METHODS IN MOLECULAR BIOLOGY

Lecturer: Ming-Hsien Tsai, PHD.
Assistant researcher
Center for Lipid bioscience, KMUH
Lipid Science and Aging Research Center, KMU

Watson and Crick (1953)

Watson, Crick, and Maurice Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"

James Dewey Watson

- James Dewey Watson (born April 6, 1928) is an American molecular biologist, geneticist and zoologist, best known as one of the co-discoverers of the structure of DNA in 1953 with Francis Crick and Rosalind Franklin.
- Watson, Crick, and Maurice Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".
- Watson has written many science books, including the textbook Molecular Biology of the Gene (1965) and his bestselling book The Double Helix (1968).

James Watson

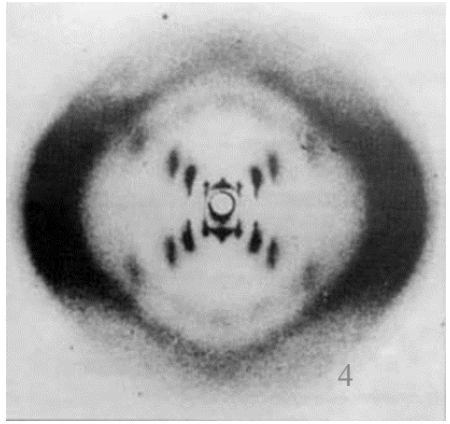
Born	James Dewey Watson April 6, 1928 (age 89)[1] Chicago, Illinois, United States
Nationality	United States
Fields	Genetics
Institutions	Indiana University Cold Spring Harbor Laboratory Laboratory of Molecular Biology Harvard University University of Cambridge National Institutes of Health
Alma mater	University of Chicago (B.S., 1947) Indiana University (Ph.D., 1950)
Thesis	*The Biological Properties of X-Ray Inactivated Bacteriophage* (1951)

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

Rosalind Elsie Franklin

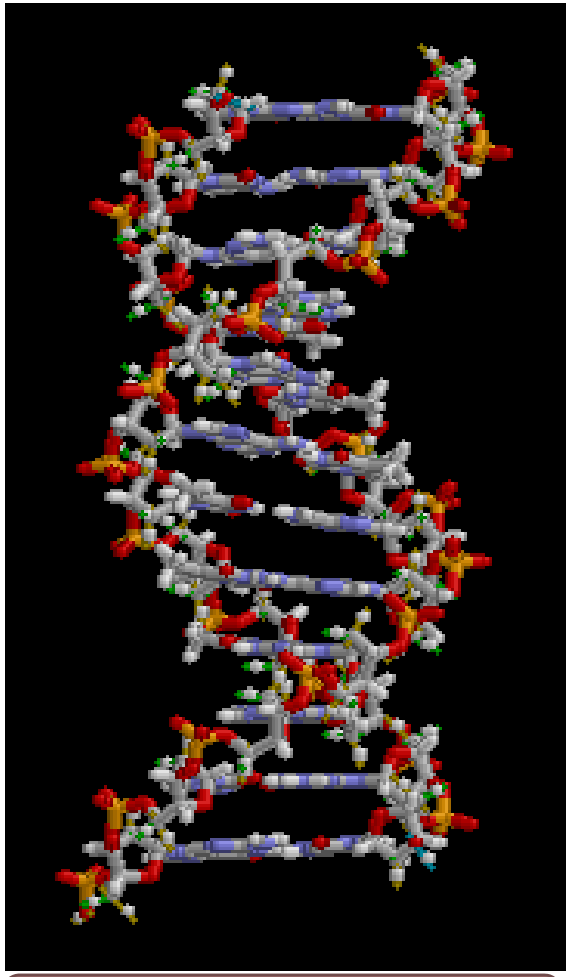
- Rosalind Elsie Franklin (25 July 1920 – 16 April 1958) was an English chemist and X-ray crystallographer who made contributions to the understanding of the molecular structures of DNA (deoxyribonucleic acid), RNA (ribonucleic acid), viruses, coal, and graphite.
- Although her works on coal and viruses were appreciated in her lifetime, her contributions to the discovery of the structure of DNA were largely recognized posthumously.
- Franklin is best known for her work on the X-ray diffraction images of DNA, particularly Photo 51.

Born	Rosalind Elsie Franklin 25 July 1920 Notting Hill, London, UK
Died	16 April 1958 (aged 37) Chelsea, London, UK Ovarian cancer



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

What is DNA?

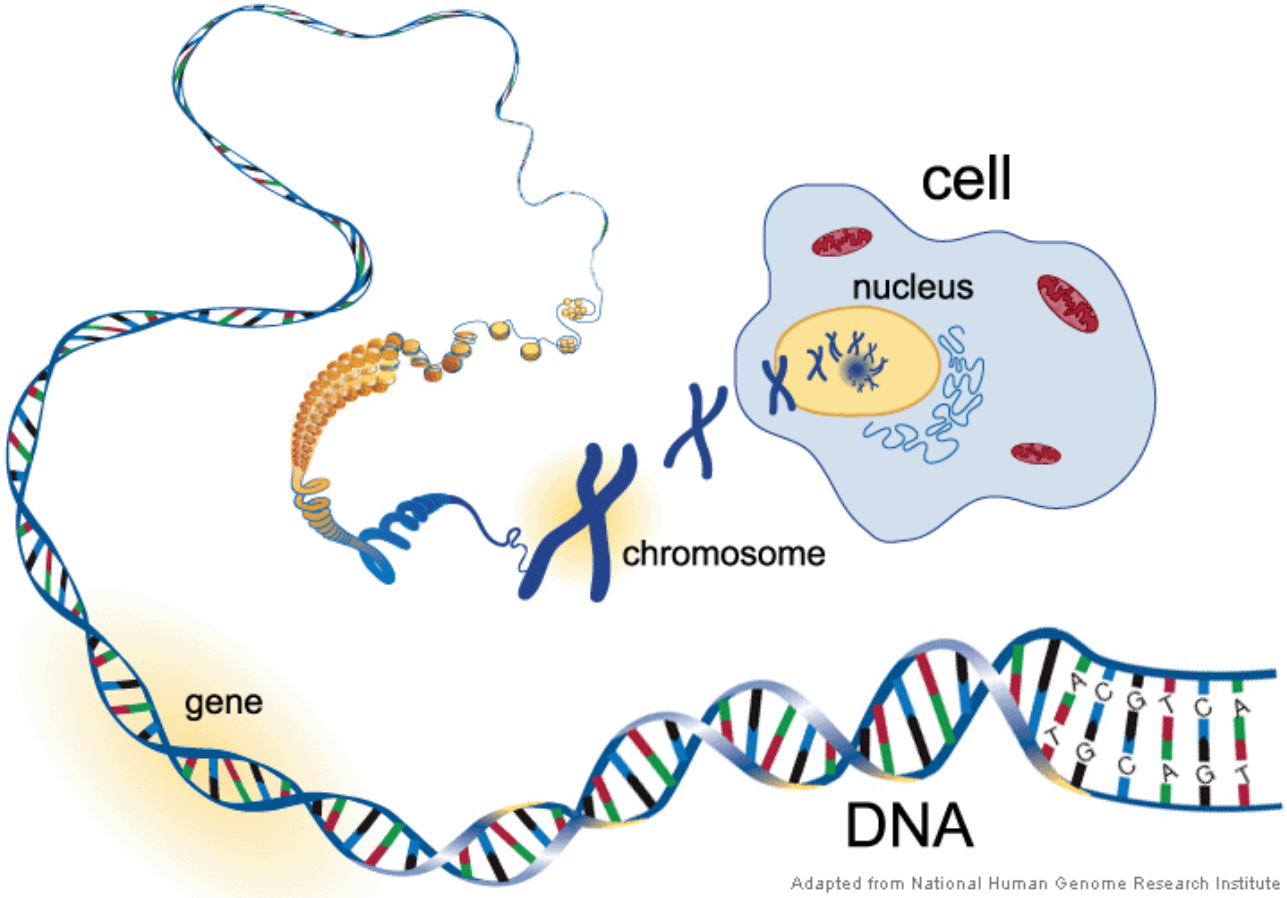


The structure of part of a DNA double helix

- Deoxyribonucleic acid
- A molecule that carries most of the genetic instructions used in the development, functioning and reproduction of all known living organisms.
- Consist of two biopolymer strands coiled around each other to form a double helix

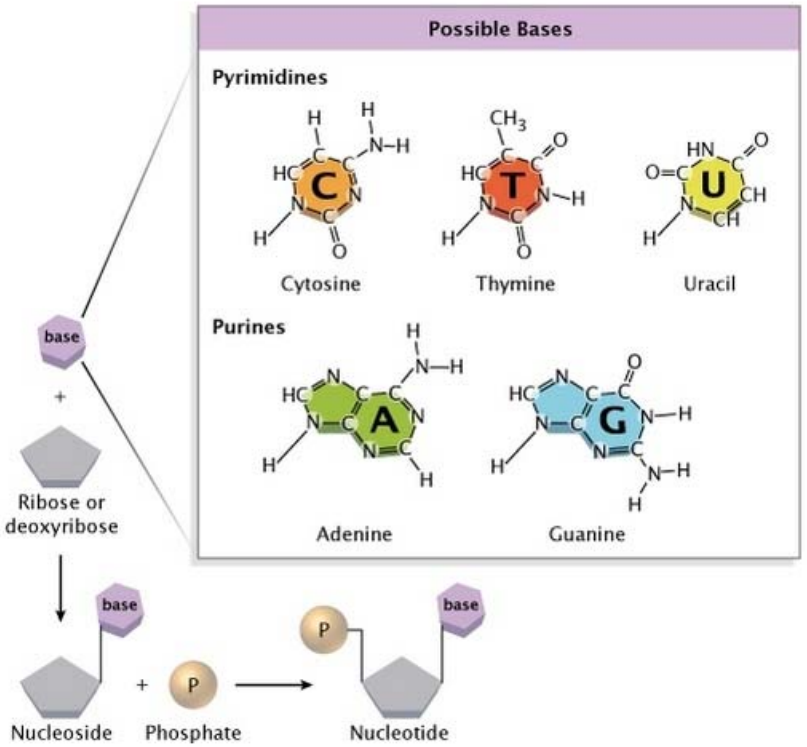
In terms of decreasing size:

Nucleus → Chromosome → Gene → DNA

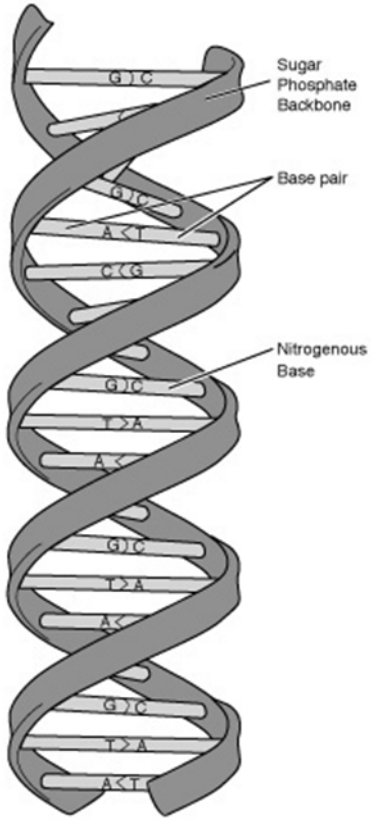


Adapted from National Human Genome Research Institute

Chemical structure of DNA



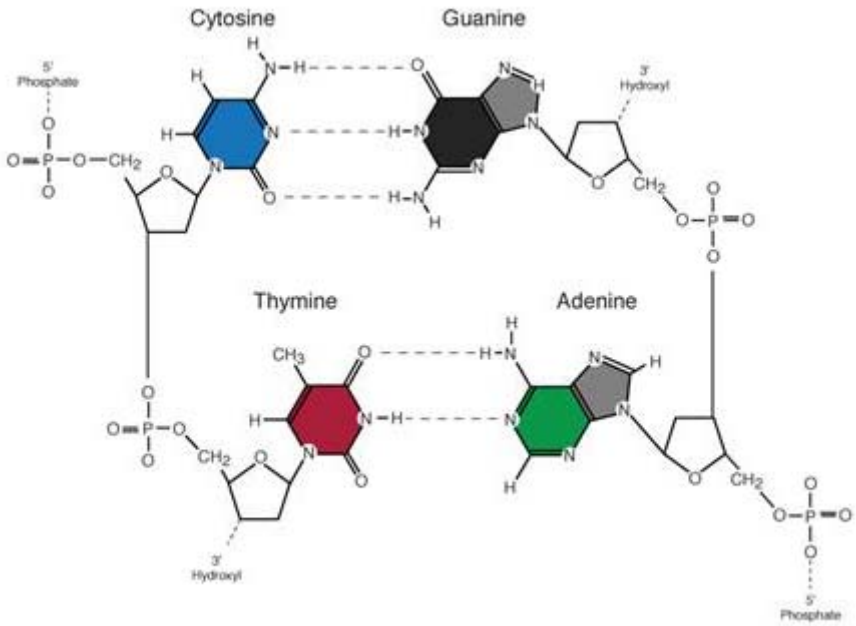
Cytosine
Thymine
Adenine
Guanine

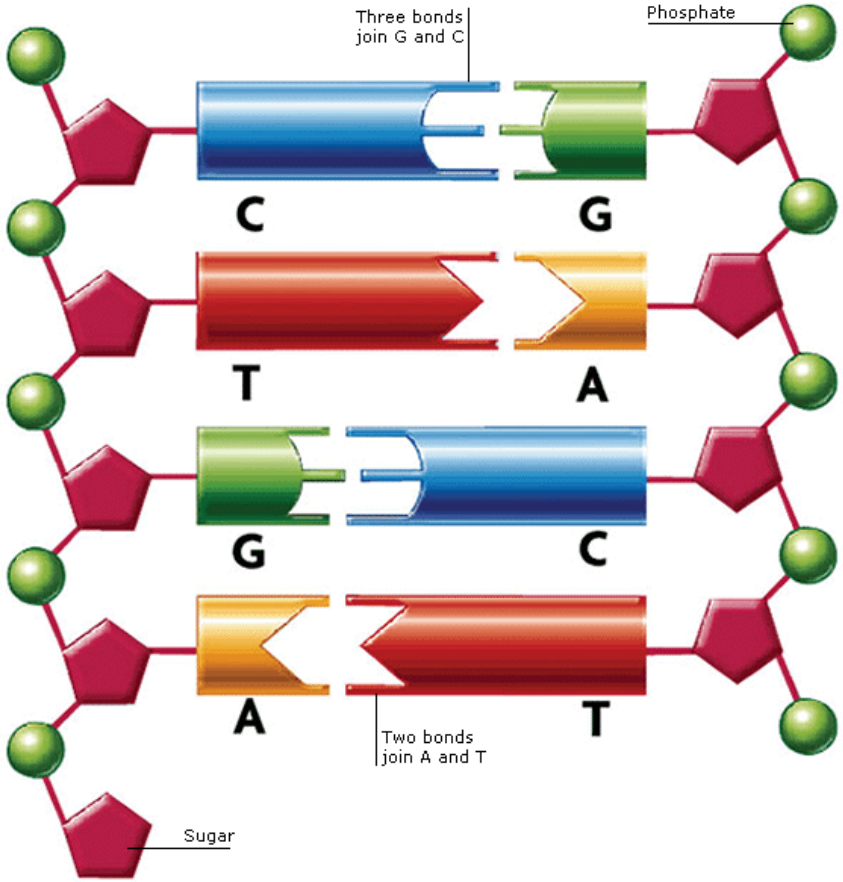


A single nucleotide is made up of three components: **a nitrogen-containing base, a five-carbon sugar**, and **a phosphate group**. The nitrogenous base is either a purine or a pyrimidine. The five-carbon sugar is either a ribose (in RNA) or a deoxyribose (in DNA) molecule.

Base Pairing Rule

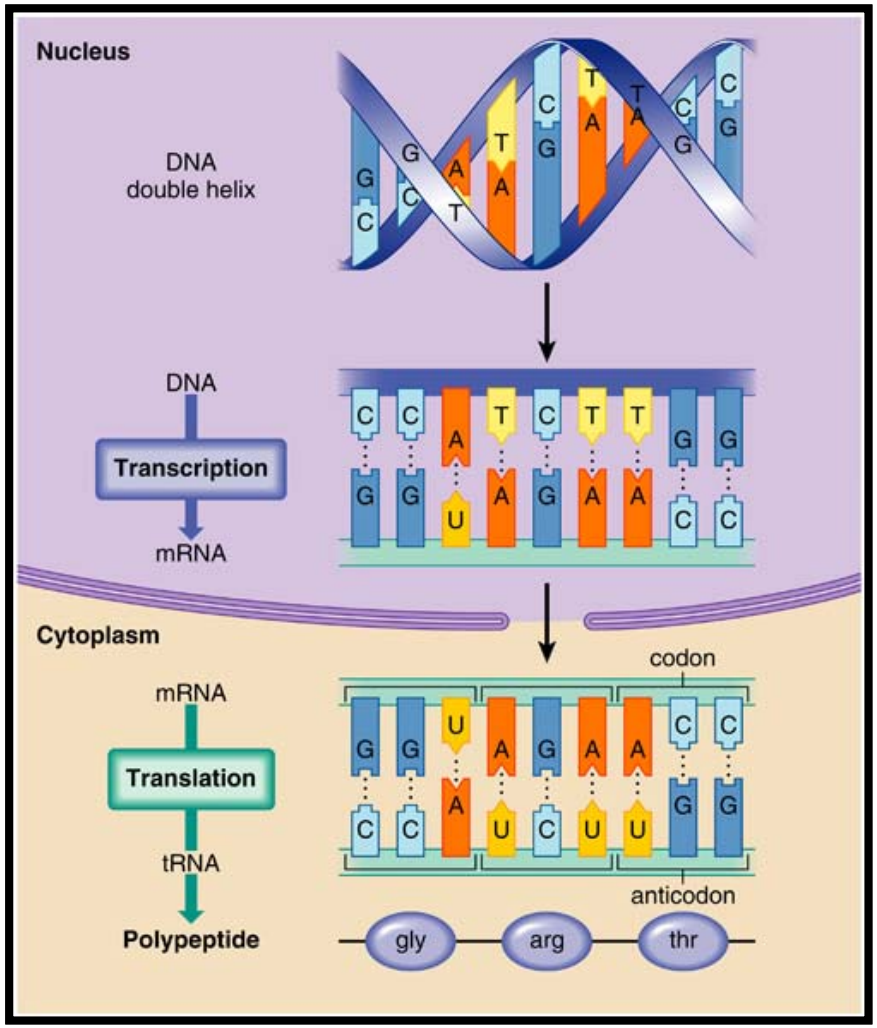
- Adenine (A) pairs with Thymine (T)
- Guanine (G) pairs with Cytosine(C)

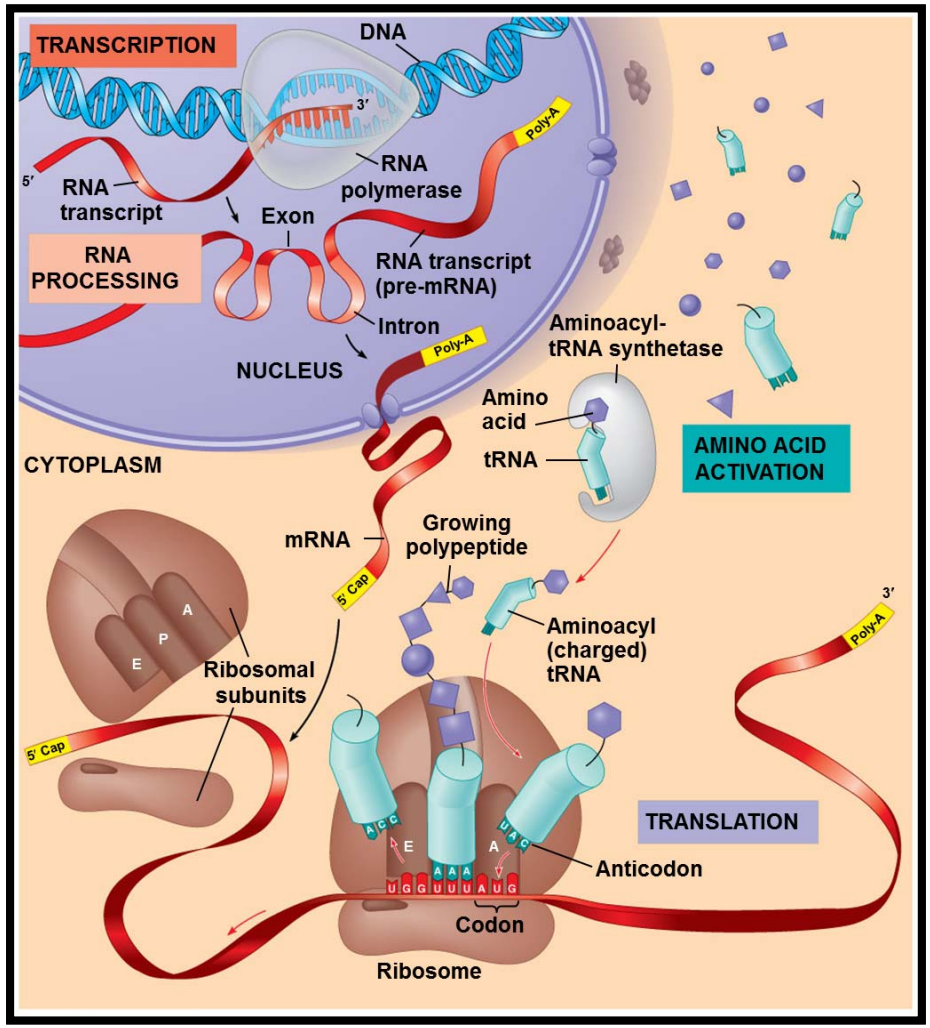




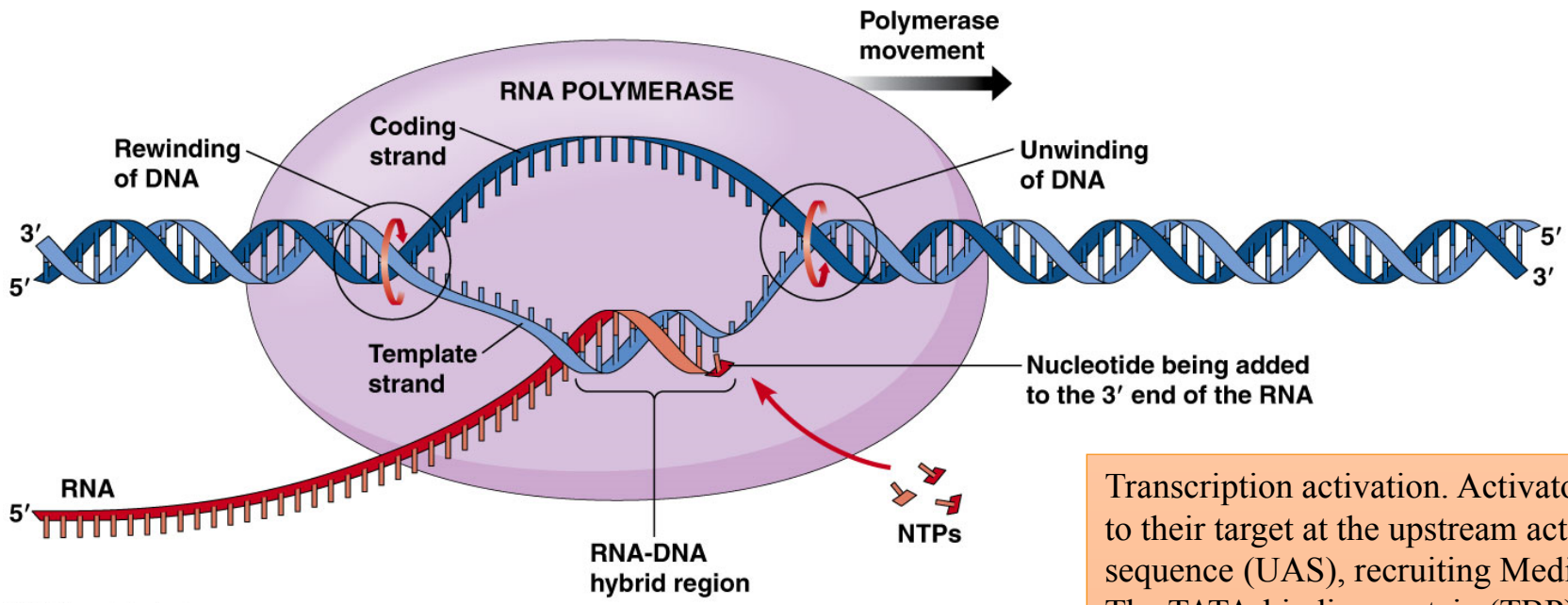
James Watson Explains DNA Base pairing

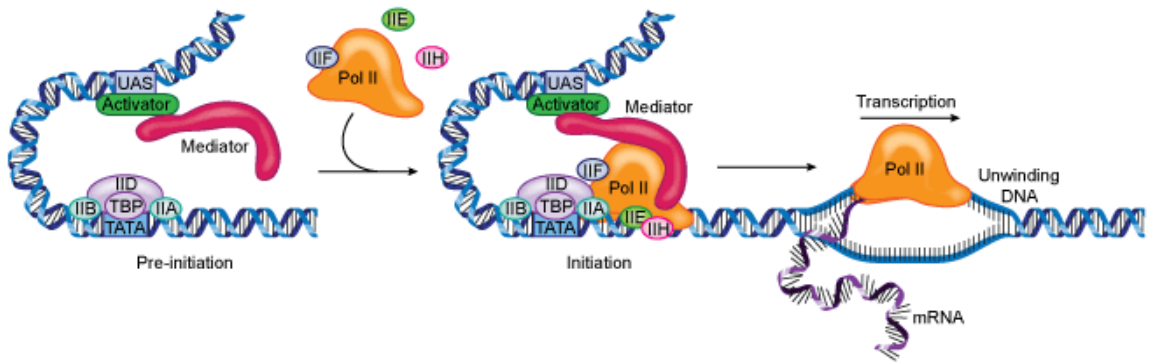
DNA → RNA → Protein





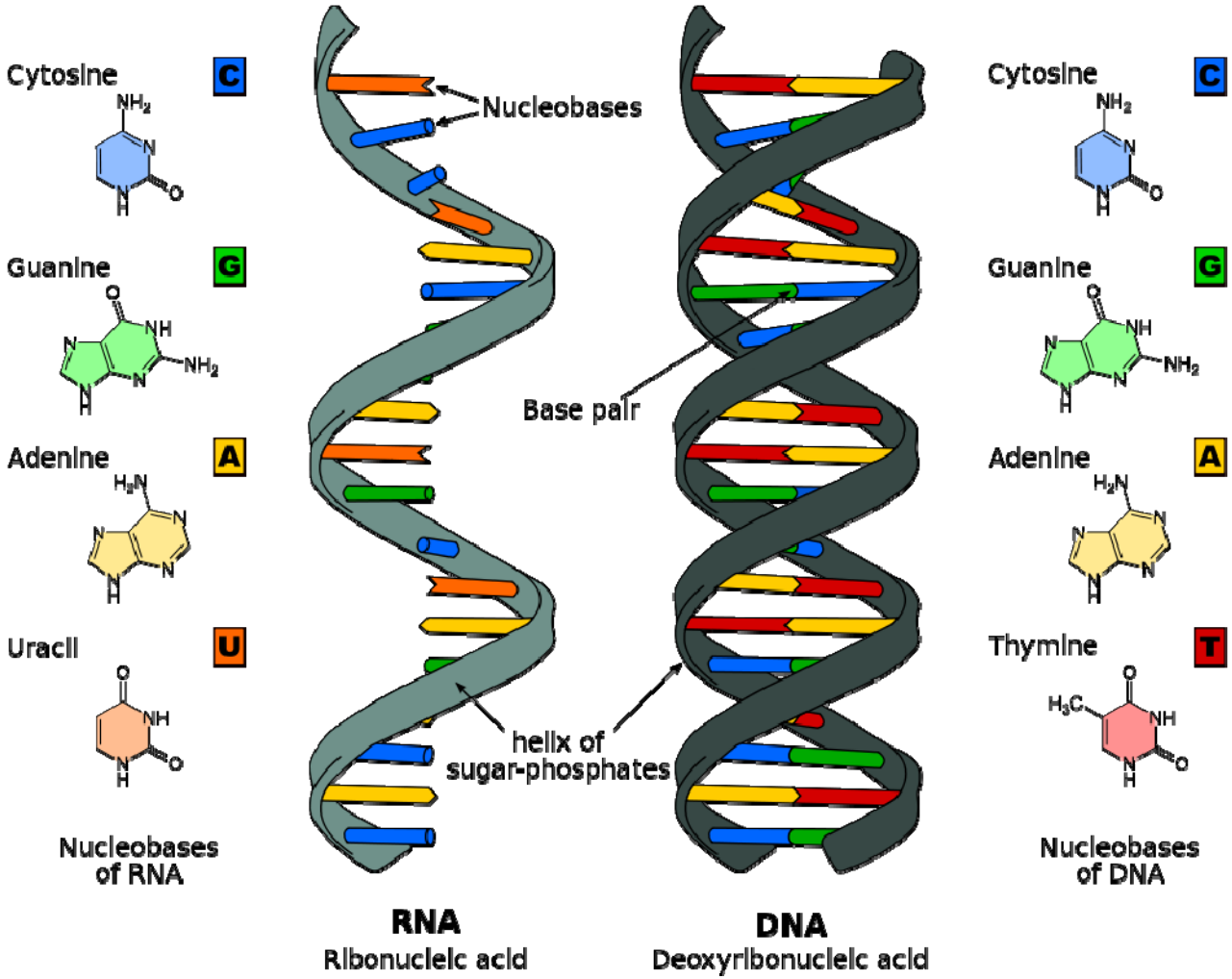
Transcription





Transcription activation. Activators bind to their target at the upstream activation sequence (UAS), recruiting Mediator. The TATA-binding protein (TBP) subunit of transcription factor IID binds to the promotor TATA box and recruits IIA and IIB. Mediator then recruits RNA polymerase II (Pol II) to the pre-initiation complex along with IIF, followed by IIE and IIH. Upon initiation, RNA polymerase II is released from the complex and begins transcription.

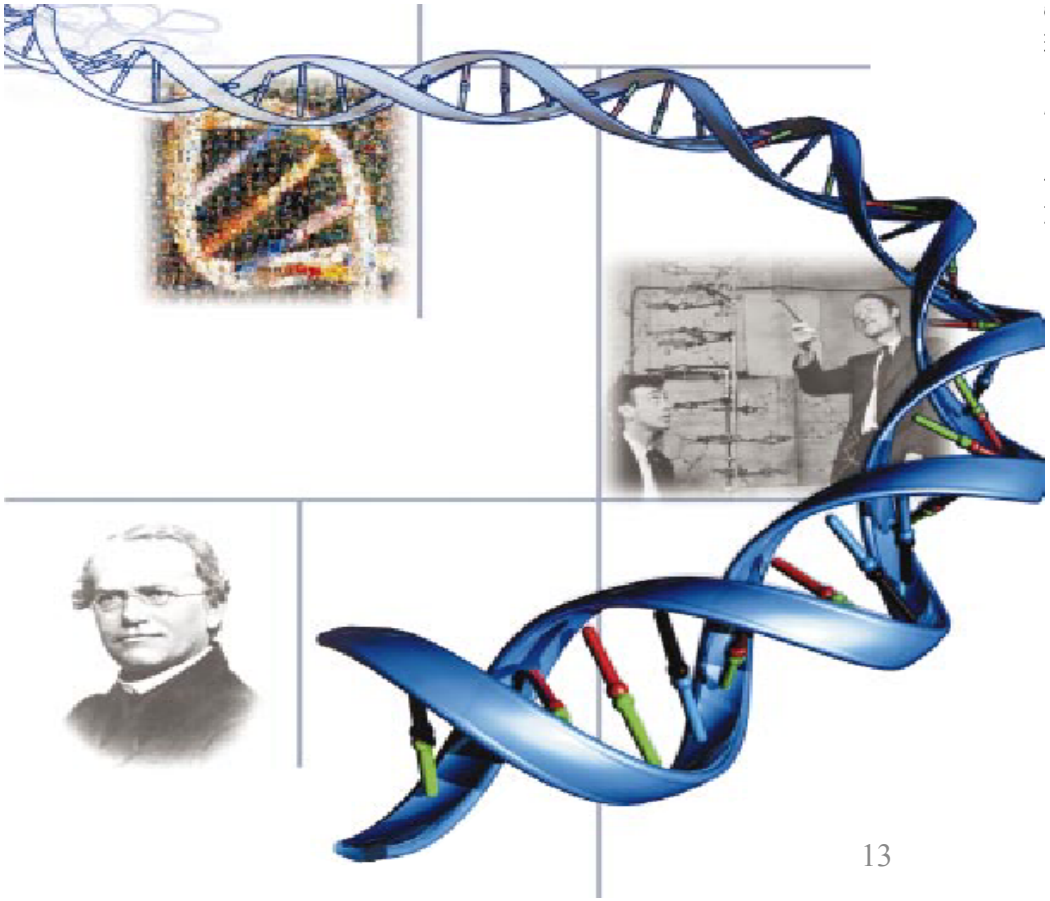
DNA verse RNA



THE DRAFT HUMAN GENOME SEQUENCE

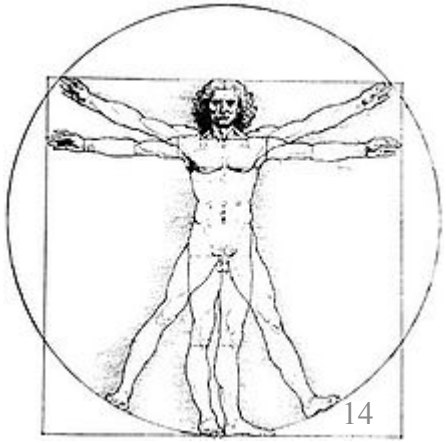
February 2001

« Finished » sequence
April 1953-April 2003

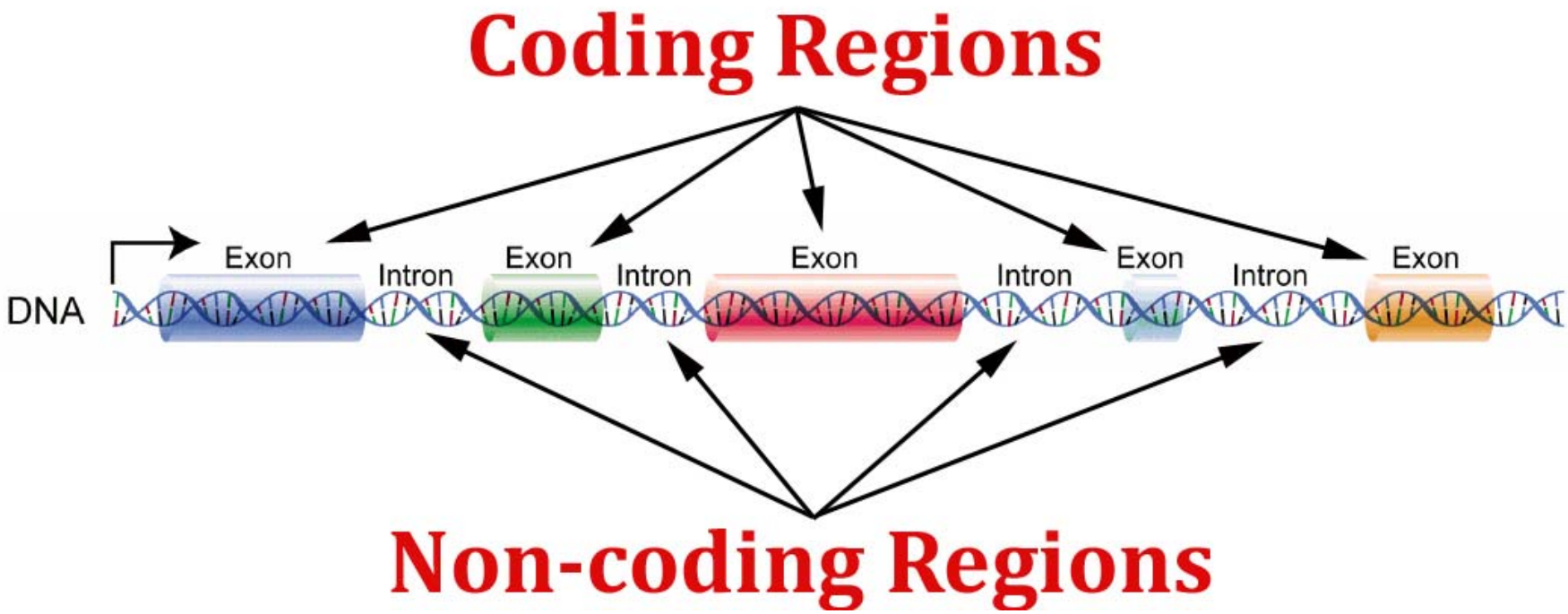


Human Genome Project (HGP)

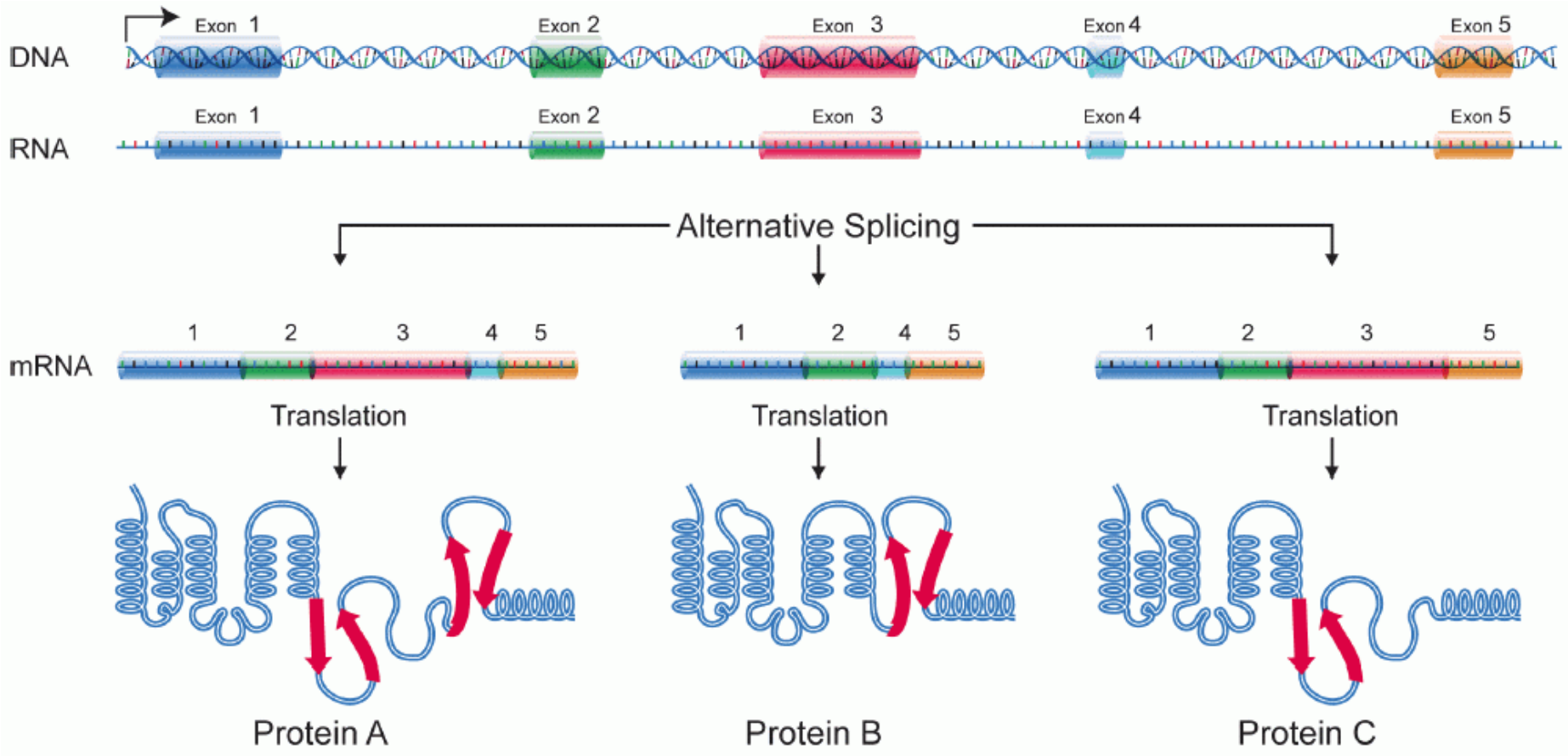
- The Human Genome Project was started in 1990 with the goal of sequencing and identifying all three billion chemical units in the human genetic instruction set, finding the genetic roots of disease and then developing treatments.
- It is considered a Mega Project because the human genome has approximately 3.3 billion base-pairs. With the sequence in hand, the next step was to identify the genetic variants that increase the risk for common diseases like cancer and diabetes.



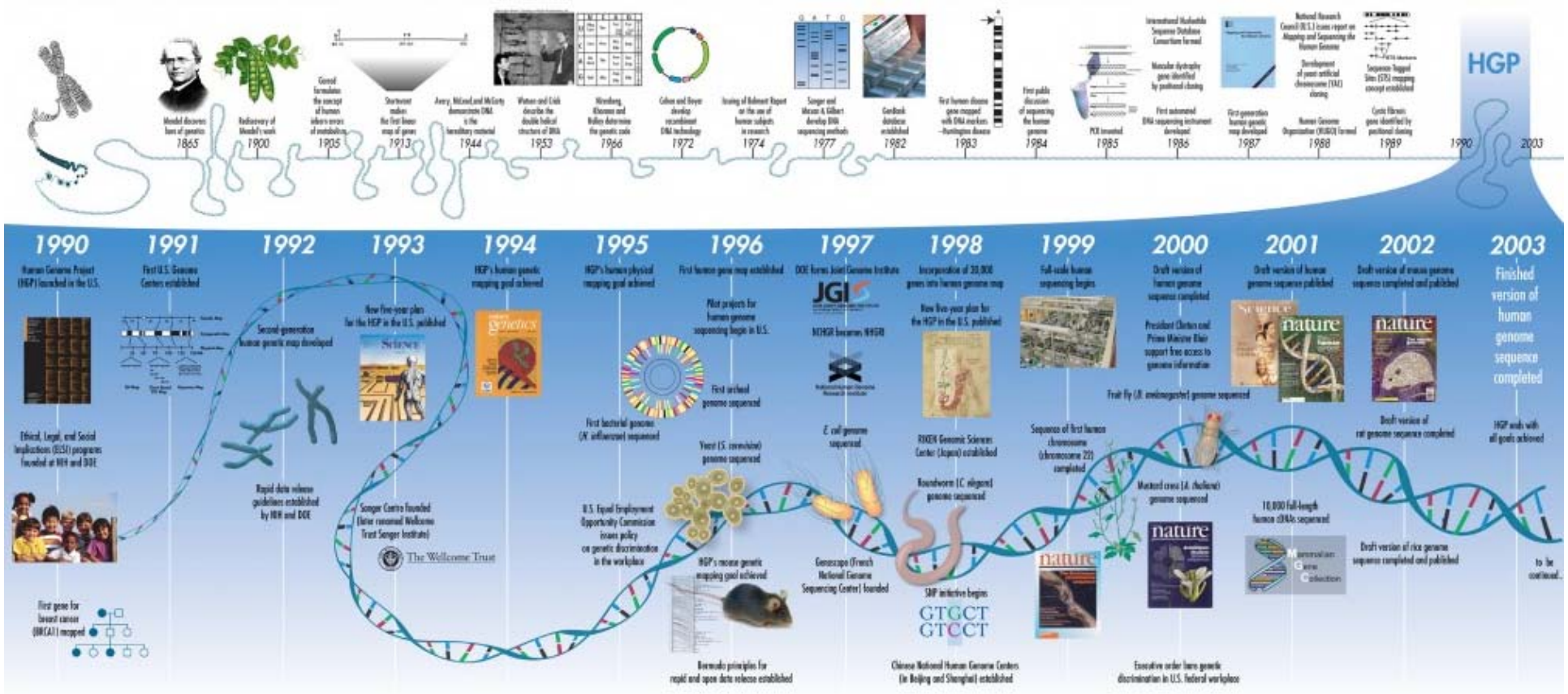
Coding and non-coding Regions

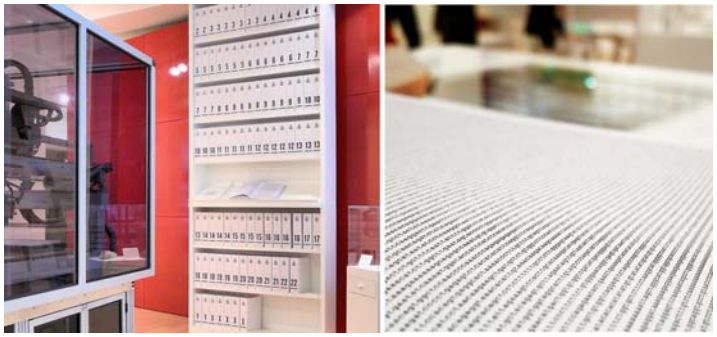


Alternative splicing

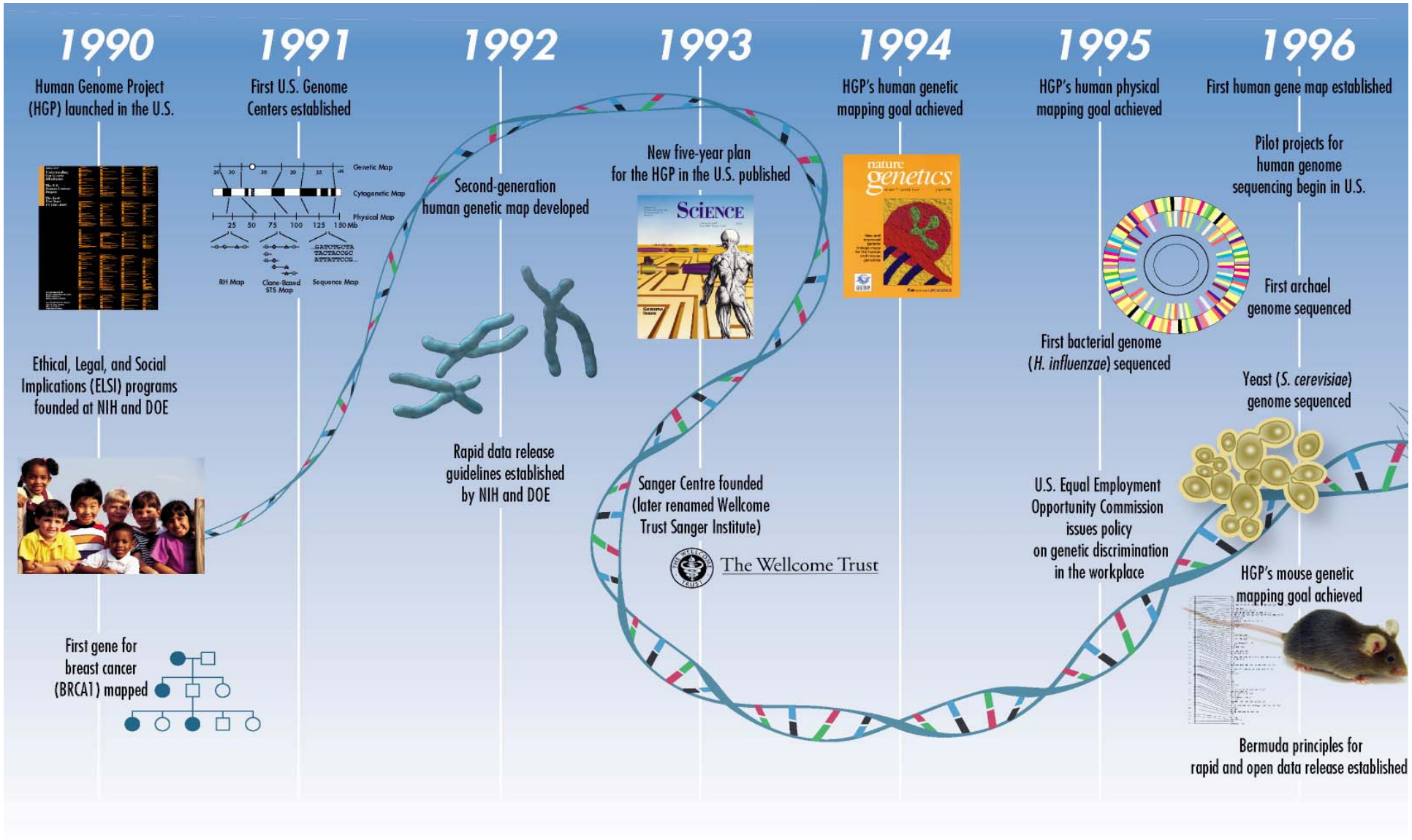


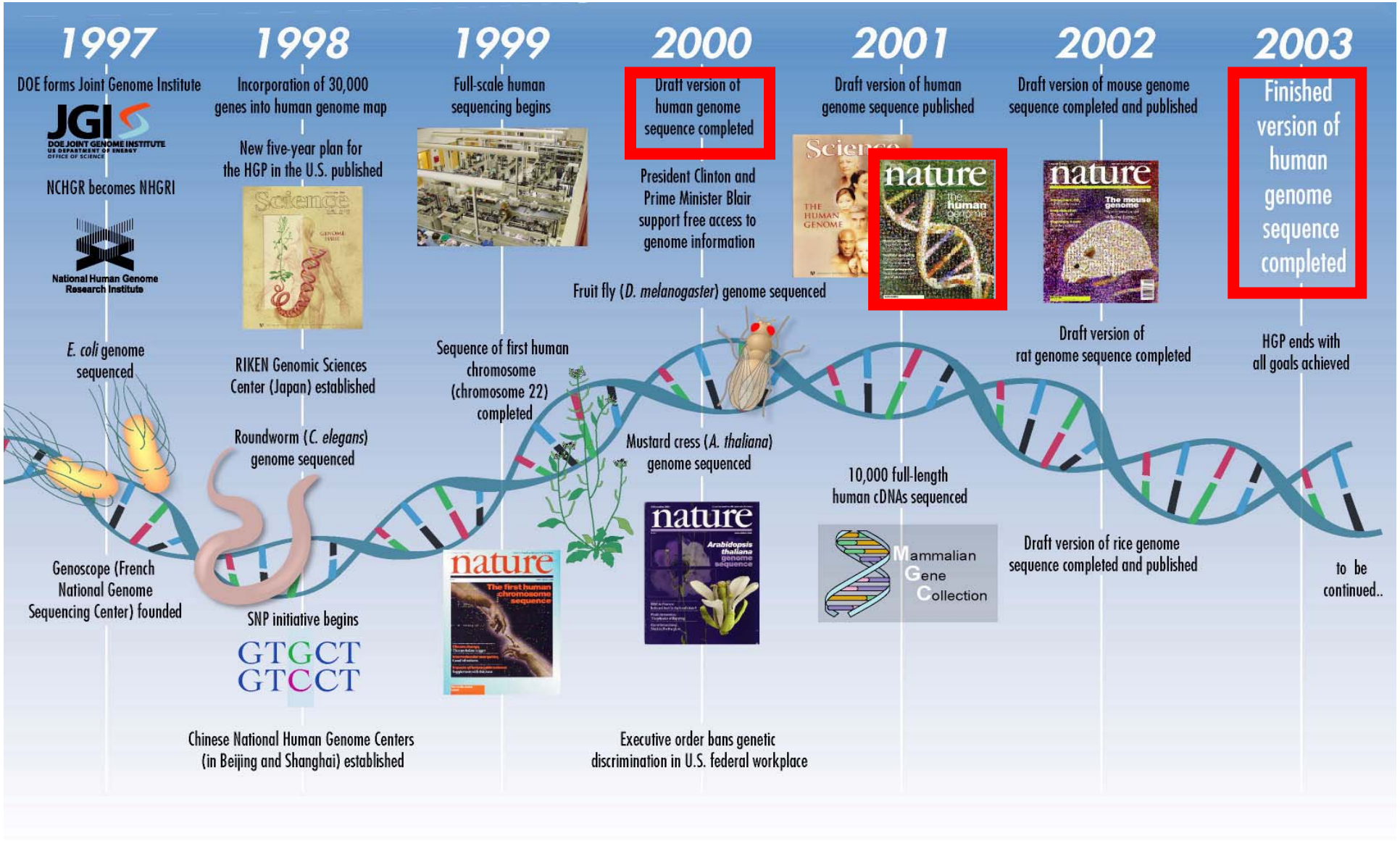
History of The Human Genome Project





The Wellcome Human Genome Library in London (Left CC: Russ London, Right Source: Wellcome Collection)

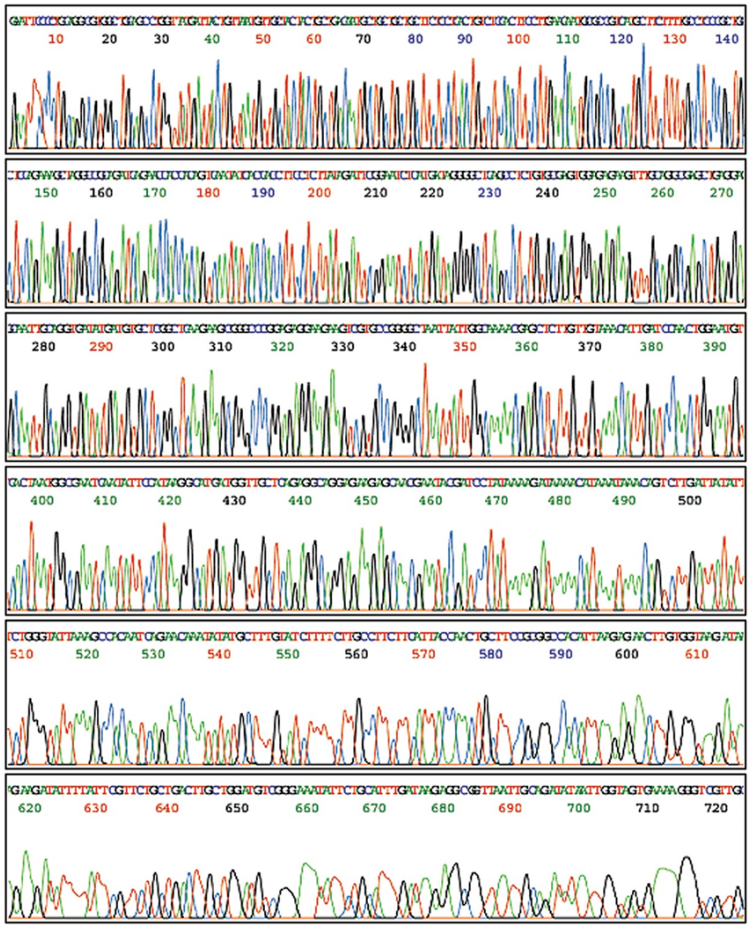

1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
Genetic Map
Cytogenetic Map
Physical Map
RH Map
Clone-Based STS Map
Sequence Map
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
SCIENCE
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established


1997
DOE forms Joint Genome Institute
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
Science
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
The first human chromosome sequence
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Arabidopsis thaliana genome sequence
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
THE HUMAN GENOME
nature
the human genome
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
The mouse genome
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued..

Findings

- Key findings of the draft (2001) and complete (2004) genome sequences include:
  1. There are approximately 20,500 genes in human beings, the same range as in mice.
  2. The human genome has significantly more segmental duplications (nearly identical, repeated sections of DNA) than had been previously suspected.
  3. At the time when the draft sequence was published fewer than 7% of protein families appeared to be vertebrate specific.

Automated Sequencing



http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v10n3/images/megabaces.jpg
http://www.ornl.gov/sci/techresources/Human_Genome/education/images.shtml

Single Nucleotide Polymorphism

- A Single Nucleotide Polymorphisms (SNP), pronounced "snip," is a genetic variation when a single nucleotide (i.e., A, T, C, or G) is altered and kept through heredity.
  - **SNP: Single DNA base variation found >1%**
  - **Mutation: Single DNA base variation found <1%**

94% → C T T A G **C** T T

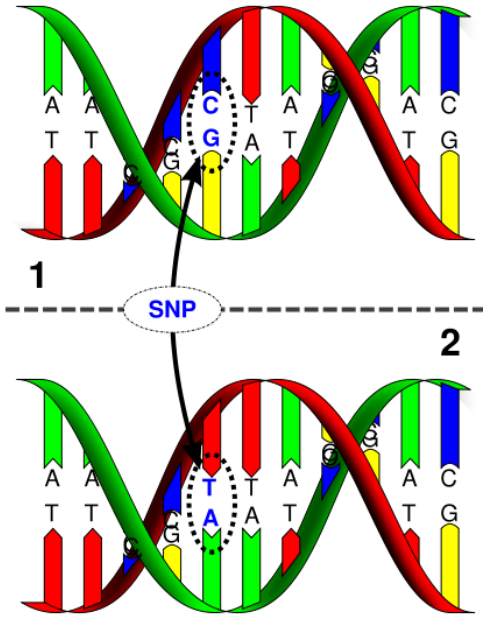
6% → C T T A G **T** T T

SNP

99.9% → C T T A G **C** T T

0.1% → C T T A G **T** T T

Mutation



International HapMap Project

http://www.hapmap.org/

International HapMap Project

Home I About the Project I Data I Publications I Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.

Project Information

- About the Project
- HapMap Publications
- HapMap Tutorial
- HapMap Mailing List
- HapMap Project Participants
- HapMap Mirror Site in Japan

Project Data

- HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)
- HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)
- HapMap Genome Browser (Phase 1 & 2 - full dataset)
- GWAs Karyogram
- HapMart
- HapMap FTP
- Bulk Data Download
- Data Freezes for Publication
- ENCODE Project
- Guidelines For Data Use

News

- 2009-12-14: **Notice to Haploview users**

  Recently, there are several questions about Haploview data format errors, and these errors were observed when users tried to analyze HapMap release 27 data dumped from HapMap. The current Haploview version (4.1) does not work with release 27 data. Haploview will generate a software error similar to "Hapmap data format error: NA06984" when trying to open the data.

  The r27 data format will be supported by next Haploview version. There is a beta test version that is supposed to work and it can be obtained from http://www.broadinstitute.org/haploview/haploview-downloads. But since it is NOT an official release version, please use it base on your own judgment.

- 2009-12-10: **Corrected HapMap3 phased haplotypes available for chromosome X**

  Phased haplotypes for consensus HapMap3 release 2 data for chromosome X has been corrected and the new data are now **available for bulk download**. Sorry for any inconvenience this might have caused.

- 2009-12-02: **HapMap3 phased haplotypes available for chromosome X**

  Phased haplotypes for consensus HapMap3 release 2 data has been phased for chromosome X and are now available for bulk download. [Update: The downloading was disabled because several users have found that there are repeating data in some of the chrX phasing data files. The data source is being contacted and the downloading will be enabled as soon as the problem is cleared.]

NPs in the dbSNP

International HapMap Project

Home I About the Project I Data I Publications

f Medicine (USA)
ome Québec Innovation Centre (Canada)
SA)

of Hong Kong (China)
Kong (China)
San Francisco (USA)
eria)
)
ellcome Trust Centre for Human Genetics (UK)
an)
)
St. Louis (USA)
Institute (UK)

T ation

G ases:

Nature 437: 1299-320, 2005

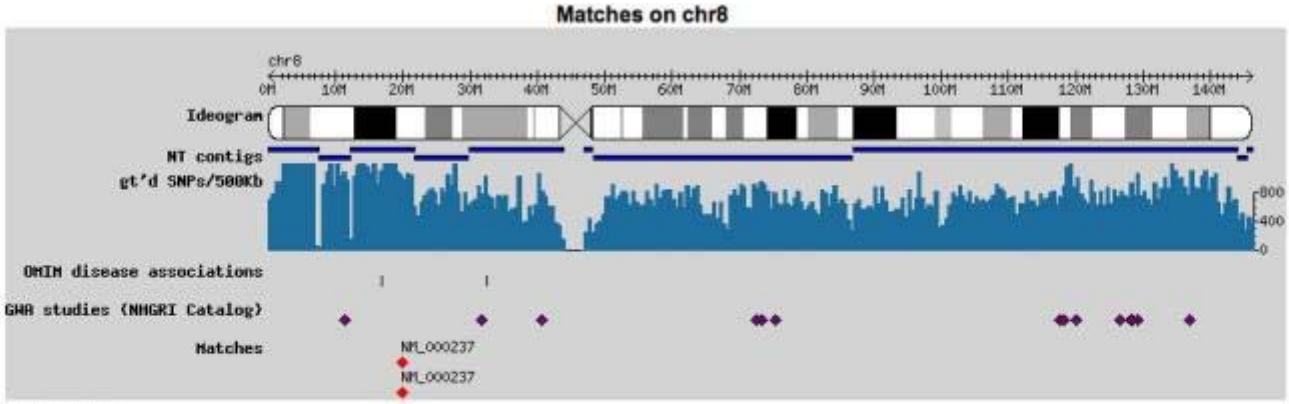
www.hapmap.org

International HapMap Project

http://www.hapmap.org/

International HapMap Project

http://www.hapmap.org/



NM_000237 **lipoprotein lipase** precursor — chr8:19.84..19.87 Mbp (27.99 kbp) score=27.49

NM_000237 LPL encodes **lipoprotein lipase**, which is expressed in heart, muscle, and adipose tissue. LPL functions as a homodimer, and has the dual functions of triglyceride hydrolase and ligand/bridging factor for receptor-mediated **lipoprotein** uptake. Severe mutations that cause LPL deficiency result in type I hyper**lipoprotein**emia, while less extreme mutations in LPL are linked to many disorders of **lipoprotein** metabolism. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Entrez Gene record to access additional publications. — chr8:19.84..19.87 Mbp (27.99 kbp) score=26.49

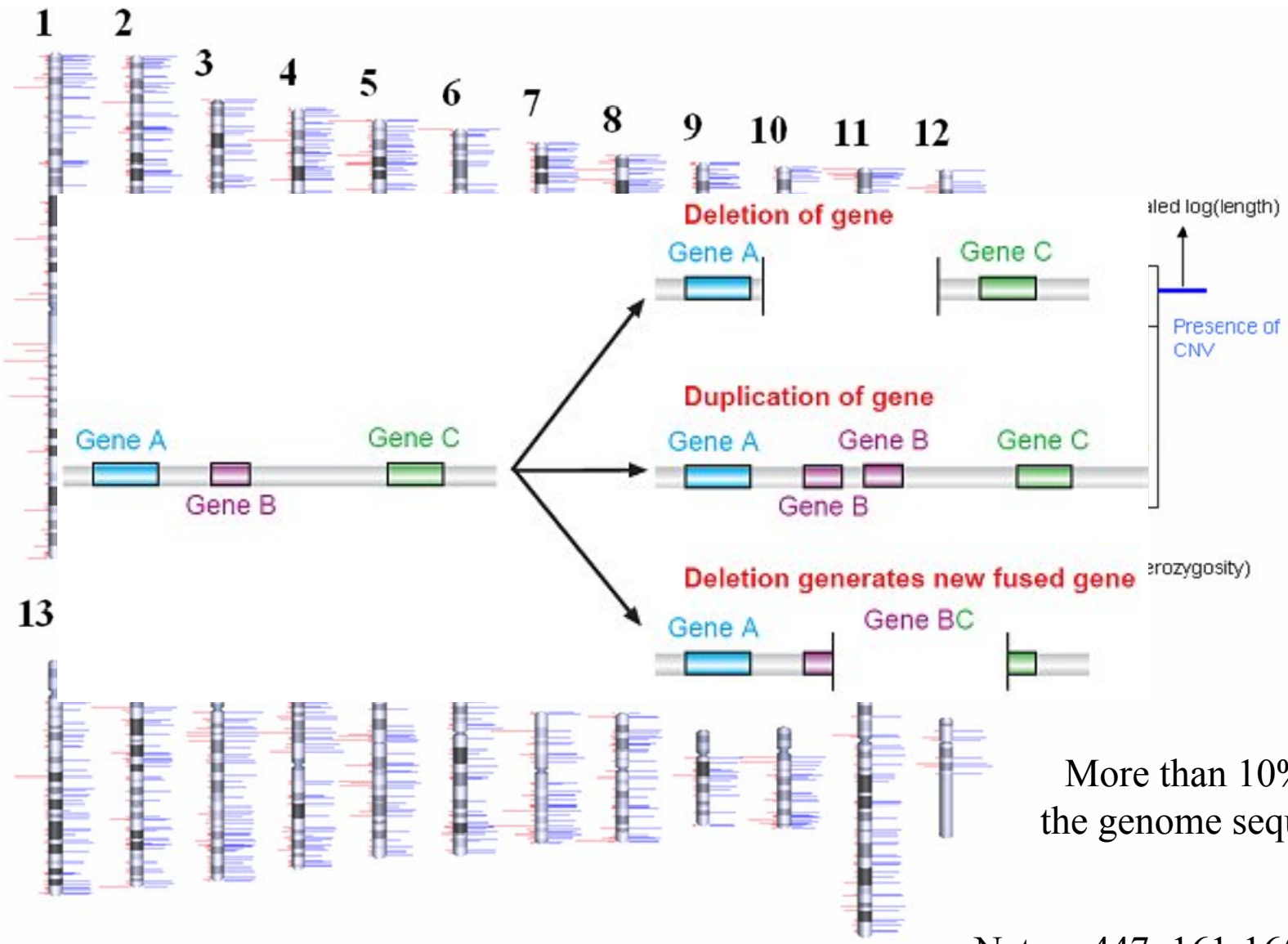
Associated SNPs can be diagnostic/predictive but finding functional SNPs to understand mechanism will take time but offers the promise of new therapies

ENCODE PROJECT - Identify the functional elements in the Human Genome - 1% now and soon all

Nature 447: 799, 2007

Transcriptional Regulatory Elements
Expressed Sequences
Chromatin Structure
Replication
Multi-species Conservation
…….

Structural Variation Project



More than 10% of the genome sequence

Nature 447: 161-165, 2007

All of the original goals of the Human Genome Project have been accomplished

What's next?

The Next Challenge

Understanding the link between -

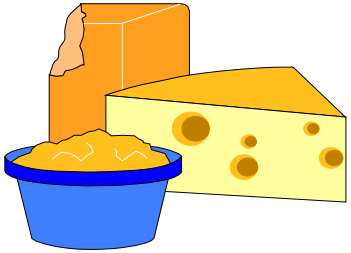
DNA sequence (Genotype) ⟷ **Biology/Disease (Phenotype)**

ATTCGCATGGACC

——C——

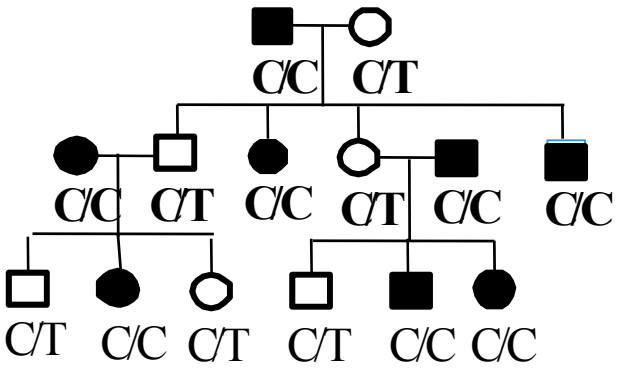
——A——

Environment



Human Genetic Analysis

Families
Linkage Studies



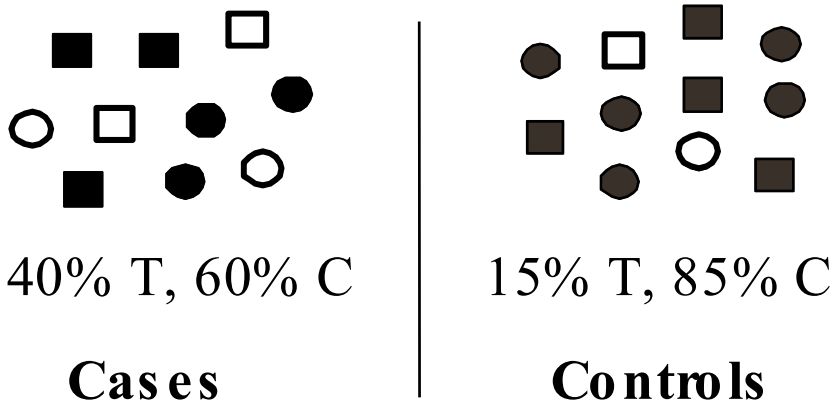
Simple Inheritance (Segregate)

Single Gene with Major Effect

Variant Rare in the Population

~600 Short Tandem Repeat Markers

Populations
Association Studies



Complex Inheritance (Aggregate)

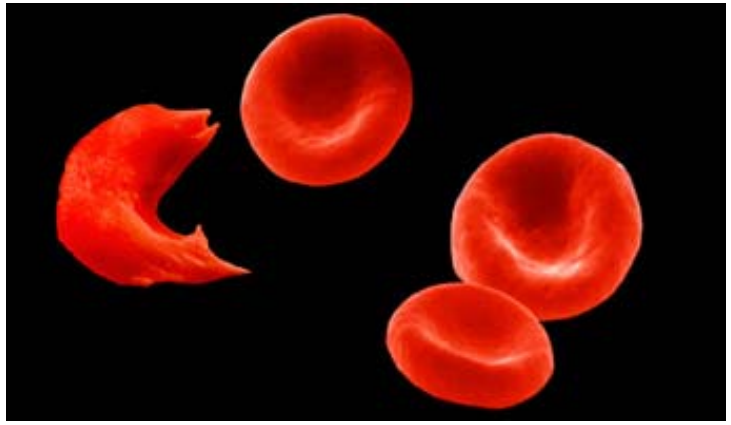
Multiple Genes with Small Contributions and Environmental Contexts

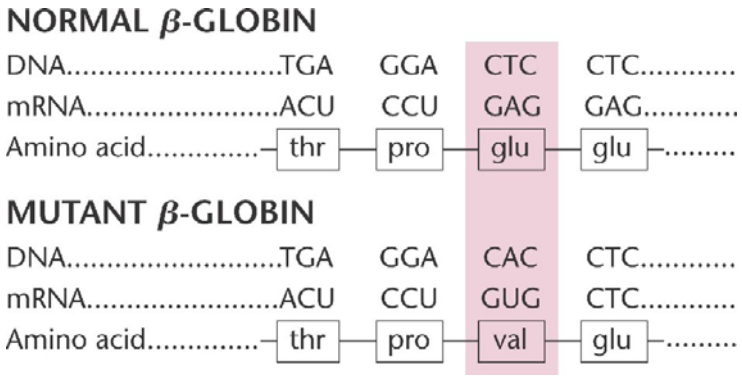
Variant(s) Common in the Population

Polymorphic Markers > 500,000 -1,000,000 Single Nucleotide Polymorphisms (SNPs)

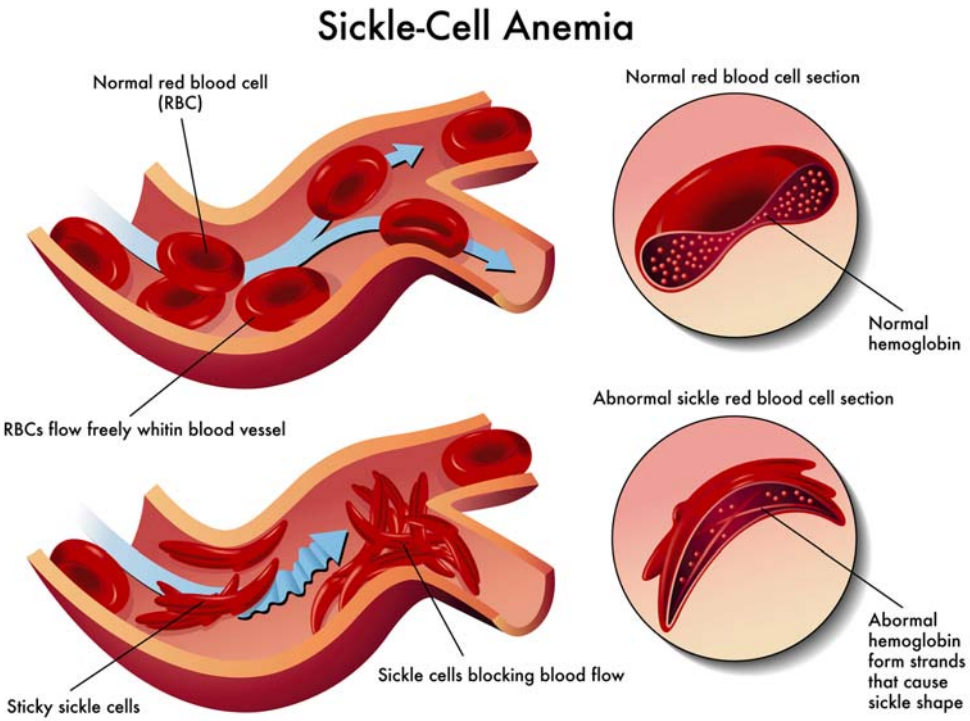
Sickle-cell disease

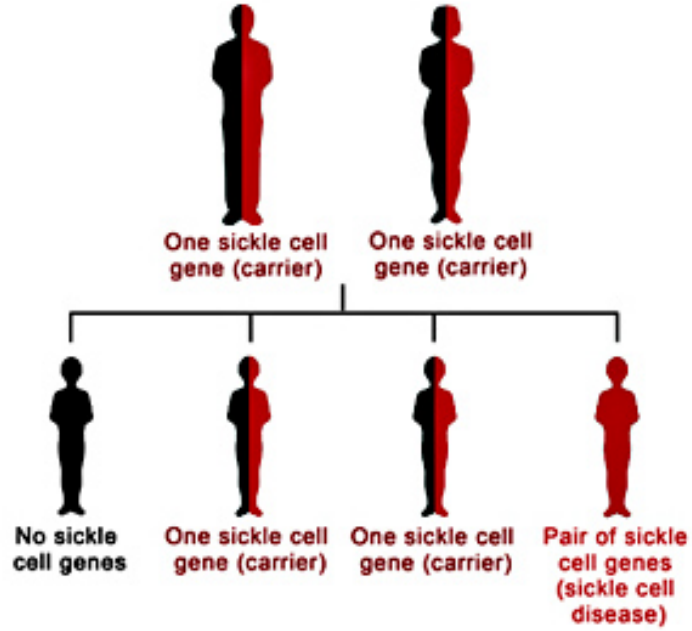
- Sickle-cell disease (SCD), also known as sickle-cell anaemia (SCA), is a hereditary blood disorder, characterized by an abnormality in the oxygen-carrying haemoglobin molecule in red blood cells.
- This leads to a propensity for the cells to assume an abnormal, rigid, sickle-like shape under certain circumstances.
- Sickle-cell disease is associated with a number of acute and chronic health problems, such as severe infections, attacks of severe pain ("sickle-cell crisis"), and stroke, and there is an increased risk of death.

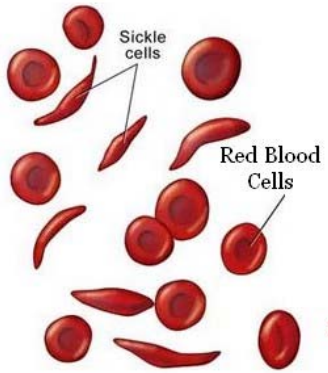




Sickle-cell disease



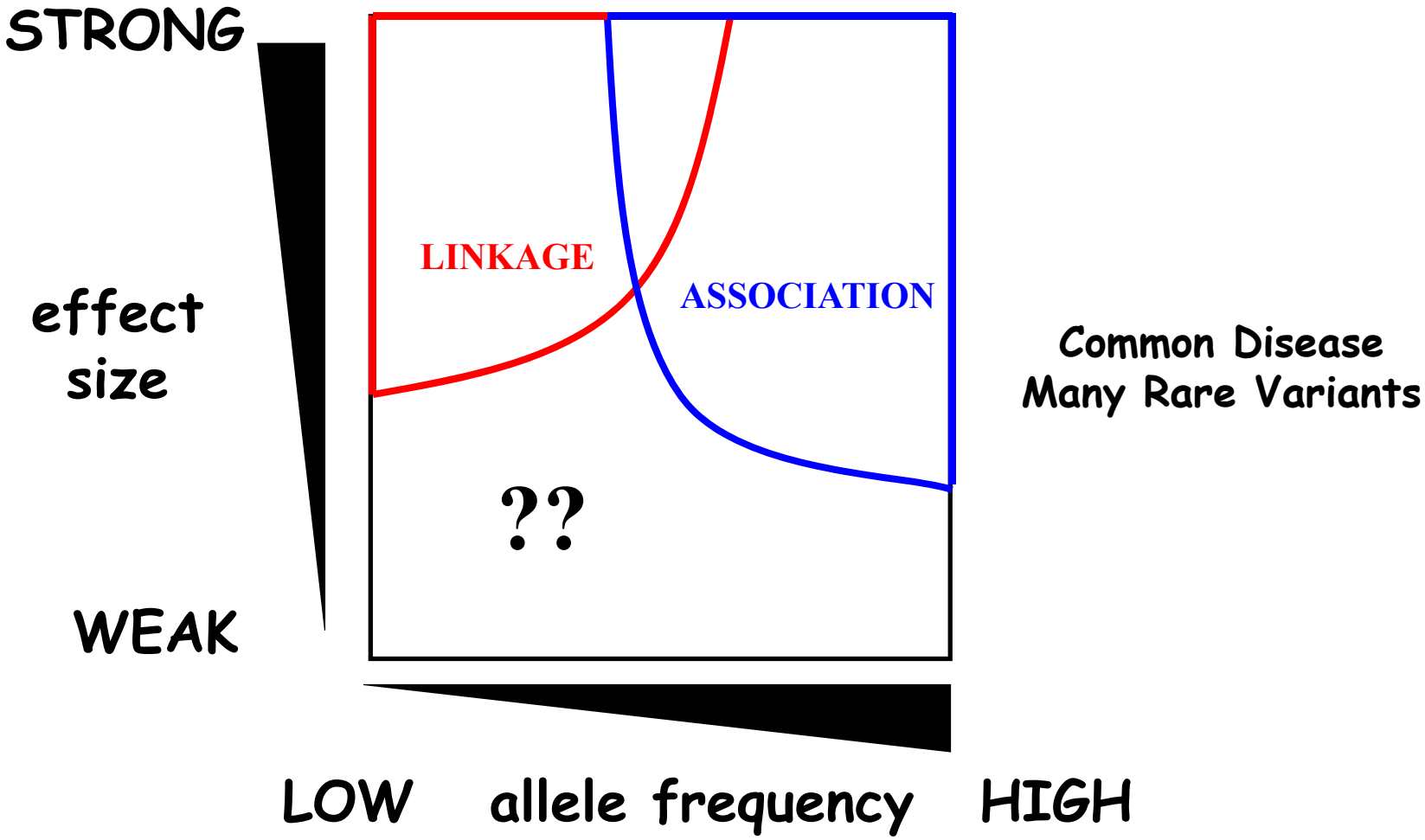




	A	a
A	AA	Aa Carrier
a	Aa Carrier	aa Child has Sickle Cell Anemia

a — Sickle cell allele

Genetic Strategy - New Insights

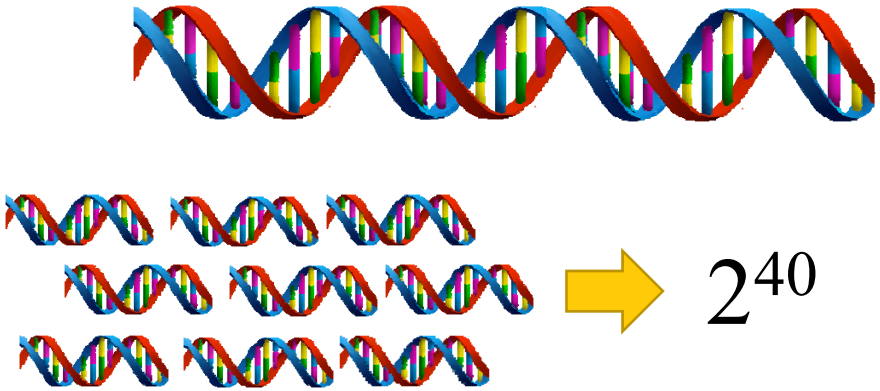


Common Disease
Many Rare Variants

Ardlie, Kruglyak & Seielstad (2002) Nat. Genet. Rev. 3: 299-309
Zondervan & Cardon (2004) Nat. Genet. Rev. 5: 89-100

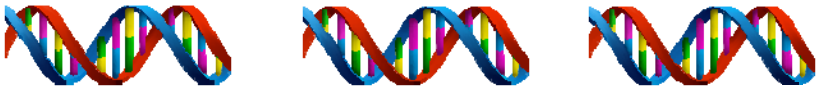
PCR-based restriction fragment length polymorphism (PCR-RFLP)

1) PCR



2^{40}

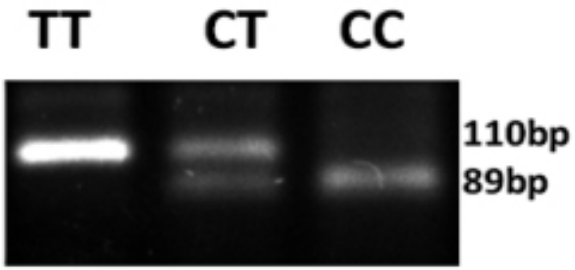
1) RESTRICTION ENZYME DIGEST



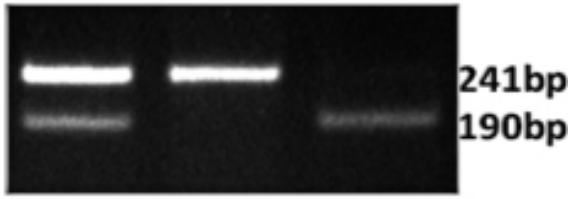
TT 110bp

T 110bp
C 89bp

C 89bp

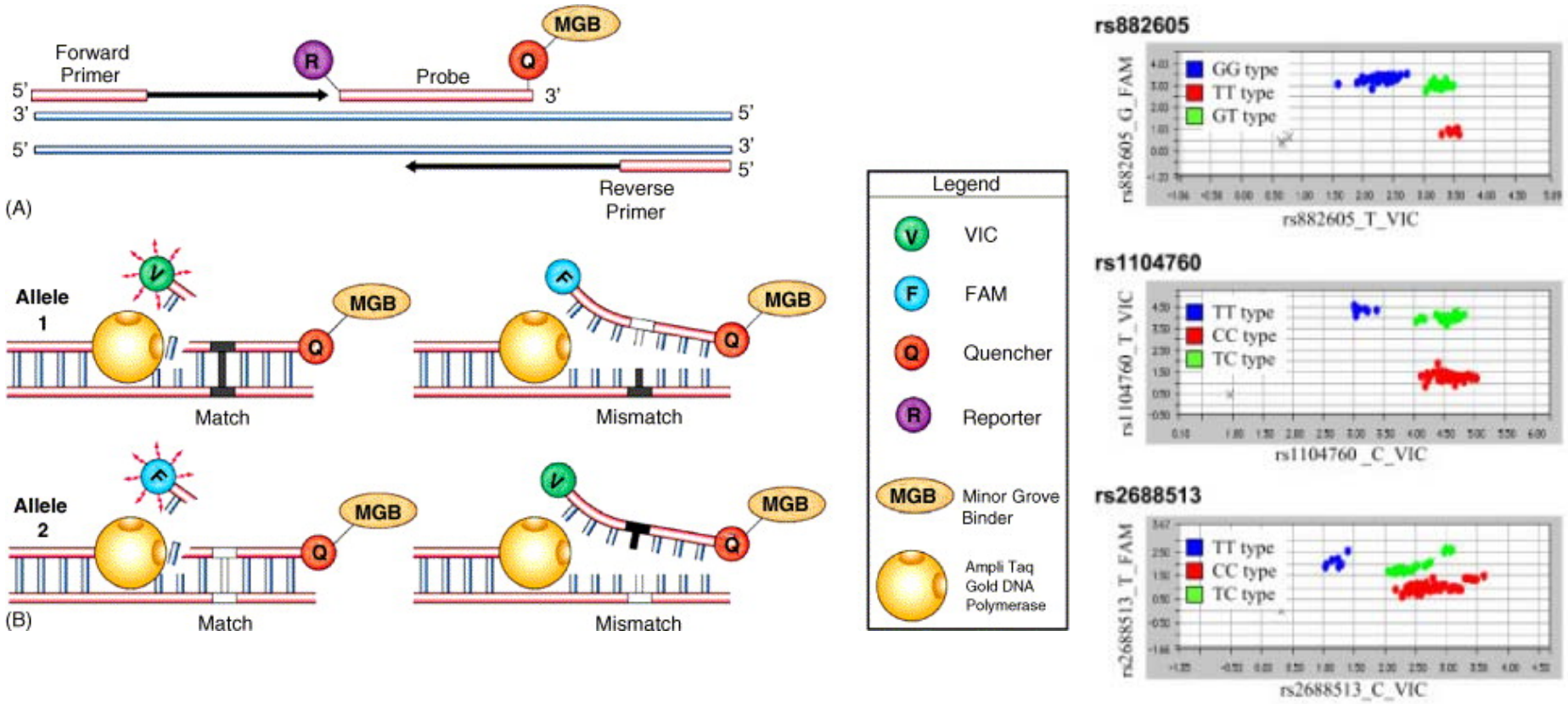


TLR4 (rs4986791)

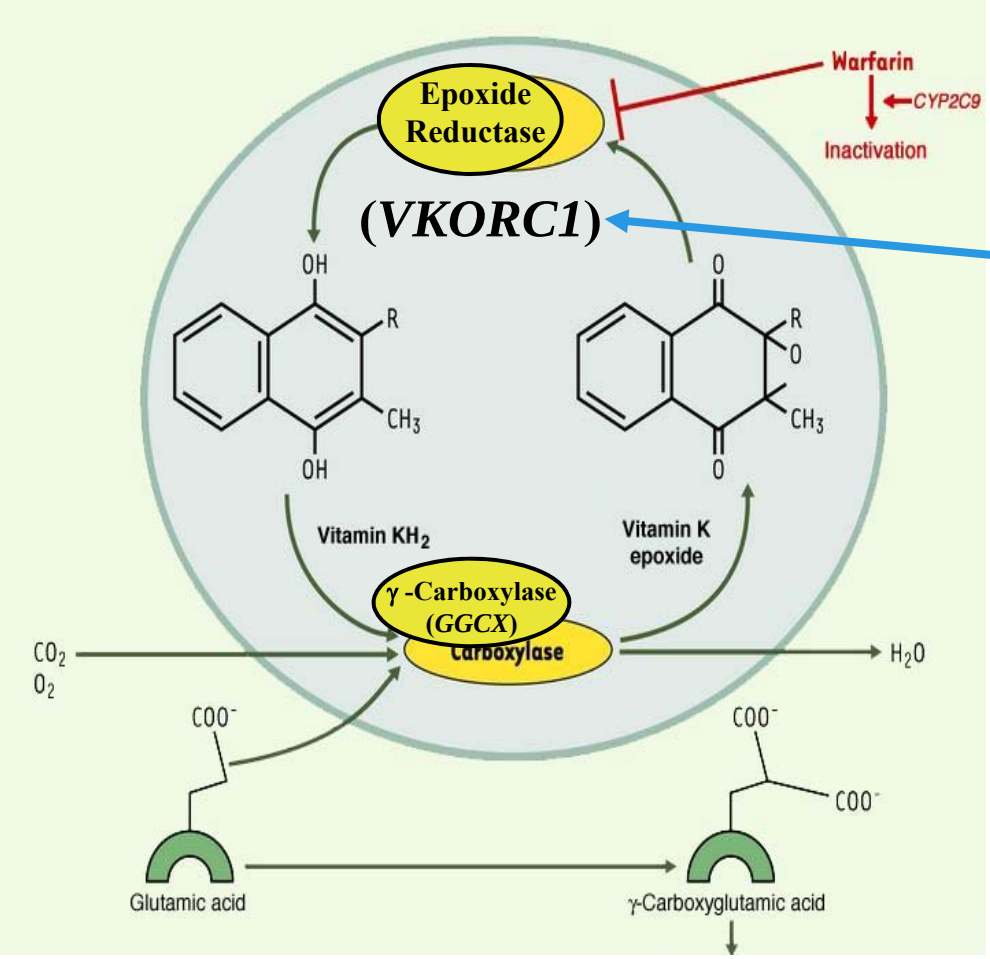


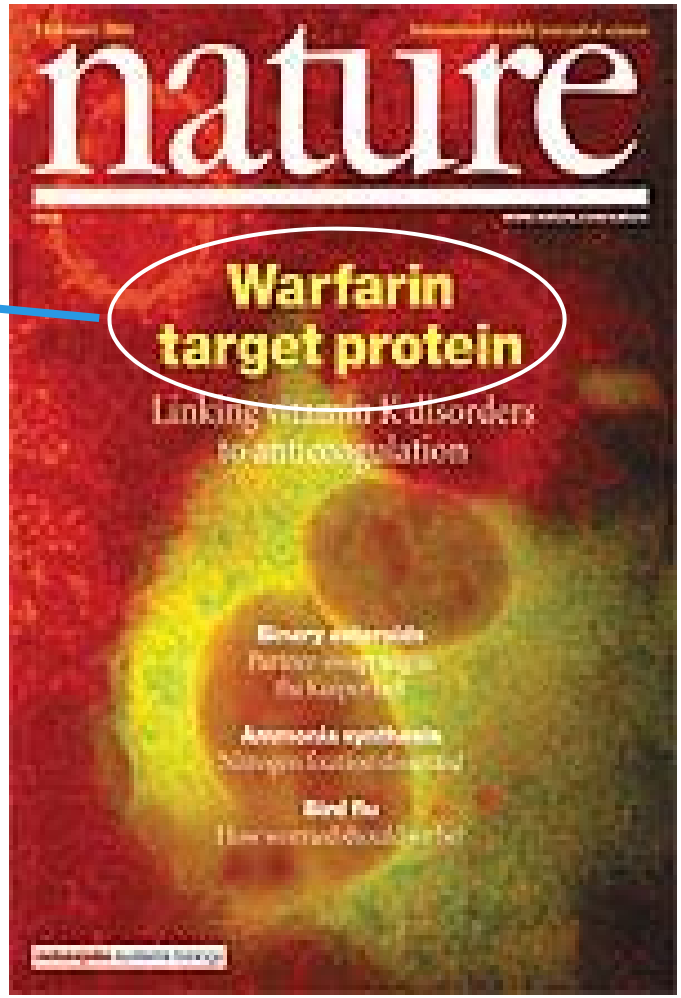
TLR4 (rs1927911)

Principle of TaqMan SNP assay



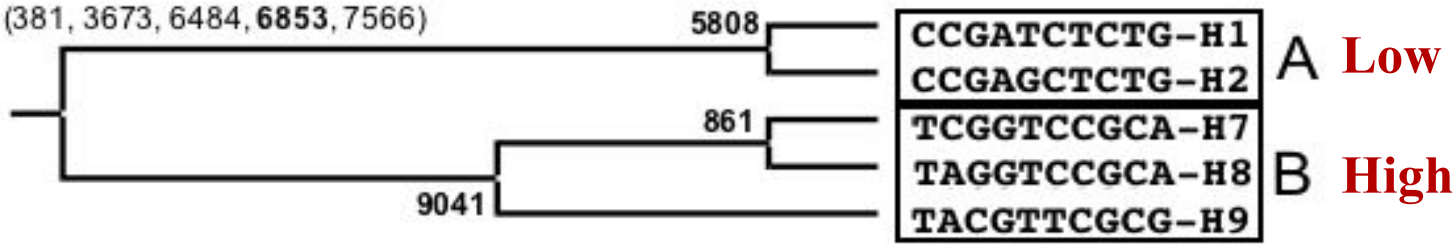
New Target Protein for Warfarin

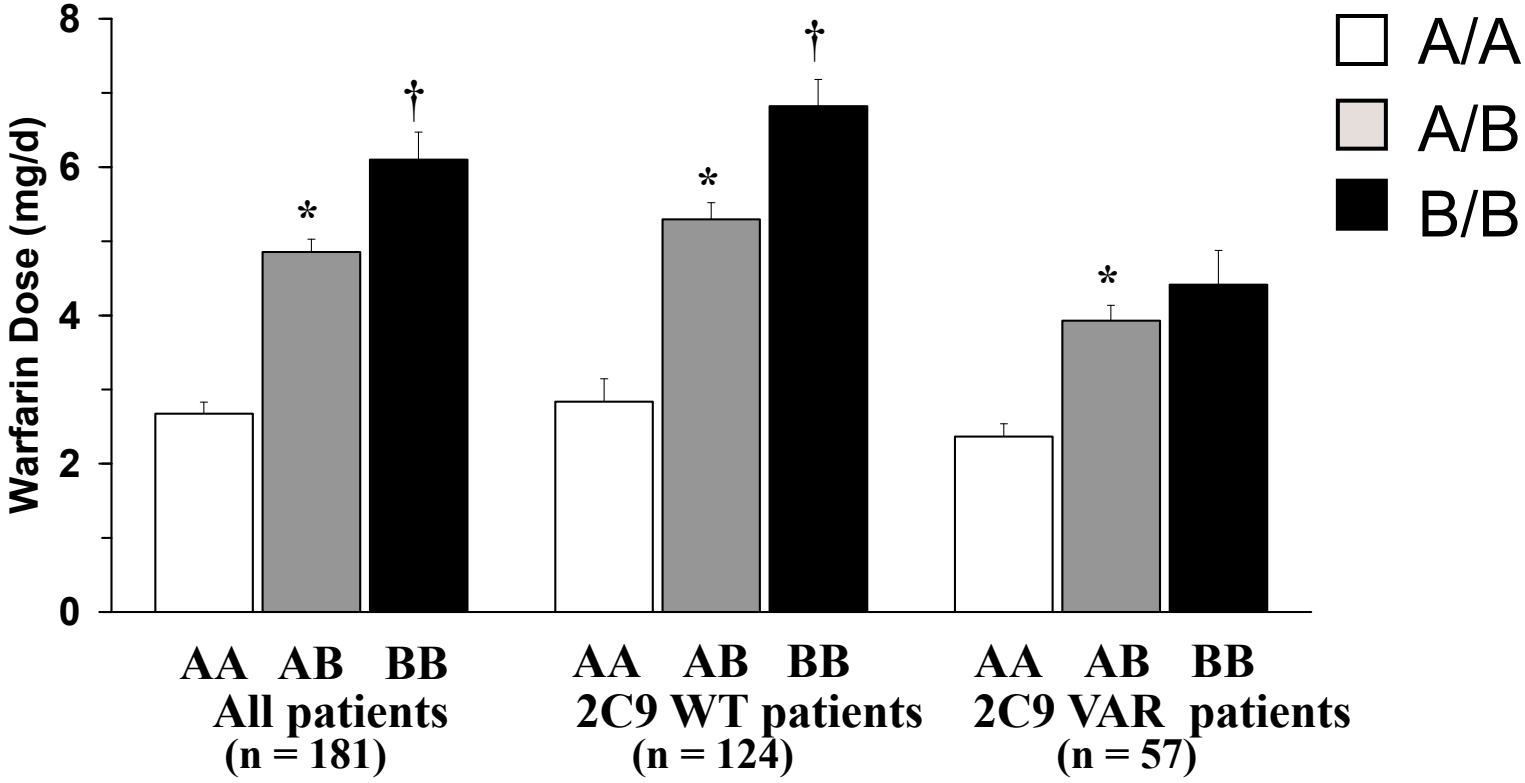




Rost et al. & Li, et al., *Nature* (2004)

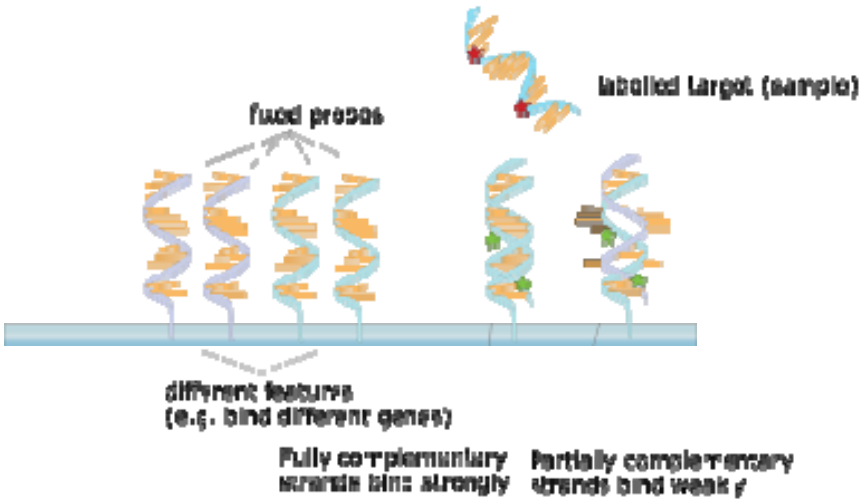
VKORC1 SNPs and haplotypes show a strong association with warfarin dose

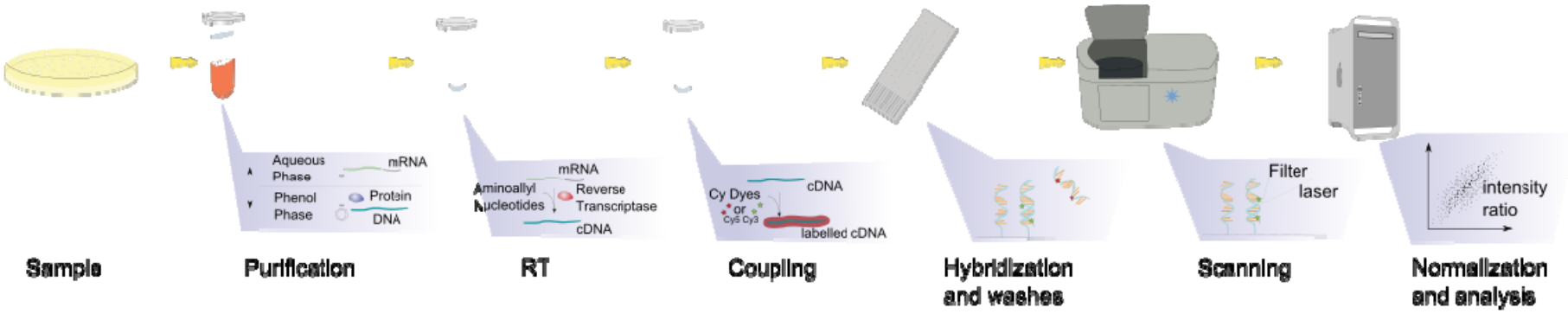




Rieder et al N Engl J Med 352: 2285-93, 2005

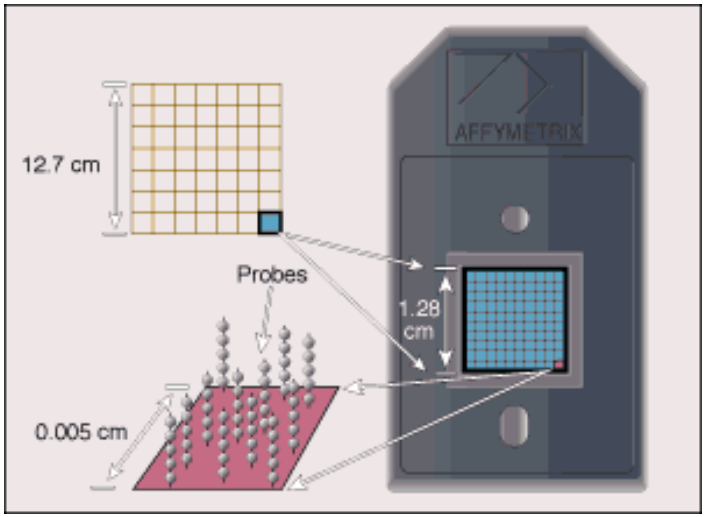
Principle of Microarray





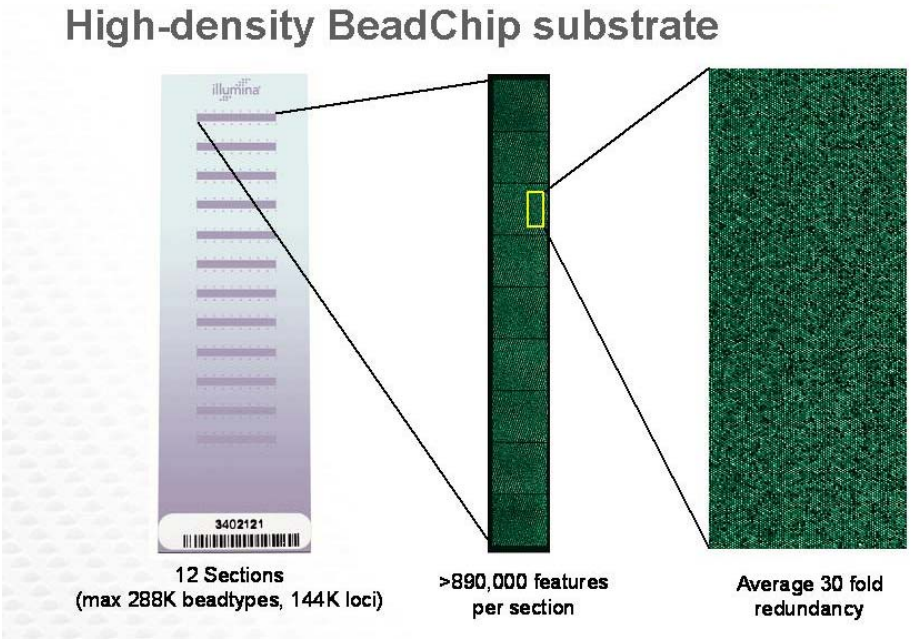
Genotyping Systems

Affymetrix



100K or 500K Quasi-Random SNPs

Illumina



100K; 317K; 550K; 650K SNPs

A significant proportion of common SNPs can be captured

Genome-wide association study, GWAS

Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia

nature genetics

Qing Lan[1,68], Chao A Hsiung[2,68], Keitaro Matsuo[3,68], Yun-Chul H
H Dean Hosgood III[1,7,68], Kexin Chen[8,68], Jiu-Cun Wang[9,10,68], Nil
Wei Zheng[12], Neil Caporaso[1], Jae Yong Park[13], Chien-Jen Chen[14],
Maria Teresa Landi[1], Hongbing Shen[17,18], Charles Lawrence[19], La
Jeffrey Yuenger[6], Kevin B Jacobs[6], I-Shou Chang[20], Tetsuya Mitsu
Bryan A Bassig[1,25], Margaret Tucker[1], Fusheng Wei[26], Zhihua Yin
Victor Ho Fun Lee[31], Daru Lu[9,10], Jianjun Liu[32,33], Hyo-Sung Jeon
Jin Hee Kim[35], Yu-Tang Gao[36], Ying-Huang Tsai[37], Yoo Jin Jung[16
Amy Hutchinson[6], Wen-Chang Wang[2], Robert Klein[39], Charles C
Sonja I Berndt[1], Xingzhou He[43], Wei Wu[27], Jiang Chang[28,29], Xu-
Hong Zheng[8], Junwen Wang[45,46], Yueying Zhao[9,10], Yuqing Li[32], Jin

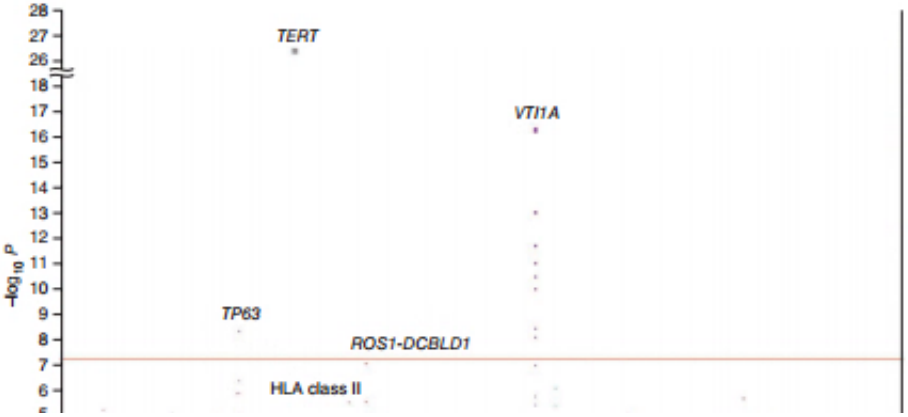


Table 2 New loci associated with adenocarcinoma and squamous carcinoma of the lung in a GWAS of never-smoking Asian females

				MAF[b]			Adenocarcinoma				Squamous carcinoma				
							Subjects				Subjects				
SNP	Putative gene	Chromosome position	Allele[a]	1	2	3	Control	Case	OR (95% CI)	P_{trend}	Control	Case	OR (95% CI)	P_{trend}	$P_{heterogeneity}$[c]
rs7086803	*VTI1A*	10q25.2	G/A	0.27	0.31	0.34	7,035	4,666	1.24 (1.17–1.32)	1.19×10^{-11}	6,714	756	1.36 (1.21–1.54)	7.11×10^{-7}	0.014
rs9387478	*ROS1, DCBLD1*	6q22.2	C/A	0.50	0.46	0.48	7,089	4,726	0.84 (0.80–0.89)	1.55×10^{-9}	6,768	755	0.90 (0.81–1.01)	0.078	0.060
rs2395185[d] (rs28366298)	HLA class II region	6p21.32	Meta				7,390	4,696	1.20 (1.13–1.28)	9.47×10^{-10}	7,211	742	1.05 (0.93–1.18)	0.42	0.56

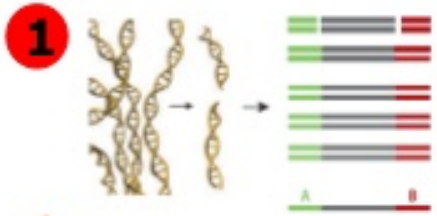
[a]Minor allele listed second. [b]Minor allele frequency. 1, MAF in controls; 2, MAF in adenocarcinoma; 3, MAF in squamous carcinoma. [c]Tested by case-case analysis. [d]For the HLA class II region, because a TaqMan assay could not be designed for rs2395185, we instead genotyped rs28366298, its perfect surrogate ($r^2 = 1.0$), by TaqMan. The reported *P* value is based on meta-analysis of the rs2395185 results in the GWAS and the rs28366298 results in the TaqMan set.

Nature Genetics 44, 1330–1335 (2012) doi:10.1038/ng.2456

Next Generation Sequencing

Next-generation DNA sequencing

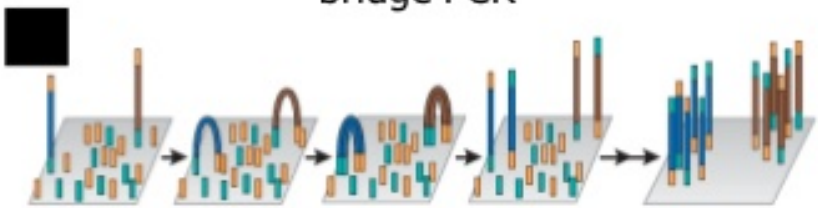
1. Library preparation
2. Clonal amplification
3. Cyclic array sequencing



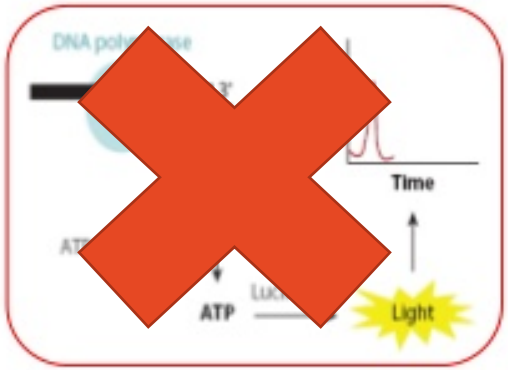
DNA fragmentation and in vitro adaptor ligation

emulsion PCR

bridge PCR



Pyrosequencing

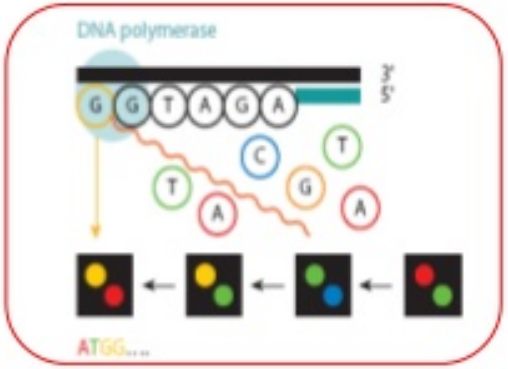


454 sequencing

Sequencing-by-ligation

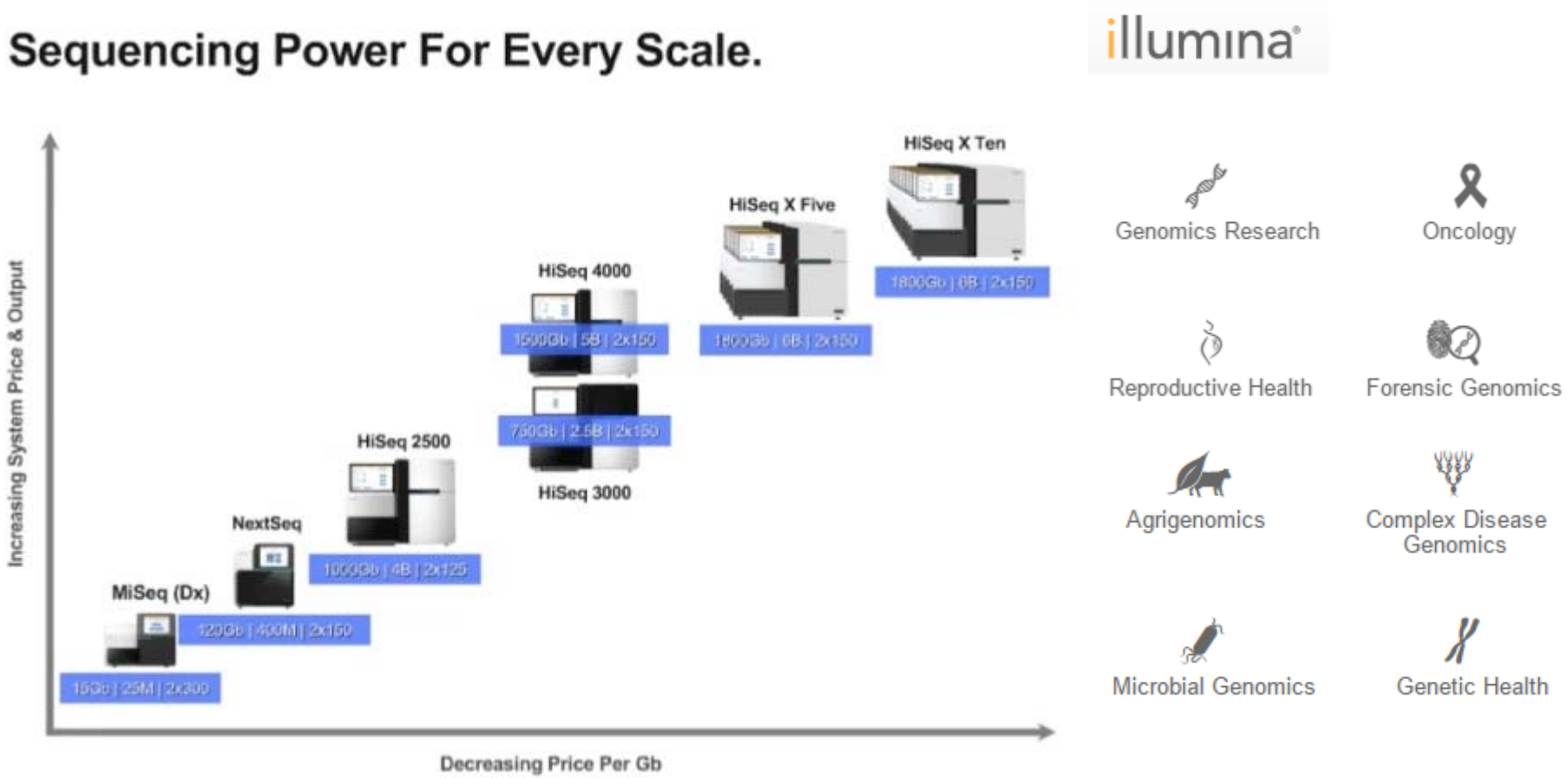
SOLiD platform

Sequencing-by-synthesis



Solexa technology

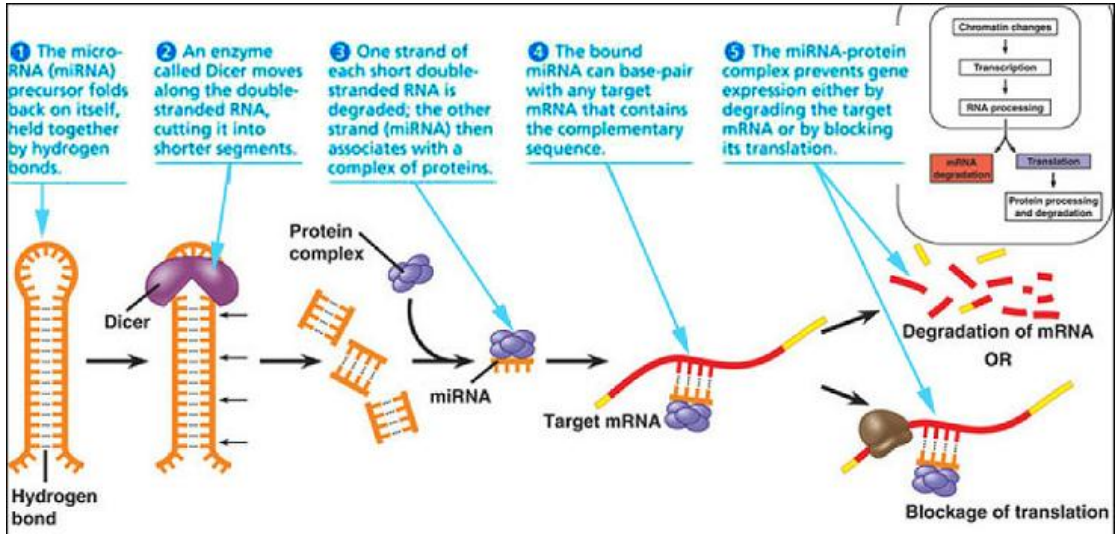
Application of NGS



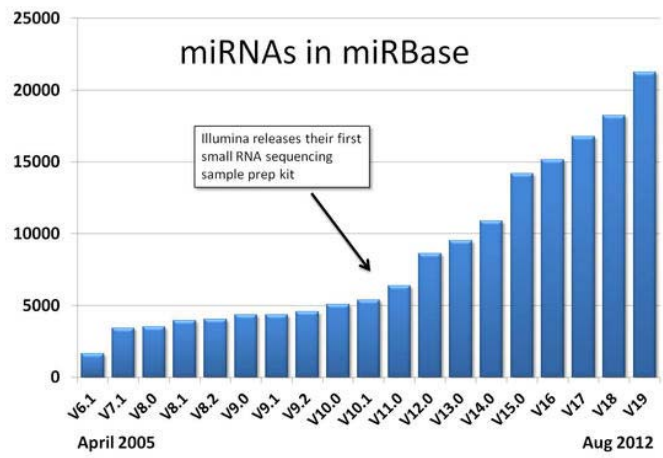
*Price per 30X human genome according to Illumina. We're not aware of any sequencing facility currently offering human genomes for $1,000.

RNA Sequencing

- RNA Sequencing (RNA-Seq) has quickly become the method of choice for discovery of new microRNAs (miRNAs) and other forms of small RNAs.
- Transcriptomics: the profiling of the transcriptome—aims to catalog the complete set of RNA transcripts produced by the genome, including mRNAs, non-coding RNAs, miRNAs, and other small RNAs

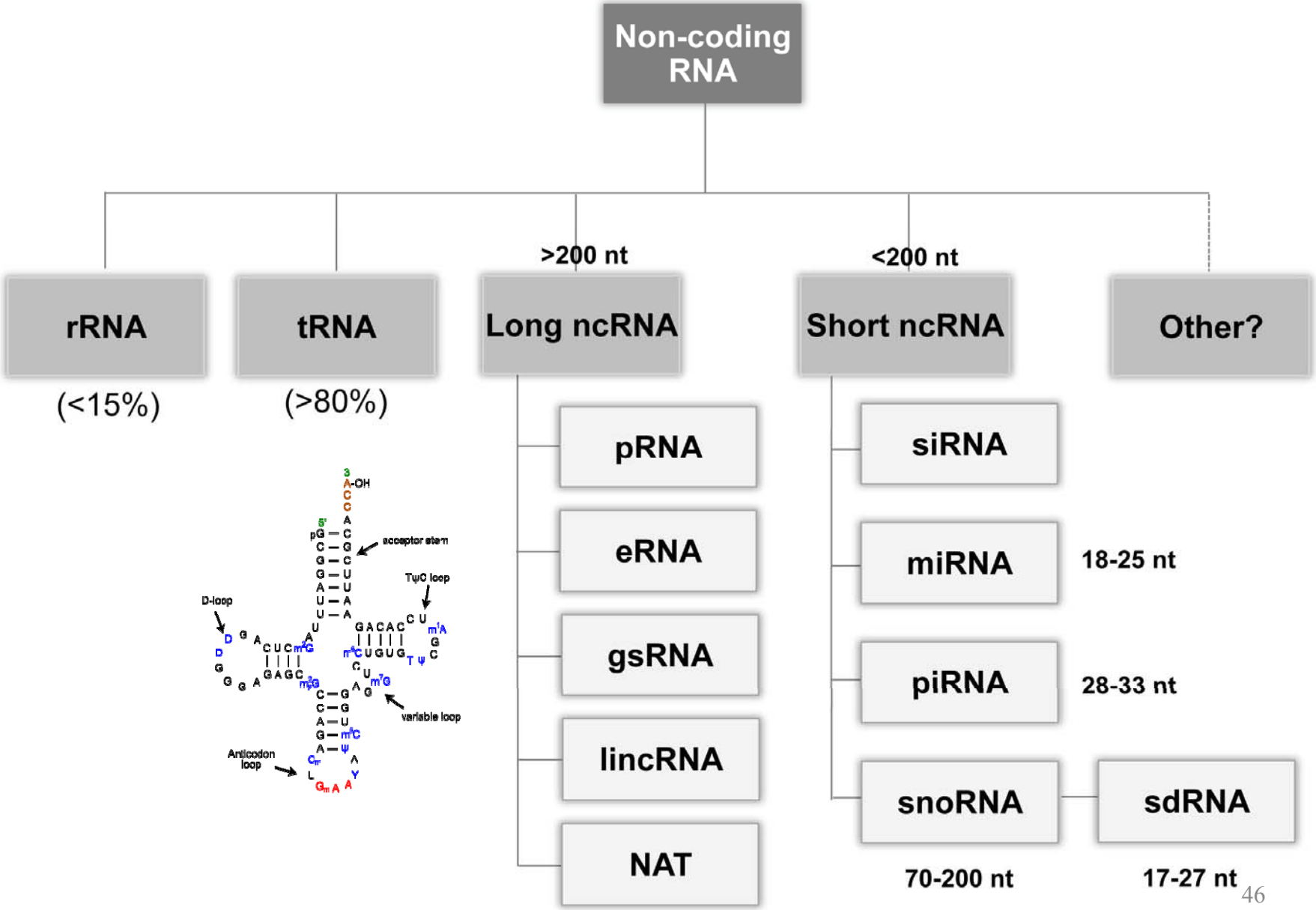


The schematic shows the major steps in miRNA processing and function. Image courtesy of Charles Mallery, University of Miami.



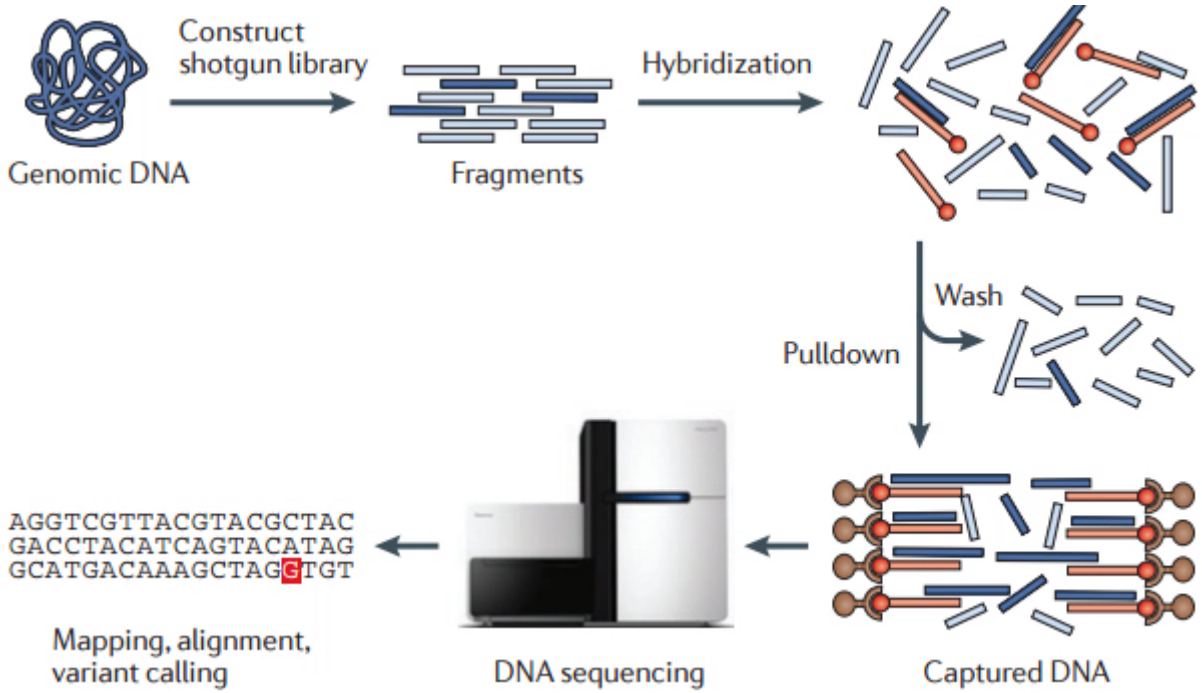
Release 21, miRNA count: 28645 entries

Classes of no-coding RNA



Exome sequencing

- Since 2007, there has been tremendous progress in the development of diverse technologies for capturing arbitrary subsets of a mammalian genome at a scale commensurate with that of massively parallel.
- To capture all protein-coding sequences, which constitute less than 2% of the human genome, the field has largely converged on the aqueous-phase



Nature Reviews Genetics 12, 745-755 (November 2011) | doi:10.1038/nrg3031

Applying Genome Variation - Will it work? YES!!

Hits:

Macular Degeneration, Obesity, Cardiac Repolarization, Inflammatory Bowel Disease, Diabetes T1 and T2, Coronary Artery Disease.Rheumatoid Arthritis, Breast Cancer, Colon Cancer, ……

-There are misses as well unclear why - Phenotype, Coverage, Environmental Contexts?

Example of a miss - Hypertension

-There are lots more hits in these data sets - sample size, low proxy coverage with other SNPs …..

-Analysis of associations between phenotype(s) and even individual sites is daunting and this will just be the first stage, and this does even consider multi-site interactions.

Thank you for your attention!!!