N Bases / Nucleosides / Nucleotides / Nucleic Acid Structures (Review)

Goals for this review unit:

1. ~~Recognize the common building blocks of nucleic acids: names / 1-letter abbrev.~~
2. ~~Nomenclature for nucleosides and nucleotides (structure ofATP)~~
3. ~~Primary structures of RNA and DNA~~
4. ~~Conformations in DNAs~~
5. ~~Characteristics of B-DNA, A-DNA and Z-DNA~~
6. ~~Denaturation of DNA~~
7. ~~Features of RNA / Functions of RNA~~
8. DNA Sequencing (Maxam – Gilbert vs. Sanger Dideoxy)

ENCODE (Encyclopedia of DNA Elements)

ENCODE involved 440 scientists from 32 labs in the United States, United Kingdom, Spain, Singapore, and Japan. Since 2007, they have collected more than 15 terabytes of raw data that describes places in the genome that contain regulatory binding sites, areas of frequent DNA modification, or roles in managing the larger chromatin structure of DNA.

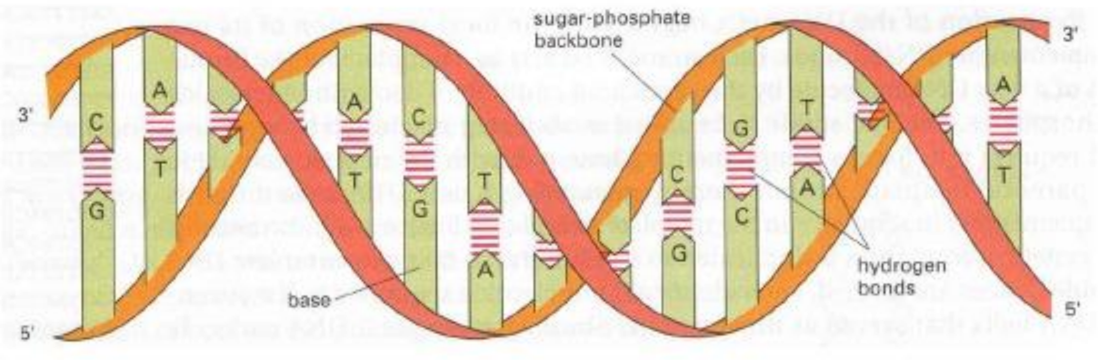
Researchers from around the world have been collaborating for the past five years to understand the non-coding regions of the human genome—the more than 95% of the genome that's been dubbed "junk DNA" . Now, with the simultaneous publication of 30 papers describing their findings, the team has reported that more than 80% of the human genome does indeed have a function.

Sites with high levels of cleavage by DNAse1—called DNAse I hypersensitive regions (DHS)—are known to contain DNA regulatory elements. They determined the placement of these DHSs and then aligned them with more than 5,000 gene variants associated with 207 diseases and 447 traits identified in GWAS (genome-wide association studies).

In a paper published September 5 in *Science*, the team reported that 76% of these disease-associated gene variants fell within DHSs.

Maurano, MT, R. Humbert, and E. Rynes. 2012. Systematic localization of common disease-associated variation in regulatory DNA.Science Vol. 337: 1190-1196.

Genetic information



... G T A C T G A A C G C A G G T...

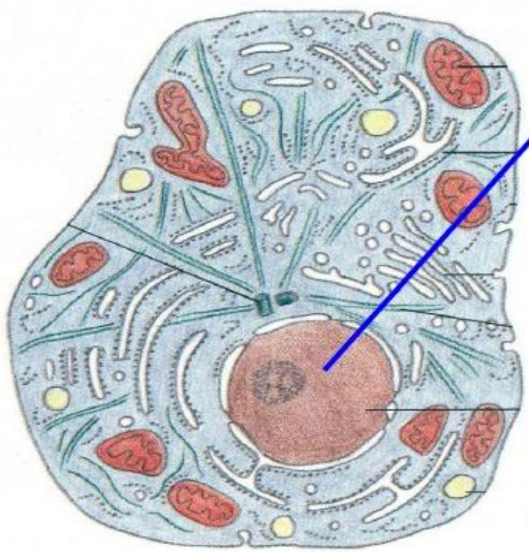
Genetic code

Human being: ~ 3,000,000,000 base-pairs

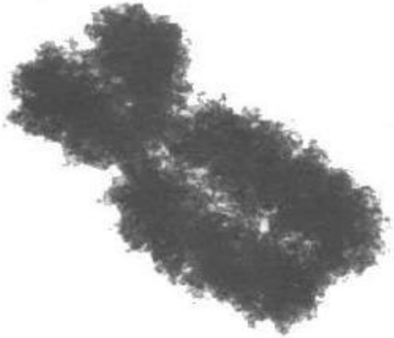
~ 30,000 – 40,000 Genes

(Public Human Genome Project and Celera Genomics)

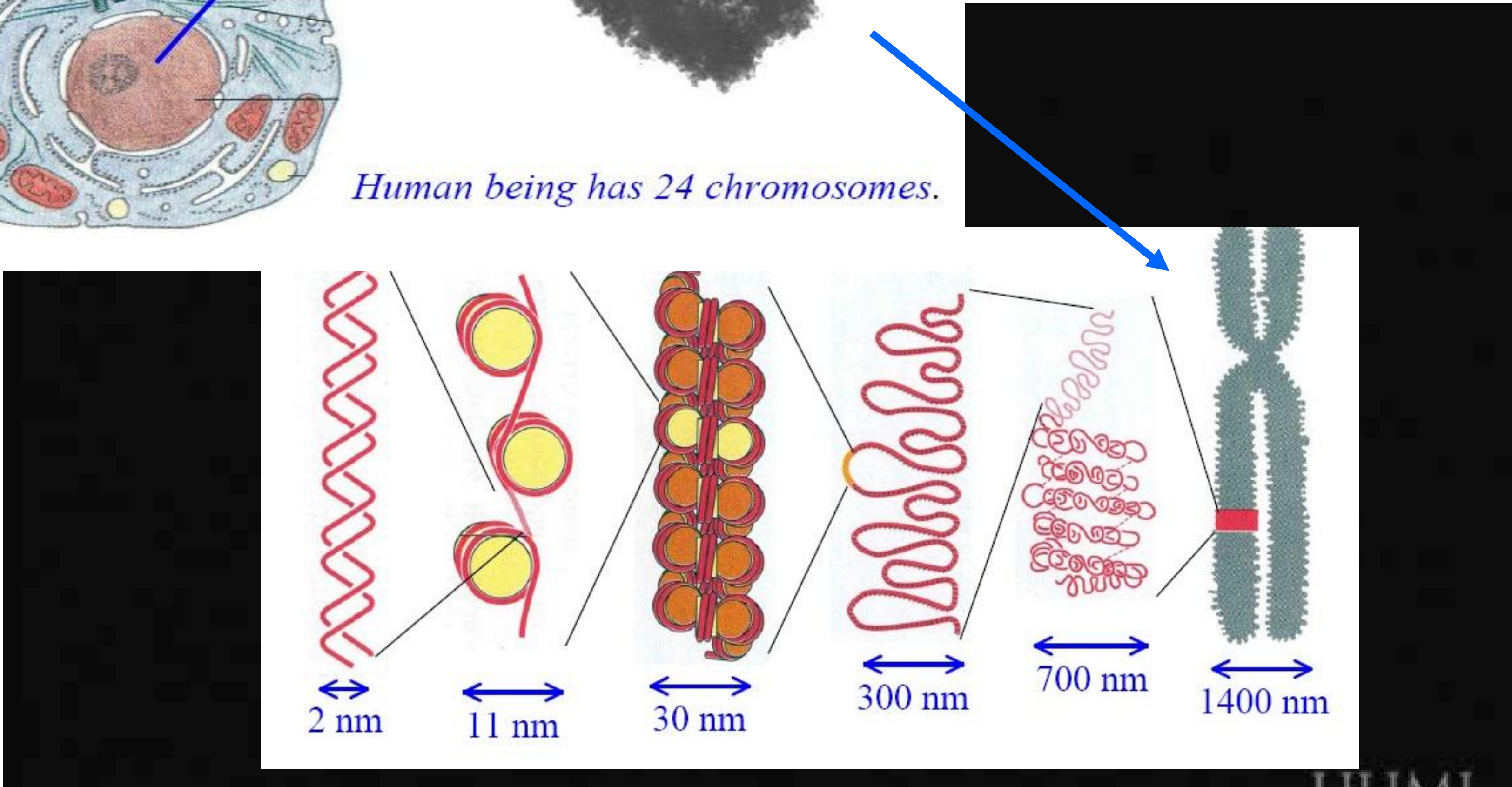
Chromosome



Chromosomes



Human being has 24 chromosomes.

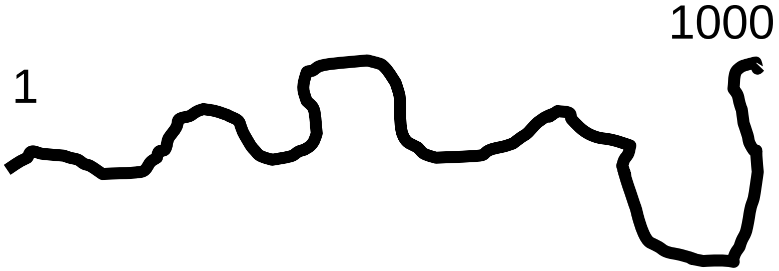


HHMI: http://www.hhmi.org/biointeractive/dna/DNAi_packaging_vo2.html

iClicker Question #1:

Consider the world at 1,000,000X, estimate how long a piece of DNA containing 1000 bp (333 a.a.) would be at this magnification?

A) 0.00034 mm
B) 0.034 mm
C) 3.4 mm
D) 340 mm
E) 3.4 m



Size of the Human Genome

The human genome comprises the information contained in one set of human chromosomes which themselves contain about 3 billion base pairs (bp) of DNA in 46 chromosomes (22 autosome pairs + 2 sex chromosomes).

The length on an average gene (330 a.a.) can be estimated as:

(length of 1 bp)(number of nucleotides per gene)

(0.34 nm)(1000) = 340 nm = 3.4 × 10^{-4} **mm** **(~13 inches at 1,000,000X)**

The length on the human genome can be estimated as:

(length of 1 bp)(number of nucleotides per cell)

(0.34 nm)(6 × 10^{9}) = **2 m** **(~1300 miles at 1,000,000X)**

The total length of DNA present in one adult human is estimated by:

(length of 1 bp)(number of bp per cell)(number of cells in the body)

(0.34 × 10^{-9} m)(6 × 10^{9})(10^{13}) = **2.0 × 10^{13} meters**

(The equivalent of nearly 70 trips from the earth to the sun and back.)

Sequencing DNA

Prior to the mid-1970's no method existed by which DNA could be directly sequenced. Knowledge about gene and genome organization was based upon studies of prokaryotic organisms and the primary means of obtaining DNA sequence was so-called reverse genetics in which the amino acid sequence of the gene product of interest is back-translated into a nucleotide sequence based upon the appropriate codons.

- **Maxam-Gilbert DNA Sequencing** – chemical sequencing
- **Sanger (didexoy) DNA Sequencing** – dideoxy sequencing
- **Next Generation DNA Sequencing (Applications)**
  - *Illumina* – bridging PCR / reversible dye terminator
  - *454 sequencing* – emulsion PCR / pyrosequencing

The Nobel Prize in Chemistry 1958

"for his work on the structure of proteins, especially that of insulin"

Frederick Sanger

United Kingdom

University of Cambridge
Cambridge, United Kingdom

b. 1918

The Nobel Prize in Chemistry 1980

"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"

"for their contributions concerning the determination of base sequences in nucleic acids"

Paul Berg

◑ 1/2 of the prize

USA

Stanford University
Stanford, CA, USA

b. 1926

Walter Gilbert

◔ 1/4 of the prize

USA

Harvard University,
Biological Laboratories
Cambridge, MA, USA

b. 1932

Frederick Sanger

◔ 1/4 of the prize

United Kingdom

MRC Laboratory of
Molecular Biology
Cambridge, United Kingdom

b. 1918

Maxam-Gilbert DNA Sequencing

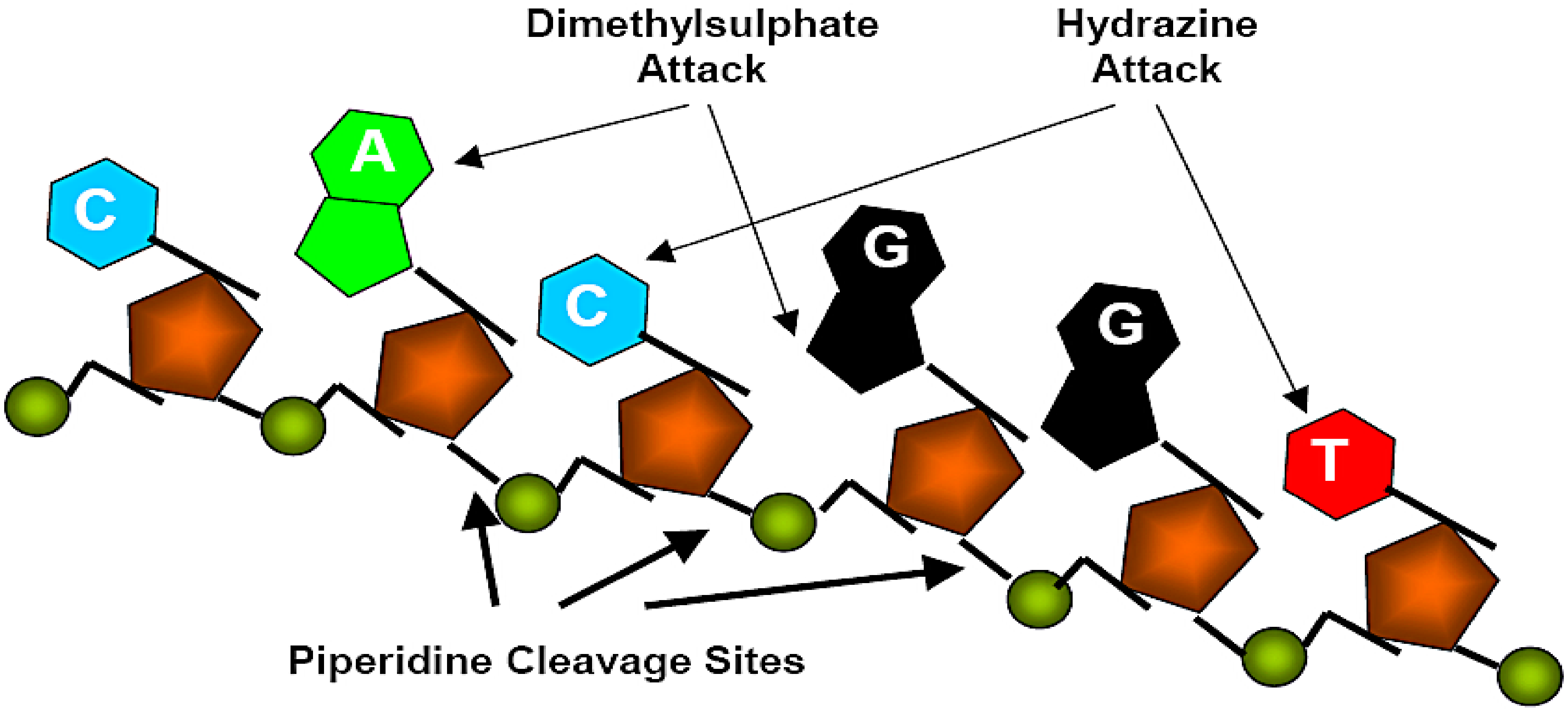


Figure 1. Chemical targets in the Maxam-Gilbert DNA sequencing strategy. Dimethylsulphate or hydrazine will attack the purine or pyrimidine rings respectively and piperidine will cleave the phosphate bond at the 3' carbon.

http://www.idtdna.com/support/technical/TechnicalBulletinPDF/DNA_Sequencing.pdf

INNOVATION & PRECISION IN NUCLEIC ACID SYNTHESIS

IDT®

INTEGRATED DNA TECHNOLOGIES

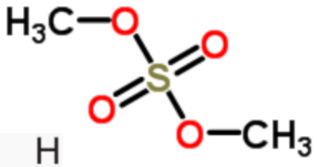
IDTutorial: DNA Sequencing

Allan Maxam / Walter Gilbert DNA Sequencing

Sequencing single-stranded DNA

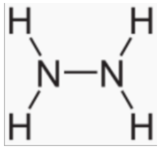
Two-step catalytic process:

1) Break glycoside bond between the ribose sugar and the base / displace base

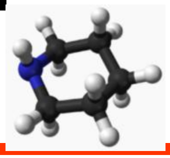


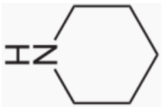
Purines react with dimethyl sulfate

Pyrimidines react with hydrazine (toxic)



2) Piperidine catalyzes phosphodiester bond cleavage where base displaced





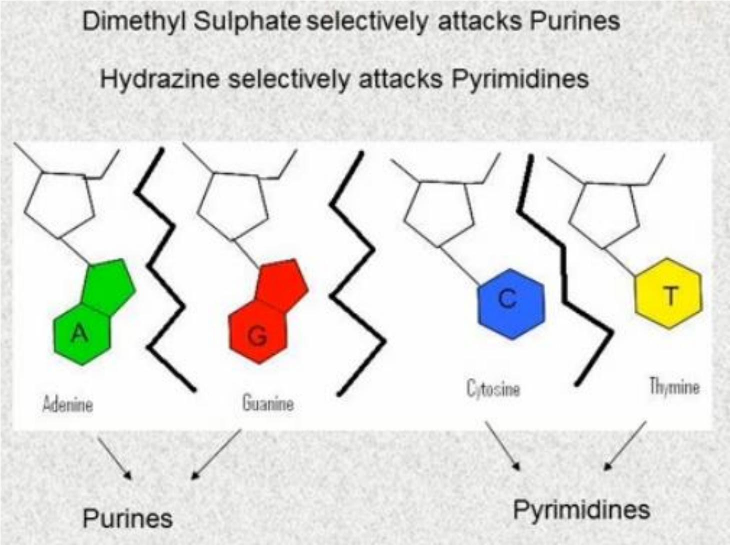
“G” - dimethyl sulfate and piperidine

“A + G” - dimethyl sulfate and piperidine in formic acid

“C” - hydrazine and piperidine in 1.5M NaCl

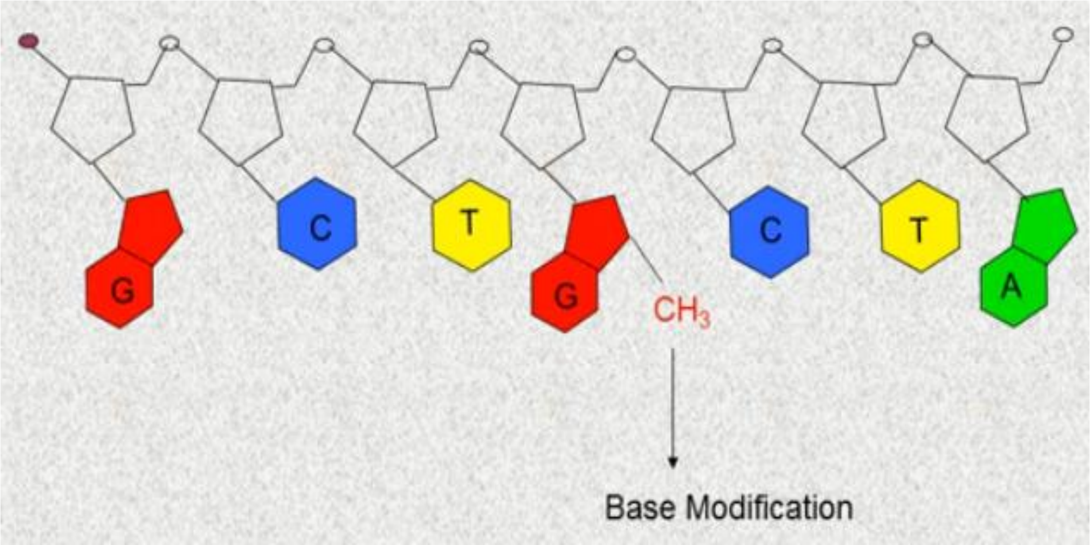
“C + T” - hydrazine and piperidine

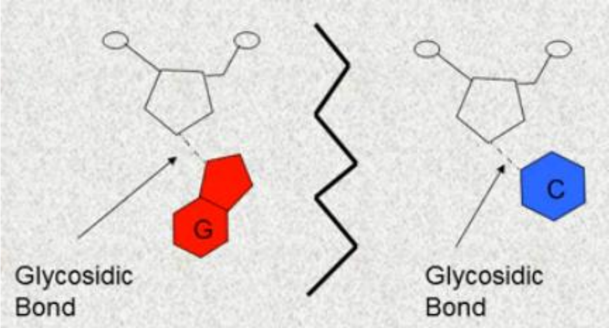
Maxam Gilbert Sequencing

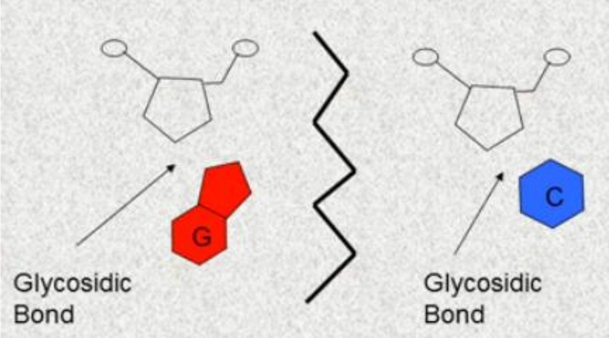


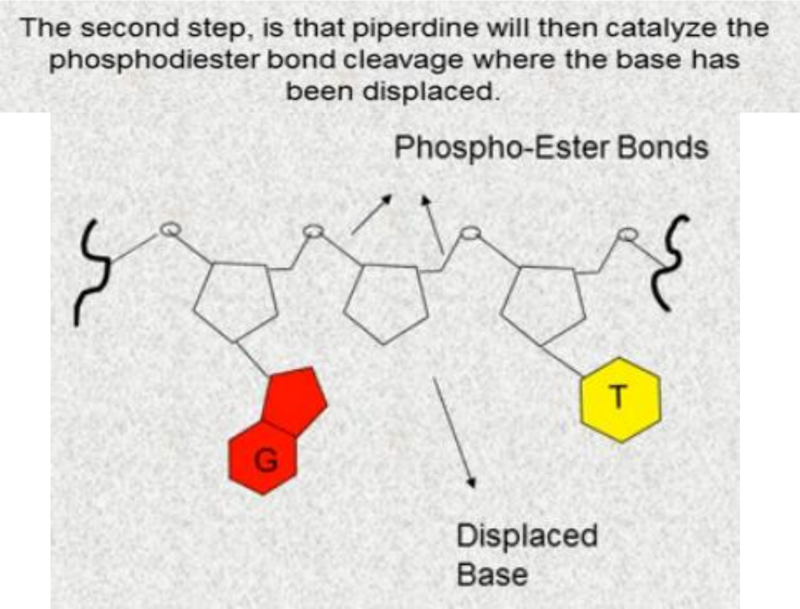
Dimethyl Sulphate selectively attacks Purines

Hydrazine selectively attacks Pyrimidines









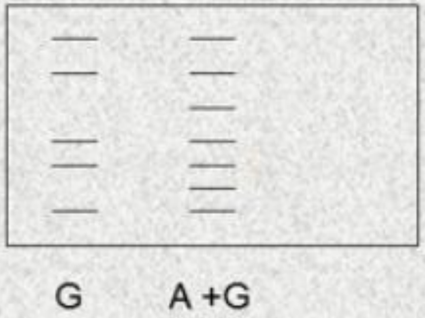
The second step, is that piperdine will then catalyze the phosphodiester bond cleavage where the base has been displaced.

Maxam Gilbert Sequencing by ChurchStreet105
http://www.youtube.com/watch?v=lqWZ-duHfu8&feature=related

Chemical Reagents and Conditions Employed For Maxam-Gilbert Sequencing.

Guanine – Dimethyl Sulphate followed by Piperdine

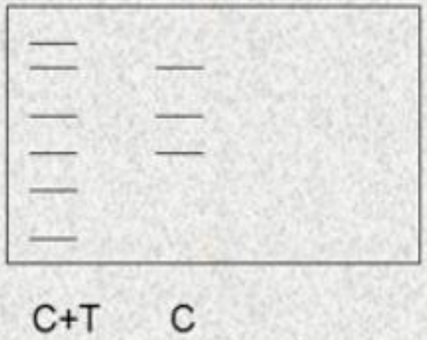
Guanine & Adenine – Dimethyl Sulphate in Formic Acid followed by Piperdine



Chemical Reagents and Conditions Employed For Maxam-Gilbert Sequencing.

Cytosine & Thymine – Hydrazine followed by perperdine

Cytosine – Hydrazine in 2M NaCl followed by perperdine



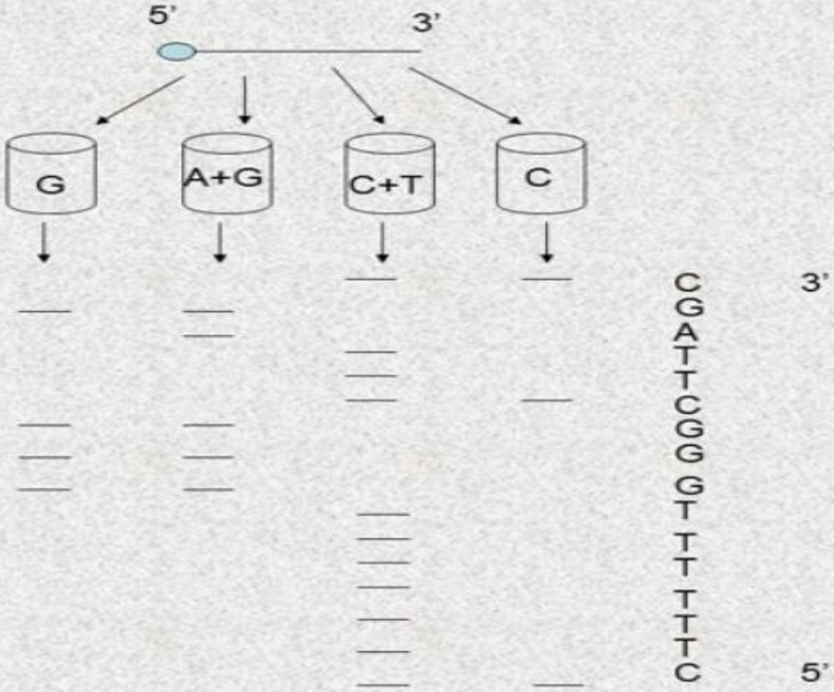
Sequenced Chain

^{32}PGpCpTpGpCpTpApGpGpTpGpCpCpGpApGpC

G G G G G G G

Cleaved Fragments

^{32}P
^{32}PGpCpTp
^{32}PGpCpTpGpCpTpAp
^{32}PGpCpTpGpCpTpApGp
^{32}PGpCpTpGpCpTpApGpGpTp
^{32}PGpCpTpGpCpTpApGpGpTpGpCpCp
^{32}PGpCpTpGpCpTpApGpGpTpGpCpCpGpAp
^{32}PGpCpTpGpCpTpApGpGpTpGpCpCpGpApGpC



Proc. Natl. Acad. Sci. USA
Vol. 74, No. 2, pp. 560–564, February 1977
Biochemistry

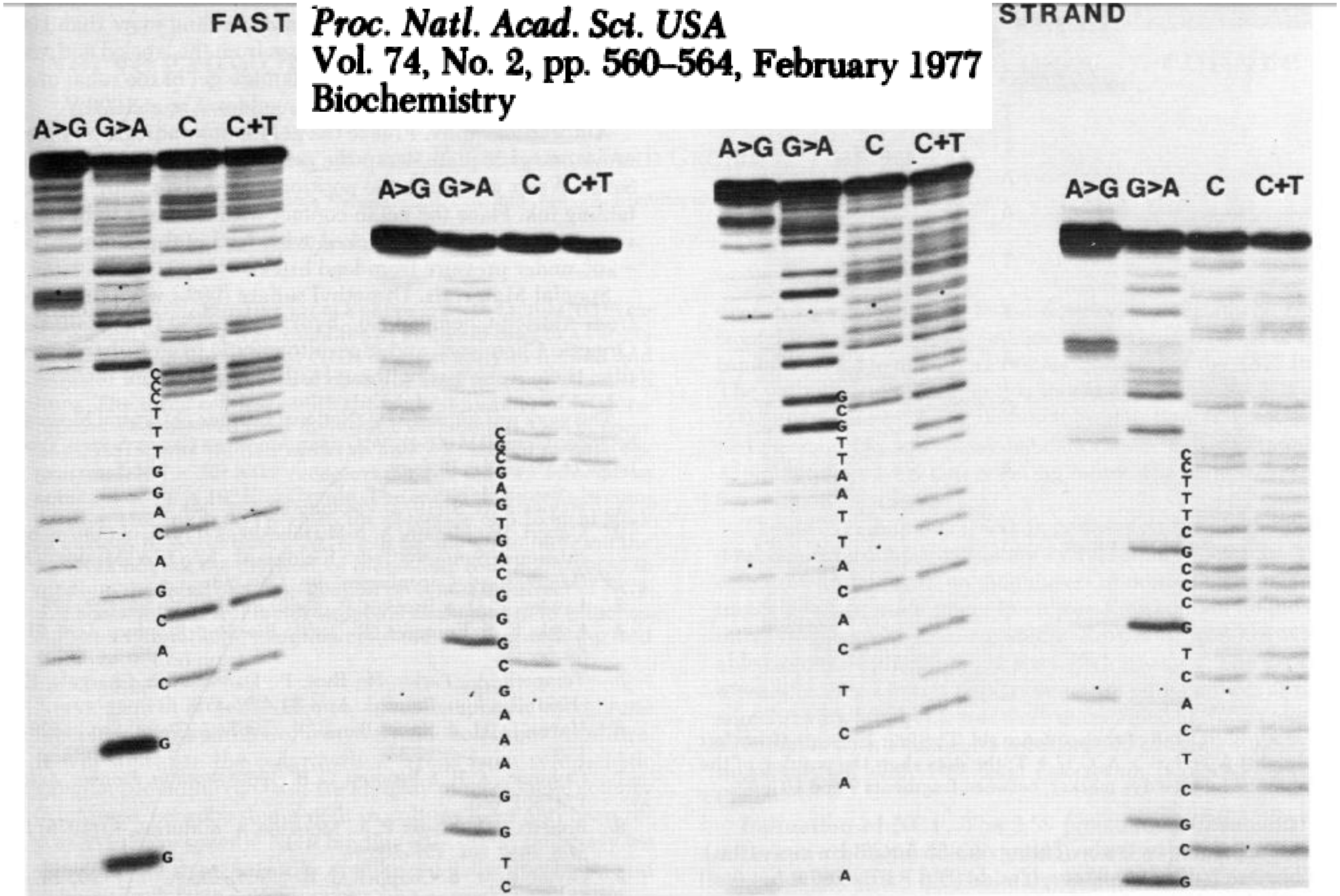


FIG. 2. Autoradiograph of a sequencing gel of the complementary strands of a 64-base-pair DNA fragment. Two panels, each with four reactions, are shown for each strand; cleavages proximal to the 5′ end are at the bottom on the left. A strong band in the first column with a weaker band in the second arises from an A; a strong band in the second column with a weaker band in the first is a G; a band appearing in both the third and fourth columns is a C; and a band only in the fourth column is a T. To derive the sequence of each strand, begin at the bottom of the left panel and read upward until the bands are not resolved; then, pick up the pattern at the bottom of the right panel and continue upward. One-tenth of each strand, isolated from the gel of Fig. 1, was used for each of the base-modification reactions. The dimethyl sulfate treatment was 50 mM for 30 min to react with A and G; hydrazine treatment was 18 M for 30 min to react with C and T and 18 M with 2 M NaCl for 40 min to cleave C. After strand breakage, half of the products from the four reactions were layered on a 1.5 × 330 × 400 mm denaturing 20% polyacrylamide slab gel, pre-electrophoresed at 1000 V for 2 hr. Electrophoresis at 20 W (constant power), 800 V (average), and 25 mA (average) proceeded until the xylene cyanol dye had migrated halfway down the gel. Then the rest of the samples were layered and electrophoresis was continued until the new bromphenol blue dye moved halfway down. Autoradiography of the gel for 8 hr produced the pattern shown.

Maxam-Gilbert DNA Sequencing

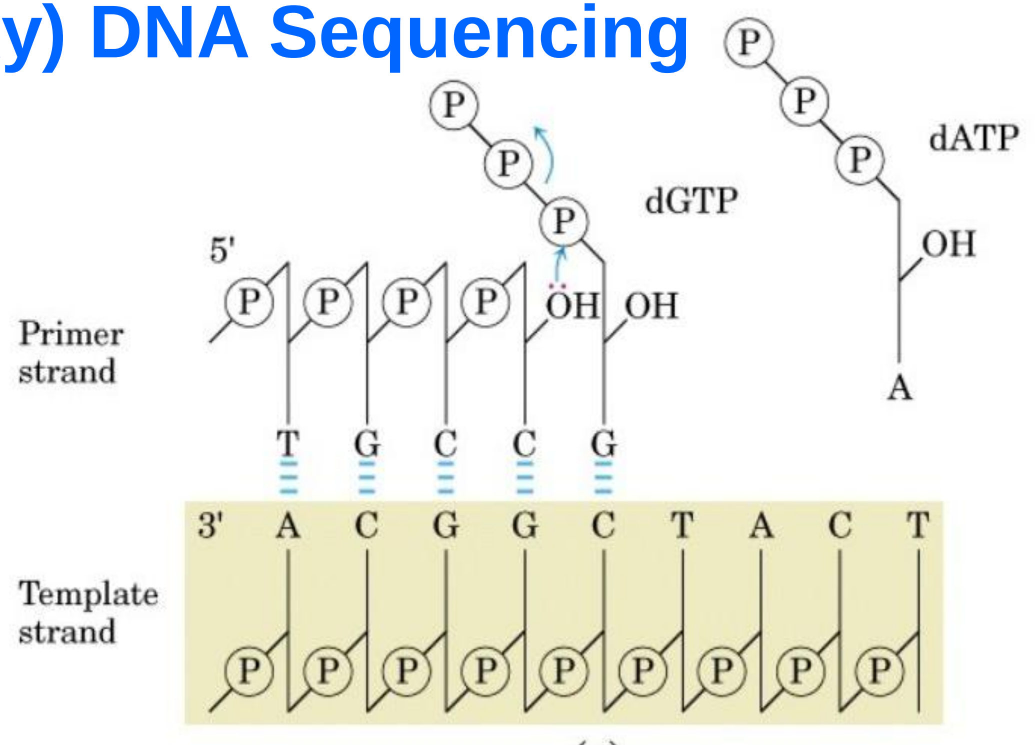
- 200-300 bases of DNA sequence every few days
- Use large amounts of radioactive material, ^{35}S or ^{32}P
- Constantly pouring large, paper thin acrylamide gels
- Hydrazine is a neurotoxin

Early Benefits -

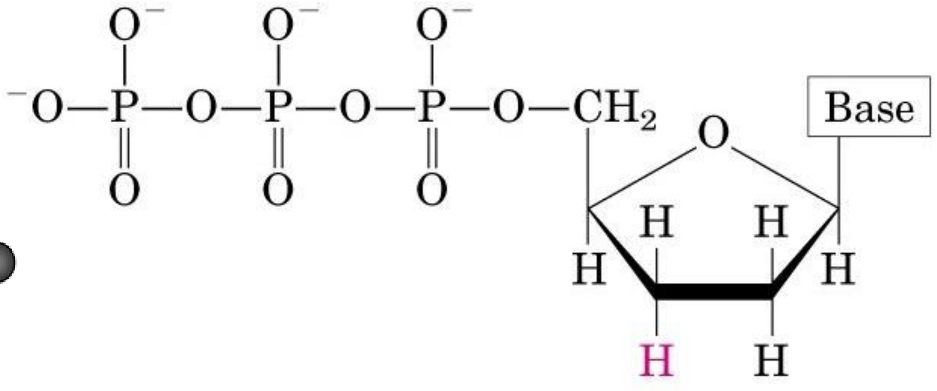
Discovery that the gene for ovalbumin in chicken and the gene encoding β-globin in rabbit contained non-coding gaps in the coding regions. These gaps<were flanked by the same dinucleotides in the two genes; GT on the 5' end of the gaps and AG on the 3' end of the gaps. Soon, the terms intron and exon were added to the genetic lexicon to describe the coding and non-coding regions of eukaryotic genes (1977).

Fred Sanger (dideoxy) DNA Sequencing

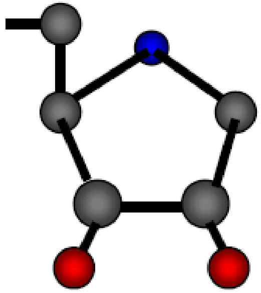
Sanger knew that, whenever a dideoxynucleotide was incorporated into a polynucleotide, the chain would irreversibly stop, or terminate. Thus, the incorporation of specific dideoxynucleotides in vitro would result in selective chain termination.



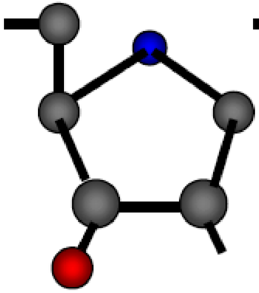
(a)



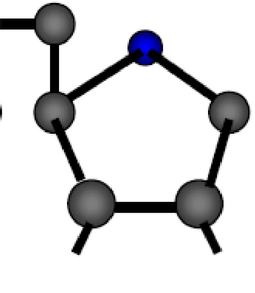
ddNTP analog



Ribose

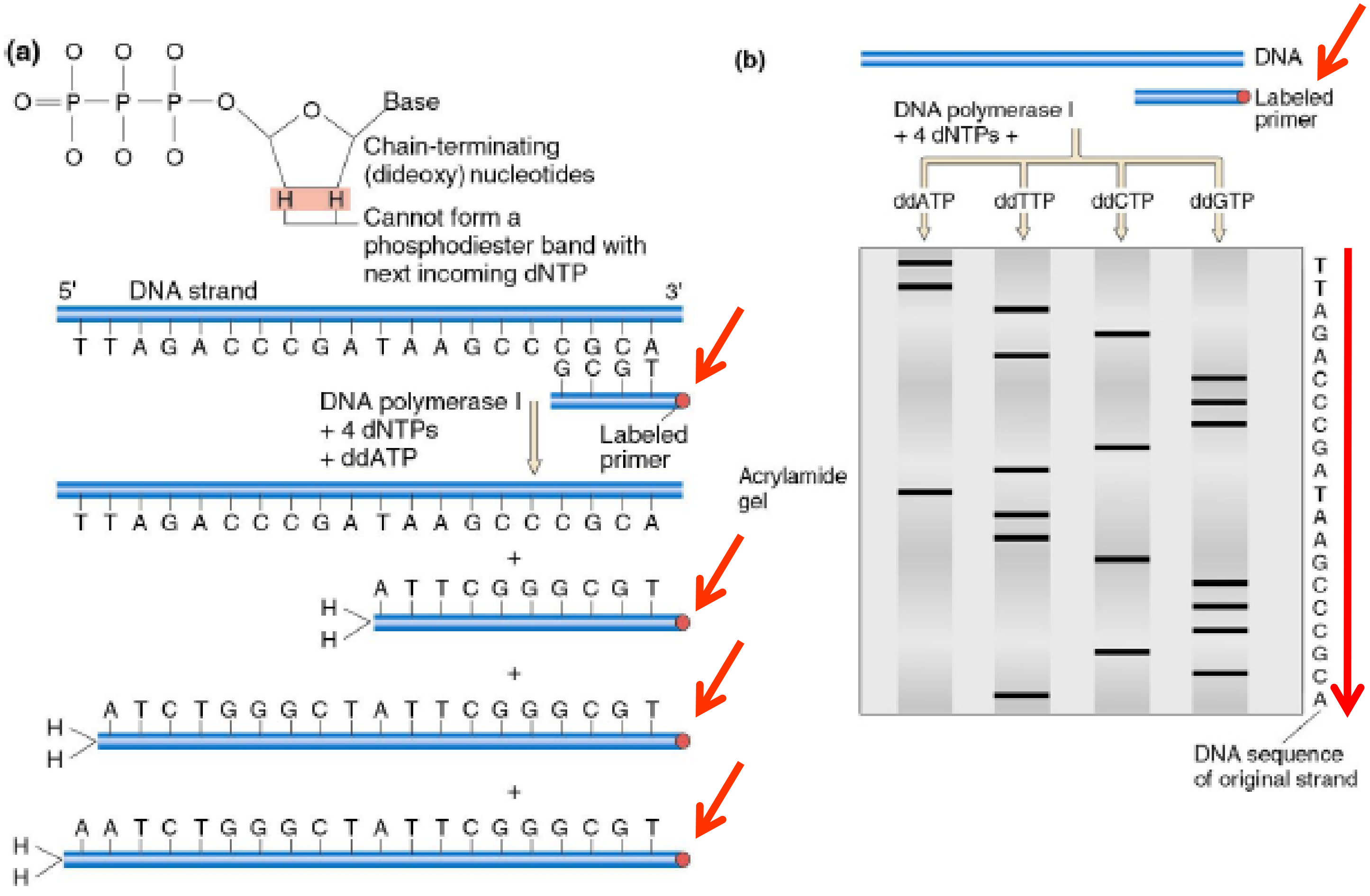


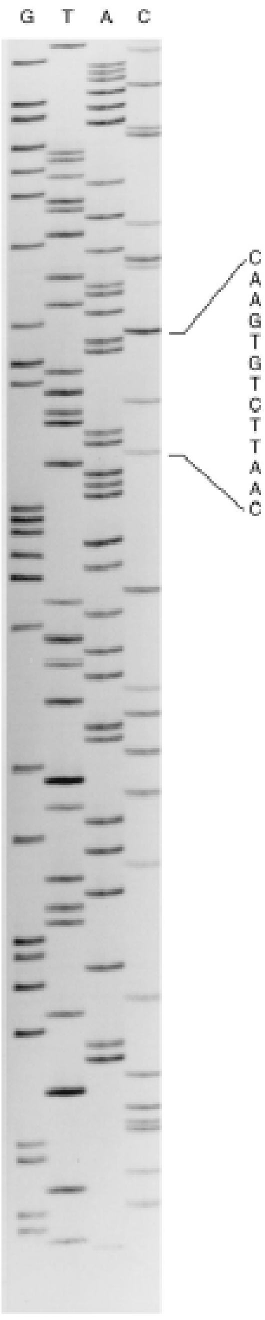
Deoxyribose



Dideoxyribose

Sanger (dideoxy) DNA Sequencing

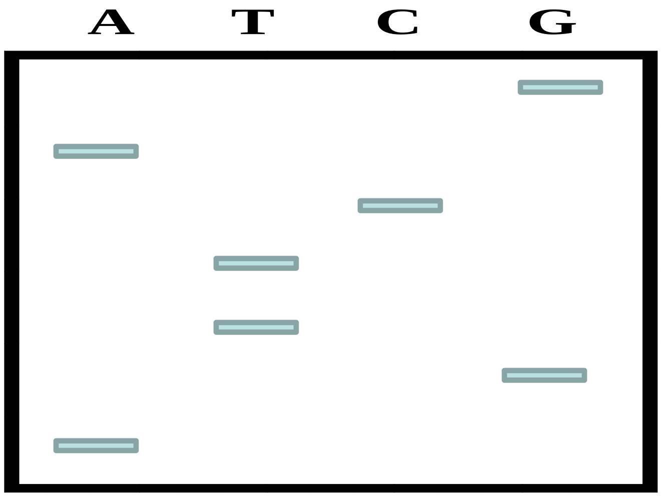

(a)
O O O
O=P—P—P—O
O O O
O
Base
Chain-terminating (dideoxy) nucleotides
H H
Cannot form a phosphodiester band with next incoming dNTP
5' DNA strand 3'
T T A G A C C C G A T A A G C C C G C A
G C G T
DNA polymerase I
+ 4 dNTPs
+ ddATP
Labeled primer
T T A G A C C C G A T A A G C C C G C A
+
A T T C G G G C G T
H
H
+
A T C T G G G C T A T T C G G G C G T
H
H
+
A A T C T G G G C T A T T C G G G C G T
H
H
(b)
DNA
Labeled primer
DNA polymerase I
+ 4 dNTPs +
ddATP
ddTTP
ddCTP
ddGTP
Acrylamide gel
T T A G A C C C G A T A A G C C C G C A
DNA sequence of original strand



Consider the following nucleic acid sequencing gel experiment using the Sanger dideoxy sequencing method:

What is the expected sequence (5’ -> 3’) of the original DNA sample assuming the primer was labeled with a 5’-prime fluorescent label?

[DNA polymerase I + 4 dNTPs + ddATP ddTTP ddCTP ddGTP]



iClicker Question #2:

A) GACTTGA
B) CTGAACT
C) AGTTCGA
D) TCAAGTC
E) None of the above

Advantages of dideoxy DNA Sequencing

- Elimination of dangerous chemicals (hydrazine)
- Greater efficiency

Taq polymerase makes DNA strands off of a template at rate of about 500 bases per minute

Chemical synthesis of a 25-mer oligonucleotide takes more than two hours.

→ High Throughput Methods (Human Genome Project)

Automated Fluorescence Sequencing

In 1986, Leroy Hood and colleagues reported on a DNA sequencing method in which the radioactive labels, autoradiography, and manual base calling were all replaced by fluorescent labels, laser induced fluorescence detection, and computerized base calling.

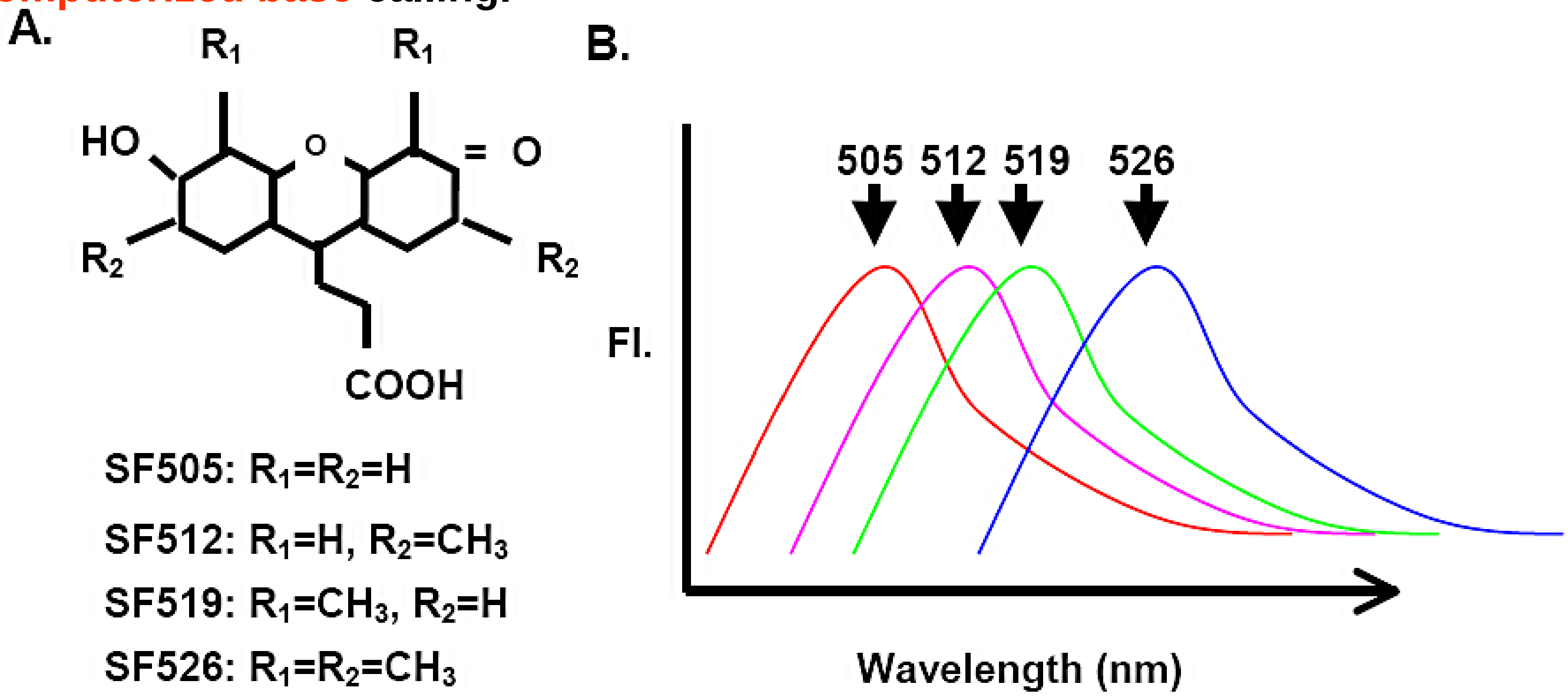
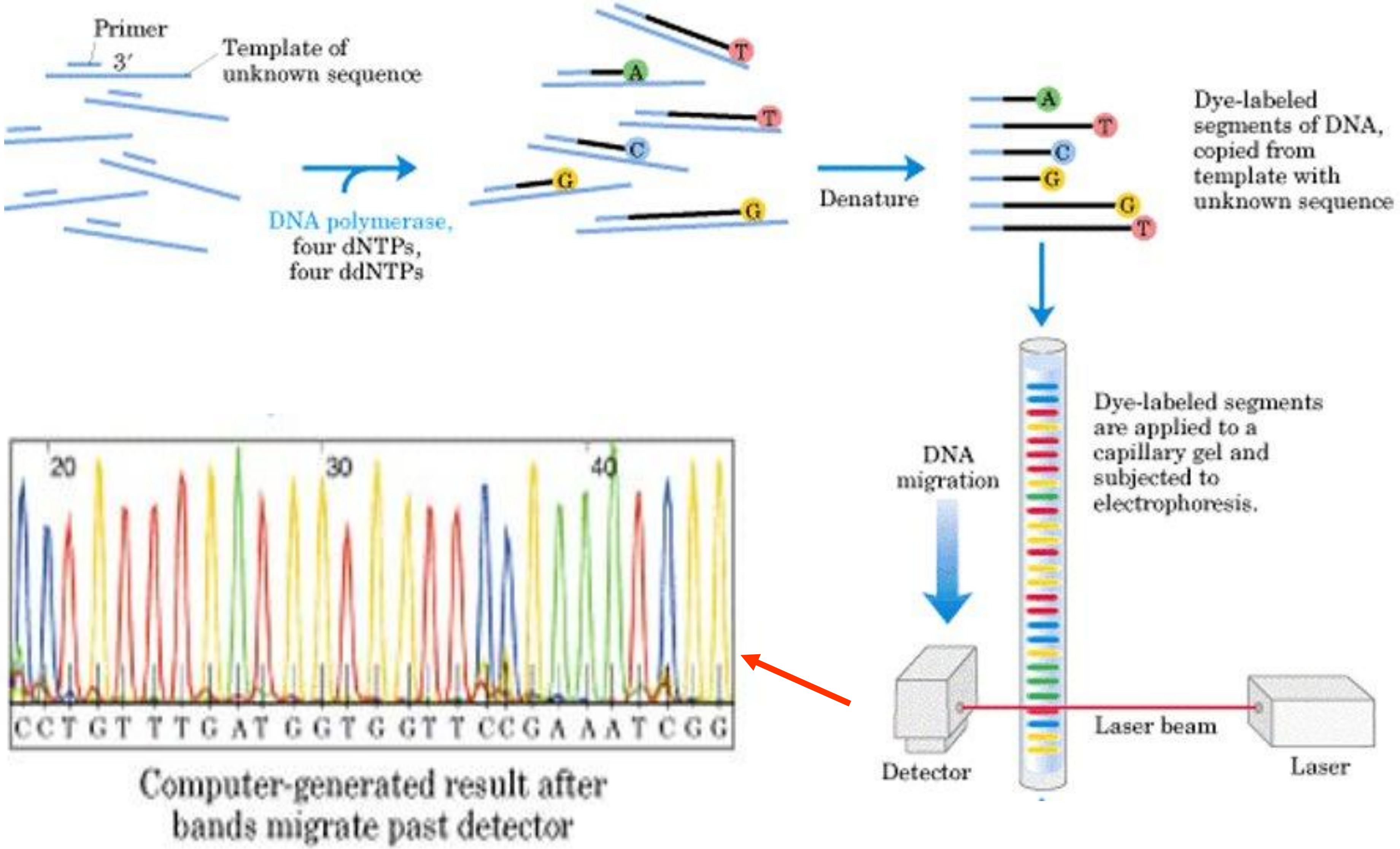


Figure 5. A. Chemical structure of the four succinylfluorescein dyes developed at DuPont. B. Normalized fluorescence emission spectra for each of the four dyes following excitation at 488nm. Shifts in the spectra were achieved by changing the side groups R_1 and R_2.

Automated DNA sequencing


Primer
3′
Template of unknown sequence
DNA polymerase, four dNTPs, four ddNTPs
T
A
T
C
G
G
Denature
A
T
C
G
G
T
Dye-labeled segments of DNA, copied from template with unknown sequence
Dye-labeled segments are applied to a capillary gel and subjected to electrophoresis.
DNA migration
Detector
Laser beam
Laser
20
30
40
C C T G T T T G A T G G T G G T T C C G A A A T C G G

Computer-generated result after bands migrate past detector

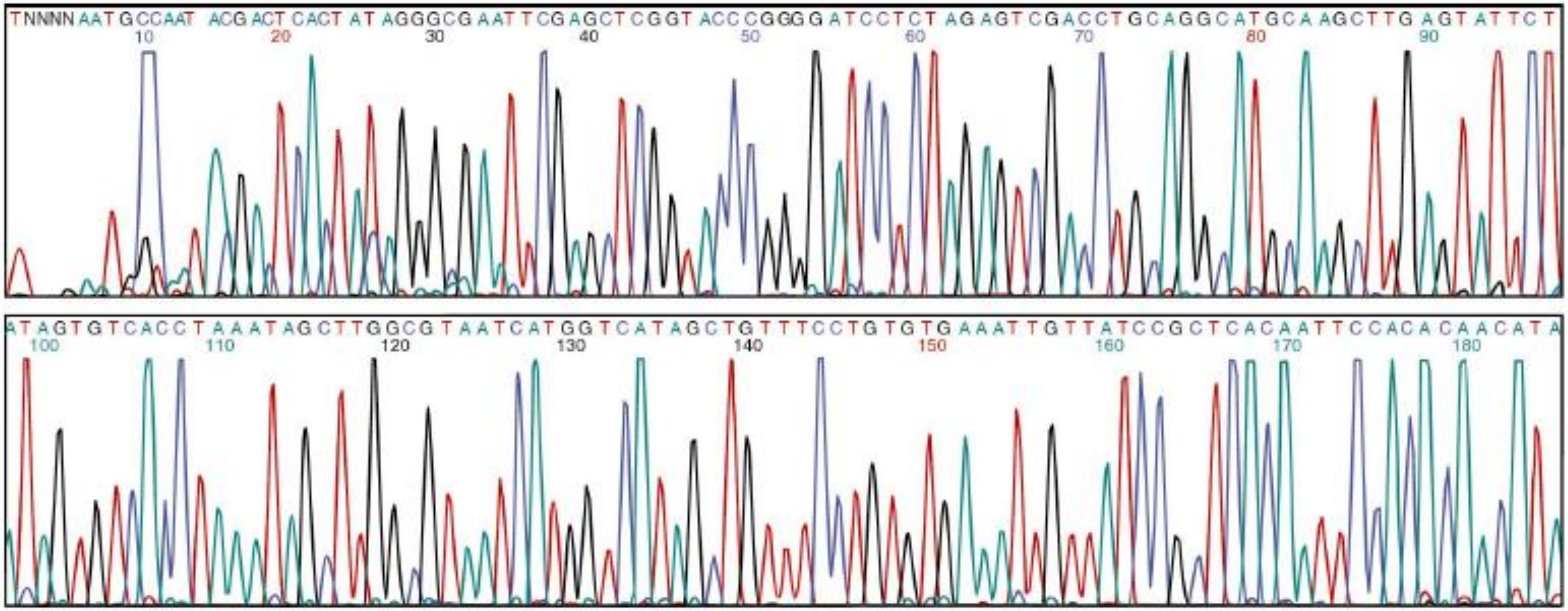
Automated dye-terminator sequencing

4-fluorescently labelled dideoxy dye terminators

ddATP
ddGTP
ddCTP
ddTTP

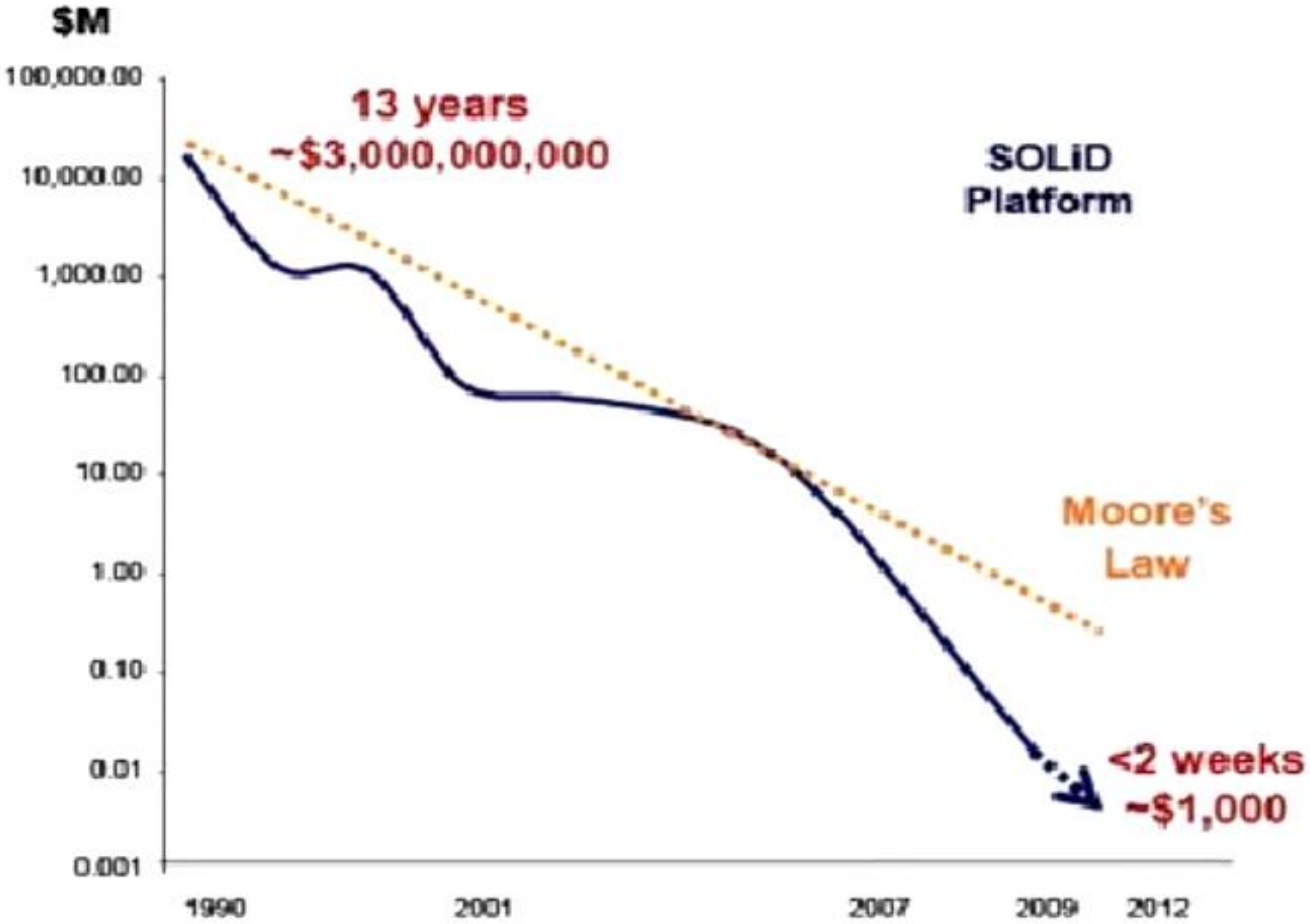
} pool and load in a single well or capillary

- scan with laser + detector specific for each dye
- automated base calling
- very long reads (~ 1000 bases)/run



[DNA dideoxysequencing animation]

Cost per Human Genome

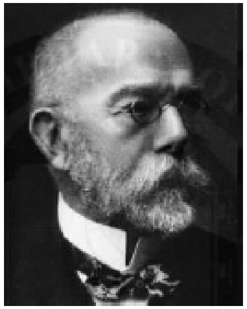


http://blogs.forbes.com/sciencebiz/2010/06/03/your-genome-is-coming/

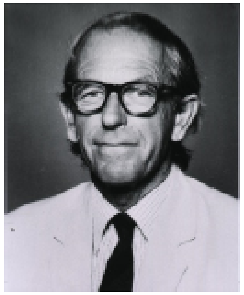
Next-Generation Sequencers

		Read Length	Gb/run	Technology
illumina	GA_{IIx}	2 x 100+ bp	20+ Gb	• Bridge amplification • Reversible terminators
454 SEQUENCING	GS FLX Titanium	1 x 400-600bp 2 x 140-200bp	0.4-0.6 Gb	• Emulsion PCR amplification • Homopolymers detected by an increase in signal proportional to length
AB applied biosystems	SOLiD 3	2 x 50bp	20+ Gb	• Emulsion PCR amplification • Ligation-based sequencing • Alignment in color space
Helicos BioSciences Corporation	Single Molecule Sequencer	2 x 25-55bp	21-28 Gb	• No amplification • Single molecule sequencing

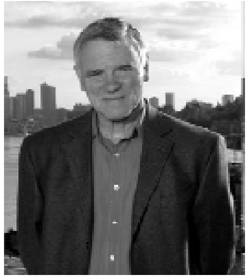
Sanger sequencing


Robert Koch

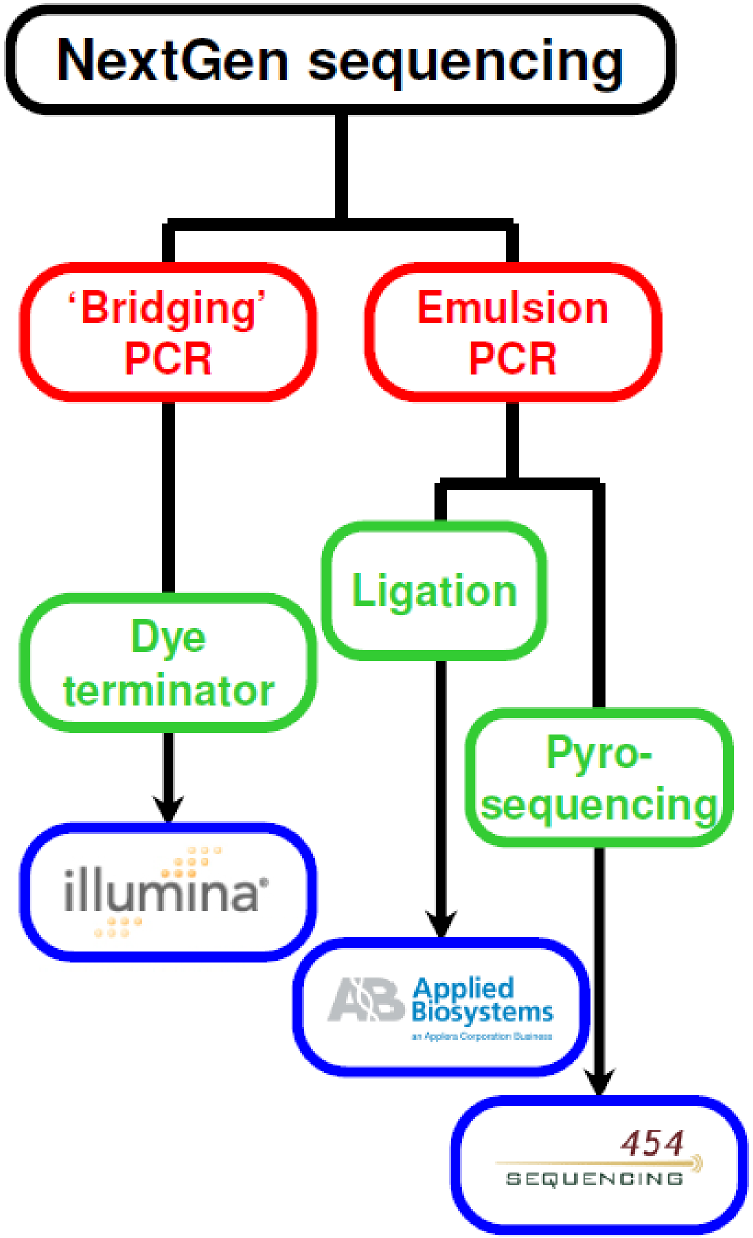
Clonal population


Frederick Sanger

Chemistry


Leroy Hood

Automation


NextGen sequencing

'Bridging' PCR

Emulsion PCR

Dye terminator

Ligation

Pyro-sequencing

illumina

Applied Biosystems

an Applera Corporation Business

454 SEQUENCING

Next Generation Sequencing Capabilities

Next Generation Sequencing | Sanger Sequencing | Sequencing Platforms

Illumina* GAIIx and HiSeq* 2000
The Illumina GAIIx and HiSeq 2000 platforms utilize reversible terminator-based sequencing by synthesis chemistry to deliver the highest sequencing output and fastest data generation rate of the next generation technologies. Ideal for whole genome re-sequencing, targeted resequencing, de novo sequencing and transcriptome sequencing.

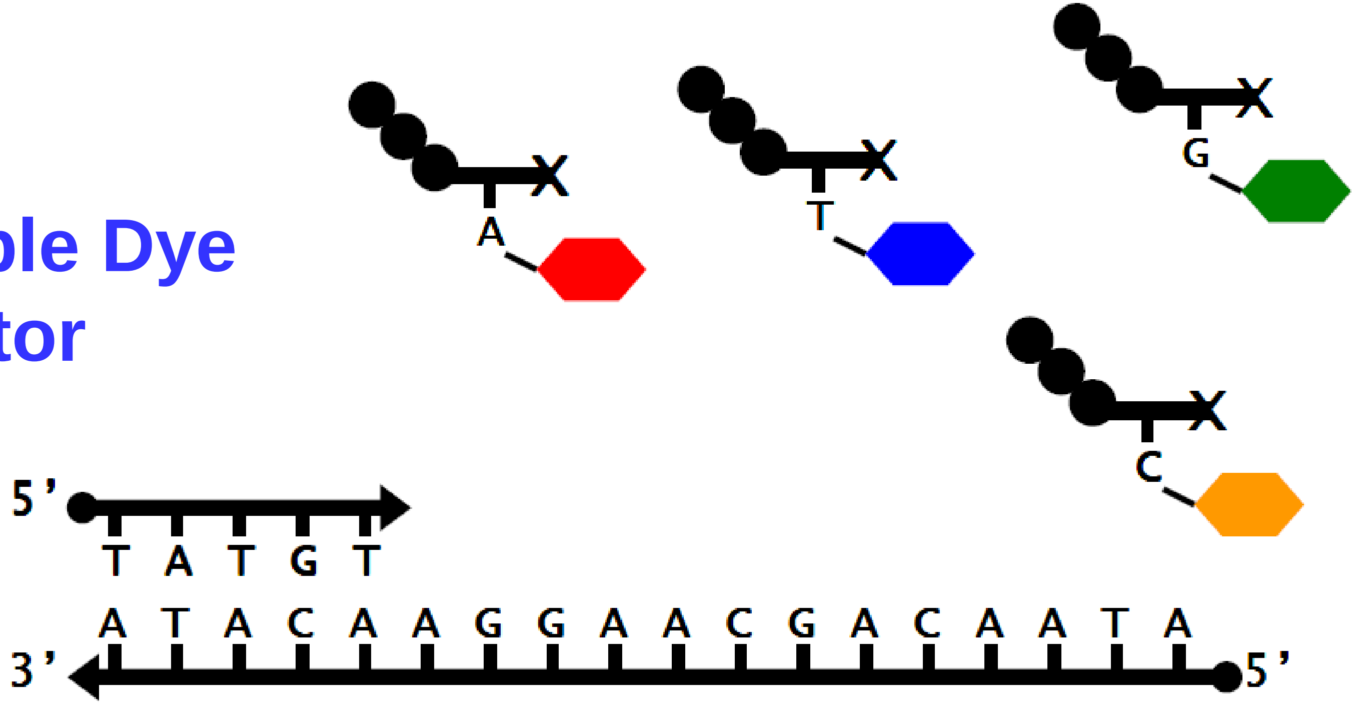
Roche* 454* GS FLX* Titanium
The Roche 454 GS FLX utilizes massively parallel pyrosequencing of bead bound templates and chemiluminescent base calling to yield long read lengths. Ideal for de novo sequencing, metagenomics and transcriptome sequencing.

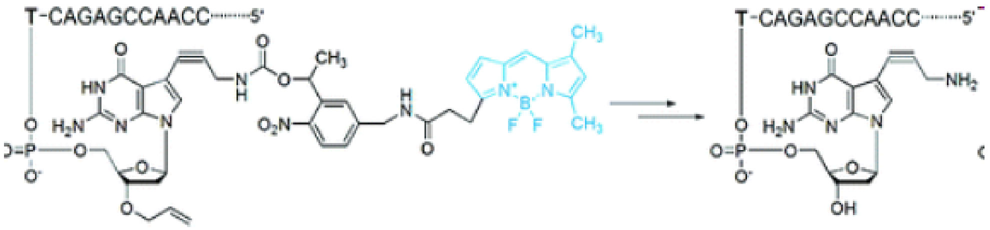
ABI* 3730XL
The ABI 3730XL Sanger sequencing platform utilizes capillary electrophoresis to generate sequences with read lengths up to 1200 bp and pass rates over 90%. Ideal for single sample sequencing, primer walking, shotgun sequencing and SNP detection.

Credit: Illumina

Illumina/Solexa method: Sequencing by synthesis

Reversible Dye Terminator





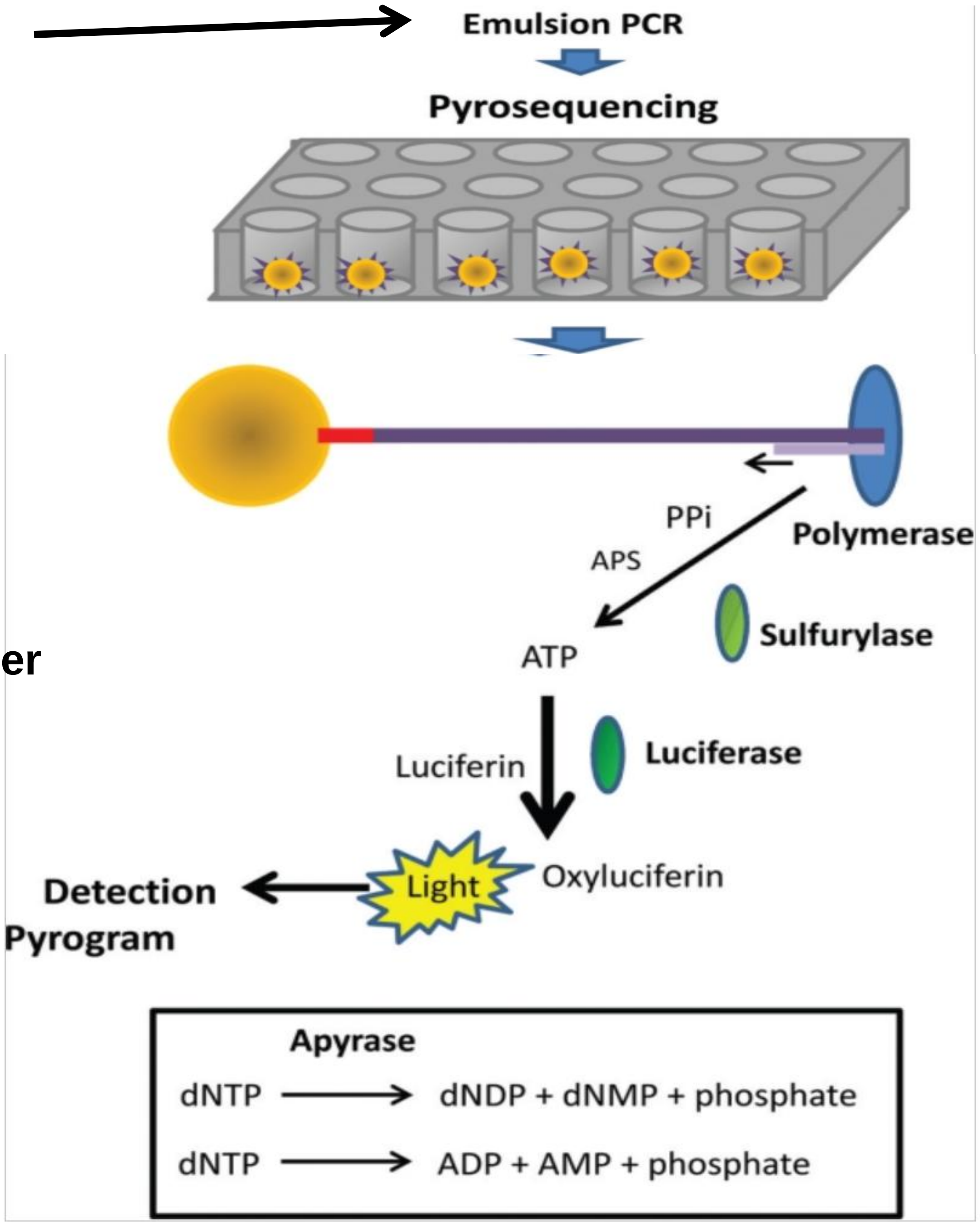
One example of reversible dye terminator (might not be Solexa's)

Preparation of DNA fragments

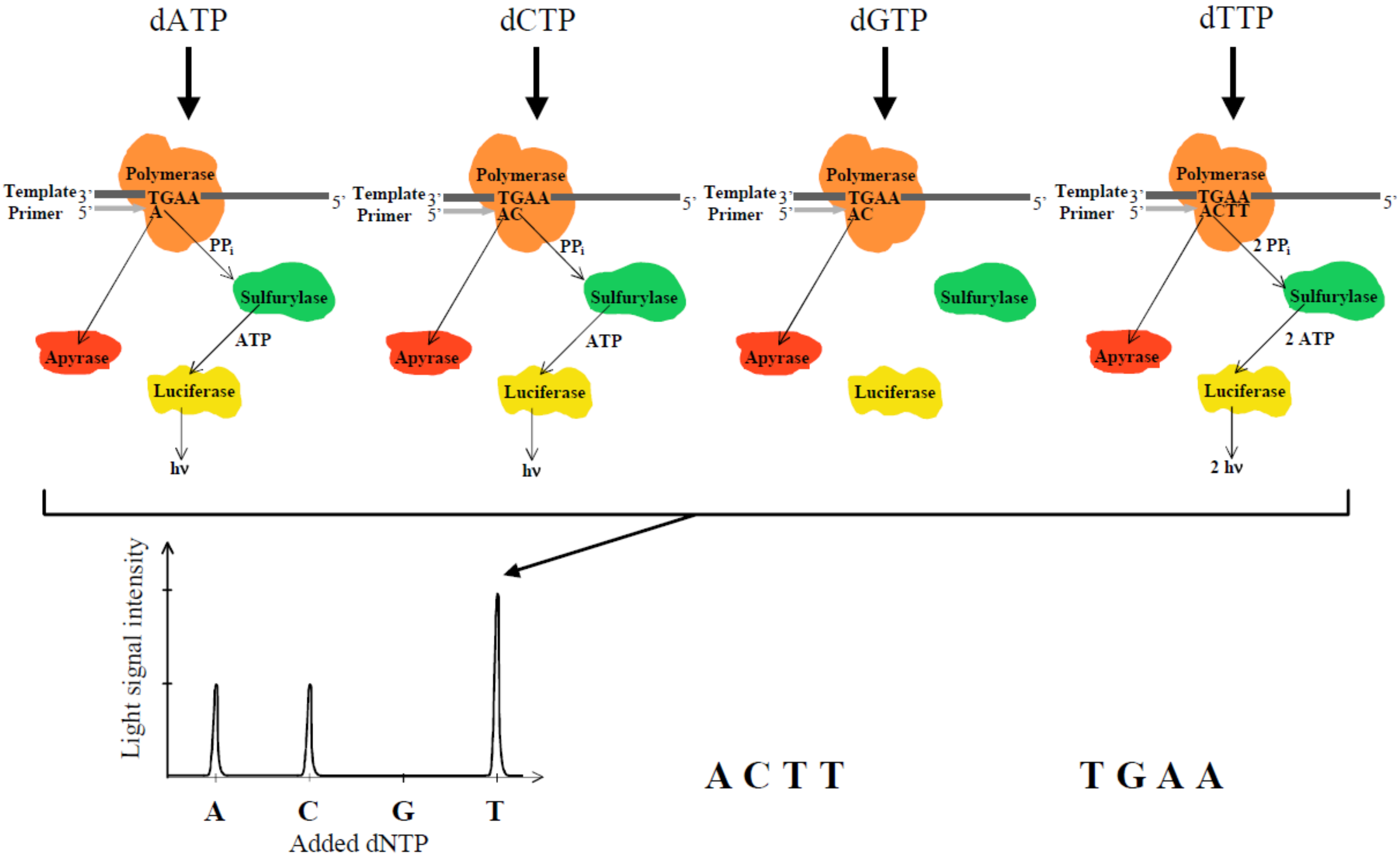
454 pyrosequencing

J Oral Microbiol. 2012; 4: 10.3402/jom.v4i0.10743.
Published online 2012 January 23. doi:

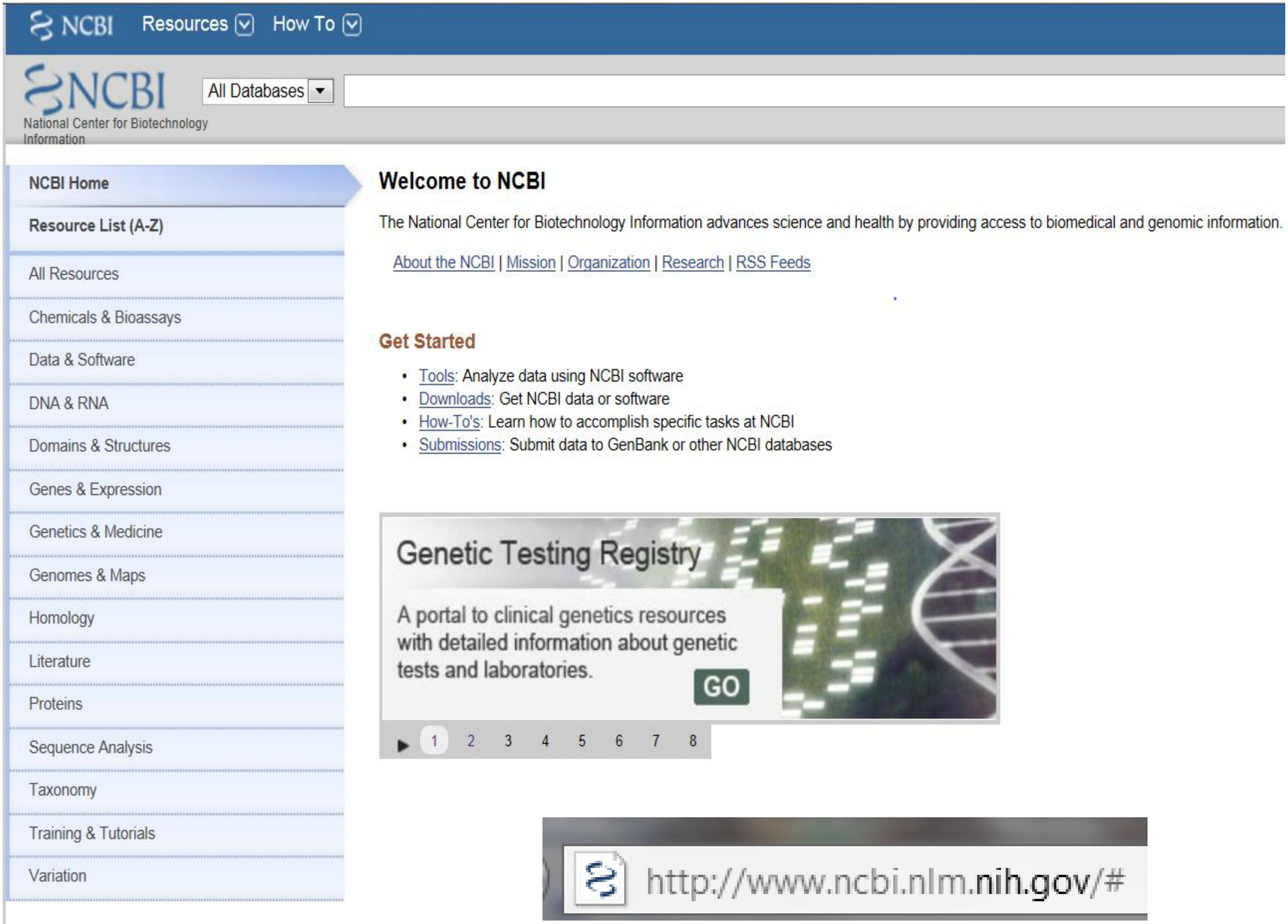
Pyrosequencing as a tool for better understanding of human microbiomes



The 454 pyrosequencing approach.



Pyrogram → Sequence of synthesised DNA → Sequence of template DNA



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

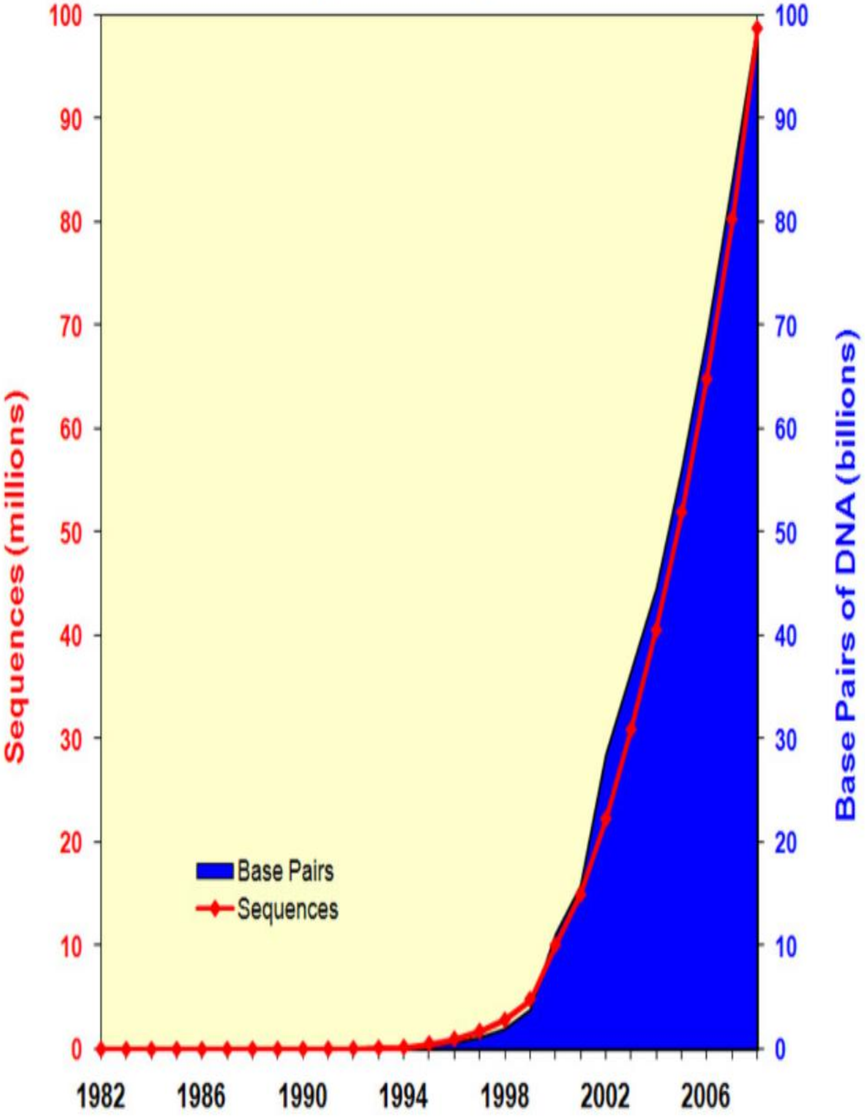
Genetic Testing Registry

A portal to clinical genetics resources with detailed information about genetic tests and laboratories.

http://www.ncbi.nlm.nih.gov/#

Growth of GenBank

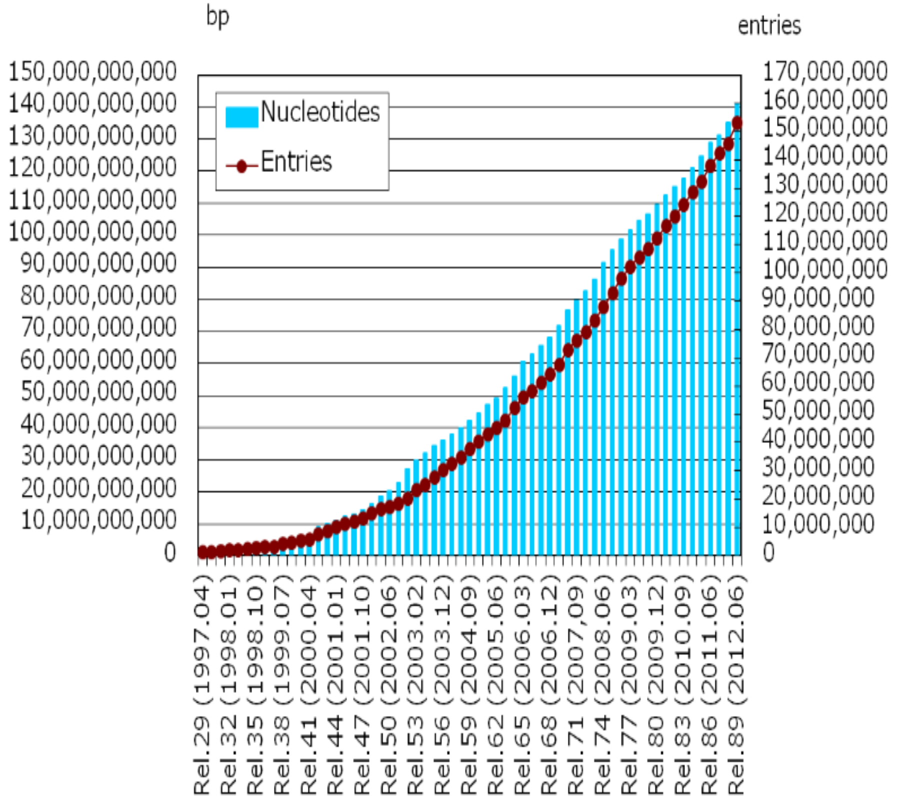
(1982 - 2008)



DDBJ Release Note

** Note : CON and TPA divisions are not counted in the following Release statistic.*

DDBJ/EMBL/GenBank database growth



Note: CON division is not counted in statistics of DDBJ

Nucleic Acids: Methods

(CSI / Law and Order / Forensic Files / House / Crossing Jordan / Quincy, M.E.)

Topics:

1. PCR – Polymerase Chain Reaction
2. Human Genome Project / Genomics
3. Use of DNA Microarrays

The Nobel Prize in Chemistry 1993

"for contributions to the developments of methods within DNA-based chemistry"

"for his invention of the polymerase chain reaction (PCR) method"

"for his fundamental contributions to the establishment of oligonucleotide-based, site-directed mutagenesis and its development for protein studies"

Kary B. Mullis

◐ 1/2 of the prize

USA

La Jolla, CA, USA

b. 1944

Michael Smith

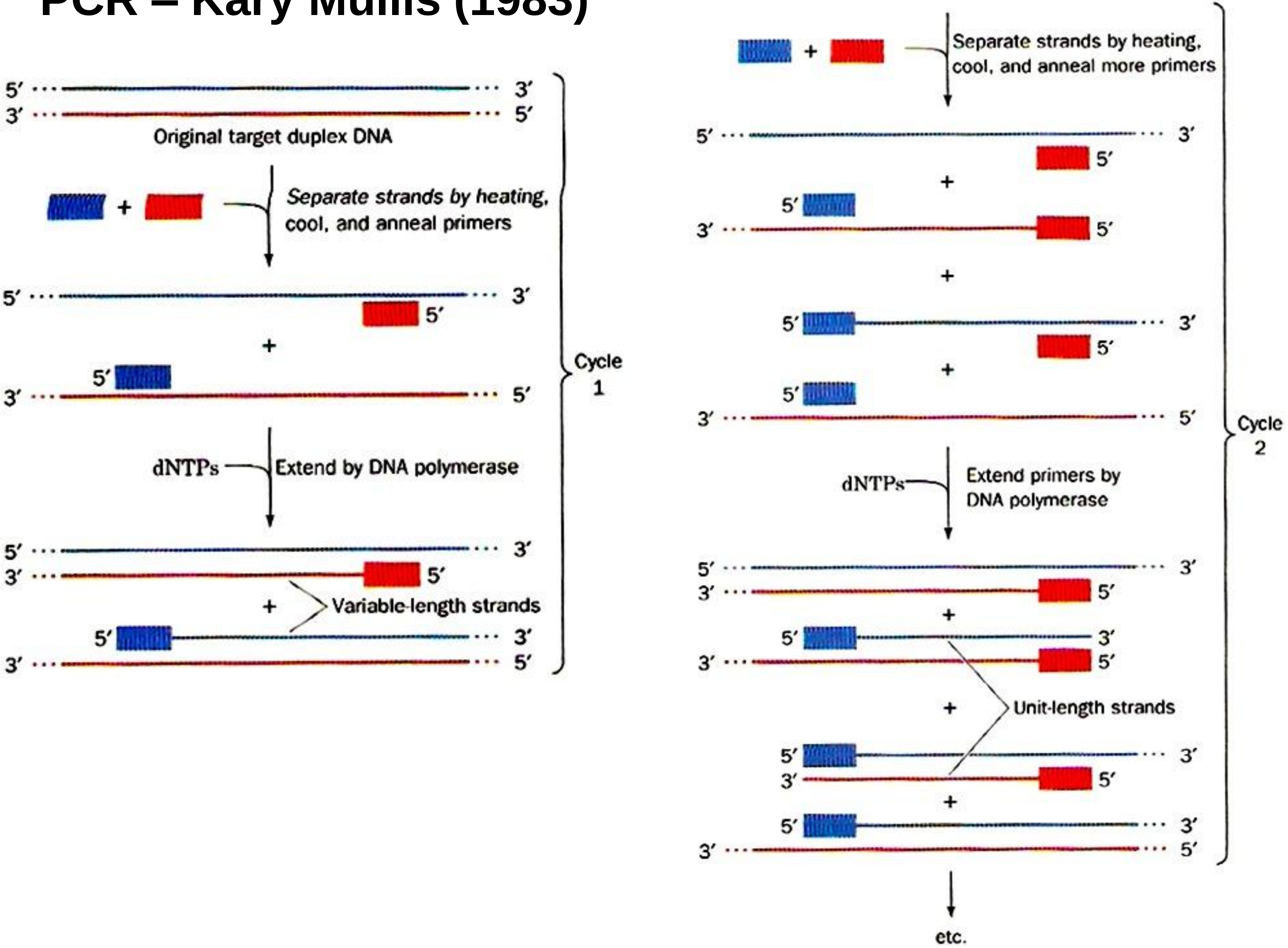
◐ 1/2 of the prize

Canada

University of British Columbia
Vancouver, Canada

b. 1932
(in Blackpool, United Kingdom)
d. 2000

PCR – Kary Mullis (1983)



Original target duplex DNA

Separate strands by heating, cool, and anneal primers

dNTPs — Extend by DNA polymerase

Variable-length strands

Cycle 1

Separate strands by heating, cool, and anneal more primers

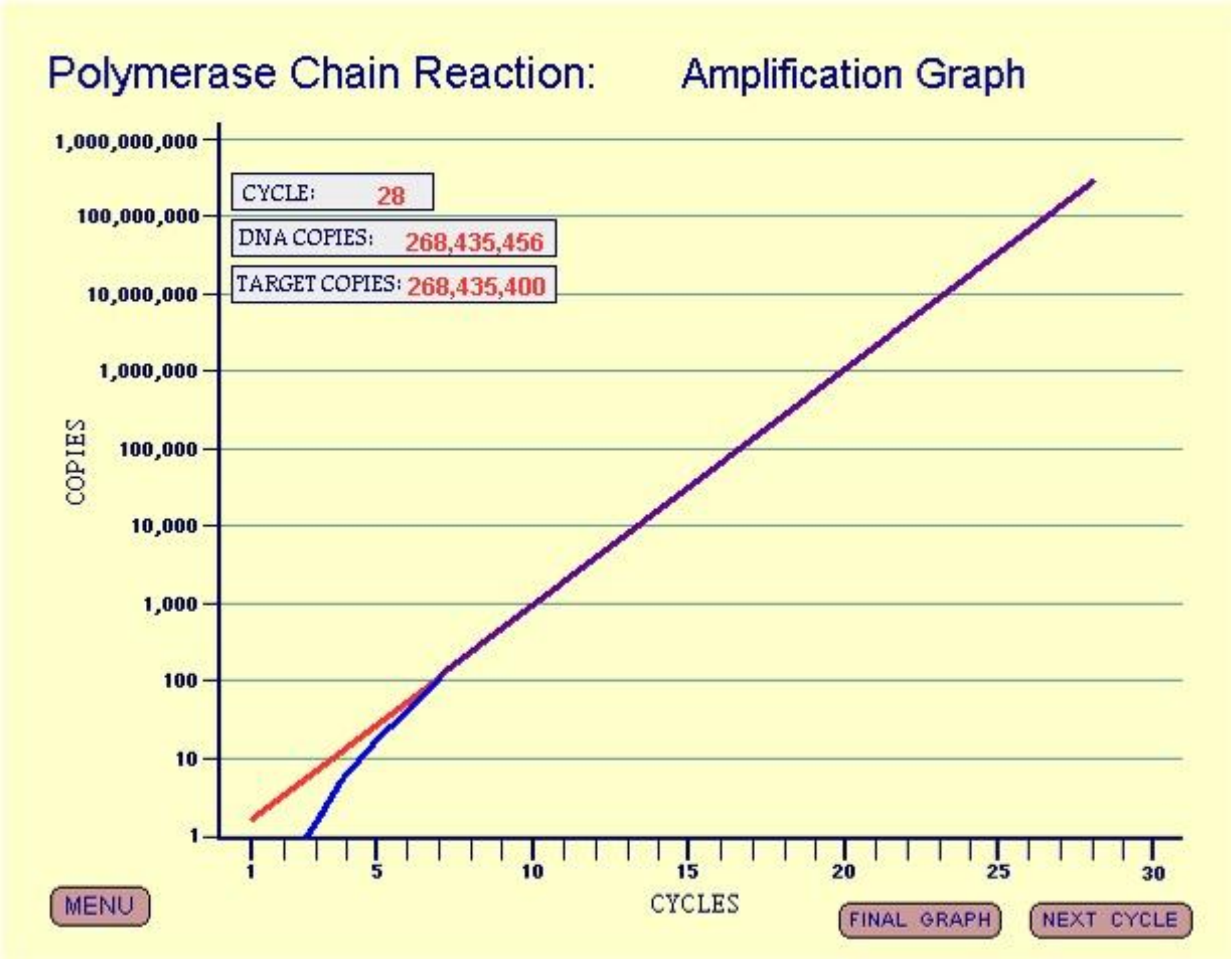
dNTPs — Extend primers by DNA polymerase

Unit-length strands

Cycle 2

etc.

http://www.dnalc.org/ddnalc/resources/animations.html



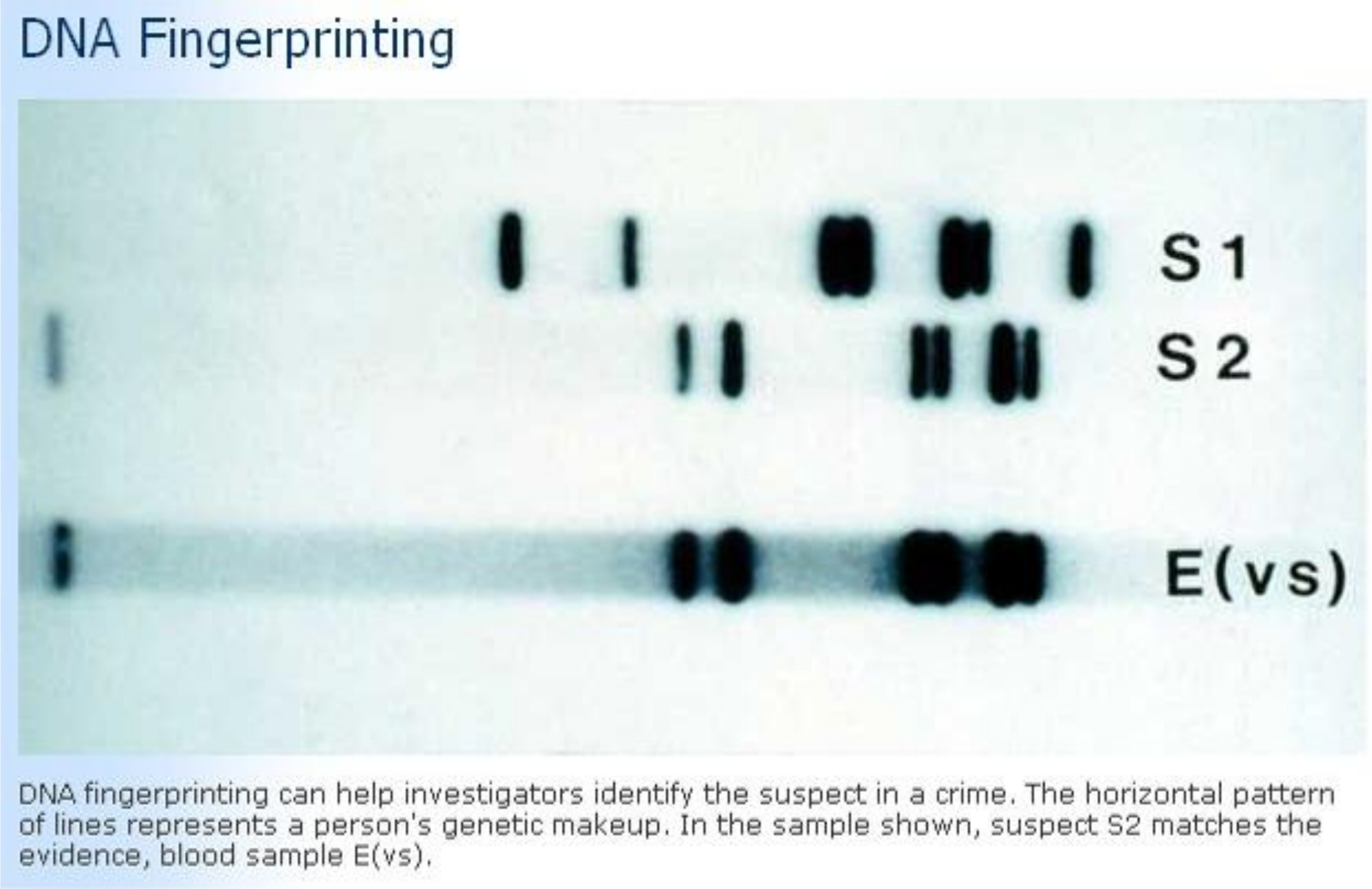
Polymerase Chain Reaction: Amplification Graph

CYCLE: 28

DNA COPIES: 268,435,456

TARGET COPIES: 268,435,400

Sir Alec Jeffreys - 1984



DNA Fingerprinting

DNA fingerprinting can help investigators identify the suspect in a crime. The horizontal pattern of lines represents a person's genetic makeup. In the sample shown, suspect S2 matches the evidence, blood sample E(vs).

Human Genome Project

Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003.

Project goals:

- ***identify*** **all the approximately 20,000-30,000 genes in human DNA,**
- ***determine*** **the sequences of ~3 billion chemical base pairs of human DNA,**
- ***store*** **this information in databases,**
- ***improve*** **tools for data analysis,**
- ***transfer*** **related technologies to the private sector, and**
- ***address*** **the ethical, legal, and social issues (ELSI) from the project.**
- ***sequence 500 Mb/year at < $0.25 per finished base***

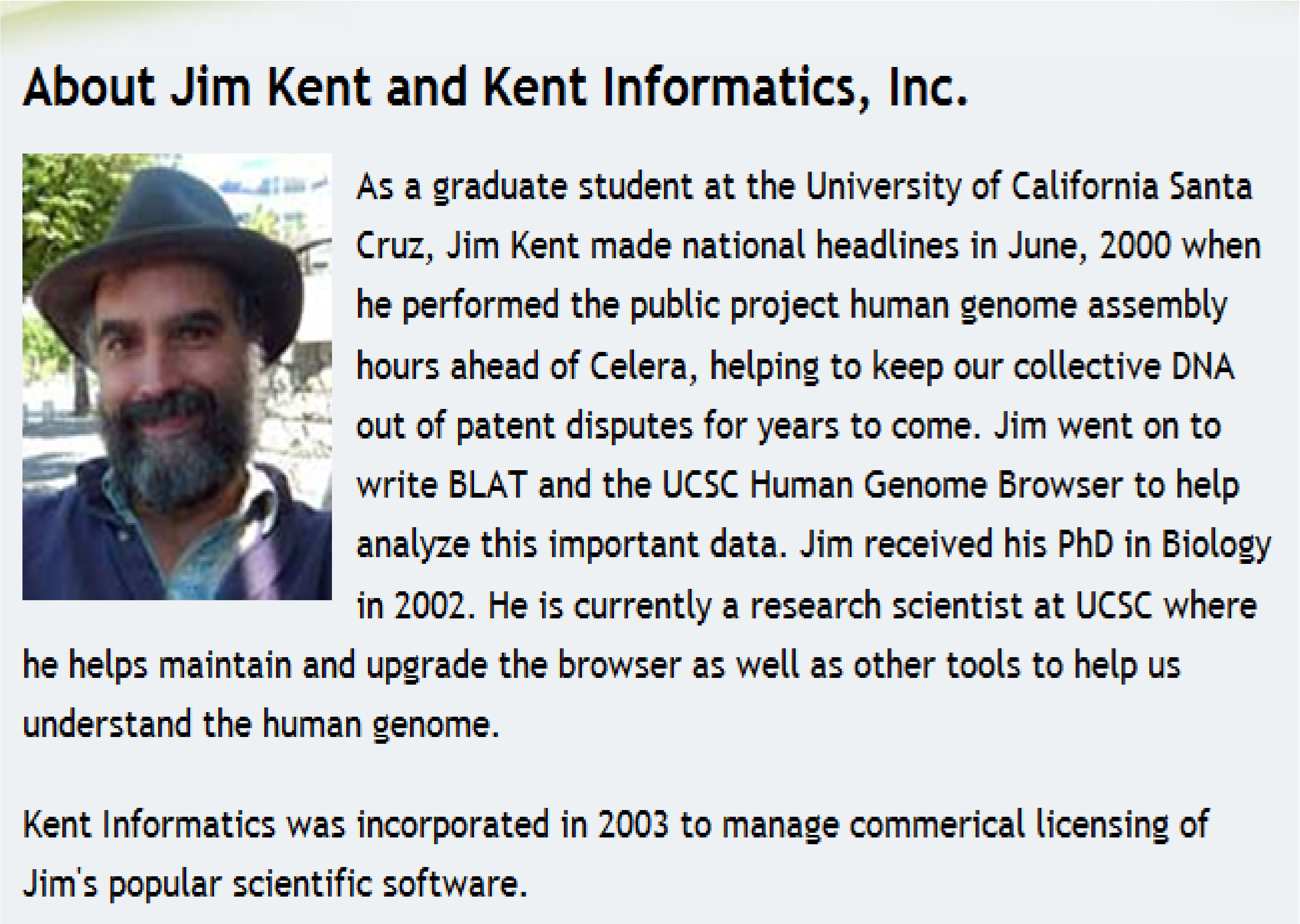
  (Sequenced >1,400 Mb/year at <$0.09 per finished base)
- ***complete genome sequences of E. coli, S. cerevisiae, C. elegans, D. melanogaster***
- ***develop genomic-scale technologies (oligo syn, DNA microarrays, 2-hybrid sys)***

HGP Hero - Jim Kent (research scientist at UC Santa Cruz)

The human genome project was ultimately a race between **Celera Genomics** and the **public** effort, with the final push being a bioinformatics problem to put all of the sequence reads together into a draft genome sequence. **Jim Kent was a grad student at UCSC**, who worked for weeks developing the algorithm and the program ***GigAssembler*** to put all of this together on **June 22, 2000**, **beating Celera by 3 days** (**June 25, 2000**) to an assembled human genome sequence.

His efforts ensured that the human genome data remained in the public domain and were not patented into private intellectual property.

Kent built a grid of cheap (~50), commodity PC's running the Linux operating system and other Freeware to beat Celera's, what was thought of then as the, world's most powerful civilian computer. In **June 2000**, thanks to the work done by Kent and several others, the **Human Genome Project** was able to publish its data in the Public Domain just hours ahead of Celera.



About Jim Kent and Kent Informatics, Inc.

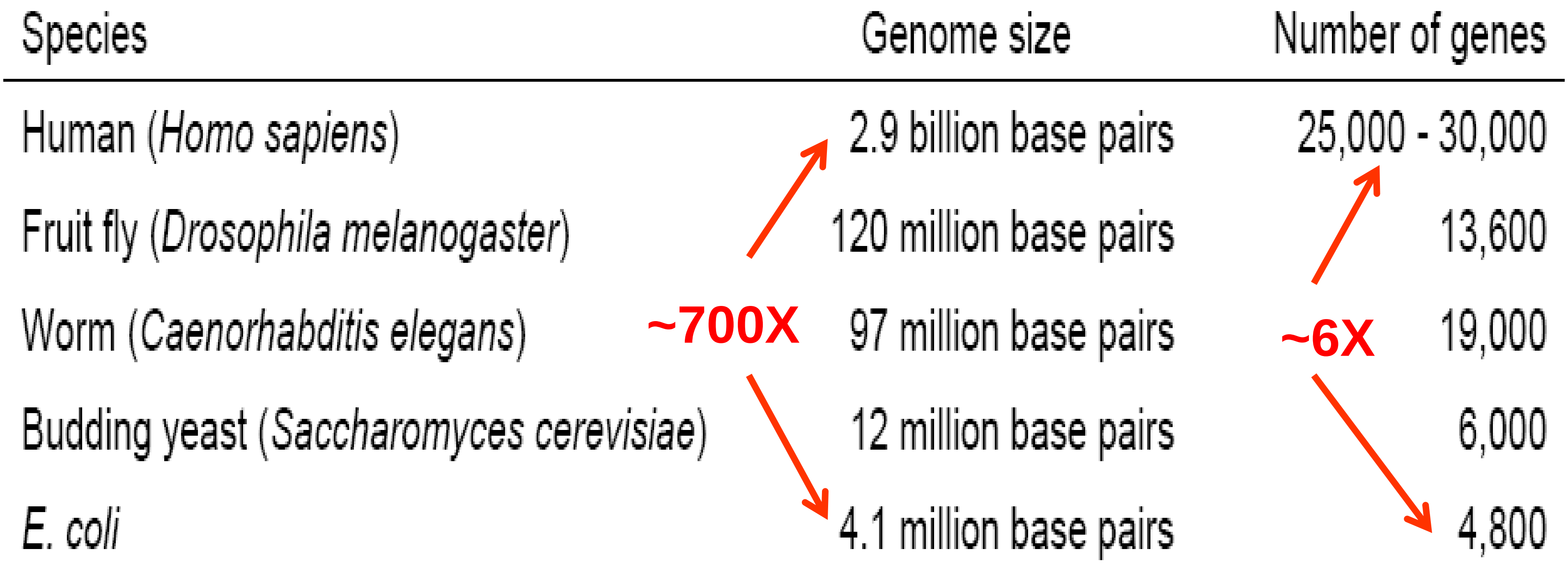
As a graduate student at the University of California Santa Cruz, Jim Kent made national headlines in June, 2000 when he performed the public project human genome assembly hours ahead of Celera, helping to keep our collective DNA out of patent disputes for years to come. Jim went on to write BLAT and the UCSC Human Genome Browser to help analyze this important data. Jim received his PhD in Biology in 2002. He is currently a research scientist at UCSC where he helps maintain and upgrade the browser as well as other tools to help us understand the human genome.

Kent Informatics was incorporated in 2003 to manage commerical licensing of Jim's popular scientific software.

The BIG QUESTIONS:

How many genes?

Why do we have so few genes?



Species	Genome size	Number of genes
Human (*Homo sapiens*)	2.9 billion base pairs	25,000 - 30,000
Fruit fly (*Drosophila melanogaster*)	120 million base pairs	13,600
Worm (*Caenorhabditis elegans*)	97 million base pairs	19,000
Budding yeast (*Saccharomyces cerevisiae*)	12 million base pairs	6,000
E. coli	4.1 million base pairs	4,800

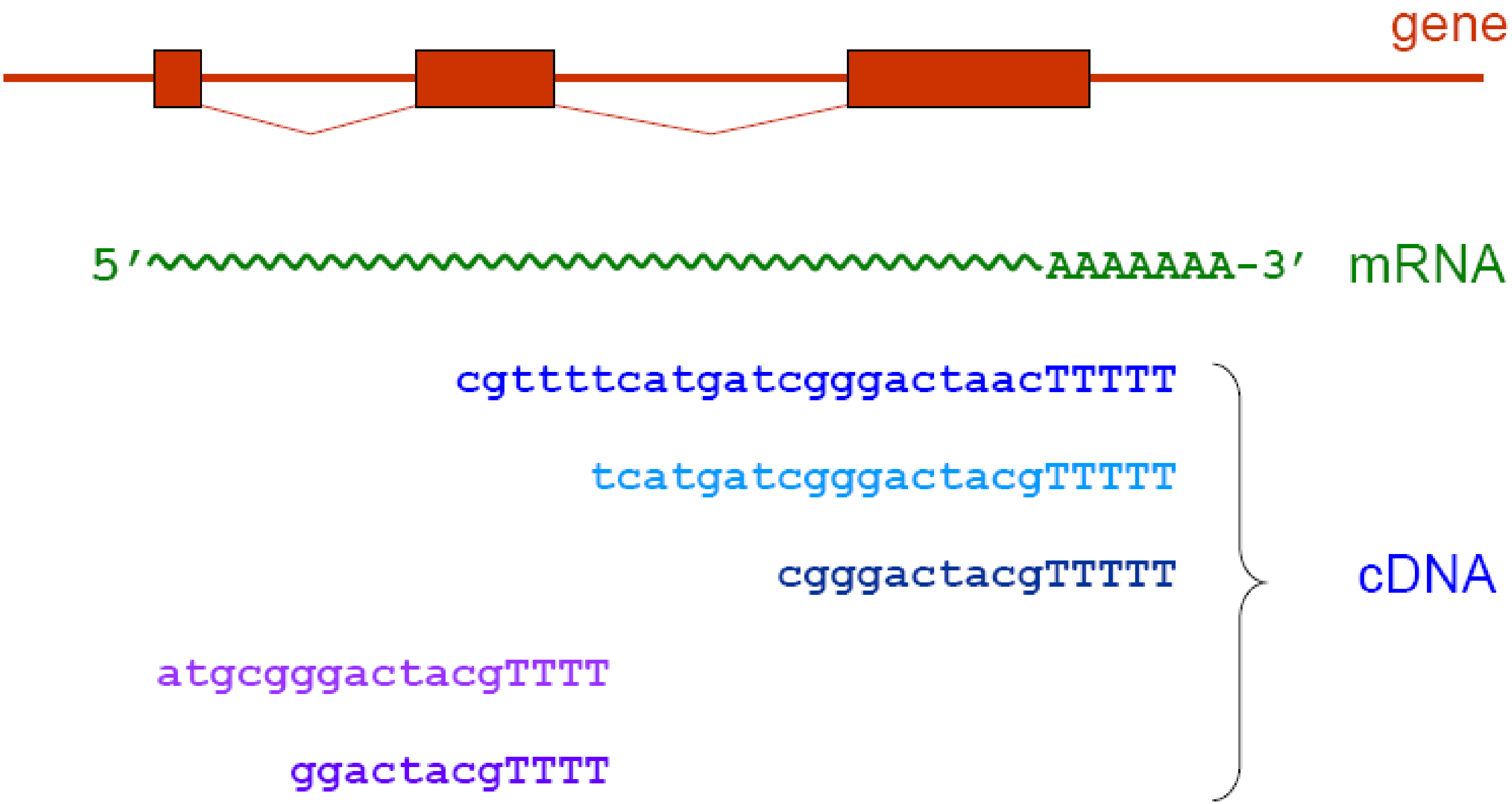
~700X

~6X

Finding genes in genomes

- compare to EST or cDNA sequence
- look for open reading frames
- similarity to other genes and proteins
- Gene prediction algorithms (identifying splice sites, coding sequence bias, etc.)

Genes can also be identified by sequencing cDNAs at random. The sequenced cDNAs are called ESTs (expressed sequence tags)



cgttttcatgatcgggactaacTTTTT

tcatgatcgggactacgTTTTT

cgggactacgTTTTT

atgcgggactacgTTTT

ggactacgTTTT

Genomics vs. Proteomics

With the completion of a rough draft of the human genome in the Spring of 2003, many researchers began looking at how genes and proteins interact to form other proteins. A surprising finding of the Human Genome Project is that there are far fewer protein-coding genes in the human genome than proteins in the human proteome (20,000 to 25,000 genes vs. about 1,000,000 proteins). The human body may contain more than 2 million proteins, each having different functions. The protein diversity is thought to be due to alternative splicing and post-translational modification of proteins. The discrepancy implies that *protein diversity cannot be fully characterized by gene expression analysis*, thus proteomics is needed for characterizing cells and tissues.

Functional genomics and proteomics

- Identify genes and proteins encoded in the genome (Gene finding)

- Measure gene expression on a genome-wide scale (microarrays)

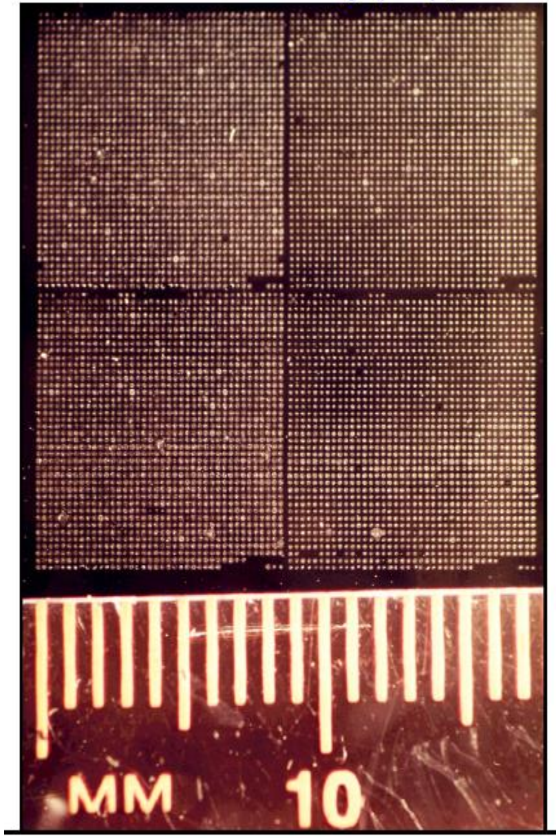
- Identify protein function
  30-50% of the genes in a genome are of unknown function

- Identify protein interactions, biochemical pathways, gene interaction networks inside cells

Methods of making microarrays

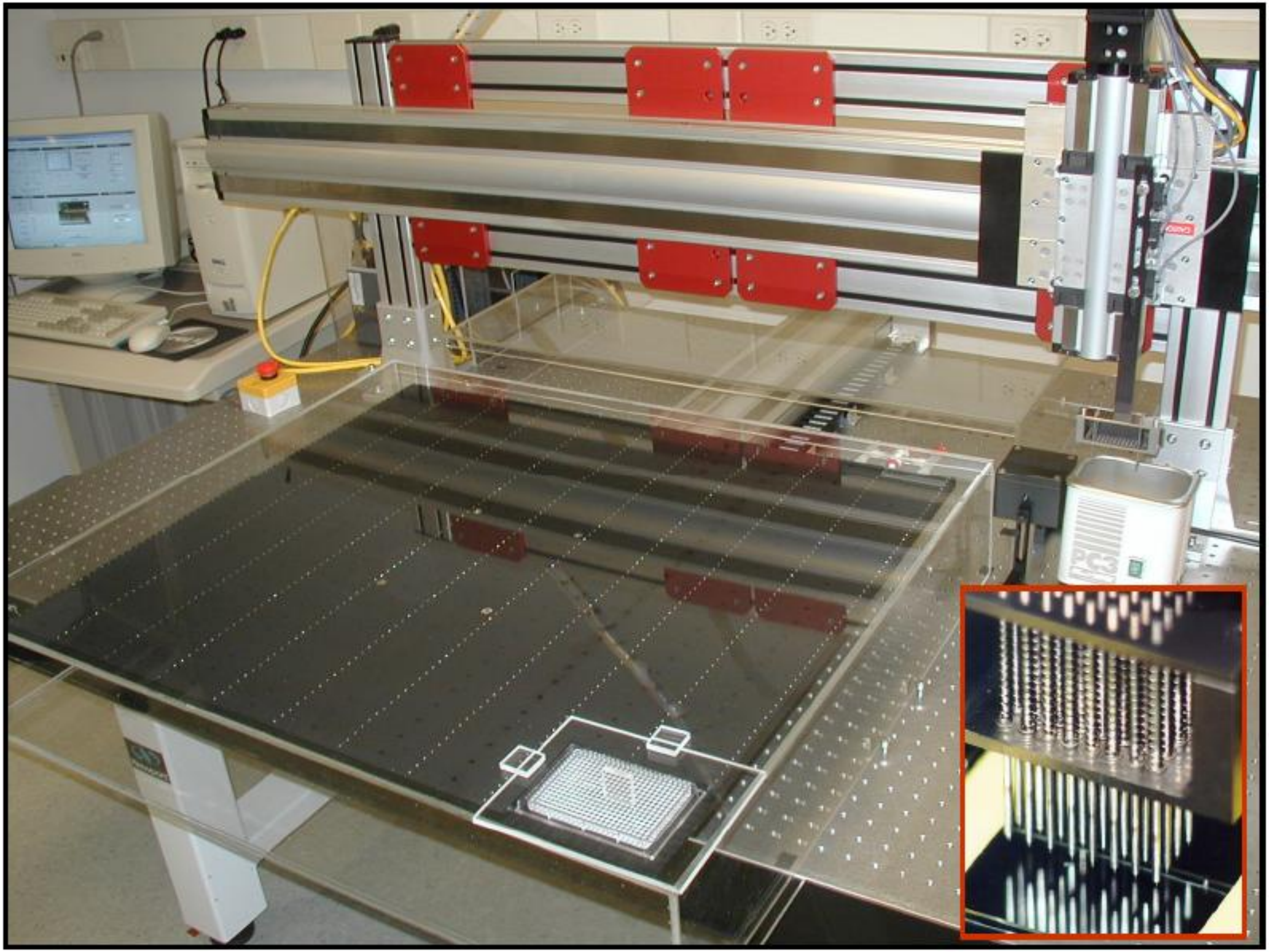
- Robotic spotting
  - using a printing tip
  - using inkjets
- Synthesis of oligonucleotides
  - photolithography (Affymetrix)
  - using inkjets
  - Digital Light Processor (DLP) or Digital Micromirror Device (DMD)

DNA microarray (chip)

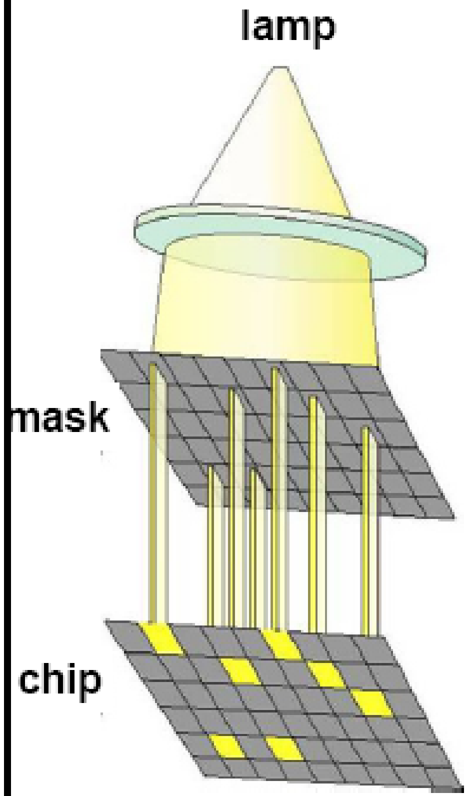


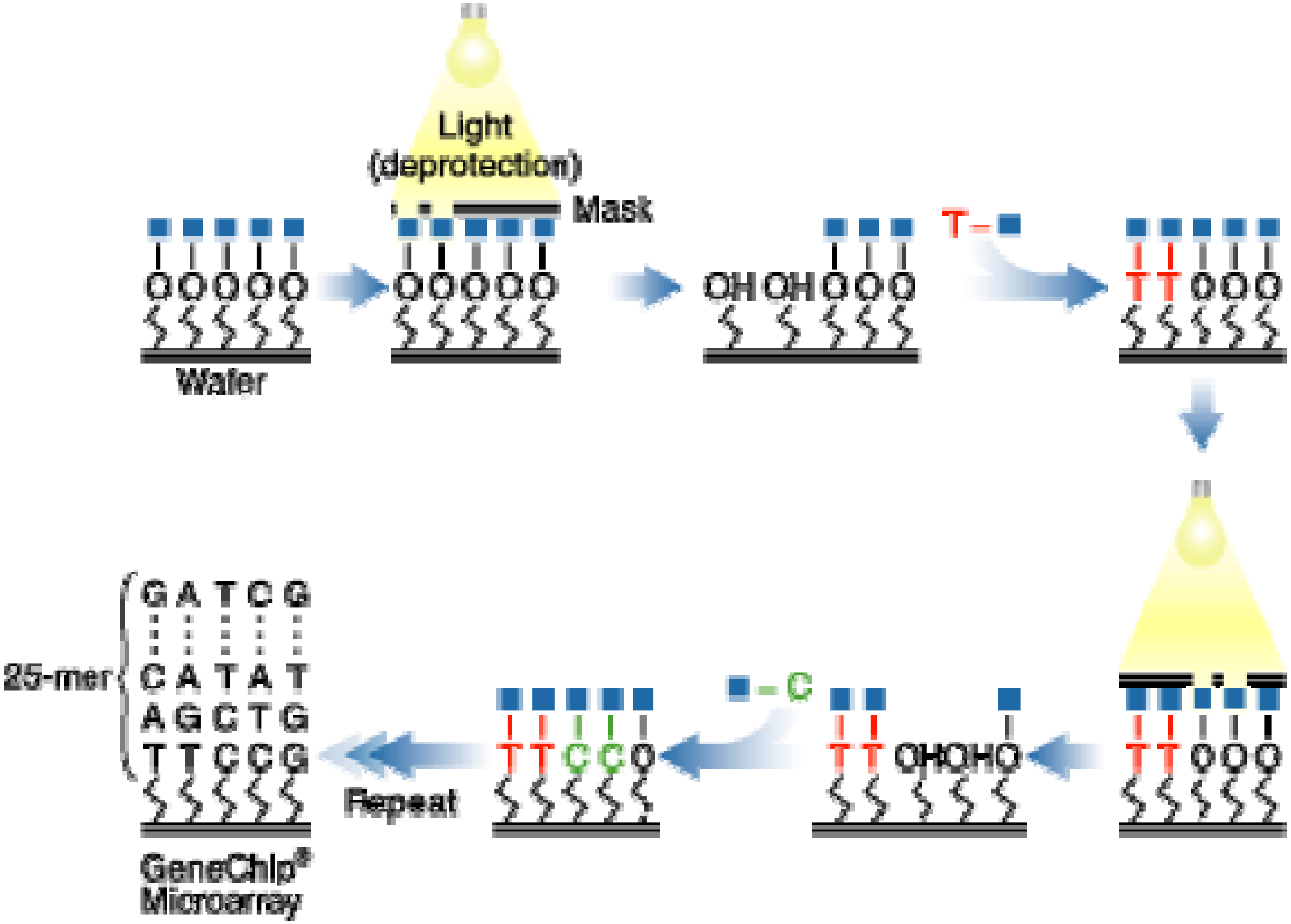
Microarrays can be used to study gene expression, DNA-protein interactions, mutations, protein-protein interactions, etc., all on a genome-wide scale

Note: Thanks to Prof. Vishy Iyer for many of these slides on microarrrays.


PC3

Affymetrix GeneChip

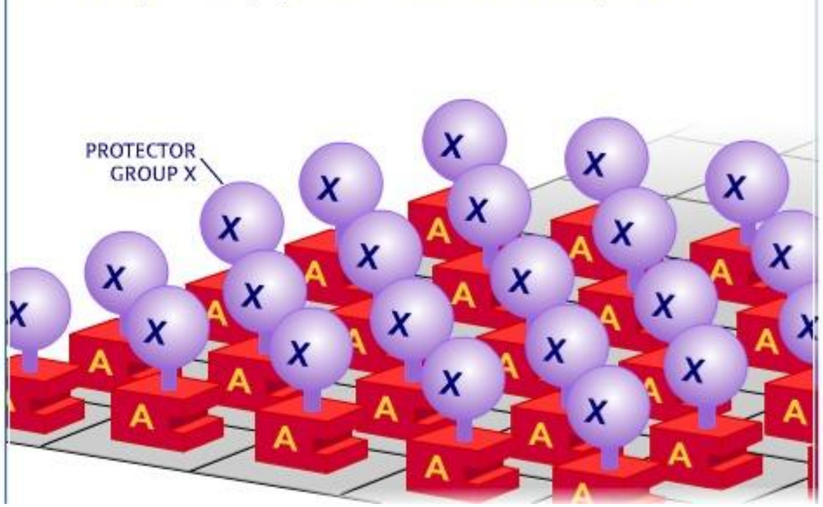




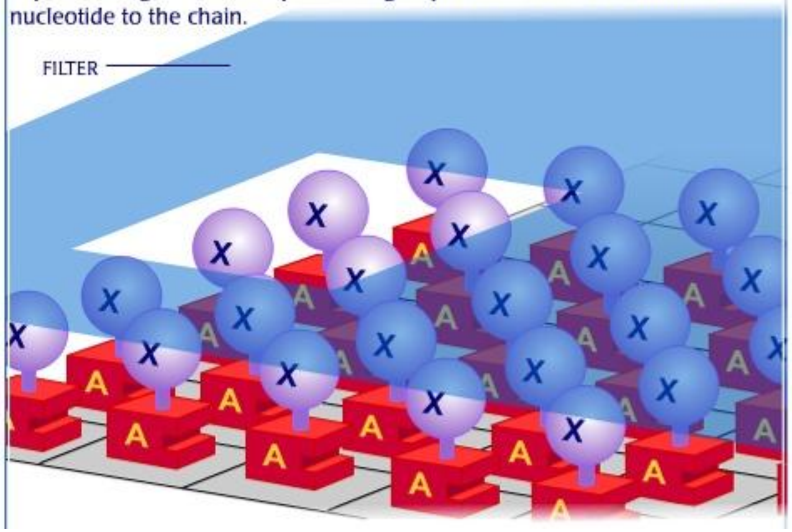
http://www.dnalc.org/ddnalc/resources/dnachip.html

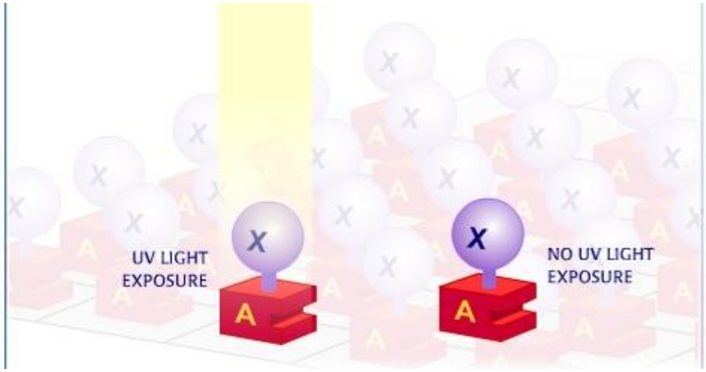
courtesy: *www.affymetrix.com*

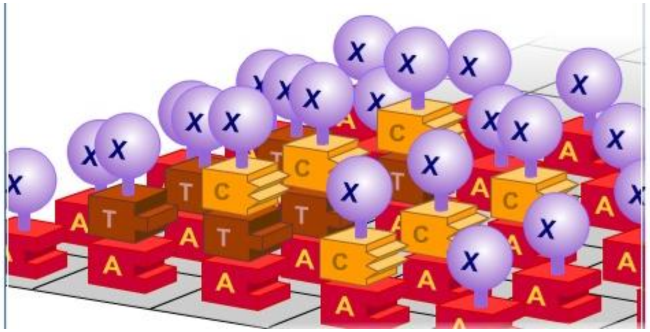
The nucleotide has a protecting group (X) that blocks polymerization. This protector group is photolabile and is released on exposure to UV light. Without the protector, polymerization and chain build-up occur.

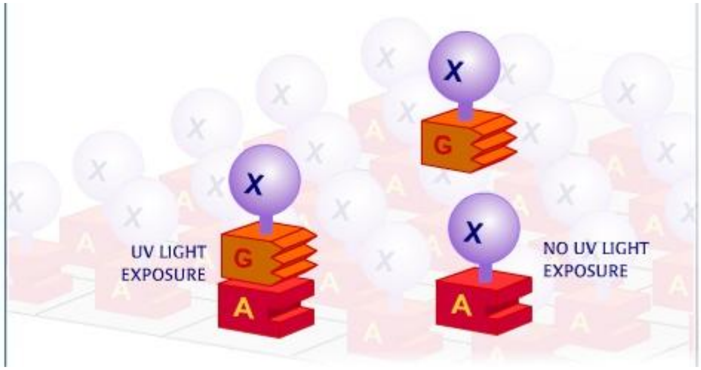


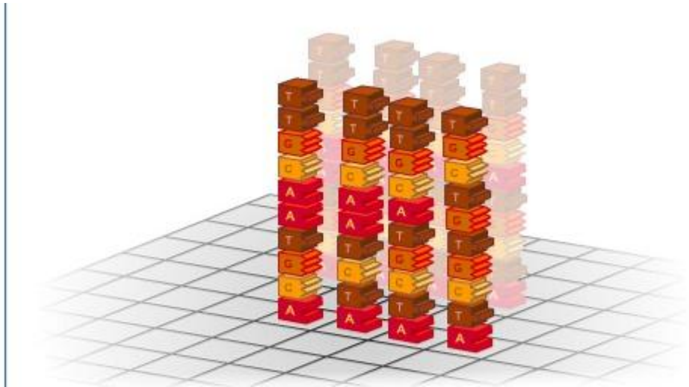
A filter is added to the chip so that only some of the nucleotides are exposed to light. These deprotected groups are then free to add the next nucleotide to the chain.

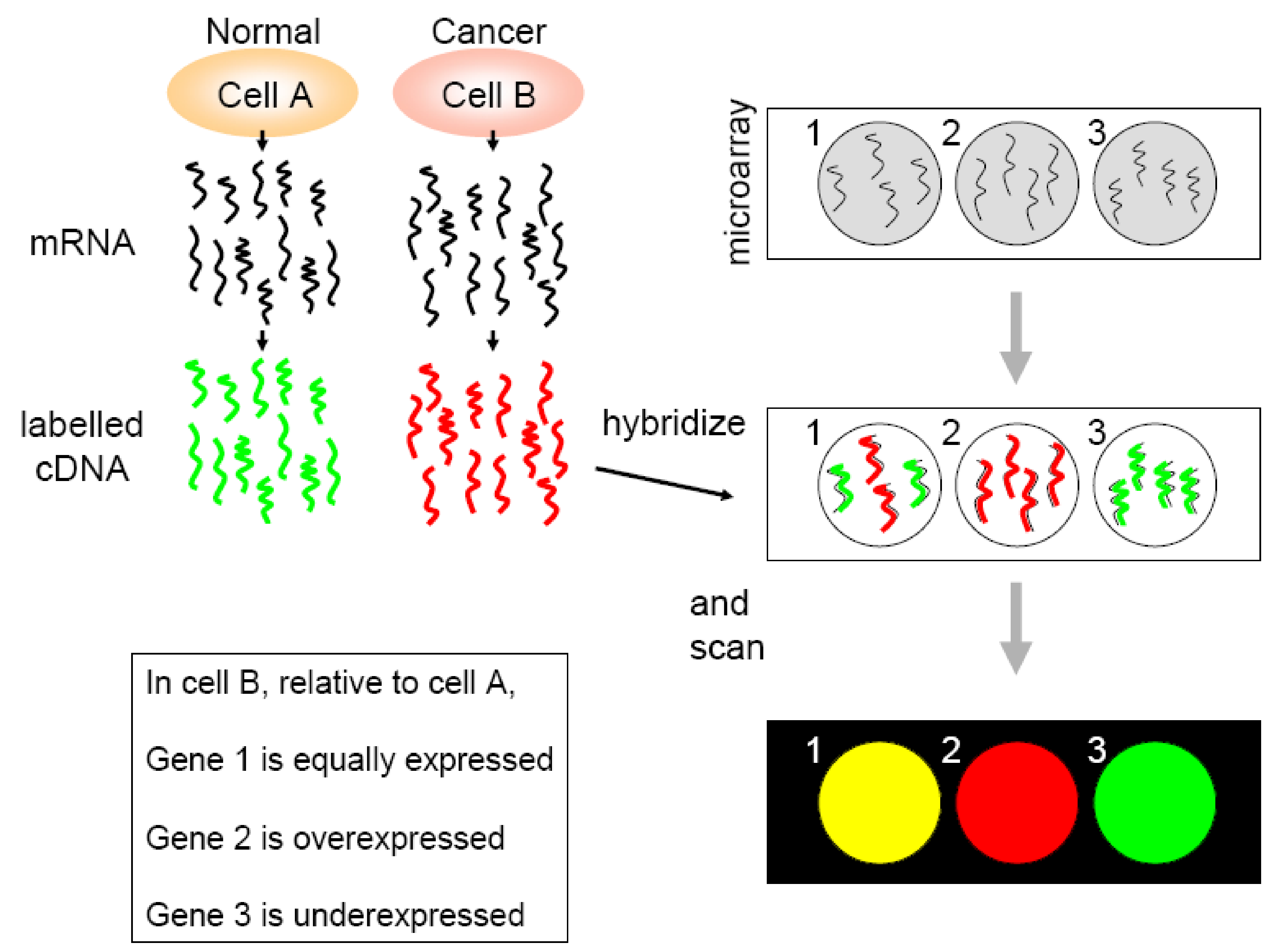












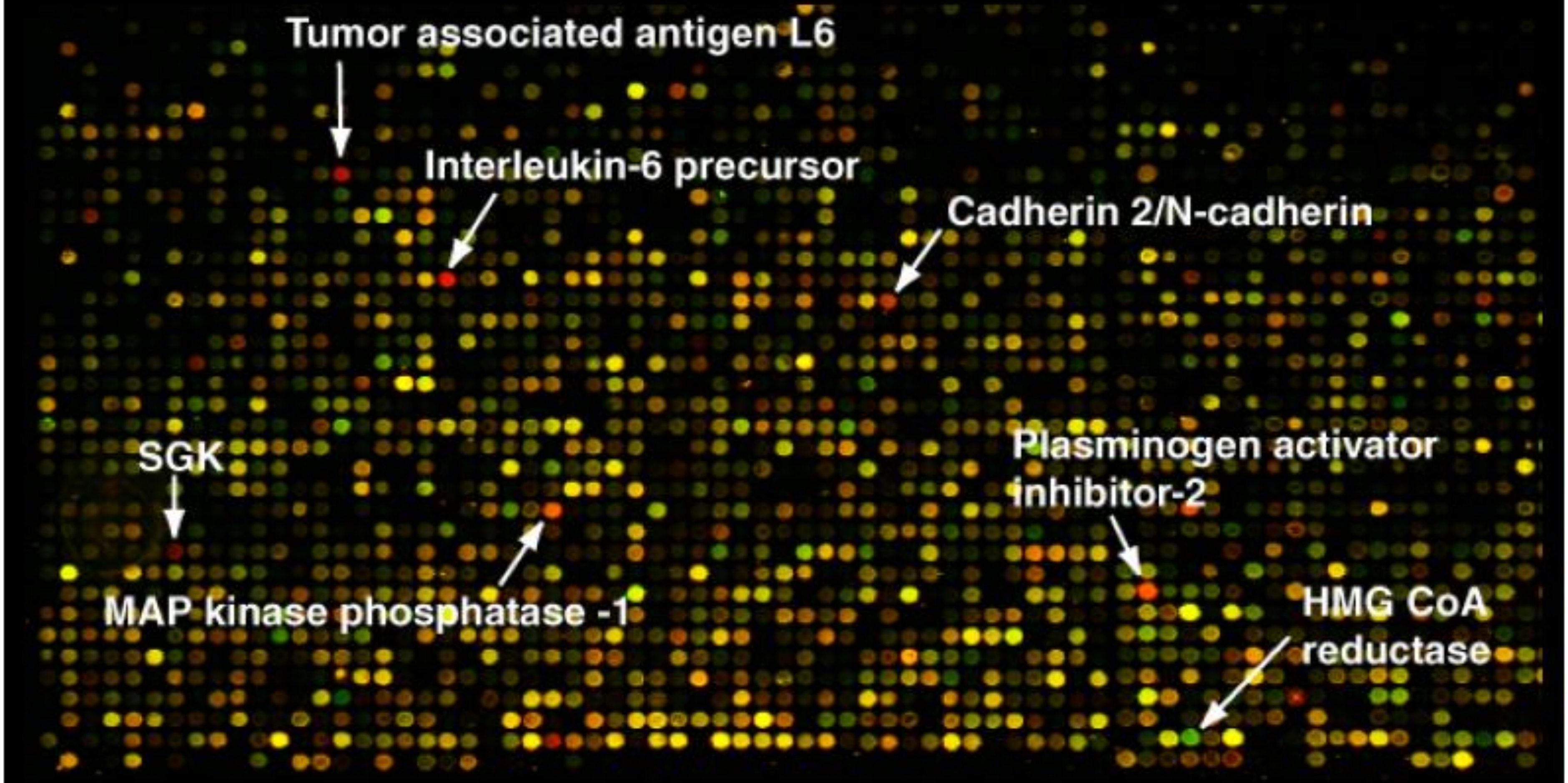
In cell B, relative to cell A,

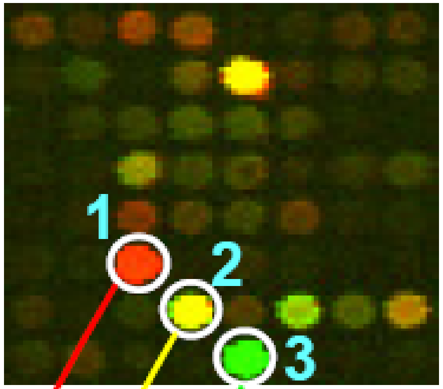
Gene 1 is equally expressed

Gene 2 is overexpressed

Gene 3 is underexpressed

DNA microarray after hybridization of fluorescent probes


Tumor associated antigen L6
Interleukin-6 precursor
Cadherin 2/N-cadherin
SGK
Plasminogen activator inhibitor-2
MAP kinase phosphatase -1
HMG CoA reductase



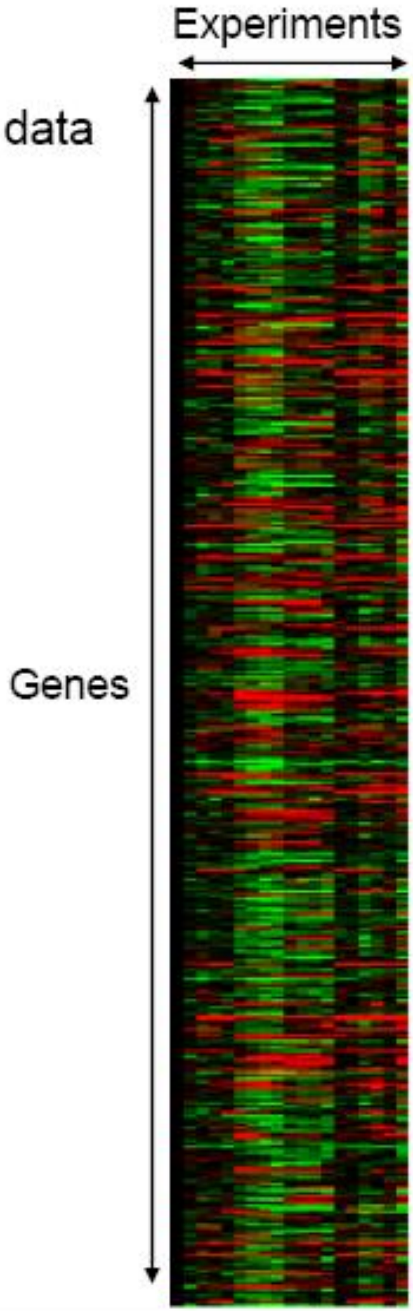
Original microarray image

Colour representation of differential gene expression

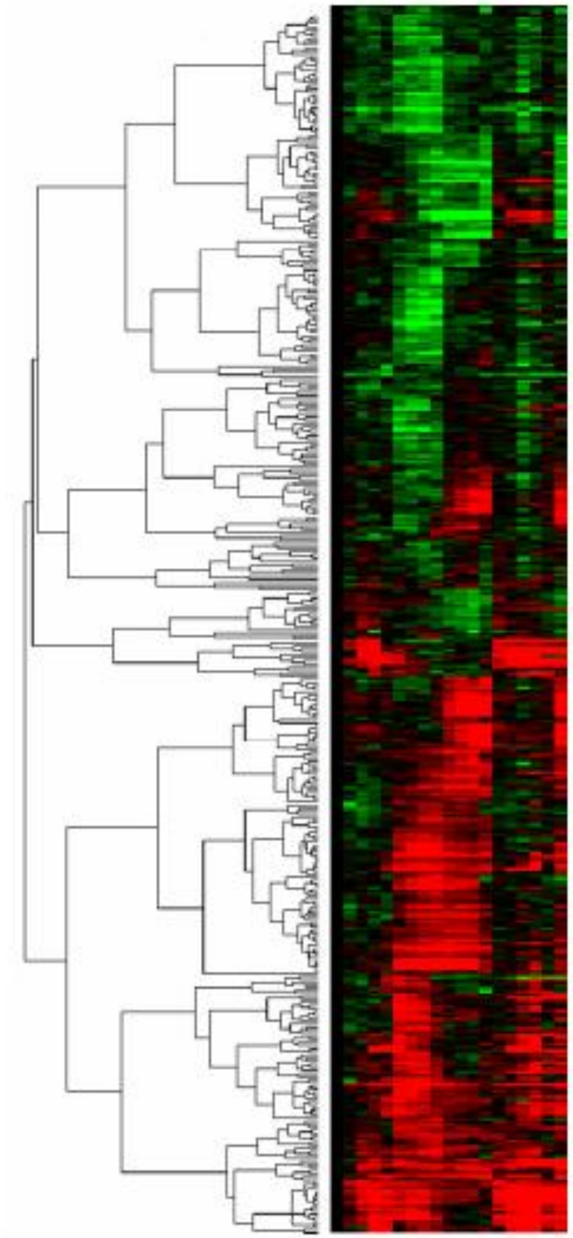
Green	Red	Red/Green		
200	10000	50.00	🟥	Gene 1
4800	4800	1.00	⬛	Gene 2
9000	300	0.03	🟩	Gene 3

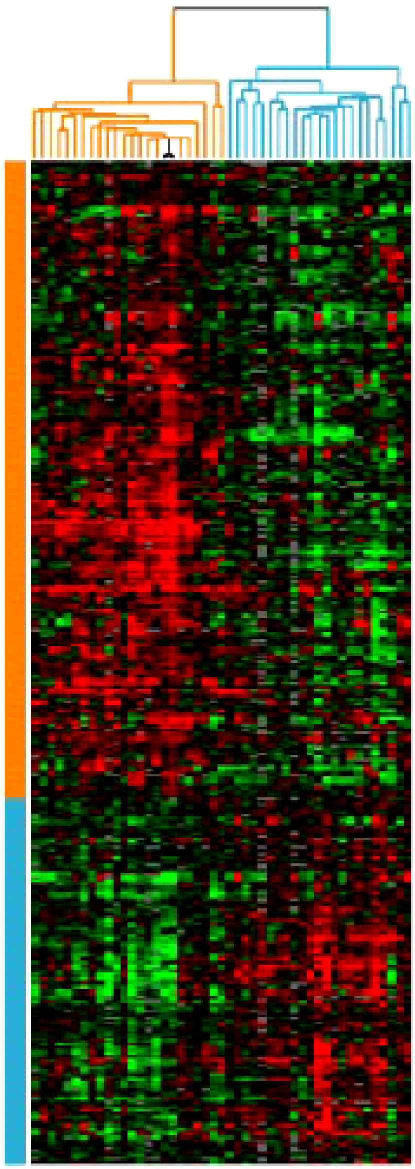
- Large amounts of data can be displayed in this manner
- Gene expression data can be computationally analyzed and organized to reveal patterns

Original data



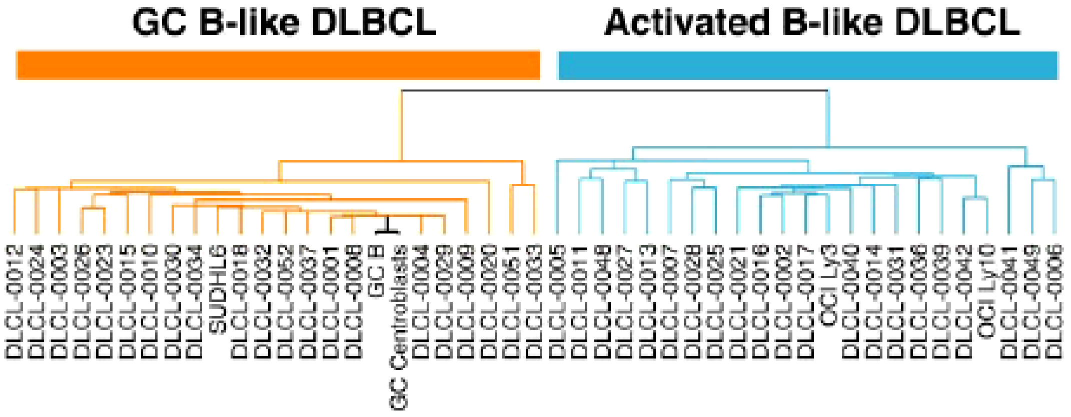
Data after hierarchical clustering





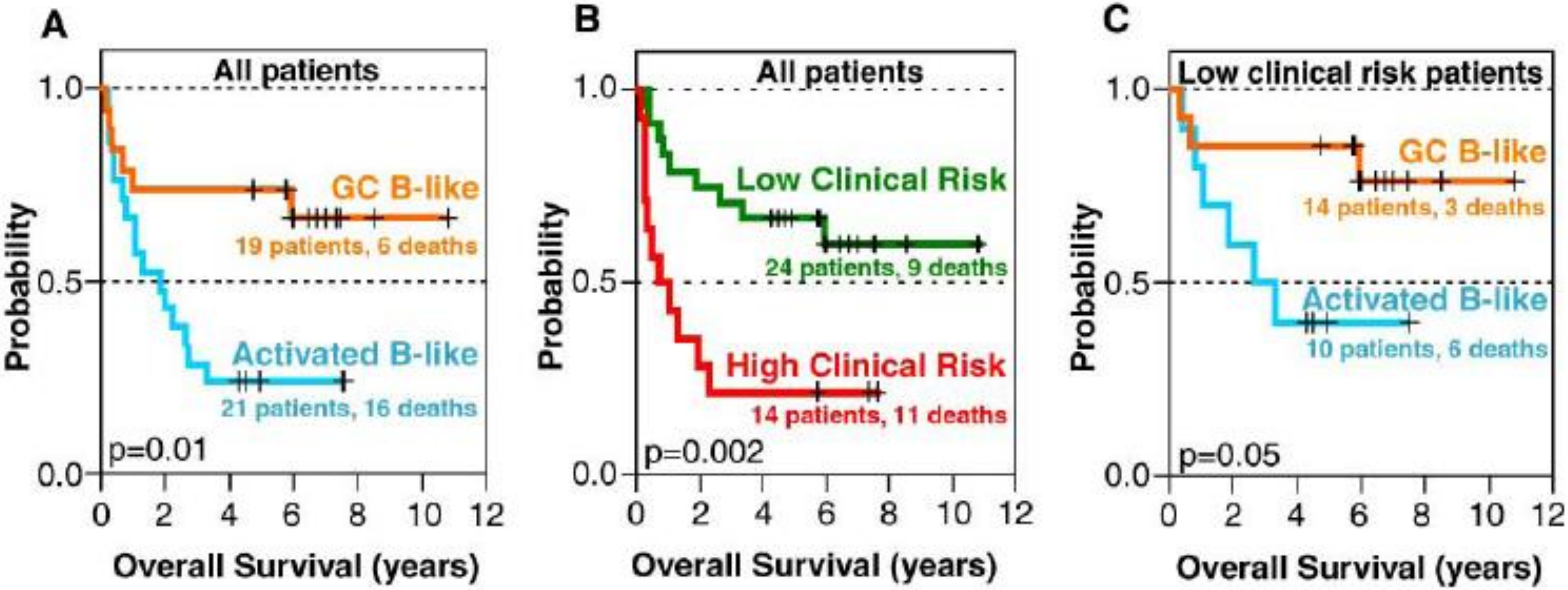
Clustering of tumour samples from cancer patients can be used for molecular classification of cancers. This may be useful for diagnosis and treatment

Subtypes of Diffuse Large B-Cell Lymphoma (DLBCL)



Nature (2000) **403**: 503

Using “clustering analysis,” Alizadeh *et al.* could separate DLBCL into two categories, which had marked differences in overall survival of the patients concerned. The gene expression signatures of these subgroups corresponded to distinct stages in the differentiation of B cells, the type of lymphocyte that makes antibodies.


A
All patients
GC B-like
19 patients, 6 deaths
Activated B-like
21 patients, 16 deaths
p=0.01
B
All patients
Low Clinical Risk
24 patients, 9 deaths
High Clinical Risk
14 patients, 11 deaths
p=0.002
C
Low clinical risk patients
GC B-like
14 patients, 3 deaths
Activated B-like
10 patients, 6 deaths
p=0.05
Probability
Overall Survival (years)