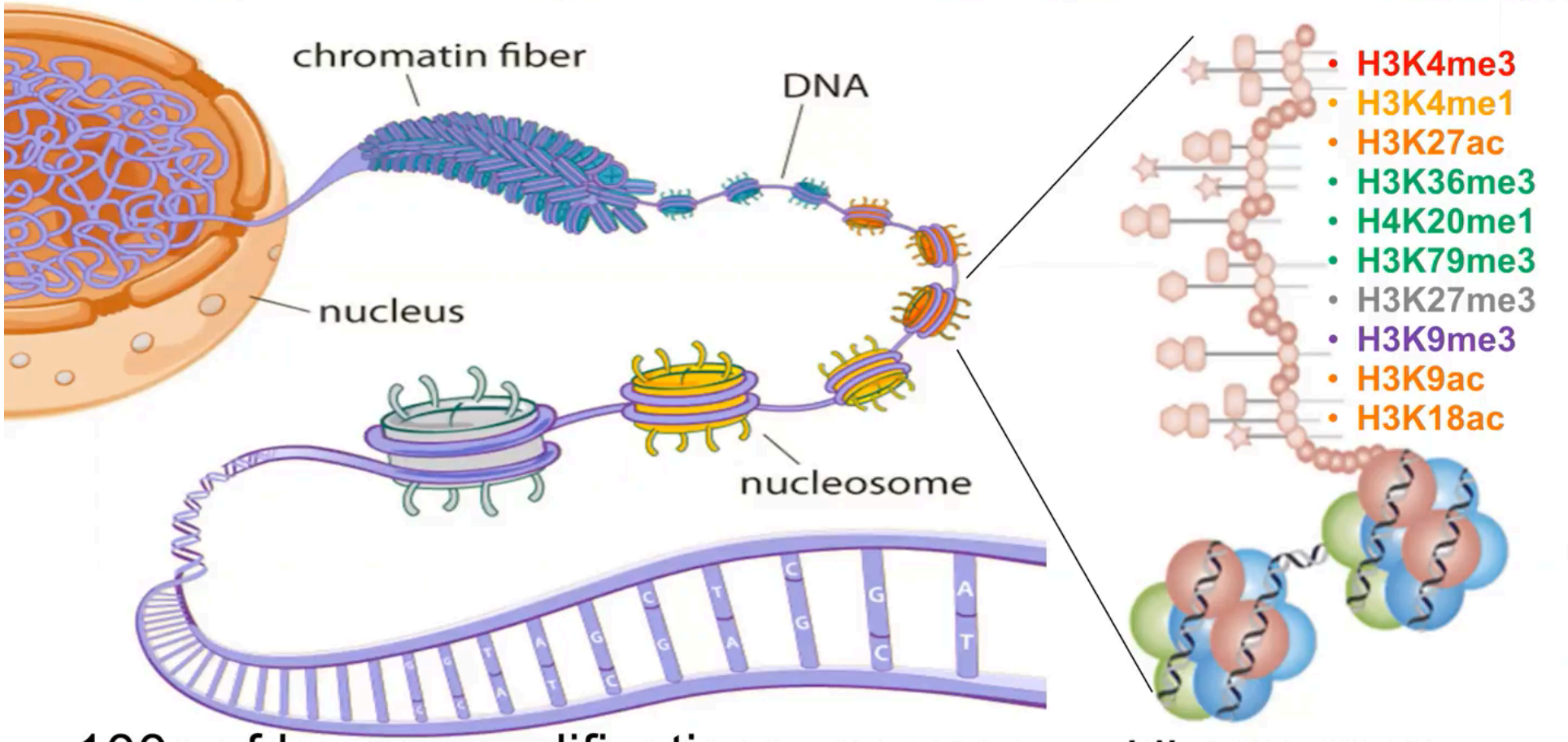


https://en.wikipedia.org/wiki/Epigenetics

Combinations of marks encode epigenomic state

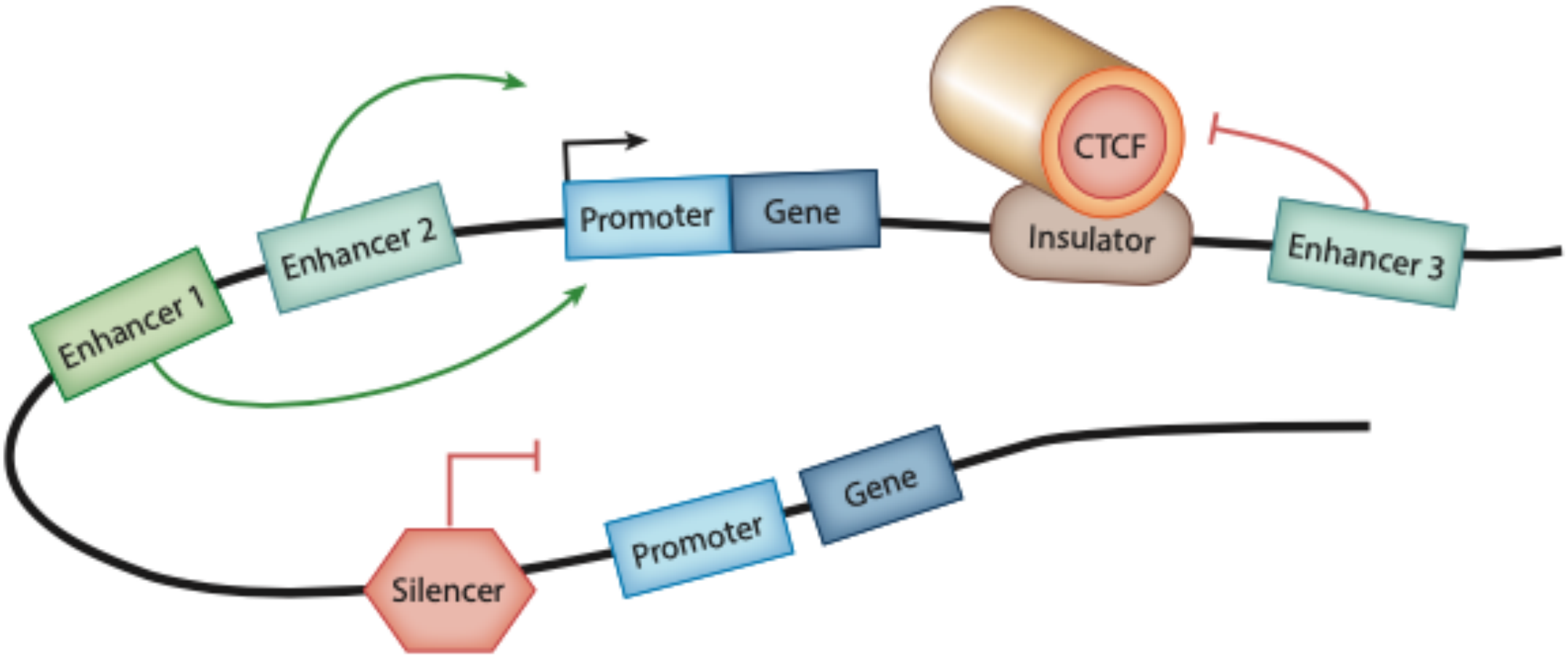


Enhancers	Promoters	Transcribed	Repressed
• H3K4me1	• H3K4me3	• H3K36me3	• H3K9me3
• H3K27ac	• H3K9ac	• H3K79me2	• H3K27me3
• DNase	• DNase	• H4K20me1	• DNAmethyl

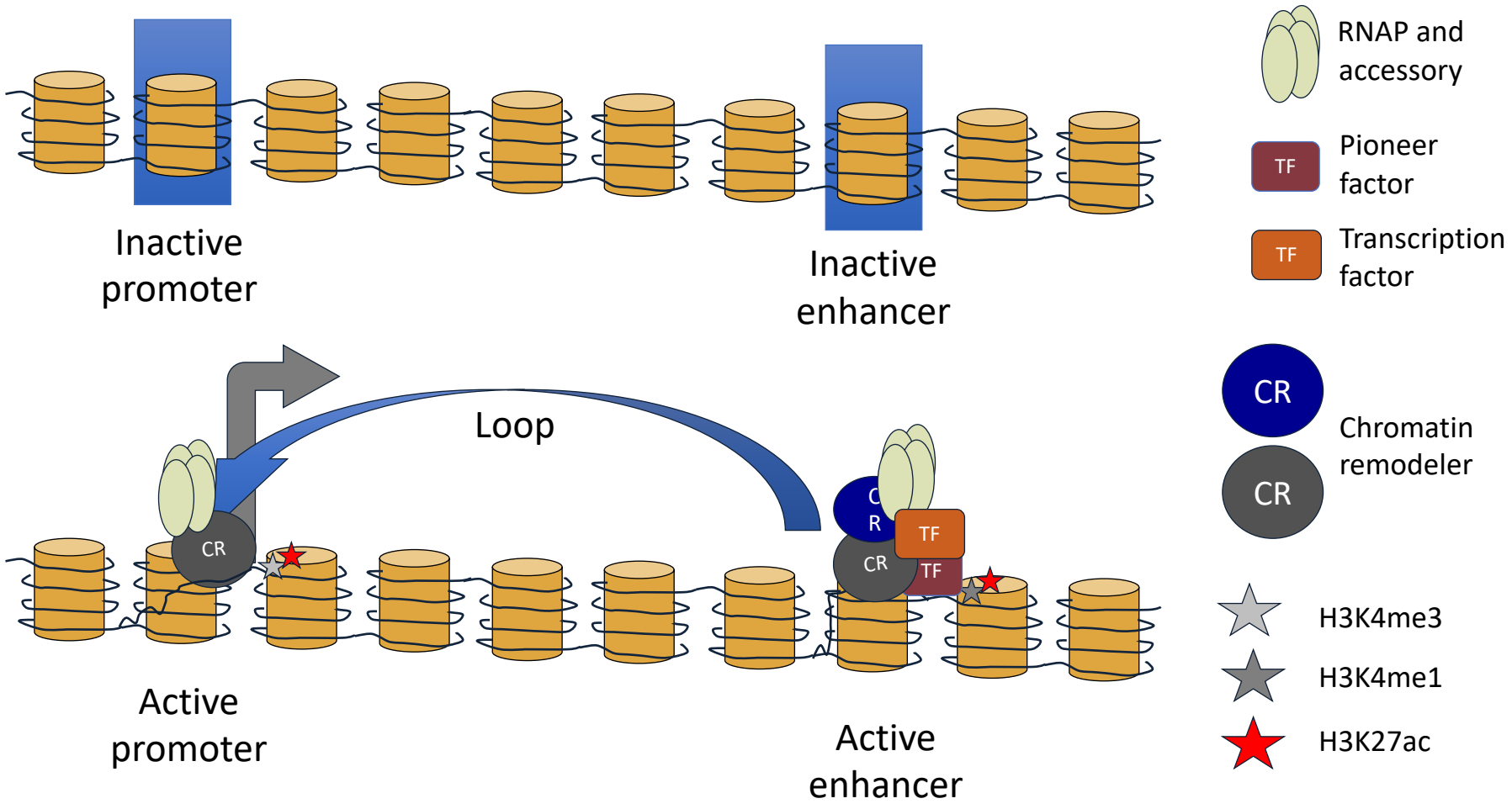


- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Types of gene regulatory elements



A model for step-wise activation of enhancers



Reviewed by Zaret & Carroll, Gene & Dev 2011;

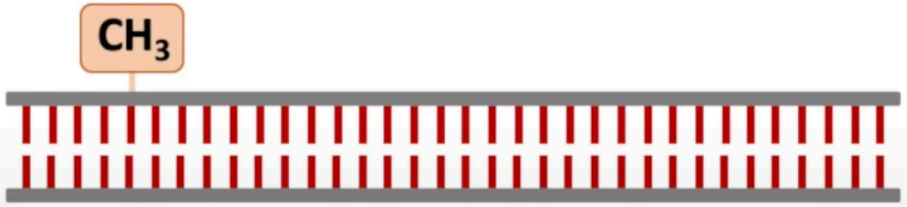
DNA Methylation

Genetics

Alteration in DNA sequence

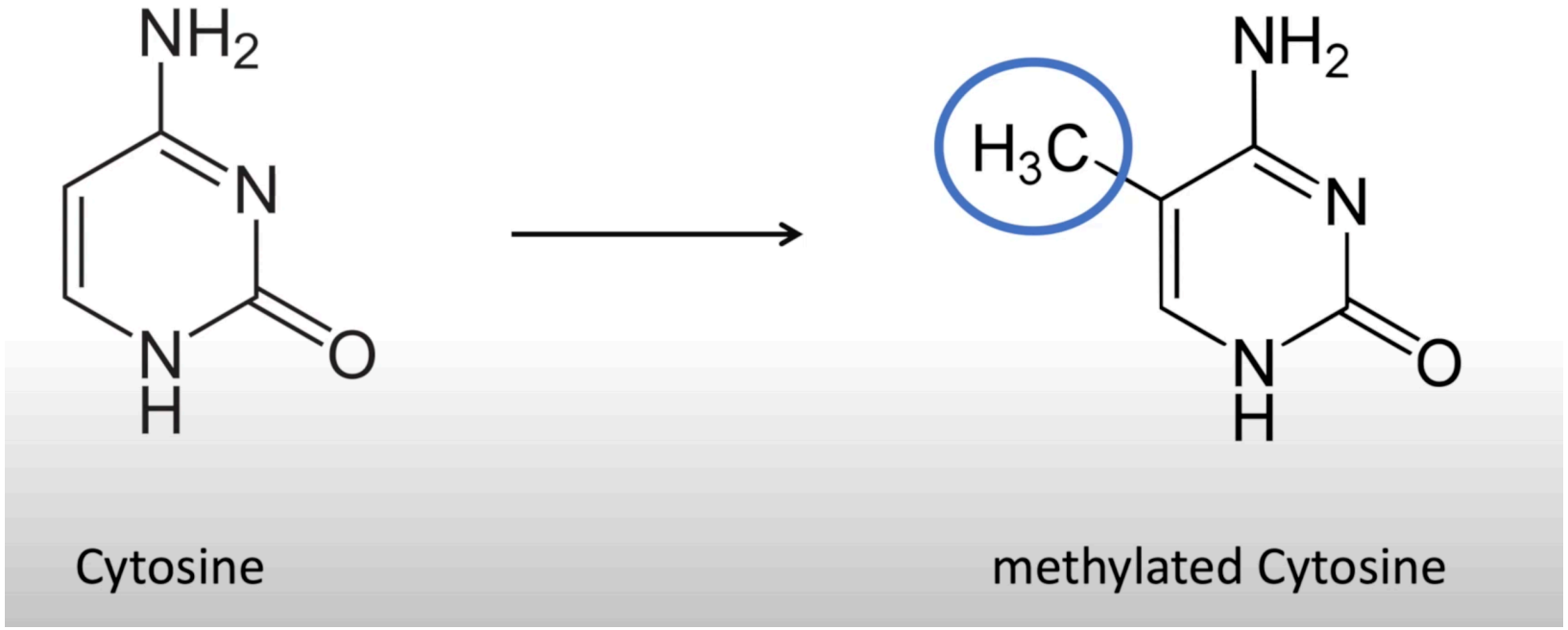
Changes in Phenotype

Epigenetics

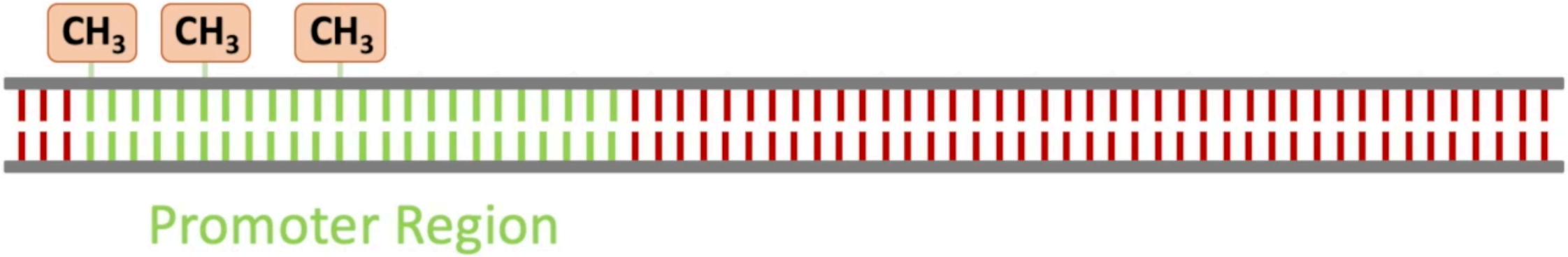


NO Alteration in DNA sequence

DNA Methylation

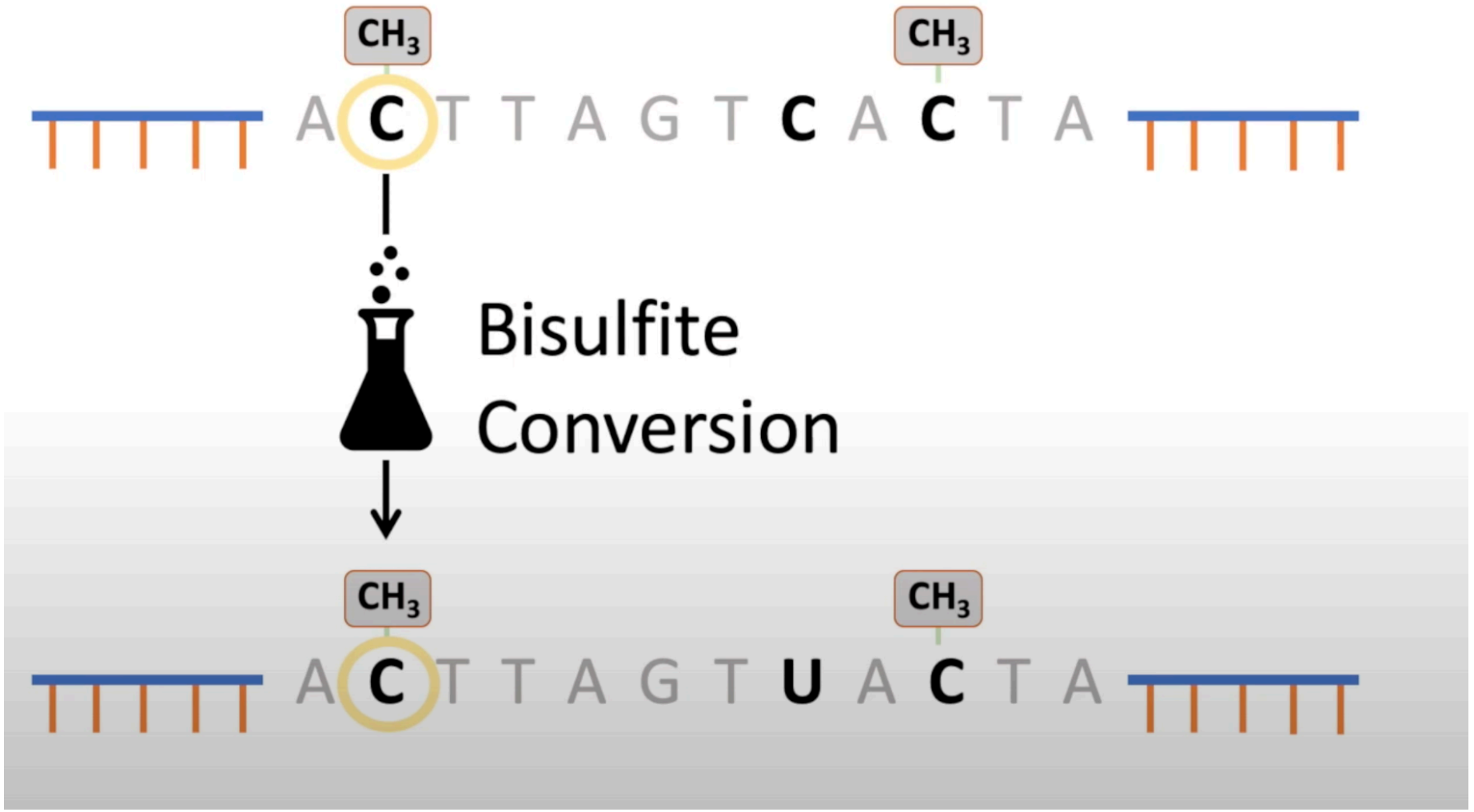


Function of DNA methylation

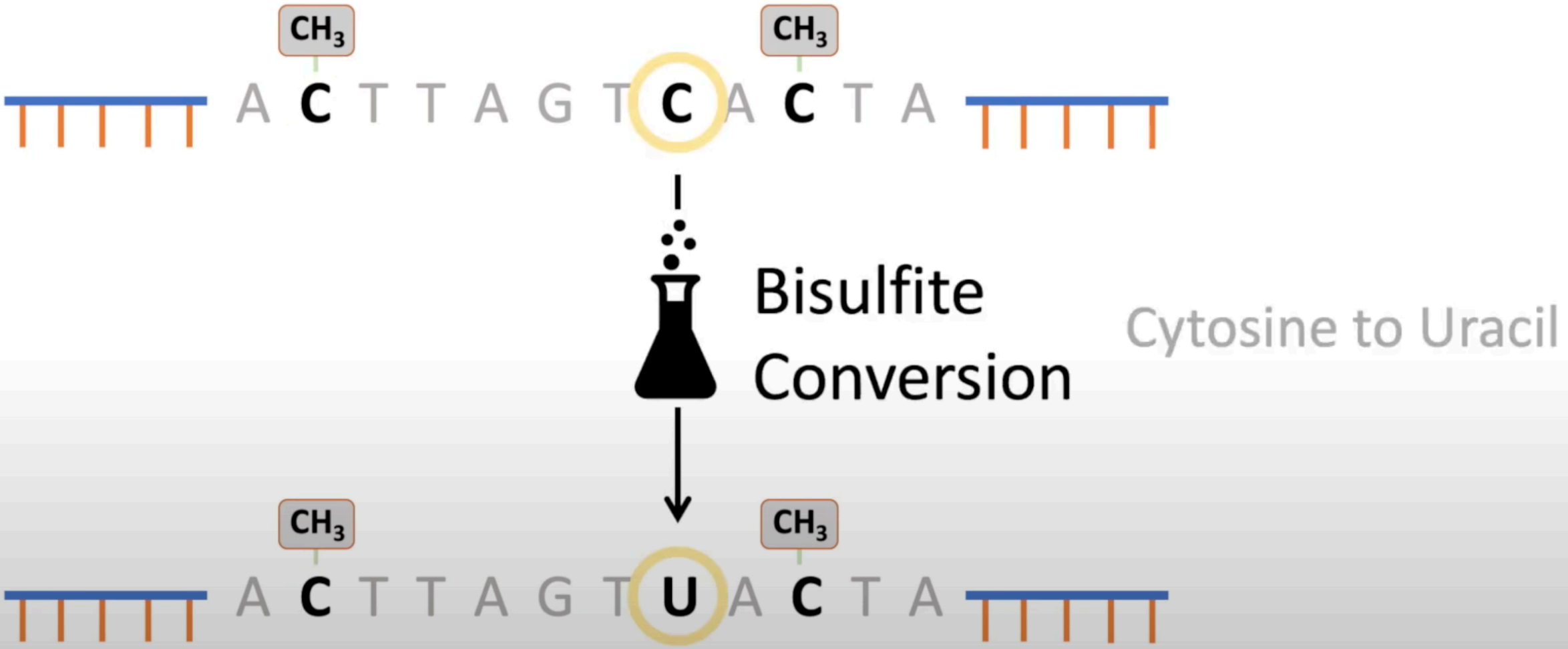


DNA Methylation → **In**activation of Genes

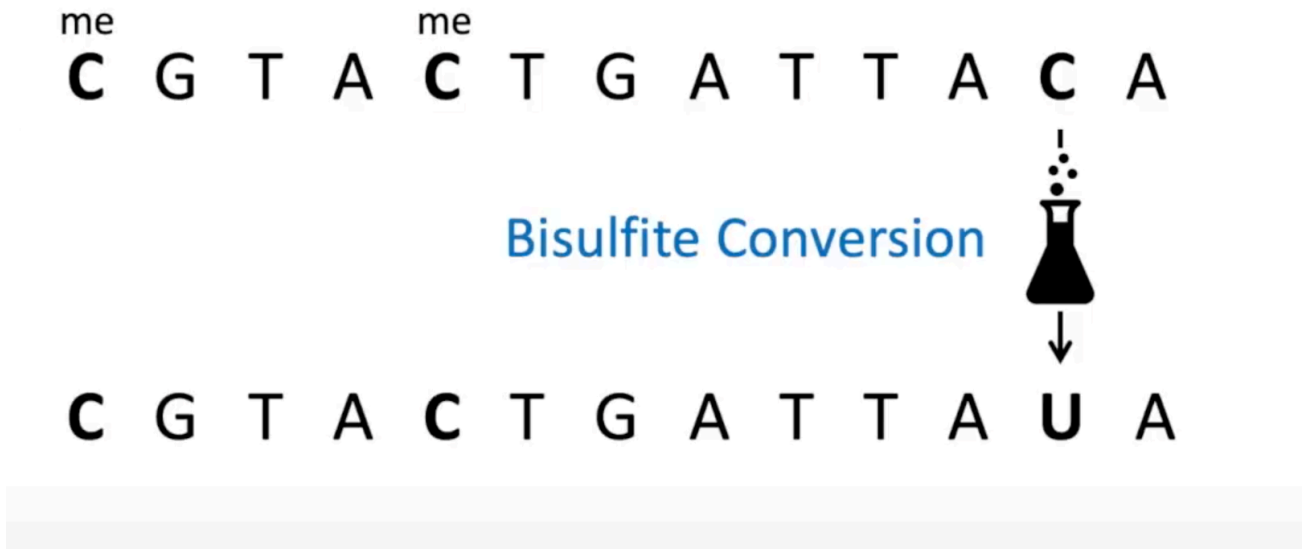
Bisulfite Conversion



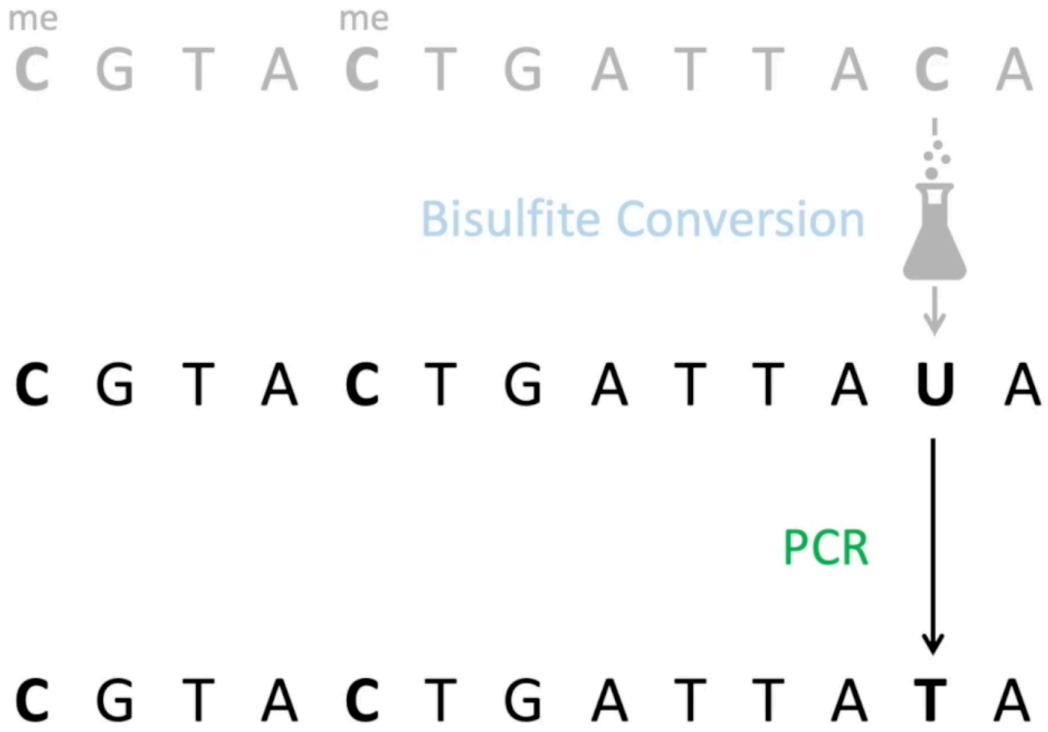
Bisulfite Conversion



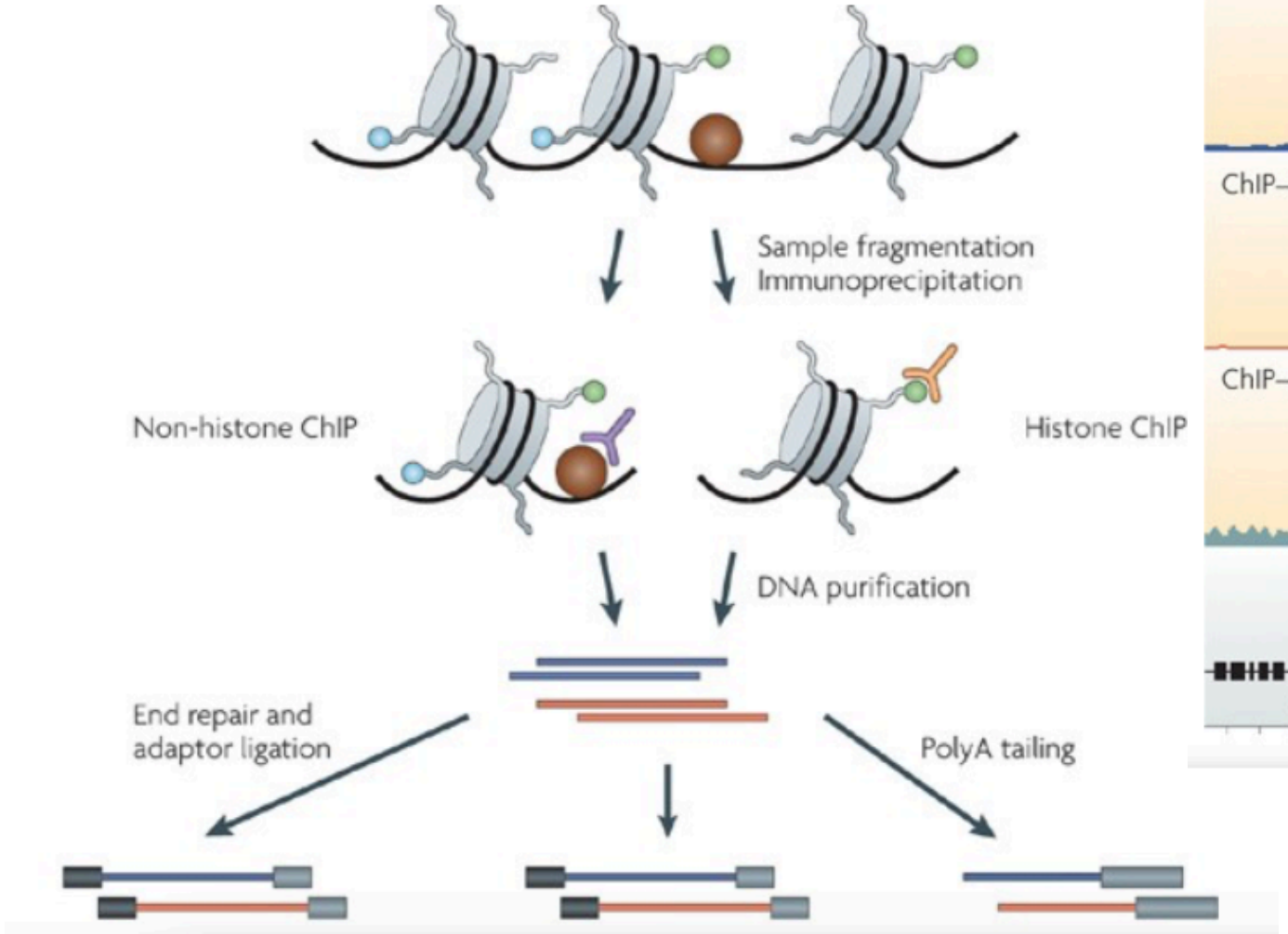
Bisulfite Conversion

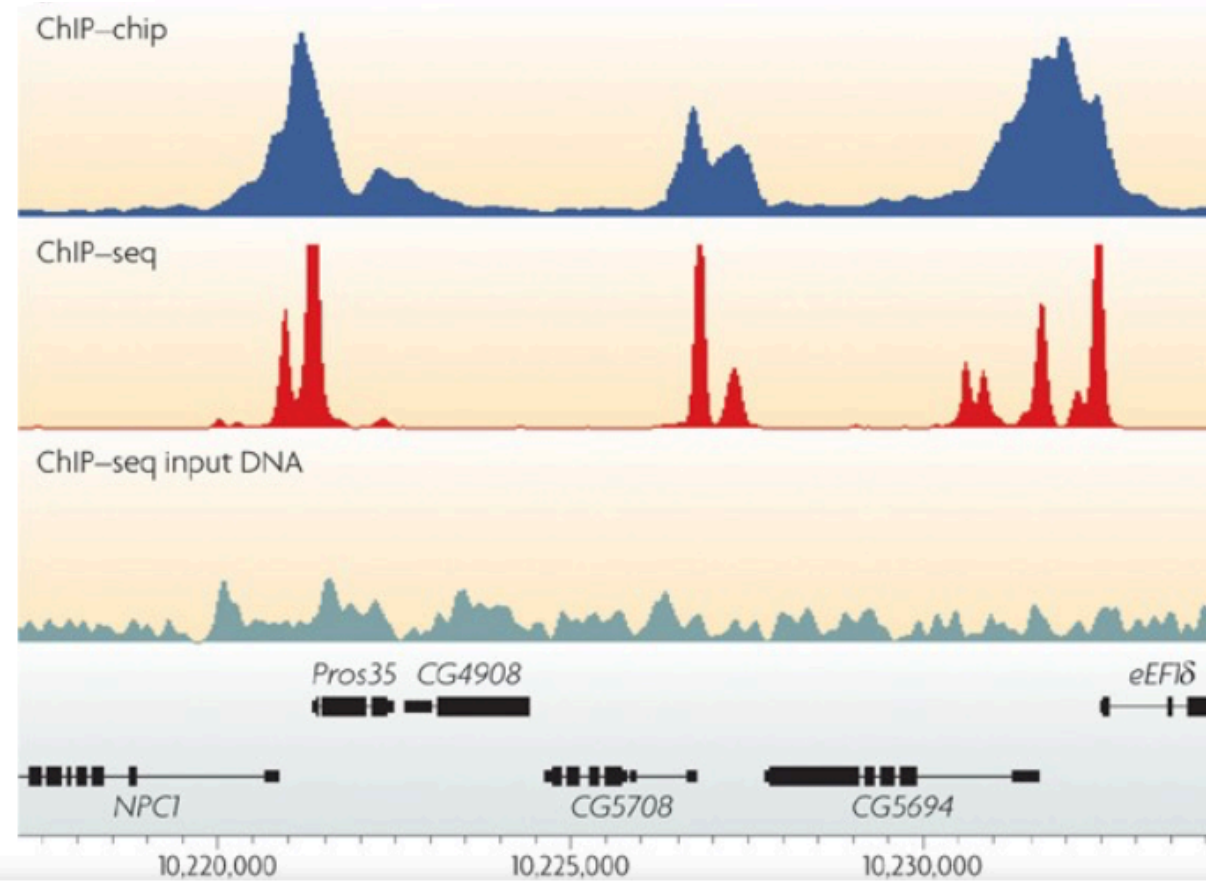


Bisulfite Conversion



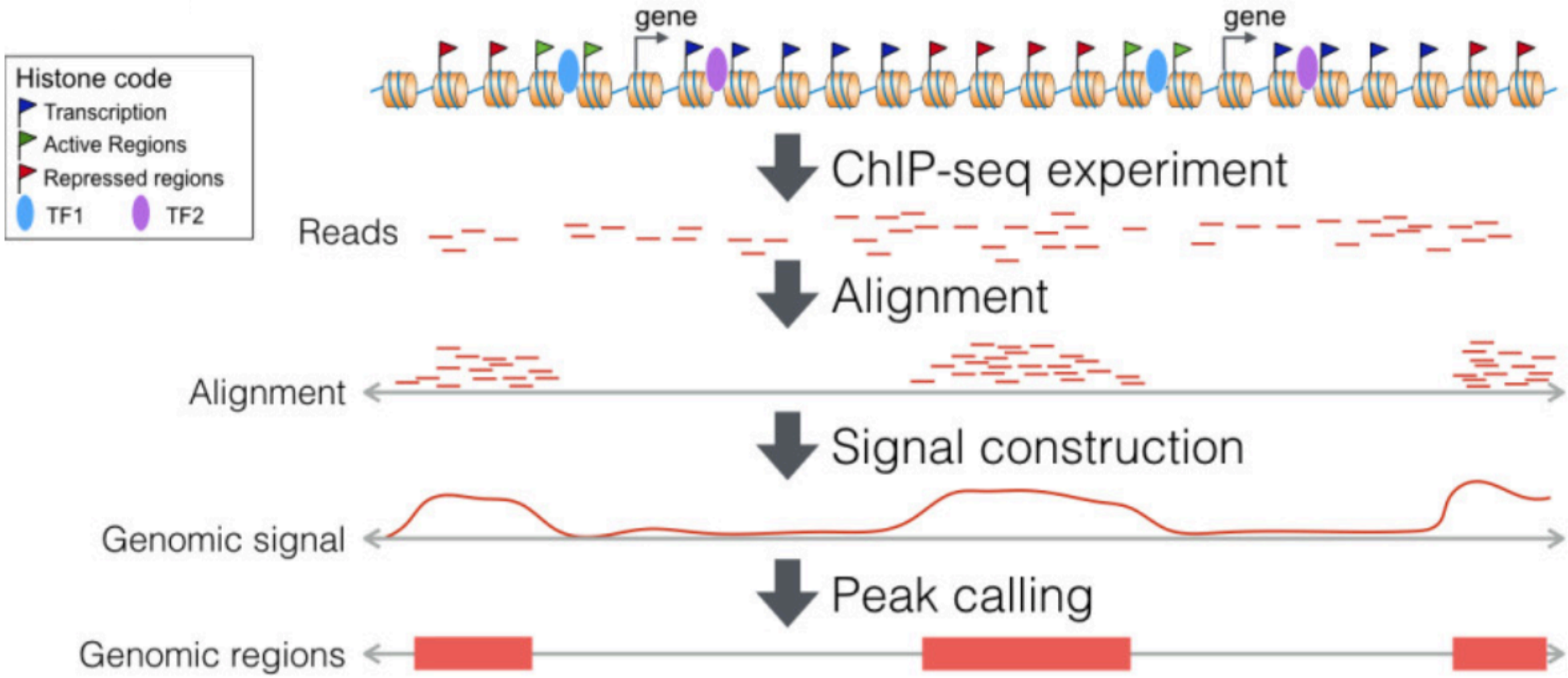
Chromatin Immunoprecipitation Sequencing (ChIP-Seq)



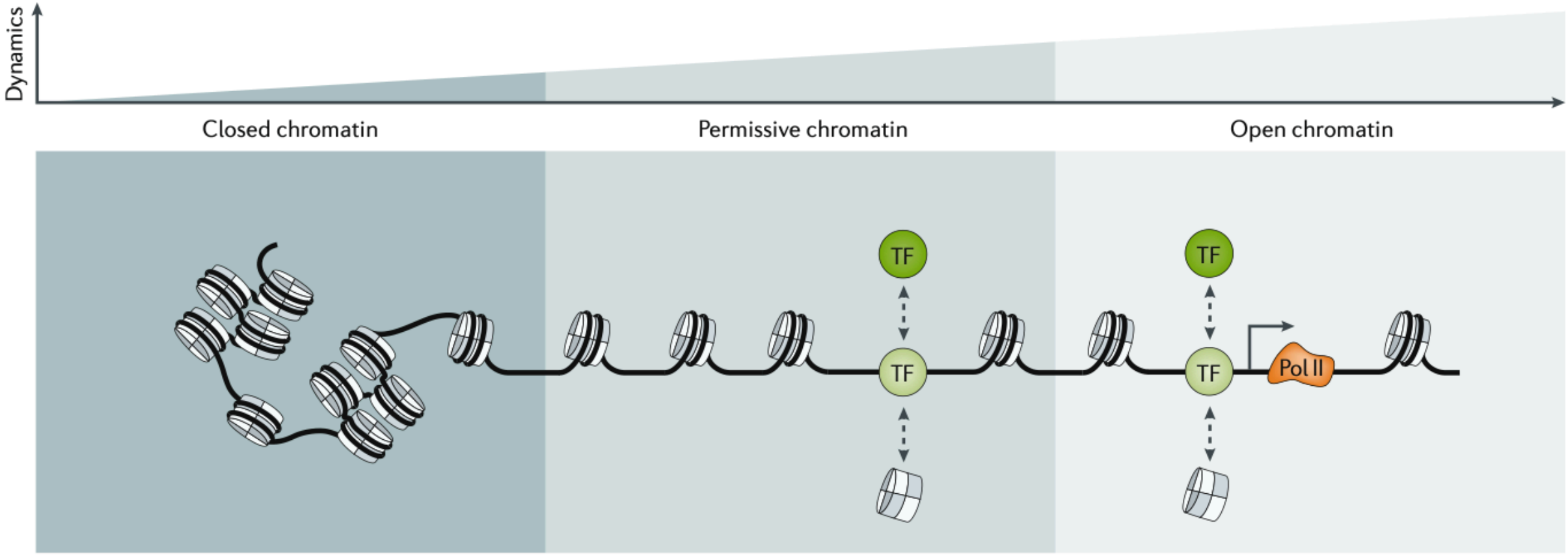


Park et al. 2009 Nature Rev Genet. 10: 669

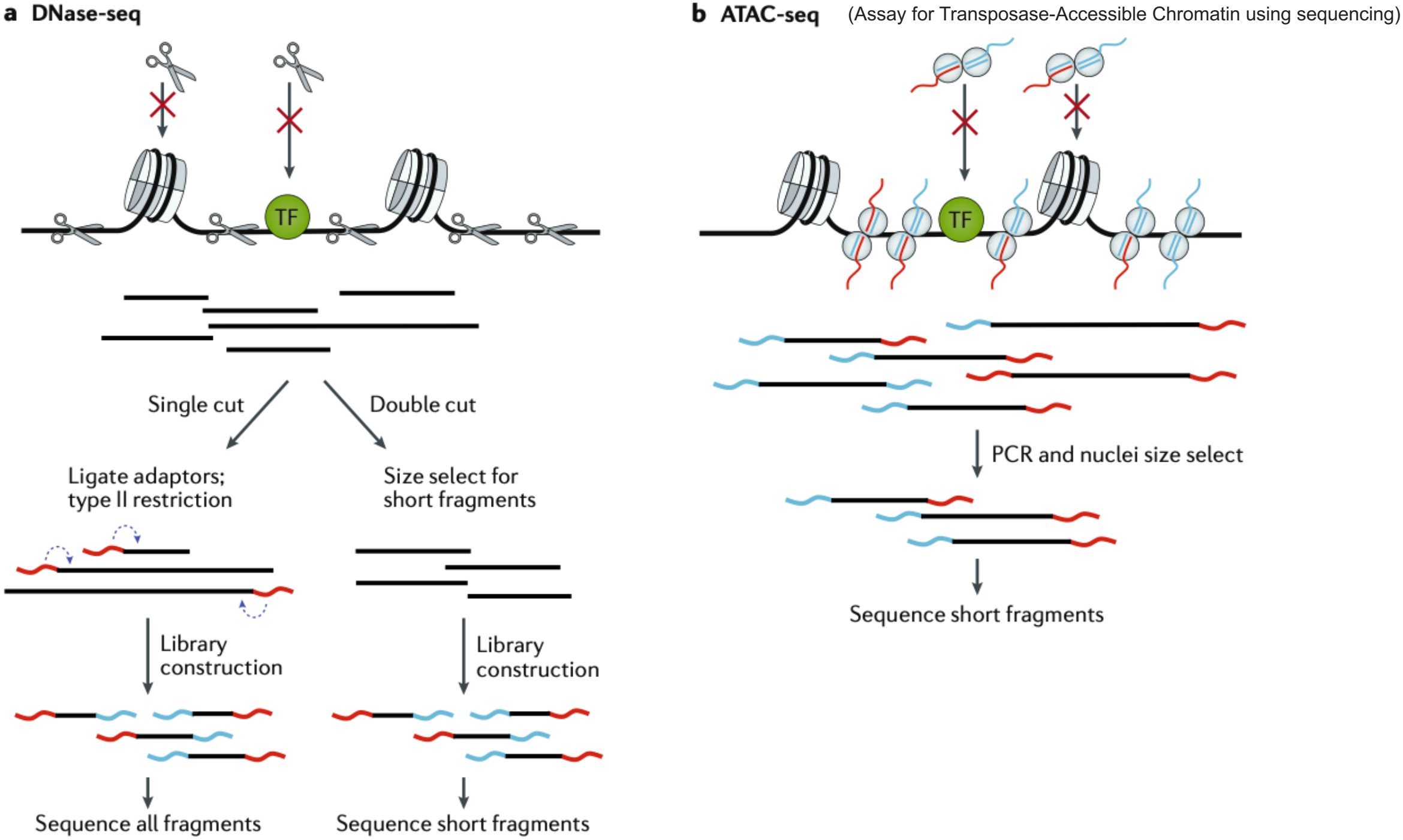
Typical pipeline for analysis of histone modification

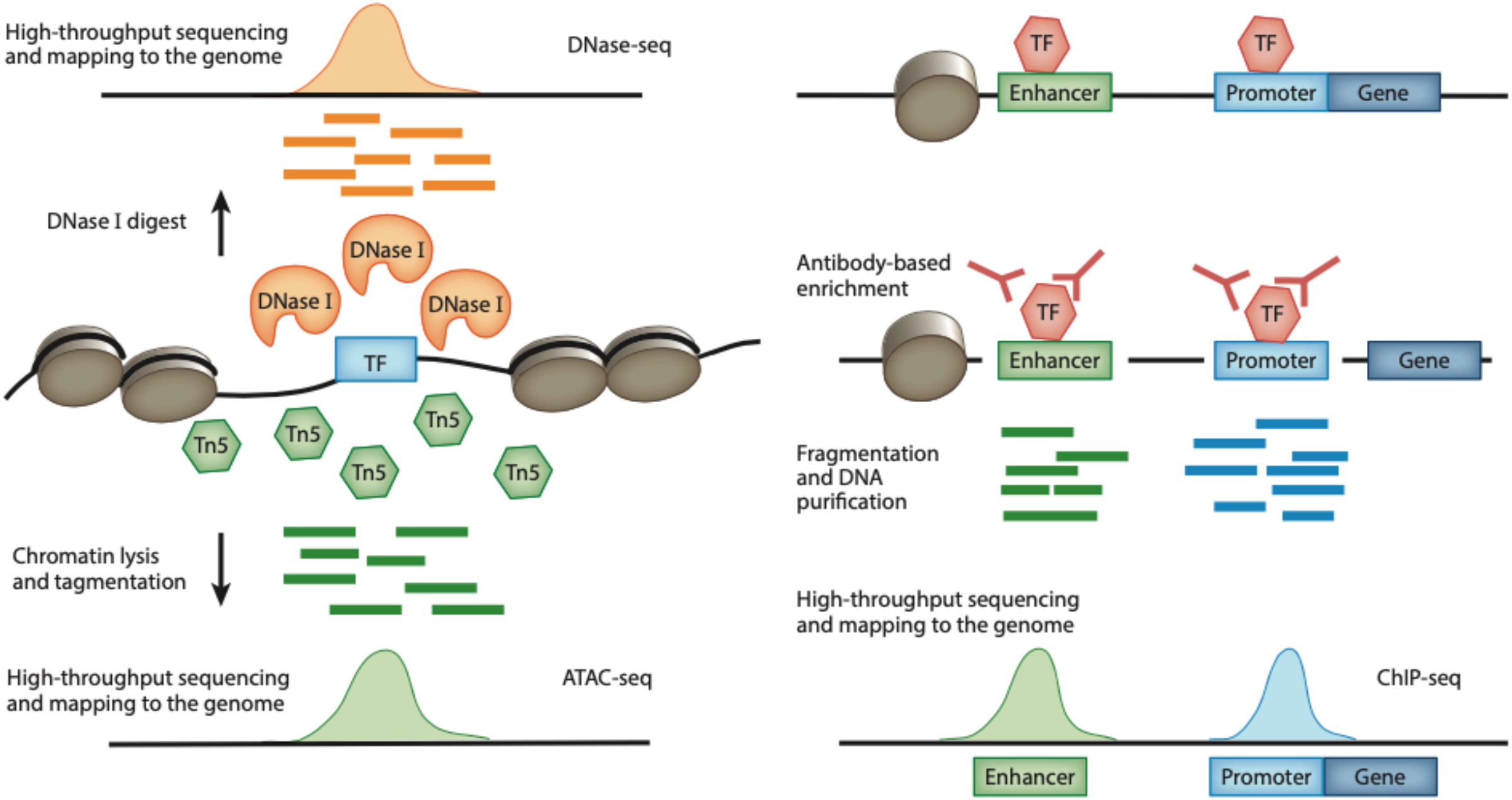


A continuum of accessibility states broadly reflects the distribution of chromatin dynamics across the genome

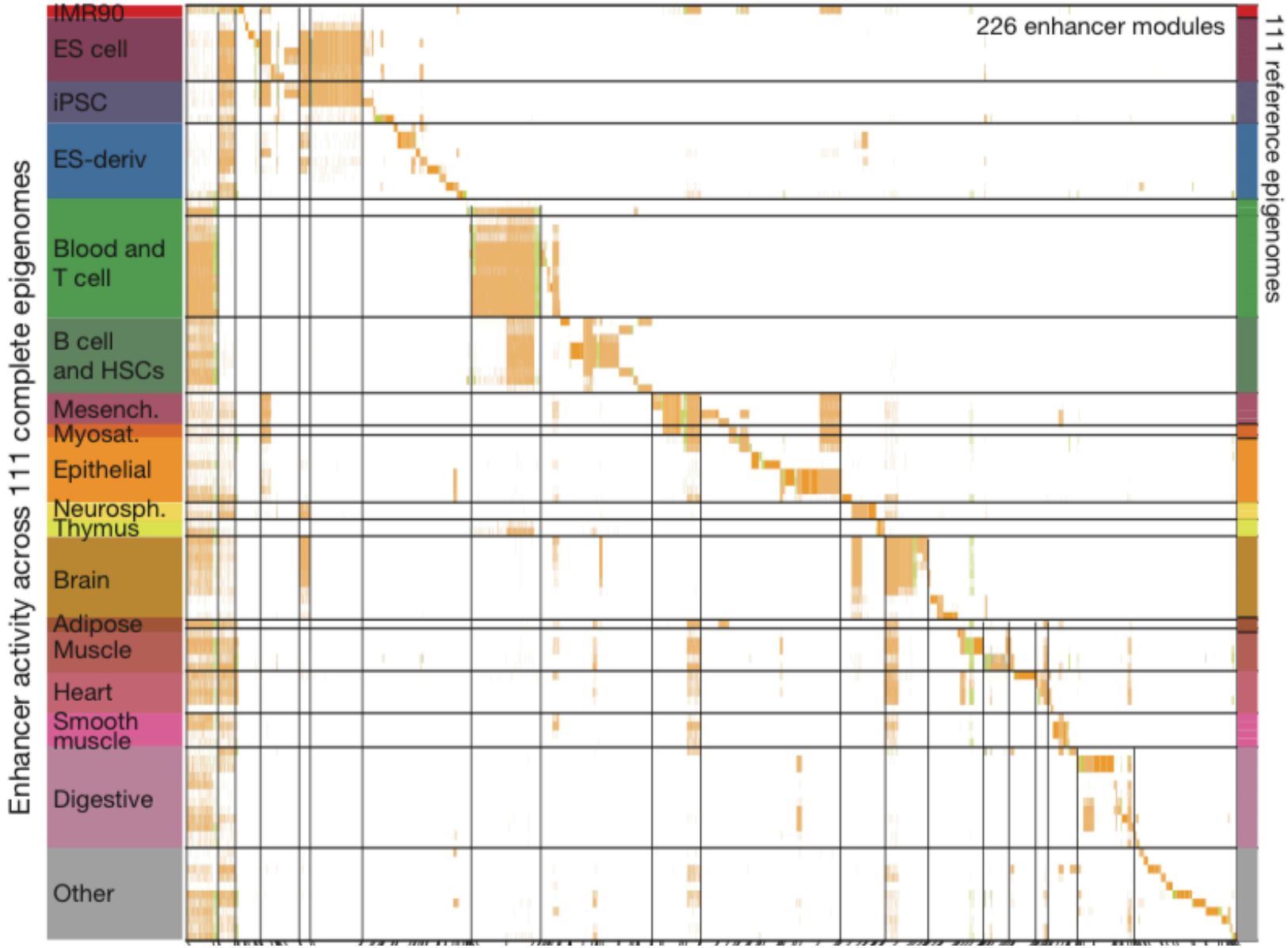


In contrast to closed chromatin, permissive chromatin is sufficiently dynamic for transcription factors to initiate sequence-specific accessibility remodeling and establish an open chromatin conformation (illustrated here for an active gene locus). Pol II, RNA polymerase II; TF, transcription factor.


a DNase-seq
TF
Single cut
Double cut
Ligate adaptors;
type II restriction
Size select for
short fragments
Library
construction
Library
construction
Sequence all fragments
Sequence short fragments
b ATAC-seq
(Assay for Transposase-Accessible Chromatin using sequencing)
TF
PCR and nuclei size select
Sequence short fragments

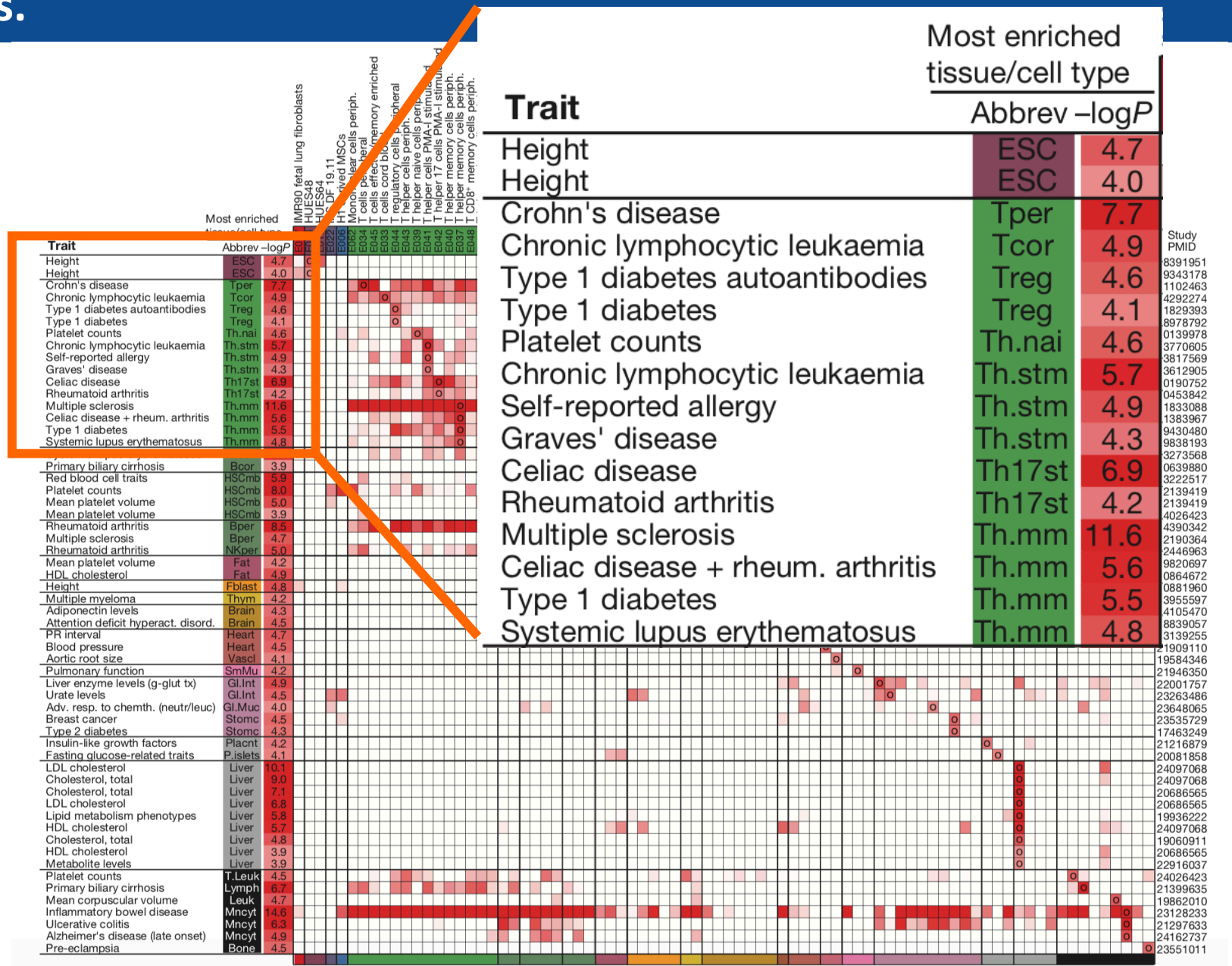

High-throughput sequencing and mapping to the genome
DNase-seq
DNase I digest
DNase I
DNase I
DNase I
TF
Tn5
Tn5
Tn5
Tn5
Tn5
Chromatin lysis and tagmentation
High-throughput sequencing and mapping to the genome
ATAC-seq
TF
TF
Enhancer
Promoter
Gene
Antibody-based enrichment
TF
TF
Enhancer
Promoter
Gene
Fragmentation and DNA purification
High-throughput sequencing and mapping to the genome
ChIP-seq
Enhancer
Promoter
Gene

Regulatory modules from epigenome dynamics



Roadmap Epigenomics Consortium Nature 2015, 518: 317-330

Epigenomic enrichments of genetic variants associated with diverse traits.



Roadmap Epigenomics Consortium
Nature 2015, 518: 317-330

Epigenomics

Epigenomics focuses on genome-wide characterization of reversible modifications of DNA or DNA-associated proteins, such as DNA methylation, chromatin accessibility or histone acetylation.

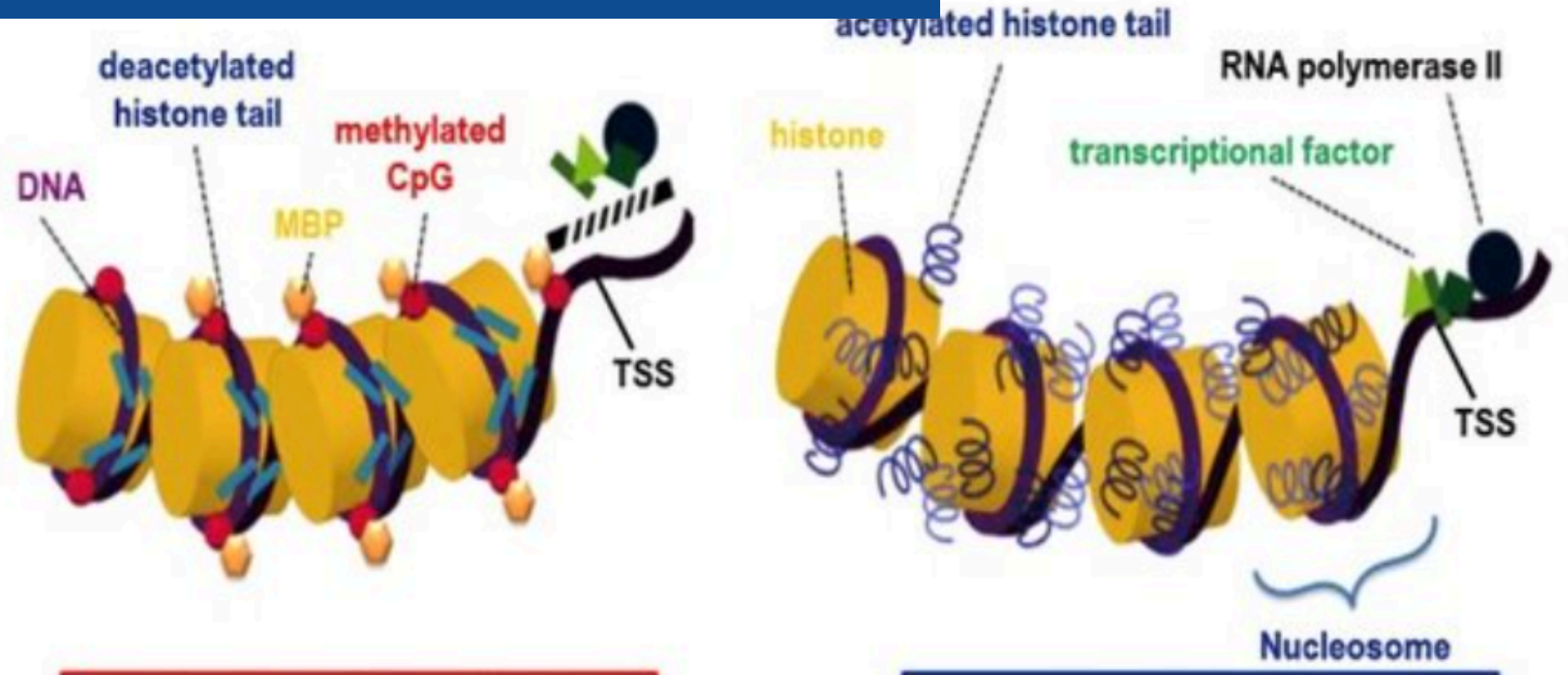
Those modifications can be influenced both by genetic and environmental factors, can be long lasting, and are sometimes heritable. While the role of epigenetic modifications as mediators of transgenerational environmental effects remains controversial

Their importance in biological processes and disease development is evident from many epigenome-wide association studies that have been reported. For example, differentially methylated regions of DNA can be used as indicators of disease status for metabolic syndrome (GenomeBiol. 2013;14:R115; Cell Metab. 2015;21:138–49), cardiovascular disease (PLoS One. 2010;5:e9692), cancer (Hum Mol Genet. 2001;10:687–92),and many other pathophysiologic states

Epigenetic signatures are often tissue-specific (Cell. 2013;152:642–54), and several large consortia are focusing on establishing comprehensive epigenomic maps in multiple human tissues (Roadmap Epigenomics (http://www.roadmapepigenomics.org/) and International Human Epigenome Consortium (http://ihec-epigenomes.org/))

Data generated by these studies has great potential to enhance our functional interpretation of genetic variants residing in those regions or of epigenetic markers associated with disease independently of genetic variation

Relationship between epigenetics and gene transcription

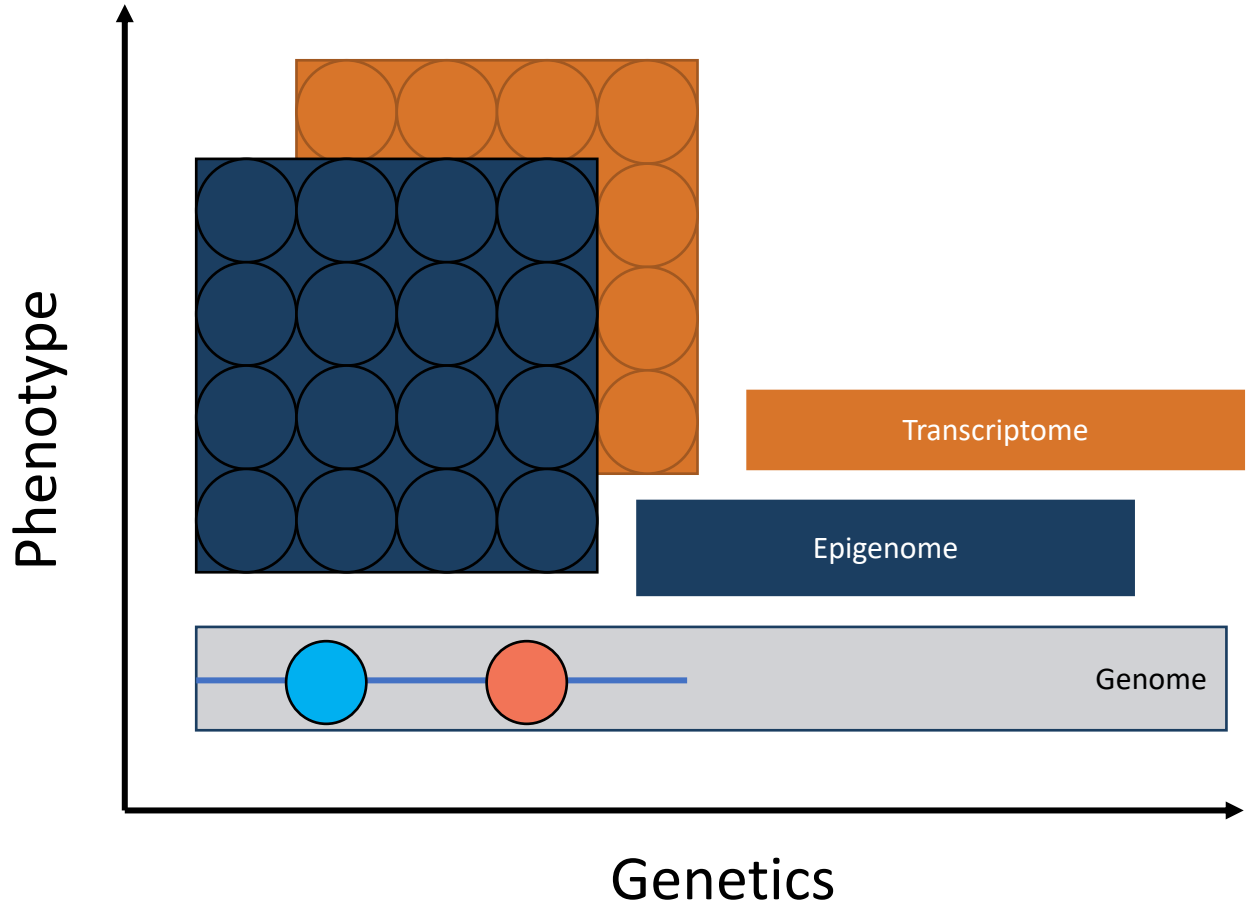


Heterochromatin
closed chromatin
conformation
: repression

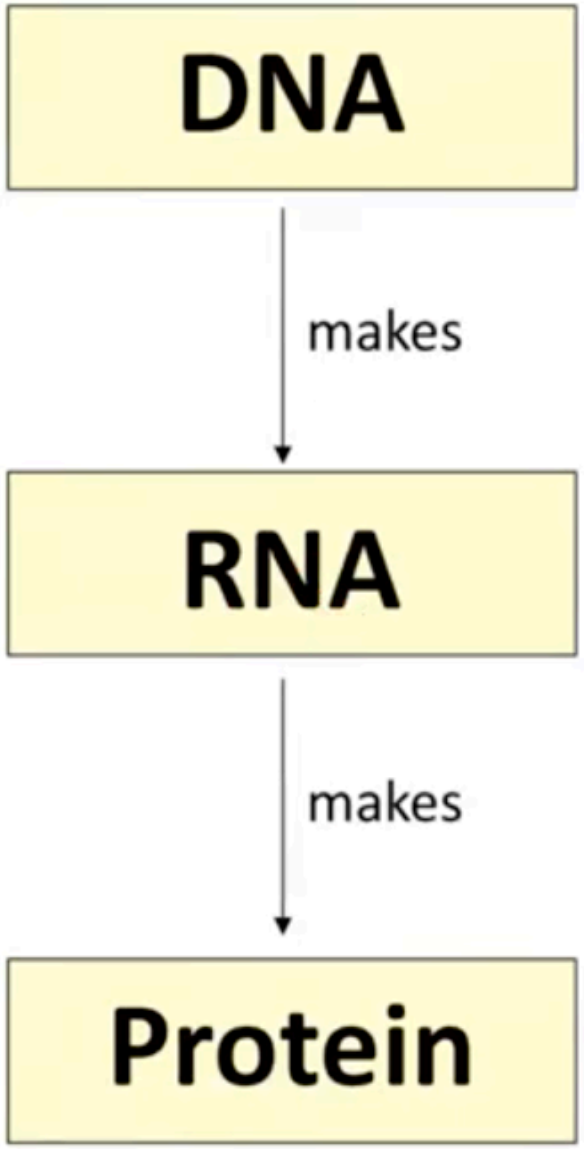
Euchromatin
open chromatin
conformation
: activation

TRANSCRIPTION

Transcriptomics



Central Dogama



Information storage and cross-generation transmission

RNA can store information

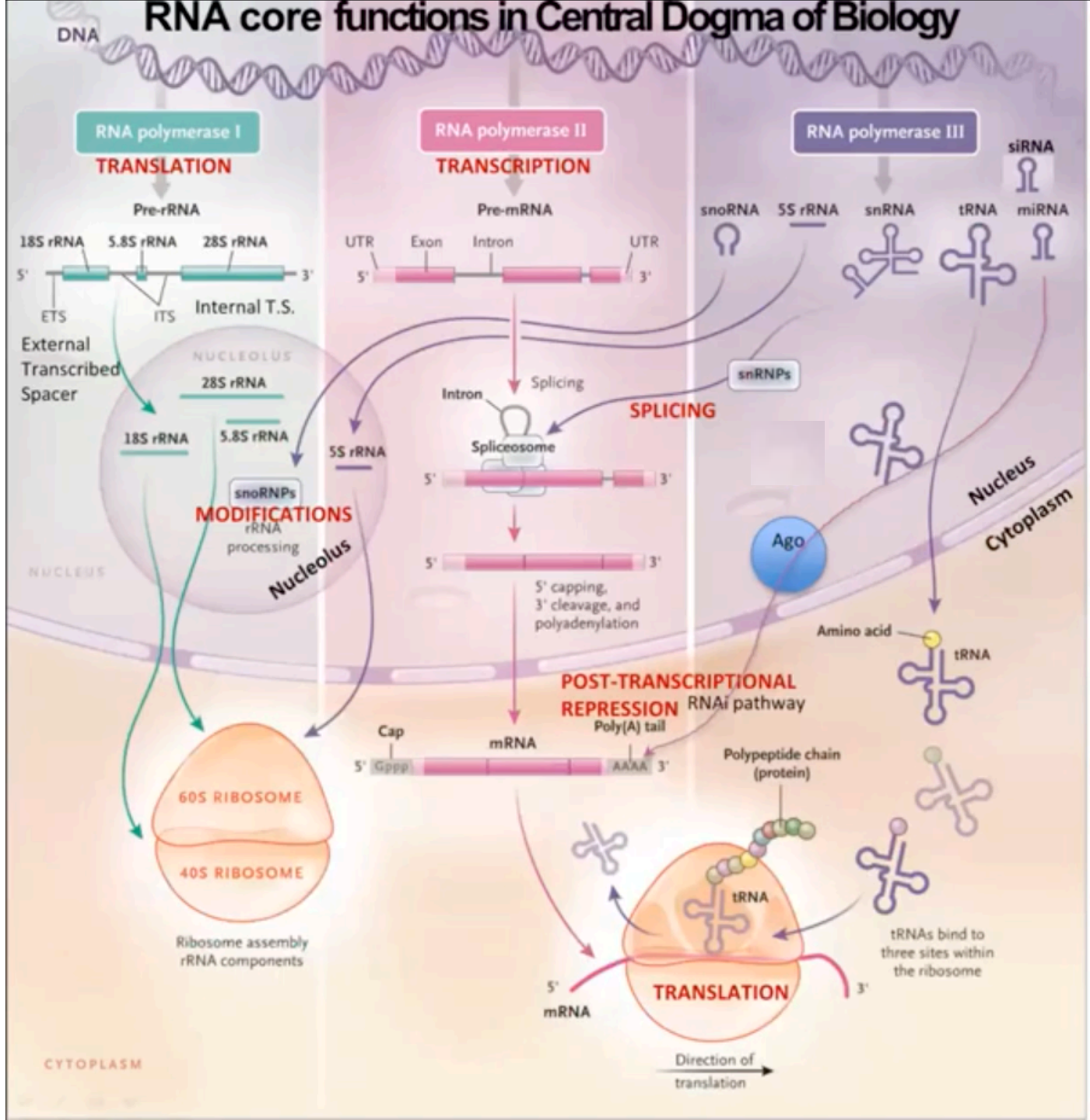
Temporary copy (boring)

RNA can transmit information

Carry out all cellular functions

RNA can carry out enzymatic reactions

RNA core functions in Central Dogma of Biology



TRANSCRIPTION

Pol II: mRNAs (messenger RNAs)

- Protein-coding genes
- lncRNAs (long non-coding RNAs)

TRANSLATION

Pol I: rRNAs (ribosomal RNAs)

- 28S, 18S, 5.8S ribosomal RNAs (these combine with 5S, from Pol III)
- Core of ribosome translation machinery

Pol III: many diverse small RNAs

TRANSLATION

tRNAs (transfer RNAs): translation

- Codon-to-AminoAcid adapter molecules

POST-TRANSCRIPTIONAL REPRESSION

miRNAs (microRNAs): ~22nt long

siRNAs (small interfering RNAs): 22-25nt

- Post-transcriptional regulation (3'UTR)
- Processing by Dicer, incorporate into Ago, form RISC complex, targets mRNAs by base-pairing(exact:siRNA, inexact:miRNA)
- Leads to cleavage; instability (polyA); less efficient translation

PROCESSING

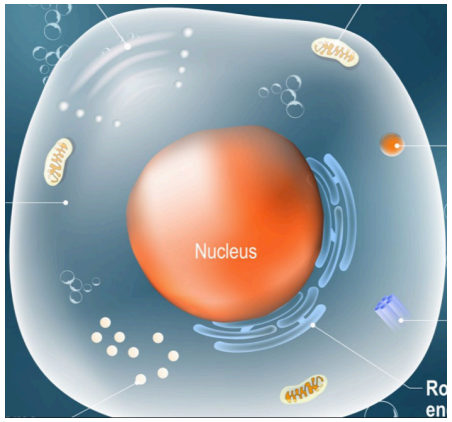
snoRNAs (small nucleolar RNAs)

- rRNA methylation (C/D box)
- rRNA pseudouridylation (H/ACA box)

SPLICING

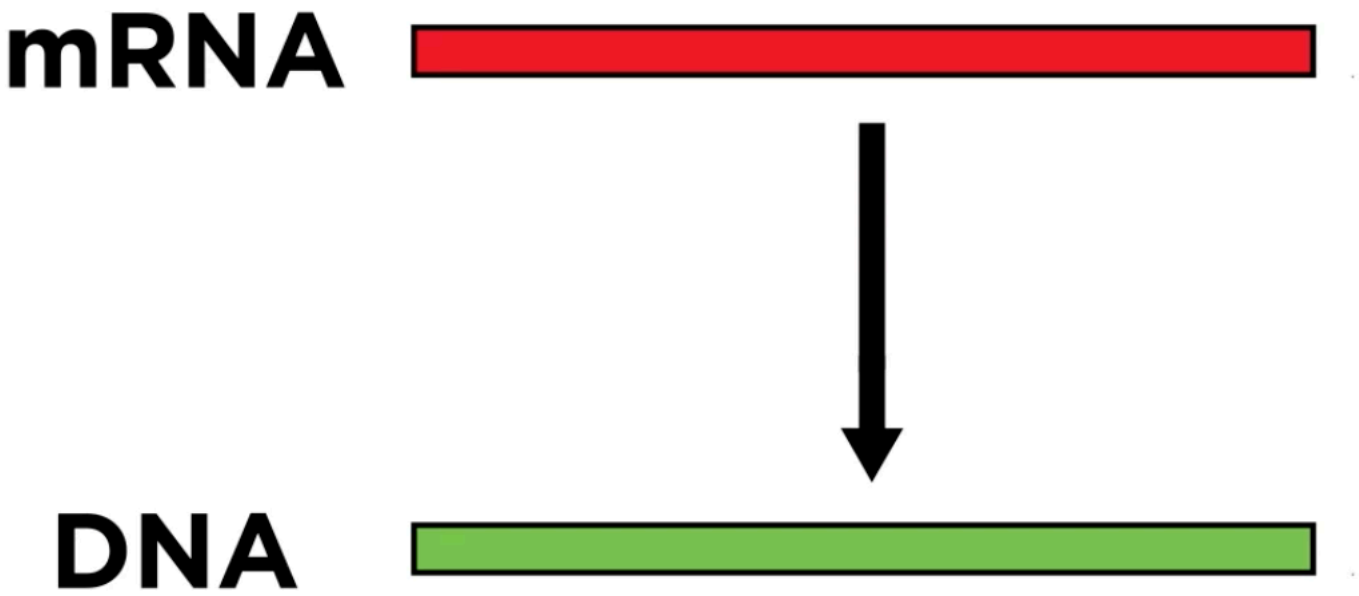
snRNAs: small nuclear RNAs (U-RNAs)

- Splicing: U1, U2, U4, U5, U6 (uridines)
- Other: TF regulation, maintain telomeres

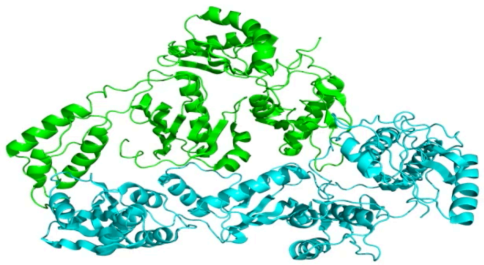


1)Isolate the **mRNAs** from a cell

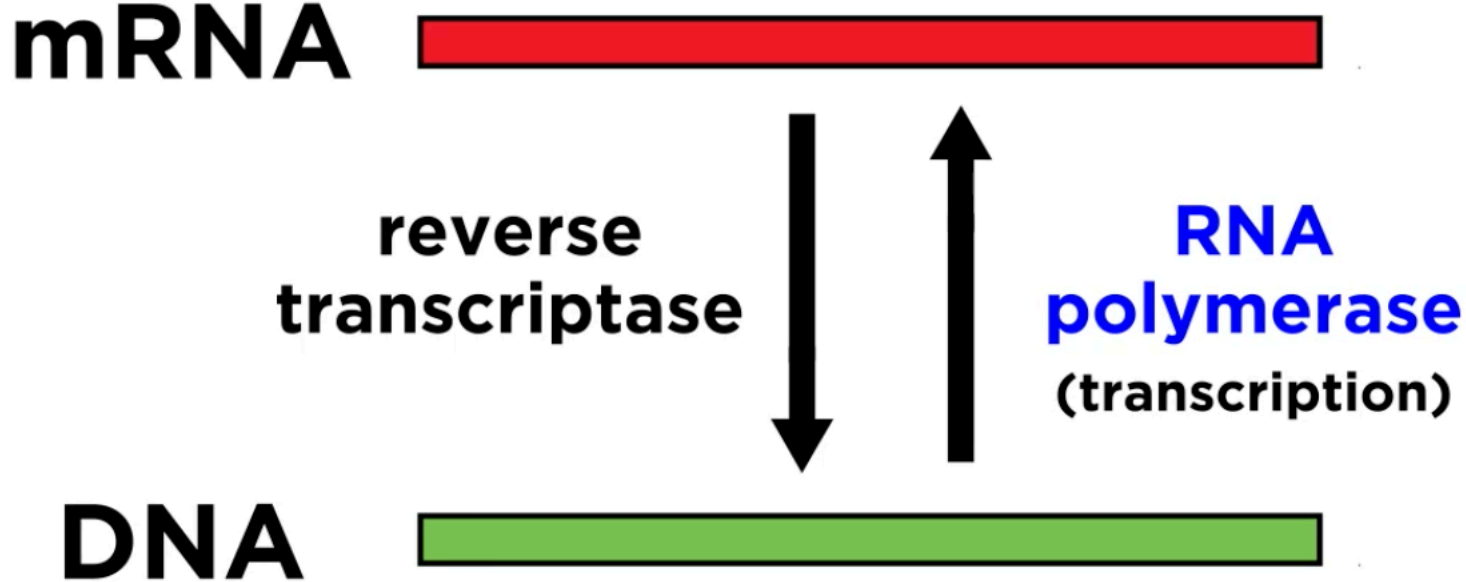
2)Performing the reverse transcriptase polymerase chain reaction (RT-PCR)

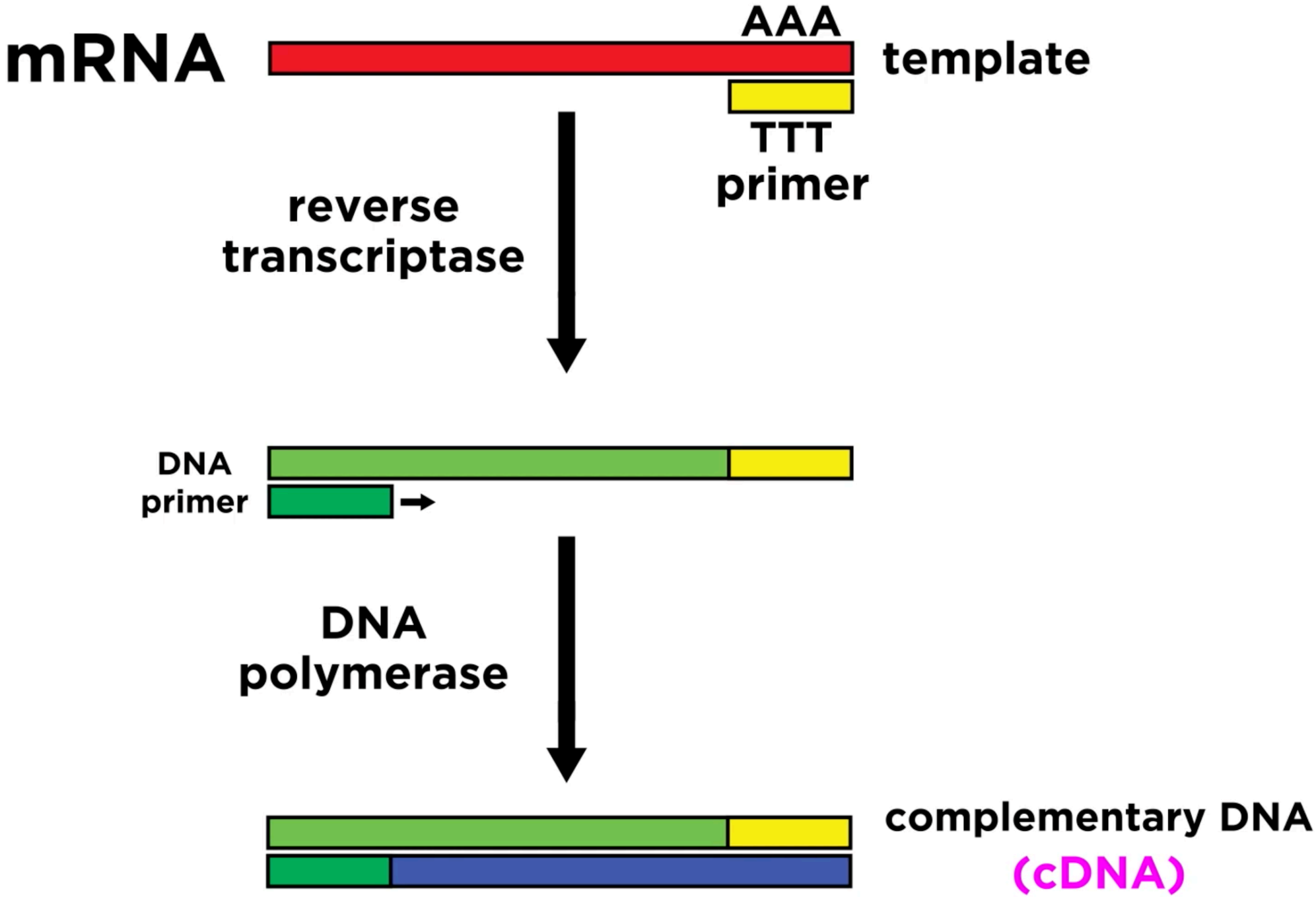


We are reverse engineering the **DNA template** that would have generated the mRNA

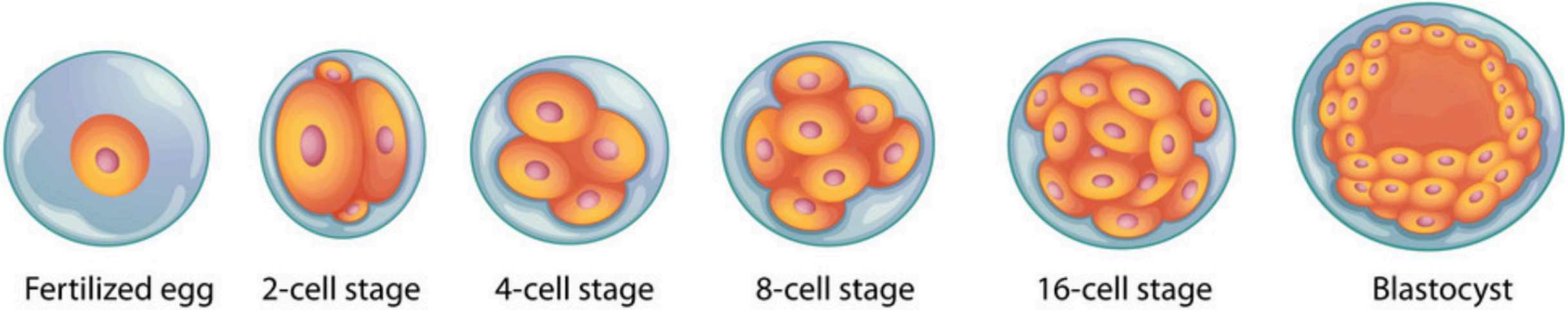


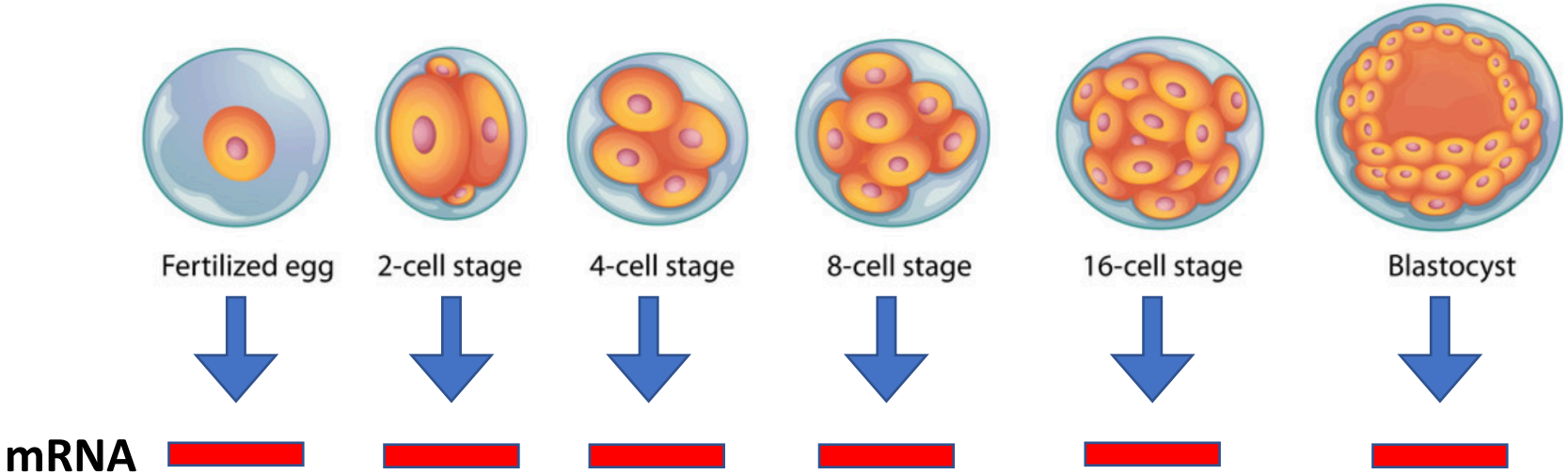
reverse transcriptase

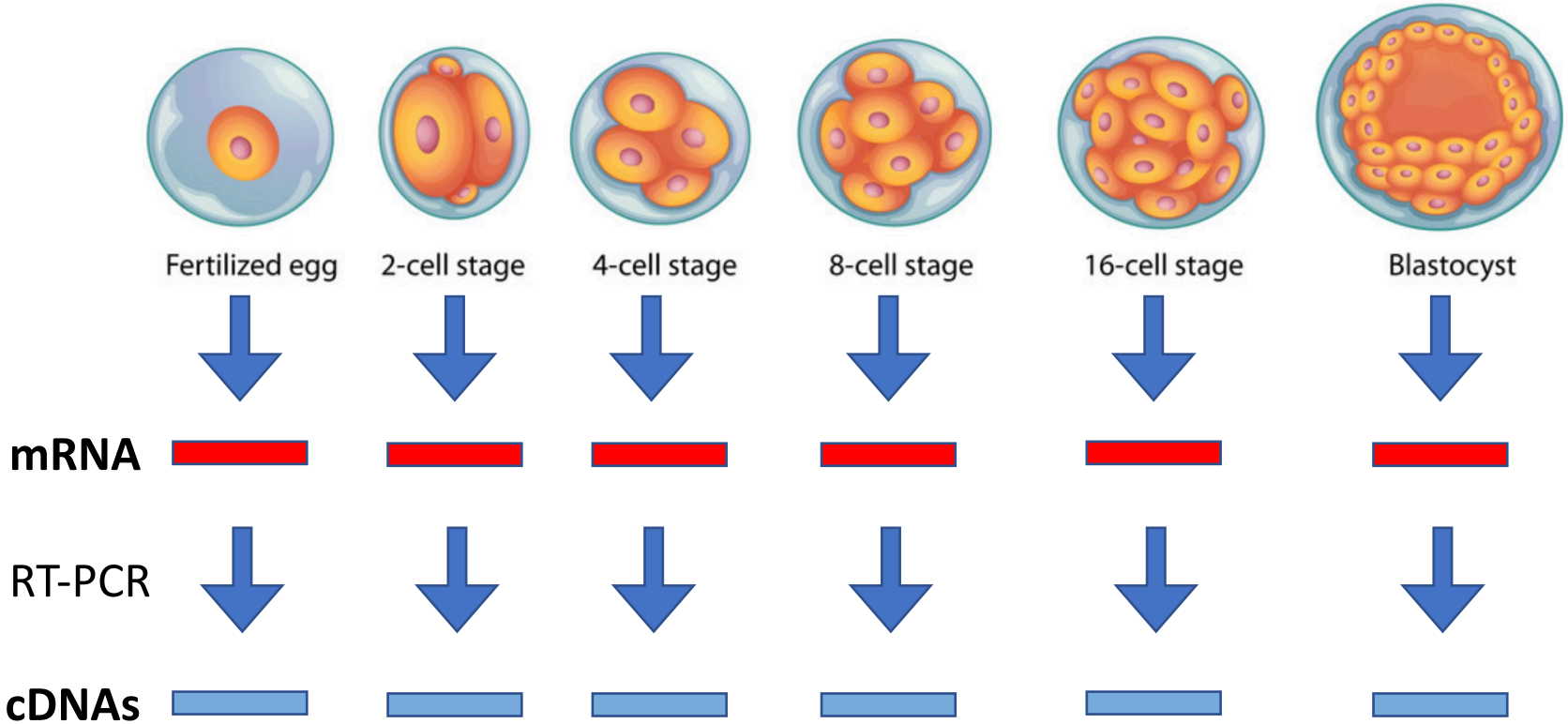


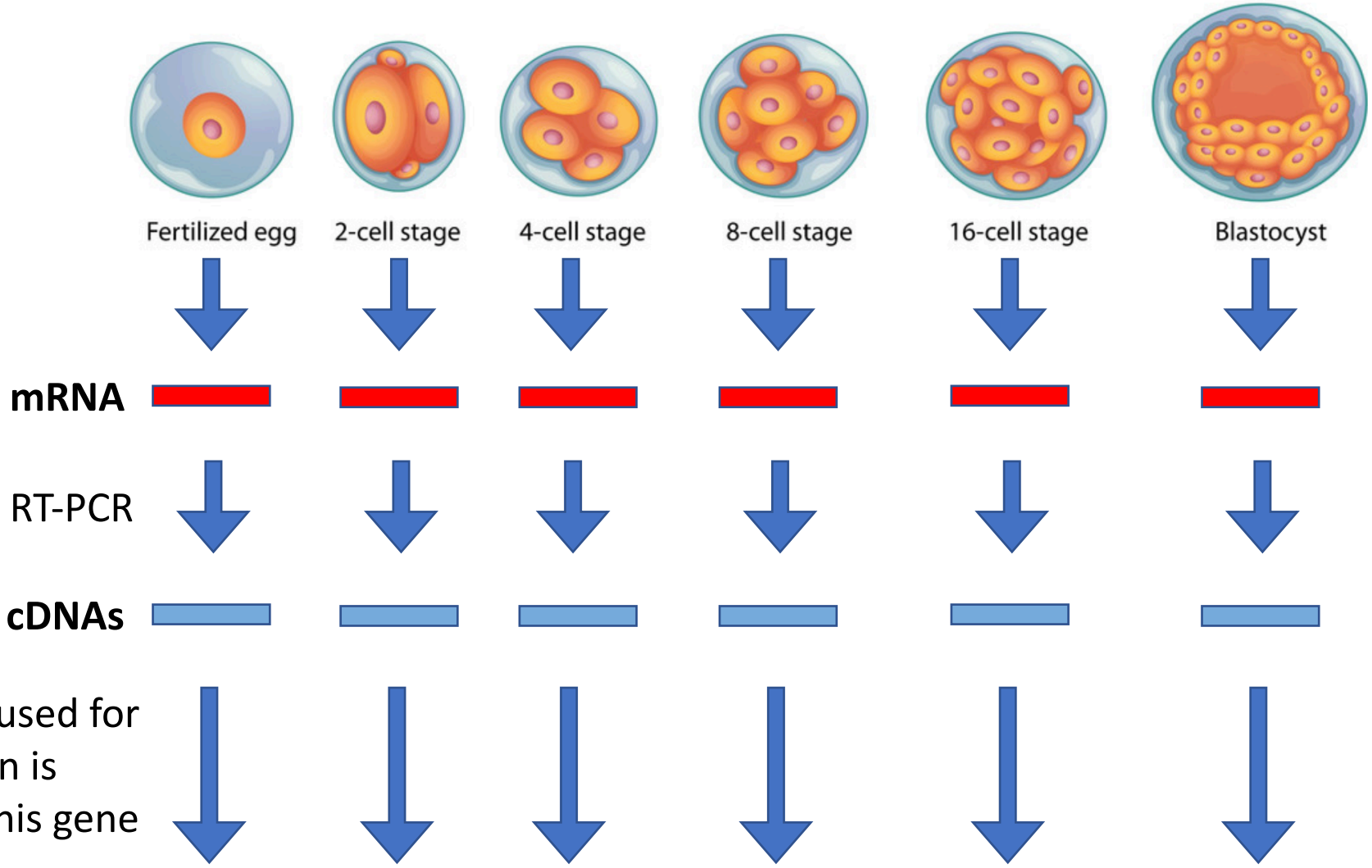

mRNA
AAA
template
TTT
primer
reverse
transcriptase
DNA
primer
DNA
polymerase
complementary DNA
(cDNA)

When a gene is being expressed during the embryonic development




Fertilized egg
2-cell stage
4-cell stage
8-cell stage
16-cell stage
Blastocyst
mRNA


Fertilized egg
2-cell stage
4-cell stage
8-cell stage
16-cell stage
Blastocyst
mRNA
RT-PCR
cDNAs



The primer used for amplification is specific to this gene

Many copies of this cDNA molecule

Gel Electrophoresis

Length of DNA fragment (base pairs)

1,000 bp

900 bp

800 bp

700 bp

600 bp

500 bp

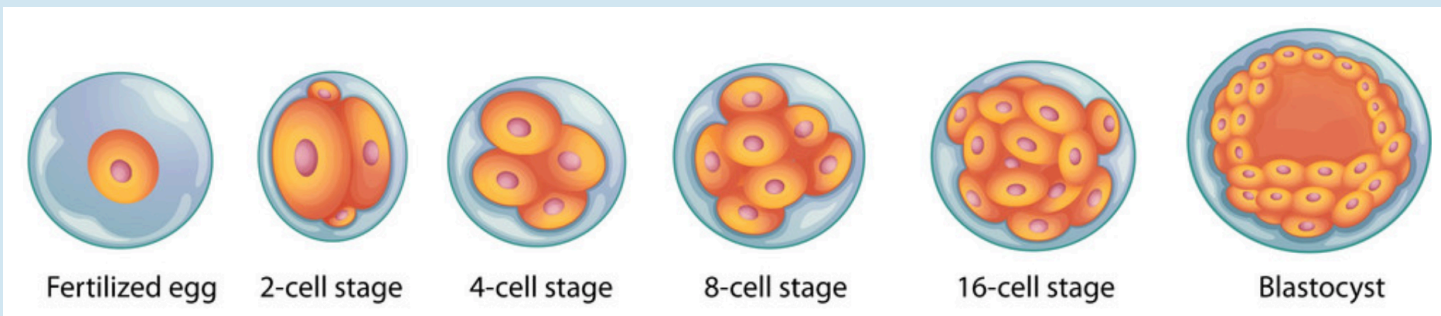
400 bp

300 bp

200 bp

100 bp

Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6



this is how we know when a gene is being expressed

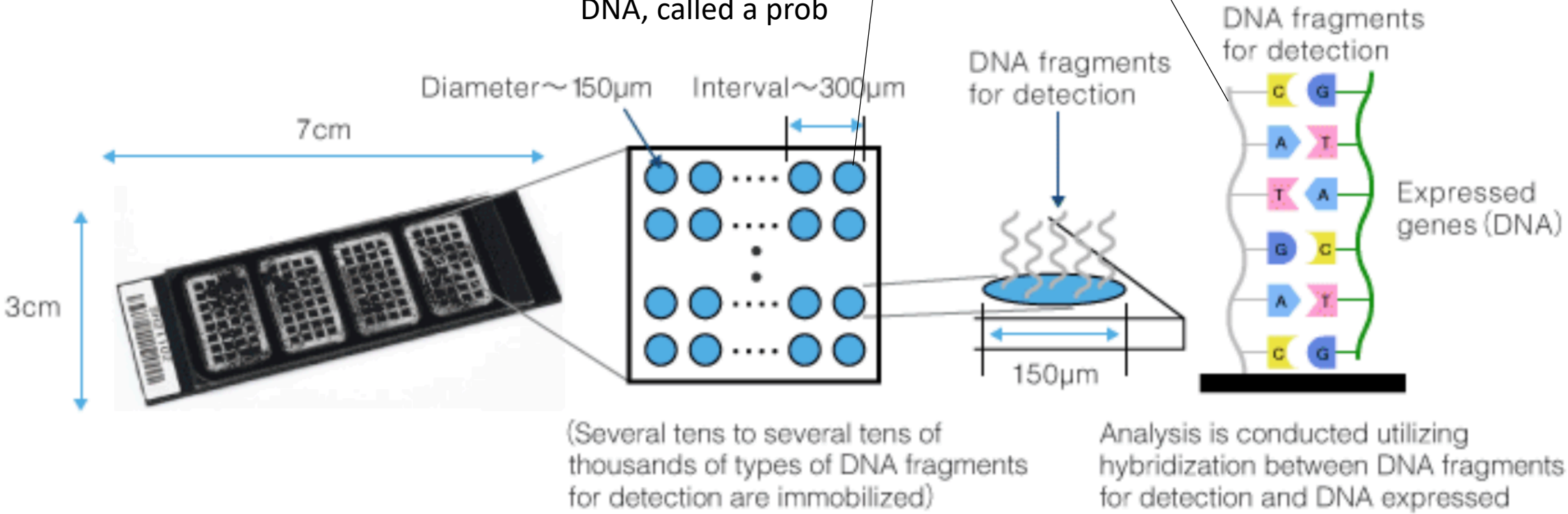
cDNA that represents the gene of interest

mRNA is produced during that stage of development

Microarray

Involved huge spots, and each well contains many copies of single copies of DNA, called a prob

Single-stranded DNA probes tethered to a solid support

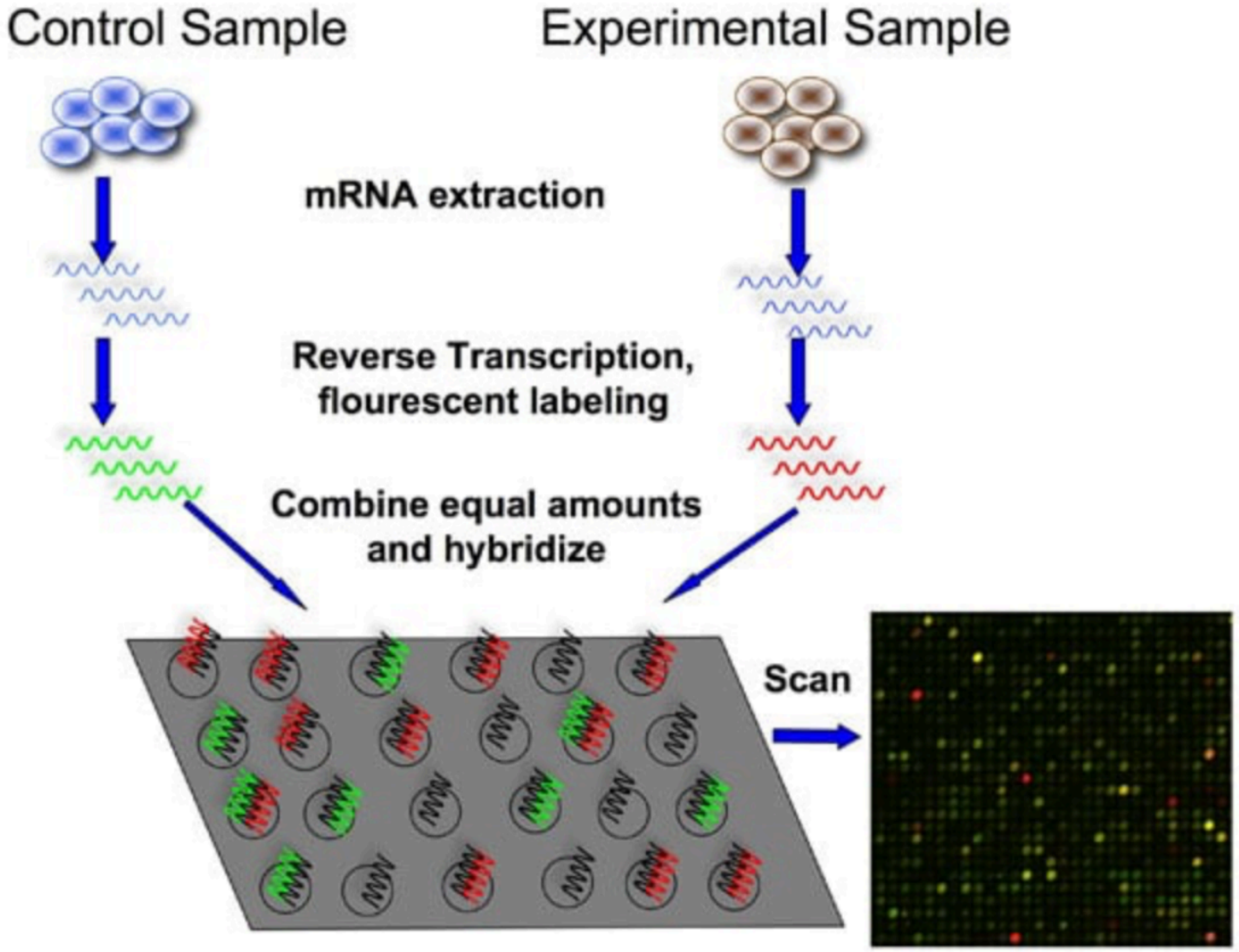


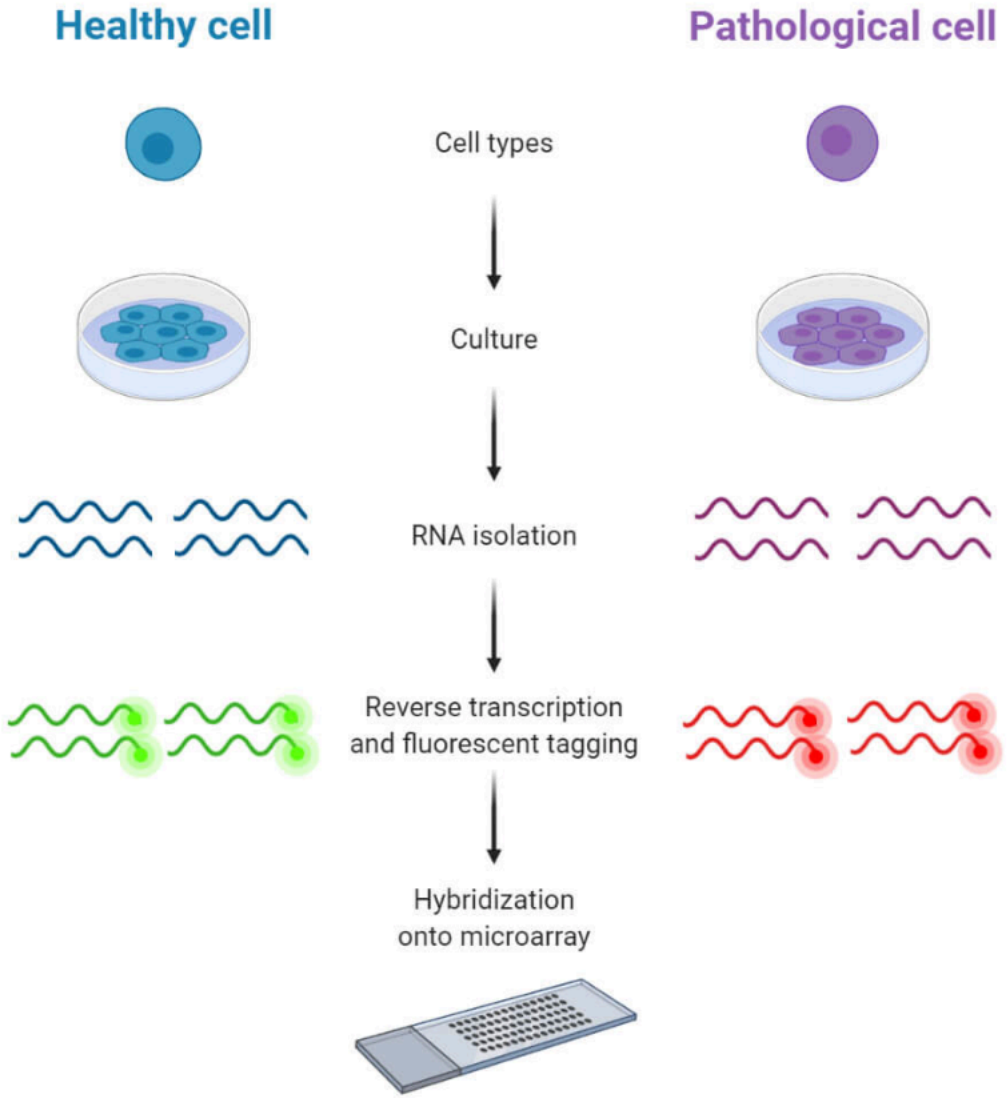
(Several tens to several tens of thousands of types of DNA fragments for detection are immobilized)

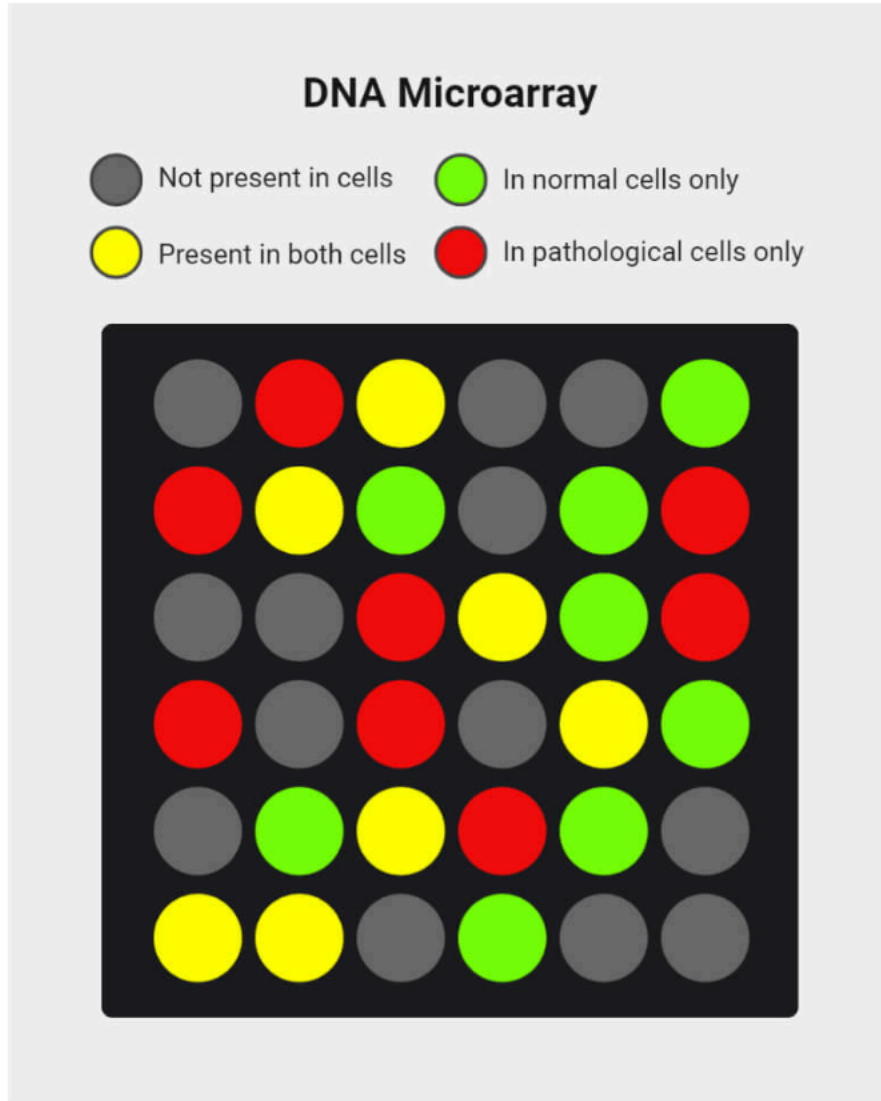
Analysis is conducted utilizing hybridization between DNA fragments for detection and DNA expressed

https://www.3d-gene.com/en/about/chip/index.html#/

Microarray

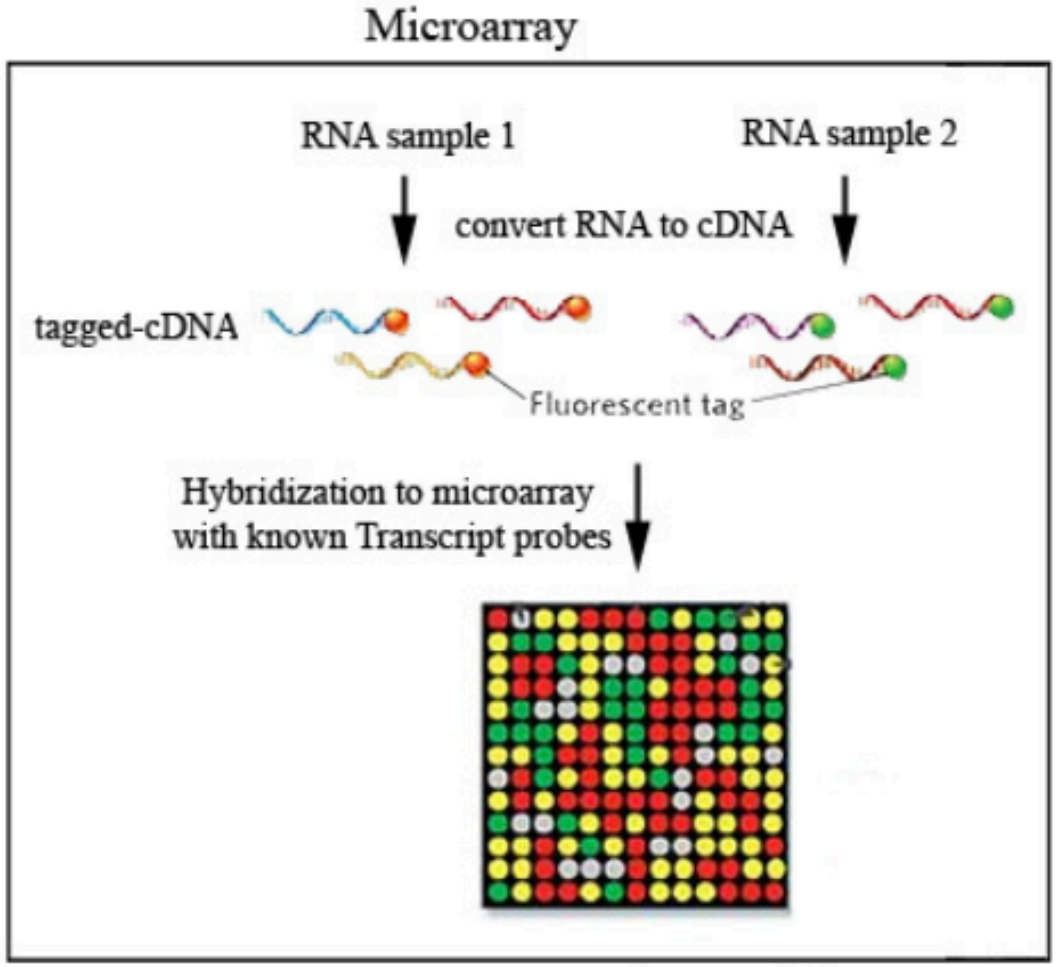

Control Sample
Experimental Sample
mRNA extraction
Reverse Transcription,
flourescent labeling
Combine equal amounts
and hybridize
Scan





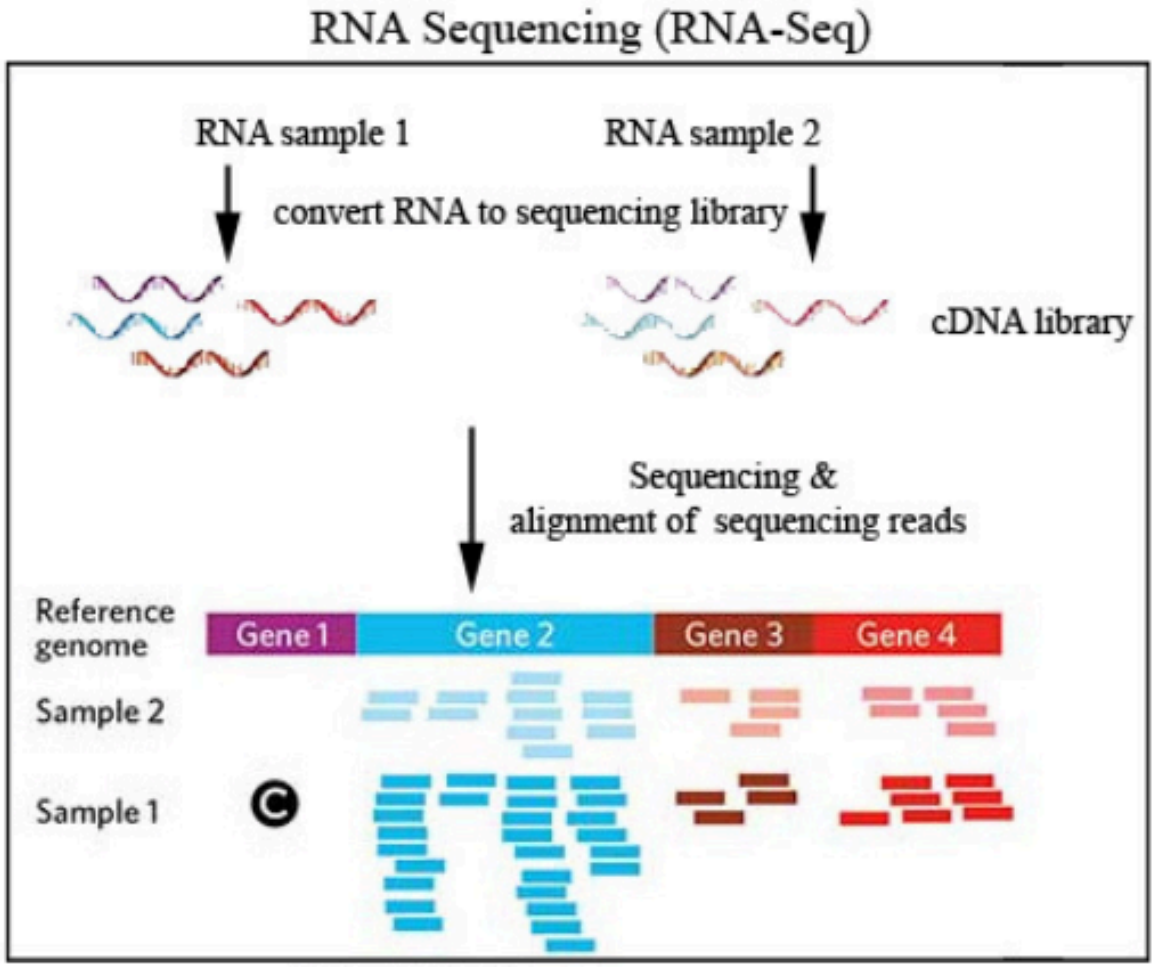
https://microbenotes.com/dna-microarray/

Microarray vs. RNA-Seq



relative intensity
=
expression levels

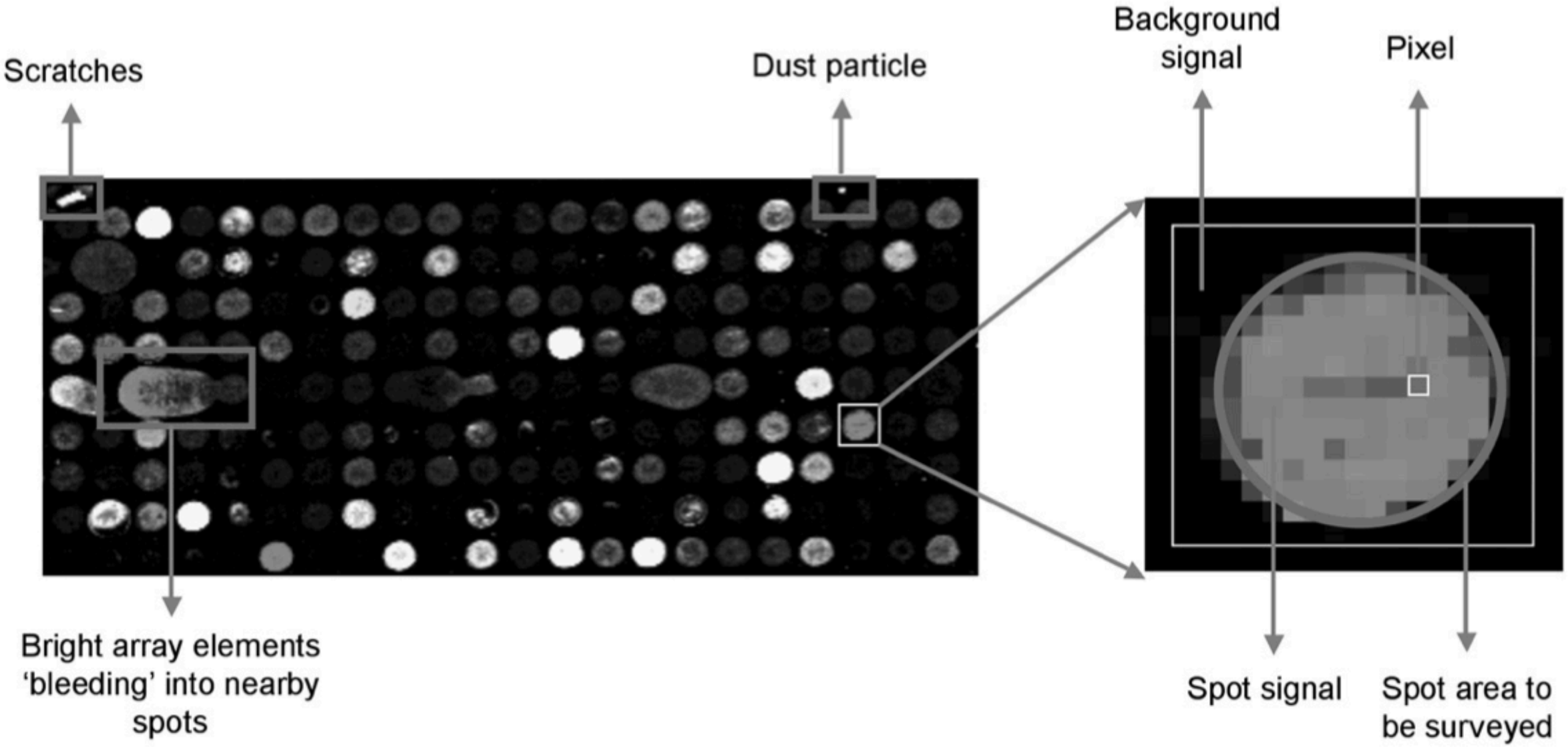
Low sensitivity
Low dynamic range
known transcript only
No alternative splicing information
lower cost



High sensitivity
High dynamic range
Novel transcripts sequences identified
structural variation & alternative splicing revealed
unlimited sample comparisons

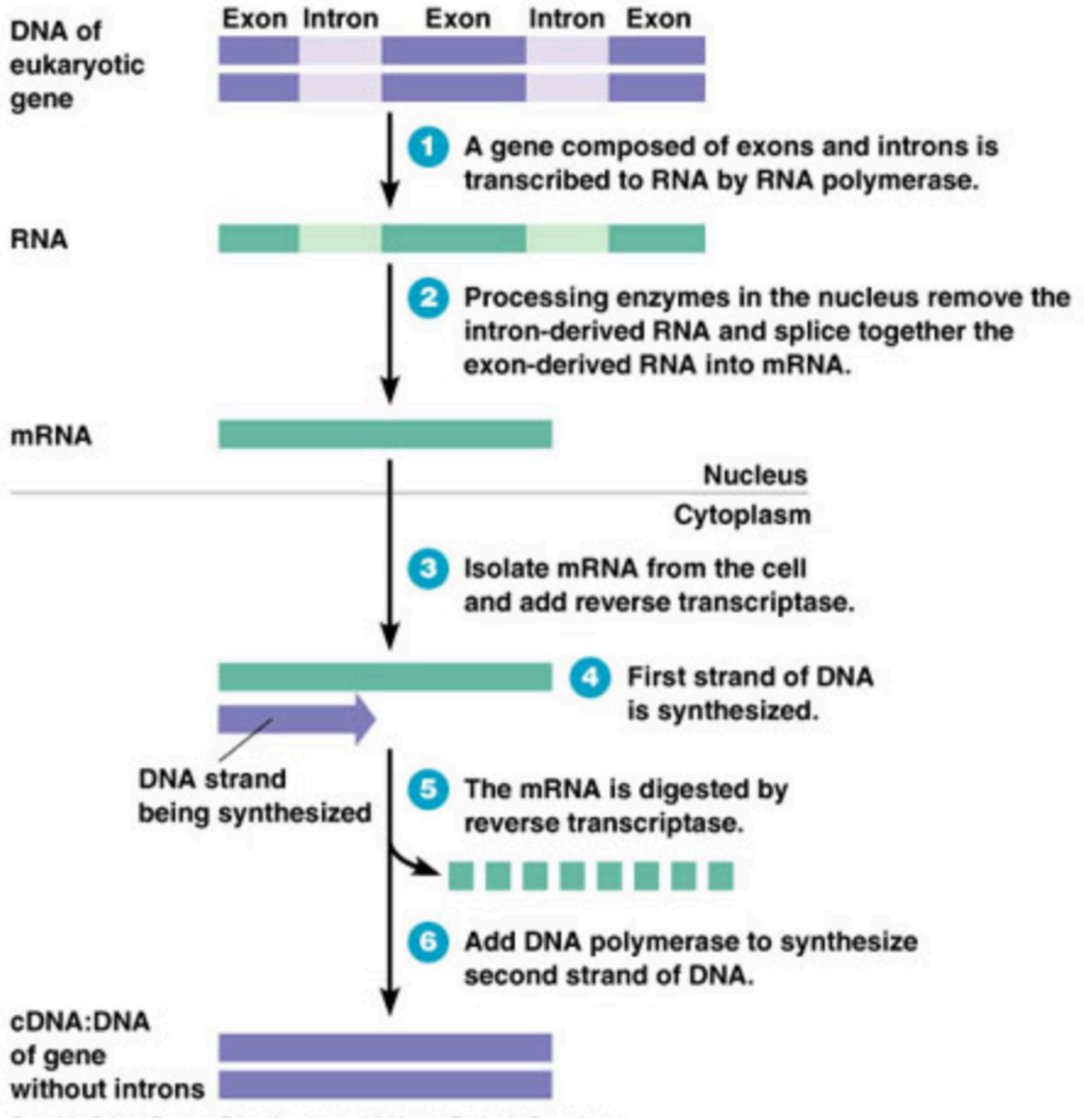
Sequencing Reads
=
expression levels

Microarray images are not perfect, as it is often the case, which leads to many problems with spurious signals from dust particles



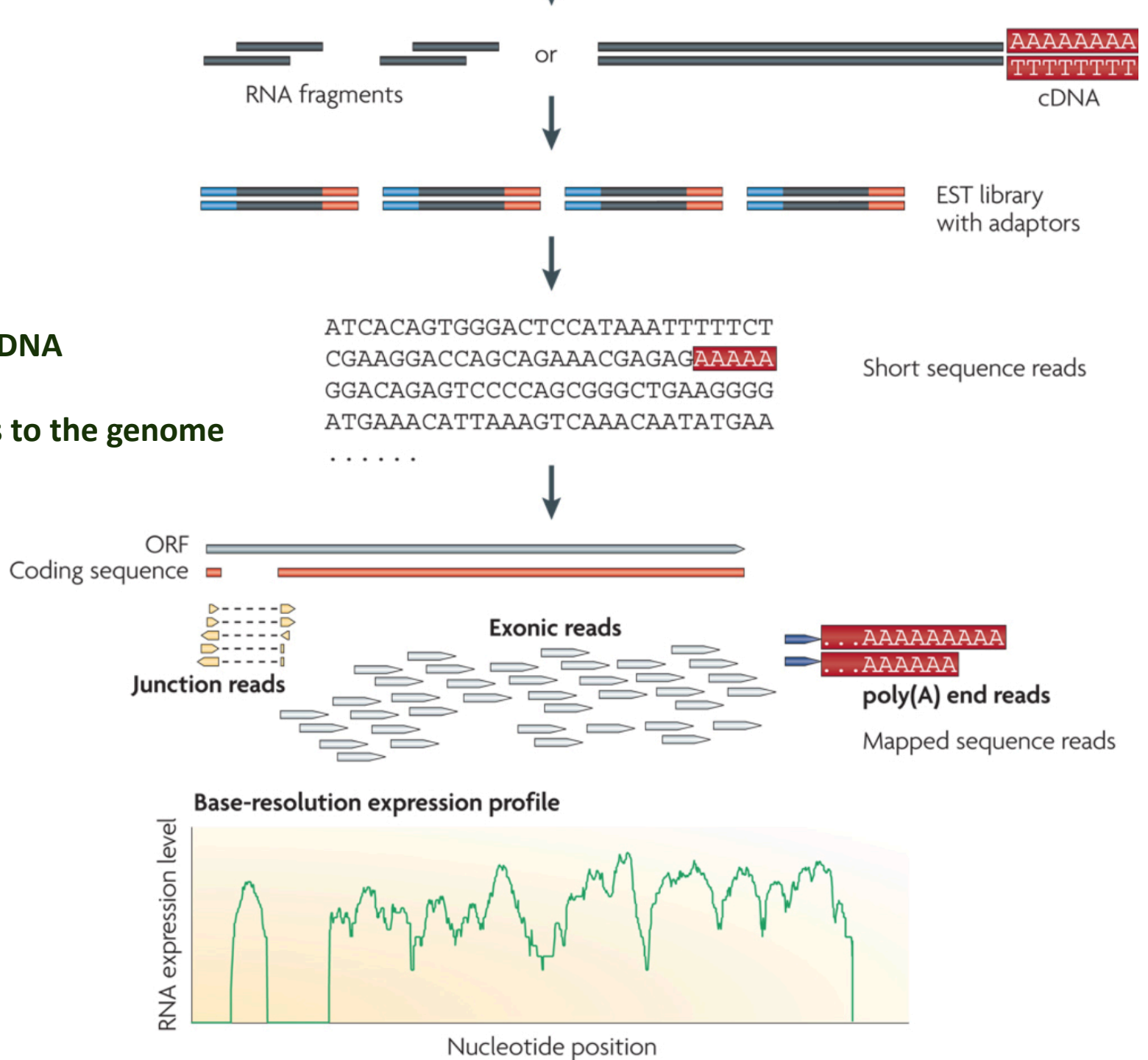
Common RNA-Seq workflow

1. Isolate all mRNA
2. Convert to cDNA using reverse transcriptase

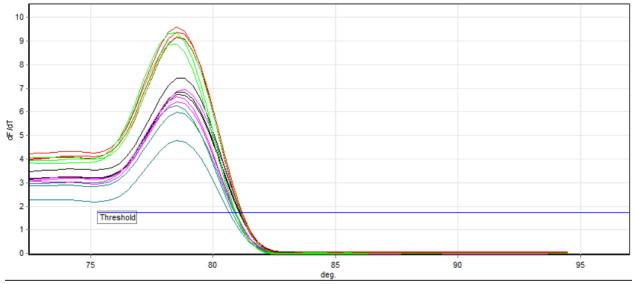


3. Sequence the cDNA

4. Map sequences to the genome

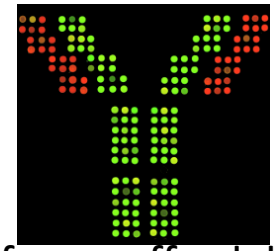


qPCR, Microarrays or RNA-sequencing: When To Choose One Over the Other?



1. qPCR

Choose qPCR for your experiment if you have to analyze the expression of a few genes with a known sequence

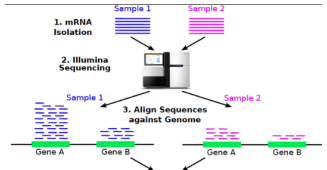


2. Microarrays

Consider going for microarrays if you are looking for an affordable and robust solution, especially if you do not know which genes you want to analyze – or for when you want to perform a whole transcriptome deferentially expressed (DE) genes analysis and have a good reference sequence for your organism.

3. RNA-Seq

In comparison to microarrays, RNA-sequencing (or RNA-seq for short) enables you to look at differential expressions at a much broader dynamic range, to examine DNA variations (SNPs, insertions, deletions) and even discover new genes or alternative splice variations using just one dataset.



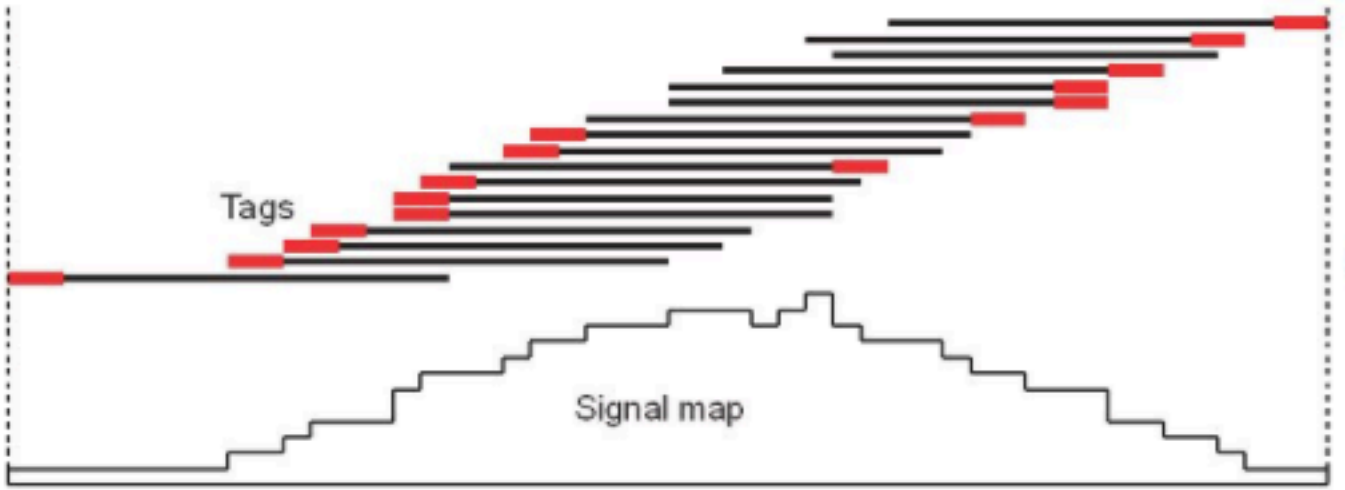
RNA-Seq Challenges

1. Exons are separated by introns – only <3% human genome codes exons.
Mapping reads to the reference genome is challenging

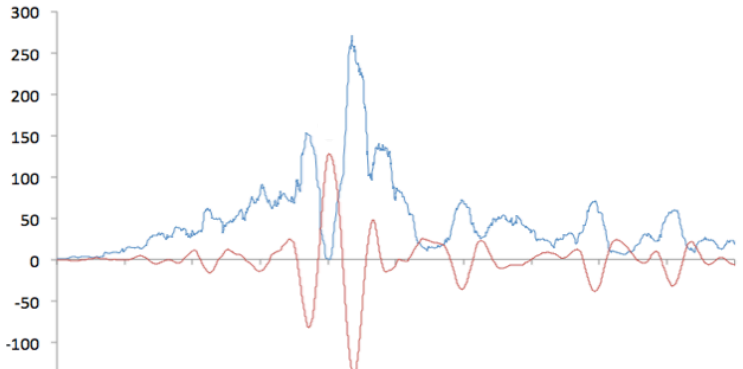
2. The relative abundance of RNAs vary over a very large range
Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may make up the majority of reads

3. RNAs come in a wide range of sizes
Small RNAs must be captured separately

RNA-Seq signal data

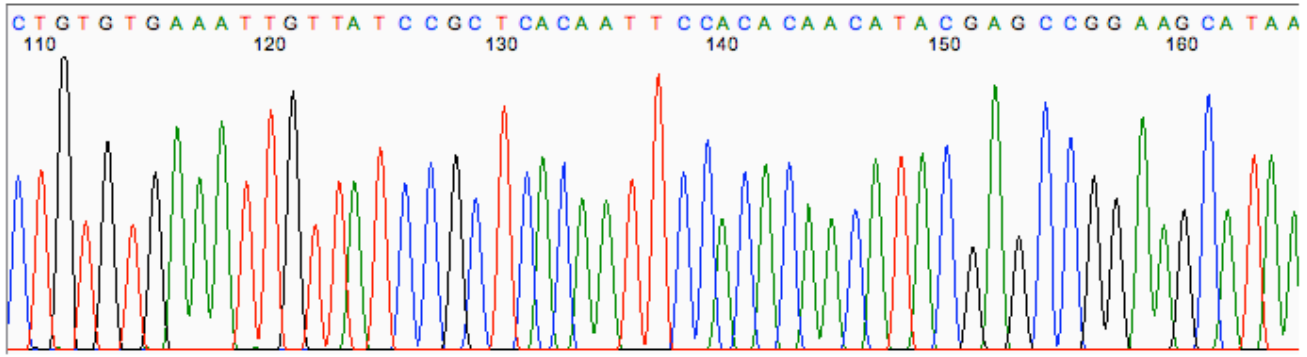


- ✓ Map reads (red) to the genome. Whole pieces of DNA are black.
- ✓ Count # of reads mapping to each DNA base → signal

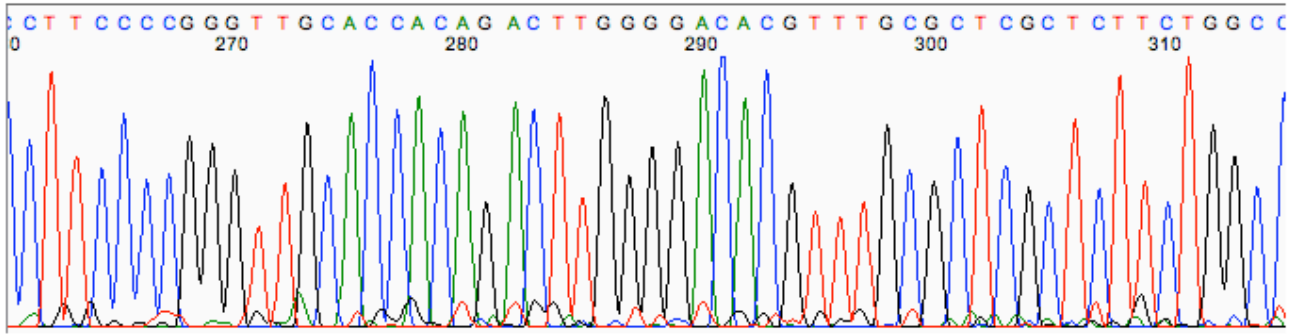


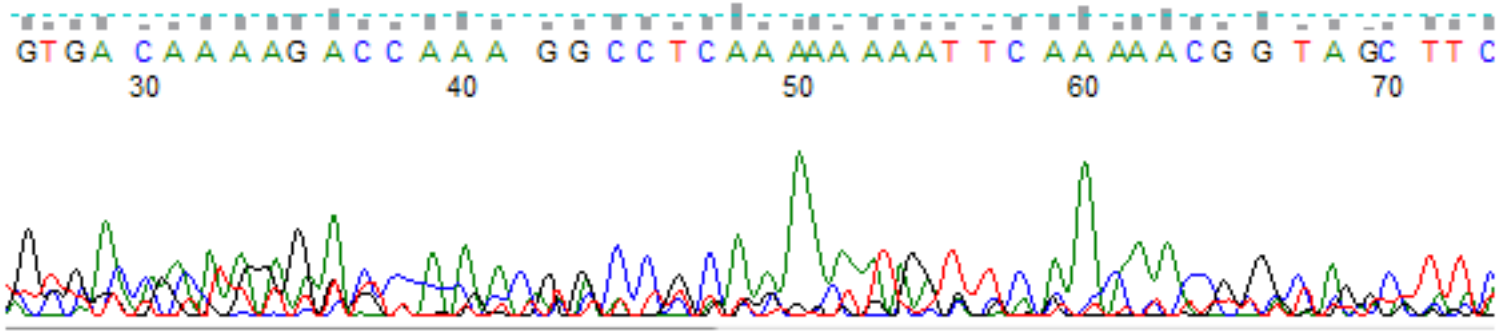
Noisy or Weak DNA Sequencing Lanes

Here's an example of excellent sequence. Note the evenly-spaced peaks and the lack of baseline 'noise' (see further down for examples of higher baseline noise):



The next example has a little baseline noise, but the 'real' peaks are still easy to call, so there's no problem with this sample:





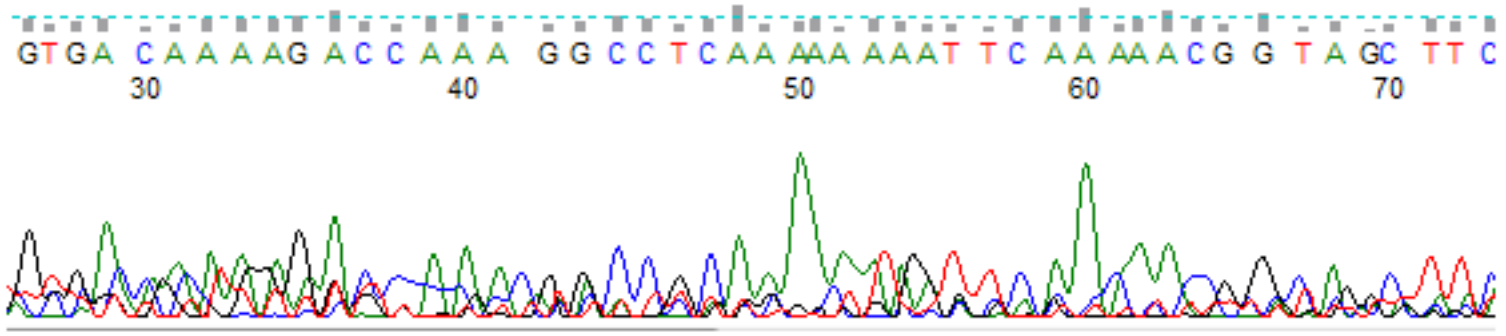
Noisy or Weak DNA Sequencing Lanes

Causes of weak signal lanes

Partially failed sequencing reaction.

Too much or too little DNA.

Partial loss of the sample during clean-up

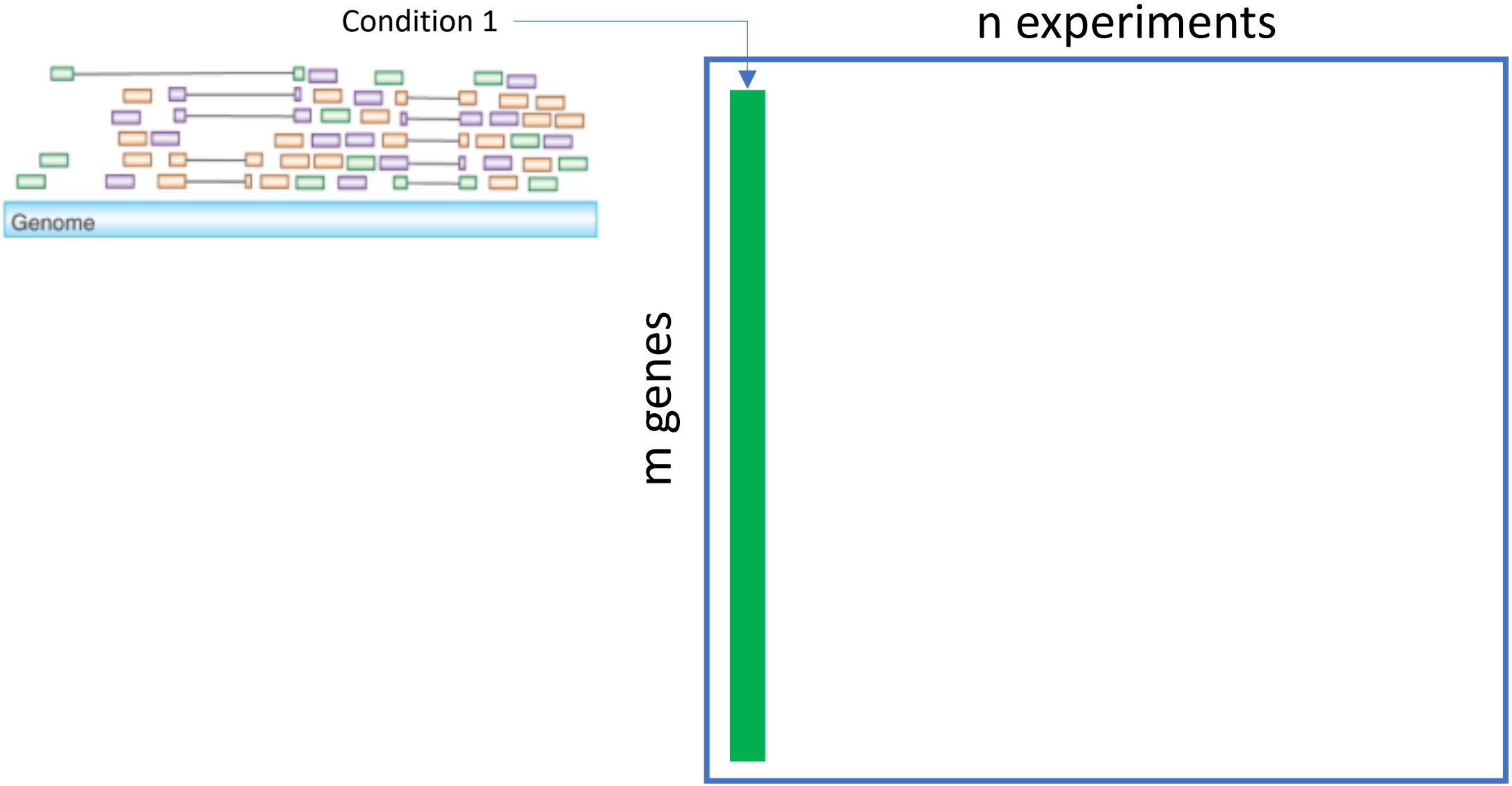


Expression analysis data matrix

Measure 20,000 genes in 100s of conditions

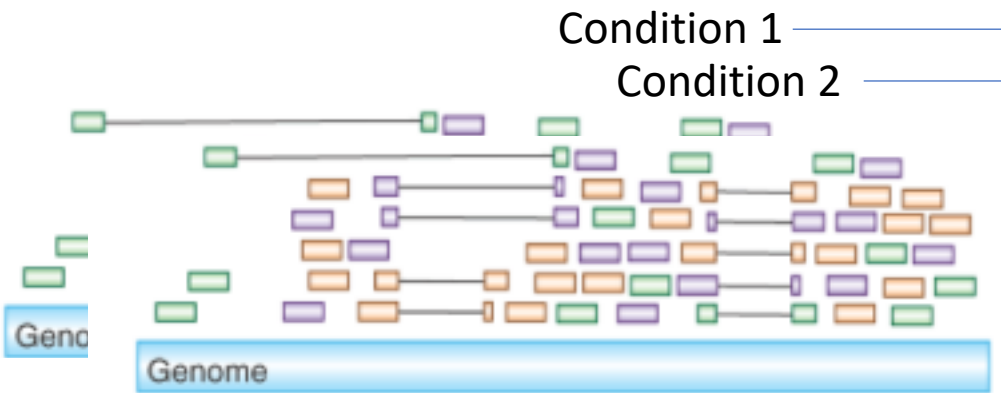
Expression analysis data matrix

Measure 20,000 genes in 100s of conditions



Expression analysis data matrix

Measure 20,000 genes in 100s of conditions



Condition 1

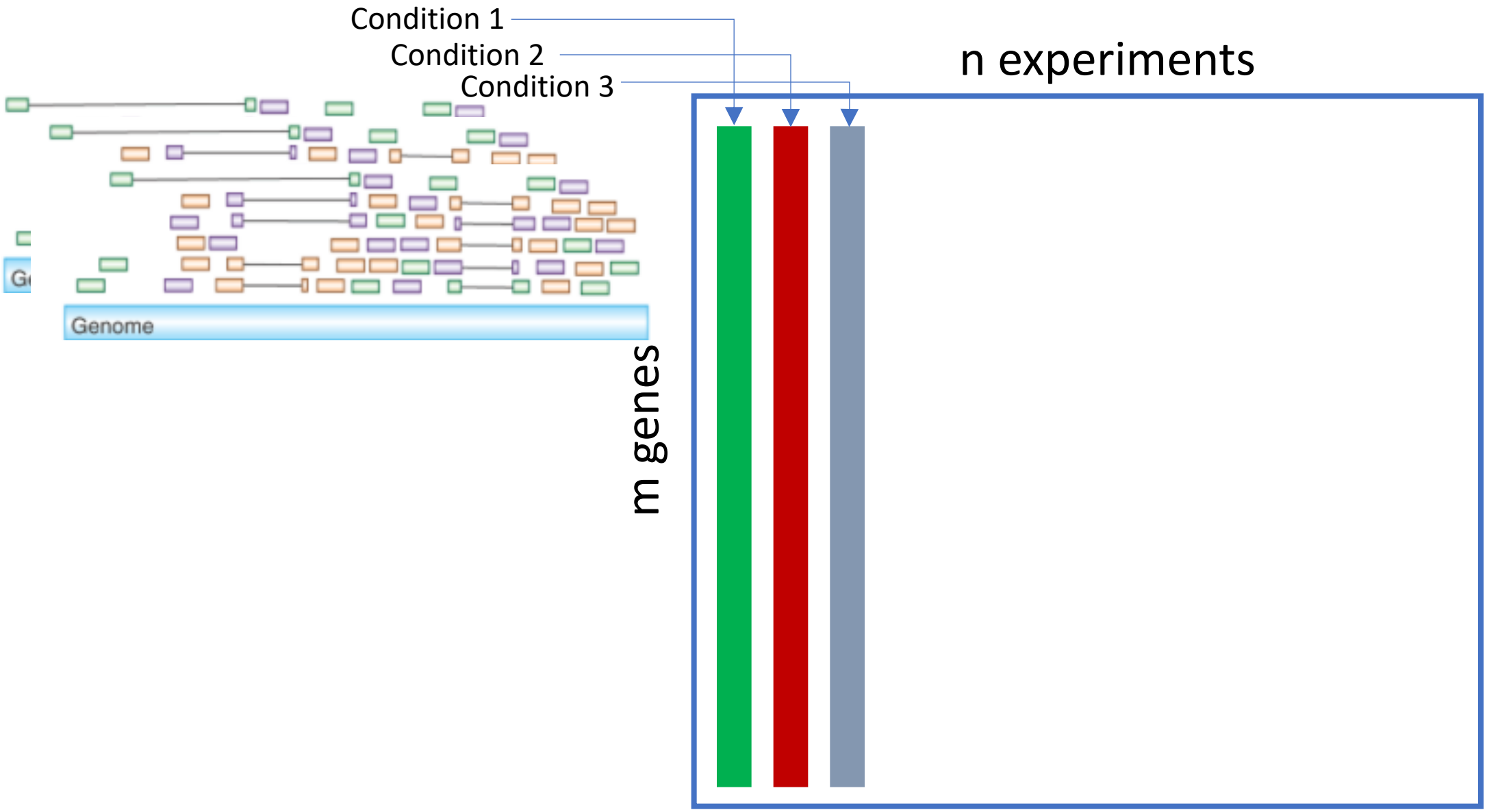
Condition 2

n experiments

m genes

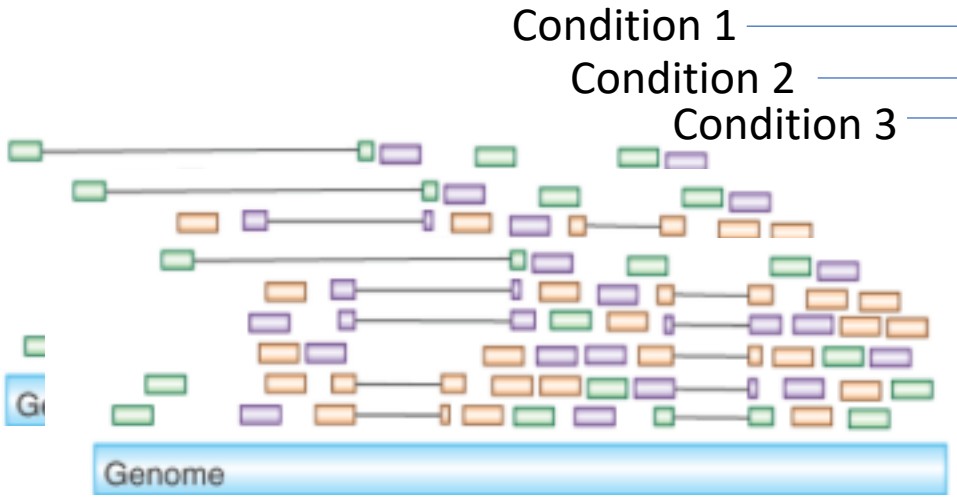
Expression analysis data matrix

Measure 20,000 genes in 100s of conditions



Expression analysis data matrix

Measure 20,000 genes in 100s of conditions



Condition 1
Condition 2
Condition 3

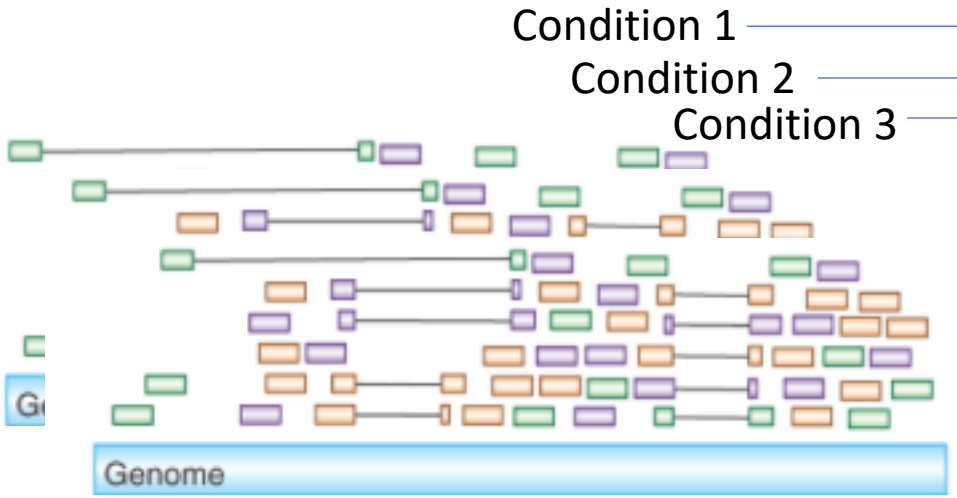
n experiments

m genes

Expression profile of a gene

Expression analysis data matrix

Measure 20,000 genes in 100s of conditions



Condition 1
Condition 2
Condition 3

n experiments

m genes

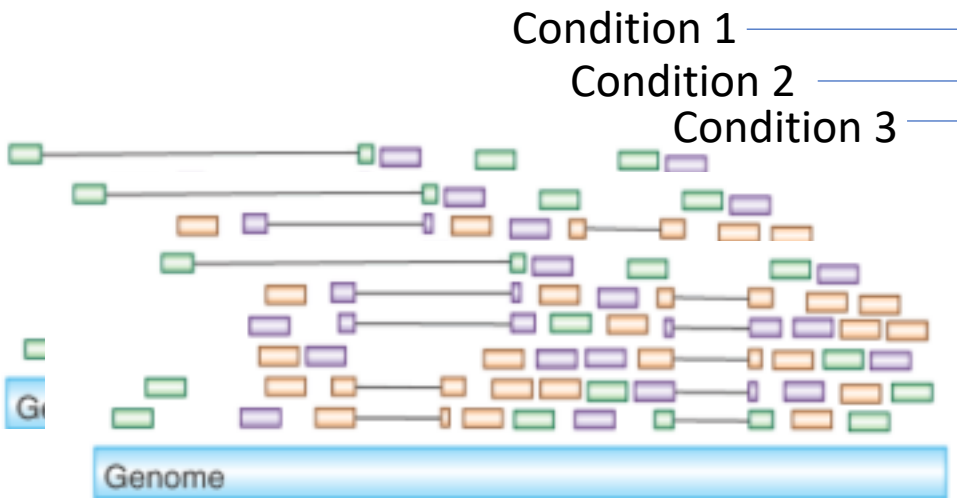
Expression profile of a gene

Expression profile of a gene

Expression profile of a gene

Expression analysis data matrix

Measure 20,000 genes in 100s of conditions



Condition 1
Condition 2
Condition 3

n experiments

m genes

Expression profile of a gene

Expression profile of a gene

Expression profile of a gene

Gene similarity

Expression analysis data matrix

Measure 20,000 genes in 100s of conditions

Condition 1
Condition 2
Condition 3

n experiments

m genes

Expression profile of a gene

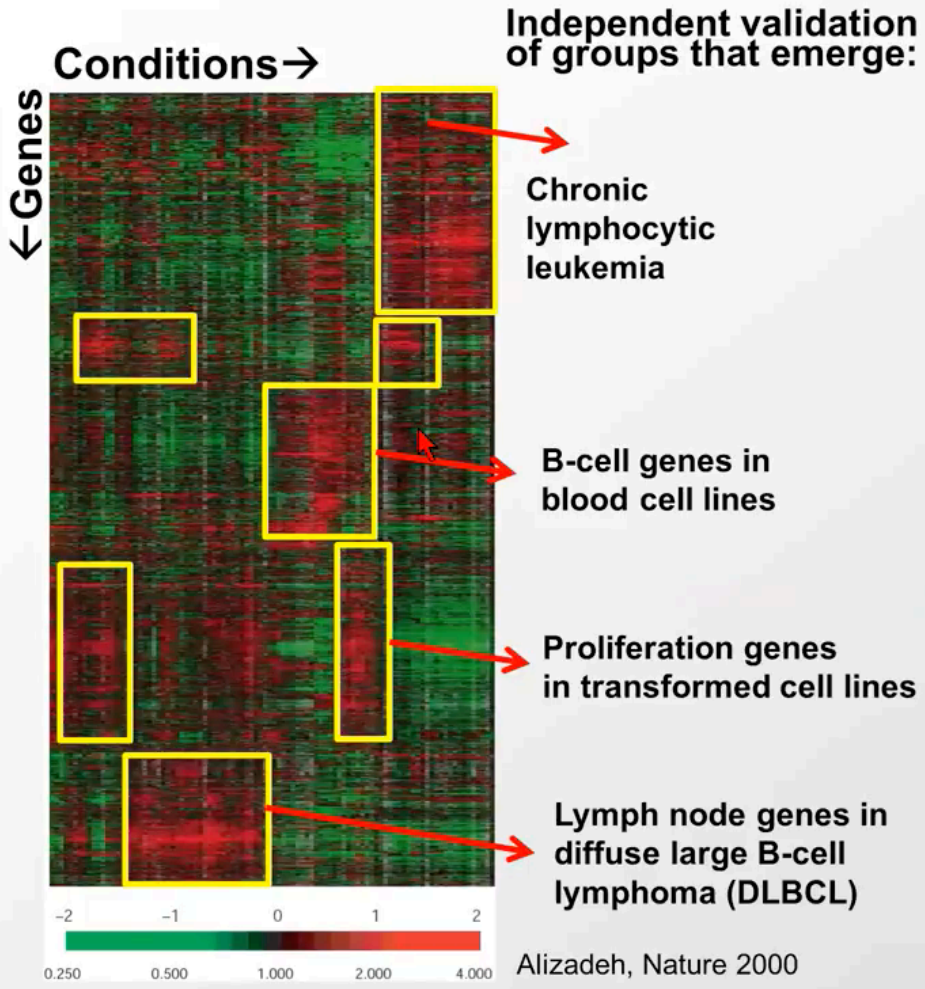
Expression profile of a gene

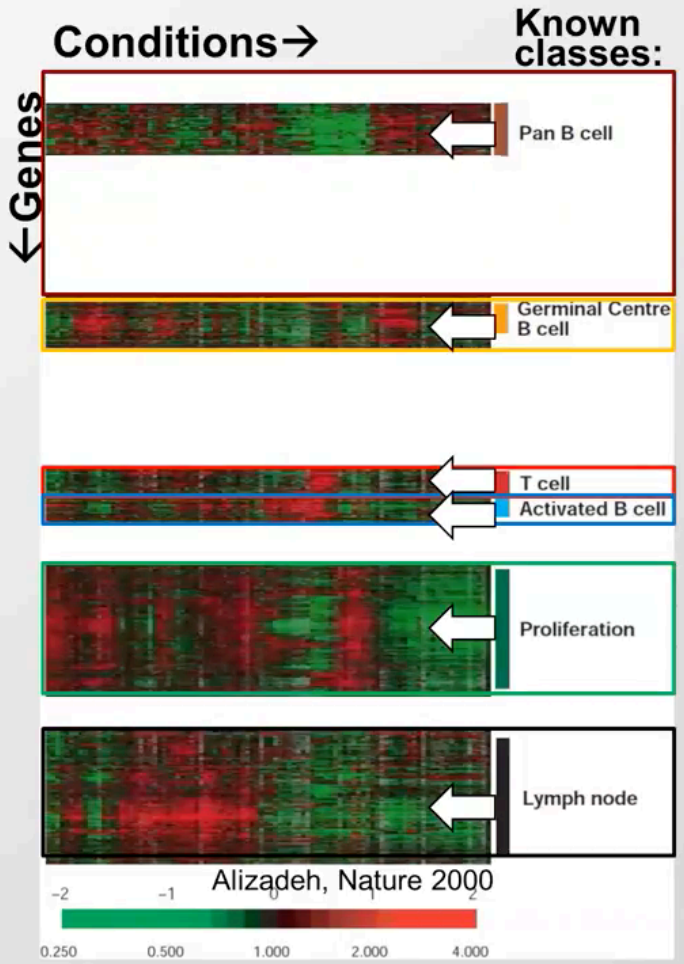
Expression profile of a gene

Gene similarity ?

Experiment (Condition) similar?

Clustering vs Classification





Goal of Clustering: **Group similar items** that likely come from the same category, and in doing so **reveal hidden structure**

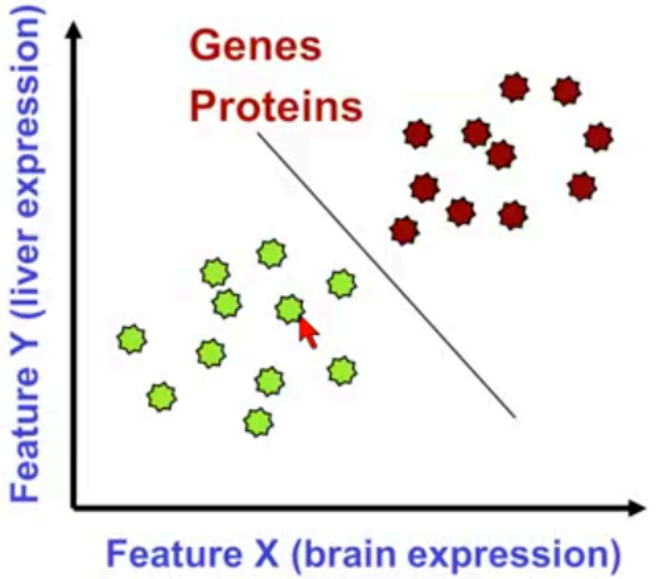
- **Unsupervised learning**

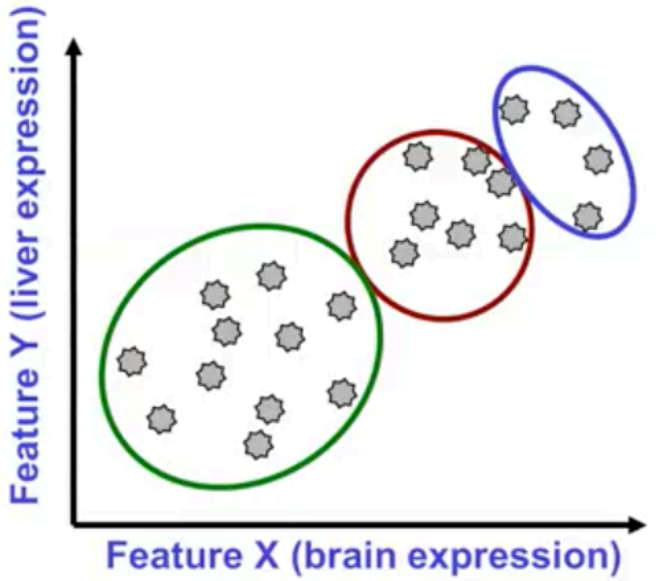
Goal of Classification: Extract features from the data that best **assign new elements** to ≥1 of **well-defined classes**

- **Supervised learning**

Clustering vs Classfication

- Objects characterized by one or more features
- **Classification (supervised learning)**
  - Have labels for some points
  - Want a "rule" that will accurately assign labels to new points
  - Sub-problem: Feature selection
  - Metric: Classification accuracy
- **Clustering (unsupervised learning)**
  - No labels
  - Group points into clusters based on how "near" they are to one another
  - Identify structure in data
  - Metric: independent validation features





Two approaches to clustering

- Partitioning (e.g. k-means)
  - Divides objects into non-overlapping clusters such that each data object is in exactly one subset
- Agglomerative (e.g. hierarchical clustering)
  - A set of nested clusters organized as a hierarchy

The Basic Idea

- Assume a fixed number K of clusters
- Partition points into K compact clusters

The Algorithm

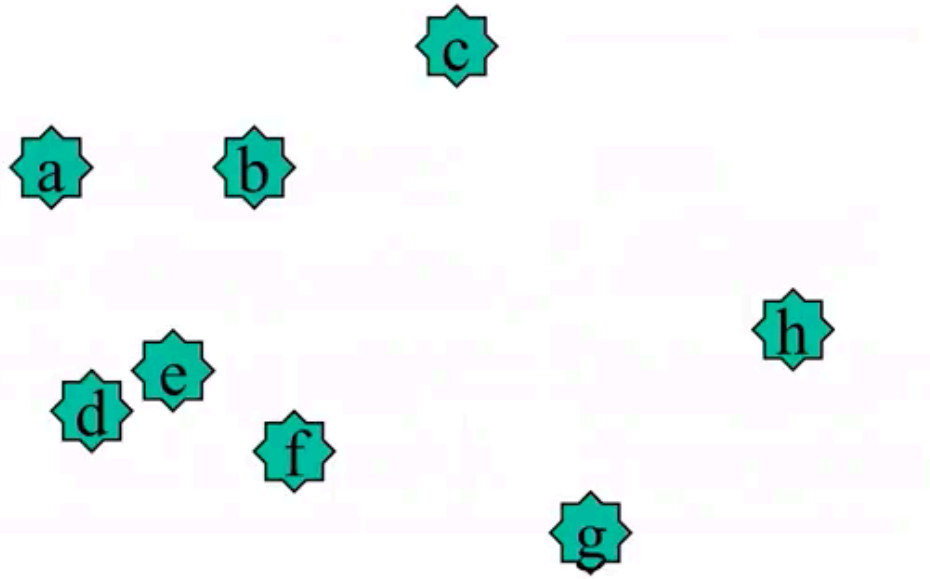
- Initialize K cluster centers randomly
- Repeatedly:
  - Assign points to nearest center
  - Move centers to center of gravity of their points
- Stop at convergence (no more reassignments)

Challenge of K-means: picking K

- How do we select K?
  - We can always make clusters “more compact” by increasing K
  - e.g. What happens is if K=number of data points?
  - What is a meaningful improvement?
- Hierarchical clustering side-steps this issue

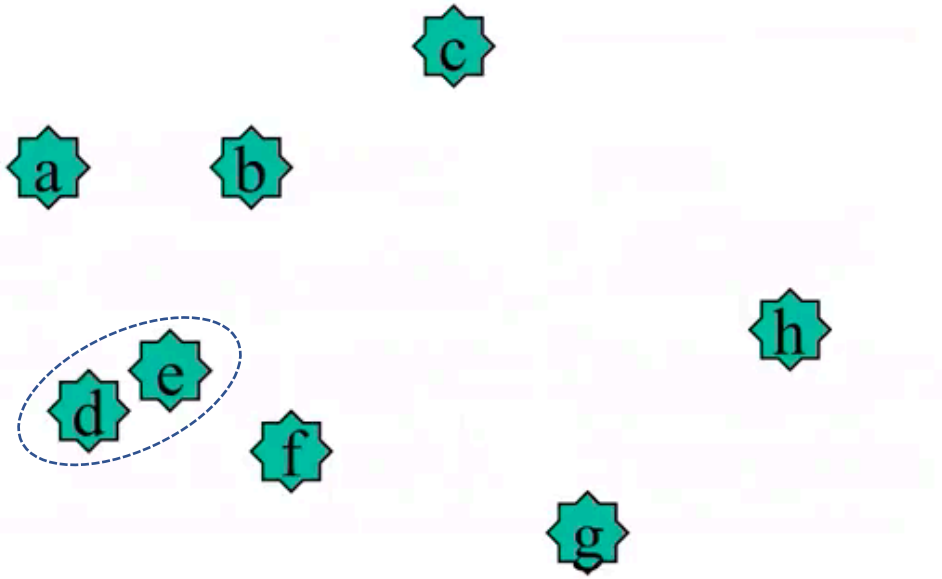
Most widely used algorithm for expression data

- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



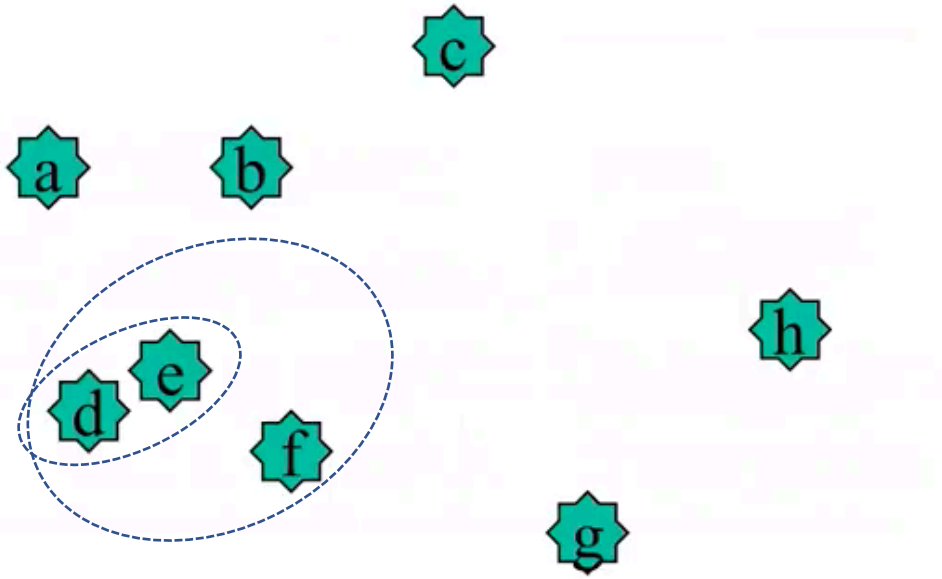
Most widely used algorithm for expression data

- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



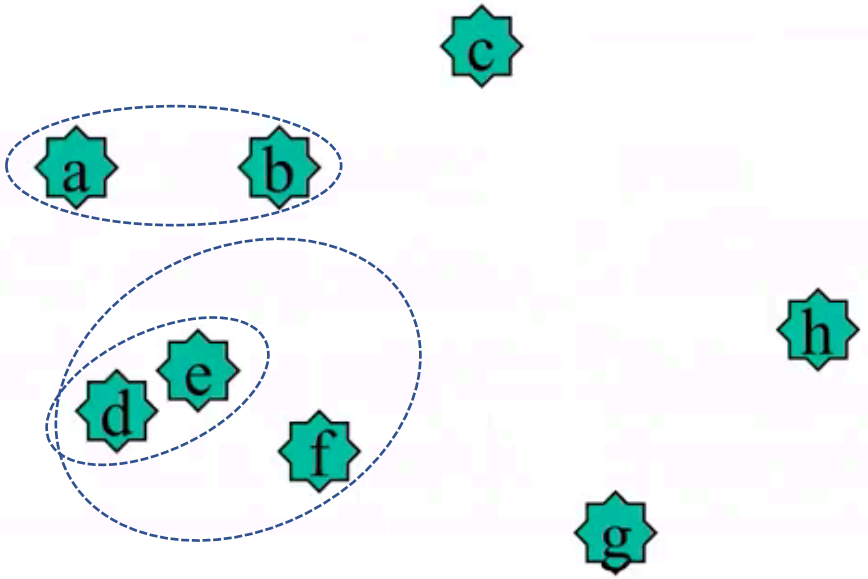
Most widely used algorithm for expression data

- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



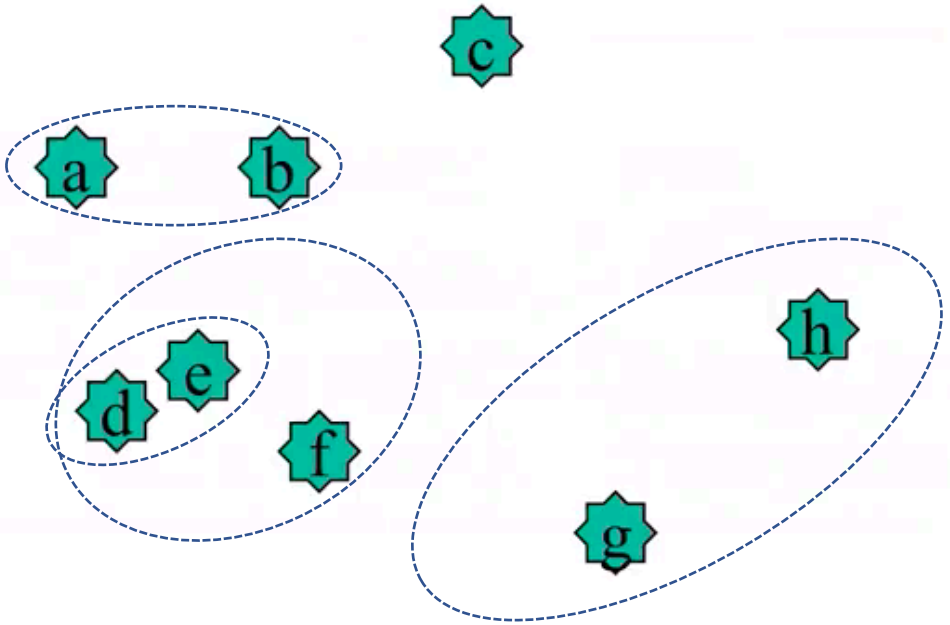
Most widely used algorithm for expression data

- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



Most widely used algorithm for expression data

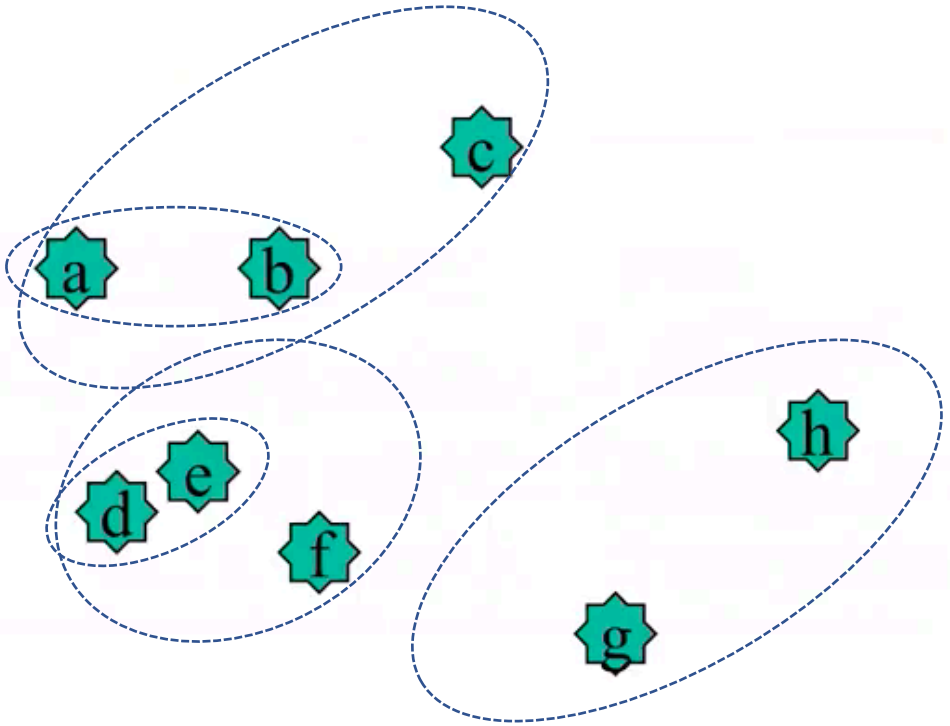
- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



Hierachical clustering

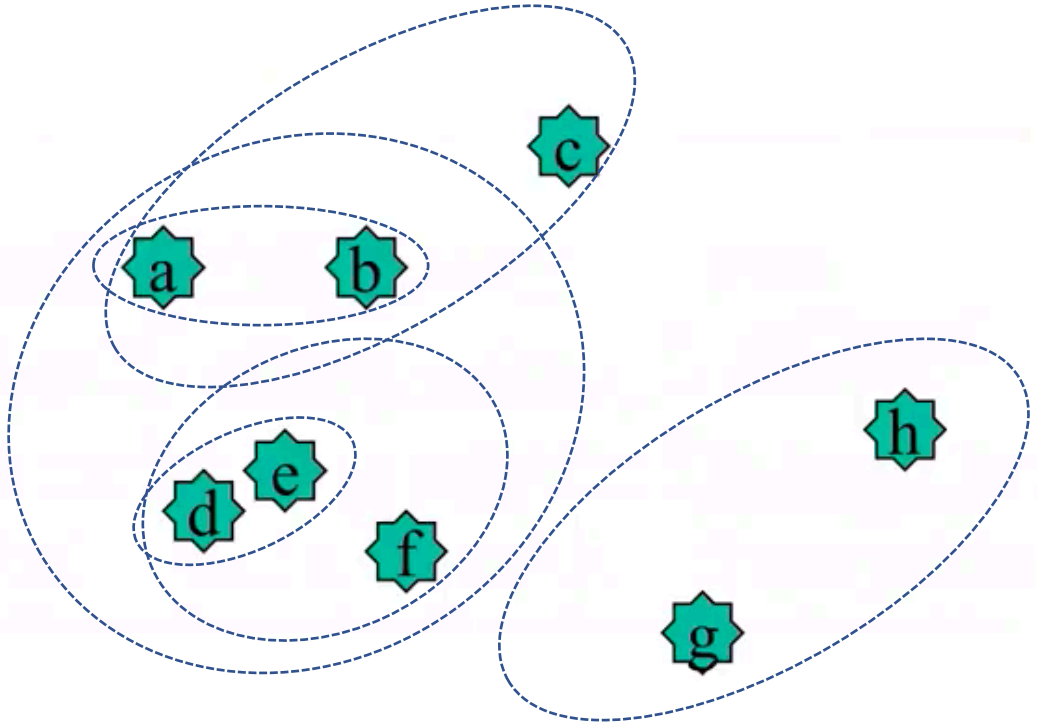
Most widely used algorithm for expression data

- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



Most widely used algorithm for expression data

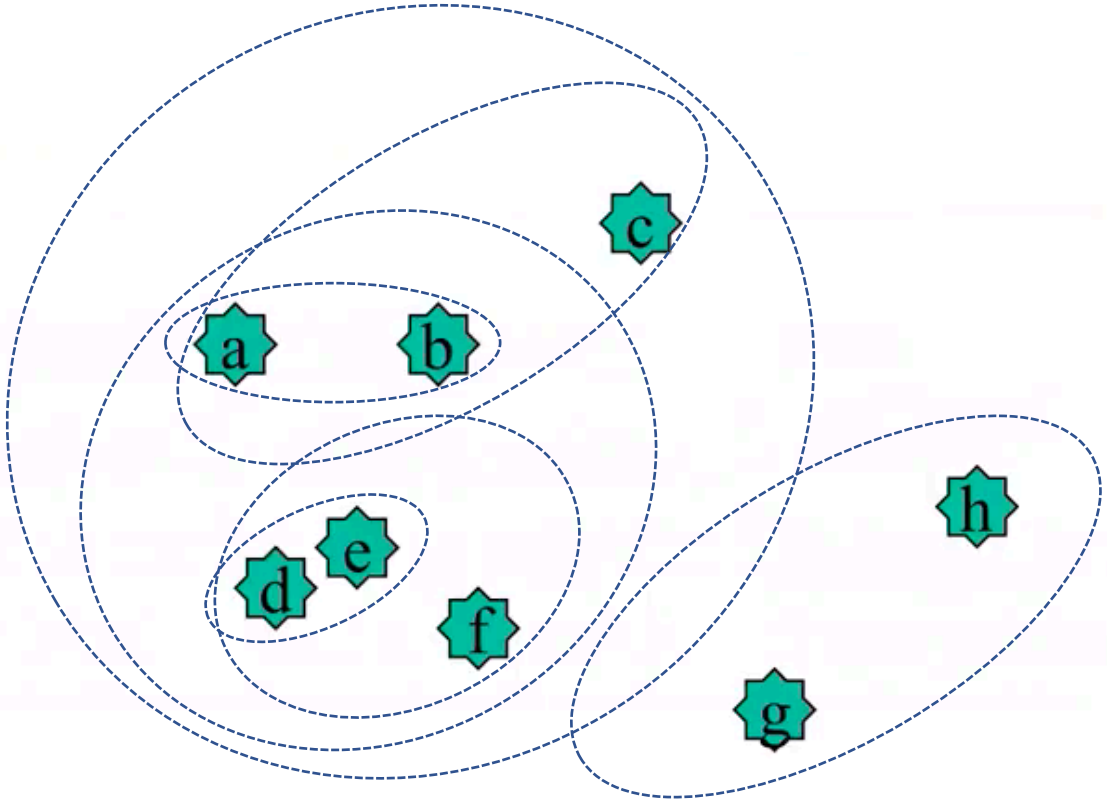
- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



Hierachical clustering

Most widely used algorithm for expression data

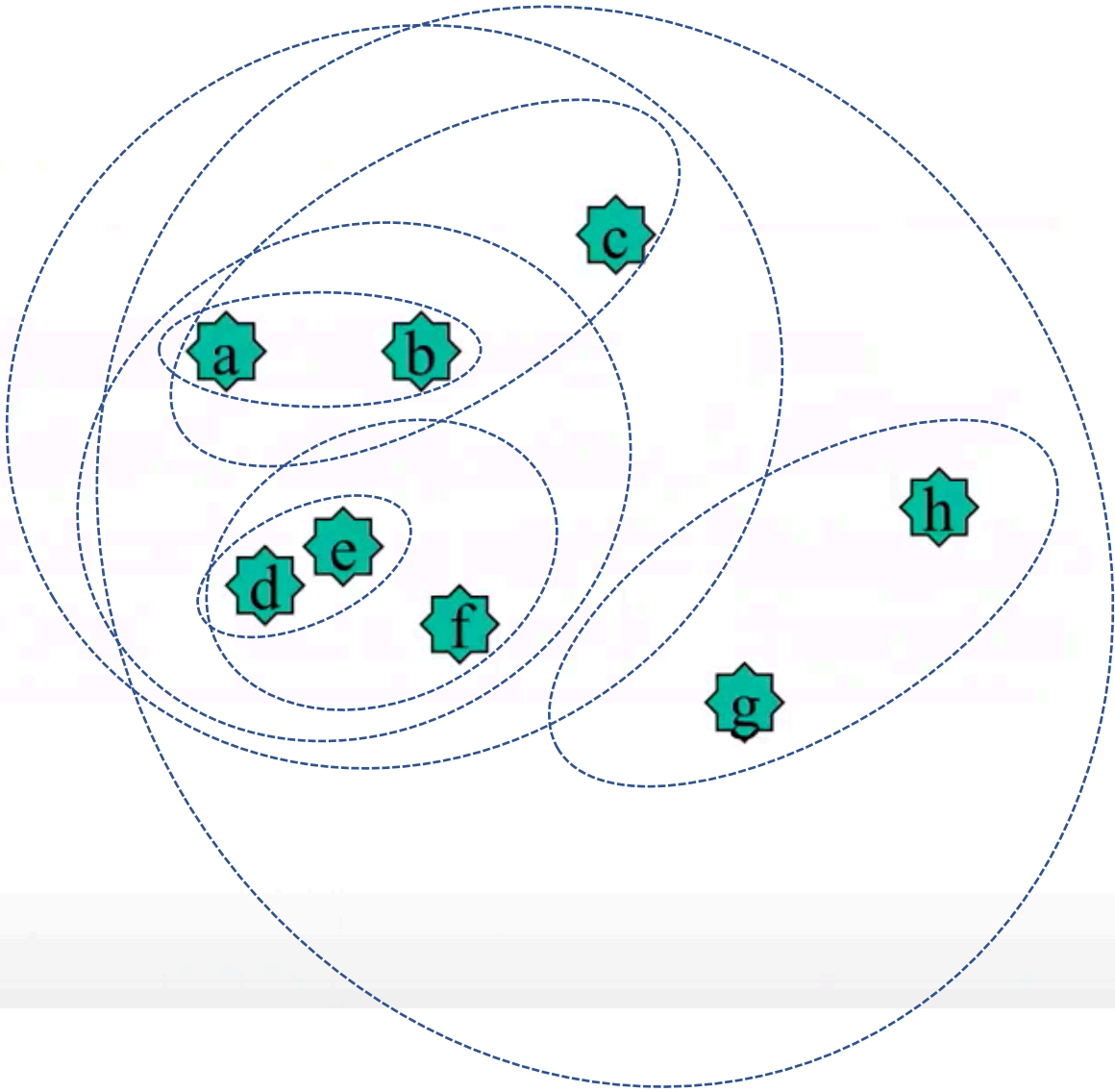
- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge



Hierachical clustering

Most widely used algorithm for expression data

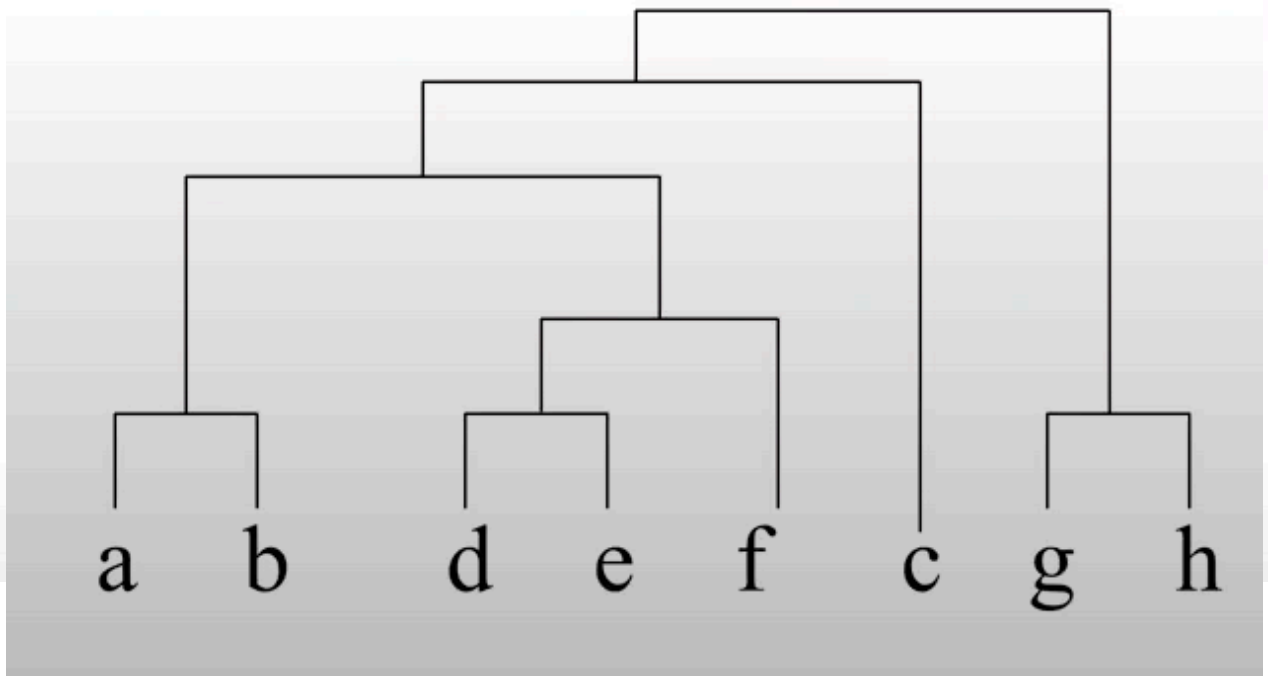
- Start with each point in a separate cluster
- At each step:
  - Choose the pair of **closest clusters**
  - Merge

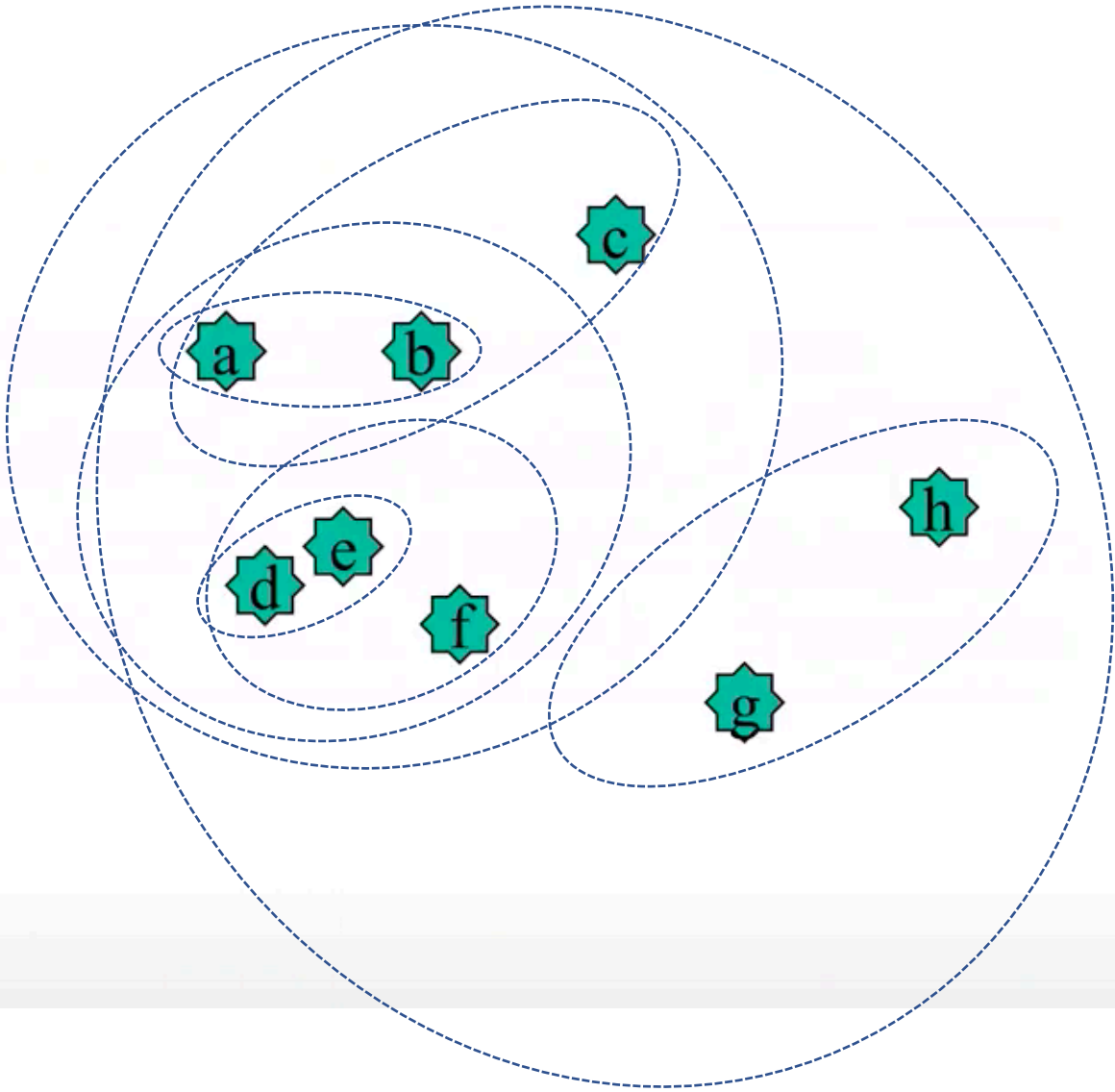


Hierachical clustering

Most widely used algorithm for expression data

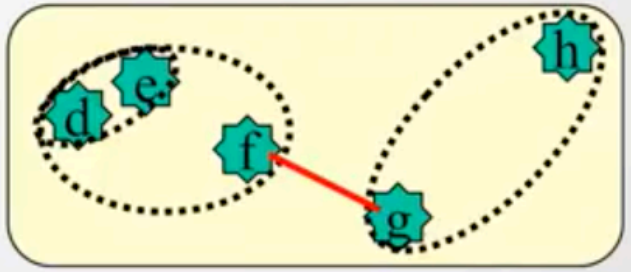
-
-



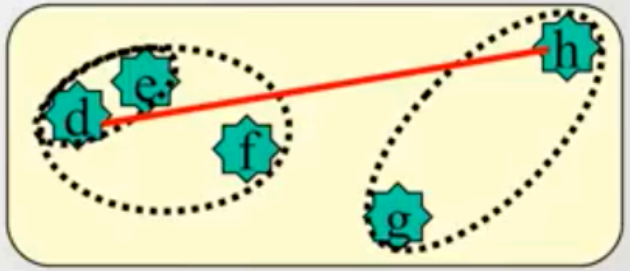


Distance between clusters

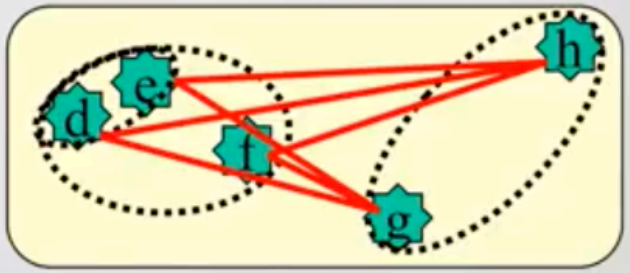
- $CD(X,Y)=\min_{x \in X,\, y \in Y} D(x,y)$
  Single-link method



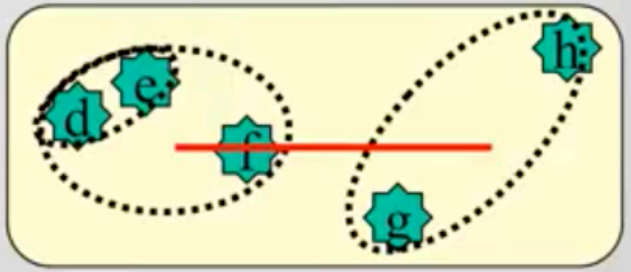
- $CD(X,Y)=\max_{x \in X,\, y \in Y} D(x,y)$
  Complete-link method



- $CD(X,Y)=\text{avg}_{x \in X,\, y \in Y} D(x,y)$
  Average-link method



- $CD(X,Y)=D(\text{avg}(X), \text{avg}(Y))$
  Centroid method



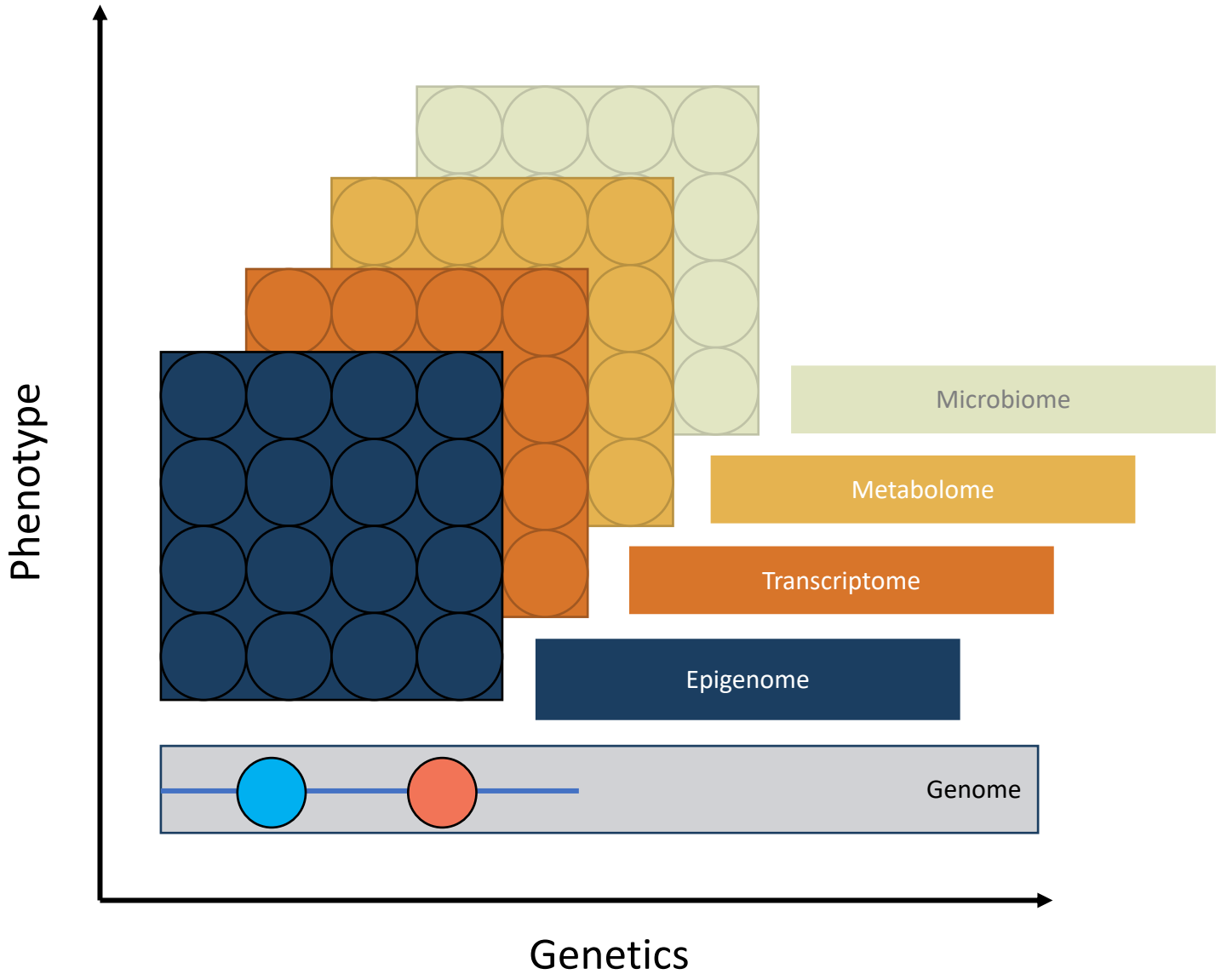
Transcriptomics

Transcriptomics examines RNA levels genome-wide, both qualitatively (which transcripts are present, identification of novel splice sites, RNA editing sites) and quantitatively (how much of each transcript is expressed). The central dogma of biology viewed RNA as a molecular intermediate between DNA and proteins, which are considered the primary functional read-out of DNA.

An even more significant contribution of these studies was the development of the non-coding RNA field. It is now clear that thousands of long non-coding RNAs transcribed in mammalian cells (http://www.gencodegenes.org/) play essential roles in many physiological processes, for example, brown adipose differentiation (Cell Metab. 2015;21:764–76), endocrine regulation (Nat Rev Endocrinol. 2015;11:151–60), and neuron development (Nat Rev Neurosci. 2016;17:537–49).

Dysregulation of long non-coding RNAs had been implicated in various diseases, such as myocardial infarction(J Hum Genet. 2006;51:1087–99),diabetes(Genes Dev. 2016;30:502–7), cancer (Nature. 2010;464:1071–6), and others (Cell Mol Life Sci. 2016;73:2491–509).

Microbiome


Phenotype
Microbiome
Metabolome
Transcriptome
Epigenome
Genome
Genetics

Standardize Microbiome Terms

Marchesi and Ravel *Microbiome* (2015) 3:31
DOI 10.1186/s40168-015-0094-5

Microbiome

EDITORIAL Open Access

The vocabulary of microbiome research: a proposal

Julian R. Marchesi[1,2] and Jacques Ravel[3,4*]

Microbiome

This term refers to the entire habitat, including the micro- organisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions. This definition is based on that of “biome,” the biotic and abiotic factors of given environments. Others in the field limit the definition of microbiome to the collection of genes and genomes of members of a microbiota. It is argued that this is the definition of metagenome, which combined with the environment constitutes the microbiome. The microbiome is characterized by the application of one or combinations of metagenomics, metabonomics, metatranscriptomics, and metaproteomics combined with clinical or environmental metadata.

Metagenome

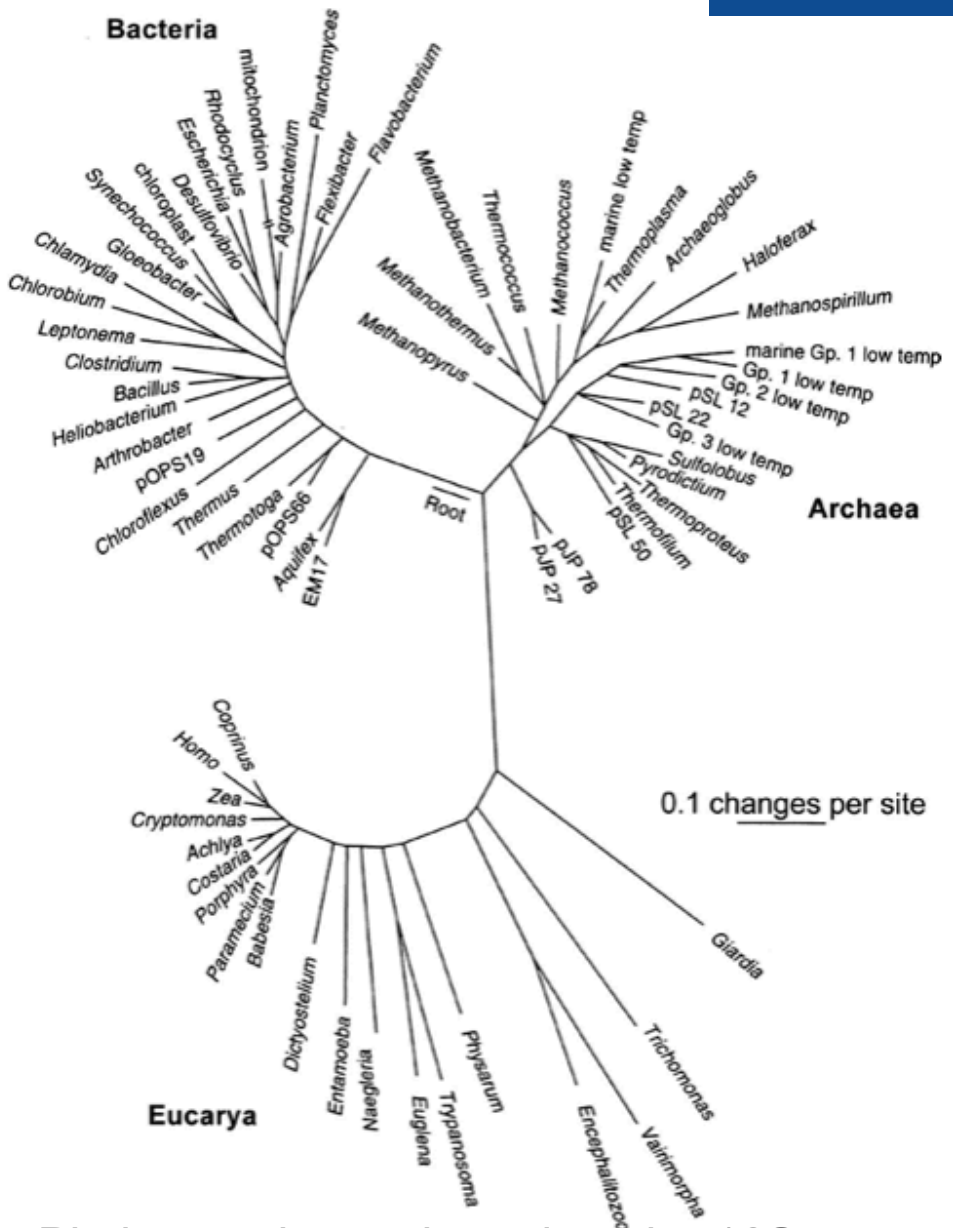
The collection of genomes and genes from the members of a microbiota. This collection is obtained through shot- gun sequencing of DNA extracted from a sample (metage- nomics) followed by assembly or mapping to a reference database followed by annotation. Metataxonomic analysis, because it relies on the amplification and sequencing of taxonomic marker genes, is not metagenomics. Metage- nomics is the process used to characterize the metagen- ome, from which information on the potential function of the microbiota can be gained.

Metataxonomics

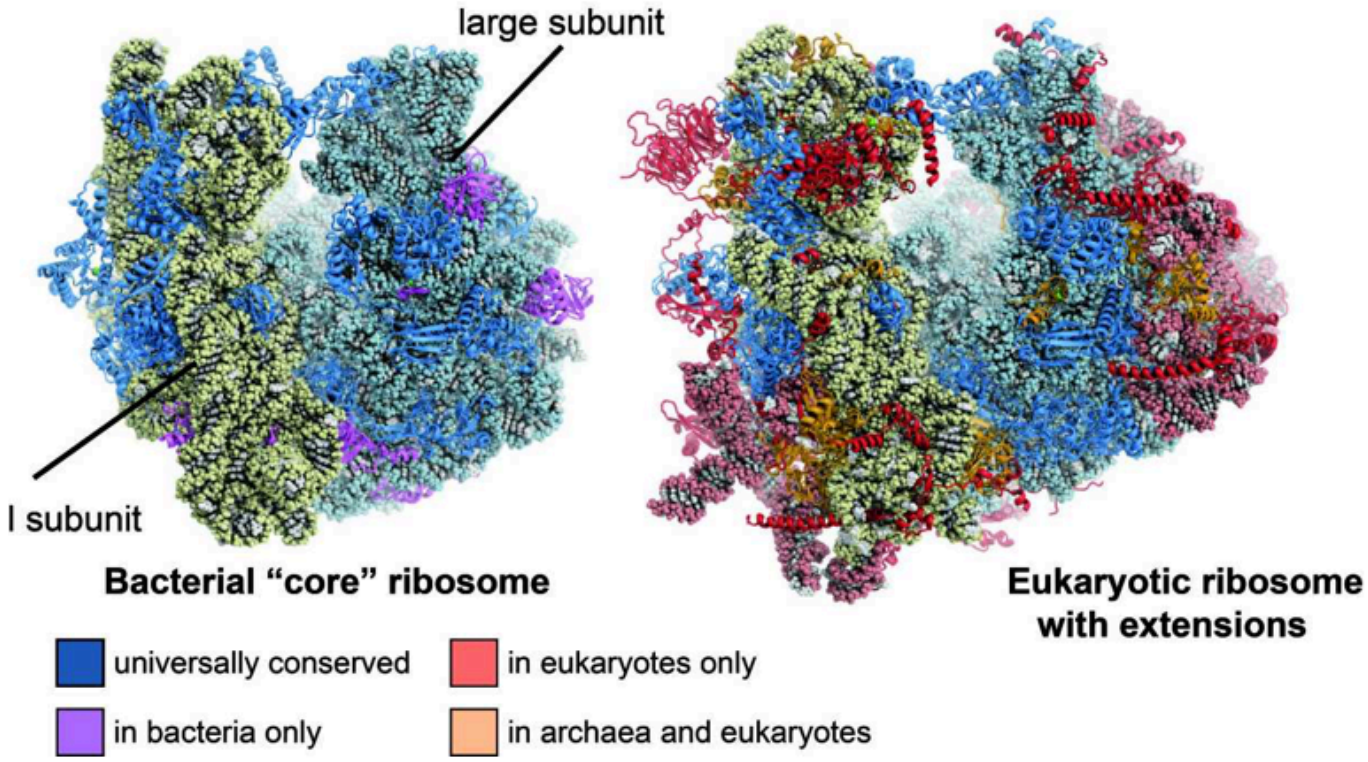
Metataxonomics is a term we propose and define as the high-throughput process used to characterize the entire microbiota and create a metataxonomic tree, which shows the relationships between all sequences obtained. While viruses are an integral part of the microbiota, no universal viral marker genes are available to perform such taxonomic assignments.

Microbiota. The assemblage of microorganisms present in a defined environment

Core marker genes are specific taxonomic clades

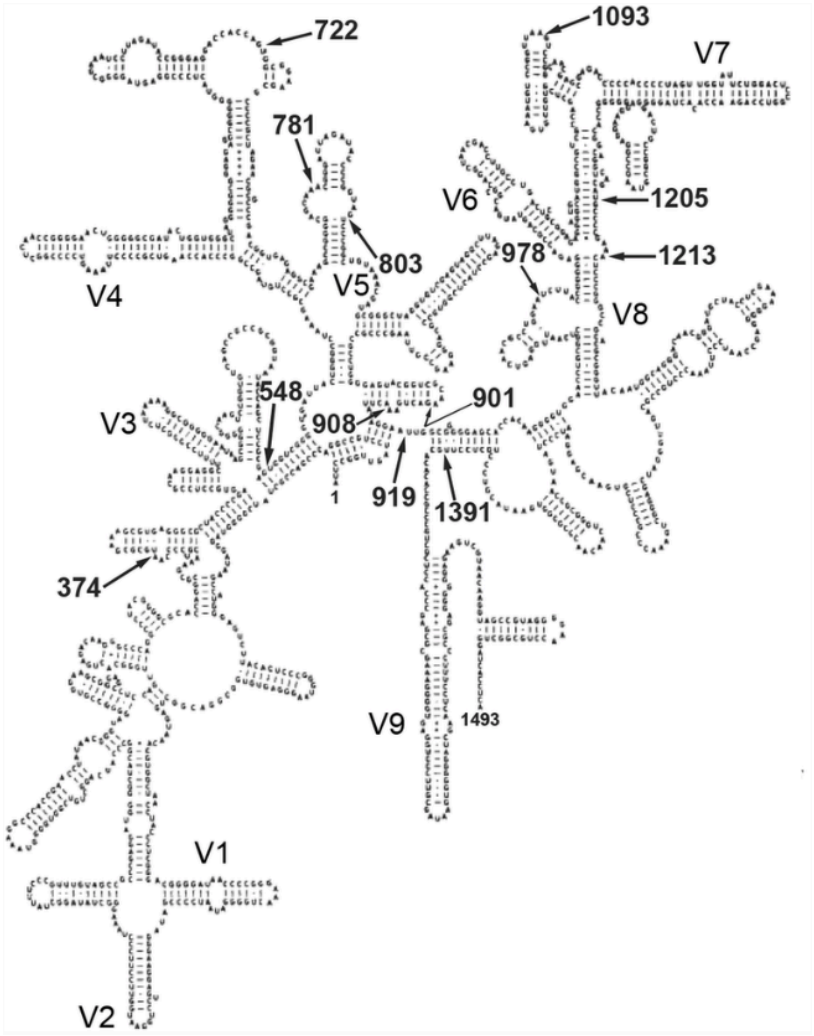


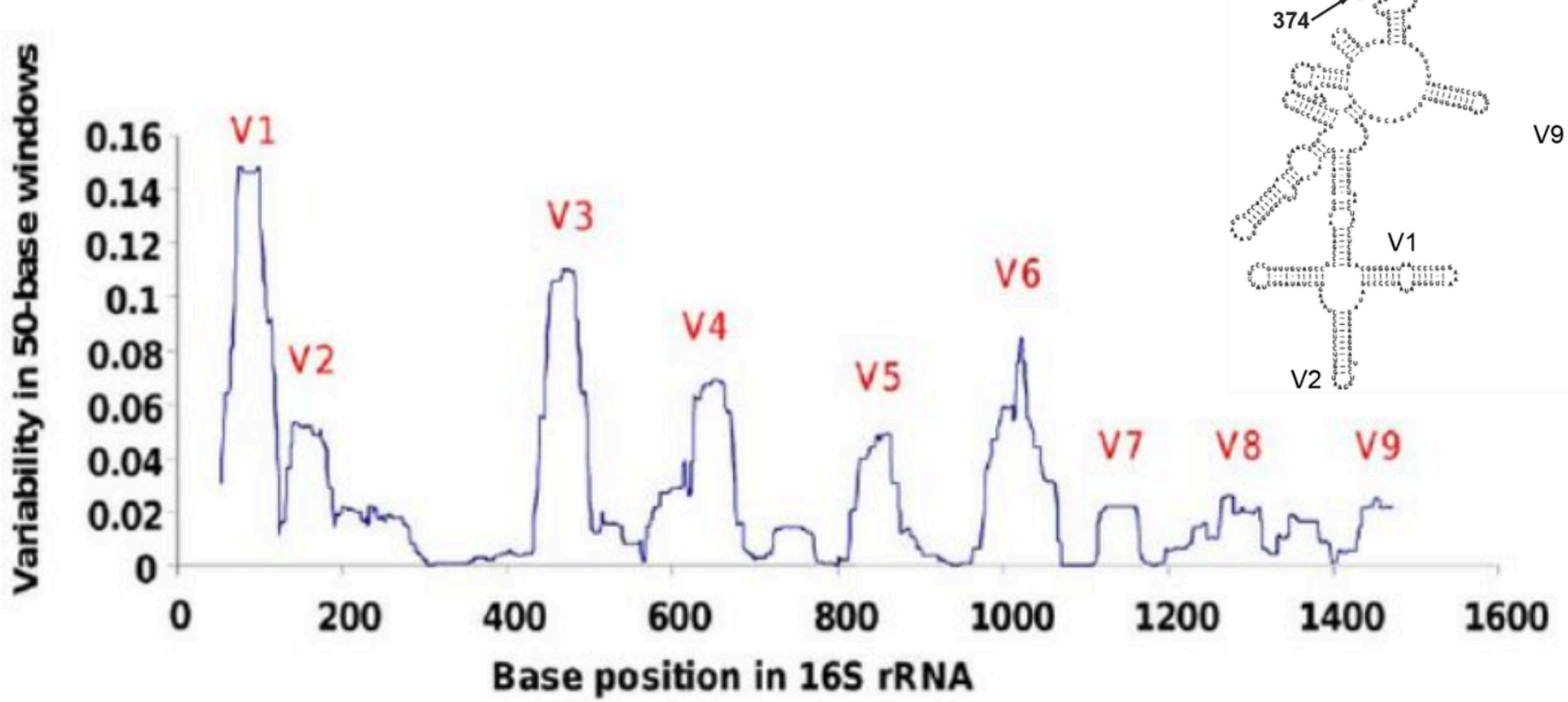
Phylogenetic tree based on the 16S rRNA gene sequence comparisons

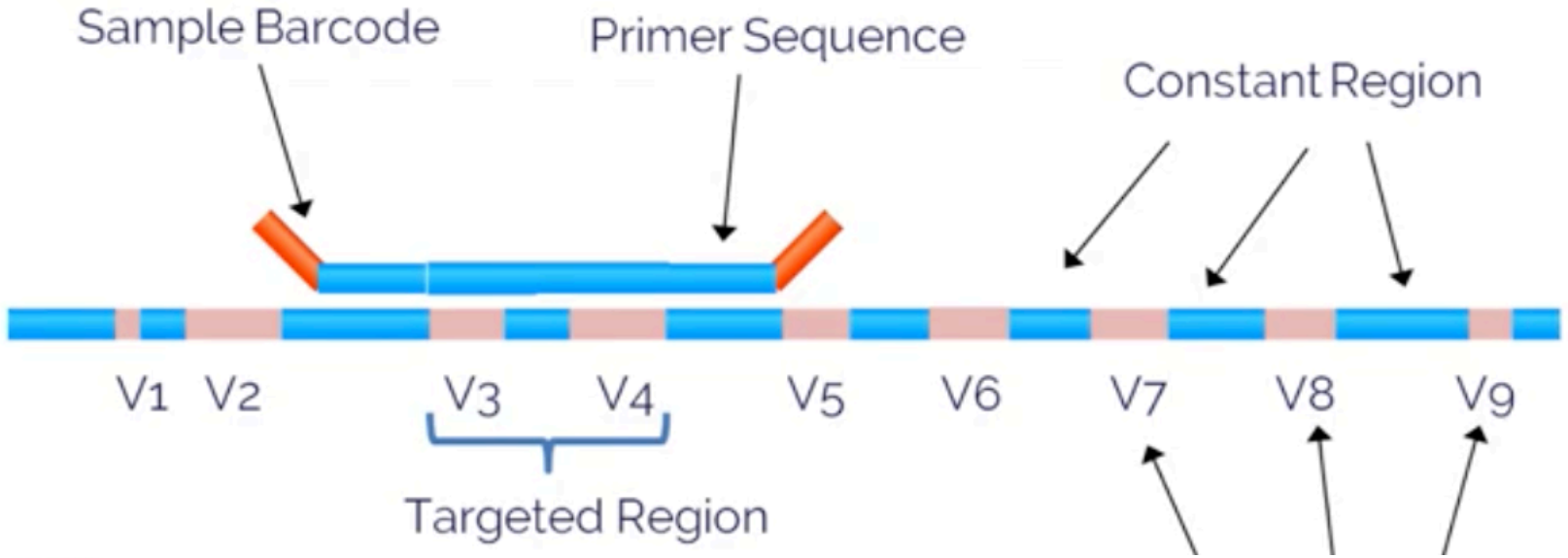


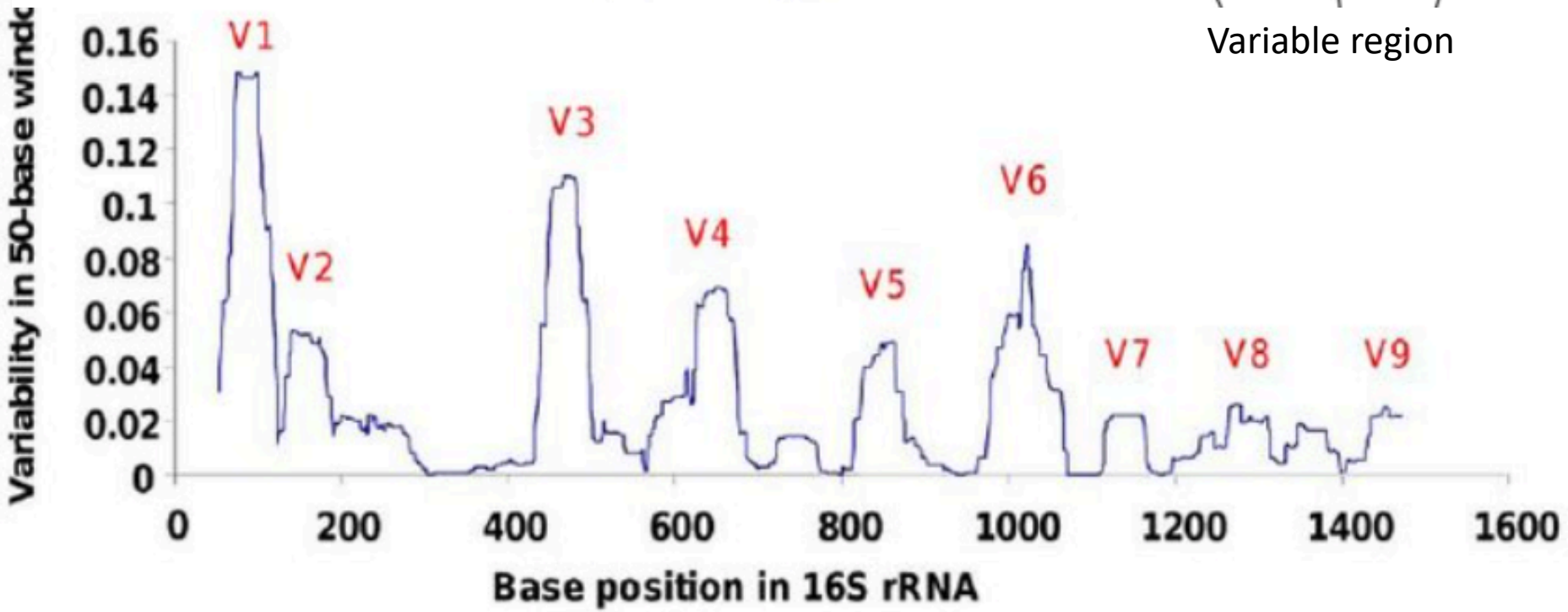
Bacterial diversity: 16S rRNA gene

Universal and variable regions of 16S rRNA used for PCR amplification and classification

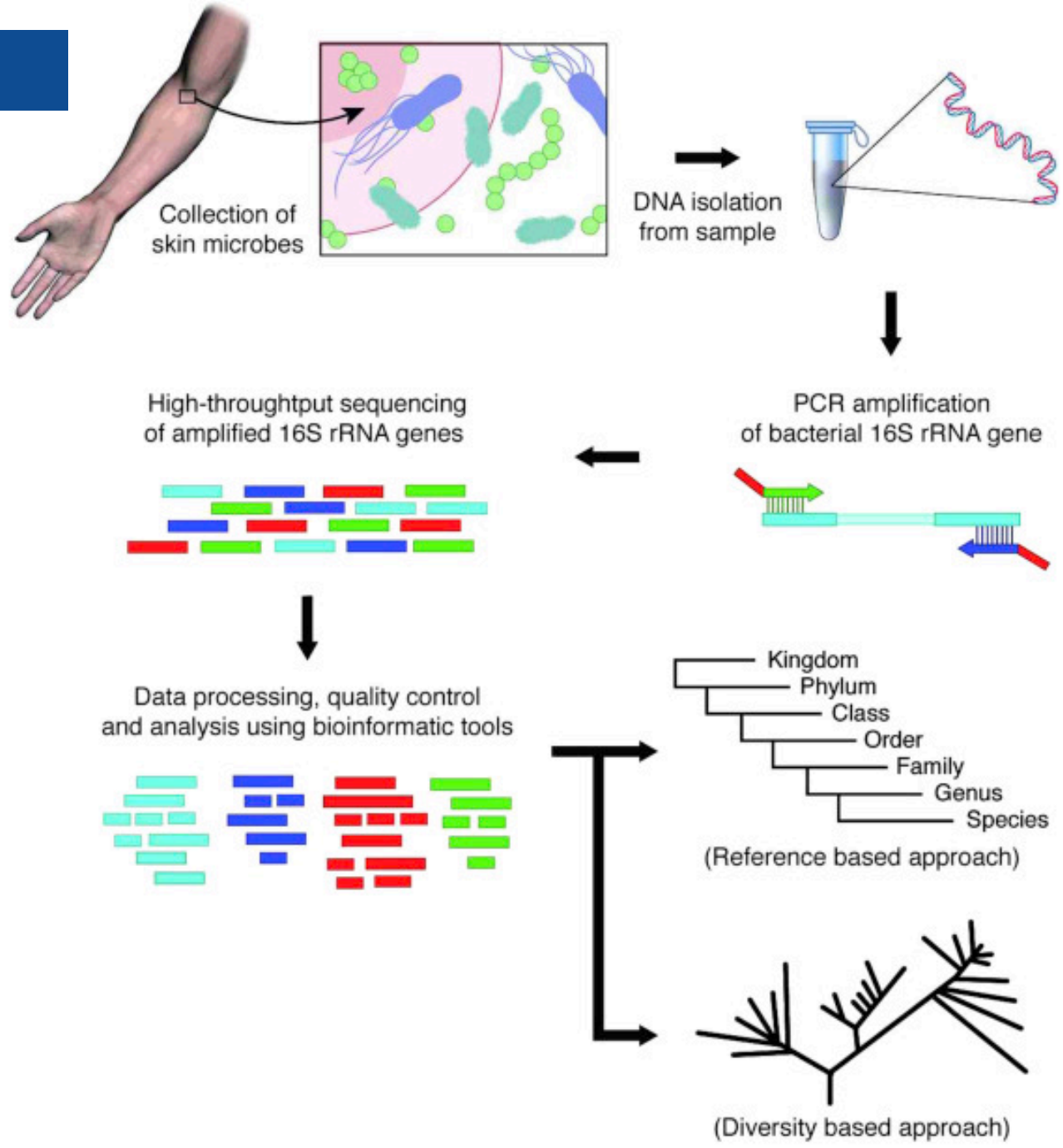





Sample Barcode
Primer Sequence
Constant Region
V1 V2 V3 V4 V5 V6 V7 V8 V9
Targeted Region
Variable region


Variability in 50-base windo
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
V1
V2
V3
V4
V5
V6
V7
V8
V9
0
200
400
600
800
1000
1200
1400
1600
Base position in 16S rRNA

Basic workflow: 16S rRNA gene

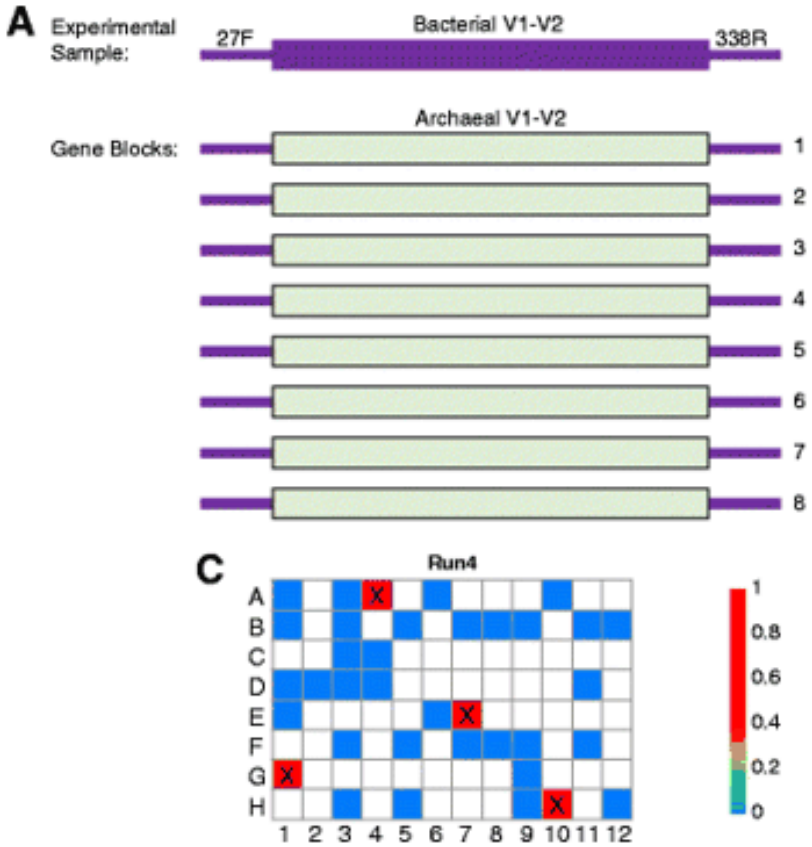


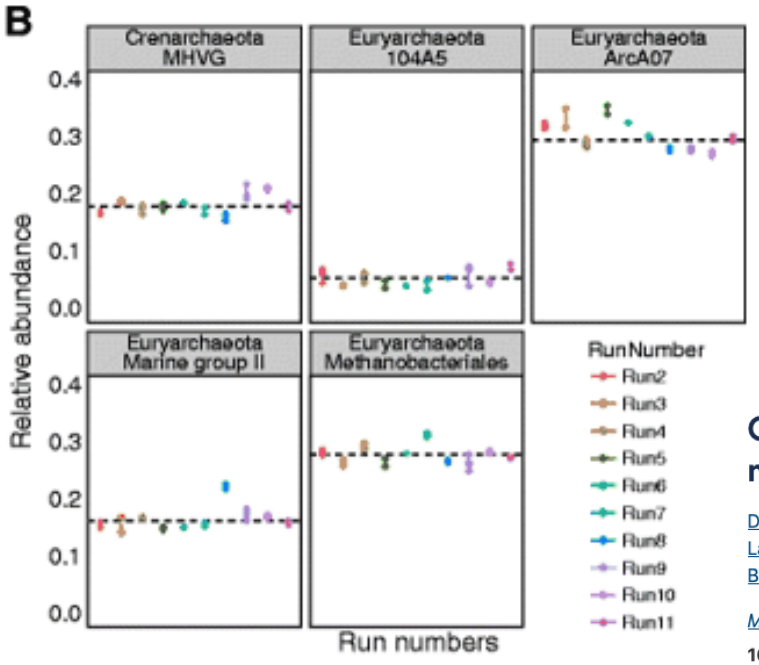
Importance of Control

Negative control: e.g. air swab processed through entire DNA prep, sequencing analysis

Positive control: e.g. Mock community (purchased from BEI or made in house). Do not underestimate primer bias or chimeras.

Mock community contains 20 bacteria. Amplified and sequenced with various methods





Optimizing methods and dodging pitfalls in microbiome research

Dorothy Kim, Casey E. Hofstaedter, Chunyu Zhao, Lisa Mattei, Ceylan Tanes, Erik Clarke, Abigail Lauder, Scott Sherrill-Mix, Christel Chehoud, Judith Kelsen, Máire Conrad, Ronald G. Collman, Robert Baldassano, Frederic D. Bushman & Kyle Bittinger

Microbiome **5**, Article number: 52 (2017) | Cite this article

16k Accesses | **140** Citations | **69** Altmetric | Metrics

Suites of analysis tools

GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2009 Jul; 19(7): 1141–1152.
doi: 10.1101/gr.085464.108
PMCID: PMC3776646
PMID: 19383763

Microbial community profiling for human microbiome projects: Tools, techniques, and challenges

Micah Hamady[1] and Rob Knight[2,3]

STUDY DESIGNS

Experimental and analytical tools for studying the human microbiome

Justin Kuczynski[1], Christian L. Lauber[2], William A. Walters[1], Laura Wegener Parfrey[3], José C. Clemente[3], Dirk Gevers[4] and Rob Knight[3,5]

OTU table

OUT table is a sample-by-observation matrix

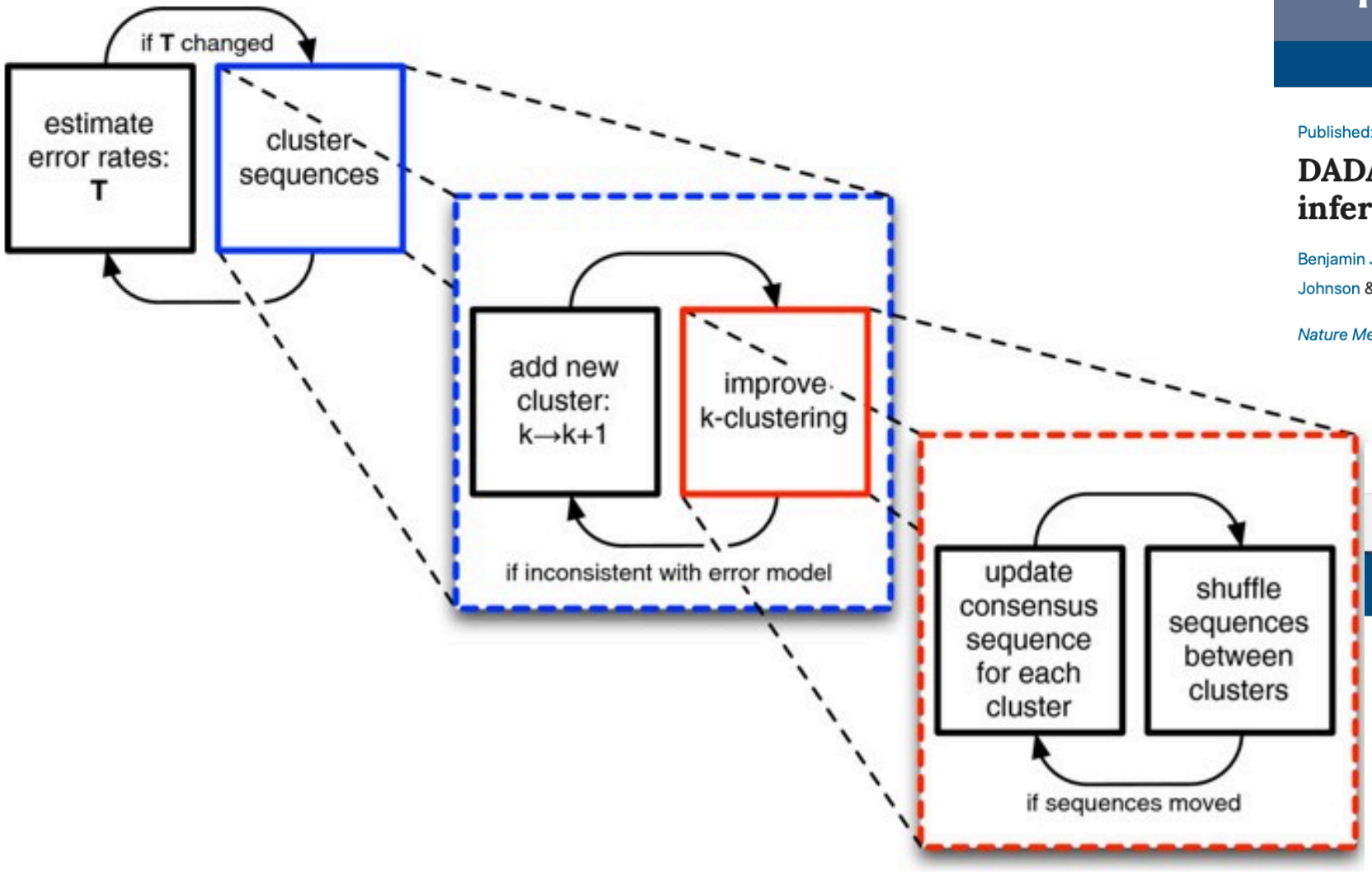
	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

Taxonomy

OTU can have corresponding taxonomy information

OTU table

OTUs were developed as a tool for classifying microbial species, but have also been repurposed to the task of correcting errors; the sequences within an OTU are typically interpreted as a taxonomic grouping without specifying whether the variation within an OTU represents errors or real diversity on a finer scale than that chosen to define the OTU. If the scale of the noise is smaller than that of the clusters, then the construction of OTUs will appropriately group error-containing sequences together with their true genotype.



DADA schematic. The basic structure of *DADA*, an algorithm to denoise amplicon sequence data. See Algorithm Algorithm 1 in the Methods section for the pseudocode and a more detailed description.

Published: 23 May 2016

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

Nature Methods **13**, 581–583(2016) | Cite this article

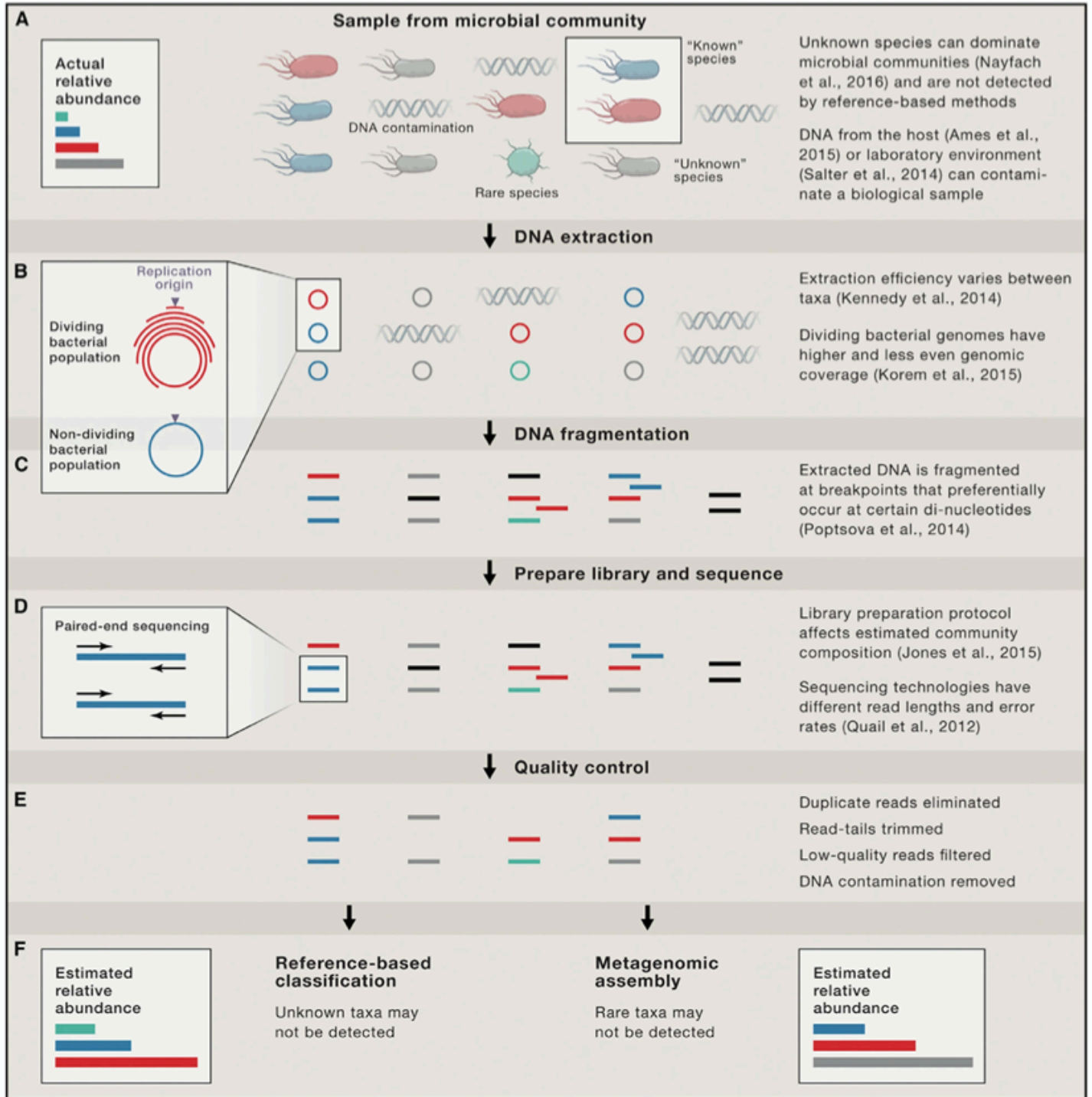
Research article | Open Access | Published: 31 October 2012

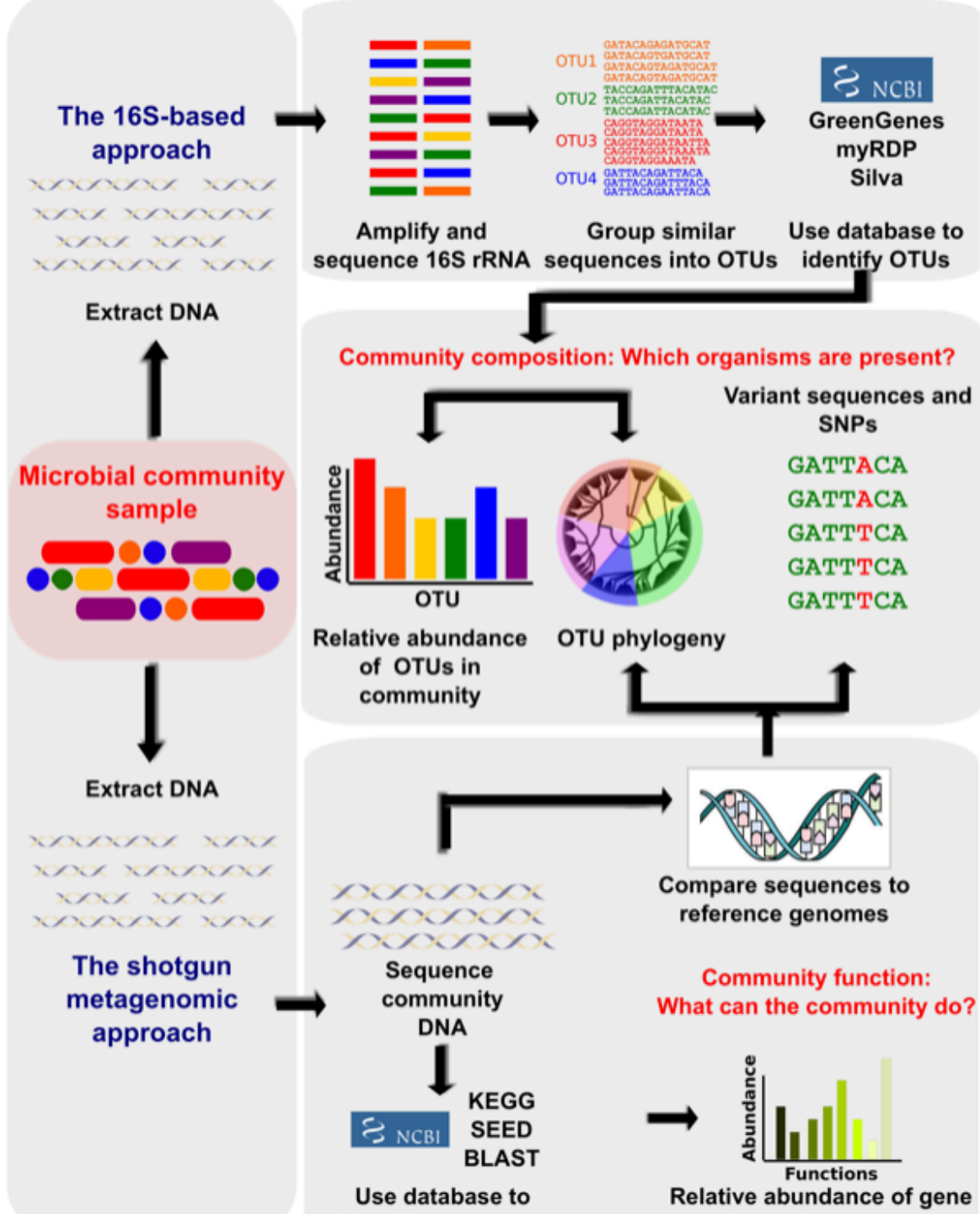
Denoising PCR-amplified metagenome data

Michael J Rosen, Benjamin J Callahan, Daniel S Fisher & Susan P Holmes

BMC Bioinformatics **13**, Article number: 283 (2012) | Cite this article

Shotgun DNA Sequencing




The 16S-based approach
Extract DNA
Microbial community sample
Extract DNA
The shotgun metagenomic approach
Amplify and sequence 16S rRNA
OTU1
OTU2
OTU3
OTU4
Group similar sequences into OTUs
NCBI
GreenGenes
myRDP
Silva
Use database to identify OTUs
Community composition: Which organisms are present?
Abundance
OTU
Relative abundance of OTUs in community
OTU phylogeny
Variant sequences and SNPs
GATTACA
GATTACA
GATTTCA
GATTTCA
GATTTCA
Compare sequences to reference genomes
Sequence community DNA
NCBI
KEGG
SEED
BLAST
Use database to
Community function: What can the community do?
Abundance
Functions
Relative abundance of gene

Microbiome

Microbiomics is a fast-growing field in which all the microorganisms of a given community are investigated together.

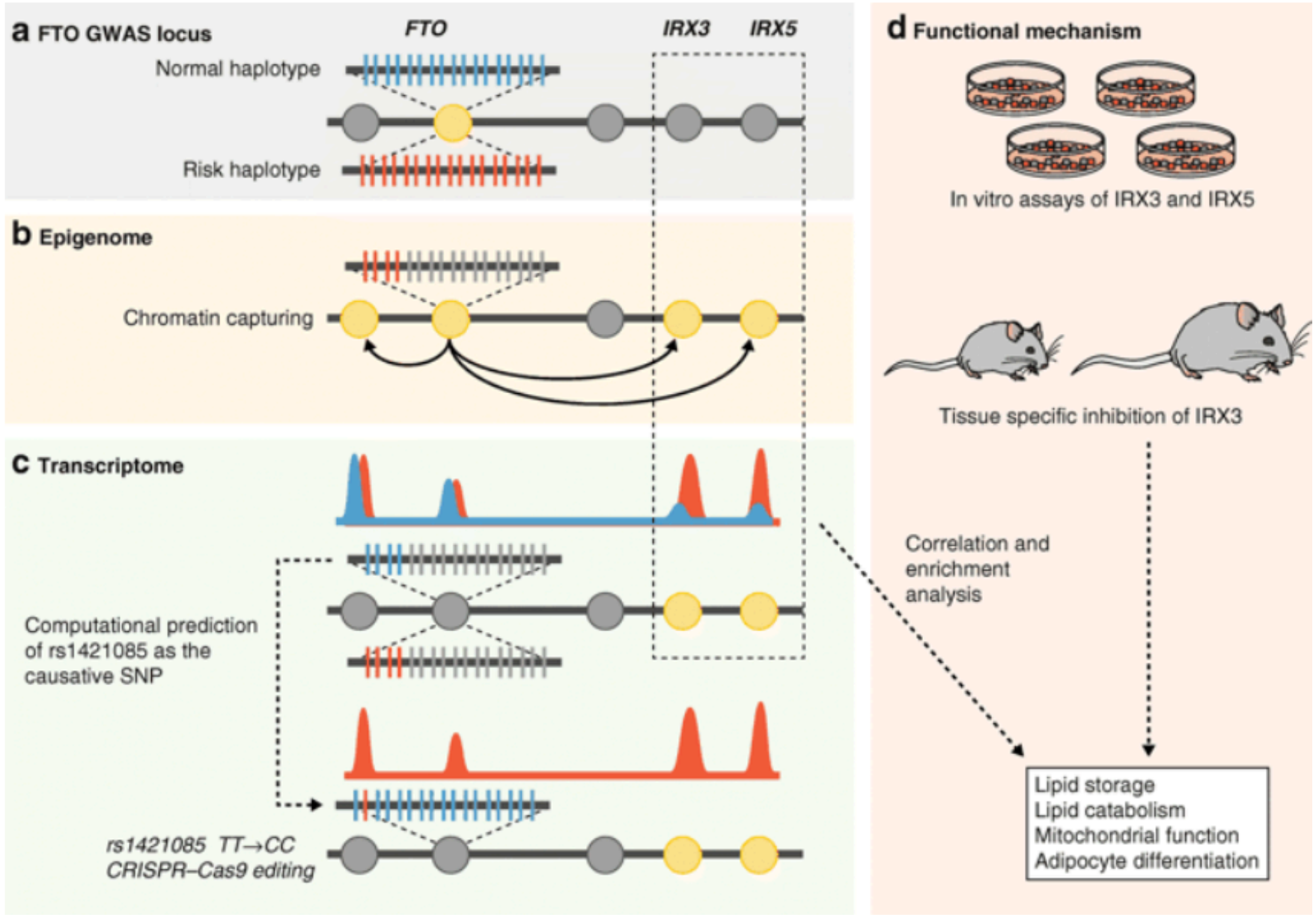
Human skin, mucosal surfaces, and the gut are colonized by microorganisms, including bacteria, viruses, and fungi, collectively known as the microbiota (and their genes constituting the microbiome).

The human microbiome is enormously complex; for example, the gut contains roughly 100 trillion bacteria from 1000 different species. There are substantial variations in microbiota composition between individuals resulting from seed during birth and development, diet and other environmental factors, drugs, and age.

Many studies have implicated perturbations in gut bacteria in a variety of disorders, including diabetes, obesity, cancer, colitis, heart disease, and autism. The microbiome can be profiled by amplifying and then sequencing certain hypervariable regions of the bacterial 16S rRNA genes followed by clustering the sequences into operational taxonomic units.

Shotgun metagenomics sequencing, in which total DNA is sequenced, can provide additional resolution for distinguishing genetically close microbial species. Several analytic tools have been developed for analyzing NGS data from targeted 16S or metagenomics analysis, such as QIIME (quantitative insights into microbial ecology).

integrative analysis of multiple omics data



N Engl J Med. 2015;373:895–907

Metataxonomic analysisVisualization

Metataxonomic analysis sOTU, ASV generation

Introduction to Microbiome Data Analysis

Epigenome-wide association study (EWAS) with methylation arrays & R

GO enrichment analysis using R

Differential gene expression analysis using R

Machine Learning in R

R tutorial and Data Visualization

Integrative Genomics Viewer

Genomic Database-an overview

RNA-sequence Analysis: Expression, Isoforms

DNA-protein Interactions; Motif analysis – HOMER2

ChIP-seq Analysis -MACS2

Genome Assembly

Sequencing quality control, Understanding genomic data file formats

Library construction, illumine sequencing technology

Linux Command

Genomic knowledge