Genomic Computing

DEIB, Politecnico di Milano

Genomic research concepts and application perspective

Marco Masseroli

The following slides were originally prepared and are kindly provided by Prof. Myriam Alcalay

Functional Genomics, Istituto Europeo di Oncologia

Patologia Generale, Università degli Studi di Milano

Outline

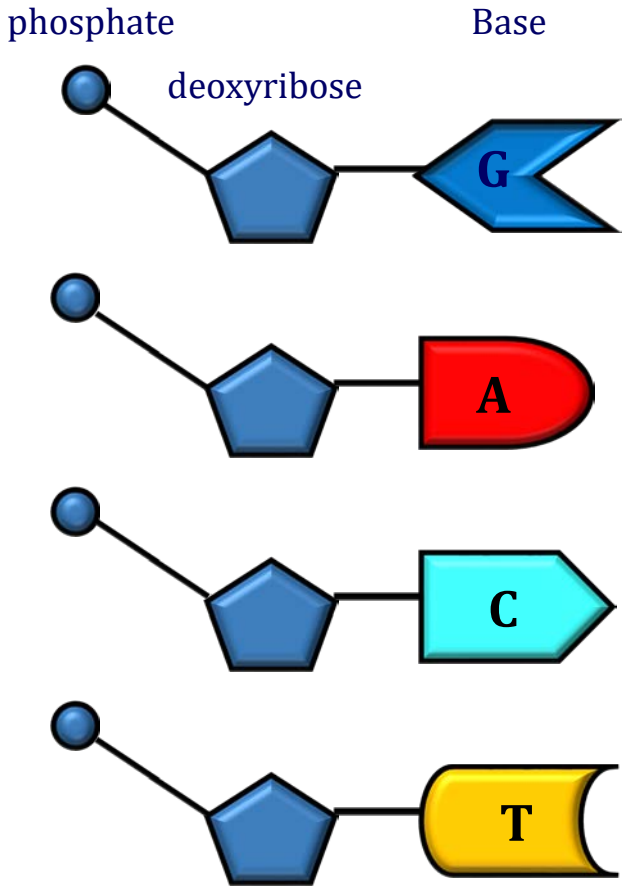
- Organization of the genome
- Genetic variations and their significance
- NGS technology
- Applications

The Genome

- The total amount of hereditary information possessed by any organism.
- It is encoded in DNA (= Deoxyribo-Nucleic Acid) and includes both genes and non-coding sequences
- *The genome is identical in all the cells of an individual, with the sole exception of germ cells (egg/sperm).*

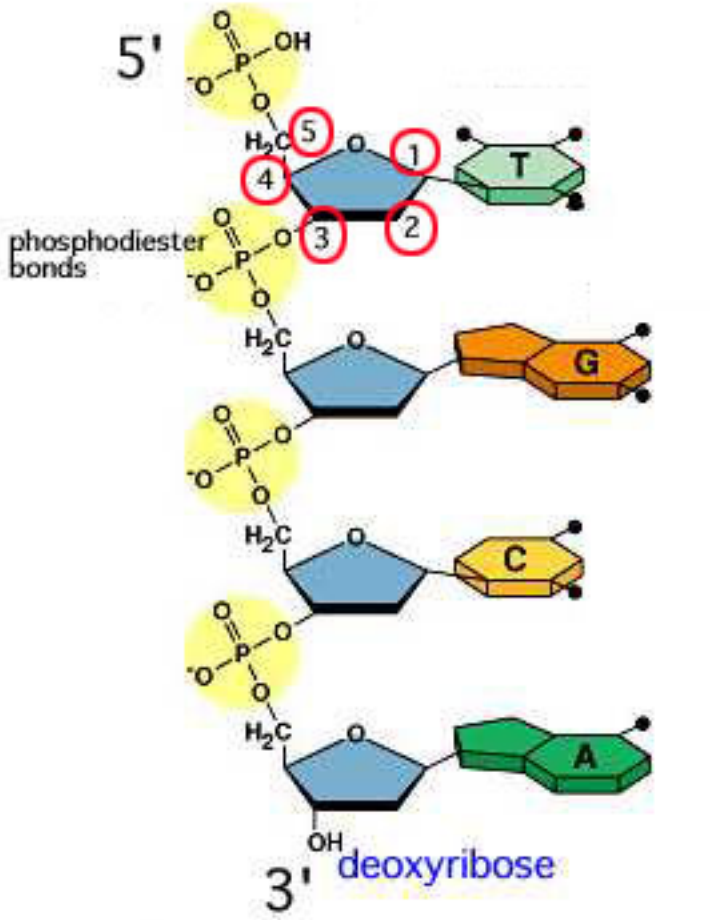
DNA – basic structure

- The basic subunits that make up the DNA molecule are called nucleotides (Guanine, Adenine, Cytosine, Thymine) and they have 3 components:
  - a pentose sugar (deoxyribose)
  - a phosphate group
  - a nitrogenous base (G, A, C, T)



DNA – basic structure

- Nucleotides are linked together though *phosphodiester* bonds to form long strands of DNA.
- These bonds are strong covalent bonds between a phosphate group and two 5-carbon ring carbohydrates.
- They involve the two carbons in position 3 and 5 → the DNA polymer is directional.

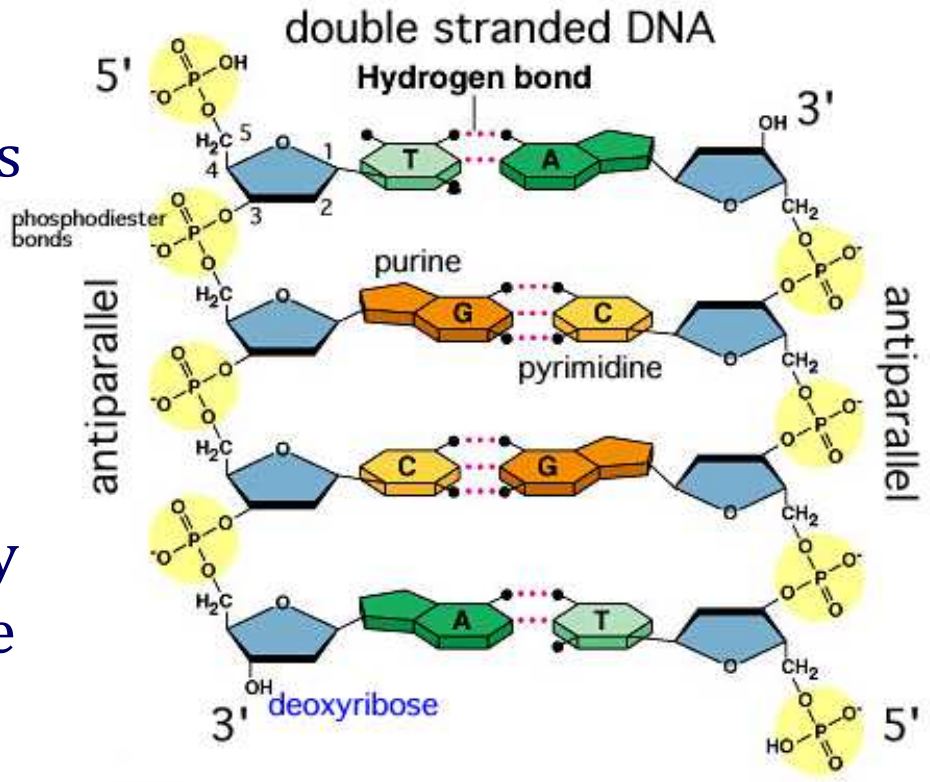


DNA – basic structure

- DNA is not present in eukaryotic genomes as a single strand, but rather as two strands that run in opposite directions (anti-parallel)
- Each specific nucleotide bonds to a complementary nucleotide on the opposite strand

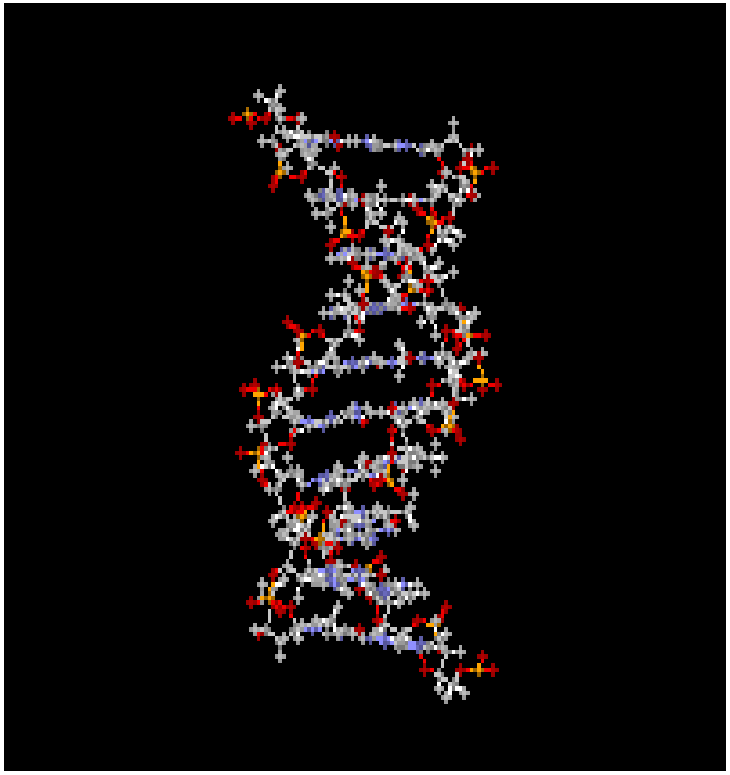
G → C

A → T



DNA – basic structure

- The two strands of DNA are tightly bonded to each other and intertwine to form a double helix.
- The DNA double helix is stabilized primarily by hydrogen bonds between complementary nucleotides



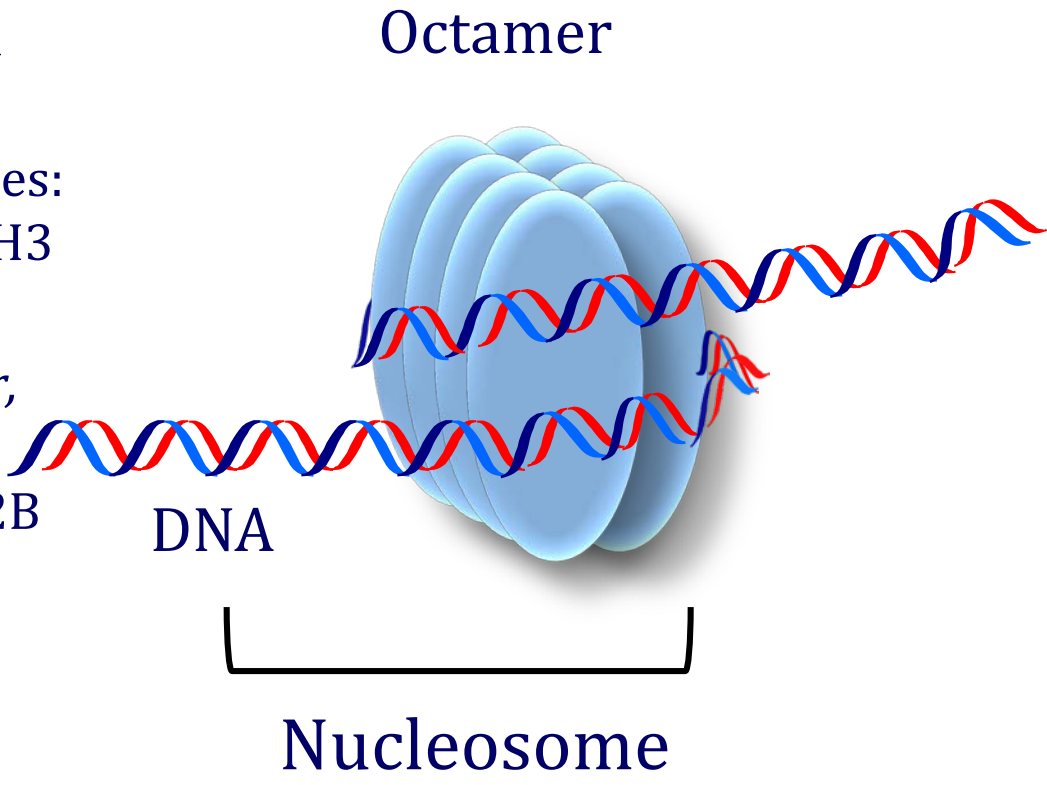
Organization of the Genome

- the human genome, if organized linearly, would form a filament 2 meters long.
- How do 2 meters of genetic material fit into the nucleus of a cell (few µm)?

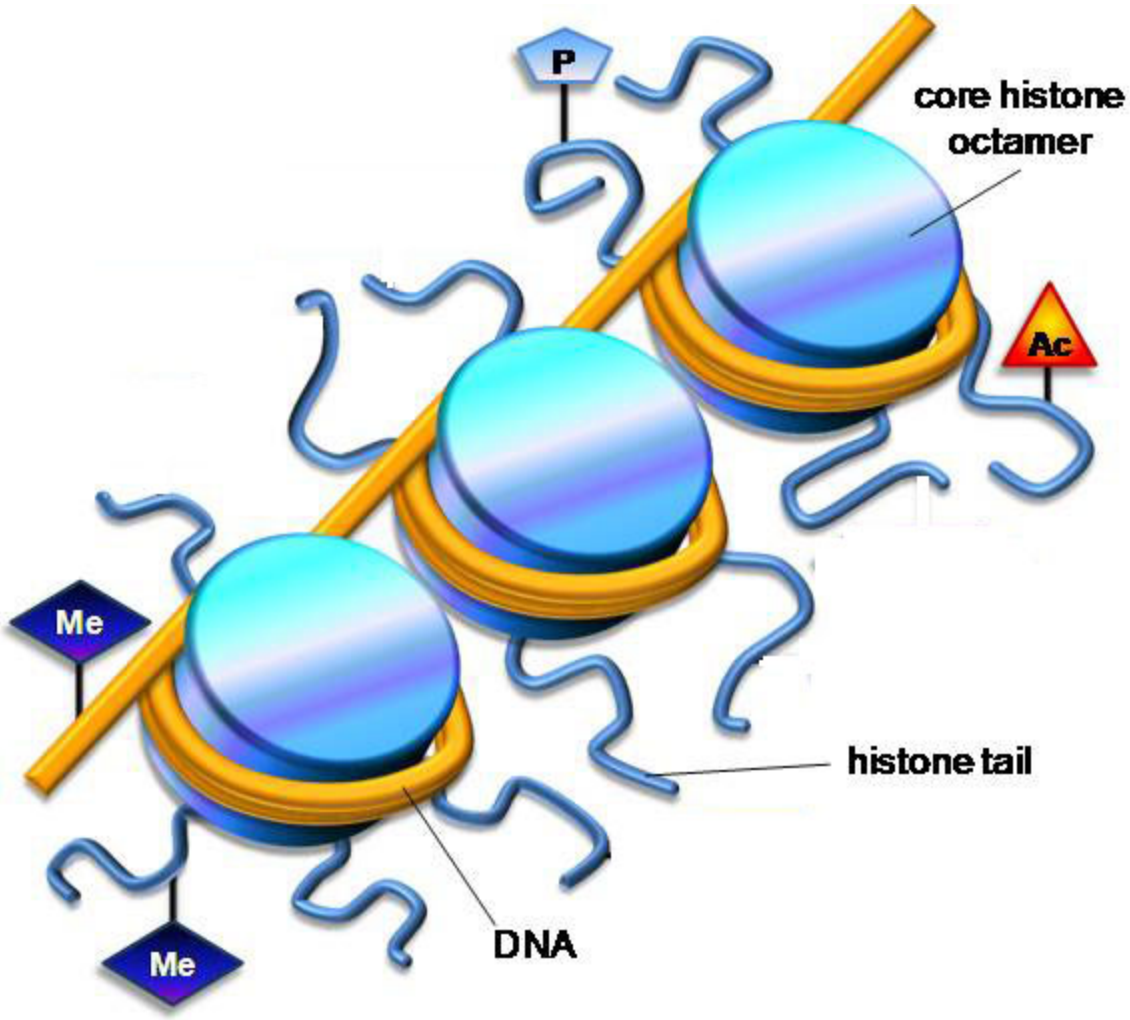
→ chromatin (DNA and proteins)

Histones and Nucleosomes

- Histones are a family of small nuclear proteins that form complexes with each other and with DNA.
- There are several classes of histones: the most abundant are H2A, H2B, H3 and H4, known as *core histones.*
- The core histones form an *octamer,* composed of a central H3-H4 tetramer and two flanking H2A-H2B dimers.
- DNA wraps around a histone octamer to form a *nucleosome*, the first order of compaction of eukaryotic chromatin.



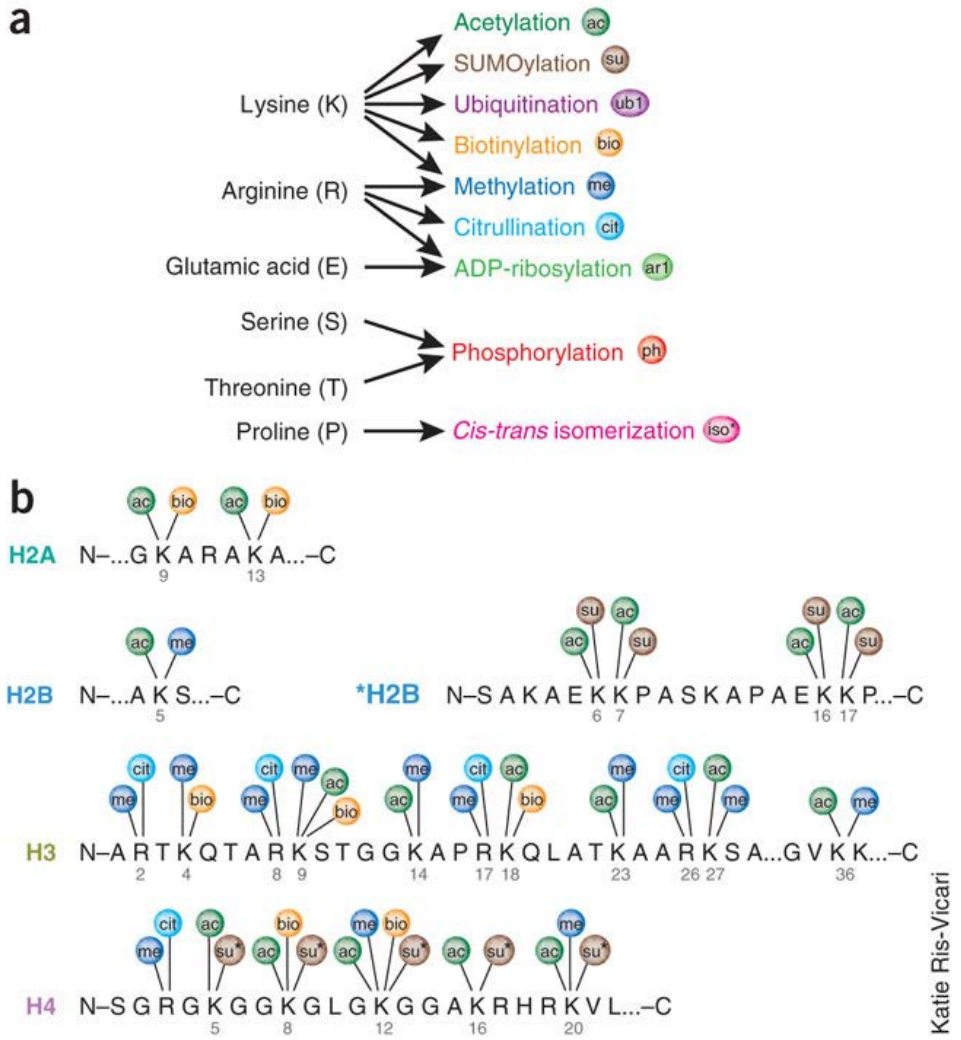
Histones tails and epigenetic modifications



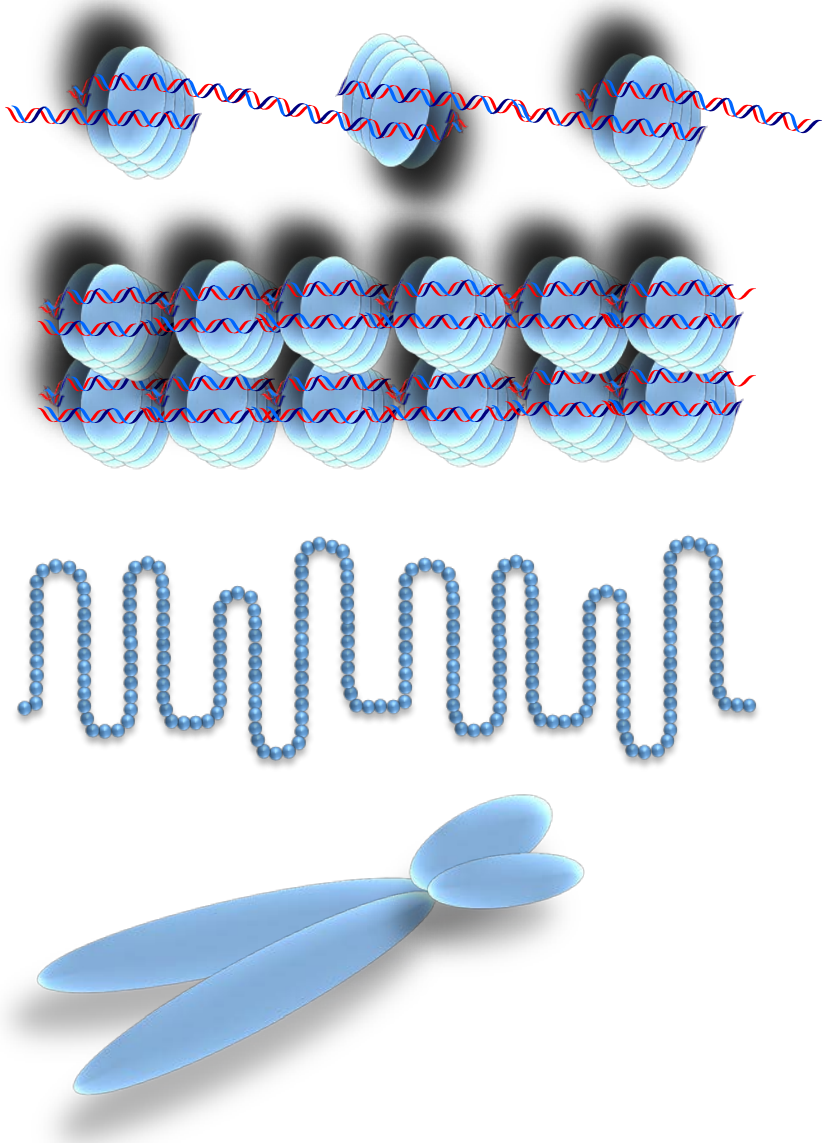
Histones tails and epigenetic modifications

A large (and growing) number of post-translational modification of histone tails has been identified.

These occur at specific positions of specific core histones and determine the structure of nucleosomes and the accessibility of DNA.



Chromatin and Chromosomes



Nucleosomes like "beads-on-a-string"

↓

30 nanometer fiber

↓

Chromatin loops

↓

Chromosome

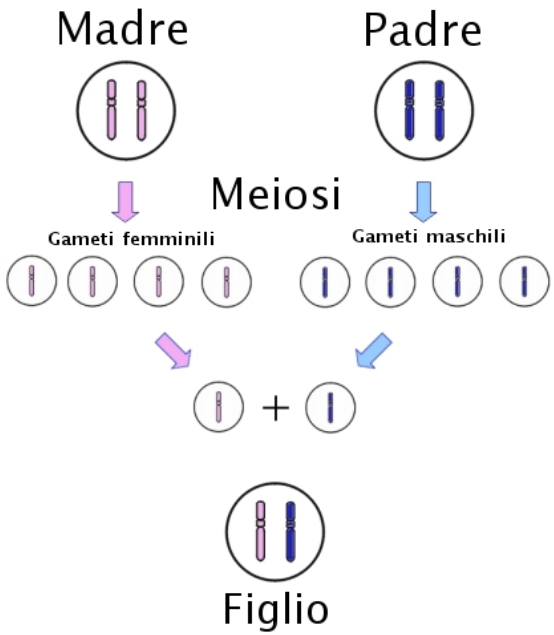
Chromosomes

- Each species has a characteristic number of chromosomes – humans have 46 chromosomes (23 pairs).

- Each chromosome is present in two copies (one of maternal and one of paternal origin) = diploid genome.

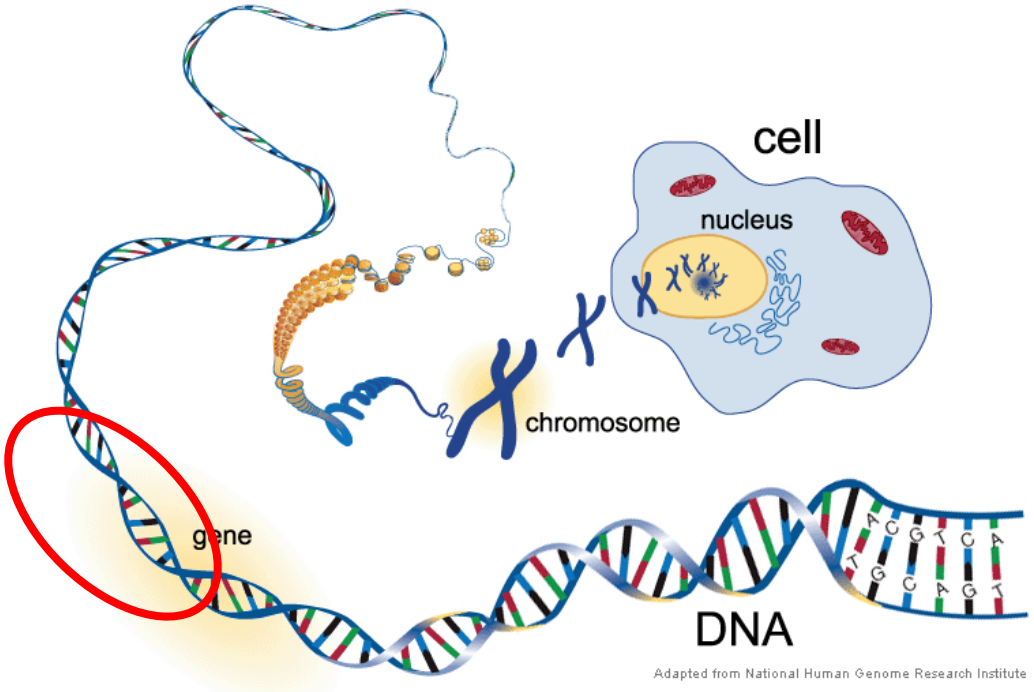
Diploid Chromosome Number

Goldfish	94
Potato	48
Human	46
Pea	14
Fruit fly	8



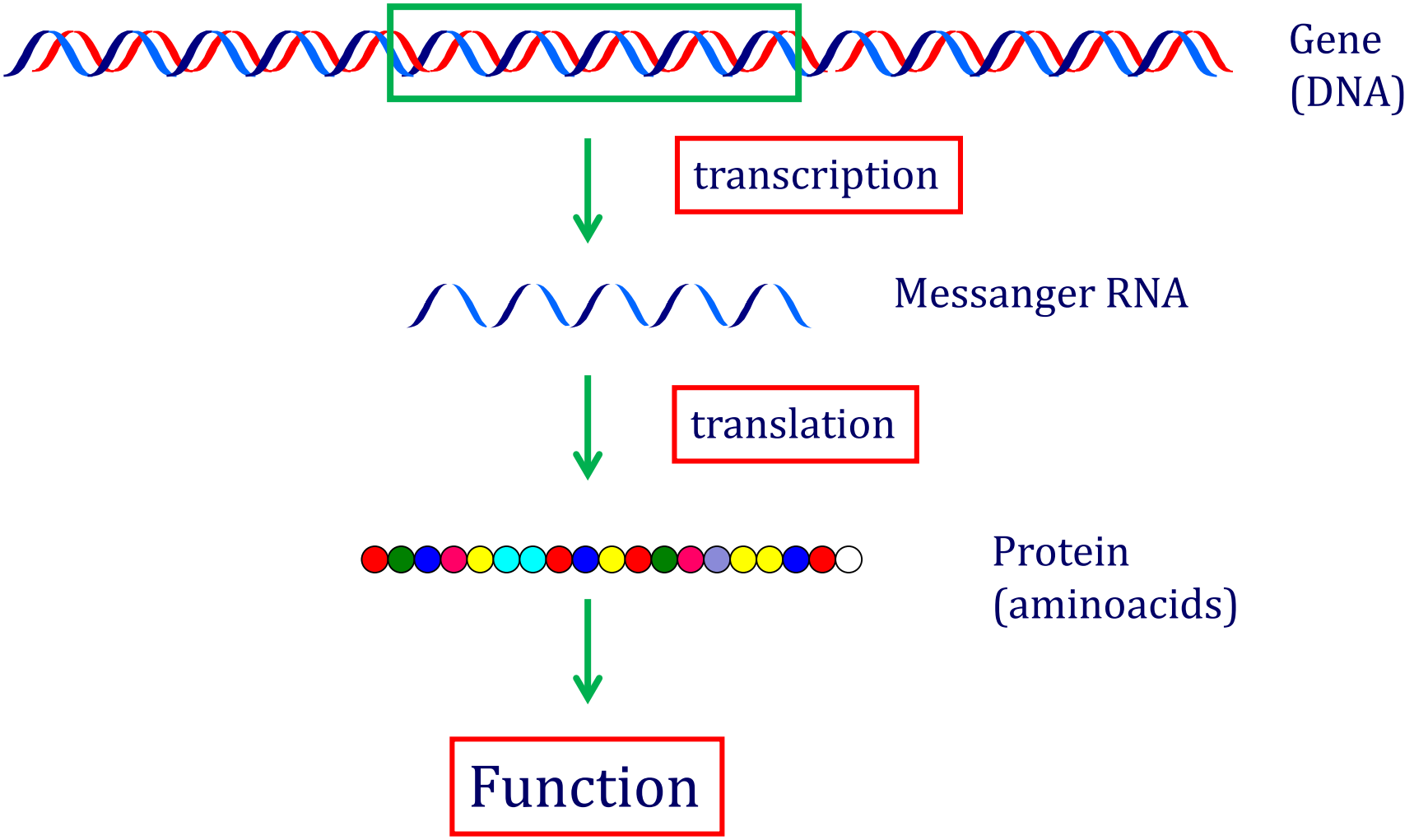
Genes

- a gene is a region of the genome, which encodes for a protein, and is associated with regulatory regions, transcribed regions, and other functional sequence regions.



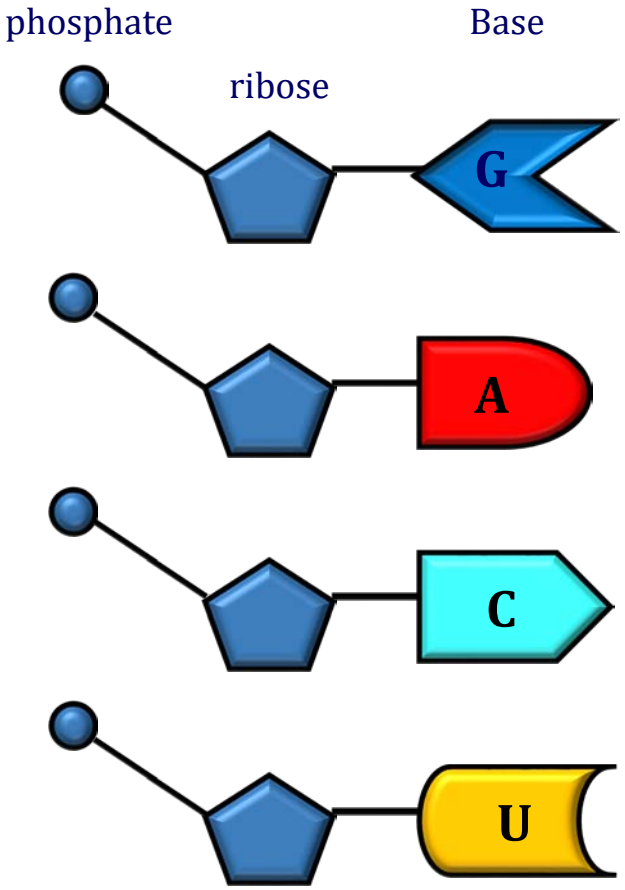
2-3% of the genome

Flow of genetic information



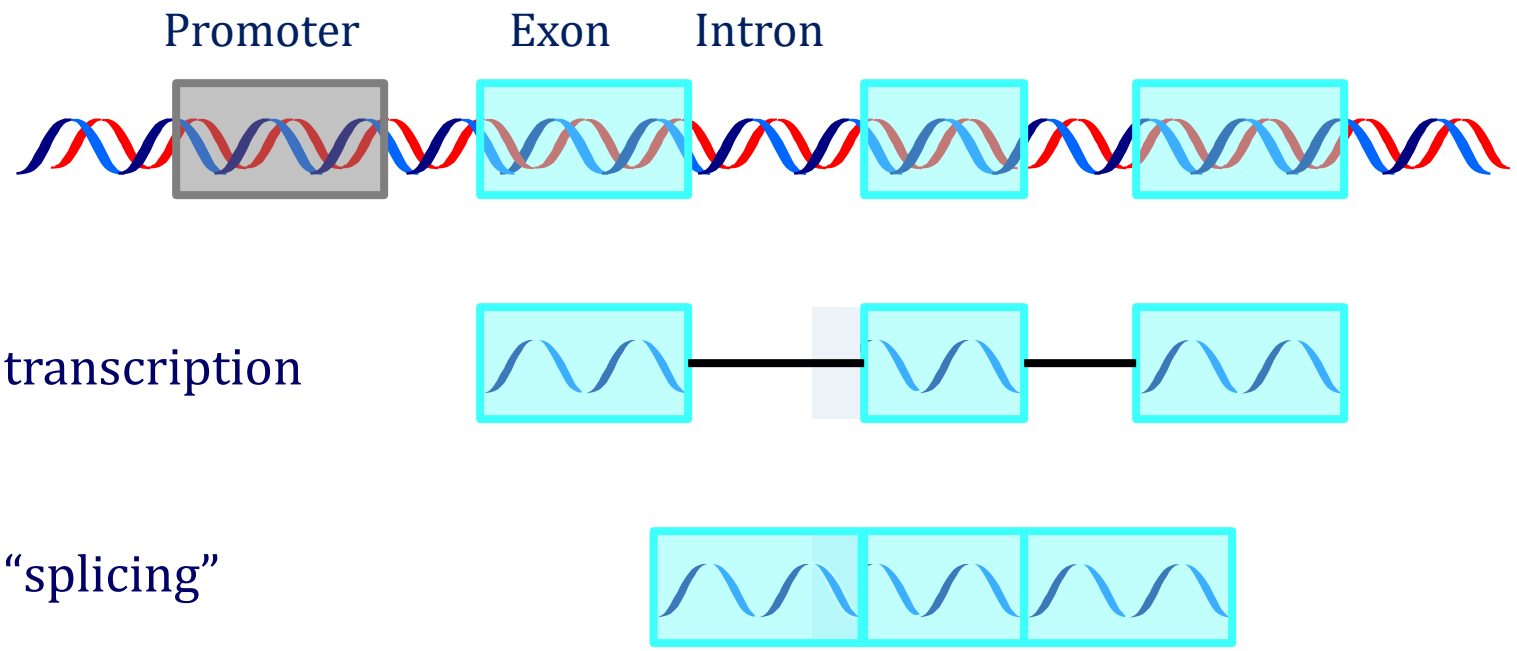
mRNA

- Synthesized by RNA polymerase – complementary to DNA
- Single stranded
- Pentose sugar is ribose
- Thymidine is replaced by Uracil
- Shuttles from the nucleus to the cytoplasm for translation into proteins.

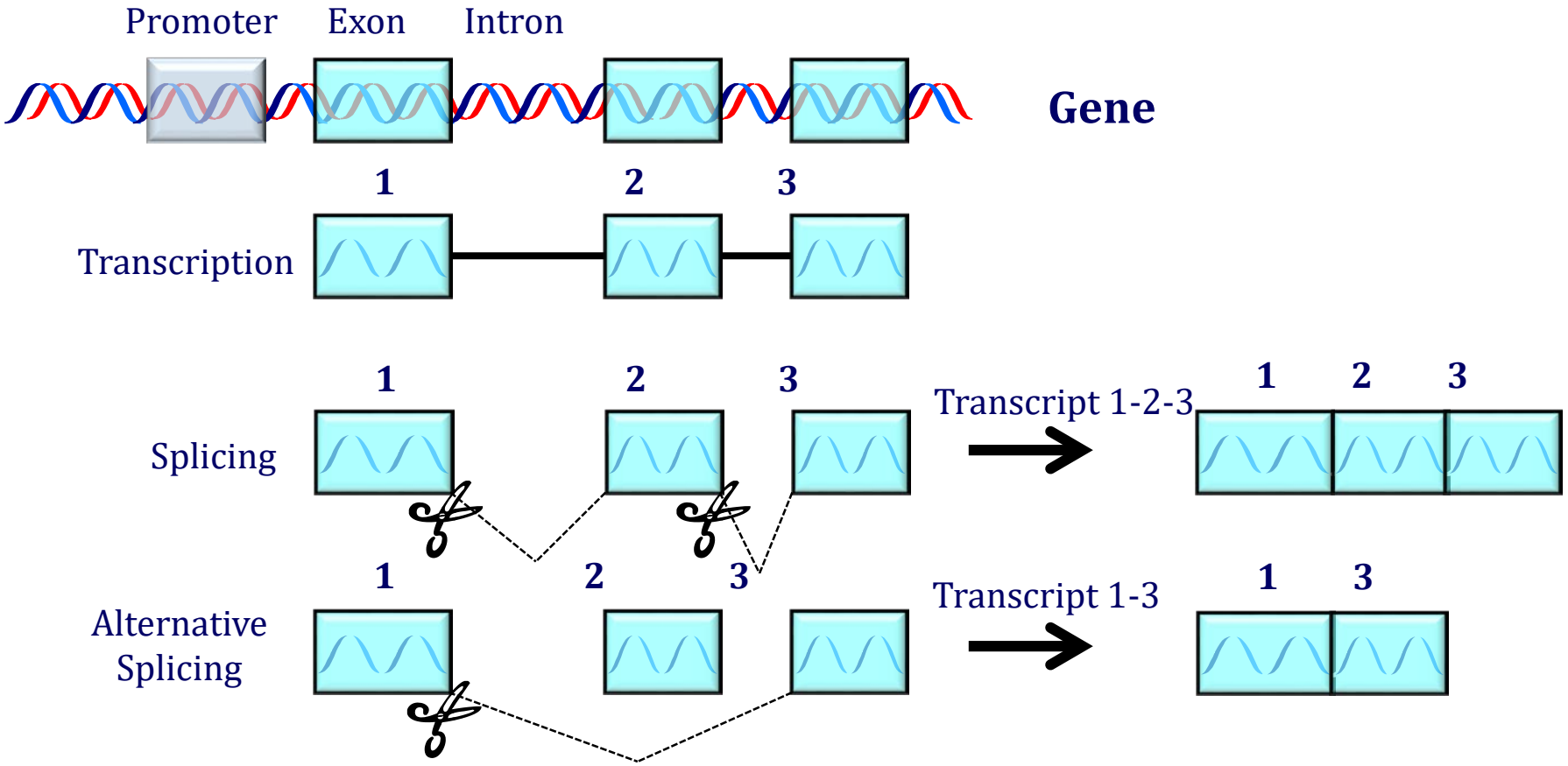


The gene - structure

- Genes contain coding and non-coding sequences:



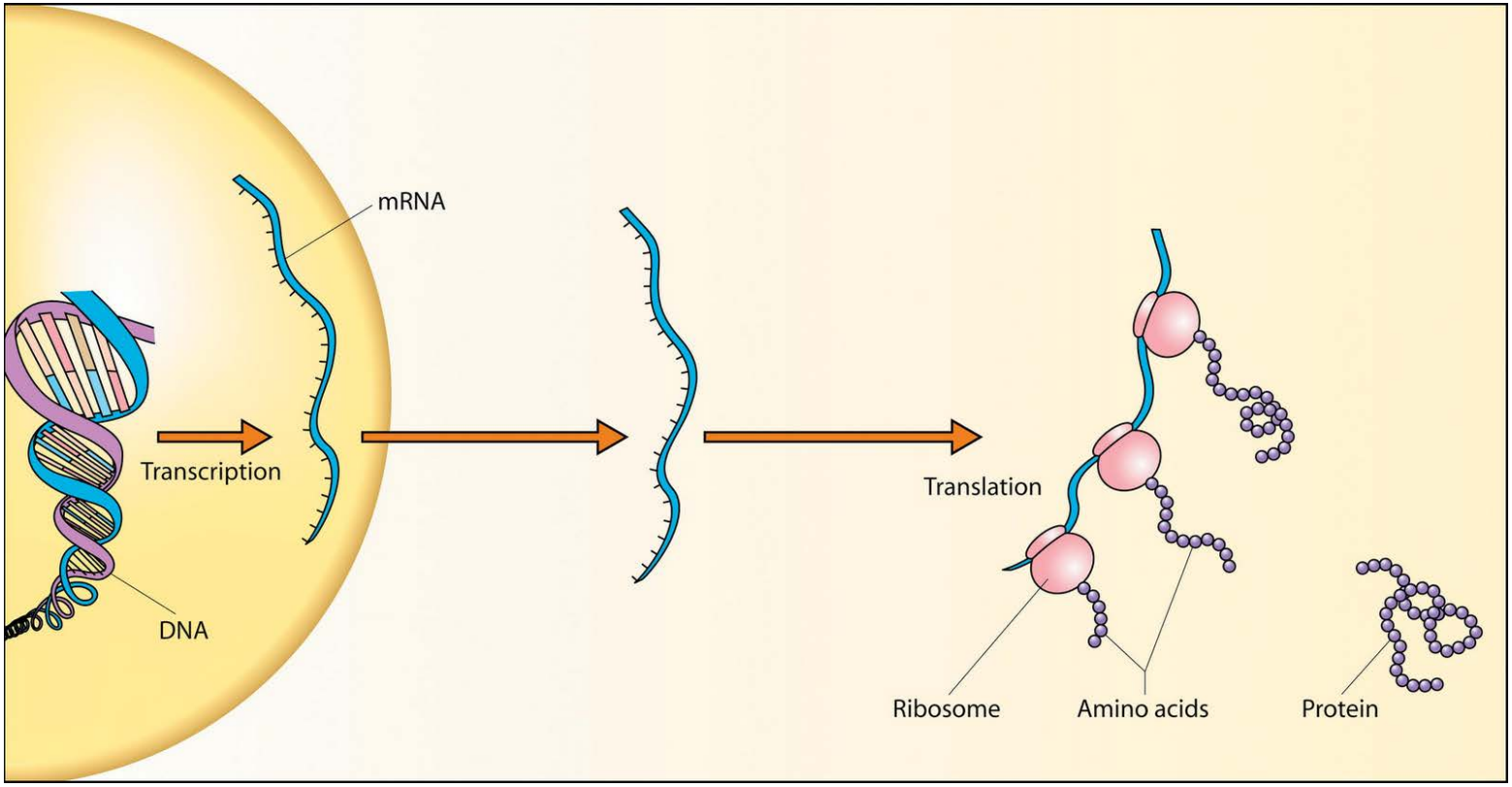
Alternative Splicing



1 gene → > 1 protein

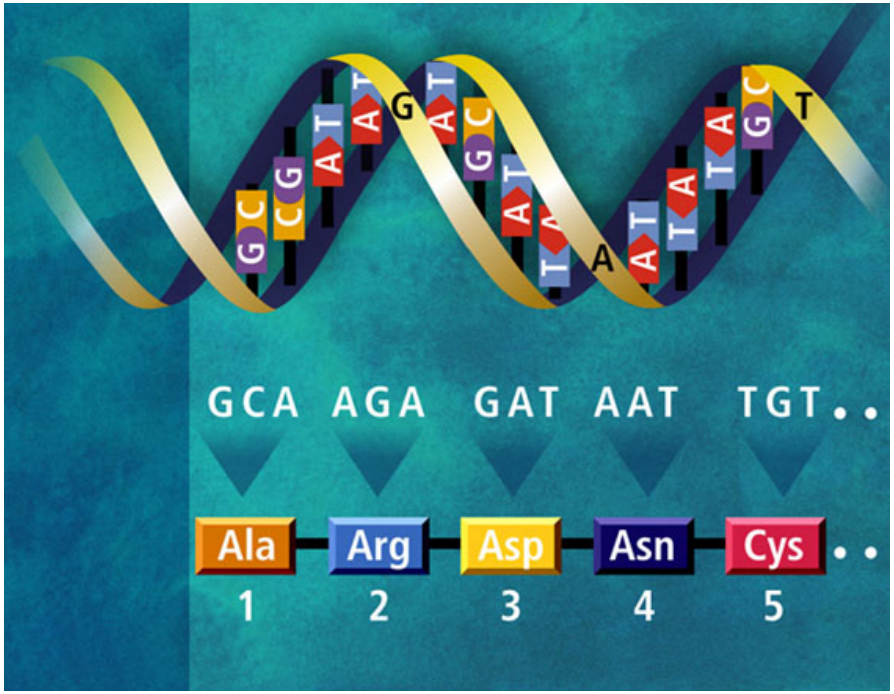
Translation

- mRNAs exit the nucleus migrate to cytoplasmic organelles known as *ribosomes*, where translation into proteins occurs.



The Genetic Code

- The genetic code is the set of rules by which information encoded within genetic material is translated into proteins.
- Nucleotides are read in *triplets*, and with few exceptions, a 3-nucleotide *codon* specifies a single amino acid.



The Genetic Code

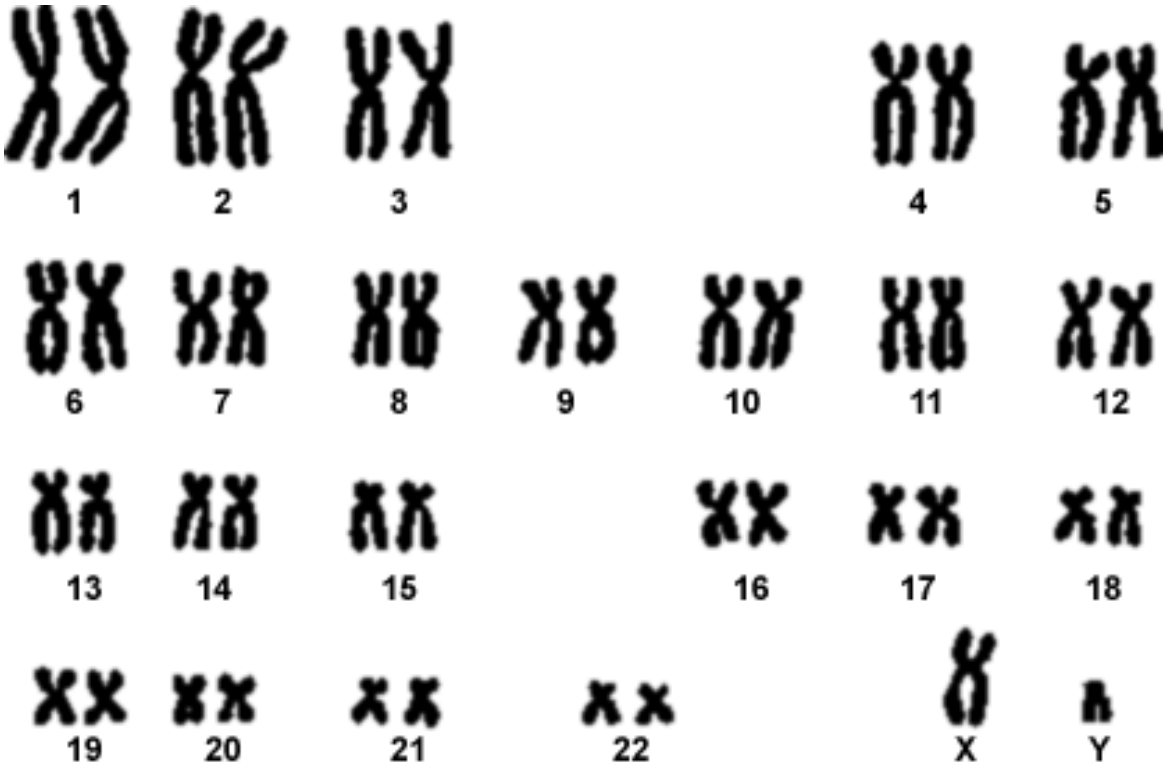
- There are 4 nucleotides, therefore 64 different triplets (4^3).

- But there are only 20 amino acids

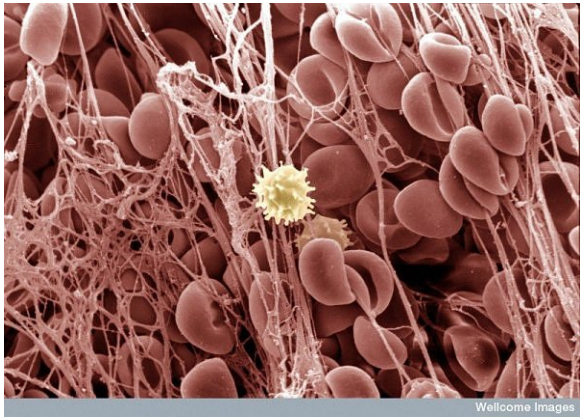
  → the genetic code is redundant (or *degenerate*), since different triplets can encode for the same amino acid.

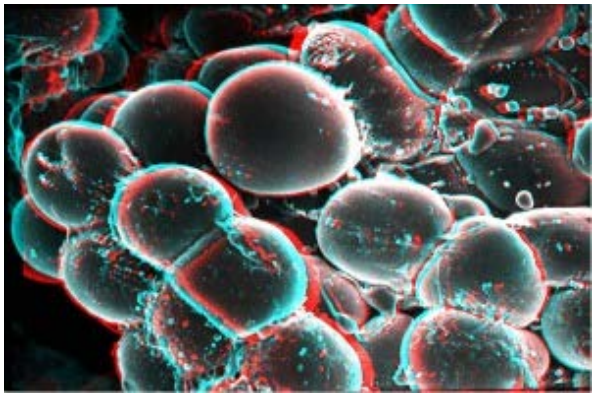
	U	C	A	G	
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

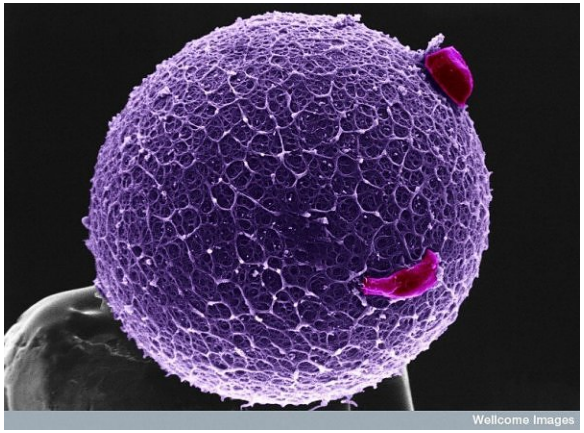
Genetic information is the same in all cells of an individual

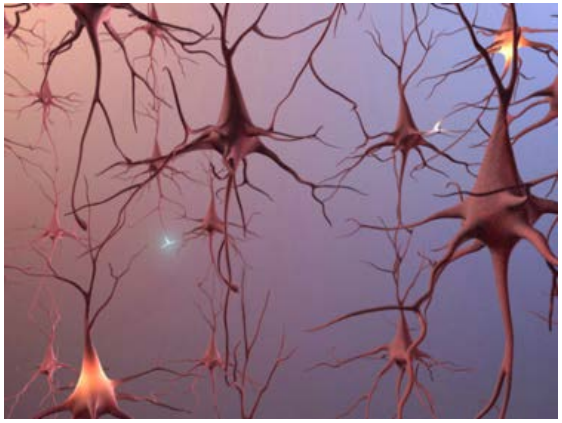


→ but cells are very different in terms of both structure and function



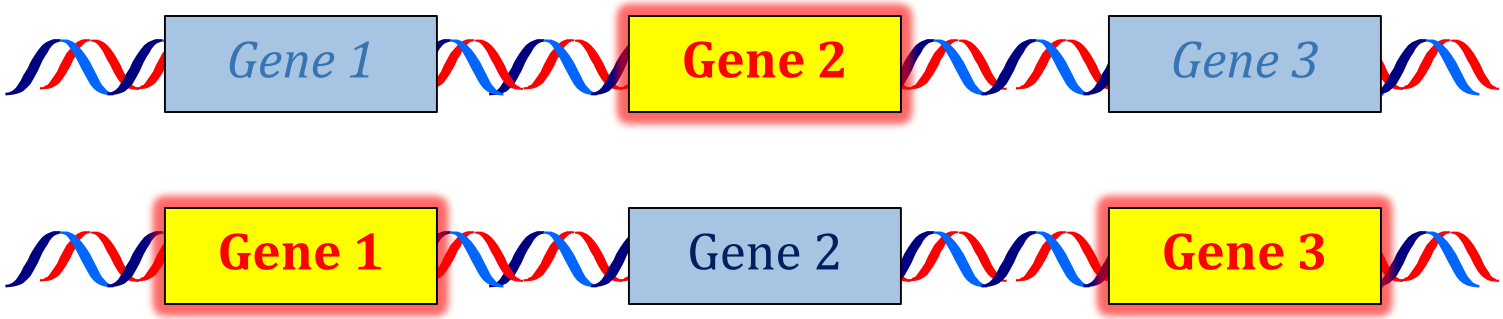




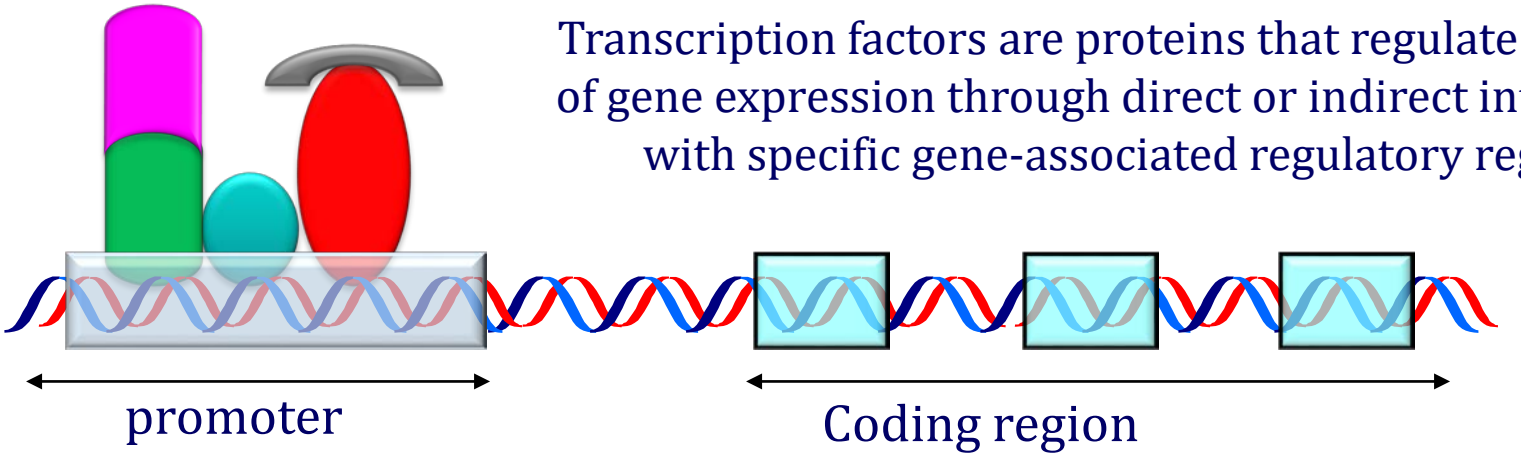


Gene activity

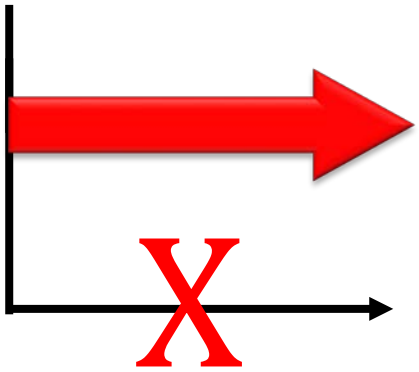
- Not all genes are active (= transcribed) at the same time or in the same cells.



Regulation of gene activity:
1) Transcription Factors



Transcription factors are proteins that regulate the levels of gene expression through direct or indirect interactions with specific gene-associated regulatory regions.

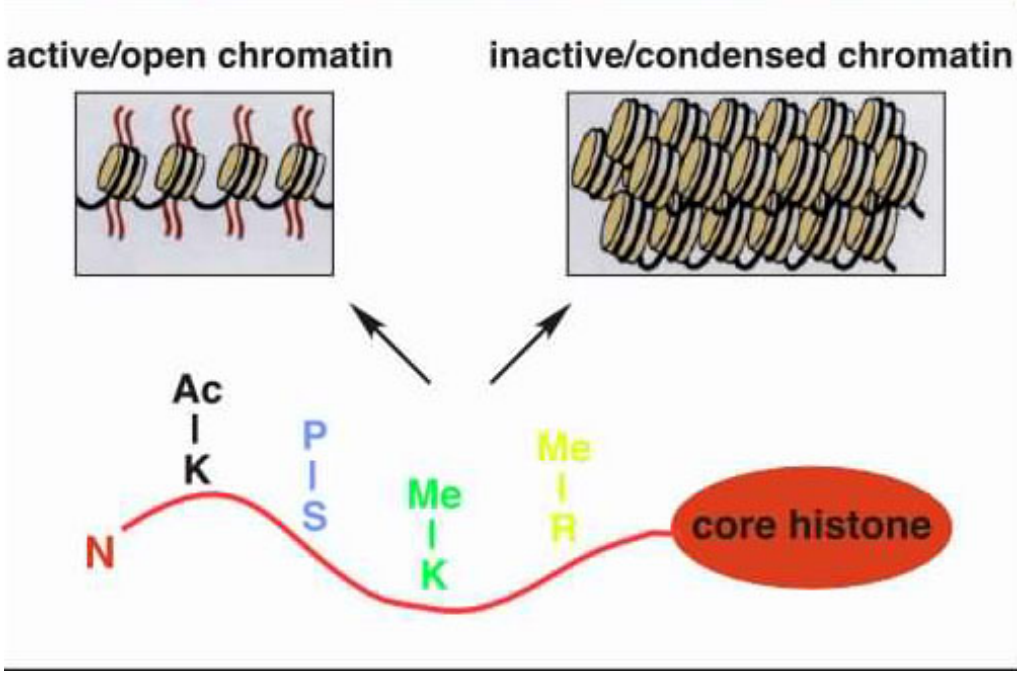


Increased transcription (= activation)

Decreased transcription (= repression or silencing)

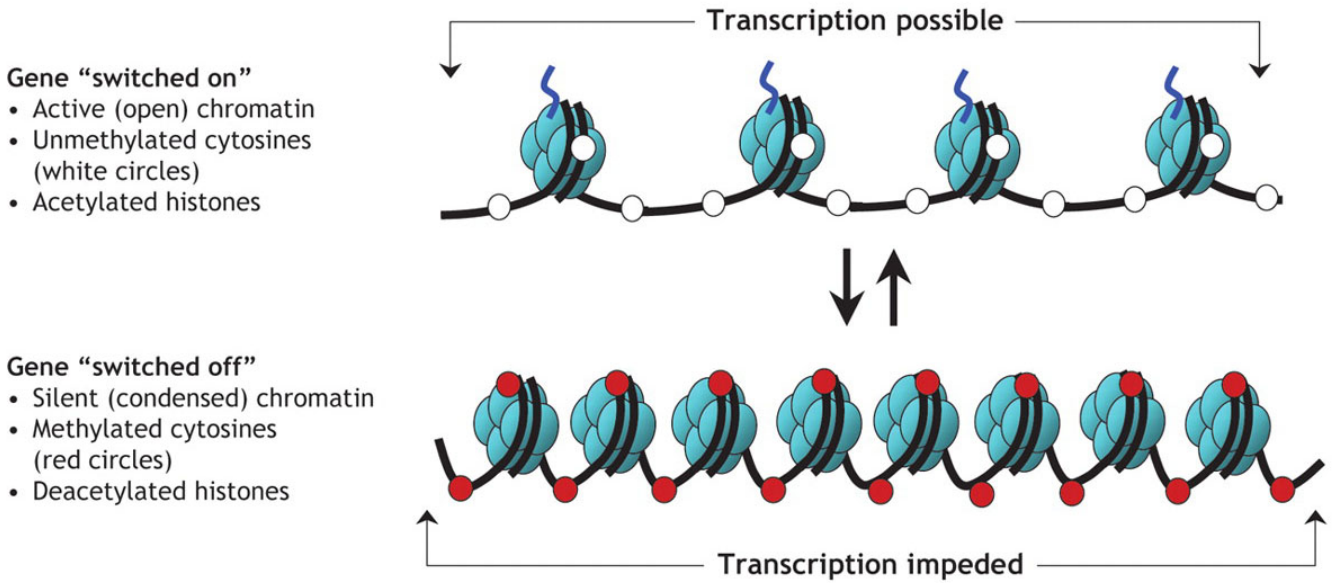
Regulation of gene activity:
2) histone modifications & DNA accessibility

- Epigenetic modifications are changes in the genome that do not involve a change in the nucleotide sequence
- Histone modifications modulate chromatin structure, which, in turn, modulates DNA accessibility.



Regulation of gene activity:
3) DNA methylation

- DNA methylation is a biochemical process involving the addition of a methyl group to the cytosine or adenine DNA nucleotides.
- DNA methylation at the 5' position of cytosine, typically occurring in a CpG dinucleotide, has the specific effect of reducing gene expression.
- DNA methylation is permanent and unidirectional and can be heritable.



In summary (1)

- DNA is organized in chromosomes - there are 23 pairs of chromosomes in each human cell.
- The genome is the same in all cells of an individual.
- All functions in an organism are performed by proteins that are encoded in specific regions of the genome (genes).
- There are two copies of each gene in our genome.
- Not all genes are active simultaneously in a cell.
- Functional and structural characteristics of each cell type are determined by the combination of genes that are actively transcribed.

In summary (2)

- The structure of proteins, and therefore their function, is determined by the specific sequence of nucleotides that compose the corresponding gene.
- Functional and structural characteristics of each cell type are determined by the combination of genes that are actively transcribed.
- Gene activity is determined by a combination of regulatory mechanisms that include transcription factors and epigenetic modifications (histone modifications and DNA methylation).

Genetic variations

The gene that encodes for a protein can exist in different versions:

=

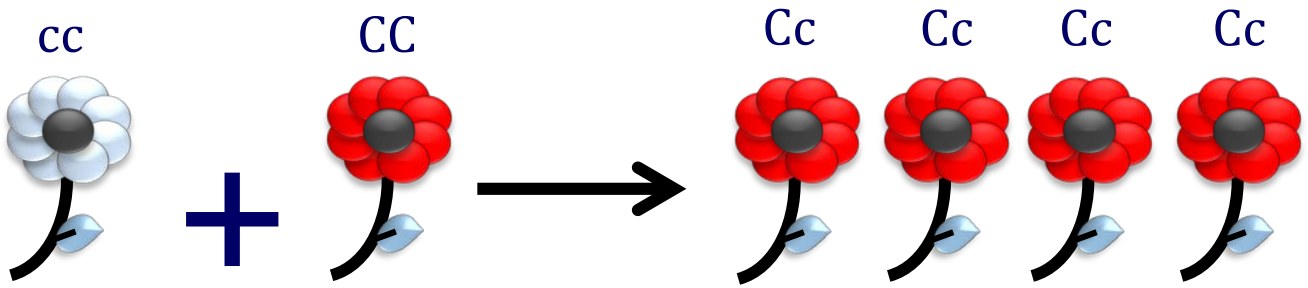
Genotype/Phenotype

- **Allele**: one of the variants of a gene that is present in a given population.

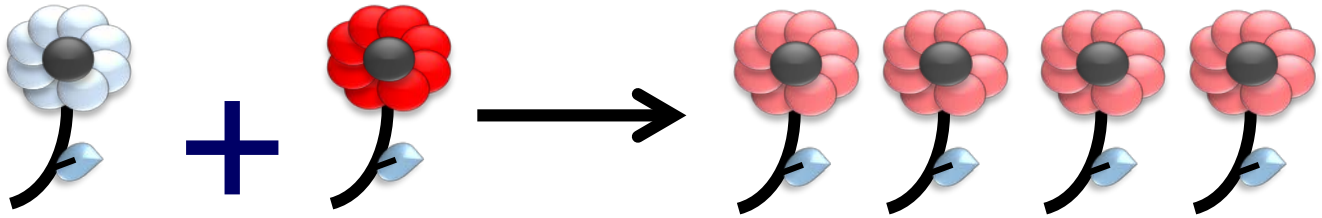
- **Genotype**: the two alleles for a given trait that are present in an individual. If the two alleles are identical, the individual is defined as homozygote, if they are different as heterozygote.

- **Phenotype**: visible trait that results from the genotype.

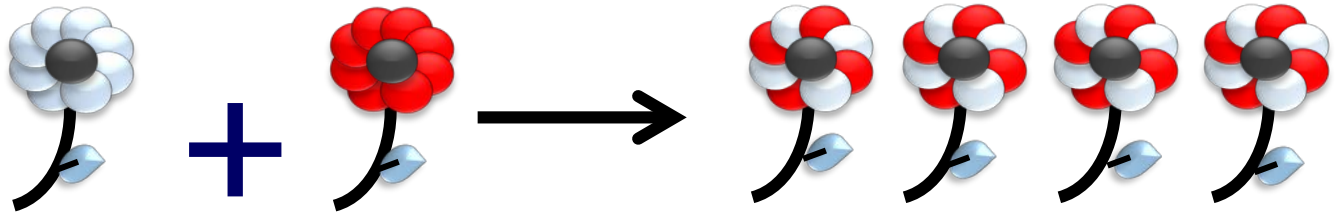
Allelic Dominance



Red is dominant



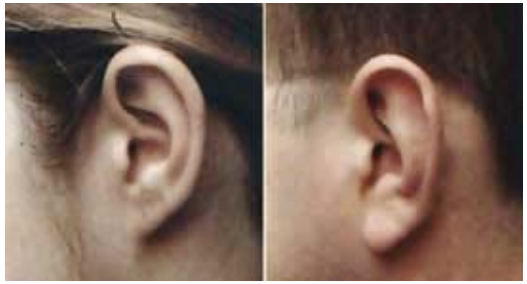
Incomplete dominance



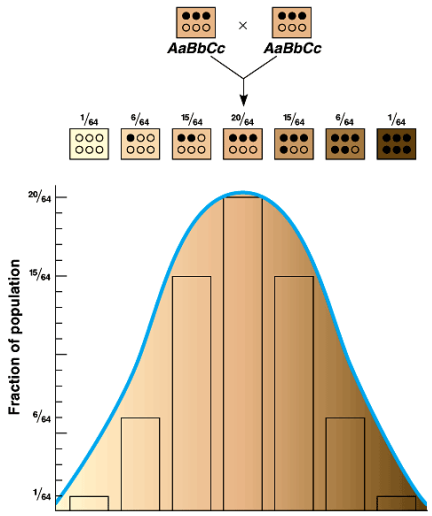
Co-dominance

Monogenic traits

- The phenotype depends on the activity of one single gene that is necessary and sufficient to express the trait.



Polygenic traits



- The trait is determined by the activity of two or more genes, each of which contributes in a certain degree to the definition of the same phenotypic

  → example: skin color (genotype = phototype)

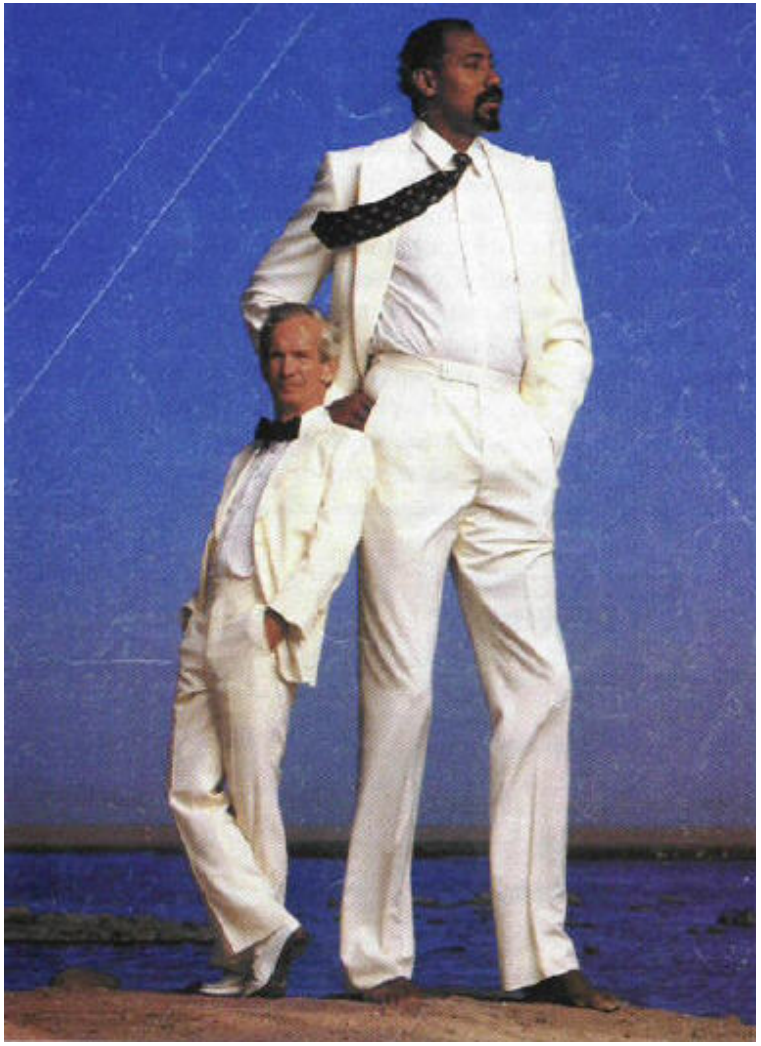
Phototype	Pigment characteristics
I	Red-haired people, with freckles, belonging to the Celtic race
II	Fair-haired people
III	Dark-haired people
IV	Latin people
V	Arabic, Asian people
VI	Black people

Multifactorial traits

- The phenotype results from the interaction of two or more genes and environmental factors. Example: skin color(phenotype)

Phototype	Sun's action on the skin	Pigment characteristics
I 	Easily get burnt, never get tanned, redden	Red-haired people, with freckles, belonging to the Celtic race
II 	Easily get burnt, do not get tanned very much	Fair-haired people
III 	Fairly get burnt, gradually get tanned	Dark-haired people
IV 	Do not get burnt very much, always get tanned very well	Latin people
V 	Rarely get burnt, are always tanned	Arabic, Asian people
VI 	Very rarely get burnt, have a large amount of melanin	Black people

Genetic variations



We are all 99.9% genetically identical to each other.
The 0.1% difference is due to **genetic variations.**

Genetic variations have a major impact on how we respond to:

- Diseases
- environmental insults (bacteria, viruses, chemicals)
- drugs and other therapies.

The analysis of genetic variants has its primary uses in:

- Forensic medicine
- Anthropology
- Pharmacogenomics
- Correlation genotype-disease

Genetic Variations

Two types of genetic mutational events create variations:

Quantitative: insertion or deletion of one or more nucleotide(s)

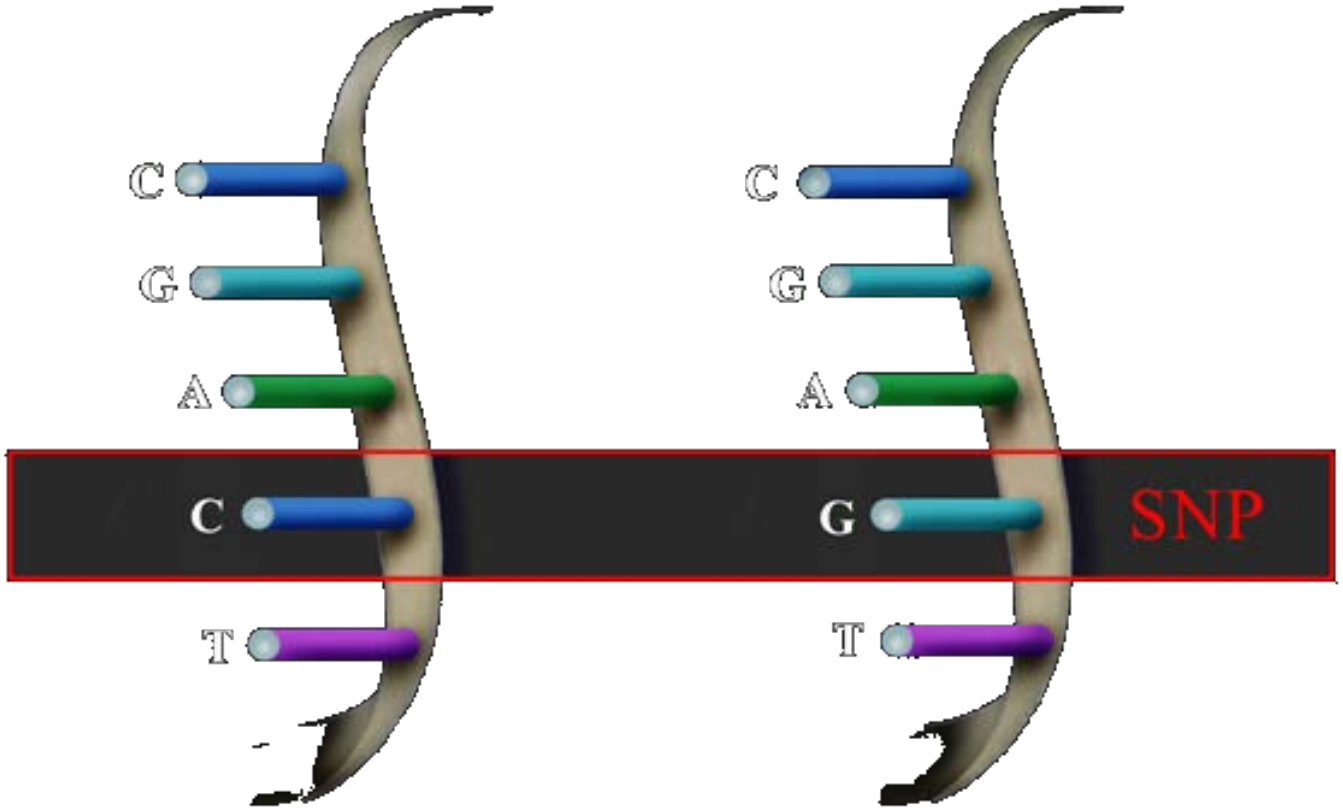
- Insertion/Deletion Polymorphisms (InDel)
- Tandem Repeat Polymorphisms (VNTR)
- Copy Number Variations (CNV)

Qualitative: Single nucleotide substitutions

- Single Nucleotide Polymorphisms (SNP)

Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNP) are DNA sequence variations that occur when a single nucleotide (A,T,C or G) in the genome sequence is substituted by another.



SNP Key Concepts

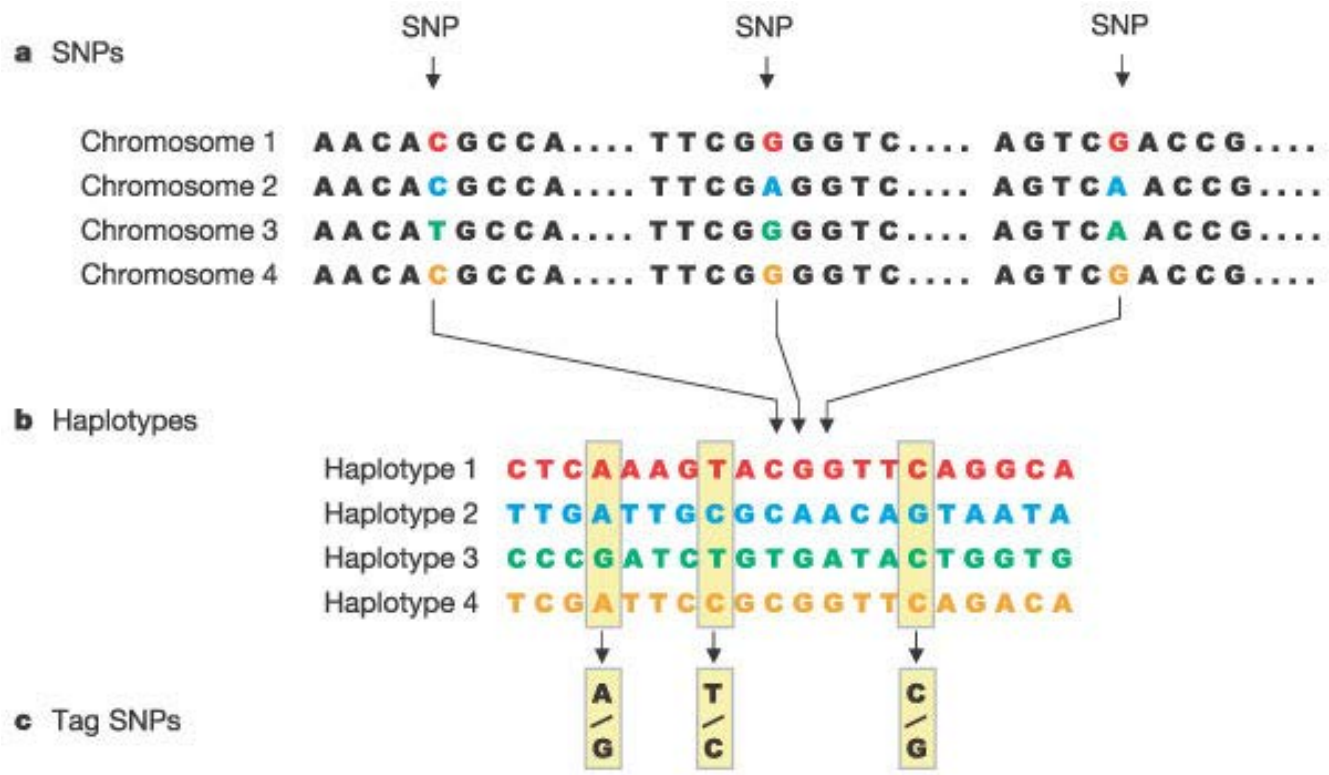
- Definition: More than one alternative bases occur at an appreciable frequency
- Availability: Over 154 million SNPs identified in human genome (dbSNP Build 149, November 2016)
- Function: Most SNPs are "neutral" (a small proportion is present in protein-coding regions)

SNPs vs. Mutations

Both terms indicate variation at a single nucleotide position. The difference is defined by allele **frequency**.
A single base change occurring in a population at a frequency of **>1%** is termed a **SNP**.
When a single base change occurs at **<1%** it is considered a **mutation**.

Haplotypes

Groups of SNPs at adjacent locations on a chromosome are inherited together. Therefore, the identification of a few alleles of a haplotype sequence, can unambiguously identify all other SNPs in its region.



SNPs and disease

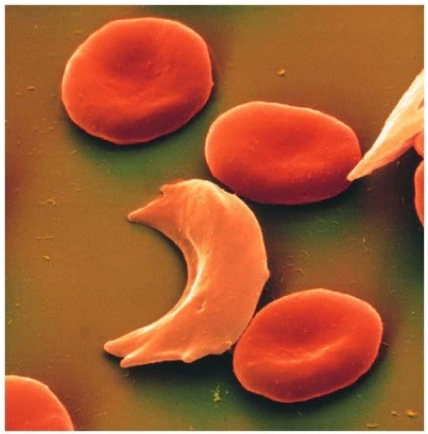
- Single nucleotide variations can lead to the formation of a pathologic protein product → associated with a specific disease.

Sickle cell anemia:

GLU

HBB ...CAC CTG ACT CCT GTG GAG AAG TCT...
HBS ...CAC CTG ACT CCT GAG GAG AAG TCT...

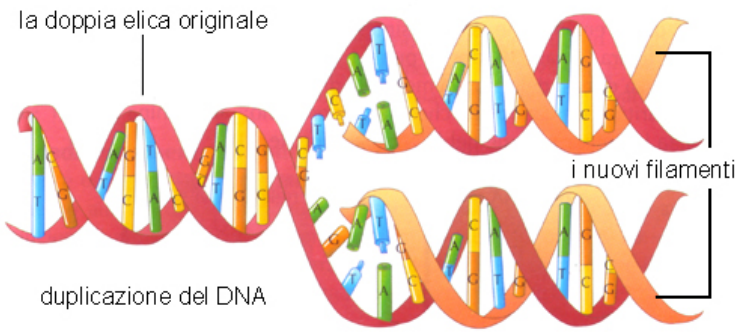
VAL



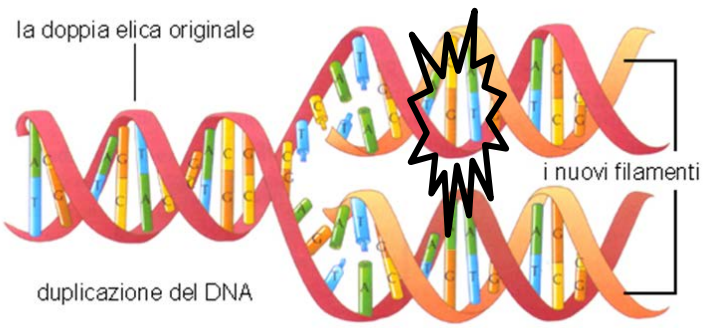
Molecular mechanism

- SNPs/mutations occur due to the insertion of a “wrong” (mismatched) nucleotide during the process of DNA duplication or DNA repair.

normal



mutation



Considerations

- Most nucleotide variations occur in non-coding regions of the genome → no protein is affected.

- Variations in coding sequences can generate different proteins, but in most cases such variations do NOT cause harm to the affected individual.

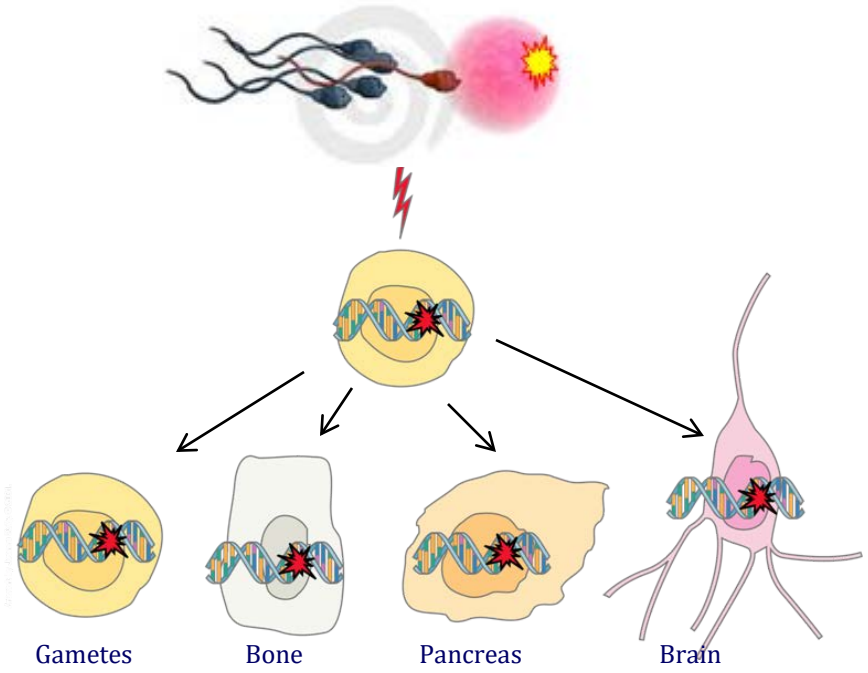
- When such variations led to abnormalities in the structure/function of the encoded protein, the consequences depend on the type of cell where the genetic variation occurs:

  - Germ cells (or sex cells): egg, sperm

  - Somatic cells: all other cells

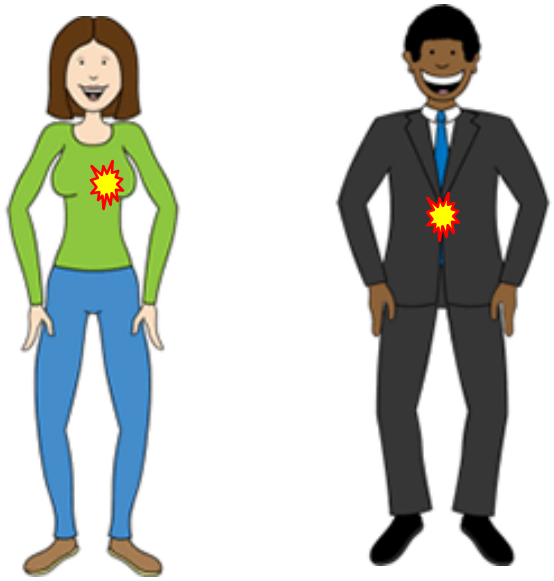
Germline mutations

- If a nucleotide substitution occurs in germ cells:
  → it will be present in all cells of the developing individual
  → transmissible to the progeny (inheritable)



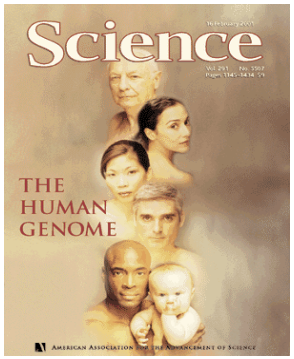
Somatic mutations

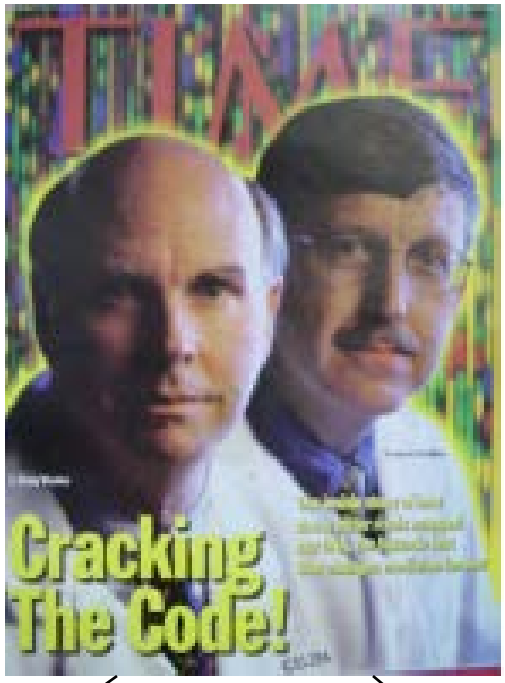
- If a nucleotide substitution occurs in somatic cells:
  → present only in the cells that derive from the one where the substitution took place
  → NOT inheritable

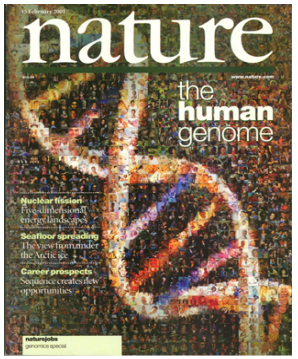


Advances in sequencing technologies

2001: Sequencing of the Human Genome

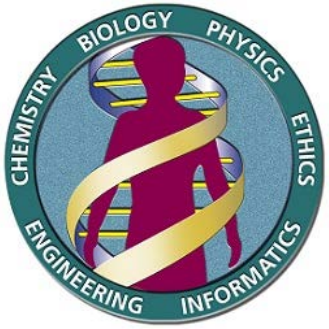






Craig Venter
Celera

Francis Collins
HGP-NIH



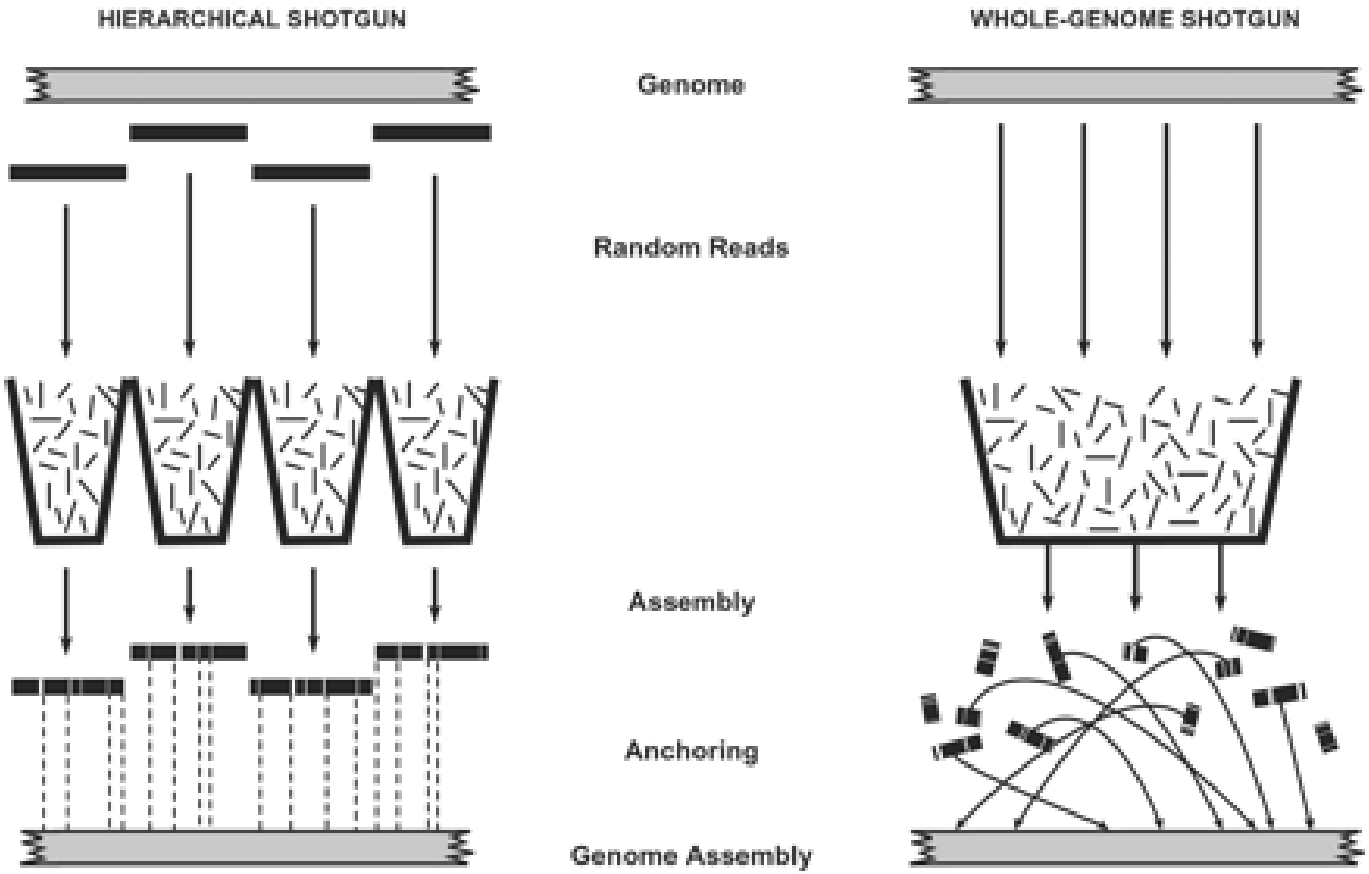
Human Genome Project

- 1986: first announcement of the human genome initiative
- 1990: a predicted 15-year project formally begins
- 1999: first billion bp sequenced
- 2000: first draft of the human genome completed (published 2001)
- 2003: HGP declared complete (*May 2006: last human chromosome (chr 1) completed*)

Time: 13 (20?) years

Cost: ~ $ 1 billion

Sequencing strategy



Human Genome Project (HGP)

Celera Genomics

Waterston, Lander, Sulston (2002) PNAS 99(6):3712-3716

Two necessary steps:

1) Amplification of DNA fragments (PCR)

2) Determination of nucleotide sequence

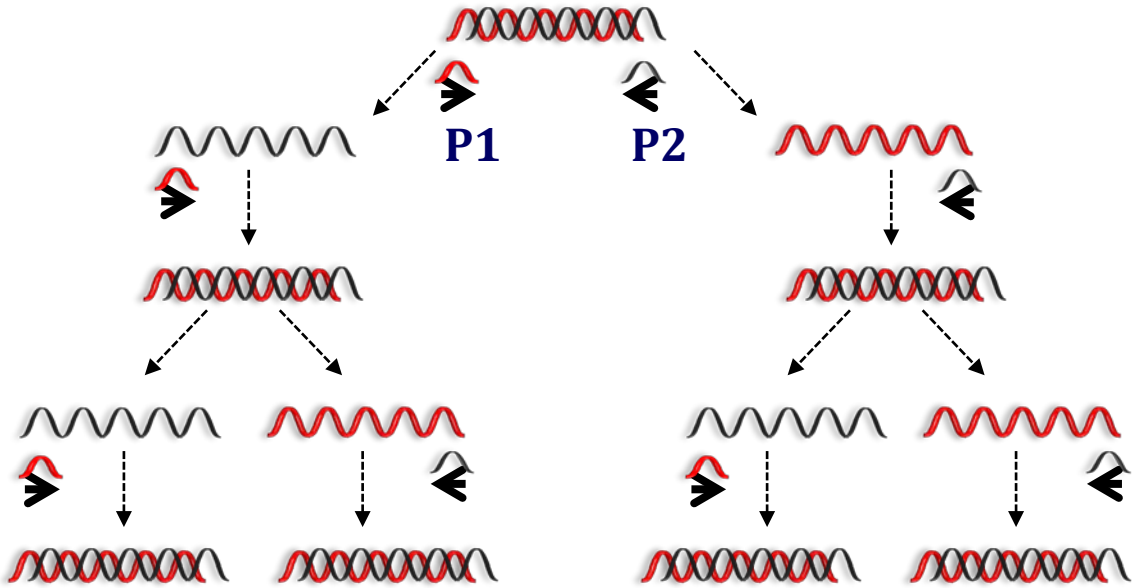
Amplifying DNA - PCR

PCR = Polymerase Chain Reaction

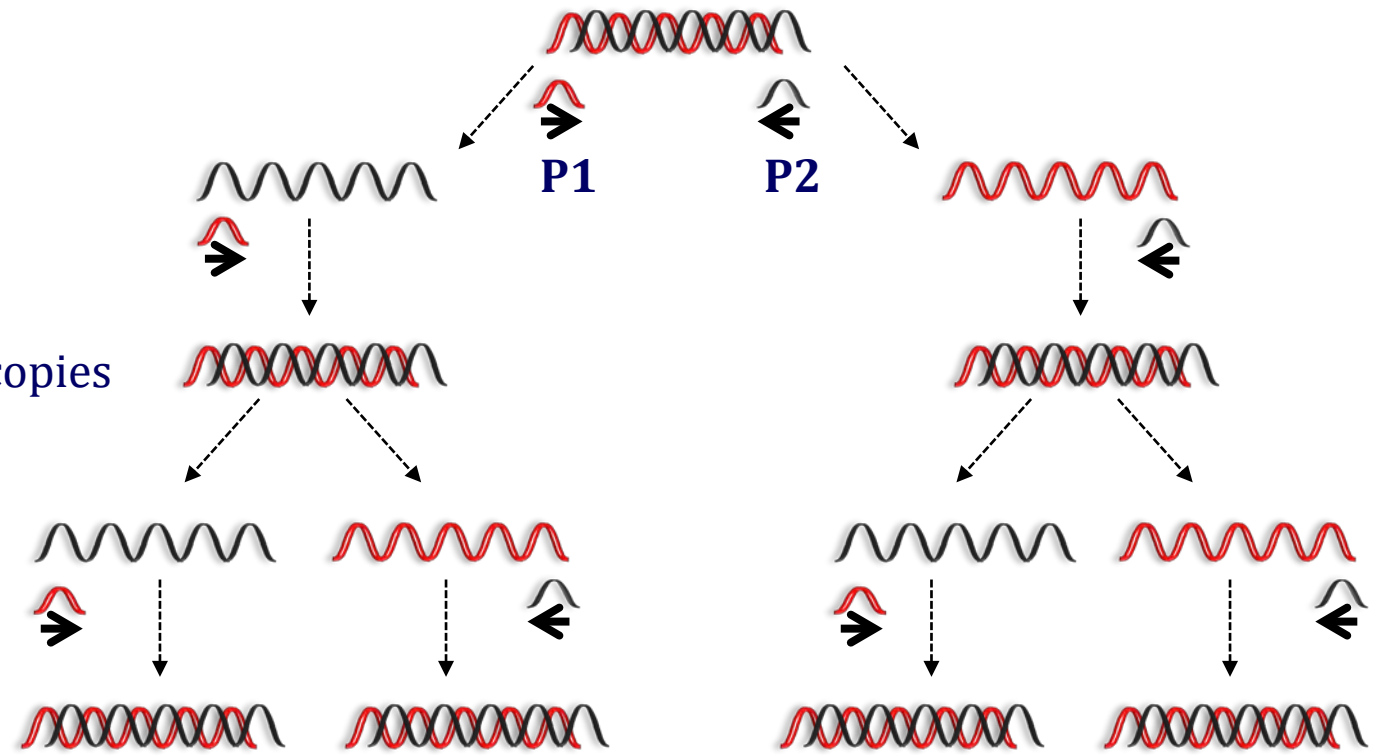
Invented in1983 by Kary Mullis → 1993, Nobel Prize in Chemistry.

Exploits a special thermostable DNA polymerase (Taq polymerase) to amplify specific portions of the genome.

Two short sequences of DNA flanking the region of interest (*primers*, P1 and P2) are necessary to initiate the reaction. DNA is *denatured* to single strands, primers recognize their complementary sequences, Taq polymerase is added to elongate the DNA strand.
The reaction is repeated for several cycles.



PCR: amplification kinetics



Cycle 1: 2^1 (=2) copies

Cycle 2: 2^2 (=4) copies

Cycle 3: 2^3 (=8) copies……

Cycle 30: 2^{30} copies

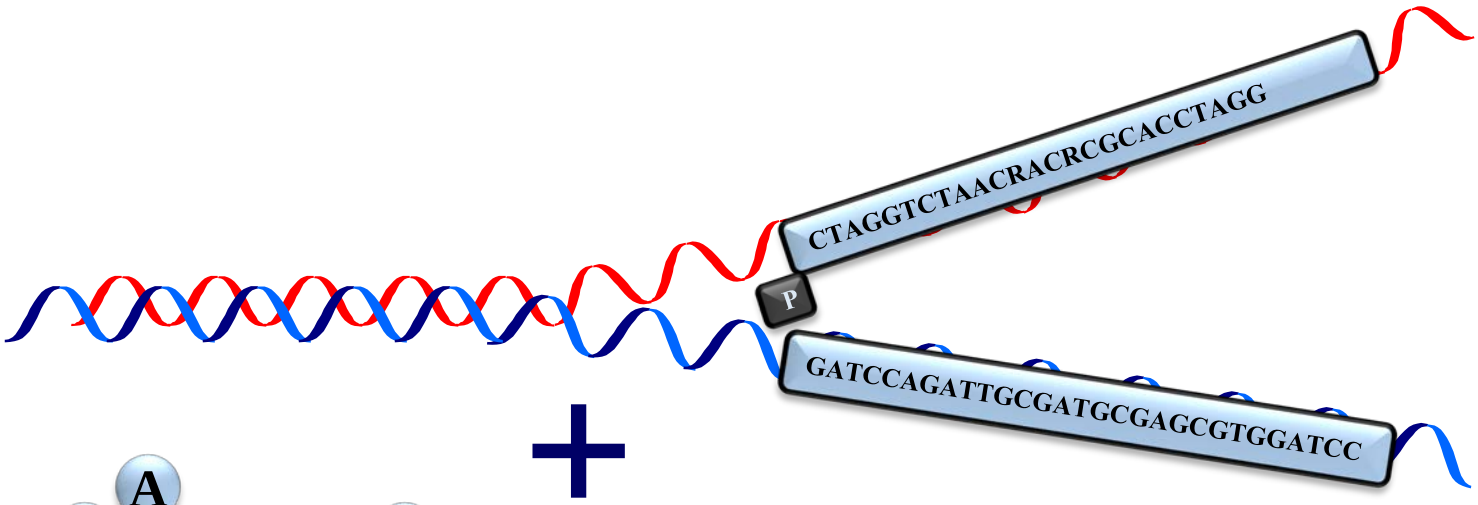
DNA sequencing

The method that underlies most sequencing approaches was originally proposed in 1977 by Frederick Sanger, who is the only chemist to have received two Nobel Prizes in Chemistry, the first as the sole recipient in 1958 for his work as the first to sequence a protein, the sequencing of insulin; and the second in 1980, shared with Paul Berg and Walter Gilbert, for the sequencing of nucleic acids.

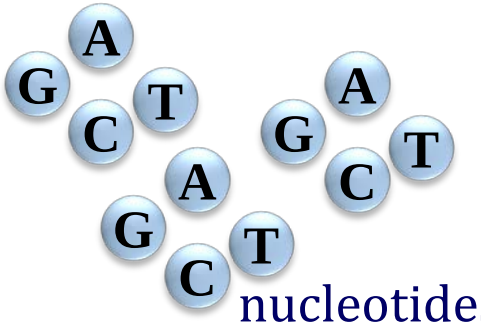
Sanger sequencing exploits the activity of a natural enzyme, DNA polymerase, which synthesizes DNA molecules from free nucleotides and is at the basis of DNA replication.

The principle underlying Sanger sequencing implies the use of modified nucleotides that do not allow further synthesis and thus terminate the elongation of the nascent DNA chain.

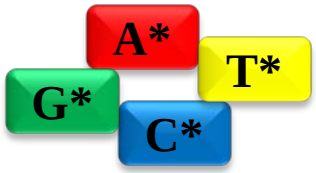
DNA sequencing



+



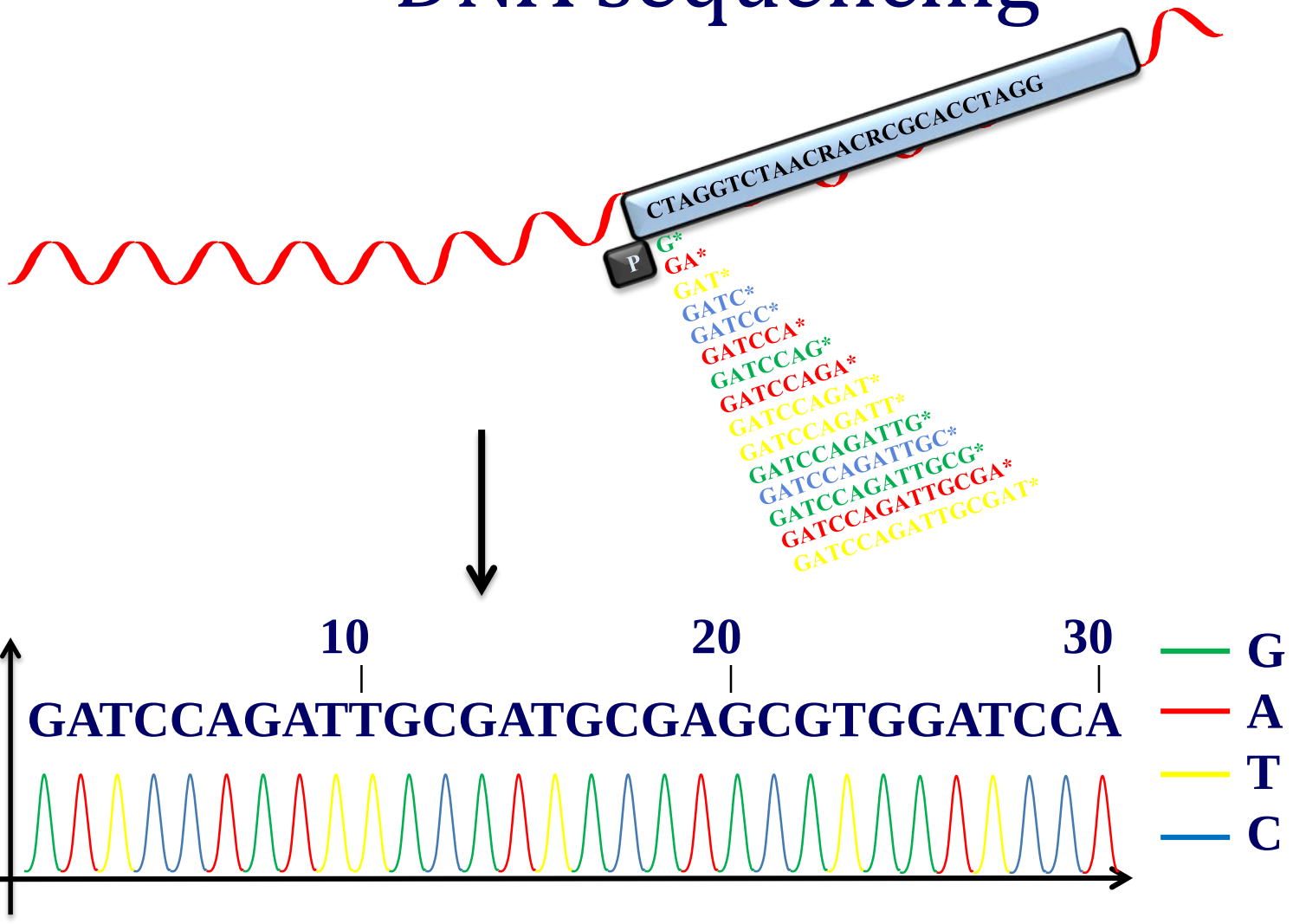
nucleotides



Terminator nucleotides

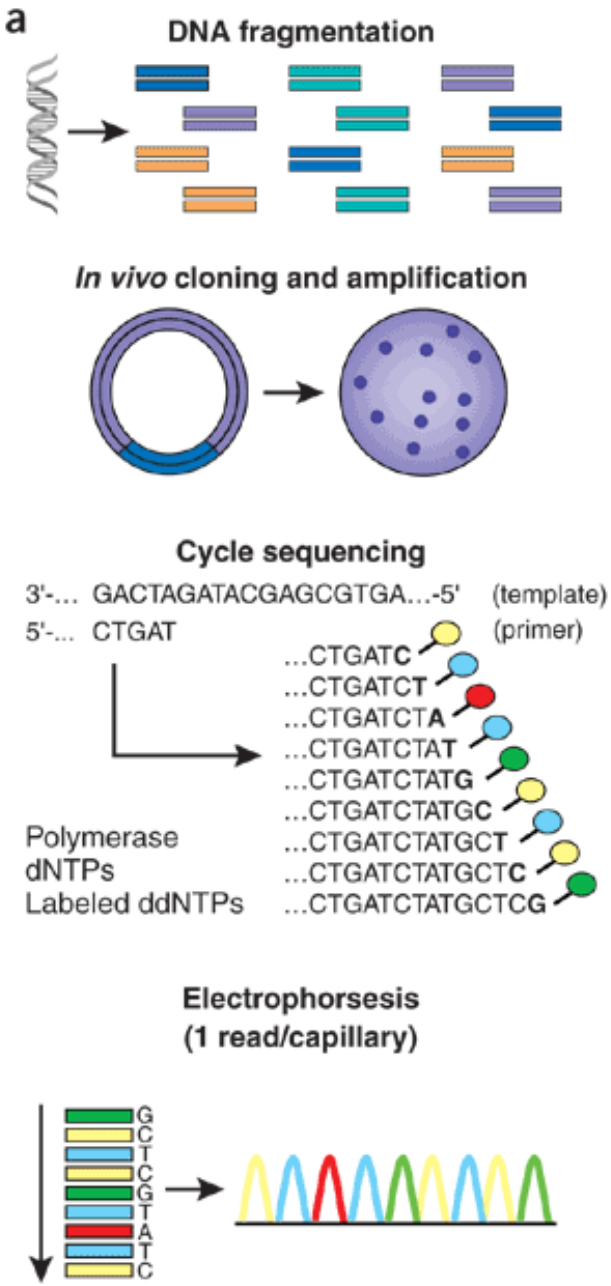
+ polymerase

DNA sequencing



Basic principles
Sanger sequencing

- DNA amplification or cloning of fragments to be sequenced
- Multiple cycles of terminator nucleotide incorporation
- Discrimination of sequence based on size of individual fragments



Jay Shendure & Hanlee Ji
Nature Biotechnology 26, 1135 - 1145 (2008)

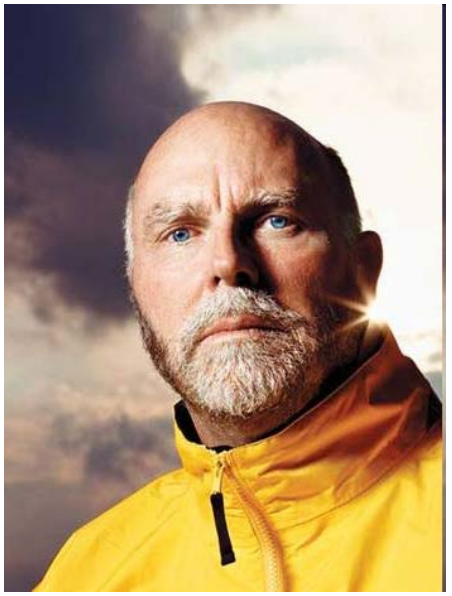
2007

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, 4 Genetics Department, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

PLoS Biology | www.plosbiology.org 2113 October 2007 | Volume 5 | Issue 10 | e254



Craig Venter's genome (Sanger method)

Cost: $ 70,000,000

Time: 3 years 2004-2007

2008

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

nature Vol 452|17 April 2008|

James Watson's genome

Cost: $ 1,000,000

Time: 2 months

November 2008 in Nature:

Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

Sequence of the complete diploid genome of male Yoruba from Nigeria
135 Gb (~25x depth of coverage)
Cost: ~ $ 250,000
Time: 8 weeks

The diploid genome sequence of an Asian individual

Jun Wang[1,2,3,4]*, Wei Wang[1,3]*, Ruiqiang Li[1,3,4]*, Yingrui Li[1,5,6]*, Geng Tian[1,7], Laurie Goodman[1], Wei Fan[1], Junqing Zhang[1], Jun Li[1], Juanbin Zhang[1], Yiran Guo[1,7], Binxiao Feng[1], Heng Li[1,8], Yao Lu[1], Xiaodong Fang[1], Huiqing Liang[1], Zhenglin Du[1], Dong Li[1], Yiqing Zhao[1,7], Yujie Hu[1,7], Zhenzhen Yang[1], Hancheng Zheng[1], Ines Hellmann[9], Michael Inouye[8], John Pool[9], Xin Yi[1,7], Jing Zhao[1], Jinjie Duan[1], Yan Zhou[1], Junjie Qin[1,7], Lijia Ma[1,7], Guoqing Li[1], Zhentao Yang[1], Guojie Zhang[1,7], Bin Yang[1], Chang Yu[1], Fang Liang[1,7], Wenjie Li[1], Shaochuan Li[1], Dawei Li[1], Peixiang Ni[1], Jue Ruan[1,7], Qibin Li[1,7], Hongmei Zhu[1], Dongyuan Liu[1], Zhike Lu[1], Ning Li[1,7], Guangwu Guo[1,7], Jianguo Zhang[1], Jia Ye[1], Lin Fang[1], Qin Hao[1,7], Quan Chen[1,5], Yu Liang[1,7], Yeyang Su[1,7], A. san[1,7], Cuo Ping[1,7], Shuang Yang[1], Fang Chen[1,7], Li Li[1], Ke Zhou[1], Hongkun Zheng[1,4], Yuanyuan Ren[1], Ling Yang[1], Yang Gao[1,6], Guohua Yang[1,2], Zhuo Li[1], Xiaoli Feng[1], Karsten Kristiansen[4], Gane Ka-Shu Wong[1,10], Rasmus Nielsen[9], Richard Durbin[8], Lars Bolund[1,11], Xiuqing Zhang[1,6], Songgang Li[1,2,5], Huanming Yang[1,2,3] & Jian Wang[1,2,3]

Sequence of the complete diploid genome of an Asian individual
117 Gb (~20x depth of coverage)
Cost: < $ 300,000
Time: 1.5 months

Summary:

- 1990: 13 years and $ 1 billion to get the reference human genome sequence (8 individuals in the HGP, 5 individuals at Celera)
- 2004-7: 3 years and $ 70 million to get Venter's sequence
- 2007-8:2 months and $ 1 million for Watson's sequence
- 2008: 2 months and $ 250,000 for other complete diploid sequences
- 2016: few hours and less than $ 1,000 for a human genome sequencing

WHAT HAPPENNED??

Next Generation Sequencing

Also known as:

- High-throughput sequencing
- Massively parallel sequencing
- Deep sequencing
- Saturation sequencing

Method of the Year

NATURE METHODS | VOL.5 NO.1 | JANUARY 2008 |

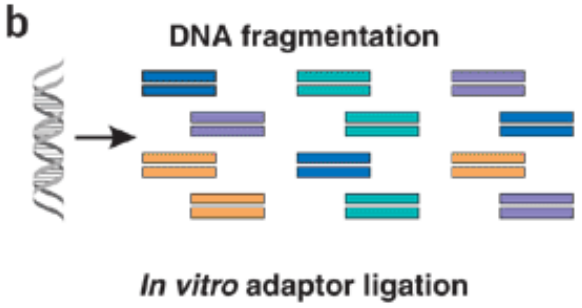
Technology

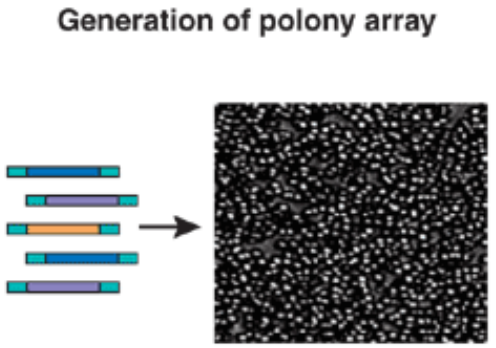
Currently, 3 platforms for NGS are reasonably widespread:

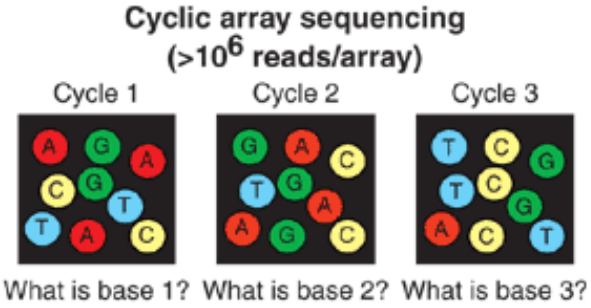
- 454 FLX (Roche) - 2000 → 2005
- Solexa Genome Analyzer (Illumina) - 2006
- SOLiD System (Applied Biosystems) - 2007

Basic principles
Next-Generation Sequencing

- Solid phase support to hold large numbers of individual DNA molecules
- Sequencing via multiple cycles of incorporation using different fluorophores for each nucleotide
- Images per cycle provide sequence data
- Based on short reads from a large number of molecules







Jay Shendure & Hanlee Ji
Nature Biotechnology 26, 1135 - 1145 (2008)

NGS main advantages

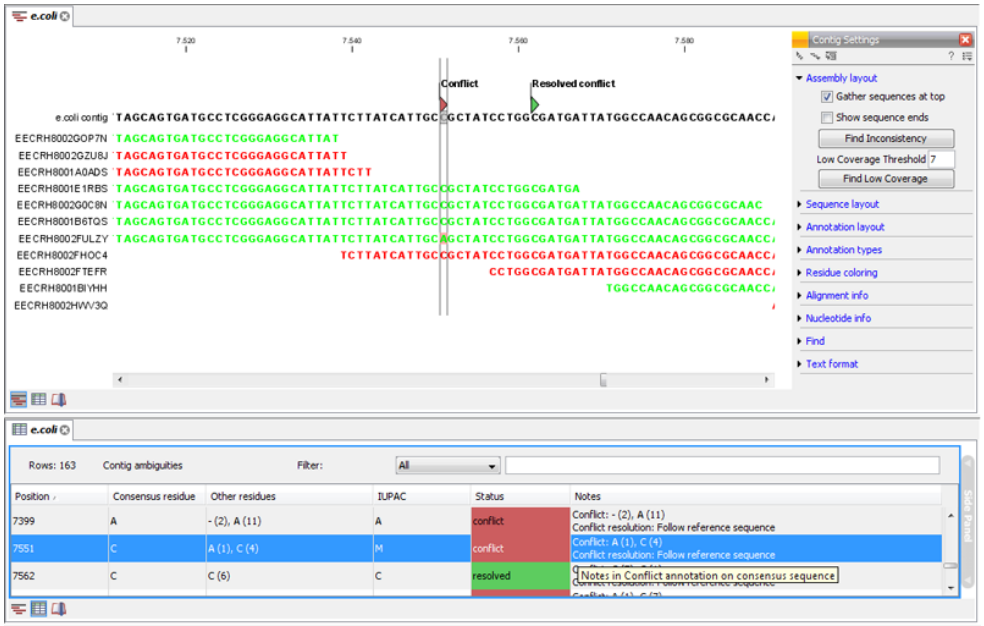
Reduction of costs due to short reads

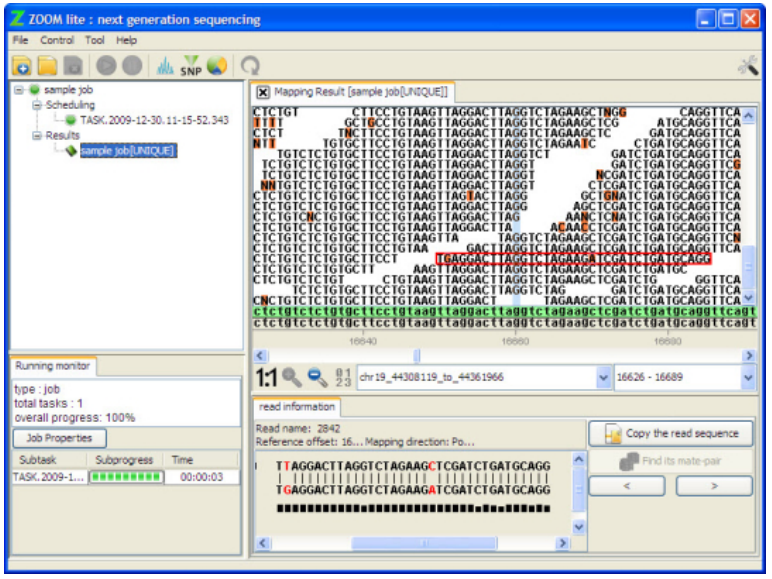
Reduction of time:

- many molecules sequenced in parallel
- no cloning
- 4 nucleotides simultaneously
- short reads

Possibility of very high coverage.

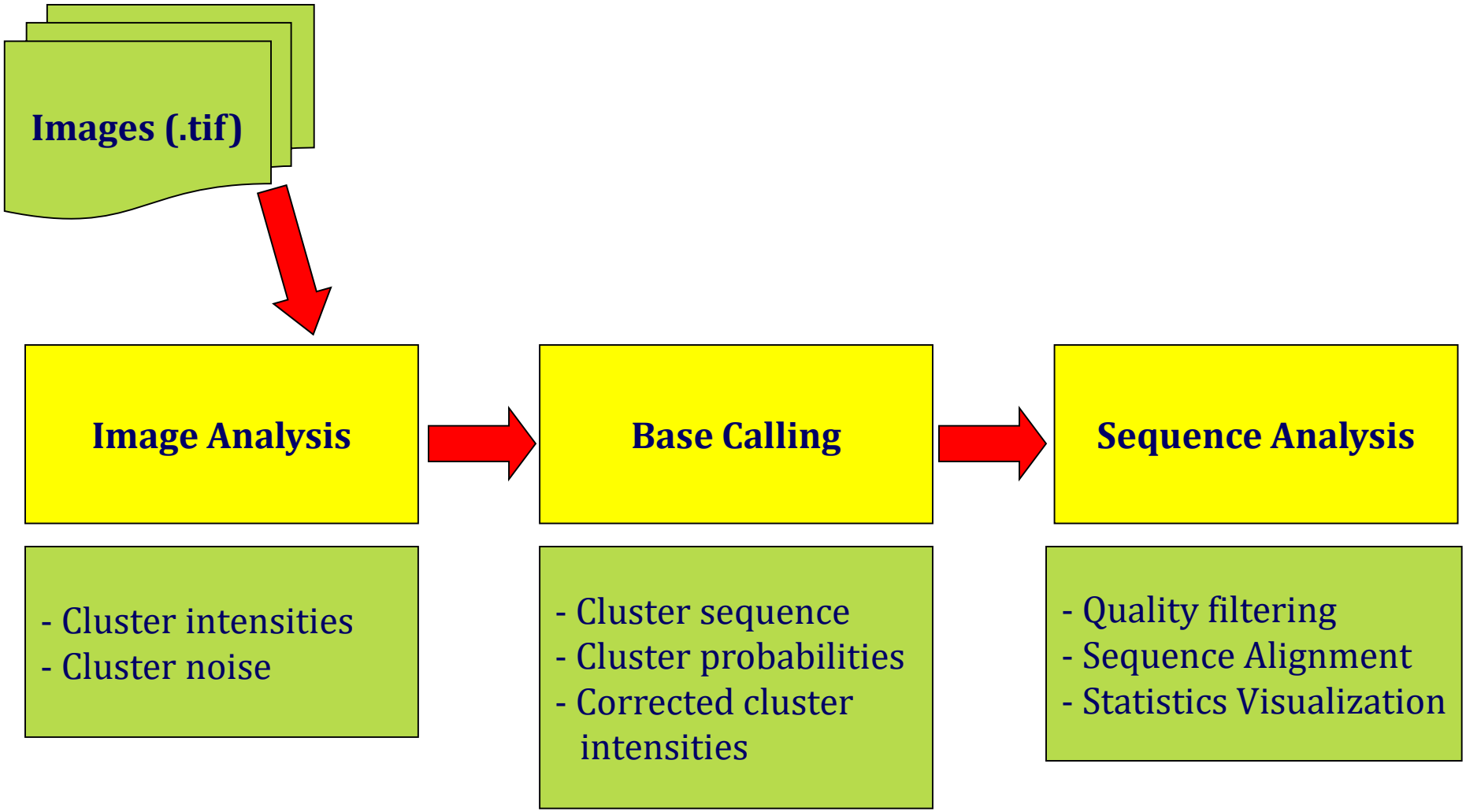
Coverage = how many times (on average) was each nucleotide in the genome sequenced?



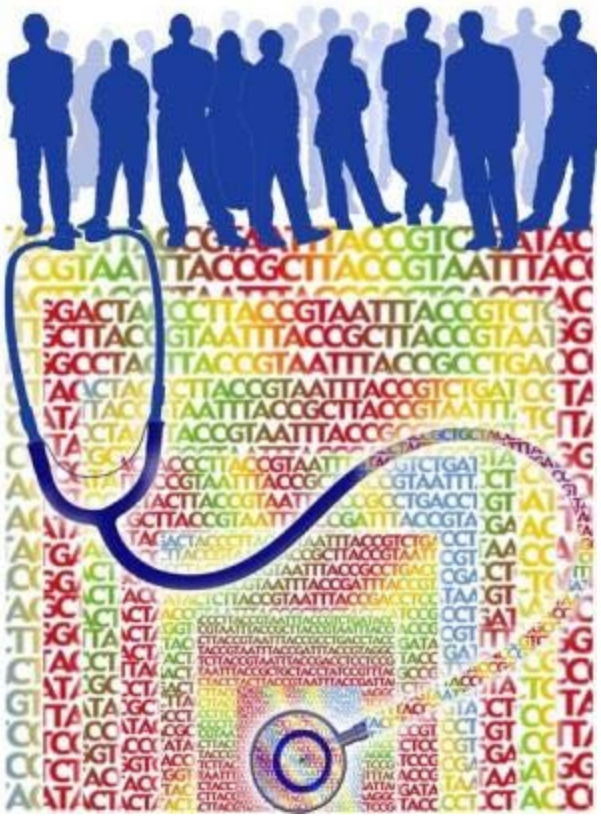


The necessary coverage depends on what you are looking at …..
For mutational analysis, minimum 30-40x

Data Analysis workflow



GA Pipeline user guide, Illumina



Applications of NGS

Next Generation Sequencing

Applications 1 – Genome Sequencing

De novo sequencing (microbial genomes)

Resequencing of genomes:

- → genetic variations (SNPs, CNV, InDels, etc)
- Mutations
- Breakpoints
- Deletions

Why is it so important to sequence individual genomes?

Genes and disease

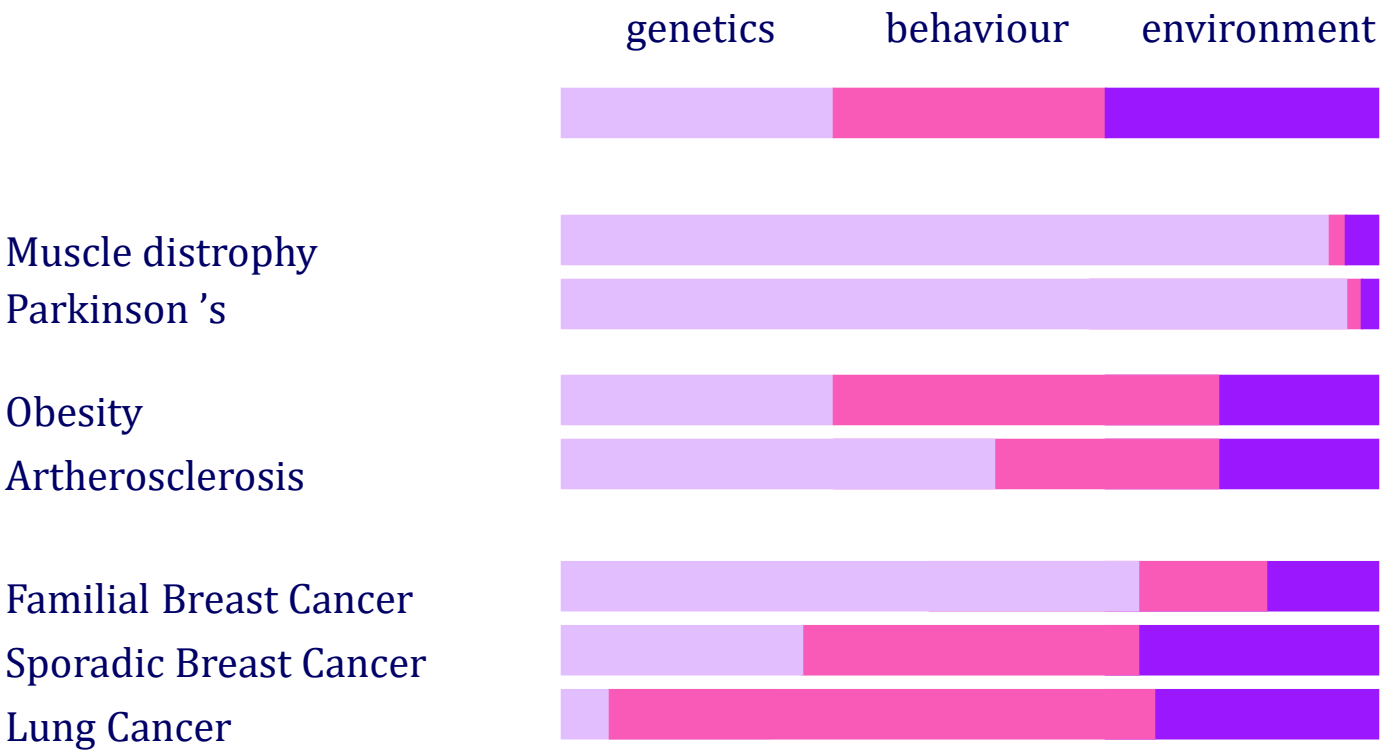
Table 1 • Relative frequency of types of mutations underlying disease phenotypes*

Change	Number	% of total
deletion	6,085	21.8
insertion/duplication	1,911	6.8
complex rearrangement	512	1.8
repeat variations	38	0.1
missense/nonsense	16,441	58.9
splicing	2,727	9.8
regulatory	213	0.8
total	27,027	100.0

*Data are from the Human Gene Mutation Database (June 2002).

Multifactorial Diseases

Most disease phenotypes have 3 components:



This rule applies to social behaviour:

Genetic predisposition

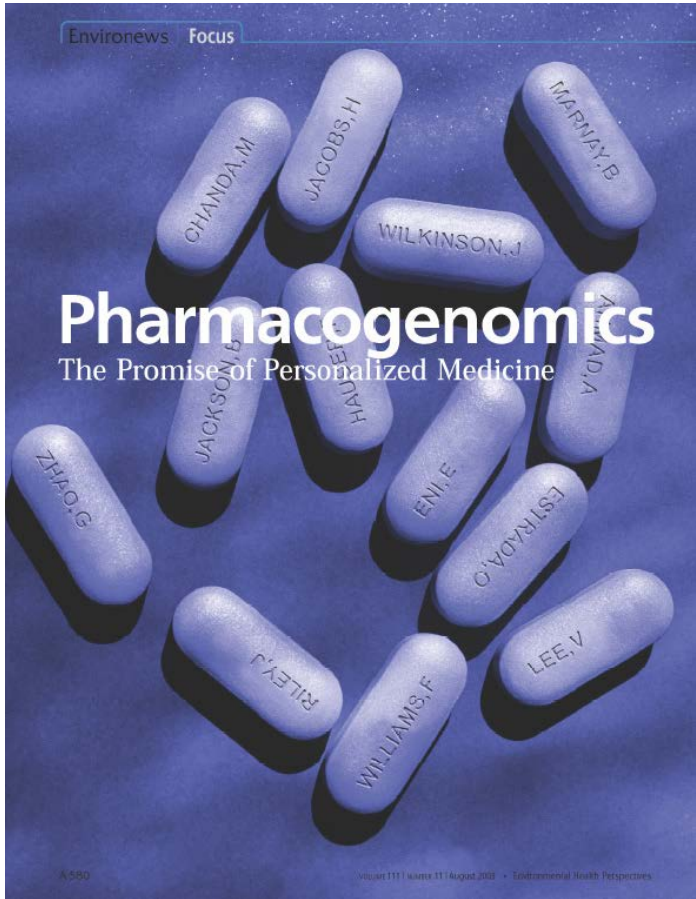
- Multifactorial diseases are the most frequent (obesity, cardiovascular diseases, diabetes, inflammatory bowel diseases, bipolar disorder, etc.), and in most cases the underlying genes are unknown.
- Knowing who carries genetic variants that predispose for specific diseases would represent a very powerful tool for implementing targeted protocols for disease prevention.

Pharmacogenomics

"Pharmacogenomics is a science that examines the inherited variations in genes that dictate drug response and explores the ways these variations can be used to predict whether a patient will have a good response to a drug, a bad response to a drug, or no response at all."

SOURCE: NCBI A Science Primer

Aim: personalized medicine



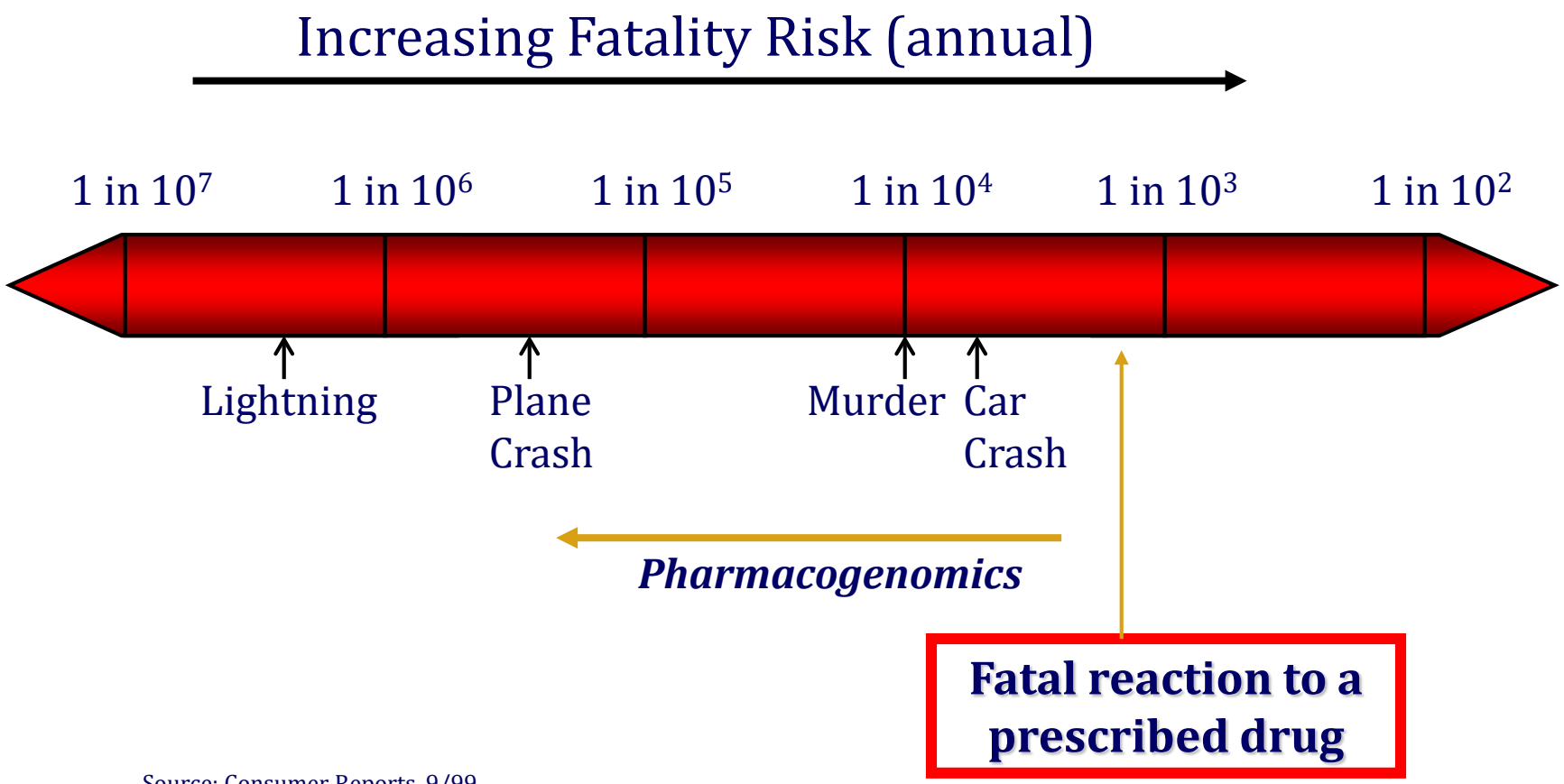
What is Personalised Medicine?

Personalised Medicine is about enabling clinicians to prescribe the:

- *Right* Drug
- At the *Right* Dose
- For the *Right* Disease
- To the *Right* Patient

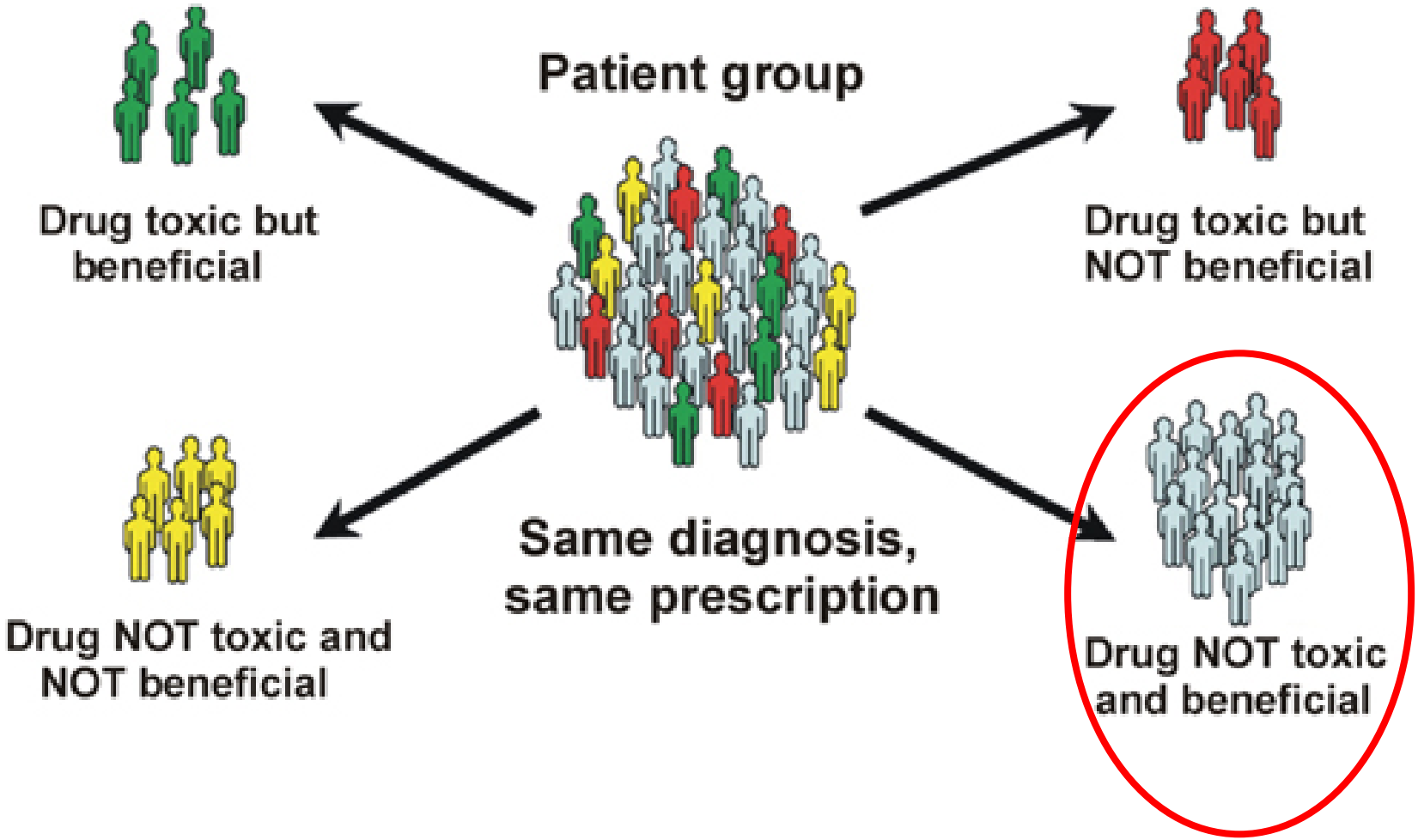
and to know all this *prior* to the patient taking the medicine

Fatality Risk Comparison



Source: Consumer Reports, 9/99

Why use genetic information in pharmacy practice?



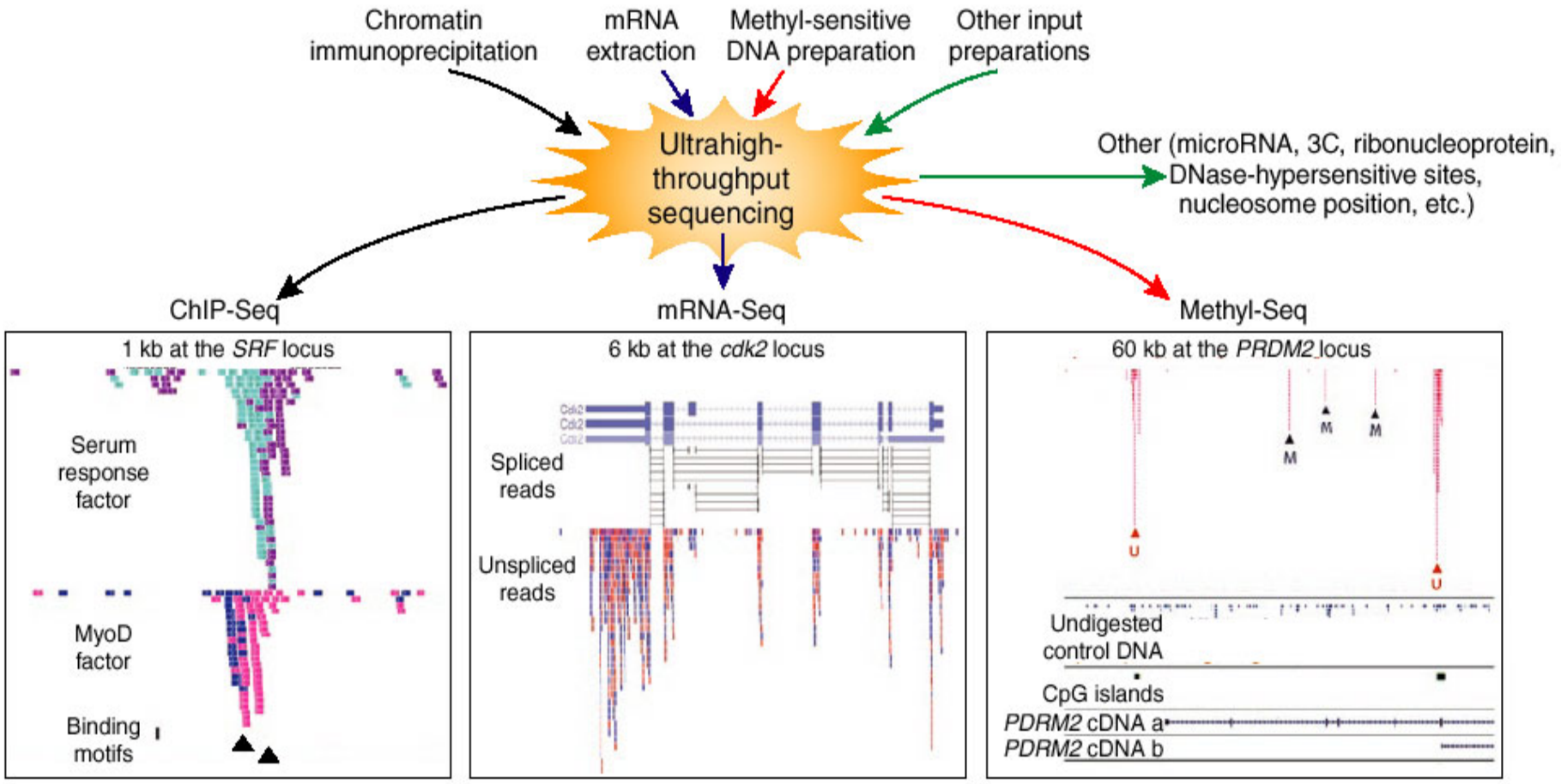
Why is it so important to sequence individual genomes?

It represents the first step towards personalized medicine.

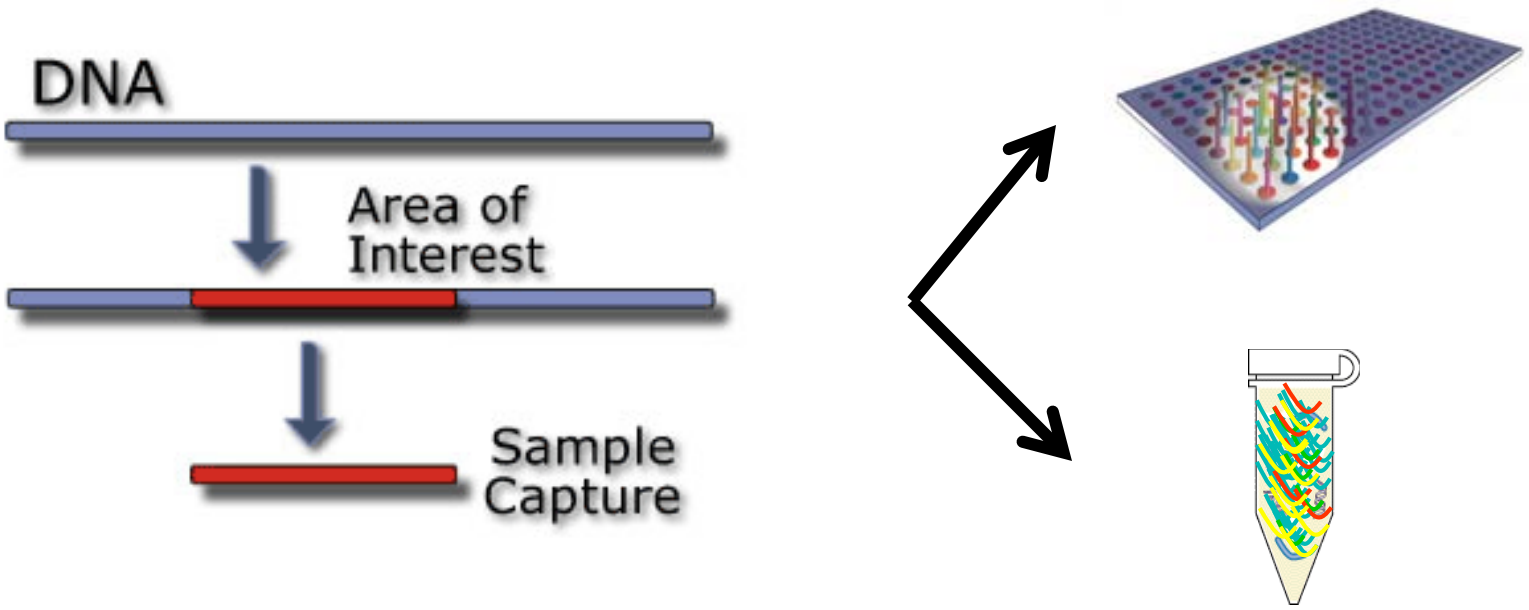
Next Generation Sequencing

Applications 2 – Functional Genomics

Sequence Census Methods: Genome Function

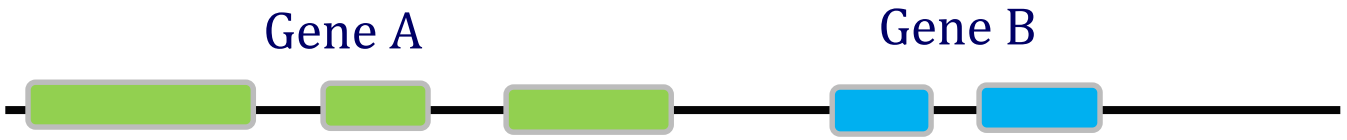


Target Enrichment



There are several companies that offer solutions for isolating specific portions of the genome: can be ready made or customized.

Exome sequencing



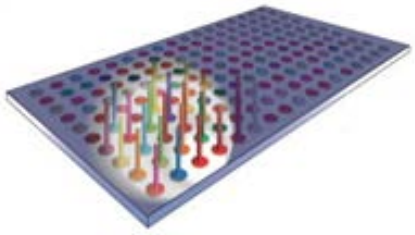
Reference DNA
(exons+introns)

determine exome and design probes

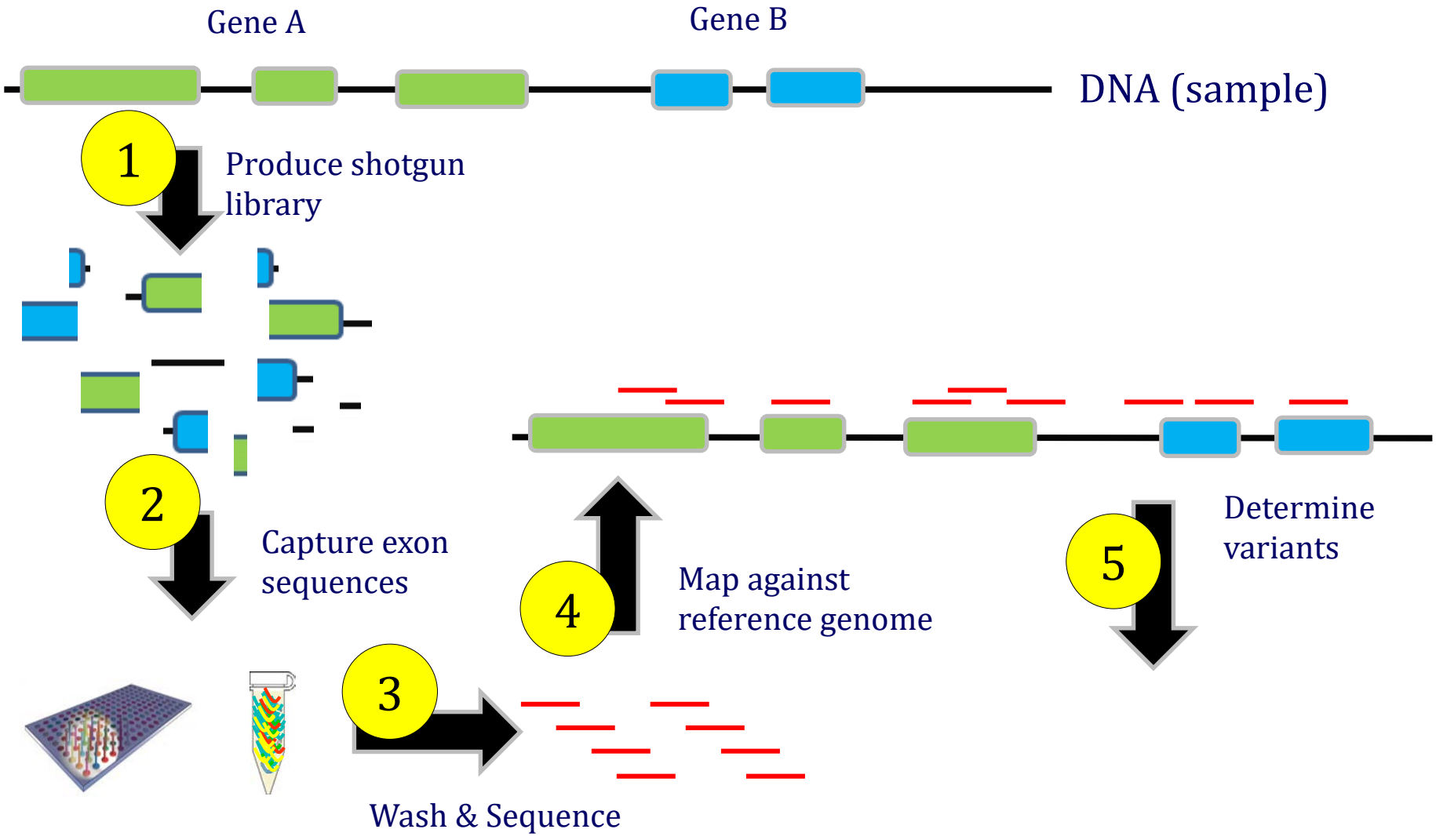
— probes

Bioinformatics can play a role with design of probes

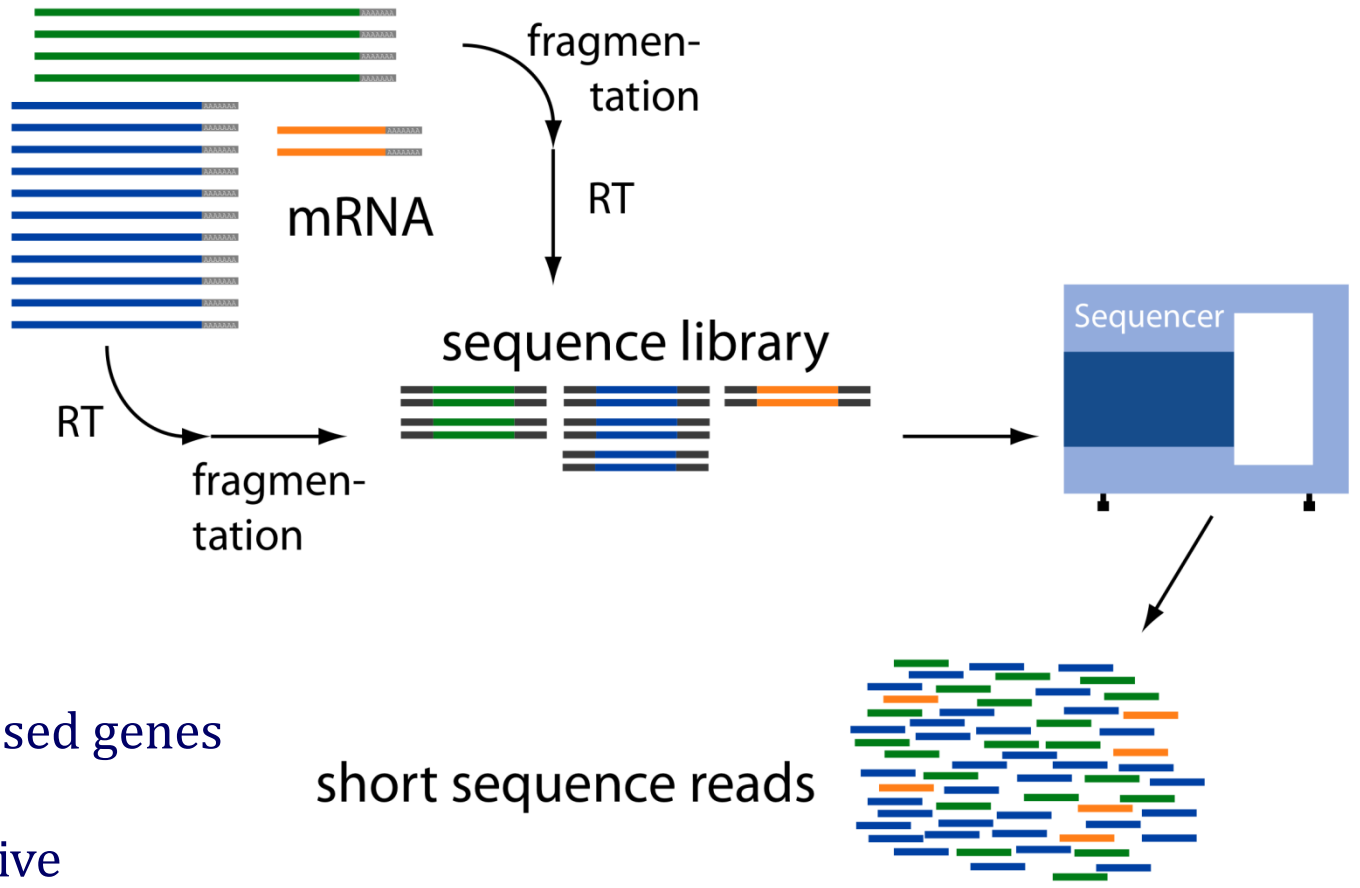
construct microarray or oligo library



Exome sequencing



mRNA sequencing

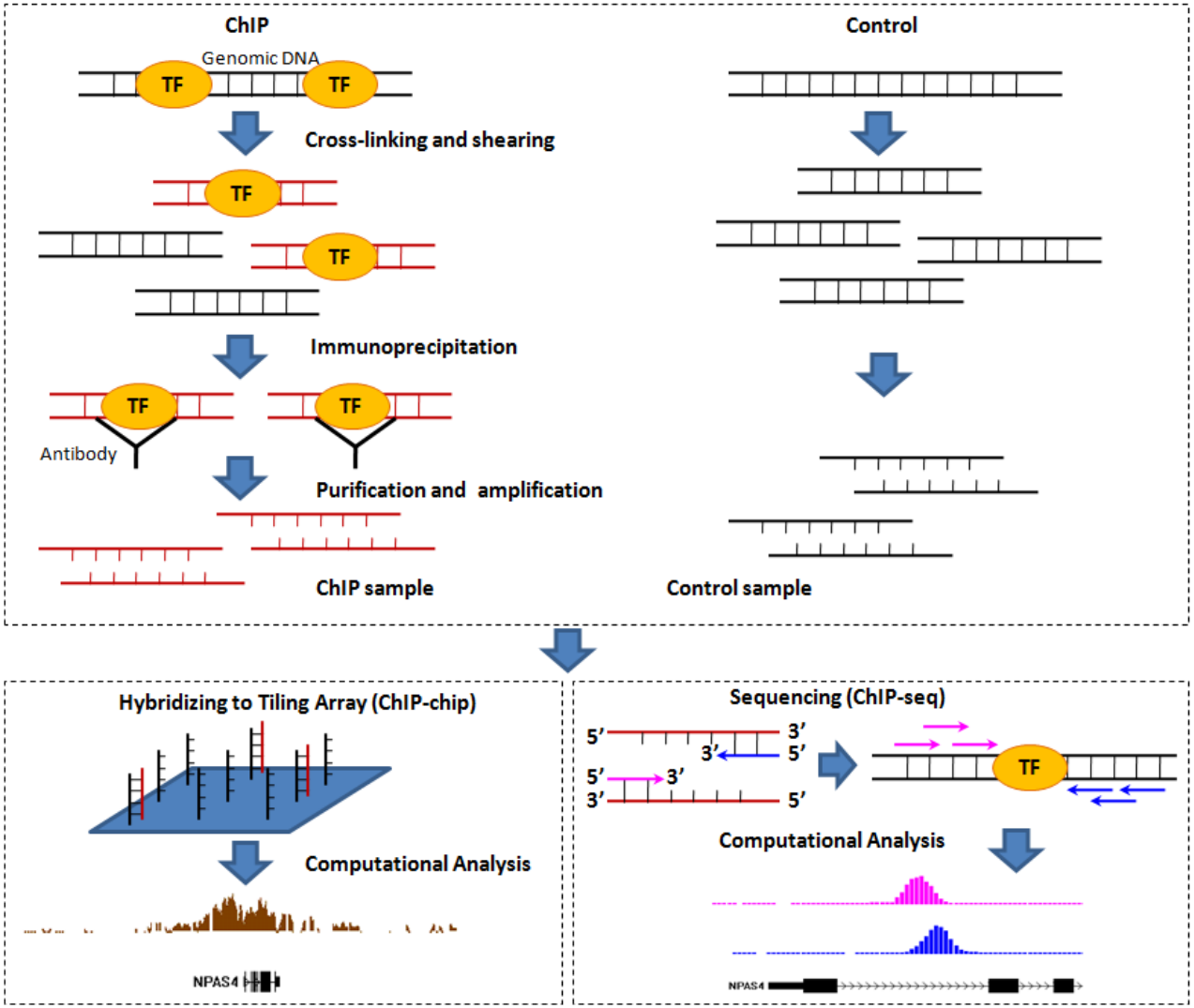


Only expressed genes

Is quantitative

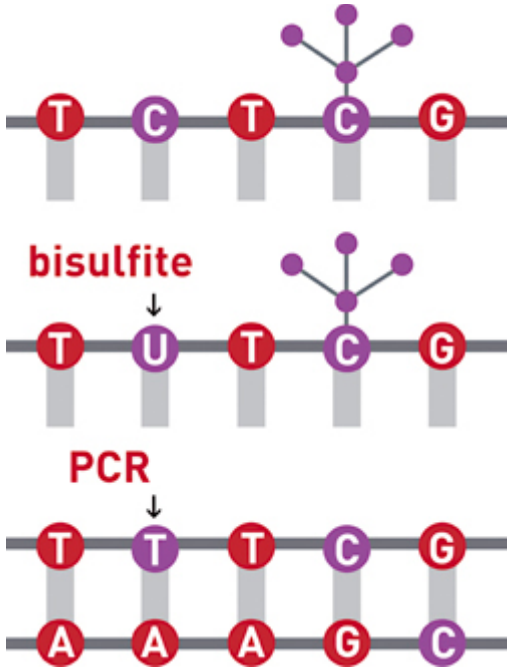
ChIP-Seq

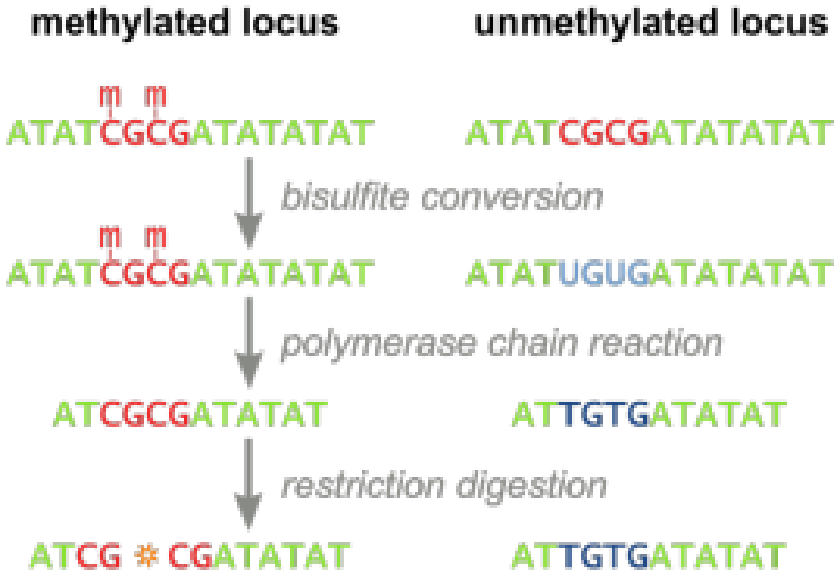
= transcription factor binding sites, distribution of histone marks



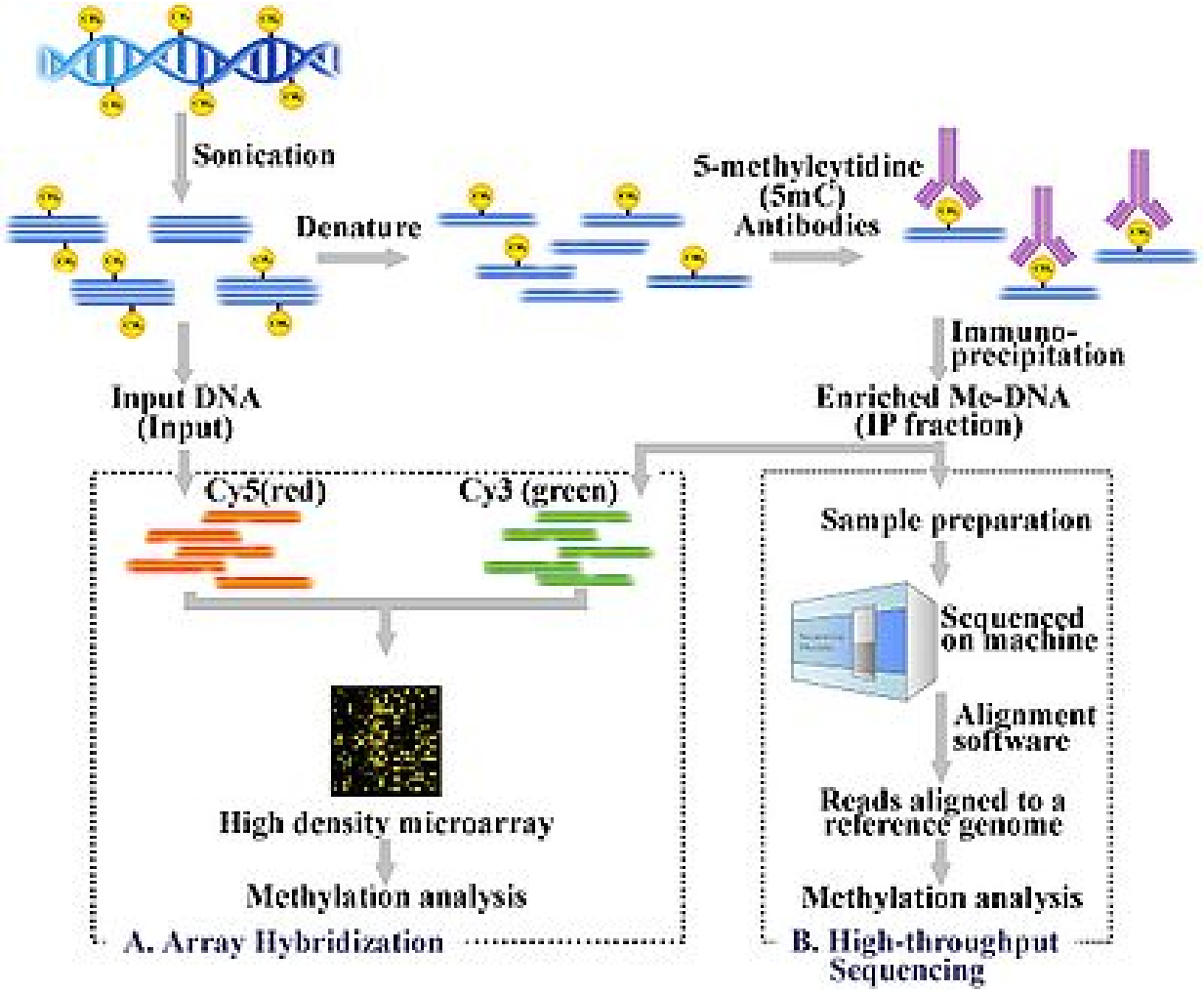
Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

Methylated DNA – Bisulfite conversion





Methylated DNA - MeDIp



The future:
Next-Next Generation Sequencing

Many companies are developing platforms that work on single molecules (no PCR step required) and should:

- Increase throughput (higher density devices)
- Increase speed
- Lower costs (different chemistry, elimination of PCR step)

Social Issues

Genetic information is
personal
powerful
potentially predictive
pedigree-sensitive
permanent
prejudicial

Y-GA 99-1395

From US Department of Energy Human Genome Program, http://www.ornl.gov/hgmis