



Genome Sequencing Technologies

Jutta Marzillier, Ph.D.

Lehigh University

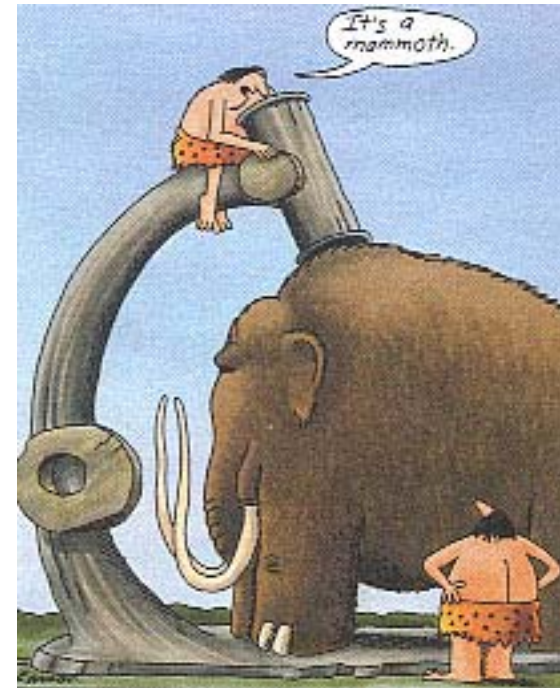
Department of Biological Sciences

Iacocca Hall

Sciences start with Observation



Sciences start with Observation and flourish with Introduction of Technology

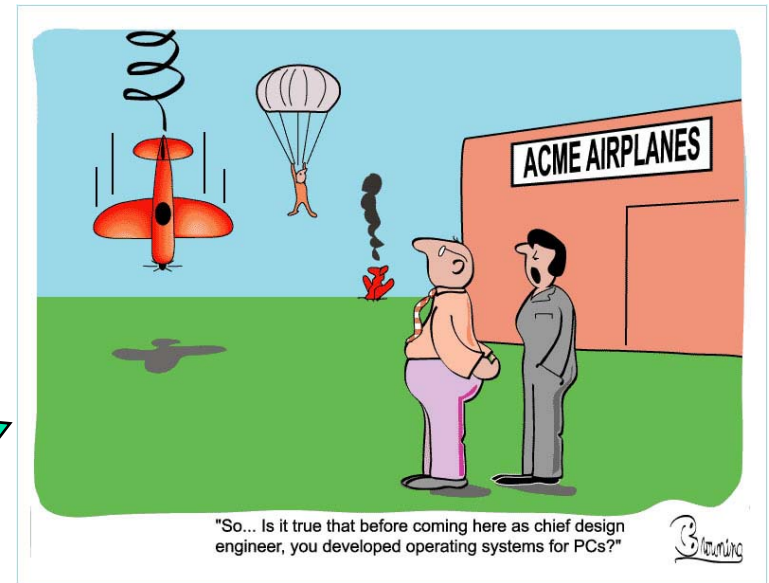


Life Sciences Thrive Through the Co-laboration between Biologists and Engineers

Biologist



Computer Engineer/ Bioinformatics



BioMedical/ Chemical/ Mechanical Engineer





Components that lead to completion of the Human Genome Project

Life Sciences:

Discovery of DNA (1953)
Invention of DNA Sequencing (1975)
Polymerase Chain Reaction (PCR, 1985)
Human Genome published (2003)

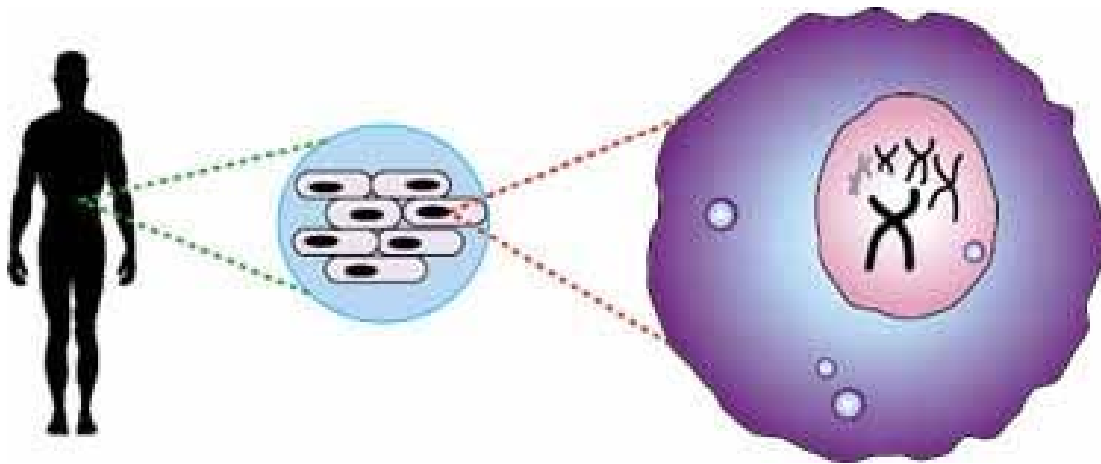
Engineering:

Dye coupling to DNA (1986)
Capillary electrophoresis (1995)
Robots
Automated Sequencer (1986)
Microfluidics
Nanotechnology

Computer Sciences/ Bioinformatics:

Computer languages (e.g. FORTRAN 1956)
Floppy disc (1970), hard drives
1st sequence database (EMBL) opens (1980)
Genbank (1982)
BLAST (Basic Local Alignment Search Tool) (1990)
World Wide Web (1989)
'PC' revolution (1st PC by IBM in 1981)

What is the Genome ?



The human body has about 100 trillion cells

Each cell carries the same genetic information (**genome**) in its nucleus

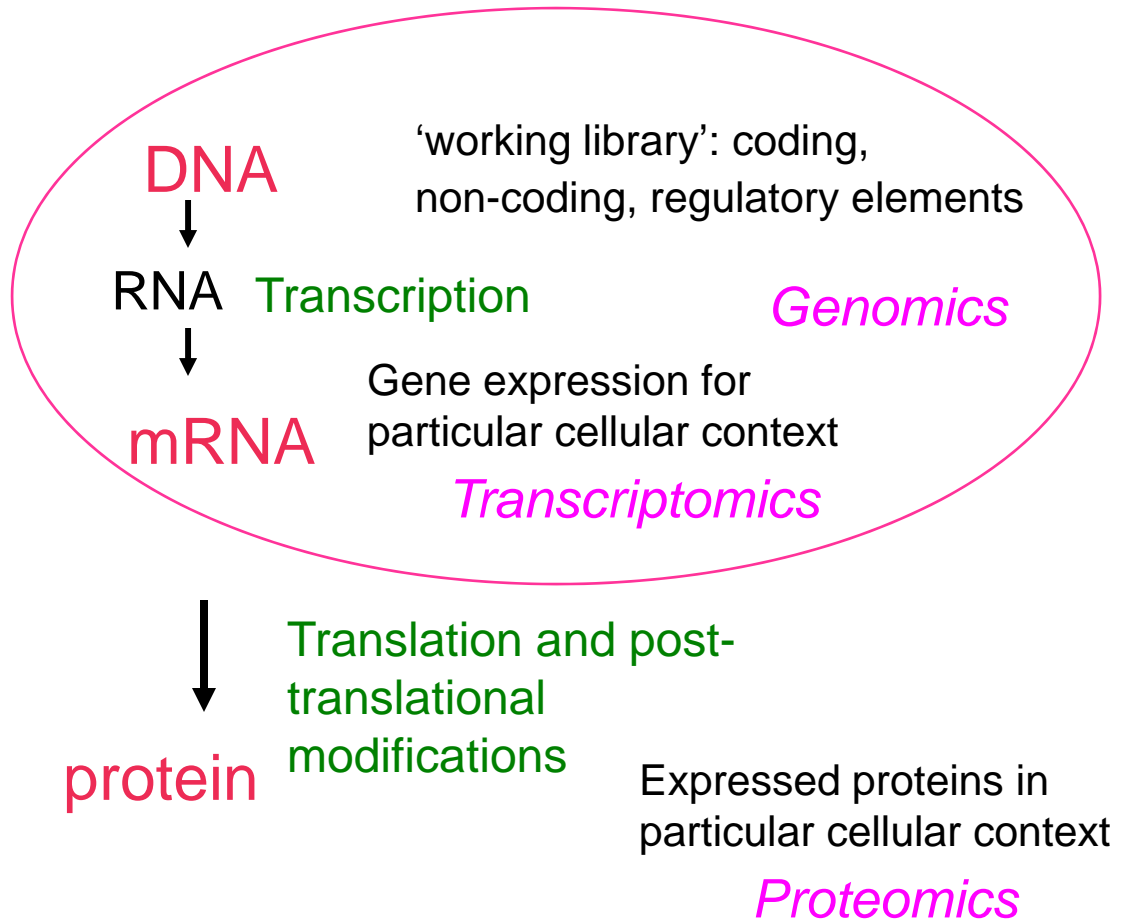
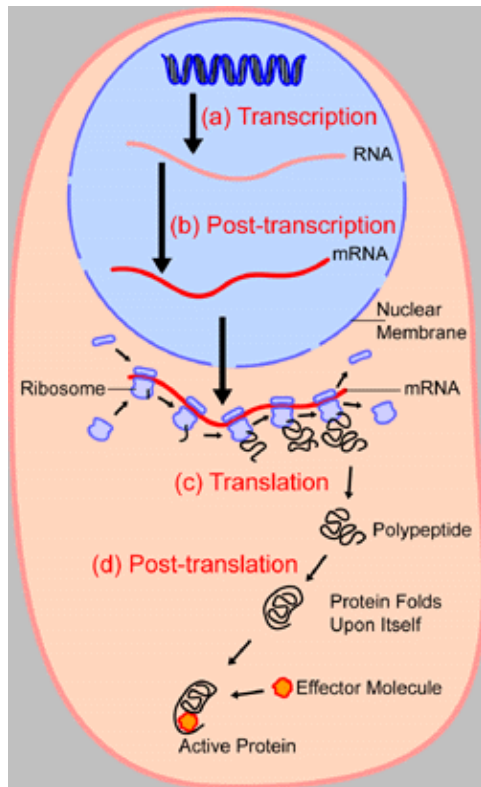
Depending on the cellular context only a subset of genes is expressed (**transcriptome**)

© 2002 The Center for the Advancement of Genomics (TCAG).

Why study genes?

The study of genes is crucial to our understanding of the cell, development and disease.

Main Biological Dogma



DNA structure and synthesis

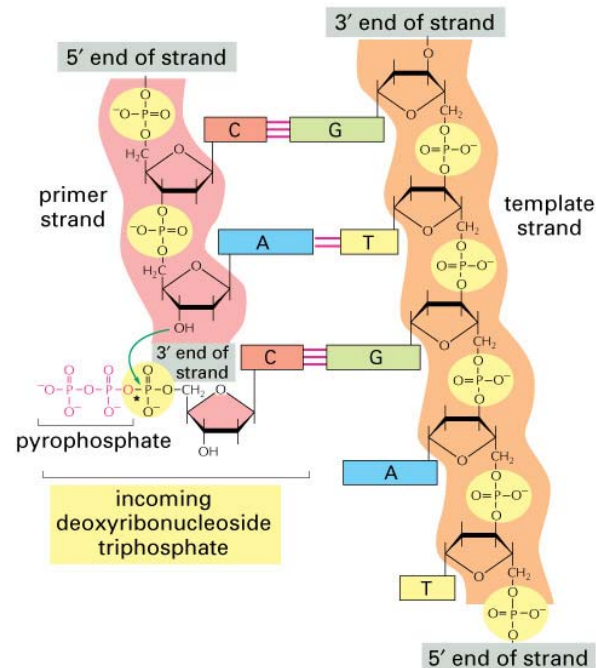
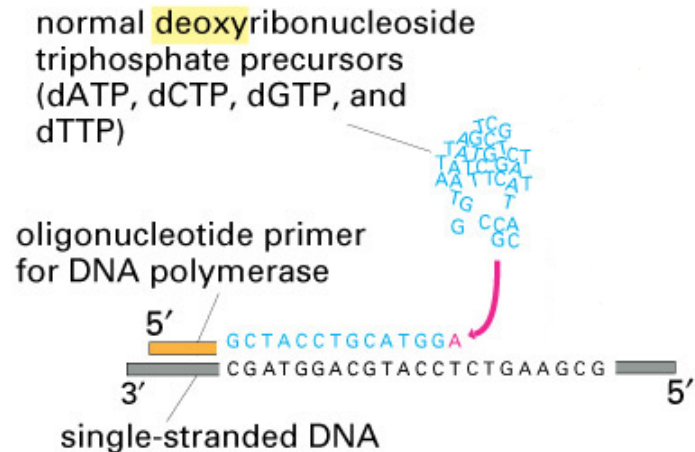


Figure 6-10 Essential Cell Biology, 2/e. (© 2004 Garland Science)

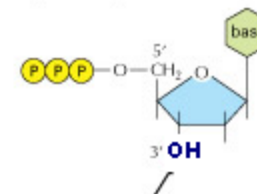
DNA: chemically linked chain of nucleotides which each consists of a phosphate, a sugar (ribose), and a nucleobase (Adenine, Guanine, Cytosine, Thymine)

DNA sequencing – Principle of DNA synthesis

- Amplification of targeted DNA strand in presence of
 - primer that only hybridizes to one complementary strand-
 - (DNA) polymerase, (or Taq)
 - deoxynucleotides (dATP, dTTP, dCTP, dGTP).



deoxyribonucleoside triphosphate



allows strand extension at 3' end



Frederick Sanger



* 13 August 1918

Two time Nobel laureate in chemistry

1958: identification of the amino acid
sequence of insulin

1980: developed the chain termination
method for DNA sequencing

Dideoxy Method of Sequencing (Sanger, 1975)

- DNA synthesis is carried out in the presence of limiting amounts of **dideoxyribonucleoside** triphosphates that results in chain termination
- Original method used **radio-labeled** primers or dideoxynucleotides

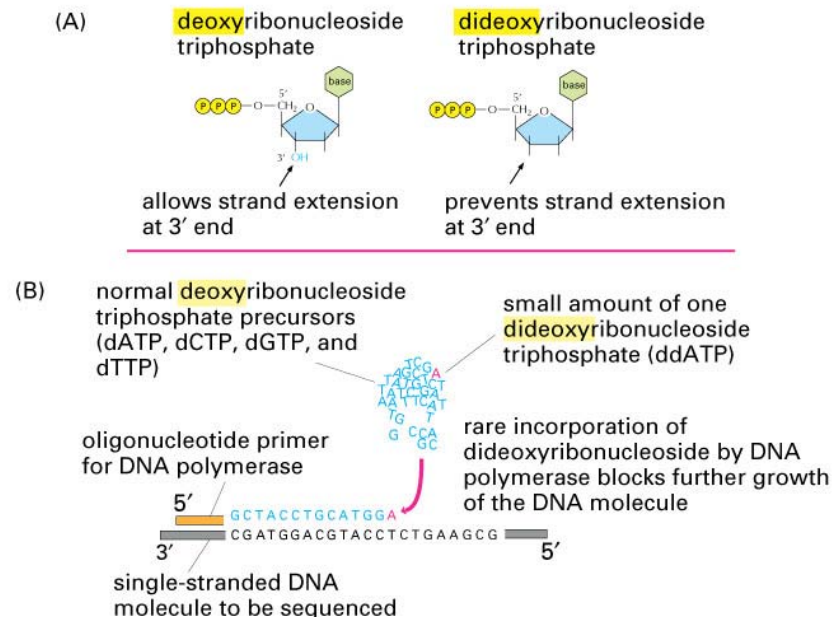


Figure 10-7 part 1 of 2 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Sanger DNA sequencing using radiolabel

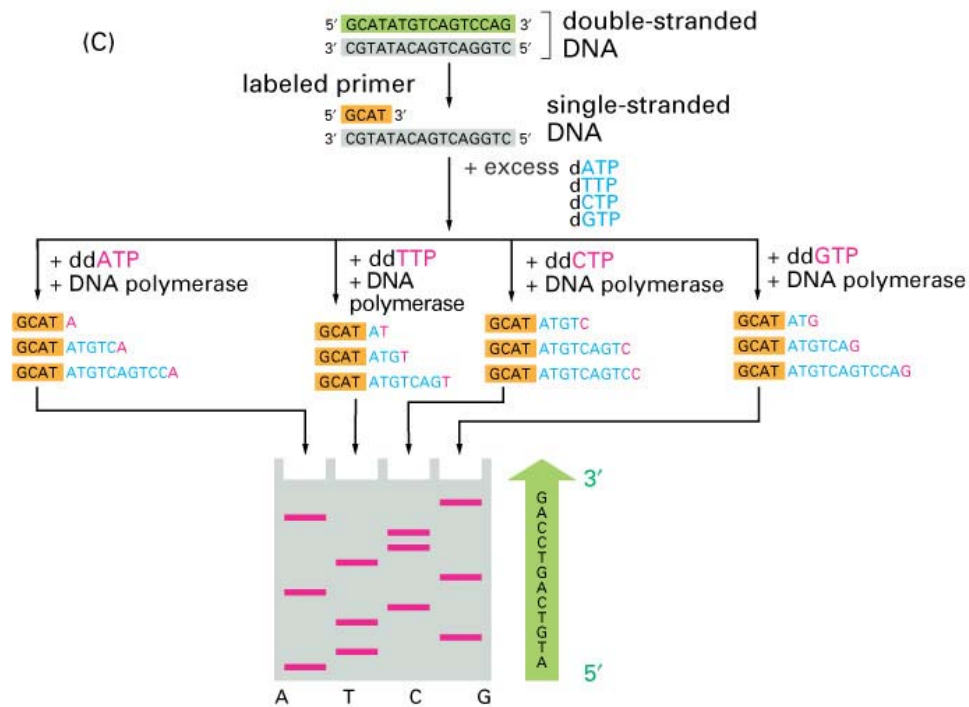


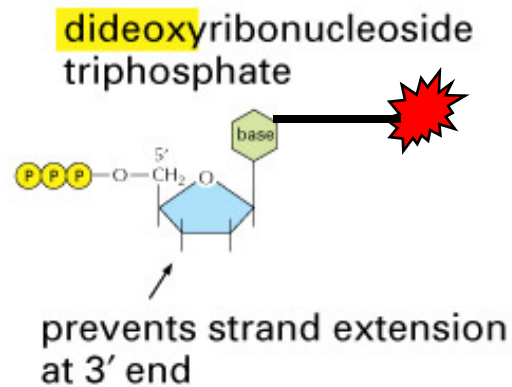
Figure 10-7 part 2 of 2 Essential Cell Biology, 2/e. © 2004 Garland Science)

Fluorescently labeled ddNTPs used for sequencing reaction

Use of thermo-stable **Taq-polymerase** allowed automation.

ddNTPs used for automated sequencing are labeled with different **fluorescent** dyes representing each of the 4 bases

Enabled single lane electrophoresis



Automated DNA Sequencing- use of labeled dideoxynucleotides

- Random incorporation of ddNTPs results in termination of elongation reaction
- This reaction set-up will produce a set of DNAs of different lengths complementary to the template DNA
- Each fluorescent label corresponds to specific nucleotide.

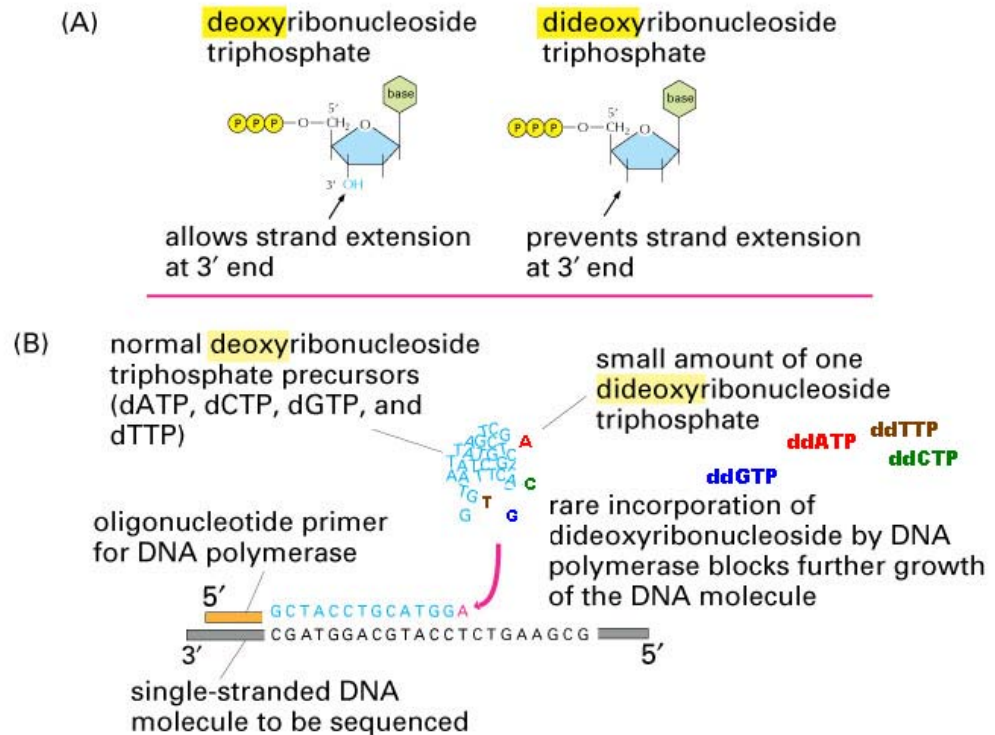
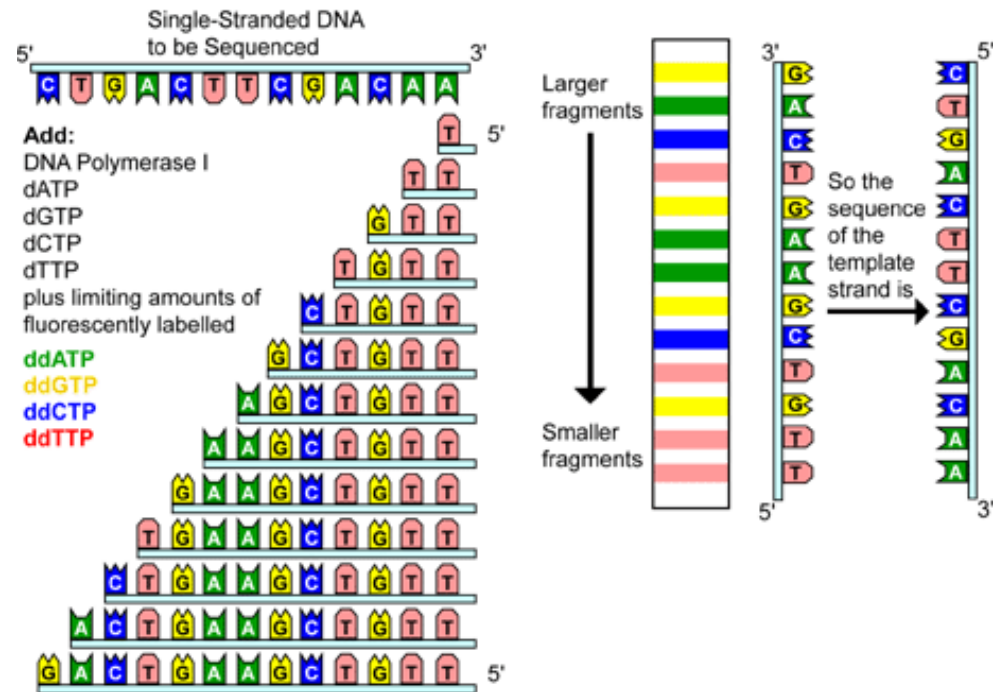


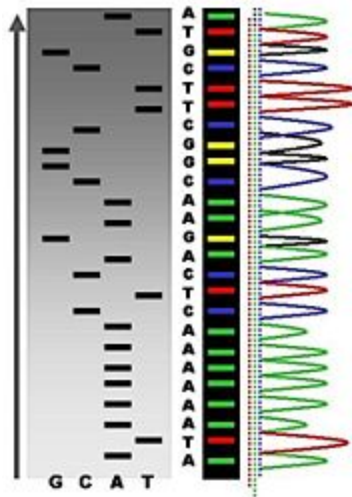
Figure 10-7 part 1 of 2 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Separation of Amplified Fragments

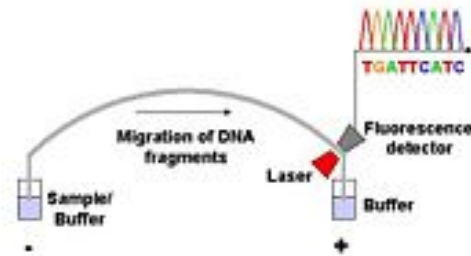
- When a fluorescence labeled, chemically modified dideoxynucleotide is integrated into the DNA strand, the reaction will be terminated, thus leaving a labeled fragment of a defined size.
- These fragments will then be separated by **capillary electrophoresis** according to their size.



Capillary Gel Electrophoresis



Sequence ladder by radioactive sequencing compared to fluorescent peaks



Capillary gel electrophoresis: Samples are excited by laser and emitted fluorescence read by CCD camera. Fluorescent signals are converted into basecalls.

Automated DNA Sequencing

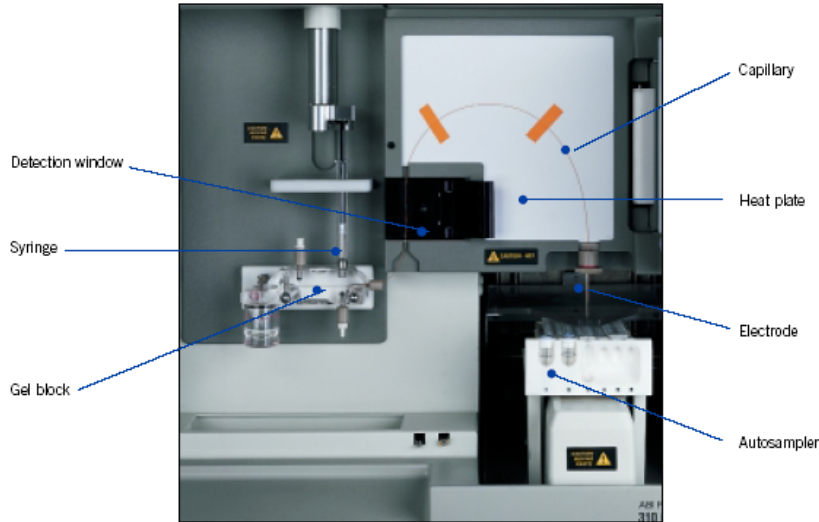


ABI 310 one capillary
DNA sequencer

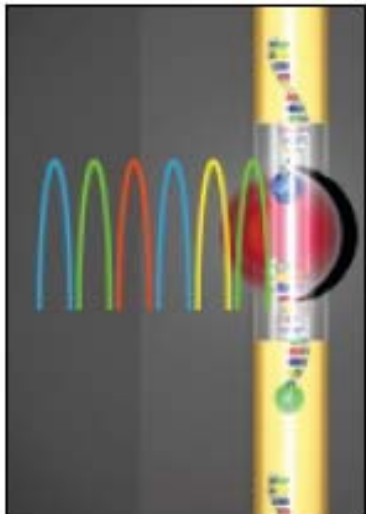
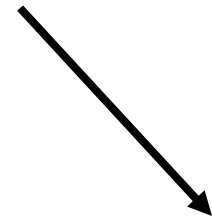
Jutta Marzillier, 2007

ABI 310 Genetic Analyzer

Inside the ABI PRISM® 310 Genetic Analyzer



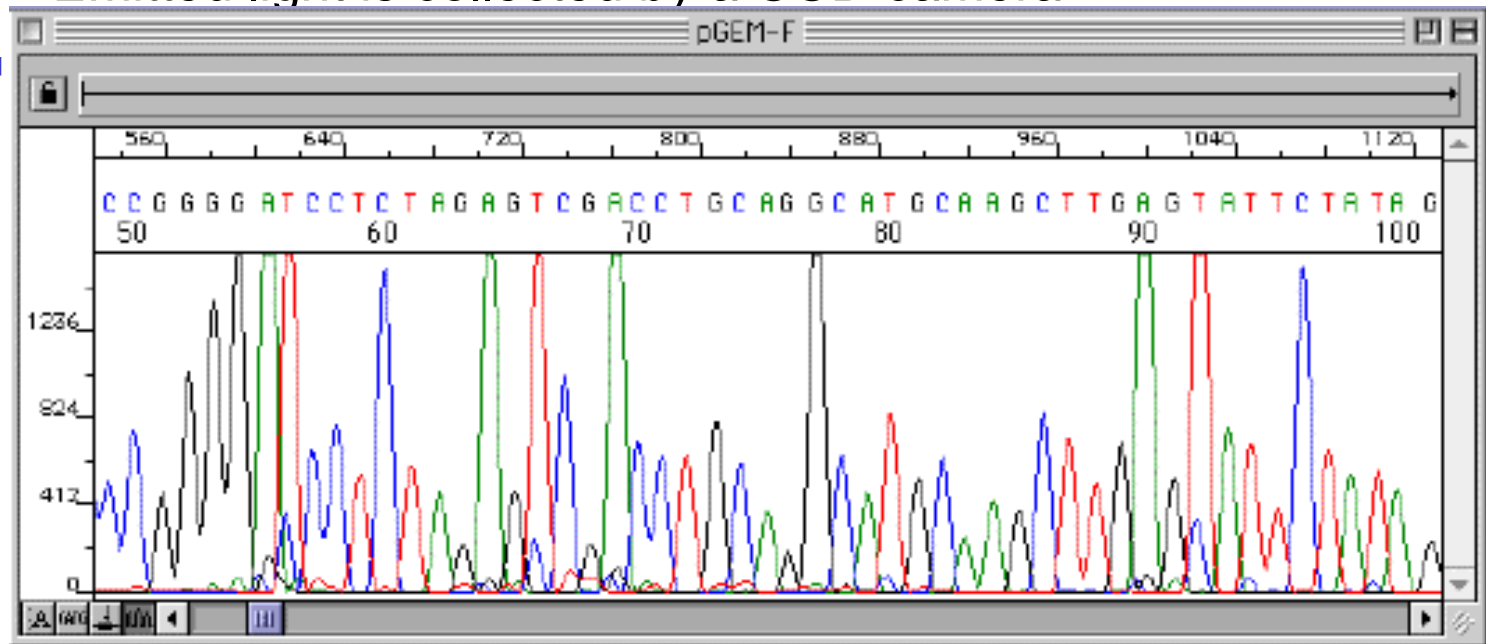
Capillary and electrode:
Negatively charged DNA
migrates towards anode



DNA fragments labeled with
different fluorescent dyes
migrate according to their
size past a laser

Sequencing Electropherogram

- Laser excites dyes, causing them to emit light at longer wavelengths
- Emitted light is collected by a CCD camera

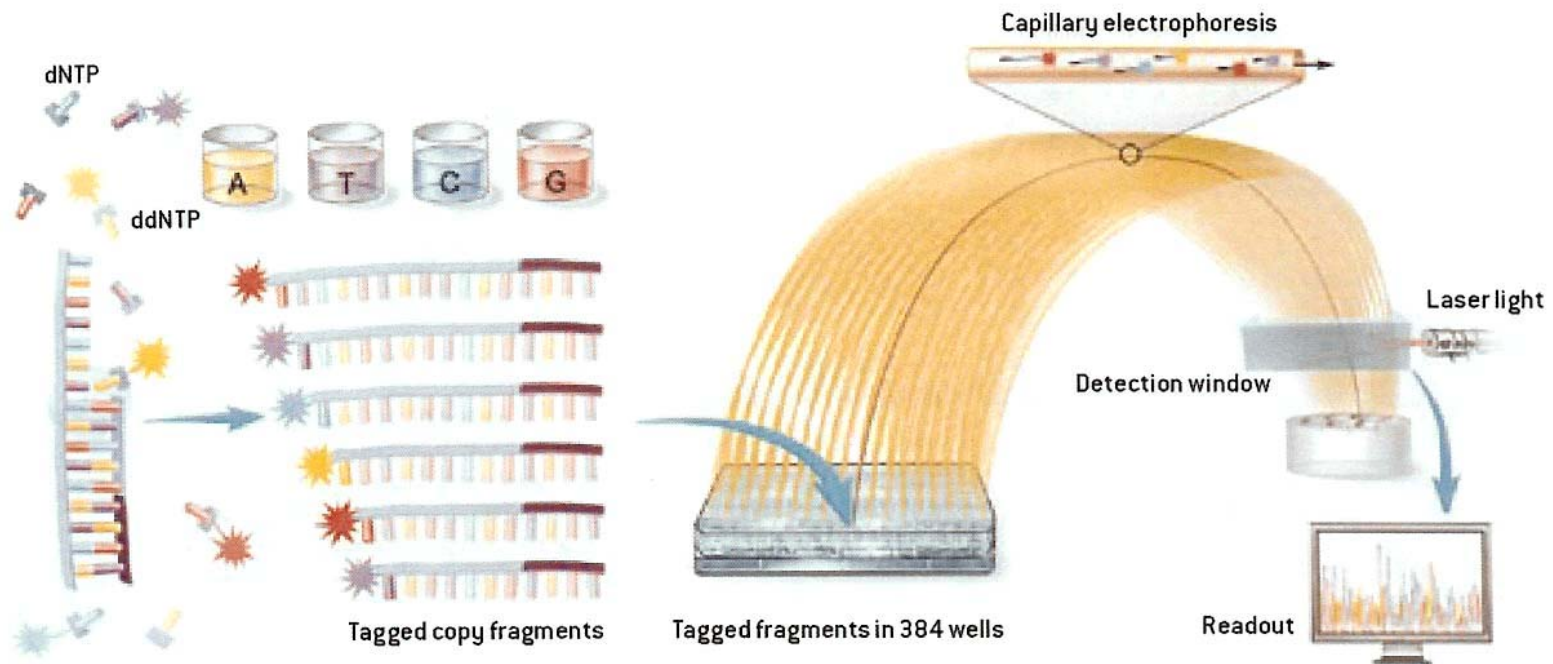


Plot of colors detected in sequencing sample scanned from smallest to largest fragment

www.contexto.info/DNA_Basics/dna_sequencing.htm

High-Throughput Whole Genome Sequencing

Analysis of 384 sequencing reactions in parallel

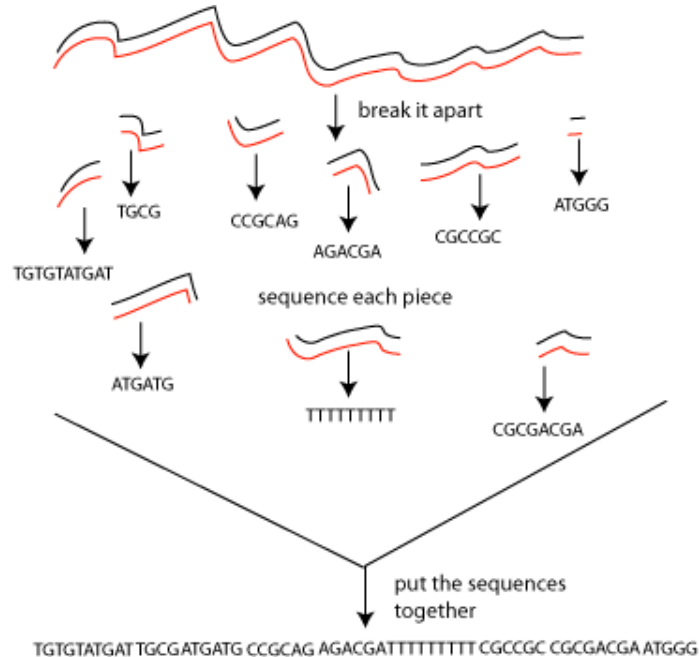


George Church, Scientific American, January 2006, pp47-54

Jutta Marzillier, 2007

How do you sequence a genome?

Mapping versus Shotgun



Mapping: break genome into smaller pieces, align the pieces to each other and sequence

Shotgun: break genome into small pieces, sequence and re-align by finding overlapping sequences



The Human Genome Project

Initial sequencing and analysis of the human genome

Nature **409**, 860 - 921 (15 February 2001)

[International Human Genome Sequencing Consortium](#)

The Sequence of the Human Genome

Science, Vol 291, Issue 5507, 1304-1351 , 16 February 2001
Venter et al. (Celera Genomics)

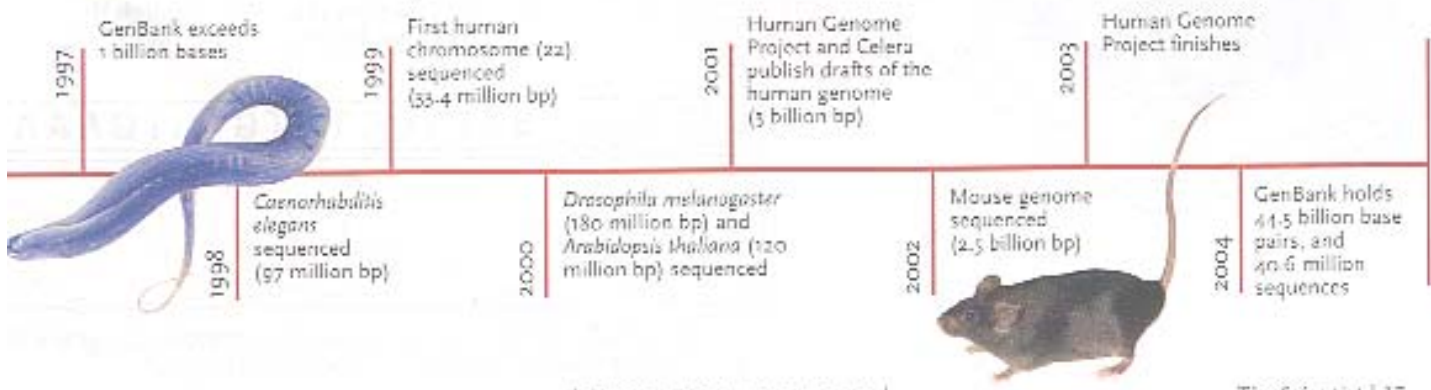
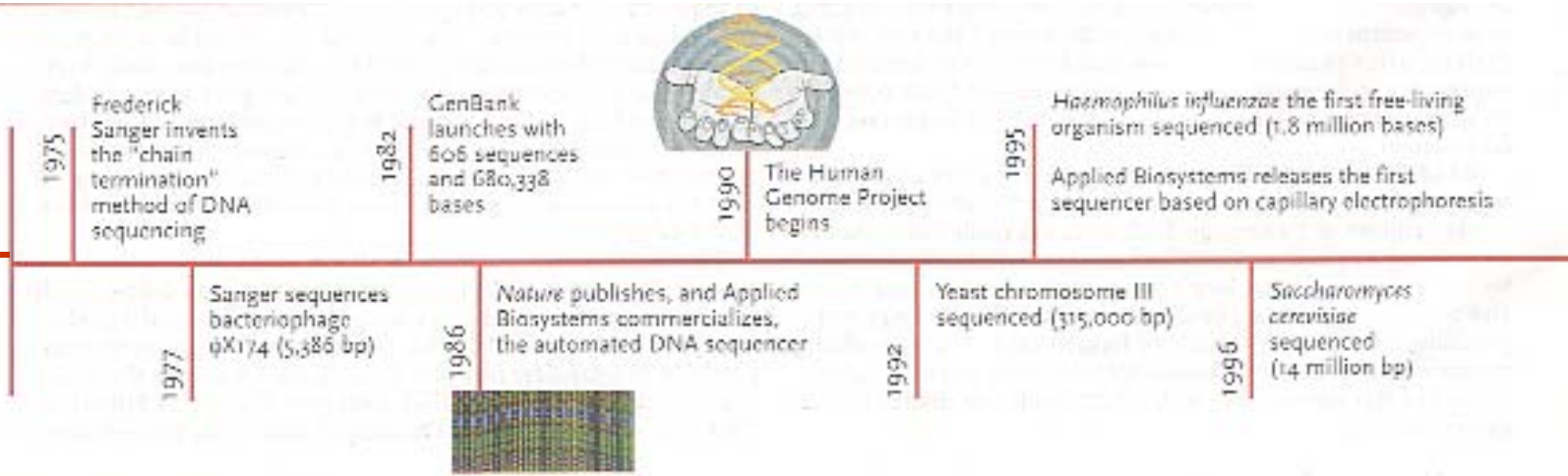
Finishing the euchromatic sequence of the human genome

Nature **431**, 931 - 945 (21 October 2004)

[International Human Genome Sequencing Consortium](#)

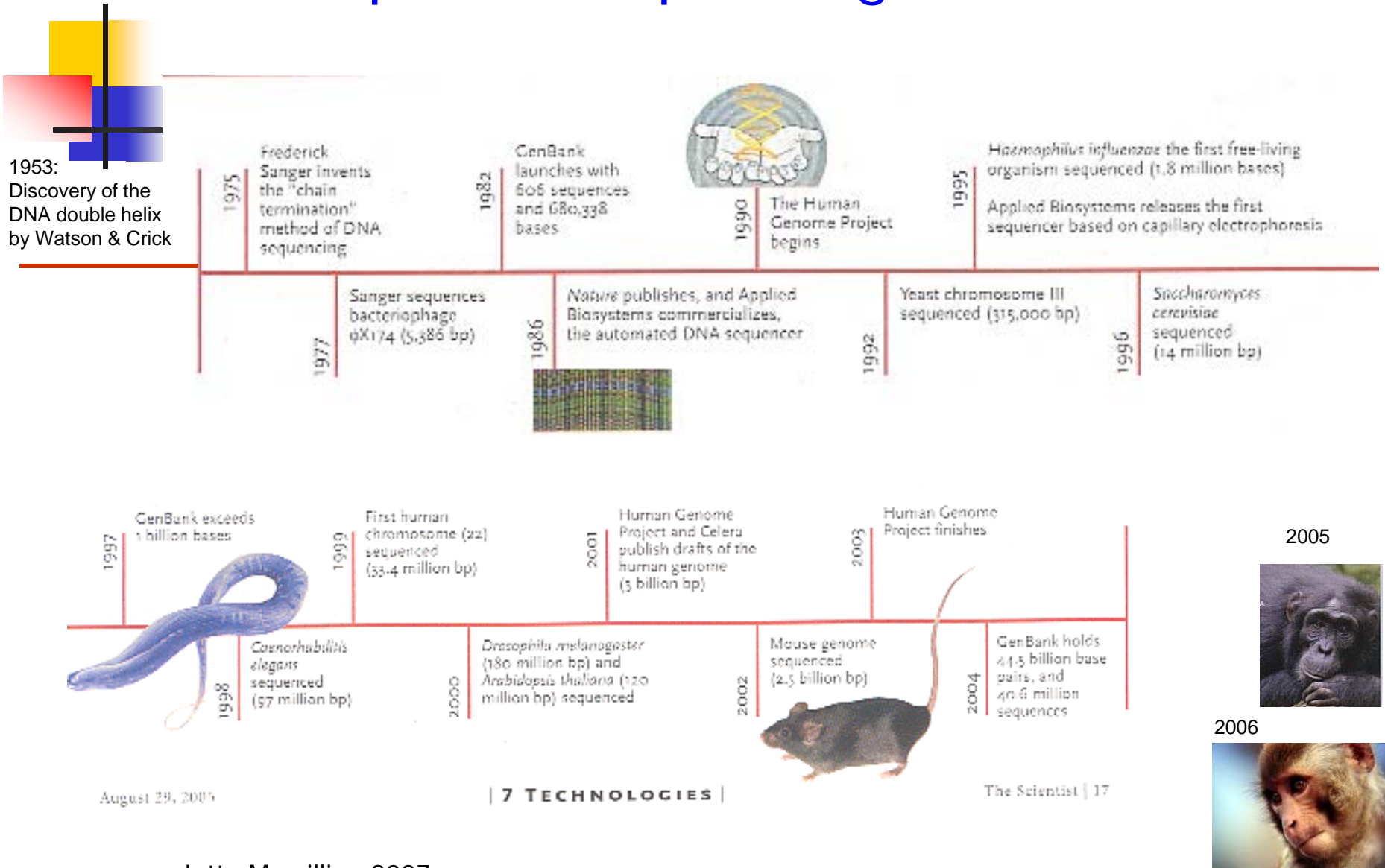
Development Sequencing Time Line

1953:
Discovery of the
DNA double helix
by Watson & Crick



Development Sequencing Time Line

1953:
Discovery of the DNA double helix by Watson & Crick



August 29, 2005

| 7 TECHNOLOGIES |

The Scientist | 17

The Human Genome

- * Current genome (build 36) contains 2.85 billion nucleotides interrupted by only 341 gaps (2004)
- * Overall error rate is less than 1 error per 100,000 bp
- * Encodes for approx. 25,000 to 35,000 protein-coding genes
- * Genes make up about 1-2 % of the total DNA, the coding portions of a gene are called **exons**, which are interrupted by intervening sequences, called **introns**.
- * Genetic variation between individuals between 0.1-0.5%.

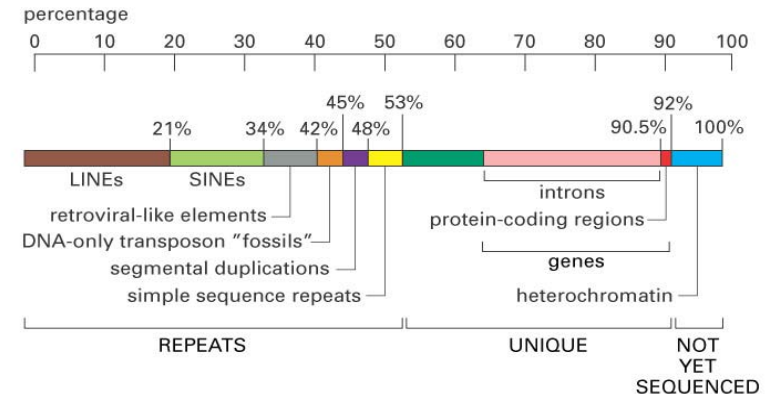
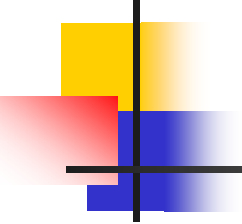


Figure 9-26 Essential Cell Biology, 2/e. (© 2004 Garland Science)



Who's sequence was used for the 'Human Genome Project'?

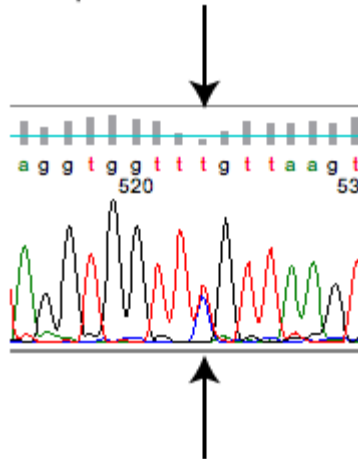
- Different sources of DNA were used for original sequencing
- Celera: 5 individuals; HGSC: 'many'
- Rationale: All humans share the same basic set of genes and genomic regulatory regions.
- The term 'genome' is used as a reference to describe a **composite** genome
- The many small regions of DNA that vary among individuals are called **polymorphisms**:
 - Mostly **single nucleotide polymorphisms** (SNPs)
 - Insertions/ deletions (indels)
 - Copy number variation
 - Inversions

Single Nucleotide Polymorphisms (SNPs)

5' AGGTGTTTGTAAAGT 3'
5' AGGTGTTCGTTAAAGT 3'

Only one strand is shown,
from each chromosome.

This is identified as a low quality
base because there are two bases
at this position.



This peak at position 523 shows
both a T and a C.

The two forms of
an SNP are called
alleles.

Most SNPs are without
a physiological effect.
They contribute to human
variety.
A smaller number correlates
with an individual's
susceptibility for certain
diseases or response to
drug treatments or disease.

http://scienceblogs.com/digitalbio/2007/09/genetic_variation_i_what_is_a.php#more



How Does the Human Genome Compare to Others?

	# of protein-coding genes
Human	25,000-35,000
<i>C. elegans</i> (roundworm)	19,000
<i>D. melanogaster</i> (fruitfly)	15,000
<i>Haemophilus</i> (bacterium)	1,738

Are we only twice as complex as the roundworm and the fruit fly?

- Many genes encode for more than one protein product (**alternative splicing, RNA modifications, microRNAs**)
- Posttranslational modifications contribute to the complexity of the human proteome.
- **Transcription factors** and **enhancers** probably provide greater flexibility of gene expression
- Additional genes not found in invertebrates, such as genes encoding for antibodies, T-cell receptor, cytokines, etc.

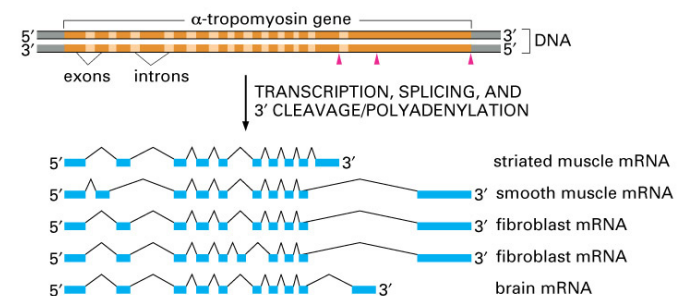


Figure 7-18 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Alternative splicing of the
alpha-tropomyosin gene

JGI Sequencing Facility

(Joint Genome Institute, US Department of Energy)



Assume 20 384-capillary
sequencers running simultaneously
approx. 700 bp per capillary run
approx. 3 hours per run

➔ Approx. 40 Million bases per day
in one facility



Hurdles to overcome

Costs

Low sensitivity, DNA needs to be amplified to be detected

Electrophoresis is slow and has limited resolution

Limited parallelization, semi high-throughput

More powerful data analysis and bioinformatics tools for analysis of short fragments

Change from bench-top robotics to microfluidic platforms



The Race for the \$1,000 Genome

Human Genome Project (2001, initial draft):
\$ 2-3 billion (includes development of technology)
“raw” expenses estimated at \$300 million

Rhesus macaque (2006)
\$ 22 million

Expected by end of 2006:
\$ 100,000 for full mammalian genome sequence

Wanted: “!!!!!! The \$ 1,000 Genome !!!!!!!”

- low cost
- high-throughput
- high accuracy



Recent Technologies

- Pyrosequencing (454 Life Sciences, Roche)
 - Emulsion PCR, Sequencing by synthesis
- SOLID System (ABI)
 - Emulsion PCR, Sequencing by ligation
- Genome Analyzer (Illumina)
 - Emulsion PCR, sequencing by synthesis using fluorescently labelled reversible chain terminators

Single-molecule sequencing

- Nanopore
- Atomic Force Microscopy

Pyrosequencing – 454 Life Sciences

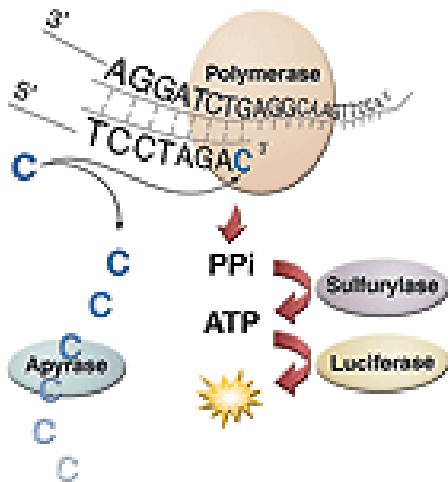
Addition of dNTP releases a pyrophosphate (PPi) stoichiometrically

Sulfurylase converts PPi to ATP

ATP and Luciferase drive conversion of luciferin to oxyluciferin that generates visible light and can be measured with a CCD camera.

Each light signal is proportional to the number of nucleotides incorporated.

Apyrase digests unincorporated nucleotides



Biotage, 2005

Real time sequencing

Amplification of DNA by emulsion PCR

Gel free, microchips with 800,000 holes

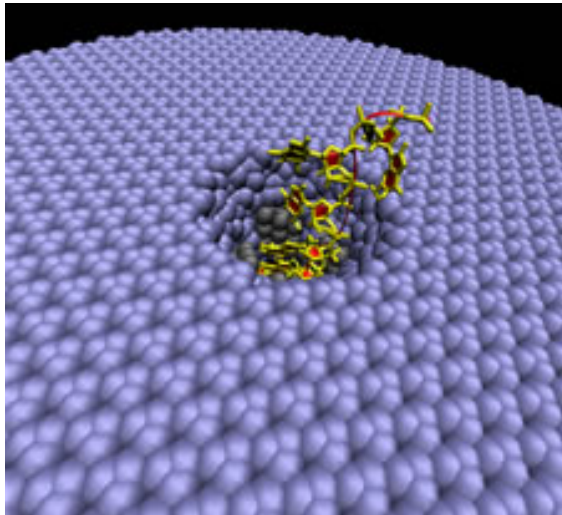
25 million bases in 4 hour run

Shorter read length, approx. 100 bp, more complicated genome assembly

Jutta Marzillier, 2007

Difficult to sequence repetitive stretches

Single molecule sequencing - Sequencing through nanopore



Johan Lagerqvist
www.ucsdnews.ucsd.edu

Measure changes in electric current as DNA is passes through a nanopore surrounded by two pairs of tiny gold electrodes.

Electrodes record electrical current perpendicular to the DNA strand
As each DNA base is structurally and chemically different, each base creates its own electronic signature.

Proposed model, not proven yet.

Single molecule sequencing models using **AFM** also proposed.



Recent achievements

May 2007: James Watson's genome deposited
454 technology, 2 month, \$ 1-2 million

Sept 2007: Craig Venter's genome deposited
Sanger technology, 4 years, 70 million

*Diploid genomes

*Approx. 3.5% of Watson's genome could not be matched to the reference genome

*Venter's genome had 4.1 million DNA variants comprising 12.3 Mb
- 3.2 million SNPs
- non-SNP variants

*These variants include a potentially increased risk of alcoholism, coronary heart disease, obesity, Alzheimer's disease, antisocial behavior and conduct disorder.



Aspects of the \$ 1,000 Genome

Triggers basic research:

- how is activation of genes regulated?
- understanding genetic links to cancer
- facilitate genetic engineering
- hunt for “disease” genes
- diagnosis and treatment of diseases?
- personalized patient treatment strategies?

However:

- How can patient privacy be protected?
- Will insurance companies and employers use genetic information to screen out those at high risk for disease?
- genetic engineering of bioterrorism agents