

The Human Genome Project – Its History and Advancements into New Research and Technology

Jutta Marzillier, Ph.D
Lehigh University
Biological Sciences

September 5th, 2014

Objectives

Techniques that enabled genome sequencing

Process of Human Genome Sequencing

New DNA sequencing Technologies

Some research spinoffs



**Guess, who turned
60 last year !!!**

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey¹. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons:

(1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β -D-deoxy-ribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow righthanded helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions.

Each chain loosely resembles Furberg's² model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's standard configuration³, the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 Å. The distance of a phosphorus atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine

bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so

that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain, does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain, is given, then the sequence on the other chain is automatically determined.

It has been found experimentally^{4,5} that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data^{5,6} on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in time following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereo-chemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at King's College, London. One of us (J.D.W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J.D. WATSON
F.H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.

¹Pauling, L., and Corey, R. B. *Nature*, 171, 346 (1953); *Proc. U.S. Nat. Acad. Sci.*, 39, 84 (1953).

²Furberg, S., *Acta Chem. Scand.*, 6, 634 (1952).

³Chargaff, E., for references see Zamenhof, S., Braverman, G., and Chargaff, E., *Biochim. et Biophys. Acta*, 9, 402 (1952).

⁴Wyatt, G.R., *J. Gen. Physiol.*, 34, 291 (1952).

⁵Astbury, W.T., *Symp. Soc. Exp. Biol.*, 1, *Nucleic Acid*, 66 (Camb. Univ. Press, 1947).

⁶Wilkins, M. H. F. and Randall, J. T. *Biochim. et Biophys. Acta*, 10, 102 (1953).

Guess, who turned 60 last year !!!

The structure of the DNA double helix:

published in **1953** by Watson and Crick in the journal of 'nature' thereby launching the '**Genomic Age**'



**Guess, who turned
10 last year !!!**

Human Genome Project

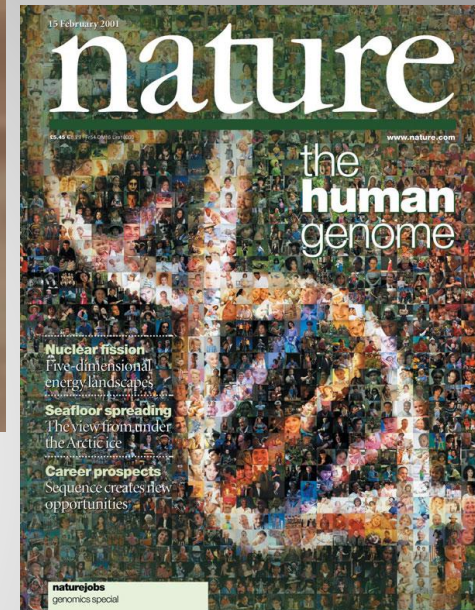
- 2001 Draft Human Genome Sequence
- 2003 Finished Human Genome
(50 years after DNA structure solved)

Two techniques published in 1977 by

- Sanger et al.
DNA sequencing by **chain termination** or **dideoxysequencing**
- Maxam & Gilbert
DNA sequencing by chemical modification

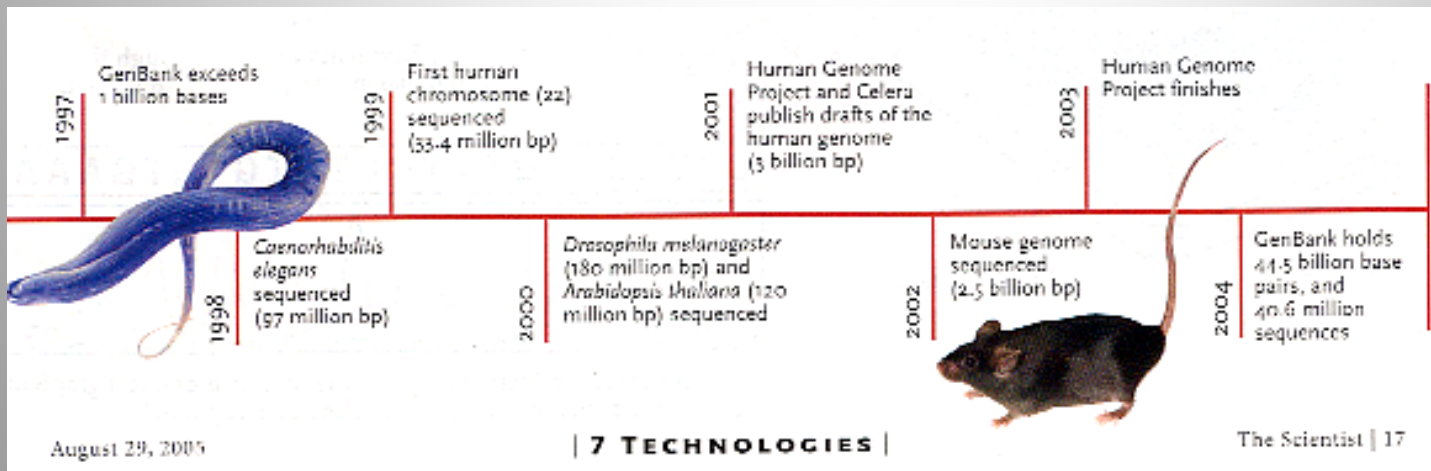
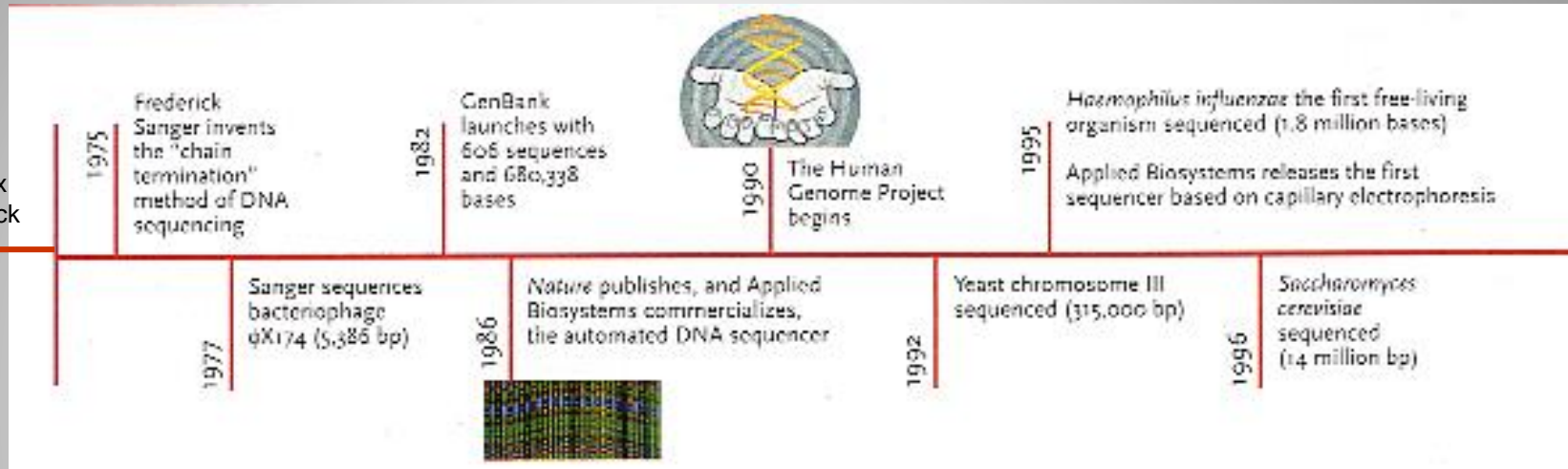
The original method of **Sanger** sequencing and multiple improvements regarding chemistry and computation lead to complete sequencing Human Genome (and many others).

Sanger sequencing is expensive and **Next-Generation-Sequencing (NGS)** technology took its place



Development Sequencing Time Line

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



2005



2006



Genome Trivia

How many base pairs (bp) are there in a human genome?

How much did it cost to sequence the first human genome?

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Why is the Human Genome information important?

to serve as a reference for human disease

Genome Trivia

How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Why is the Human Genome information important?

Genome Trivia

How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ \$ 3 billion

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Why is the Human Genome information important?

Genome Trivia

How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ \$ 3 billion

How long did it take to sequence the first human genome?

~ 13 years

When was the first human genome sequence complete?

Why is the Human Genome information important?

Genome Trivia

How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ \$ 3 billion

How long did it take to sequence the first human genome?

~ 13 years

When was the first human genome sequence complete?

~ 2000-2003

Why is the Human Genome information important?

Genome Trivia

How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ \$ 3 billion

How long did it take to sequence the first human genome?

~ 13 years

When was the first human genome sequence complete?

~ 2000-2003

Why is the Human Genome information important?

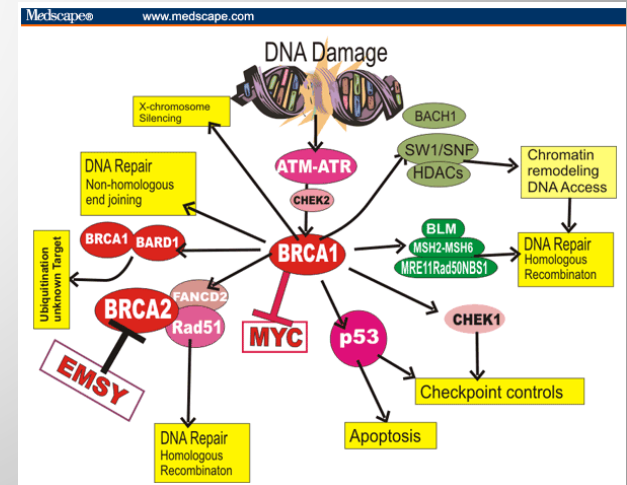
to serve as a reference for human disease

Why is Genome Sequencing Important?

- To obtain a 'blueprint' – DNA directs all the instructions needed for cell **development** and function
- DNA underlies almost every aspect of human health, both, in function and dis-function
- To study gene expression in a specific tissue, organ or tumor
- To study human variation
- To study how humans relate to other organisms
- To find correlations how genome information relates to development of **cancer**, susceptibility to certain **diseases** and drug metabolism (pharmacogenomics)
- Outlook: Personalized Genomics (David Church, Harvard)

The establishment of various sequencing technologies gave rise to many new fields in biology, medicine and engineering

- Identification of genes contributing to disease
- synthetic biology
- personalized medicine
- genomic screening
-



<http://jenniferdayan.girlshopes.com/brca1europe/#>

Outline

DNA Sequencing

Biological Dogma and Principle of DNA synthesis

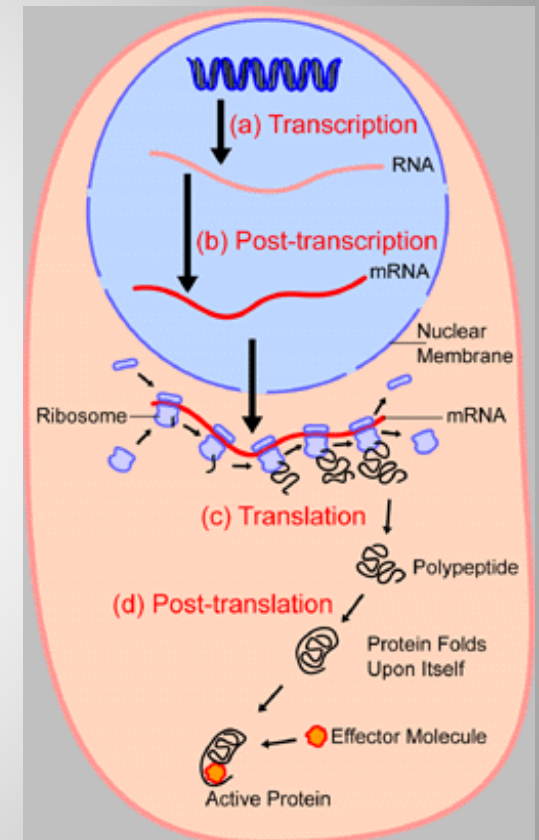
Sanger sequencing and its improvements

Next-Generation-Sequencing (NGS) Technologies

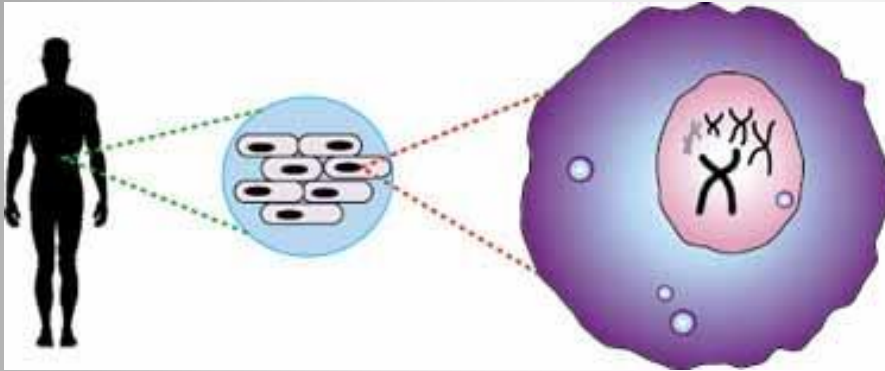
3rd Generation Technologies

Projects and 'Spinoffs'

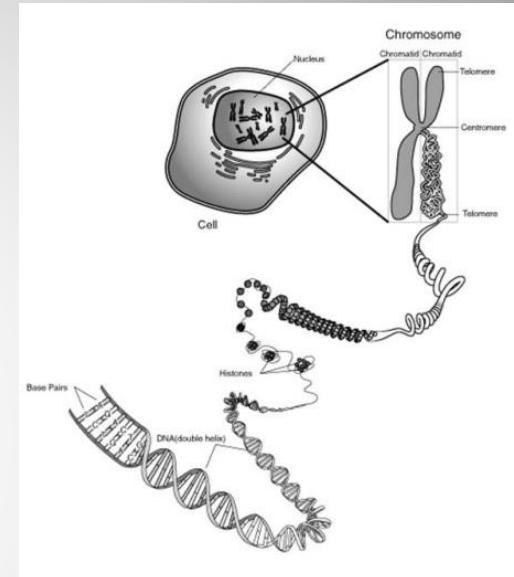
'Dogma of Biology'



Why is the Knowledge about the Human Genome interesting?



http://www.genomenewsnetwork.org/articles/06_00/sequence_primer.shtml



http://www.pharmainfo.net/files/images/stories/article_images/

The human body has about 100 trillion cells with more than 200 different cell types.

Each **cell** harbors the same genetic information in its nucleus in form of DNA containing chromosomes.

Depending on cellular, developmental, and functional stage of a cell **only a subset of genes** is expressed.

What is DNA?

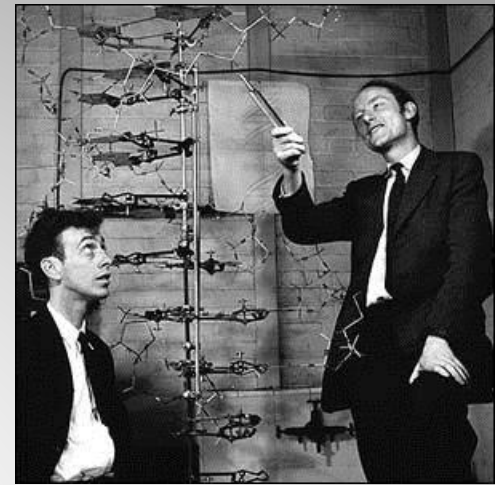
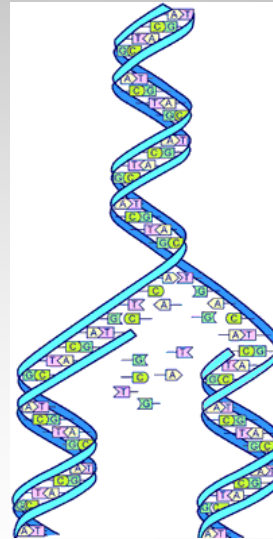
DNA constitutes the heritable genetic information that forms the basis for the developmental programs of all living organisms.

A **genome** is an organism's complete set of DNA

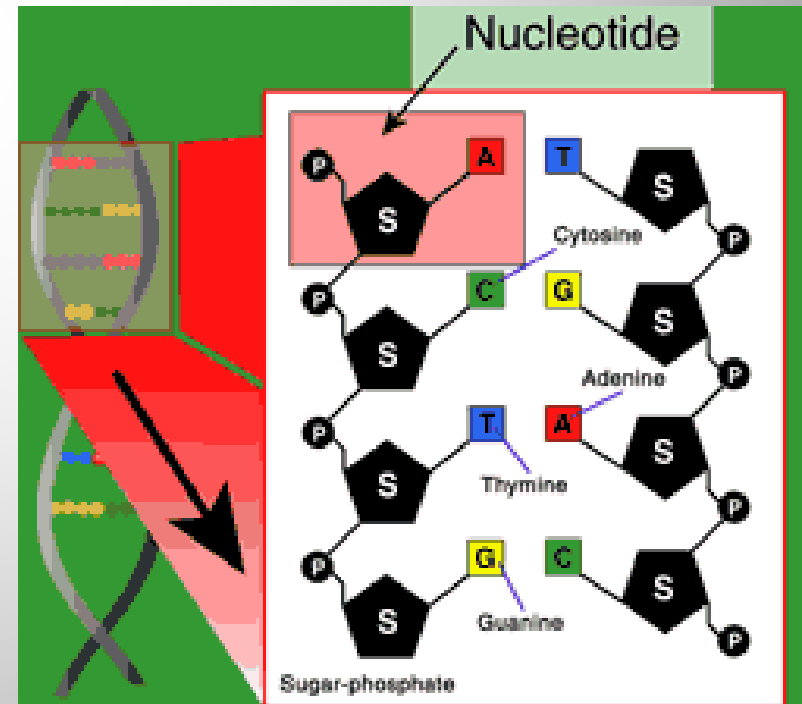
The DNA is made up of *four* building blocks called **nucleotides**.

DNA sequence is the particular side-by-side arrangement of bases along a DNA strand.

DNA sequencing is a biochemical method to determine the sequence of the nucleotide bases that make up the DNA.

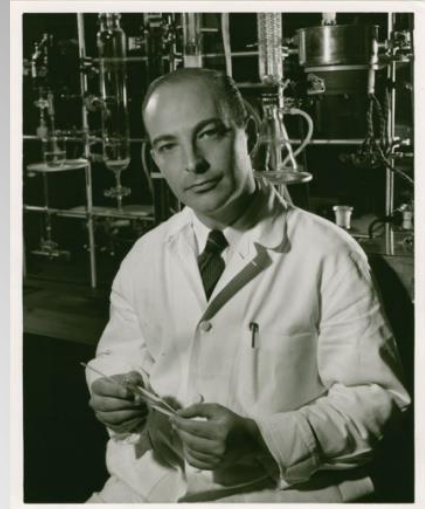


Watson & Crick, 1953

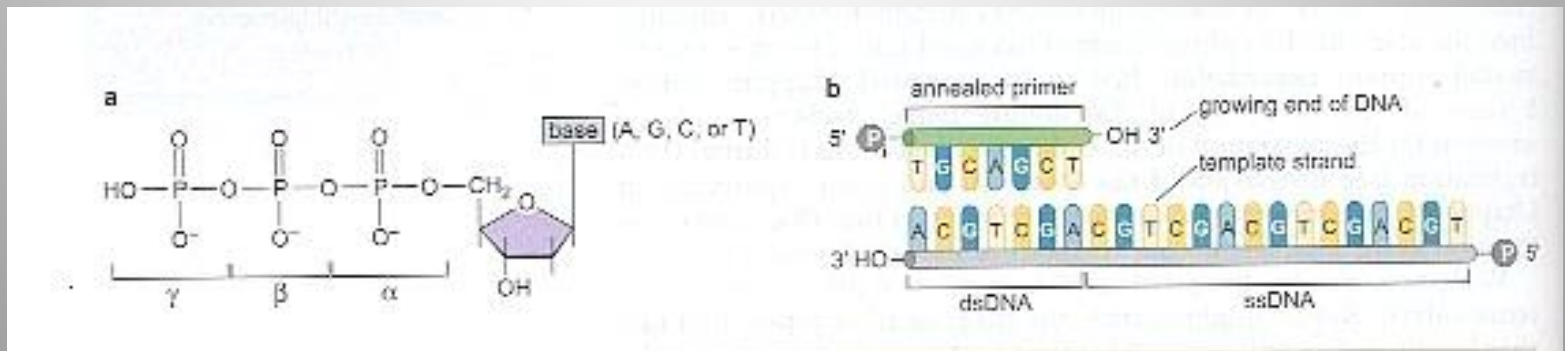


Principle of DNA Synthesis

Arthur Kornberg demonstrated DNA replication in a cell-free (*in vitro*) bacterial extract (Nobel prize, 1959)

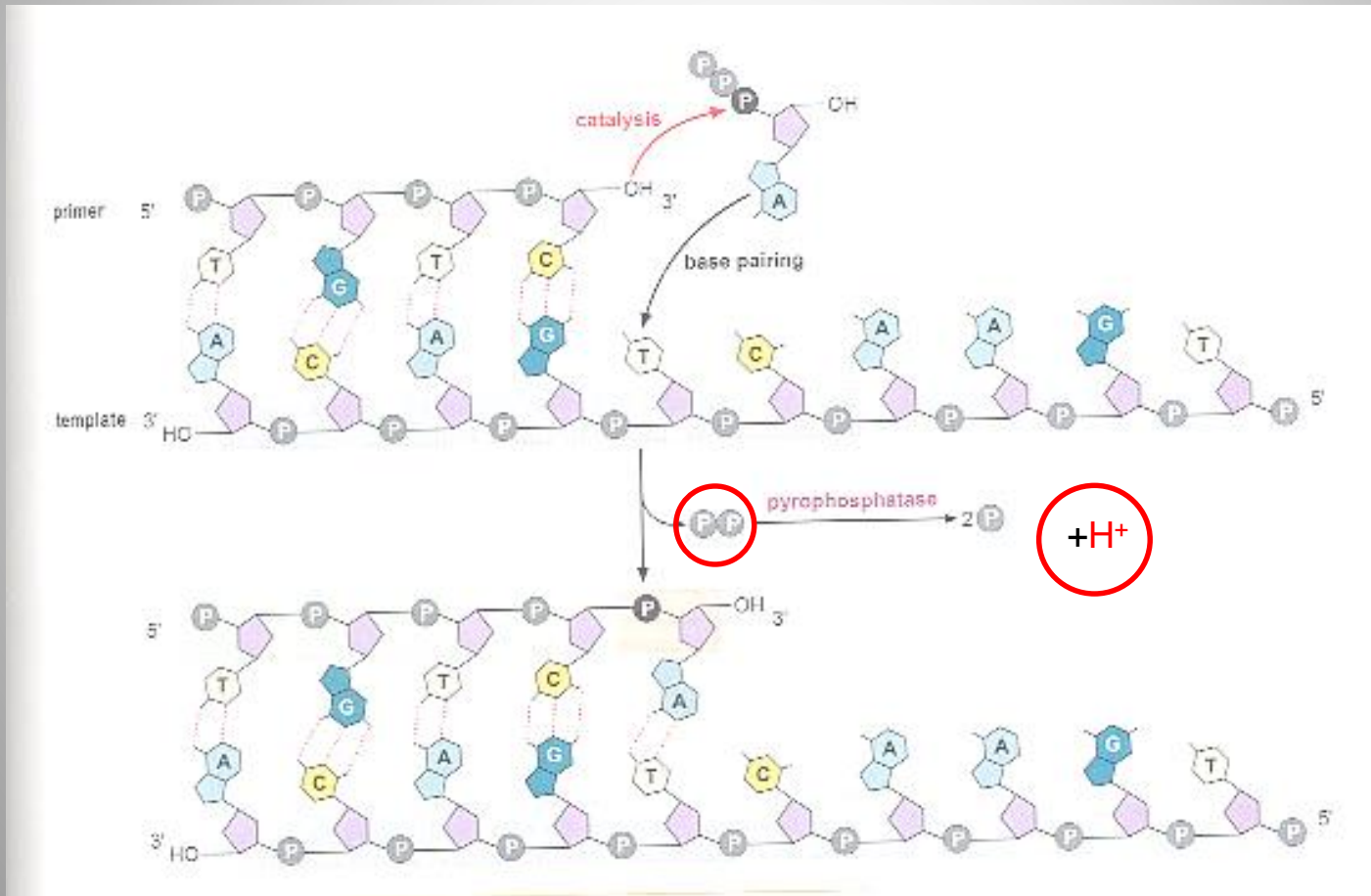


- Discovered *DNA polymerase* (Pol1) to facilitate DNA synthesis
- Unraveled the *mechanism* of DNA synthesis
- nucleotide building blocks
- a single DNA strand serves as a template
- can only extend a pre-existing chain (**primer**)
- a free 3' Hydroxyl end is required

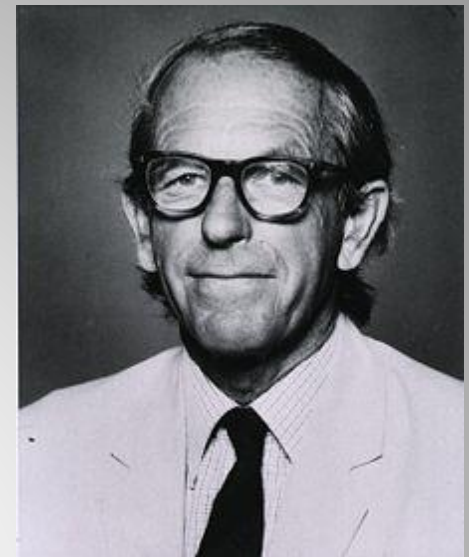


Reaction Mechanism for DNA synthesis

The 3' hydroxyl group of the primer attacks the α -phosphoryl group of the incoming nucleotide thereby forming a phosphodiester bond (S_N2 reaction).

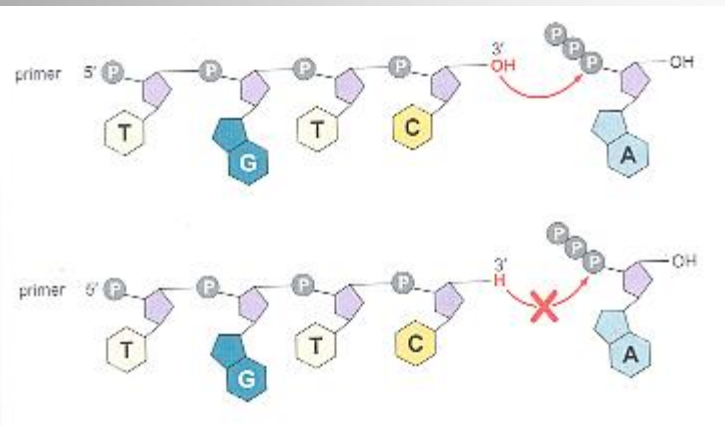


Dideoxy Sequencing according to Sanger



Frederick Sanger
Nobel Prize (1980)

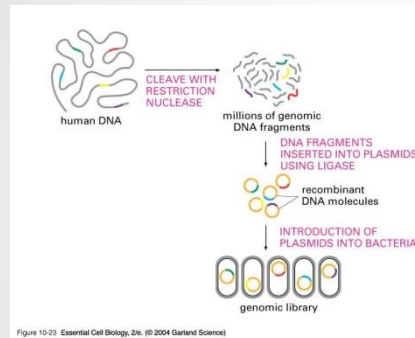
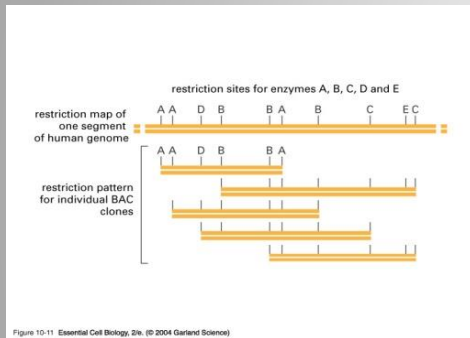
Both nucleotide types can be incorporated into growing DNA chain.



Presence of **dideoxy**-cytosine in growing chain blocks further addition of incoming nucleotides.

How was Sanger Genome Sequencing Done?

Clone by Clone

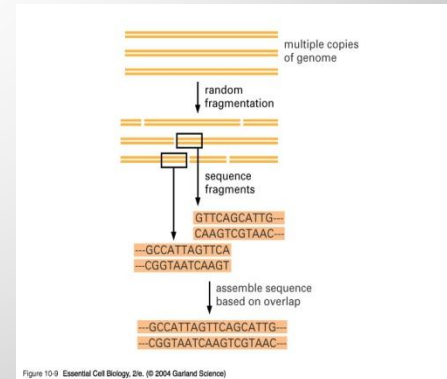


Shotgun sequencing

Break genome into random fragments, insert fragments into vector, sequence each of the fragments and assemble fragments based on sequence overlaps

Create a crude physical map of the whole genome by restriction mapping before sequencing

Break the genome into overlapping fragments and insert them into BACs and transfect into *E.coli*

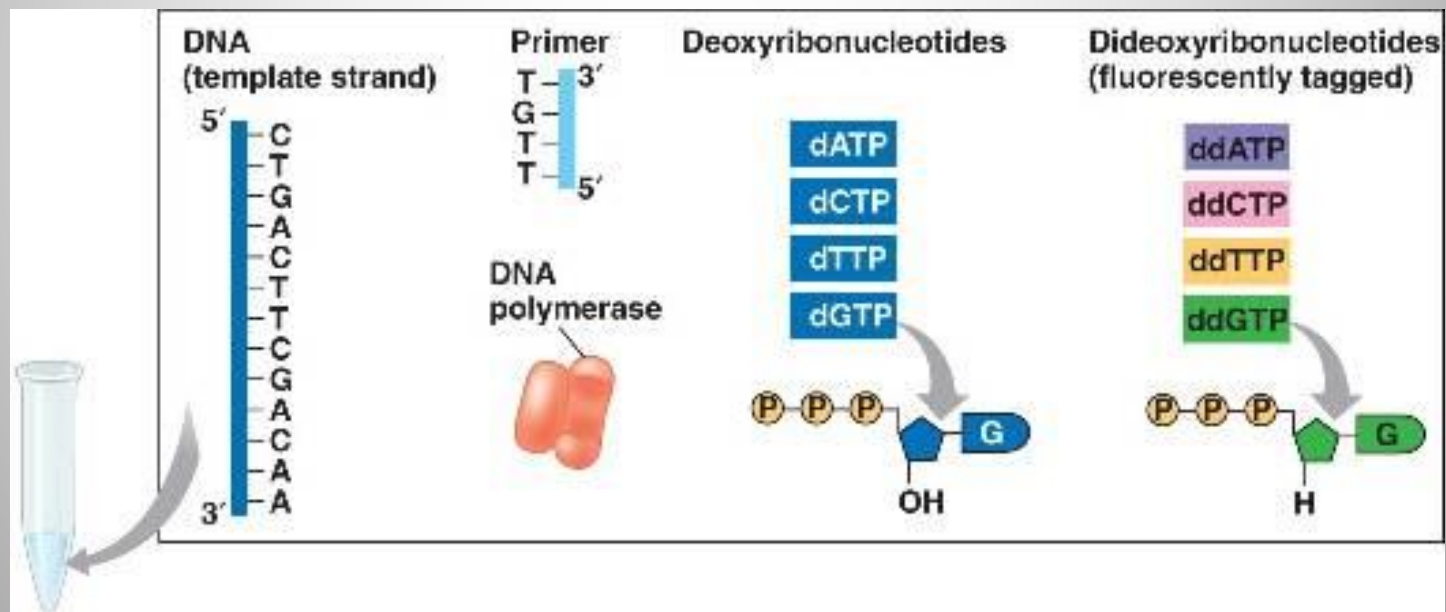


- need known flanking region to anneal primer

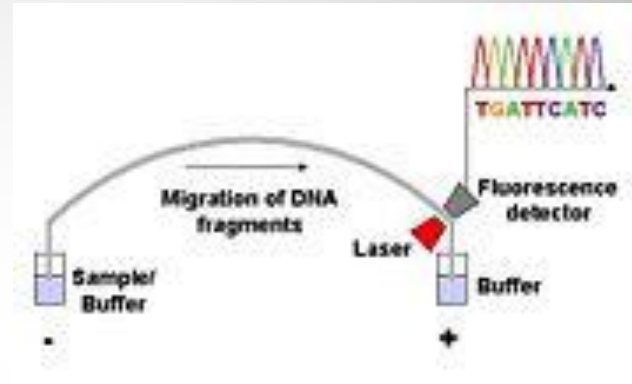
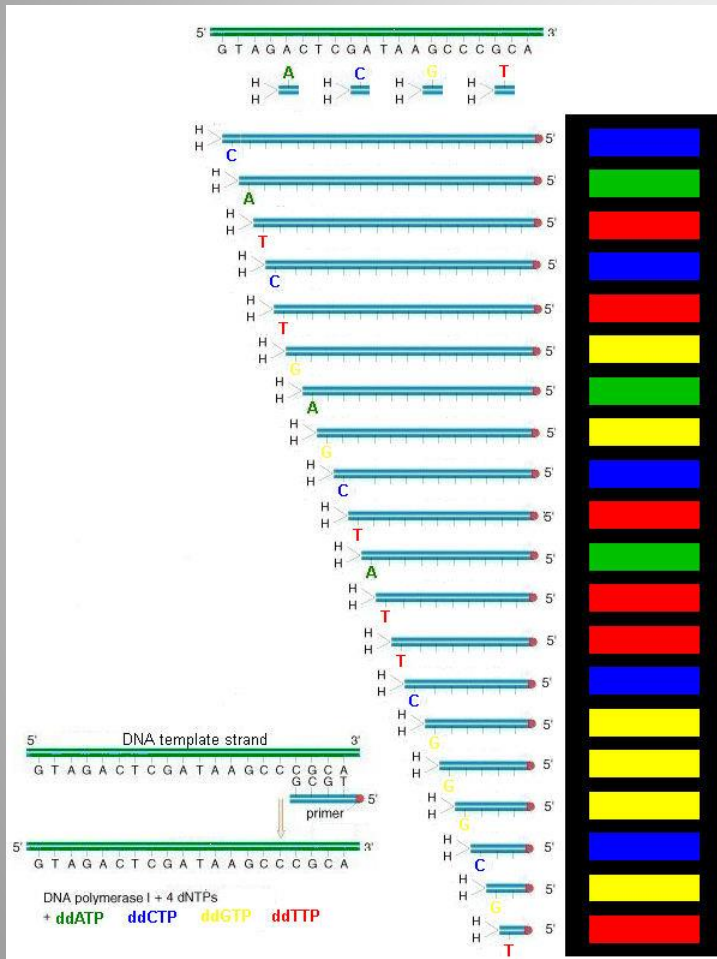
Dideoxy Method of Sequencing (Sanger, 1975)

DNA synthesis is carried out in the presence of **limited amounts of fluorescently labeled dideoxynucleoside triphosphates** that results in chain termination

Through chain termination fragments of distinct sizes are generated which can be separated by gel electrophoresis



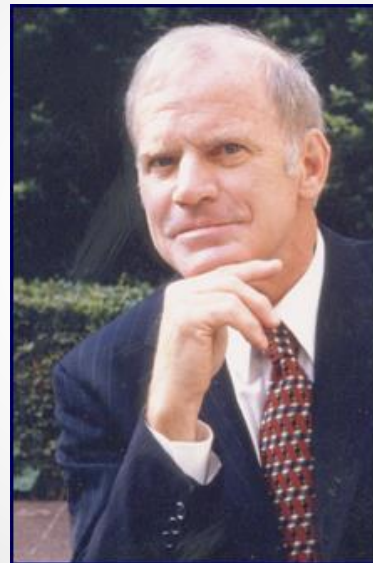
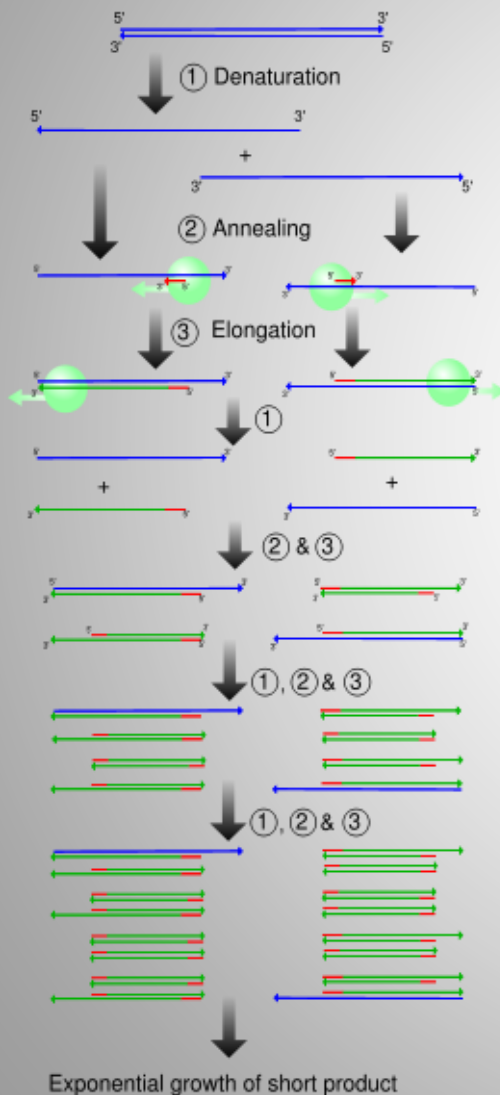
Separation of Sequencing Fragments by Capillary Gel Electrophoresis



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

High resolution
Read length up to 1,000 nucleotides

Use of heat-stable *Taq* Polymerase Enabled Automation of Sequencing Reaction

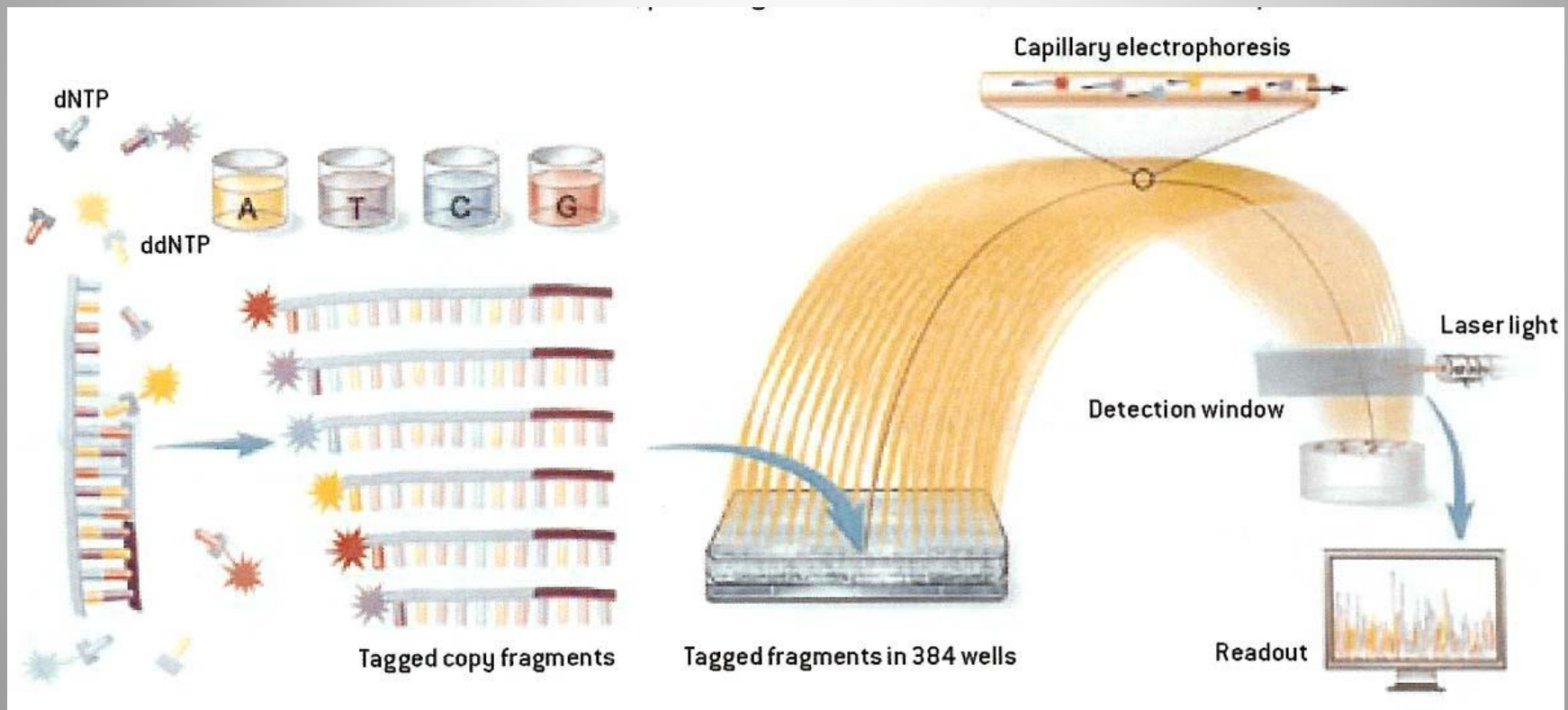


Kary Mullis invented the
Polymerase Chain Reaction



High-Throughput Whole Genome Sequencing

Analysis of 384 sequencing reactions in parallel



JGI Sequencing Facility

(Joint Genome Institute, US Department of Energy)

Engineering



Assume more than 20 384-capillary sequencers running simultaneously
approx. 1000 bp per capillary run
approx. 3 hours per run

➔ Approx. 60 Million bases per day in one facility

fasta format

> Sequence,

```
TGCGGCTGCCAGATTTTGTACGGGTTTGGAAAGTCGACGGAGAGAACAGCGCGG
GCCTAGAAGGCCCCCGTAATGCCCCCTGAGAGCCCCGTAGACGGACGAACGGTG
CGGATCGATAGATGGCACCGGAGACAAGCGAAGACGGCCGCAGAGCCGTCGC
CGGCTGACGCCCCGCGTAGGAAGATATTCGTGTGAAGTGCGTCACATTCTACGGG
TGAAACGCGAAAGTGGAAGGTTCTTACCTATGGAGGGGGTAAGGGAGCGAGCT
CCAGCGAGCGACCGCACCCCGACATAGGTTCTTGTCGGGGTAGTCGAACGGAG
AGAGACTACCCCTTTTAGCGACCTCCGGTCGCCCAGGTAGGTACCGAACGATGA
GAGAGGTACCTAGACCGTACAGGCCGGGGTTTATCCCCCGGCCGATACAGCATG
GTCATTTTGGGTAGGTACGTTACGTAAGCATCACTCACCAACAGAACCAGTGGTT
ACGTAACCGGGTACGTTACGTACGACTAGATACGTAACAGAACCCTAACCCTG
GCCGCCCGAAGGCGGCCCGCAGCGGGTTACGTTTGTGTCAGGTATGTCCTAGGG
AGGGTGAACCTATGAATCTCAAGGAGCATCAGAAGGAGCTTCTTCGTCTACTGCTC
GTGTCTGAGTCGATGACGTATGAGCGTGCCACTCGAAGGTGGAGACCATCCTG
CACCTCATCCAGTCCTTCGACGCGGAGAACTCGATAAGTGAGCTGGGAGTCATC
TGACCGGCGCGATCGTCTGCCGACCGACTGGCCTCGCATCCGCCGCGAGGTTT
TGCGGGCGGCTGGCCACCGCTGCCAGATCCGCTACGCGGACATCTGCACAGG
GATGGCTACCGAGGTTGATCACGTCCGCTACCGCGACGAGGCGTCACCTCTGC
AGGTGTCGTGCAGACCGTGCCATGCGCGGAAGTCCGCGATGGAAGGCGTTGCT
CAGCGTGCGAAGCTGCGCGCGATGAAGAAGCGGCCGCCGCCCGCCACCCGG
GGCGTAGAAGCAACTAGGAGGGACCAGGCGTCCCCGAGCCCAGGAGGCGTCA
TGCCGGGTCCAGTGCCCAAGCGATCGGACGAACGCGTCCGGCGCAAATCGCAT
ACGAGTGAGCGCGAGGCTAGC
```

Finishing Draft Sequences using Assembly Software - Bioinformatics

Primary sequencing reads:

```
cagacgtgtcagtcgactcgatatactgagctagtcgact
tagctagccggatagtagtattaccagacgtgtcagtcgactcgata
ccagatcgatcgattcgcgatagctagccggatagtagtattaccagacgt
```

Align sequences for homologies

```

cagacgtgtcagtcgactcgatatactgagctagtcgact
tagctagccggatagtagtattaccagacgtgtcagtcgactcgata
ccagatcgatcgattcgcgatagctagccggatagtagtattaccagacgt
```

Green: 3-fold coverage

Yellow: 2-fold coverage

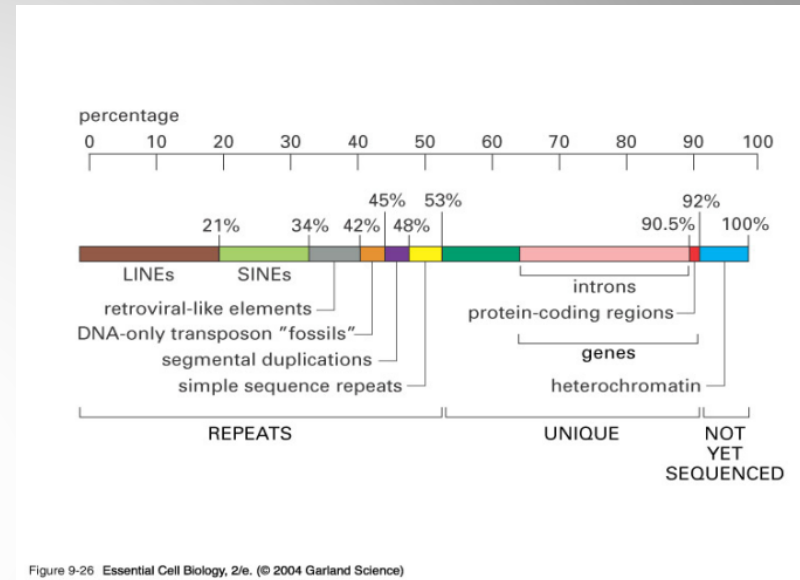
Fragment assembly to contig

```
ccagatcgatcgattcgcgatagctagccggatagtagtattaccagacgtgtcagtcgactcgatatactgagctagtcgact
```

Goal for Genome Sequencing:
Sanger: 8-fold coverage
454: 30 fold coverage

What did we learn from HGS?

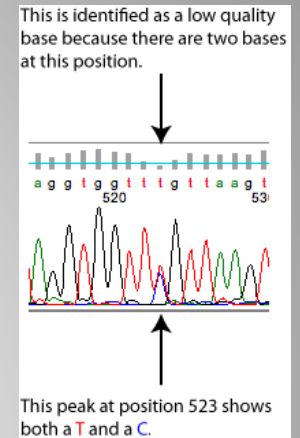
- Less than 2% of the Human Genome codes for protein
- Encodes for approx. 20,000 protein-coding genes
- The human genome sequence is almost exactly the same (99.9%) in all people
- The genome size does not correlate with the number of estimated genes



| Organism | Genome Size | Est. Gene # |
|-----------|-------------|-------------|
| E. coli | 4.6 Mb | 4,400 |
| Yeast | 12.1 Mb | 6,200 |
| Roundworm | 97 Mb | 19,700 |
| Fruit Fly | 180 Mb | 13,600 |
| Rice | 389 Mb | 37,500 |
| Human | 3200 Mb | 25,000 |

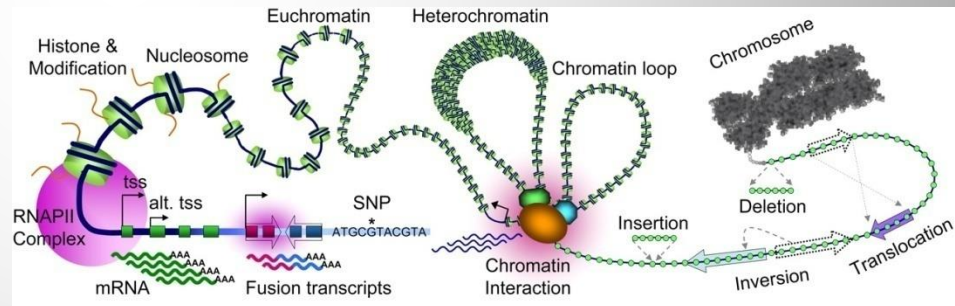
Humans have only about twice the number of genes as a fruit fly and barely more genes than a worm!

The Human Genome Project Sequence Represents a 'Composite' Genome



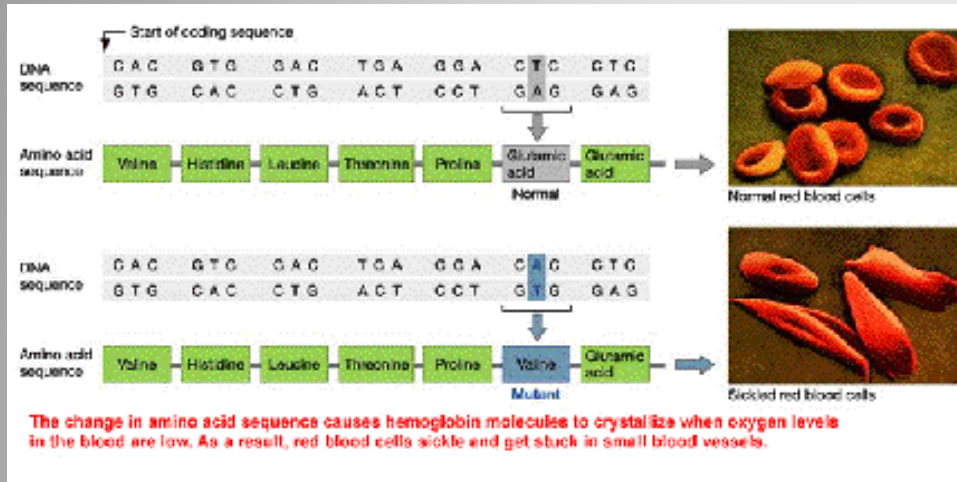
http://scienceblogs.com/digitalbio/2007/09/genetic_variation_i_what_is_a.php#more

- Different sources of DNA were used for original sequencing
- 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



- SNPs: the human genome has at least 38 million SNPs
most of these SNPs contribute to human variation
some of them may influence development of diseases, susceptibility to certain drugs, toxins, infectious agents
- 1.4 million short stretches of insertions or deletions
 - 14,000 large DNA deletions

Causes of Sickle Cell Anemia and Cystic Fibrosis have been pinpointed to specific mutations in their Protein-Coding DNA



CFTR Sequence:

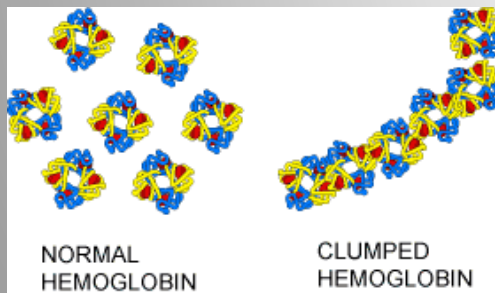
| | | | | | |
|------------|-----|-----|-----|-----|-----|
| Nucleotide | ATC | ATC | CTT | GGT | GTT |
| Amino Acid | Ile | Ile | Phe | Gly | Val |
| | 506 | | 508 | | 510 |

Deleted in $\Delta F508$

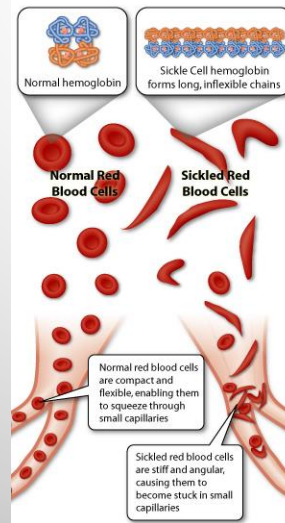
$\Delta F508$ CFTR Sequence:

| | | | | |
|------------|-----|-----|-----|-----|
| Nucleotide | ATC | ATT | GGT | GTT |
| Amino Acid | Ile | Ile | Gly | Val |
| | 506 | | | |

scienceandsociety.emory.edu



http://evolution.berkeley.edu/evolibrary/article/mutations_06



Cystic fibrosis patients have a **deletion of three base pairs** in the CFTR sequence. Protein folds incorrectly and is marked for degradation.

<http://learn.genetics.utah.edu/content/disorders/whataregd/sicklecell/>

Red blood cells carrying mutant hemoglobin are deprived of oxygen

The Race for the \$1,000 Genome

Human Genome Project (2001, initial draft):

> \$ 3 billion (includes development of technology)

“raw” expenses estimated at \$300 million

Rhesus macaque (2006)

\$ 22 million

By end of 2007:

\$ 1-2 million for full mammalian genome sequence

(Jim Watson using pyro-sequencing technology)

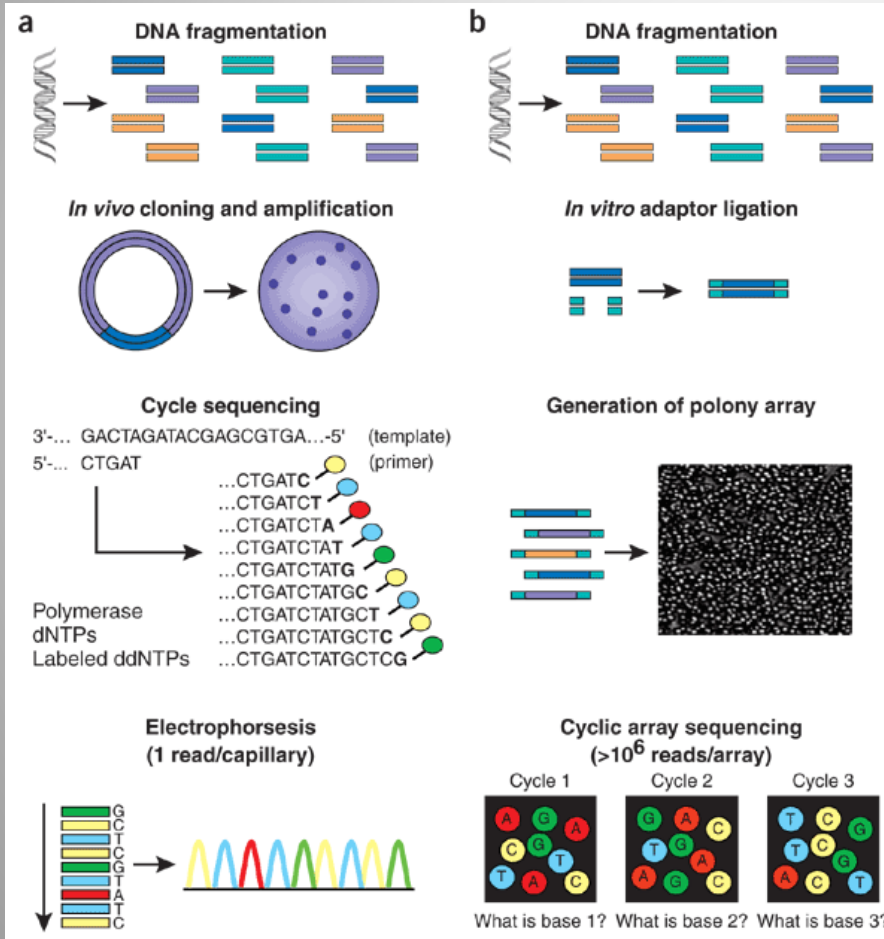
Wanted: “!!!!!! The \$ 1,000 Genome !!!!!!!!!!”

- low cost
- high-throughput
- high accuracy

Next Generation Sequencing (NGS) Focuses on Miniaturization and Parallelization

Sanger

Cyclic Array Sequencing (454 Pyrosequencing, Life Technologies SOLID Illumina Hi-Seq 2000)



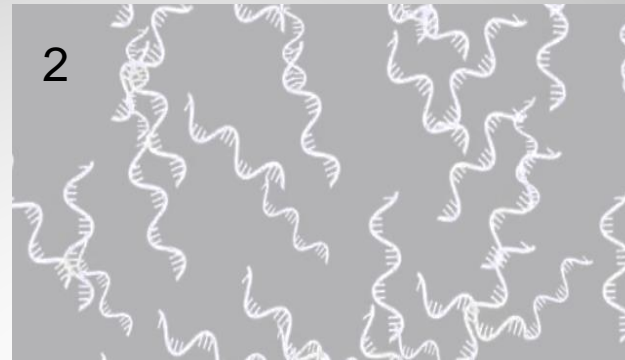
- Easier library preparation
- Fragment amplification on solid surfaces
- Nucleotide-to-nucleotide sequence determination
- Sequence from both ends

Generate dense planar array of DNA features

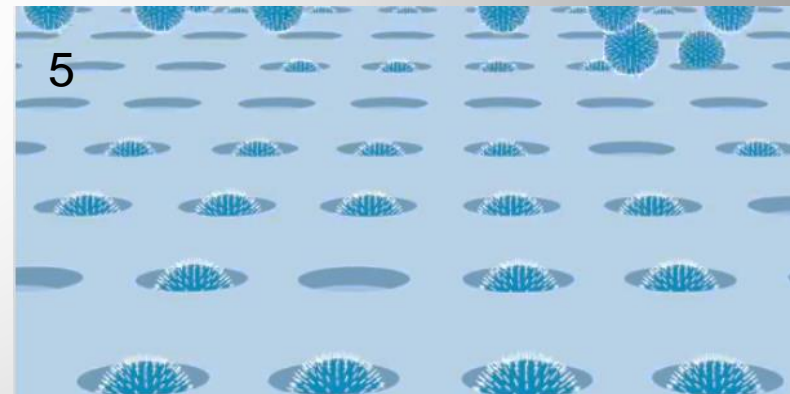
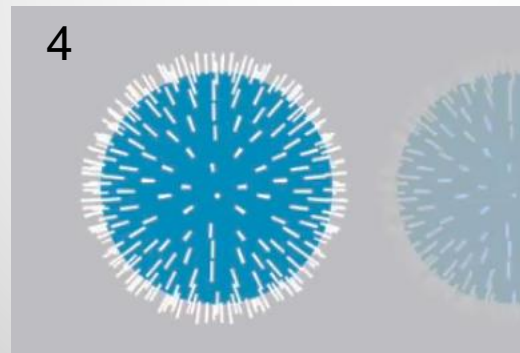
Apply cycles of enzymatic driven biochemistry

Imaging-based data collection

DNA Preparation for High Throughput Sequencing



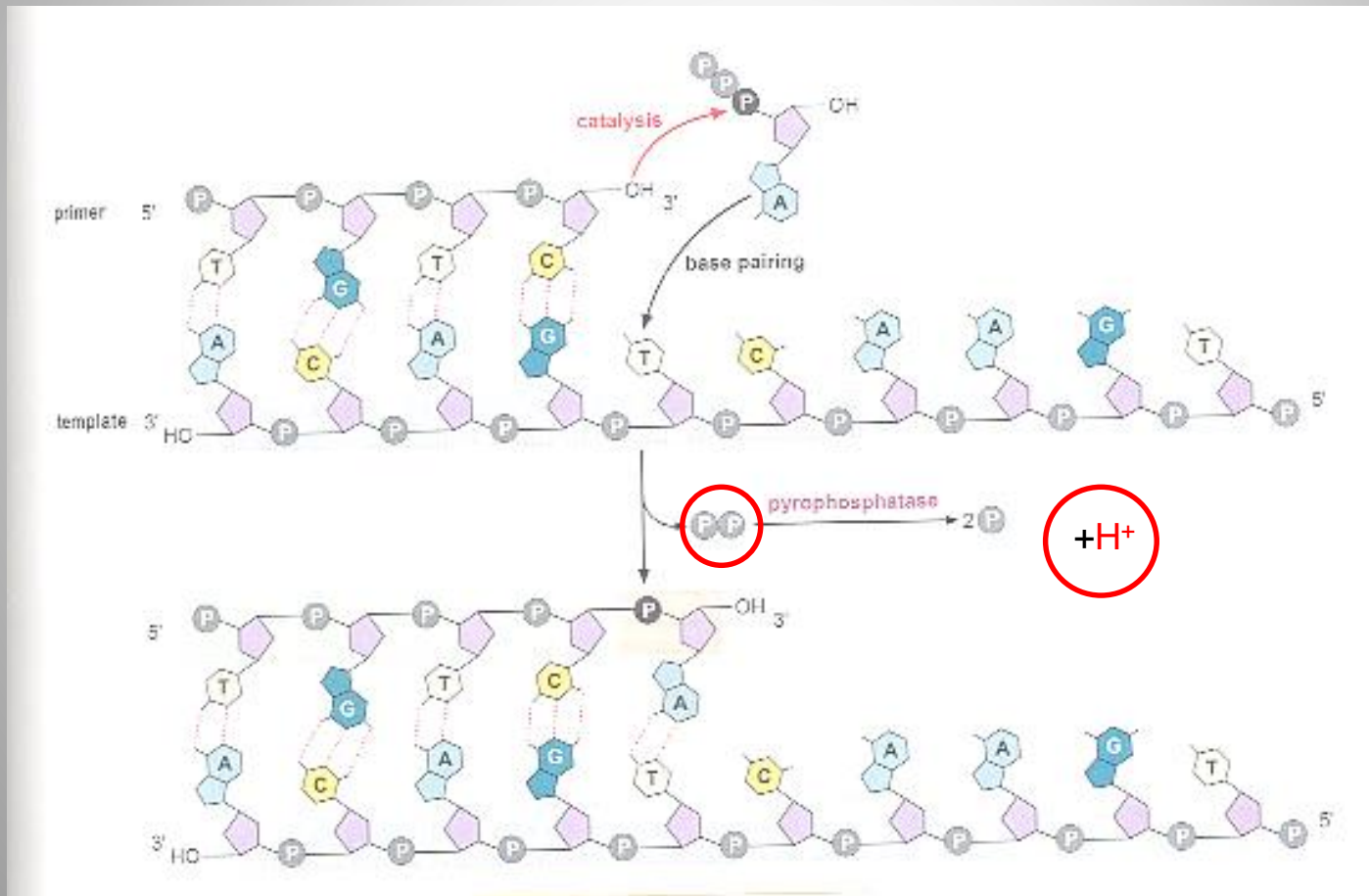
DNA is sheared into small fragments and each fragment is attached to a single bead



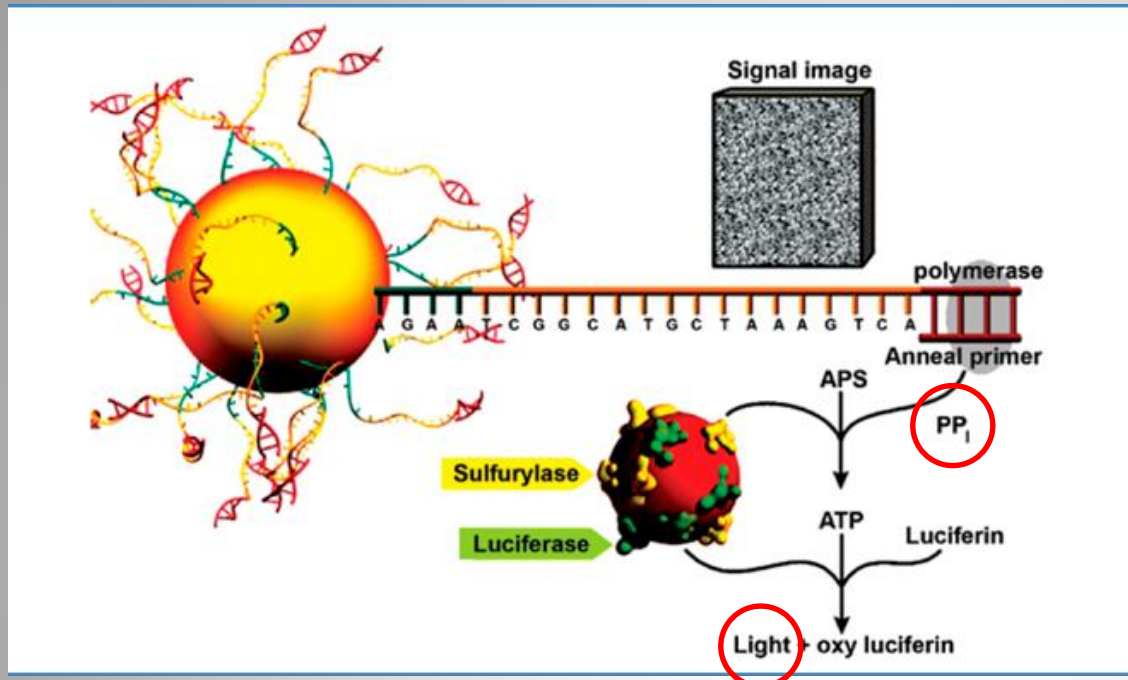
DNA fragments on beads are amplified and placed into individual reaction chambers; **Amplification** needed to achieve required signal strength.

Reaction Mechanism for DNA synthesis

The 3' hydroxyl group of the primer attacks the α -phosphoryl group of the incoming nucleotide thereby forming a phosphodiester bond (S_N2 reaction).



454 – Pyrosequencing



www.454.com

Add **one kind** of dNTP per cycle.

A nucleotide complementary to the template strand generates a light signal
The light signal is recorded by a CCD camera

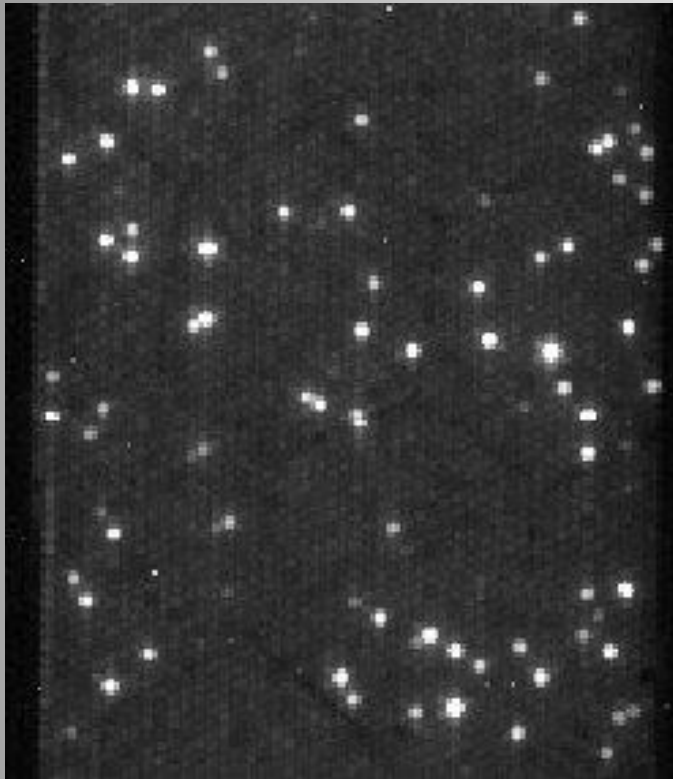
The signal strength is proportional to the number of nucleotides incorporated

Sequencing is recorded 'live'.
Read length up to **400** bases

~ 1.7 Million reactions in parallel

Diameter of single reaction chamber is approx. 44 μm

Image Capturing

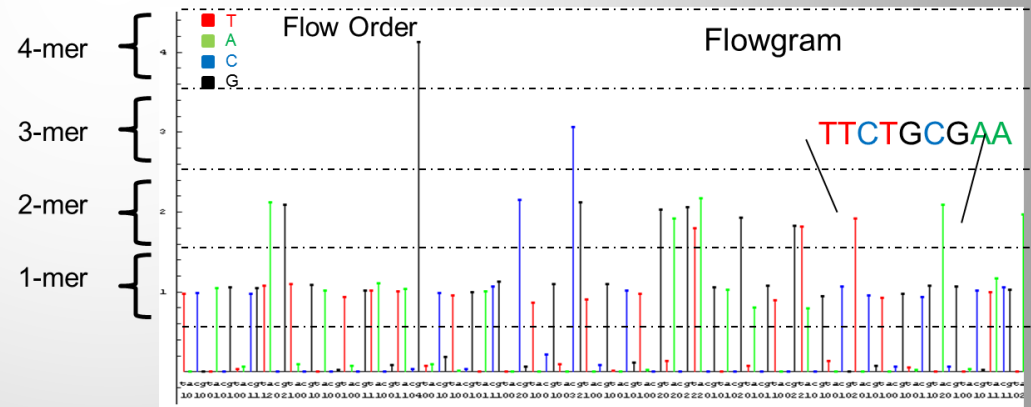


www.454.com

Example:
Addition of a C nucleotide

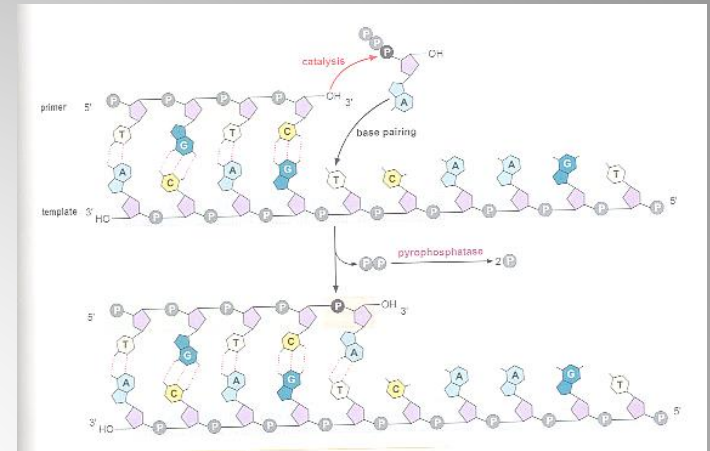
Each bright dot signals the incorporation of a C nucleotide at this position

Brighter dots mean more C nucleotides have been incorporated into the growing DNA chain



Flowgram: conversion of light intensity signals

Solexa/ Illumina Sequencing Technology



Illumina

reversible

Sanger



reversible

Use a set of deoxynucleotides that carry

- each a fluorescent label that can be cleaved off
- a reversibly terminating moiety at the 3' hydroxyl position

BGI – The Sequencing Factory

Beijing Genome Institute

Purchased 128 HiSeq2000 sequencers from Illumina in January 2010 each of which can produce **25 billion** base pairs of sequence a day



MASS PRODUCTION
Over the past decade, the BGI has gone from sequencing millions of base pairs (Mbp) to billions (Gbp) in ever shorter amounts of time. Below are some of the institute's previous achievements.

RICE GENOME
466 Mbp
Coverage: ~6x
Sequenced: April 2000-October 2001

SILKWORM GENOME
480 Mbp
Coverage: ~6x
Sequenced: June 2003-September 2003

FIRST ASIAN GENOME
3 Gbp
Coverage: ~36x
Sequenced: January 2007-October 2007

CUCUMBER GENOME
240 Mbp
Coverage: ~72x
Sequenced: January 2007-April 2009

PANDA GENOME
3 Gbp
Coverage: ~73x
Sequenced: March 2008-October 2008

ICEMAN GENOME
3 Gbp
Coverage: ~20x
Sequenced: May 2009-December 2009

Illumina Update (2013)

HiSeq X

A million dollar machine capable of sequencing 1800 human genomes
a year, i.e. 3-4 genomes a day

BUT: buyer must order 10 machines and agree only to use for
sequencing
Human genomes

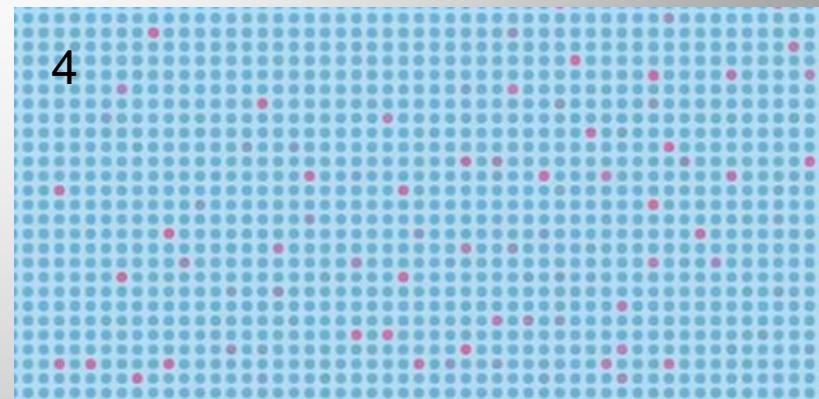
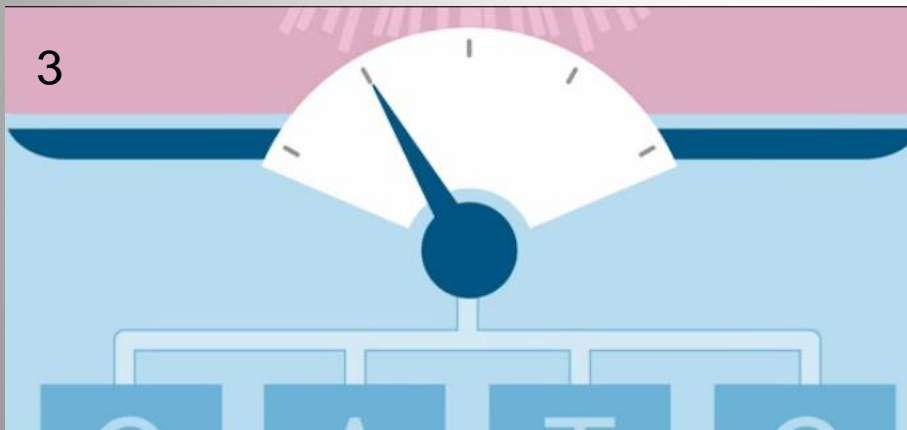
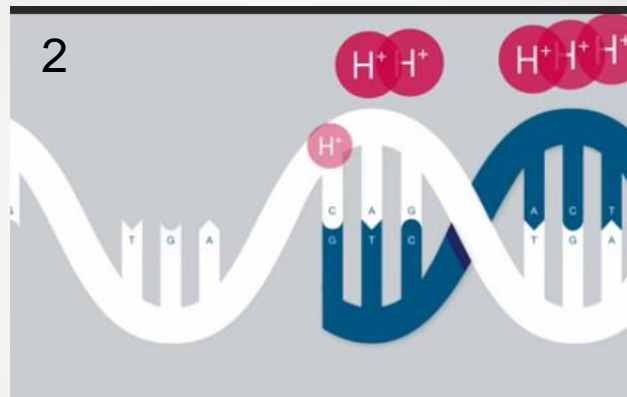
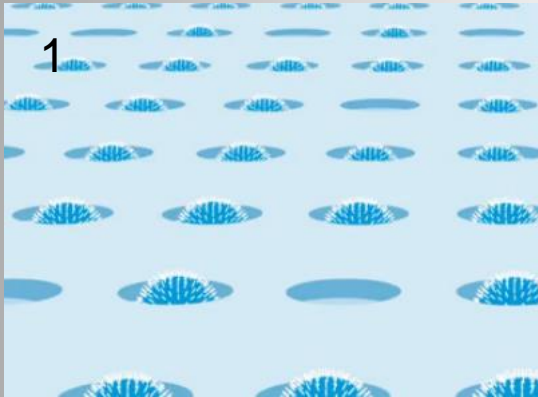
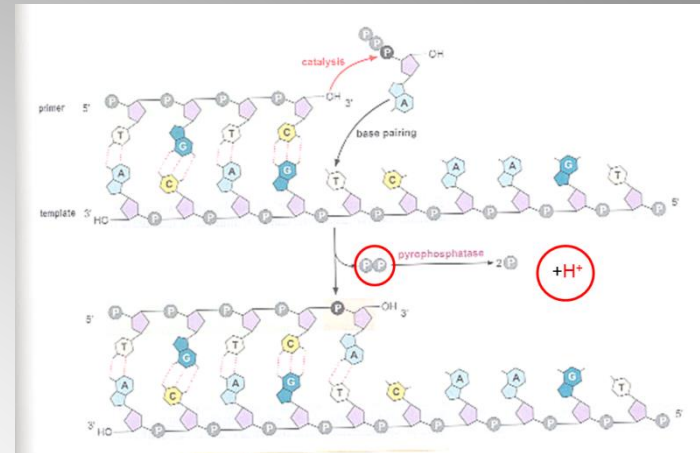
Cost will sink to \$ 1000 per genome

Ion Torrent Sequencing

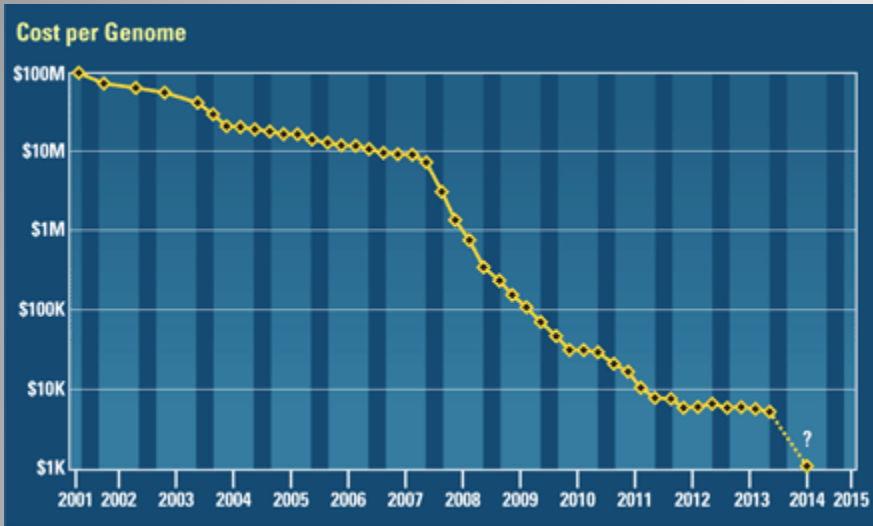
DNA to be sequenced is placed in wells of semiconductor chip

Add one nucleotide type per reaction cycle

When incorporated, a hydrogen ion is released and change in pH is recorded and converted into voltage



Oxford Nanopore Sequencing (MinION)



Targets \$ 1000 genome sequencing

Hand-held disposable sequencer

Pull a single strand DNA through nanopore

As base go through they interrupt an ionic current that reveals each base's identity

Should allow long reads without the need to piece together many short reads



Costs associated with Genome Sequencing

tracked by the National Human Genome Research Institute (NHGRI)

| | Date | Cost per Mb | Cost per Genome |
|----|--------|-------------|-----------------|
| 1 | | | |
| 2 | Sep-01 | \$5,292.39 | \$95,263,072 |
| 3 | Mar-02 | \$3,898.64 | \$70,175,437 |
| 4 | Sep-02 | \$3,413.80 | \$61,448,422 |
| 5 | Mar-03 | \$2,986.20 | \$53,751,684 |
| 6 | Oct-03 | \$2,230.98 | \$40,157,554 |
| 7 | Jan-04 | \$1,598.91 | \$28,780,376 |
| 8 | Apr-04 | \$1,135.70 | \$20,442,576 |
| 9 | Jul-04 | \$1,107.46 | \$19,934,346 |
| 10 | Oct-04 | \$1,028.85 | \$18,519,312 |
| 11 | Jan-05 | \$974.16 | \$17,534,970 |
| 12 | Apr-05 | \$897.76 | \$16,159,699 |
| 13 | Jul-05 | \$898.90 | \$16,180,224 |
| 14 | Oct-05 | \$766.73 | \$13,801,124 |
| 15 | Jan-06 | \$699.20 | \$12,585,659 |
| 16 | Apr-06 | \$651.81 | \$11,732,535 |
| 17 | Jul-06 | \$636.41 | \$11,455,315 |
| 18 | Oct-06 | \$581.92 | \$10,474,556 |
| 19 | Jan-07 | \$522.71 | \$9,408,739 |
| 20 | Apr-07 | \$502.61 | \$9,047,003 |
| 21 | Jul-07 | \$495.96 | \$8,927,342 |
| 22 | Oct-07 | \$397.09 | \$7,147,571 |
| 23 | Jan-08 | \$102.13 | \$3,063,820 |
| 24 | Apr-08 | \$15.03 | \$1,352,982 |
| 25 | Jul-08 | \$8.36 | \$752,080 |
| 26 | Oct-08 | \$3.81 | \$342,502 |
| 27 | Jan-09 | \$2.59 | \$232,735 |
| 28 | Apr-09 | \$1.72 | \$154,714 |
| 29 | Jul-09 | \$1.20 | \$108,065 |
| 30 | Oct-09 | \$0.78 | \$70,333 |
| 31 | Jan-10 | \$0.52 | \$46,774 |
| 32 | Apr-10 | \$0.35 | \$31,512 |
| 33 | Jul-10 | \$0.35 | \$31,125 |
| 34 | Oct-10 | \$0.32 | \$29,092 |
| 35 | Jan-11 | \$0.23 | \$20,963 |
| 36 | Apr-11 | \$0.19 | \$16,712 |
| 37 | Jul-11 | \$0.12 | \$10,497 |
| 38 | Oct-11 | \$0.09 | \$7,743 |
| 39 | Jan-12 | \$0.09 | \$7,666 |
| 40 | Apr-12 | \$0.07 | \$5,901 |
| 41 | Jul-12 | \$0.07 | \$5,985 |
| 42 | Oct-12 | \$0.07 | \$6,618 |
| 43 | Jan-13 | \$0.06 | \$5,671 |
| 44 | Apr-13 | \$0.06 | \$5,826 |
| 45 | Jul-13 | \$0.06 | \$5,550 |
| 46 | Oct-13 | \$0.06 | \$5,096 |
| 47 | Jan-14 | \$0.04 | \$4,008 |
| 48 | Apr-14 | \$0.05 | \$4,920 |

<http://www.genome.gov/sequencingcosts/>

Single DNA Strand Sequencing (Pacific Biosciences) – 3rd Generation Sequencing

IN A FLASH

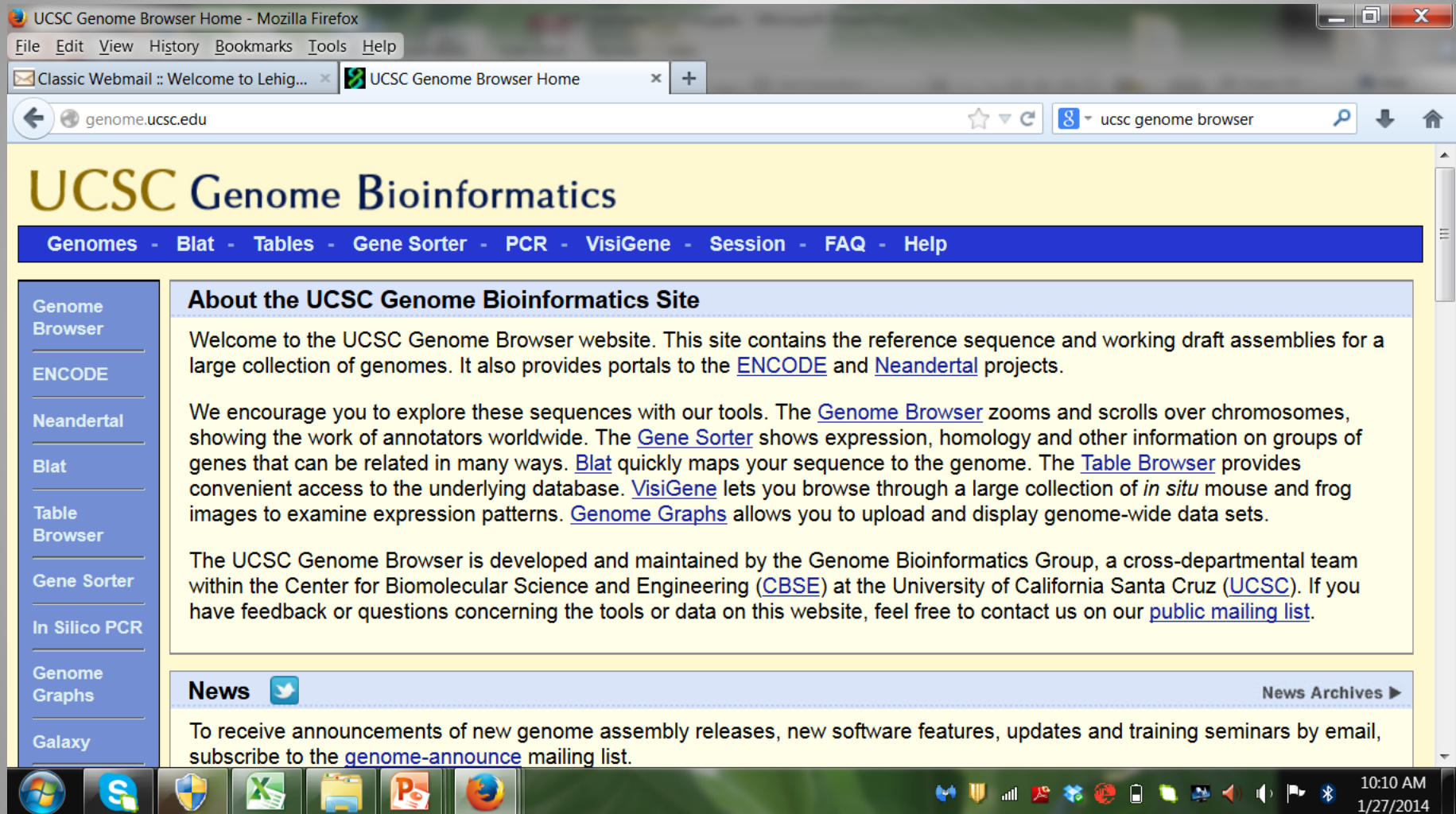
New DNA sequencers watch an enzyme called DNA polymerase as it uses fluorescently tagged bases to synthesize DNA. Each base is identified by a distinguishing colour that flashes as the base is incorporated into the DNA strand.



Reportedly generating sequences an average of 1,500 bp long

Examples of Projects and 'Spinoffs' derived from the HGP:

UCSC Genome Browser collects genome sequences



The screenshot shows a Mozilla Firefox browser window displaying the UCSC Genome Browser homepage. The address bar shows the URL genome.ucsc.edu. The page title is "UCSC Genome Bioinformatics". A navigation menu includes links for Genomes, Blat, Tables, Gene Sorter, PCR, VisiGene, Session, FAQ, and Help. A sidebar on the left lists various tools: Genome Browser, ENCODE, Neandertal, Blat, Table Browser, Gene Sorter, In Silico PCR, Genome Graphs, and Galaxy. The main content area features a section titled "About the UCSC Genome Bioinformatics Site" with the following text:

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the [ENCODE](#) and [Neandertal](#) projects.

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over chromosomes, showing the work of annotators worldwide. The [Gene Sorter](#) shows expression, homology and other information on groups of genes that can be related in many ways. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to the underlying database. [VisiGene](#) lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. [Genome Graphs](#) allows you to upload and display genome-wide data sets.

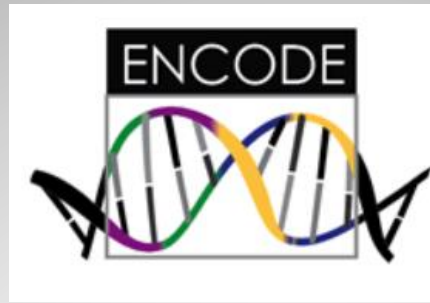
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering ([CBSE](#)) at the University of California Santa Cruz ([UCSC](#)). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our [public mailing list](#).

Below the "About" section is a "News" section with a Twitter icon and a "News Archives" link. The text reads: "To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the [genome-announce](#) mailing list."

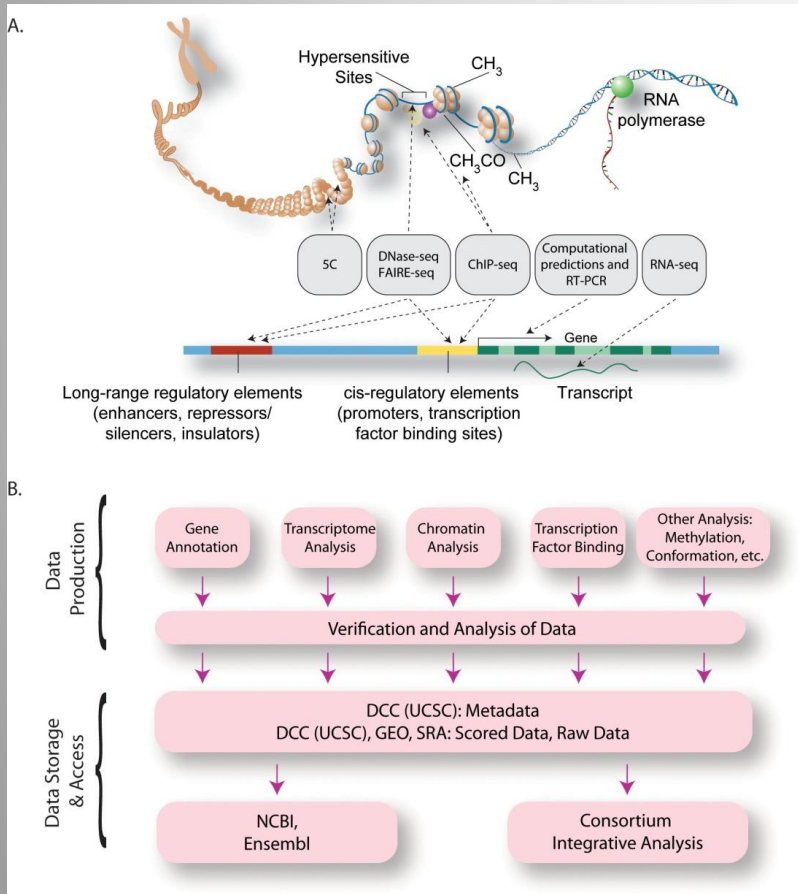
The Windows taskbar at the bottom shows the system tray with the time 10:10 AM and date 1/27/2014, along with various application icons.

<http://genome.ucsc.edu/>

ENCODE



6. September 2012



Build a comprehensive parts list of **functional elements** in the human genome, including elements that act at the **protein** and **RNA** levels, and **regulatory** elements that control cells and circumstances in which a gene is active

More than 80% of the human genome has at least one biochemical activity
Many regulatory regions transcribed into RNA with regulatory function

1000 Genomes Project



Launched in 2008

Catalog of human genetic variation

Genomic knowledge will contribute to the fields of genetics, medicine, pharmacology, biochemistry, and bioinformatics

An integrated map of genetic variation from 1,092 human genomes

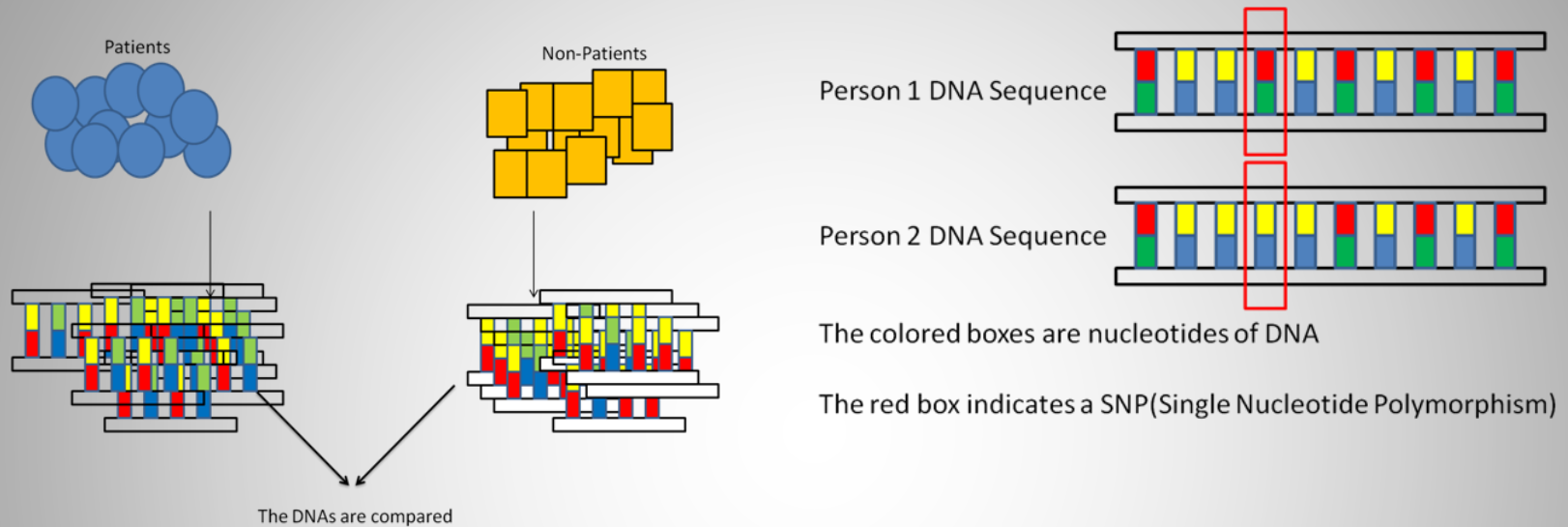
The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

Nature, Nov 1st, 2012, Vol 491, pg 56

GWAS – Genome Wide Association Study

What is GWAS?

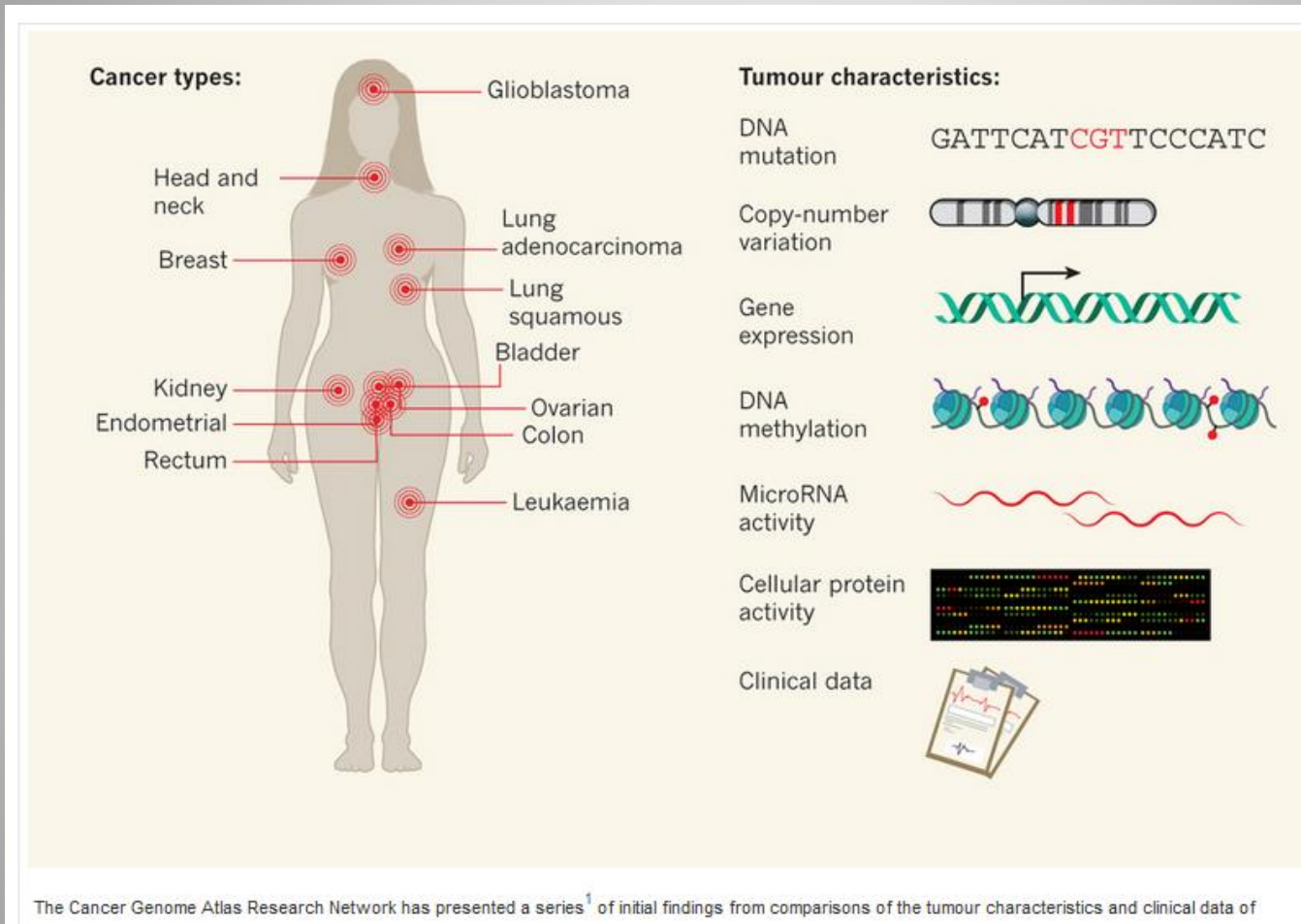


Examination of many common genetic variants associated with a trait

Association between single nucleotide polymorphisms (SNPs) and onset of major diseases, such as diabetes or heart disease, compared to healthy individuals.

Open access repository

The Cancer Genome Atlas (TCGA)



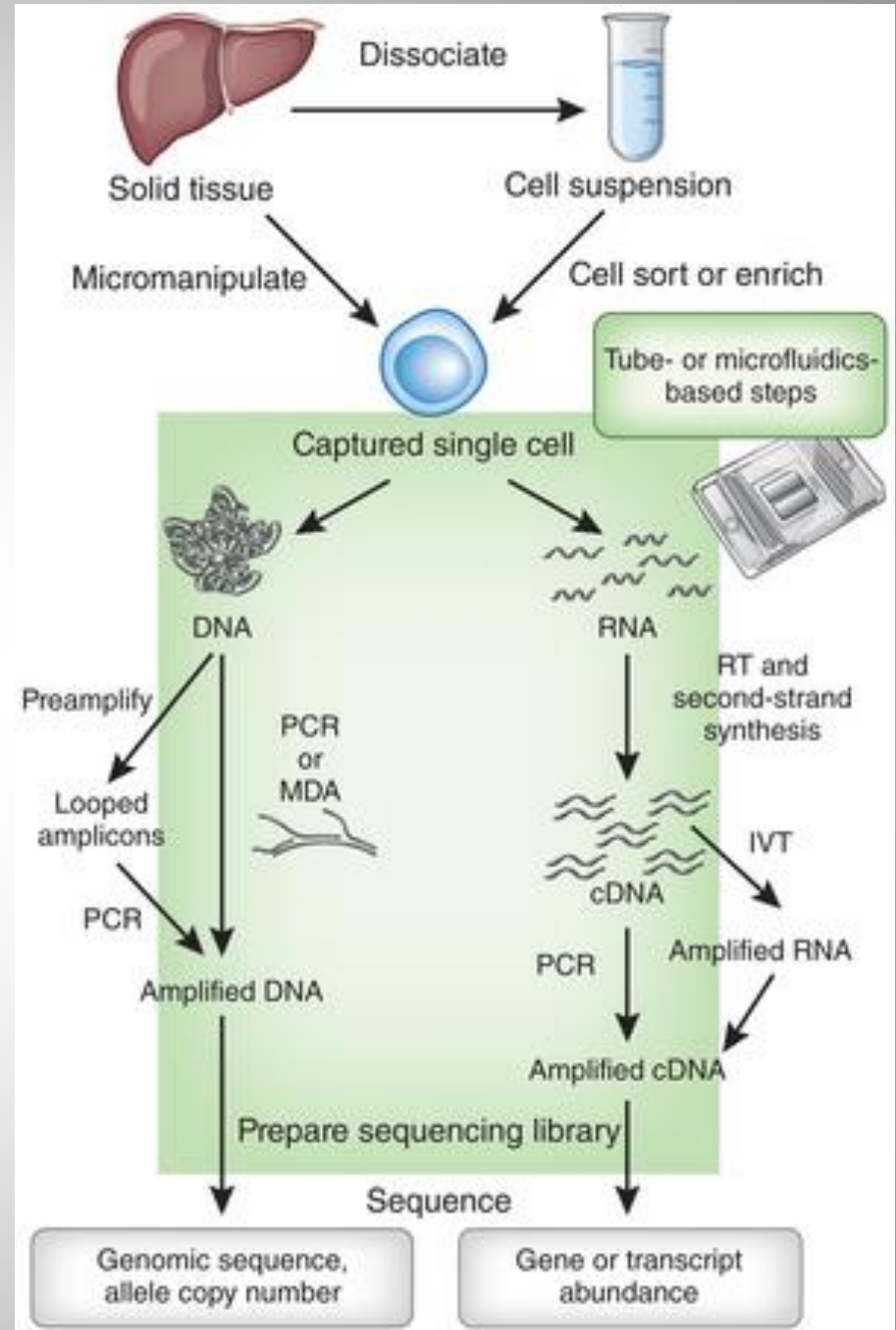
Reveal commonalities between cancer types, investigate molecular abnormalities, and define mutations that are confined to specific tumors
Development of prognostic, diagnostic and therapeutic strategies

Single Cell Sequencing – Method of the Year 2013

Single-cell sequencing can enable the discovery of clonal mutations, cryptic cell types or transcriptional features that would be diluted or averaged out in bulk tissue

Rare cell types
heterogeneous samples,
phenotypes associated with mosaicism or
variability

Technologies for single-cell amplification and sequencing are maturing. As the cost and ease of examining individual cells improves, the approach will enter the hands of more researchers as a standard tool for understanding biology at high resolution.



Summary_1

Sanger sequencing provided the basis to initiate and complete the Human Genome Project and many other genomes

Next Generation Sequencing (NGS) with high throughput and less costs took its place for genome sequencing (Pyrosequencing, Illumina, Ion Torrent)

2000: 4 eukaryotic genomes

2012: > 250 eukaryotic genomes, > 1000 human genomes

The various methods of sequencing have their advantages and disadvantages (e.g. cost, accuracy, fragment length, etc.)

Summary_2

- Data sharing (interdisciplinary approach between biology, engineering, and bioinformatics)
- Fewer protein-coding genes than thought
- Many regulatory RNA elements
- Studying of genetic variation and evolutionary origins
- Obtaining a genome 'blueprint' is not sufficient to explain the occurrence or susceptibility to disease, the combination of many factors needs to be taken into account
- The combination of various approaches (GWAS, 1000 Genomes, ENCODE, TCGA) and others will provide leads to the origin and treatment of human diseases.

