
# Sequencing Technologies

**Dr. Fayyaz ul Amir Afsar Minhas**
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

# Sequencing

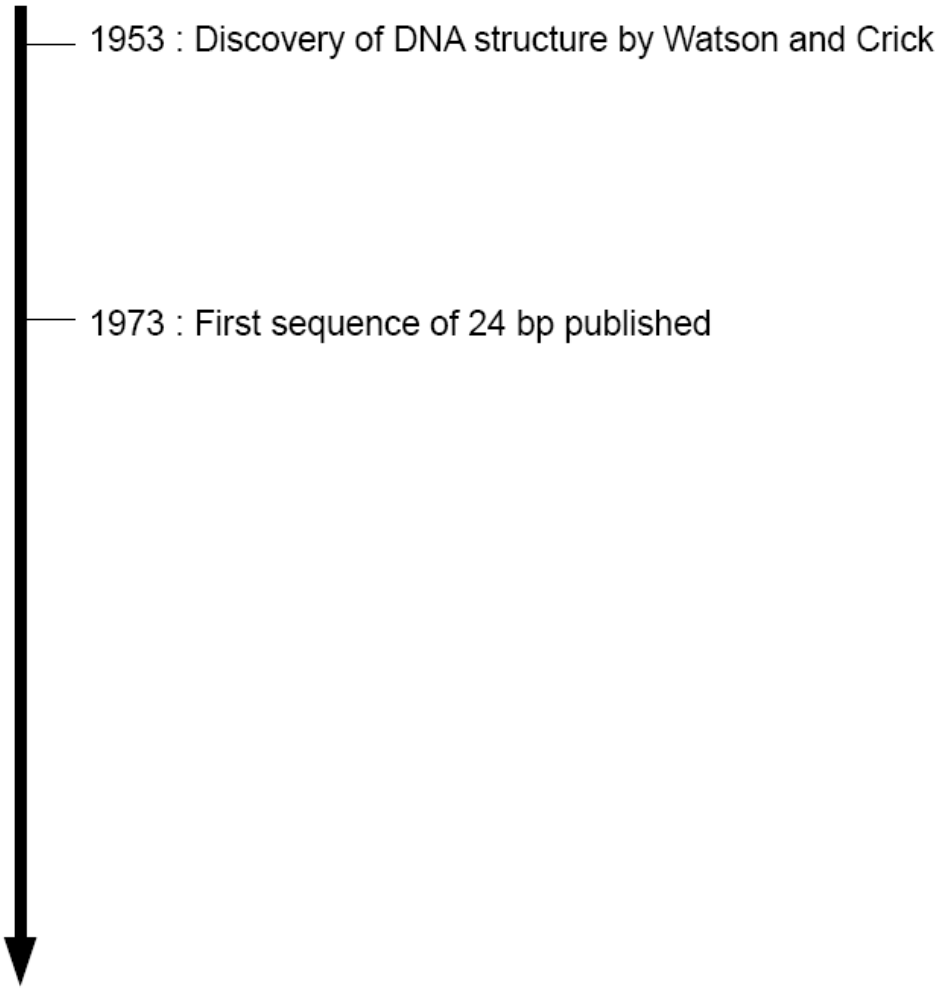
- Determine the primary structure of biopolymers
  - DNA (Genomics)
    - What's it made of?
    - DNA Sequencing
      - 3 Generations of sequencing technologies
  - RNAs (Transcriptomics)
    - What's going on?
    - RNA-Seq
  - Proteins (Proteomics)
    - Who's doing it?
      - Edman degradation
      - Mass-spectrometry
      - Peptide mass-fingerprinting
      - From DNA/RNA sequences

# DNA Sequencing

- Given
  - DNA Sample
    - Whole Genome
    - Short length
- Find
  - Its nucleotide sequence

- Types
  - Whole Genome Sequencing
  - Short-length sequencing
    - Targeted. Exome Sequencing (Expressed Sequence Tags, …)
  - De novo sequencing / Assembly
  - Mapping to reference genomes
    - Single nucleotide polymorphism (SNP) calling
    - RNA Short-read mapping (RNA-Seq)

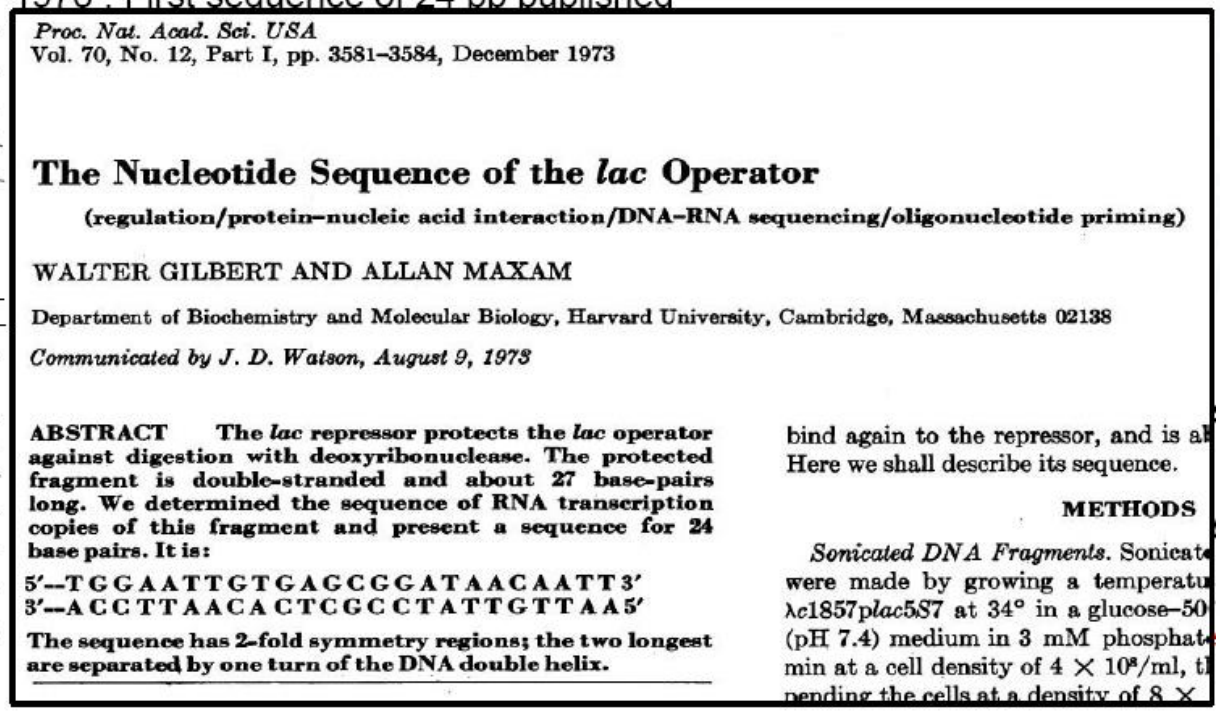
# Sequencing: History, State of the art and Future!



# Sequencing: History, State of the art and Future!

1953 : Discovery of DNA structure by Watson and Crick

1973 : First sequence of 24 bp published



*Proc. Nat. Acad. Sci. USA*
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

**The Nucleotide Sequence of the *lac* Operator**

**(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)**

WALTER GILBERT AND ALLAN MAXAM

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

*Communicated by J. D. Watson, August 9, 1973*

**ABSTRACT** **The *lac* repressor protects the *lac* operator against digestion with deoxyribonuclease. The protected fragment is double-stranded and about 27 base-pairs long. We determined the sequence of RNA transcription copies of this fragment and present a sequence for 24 base pairs. It is:**

**5′—T G G A A T T G T G A G C G G A T A A C A A T T 3′**
**3′—A C C T T A A C A C T C G C C T A T T G T T A A 5′**

**The sequence has 2-fold symmetry regions; the two longest are separated by one turn of the DNA double helix.**

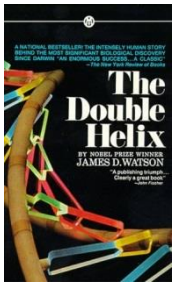
bind again to the repressor, and is ab
Here we shall describe its sequence.

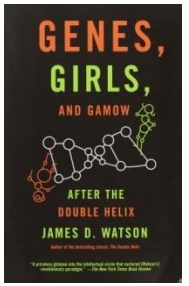
**METHODS**

*Sonicated DNA Fragments.* Sonicat
were made by growing a temperatu
λc1857p*lac*5S7 at 34° in a glucose–50
(pH 7.4) medium in 3 mM phosphat
min at a cell density of 4 × $10^8$/ml, t
pending the cells at a density of 8 ×

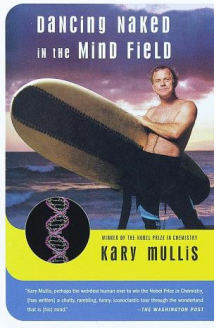
# Sequencing: History, State of the art and Future!

1953 : Discovery of DNA structure by Watson and Crick




1973 : First sequence of 24 bp published
1977 : Sanger sequencing method published
1980 : Nobel Prize Wally Gilbert and Fred Sanger
1982 : Genbank started
1983 : Development of PCR
1987 : 1st automated sequencer : Applied Biosystems Prism 373

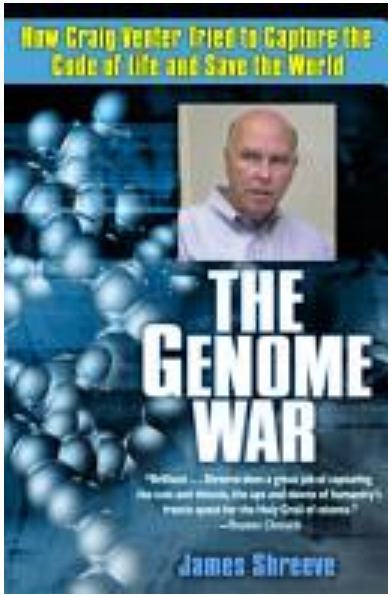



Karry Mullis: Invents PCR – a method for creating copies of a DNA molecule, used extensively in sequencing

1996 : Capillary sequencer : ABI 310
1998 : Genome of Caenorhabditis elegans sequenced
2000 : Human genome sequenced
2005 : 1st 454 Life Sciences Next Generation Sequencing system : GS 20 System
2006 : 1st Solexa Next Generation Sequencer : Genome Analyzer
2007 : 1st Applied Biosystems Next Generation Sequencer : SOLiD
2009 : 1st Helicos single molecule sequencer : Helicos Genetic Analyser System
2011 : 1st Ion Torrent Next Generation Sequencer : PGM
2011 : 1st Pacific Biosciences single molecule sequencer : PacBio RS Systems
2012 : Oxford Nanopore Technologies demonstrates ultra long single molecule reads

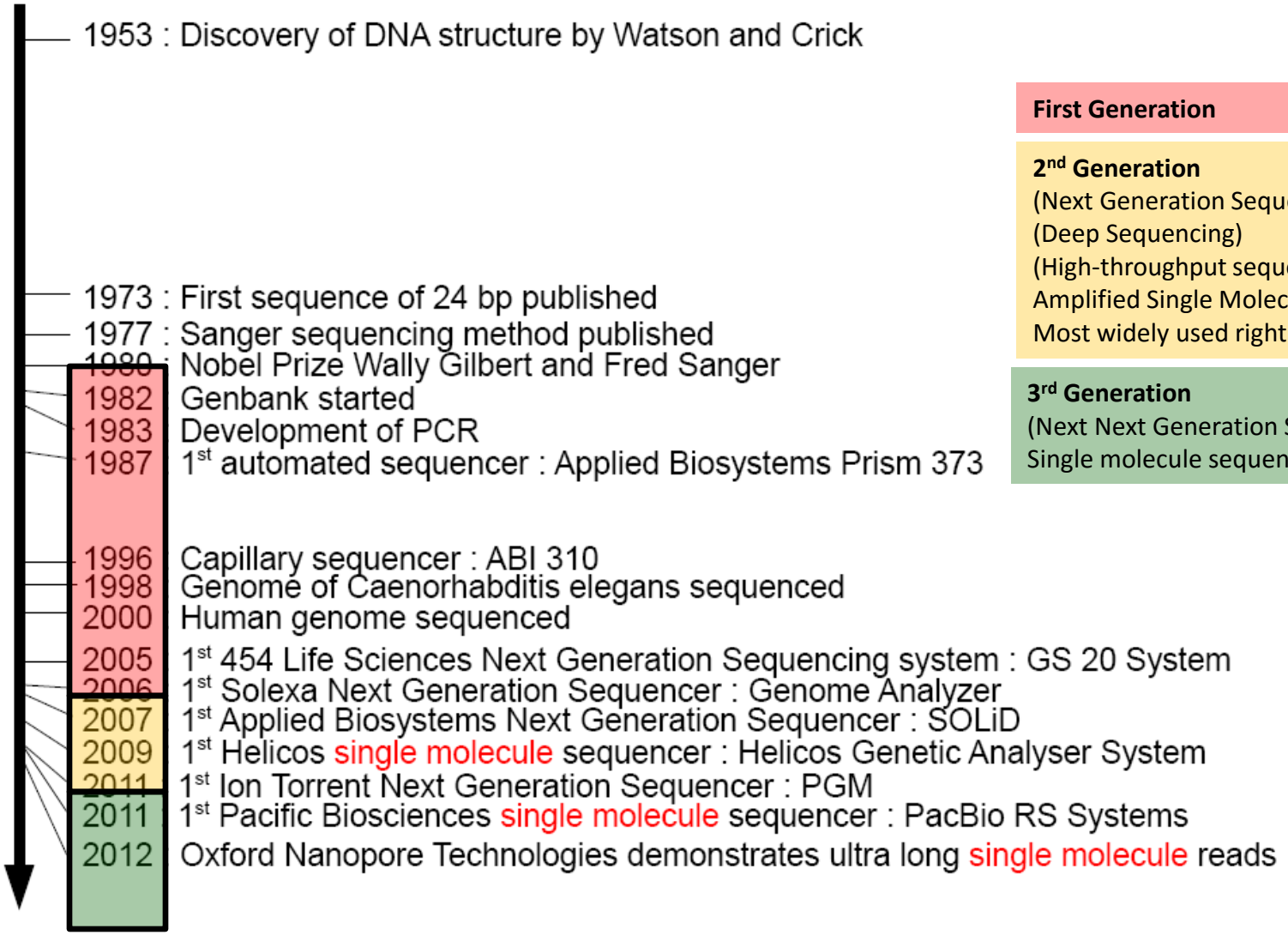
# Human Genome Project

- Started in 1990
- Objective: Sequence the human genome by 2005
- Achieved: 2000
  - Government consoritium
    - Cost: $3 Billion
  - Craig Venter's Celera / Solexa




- $1000 genome project
- 1000 genomes project

# Sequencing: History, State of the art and Future!



# Sequencing Technologies

- Sanger Sequencing

- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
- Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton

- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION

**First Generation**

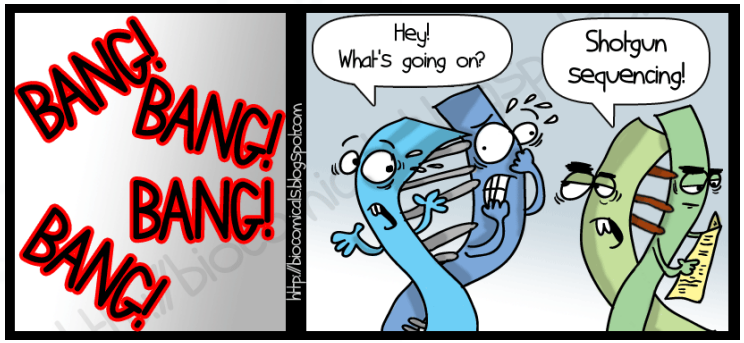
**2nd Generation**
(Next Generation Sequencing, NGS)
(Deep Sequencing)
(High-throughput sequencing)
Amplified Single Molecule Sequencing
Most widely used right now

**3rd Generation**
(Next Next Generation Sequencing)
Single molecule sequencing

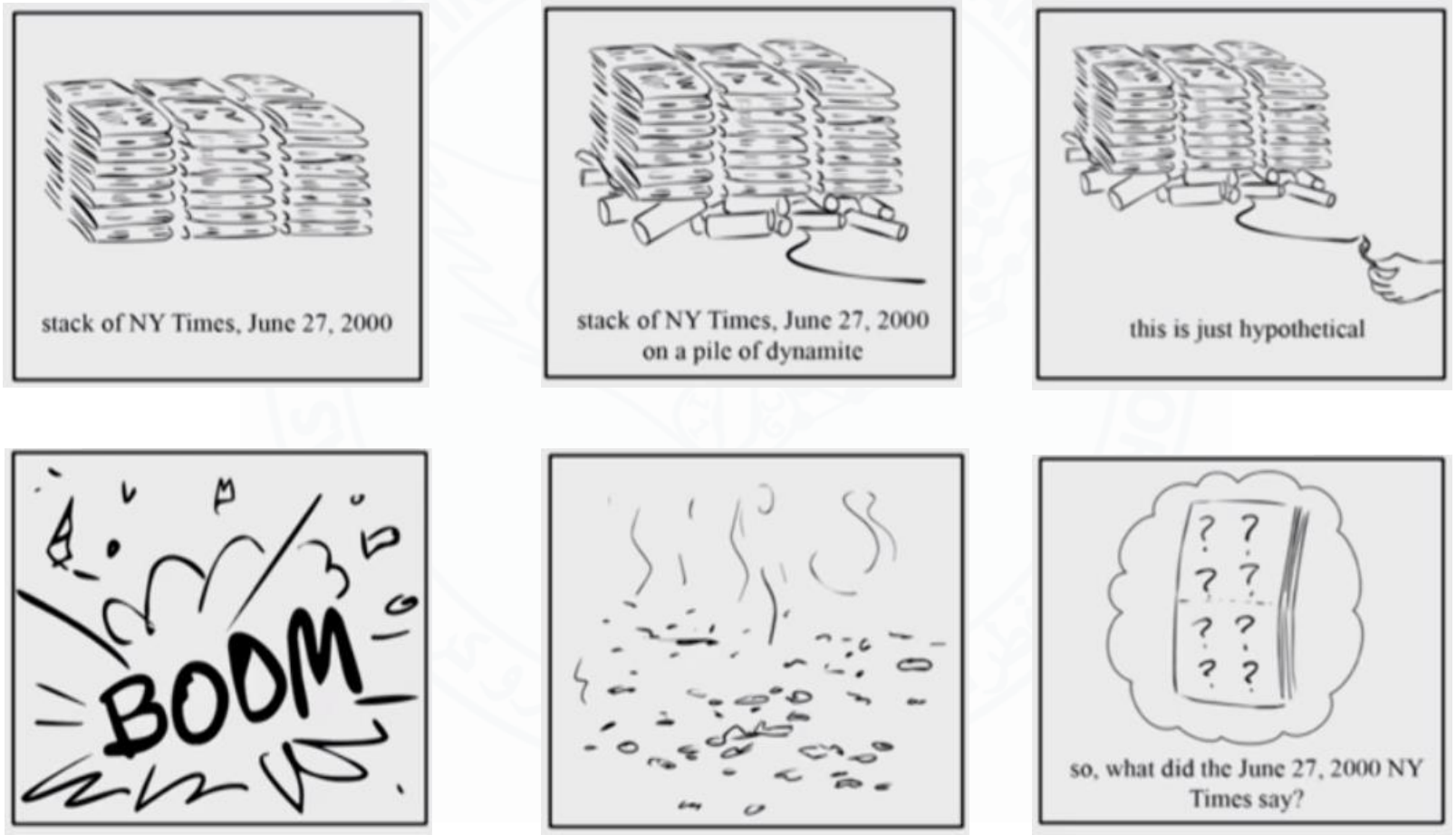
**HiSeq 2000**

# Steps in Sequencing

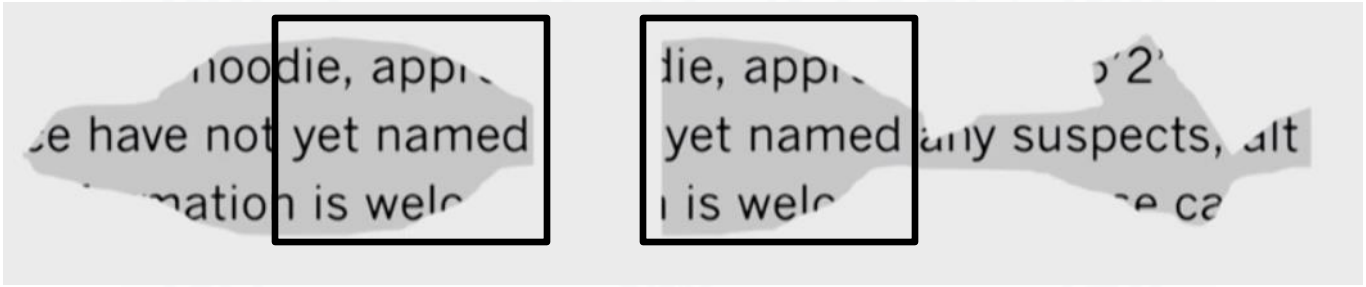
- DNA Extraction
- Preprocessing (Amplification , …)
- Sequencing
  - Shotgun sequencing
    - Reads
  - Assembly
- Data analysis

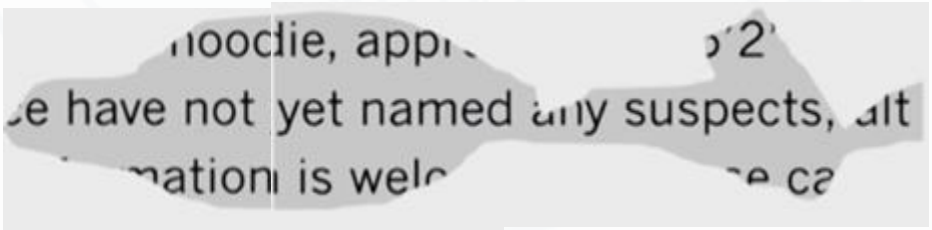



# Shotgun Sequencing: The case of exploding newspapers



# Joining overlapping reads



# Completing the overlap puzzle

# Sequencing and newspaper explosions: 1

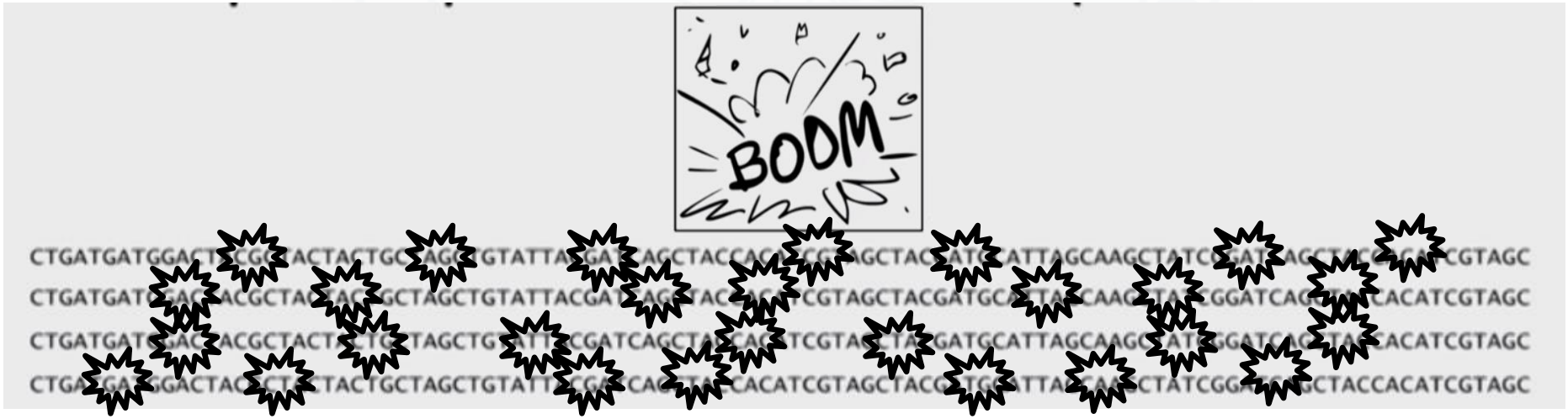
- Take (millions of) copies of the DNA you want to sequence

stack of NY Times, June 27, 2000

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

# Sequencing and newspaper explosions: 2

- Fragment the DNA into smaller pieces
  - Because our sequencing technologies can only read very short fragments reliably

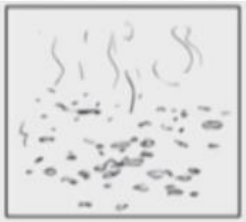


# Sequencing and newspaper explosions: 3

- The short fragments resulting from DNA fragmentation are called reads

CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC
CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC
CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

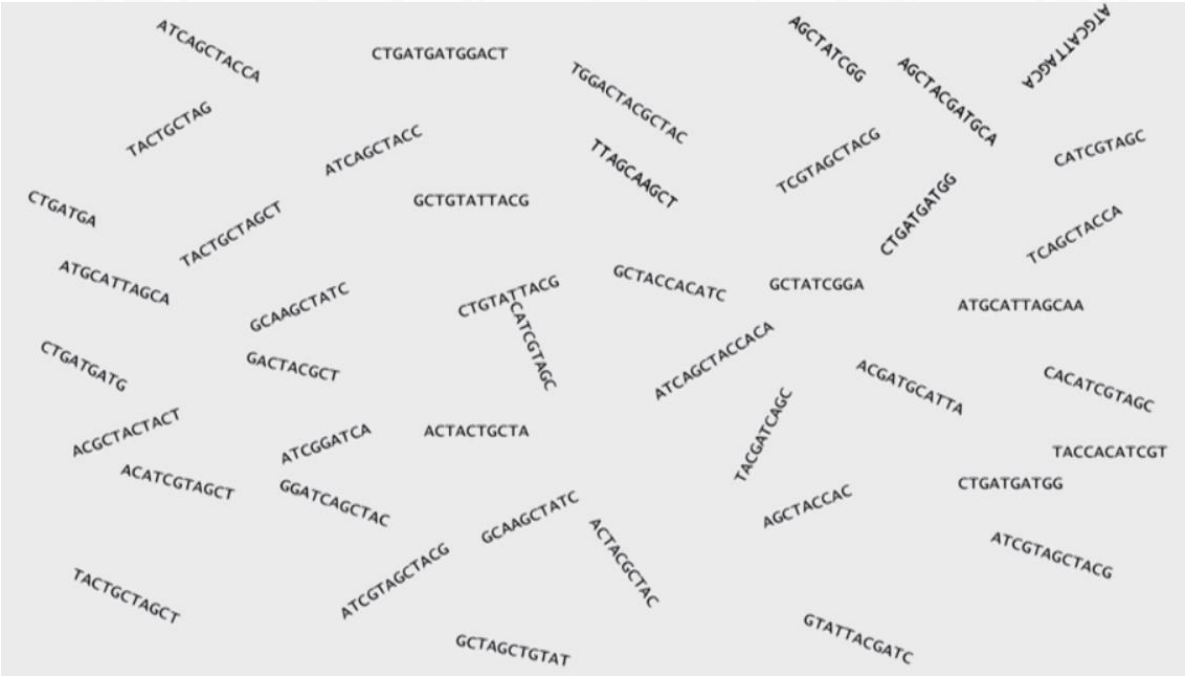
- Some reads get disappear



CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC
CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC
CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

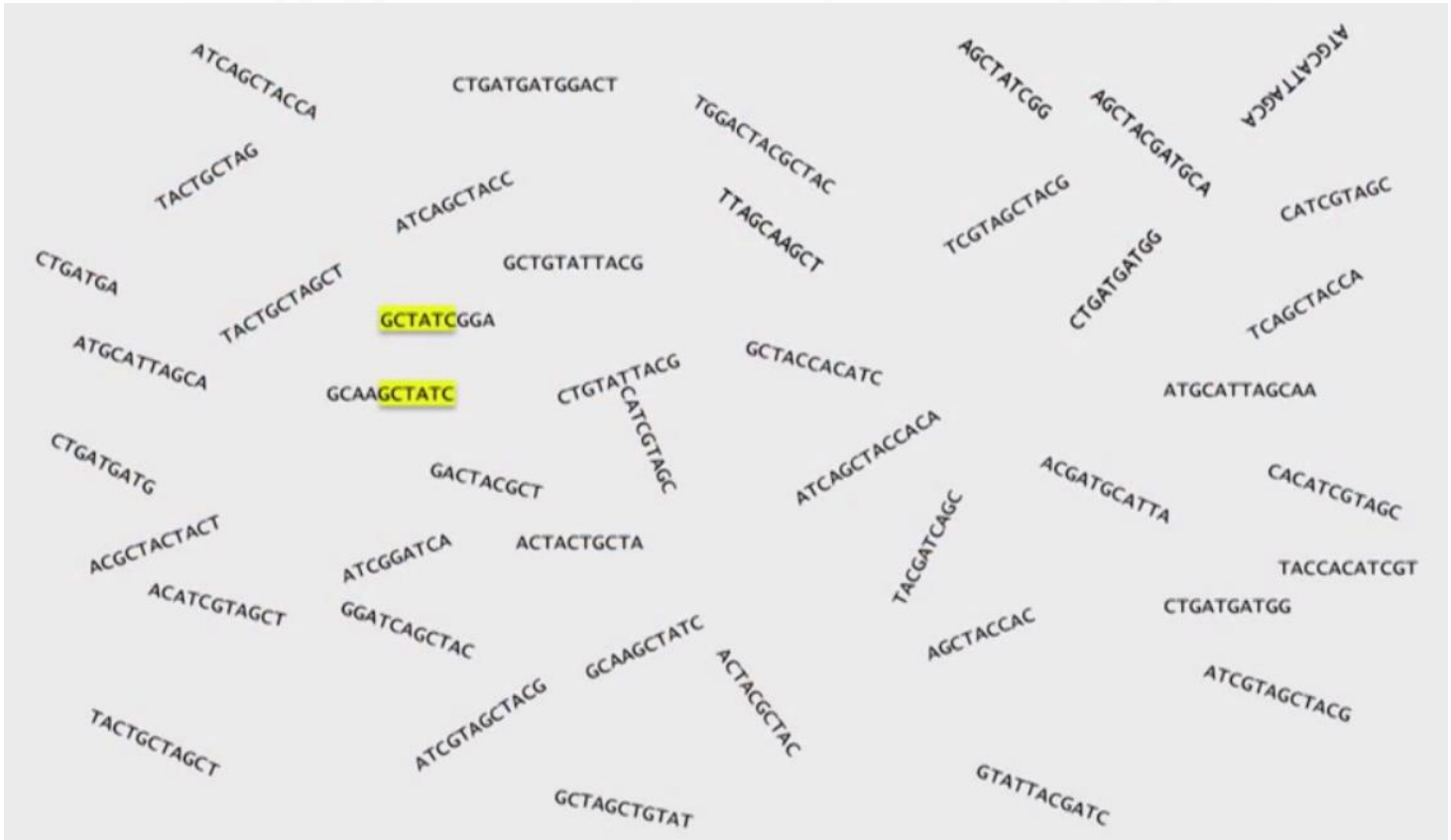
# Sequencing and newspaper explosions: 3

- We get the reads but we have no idea where they came from in the DNA
  - No position information
  - Need to reconstruct the DNA sequence



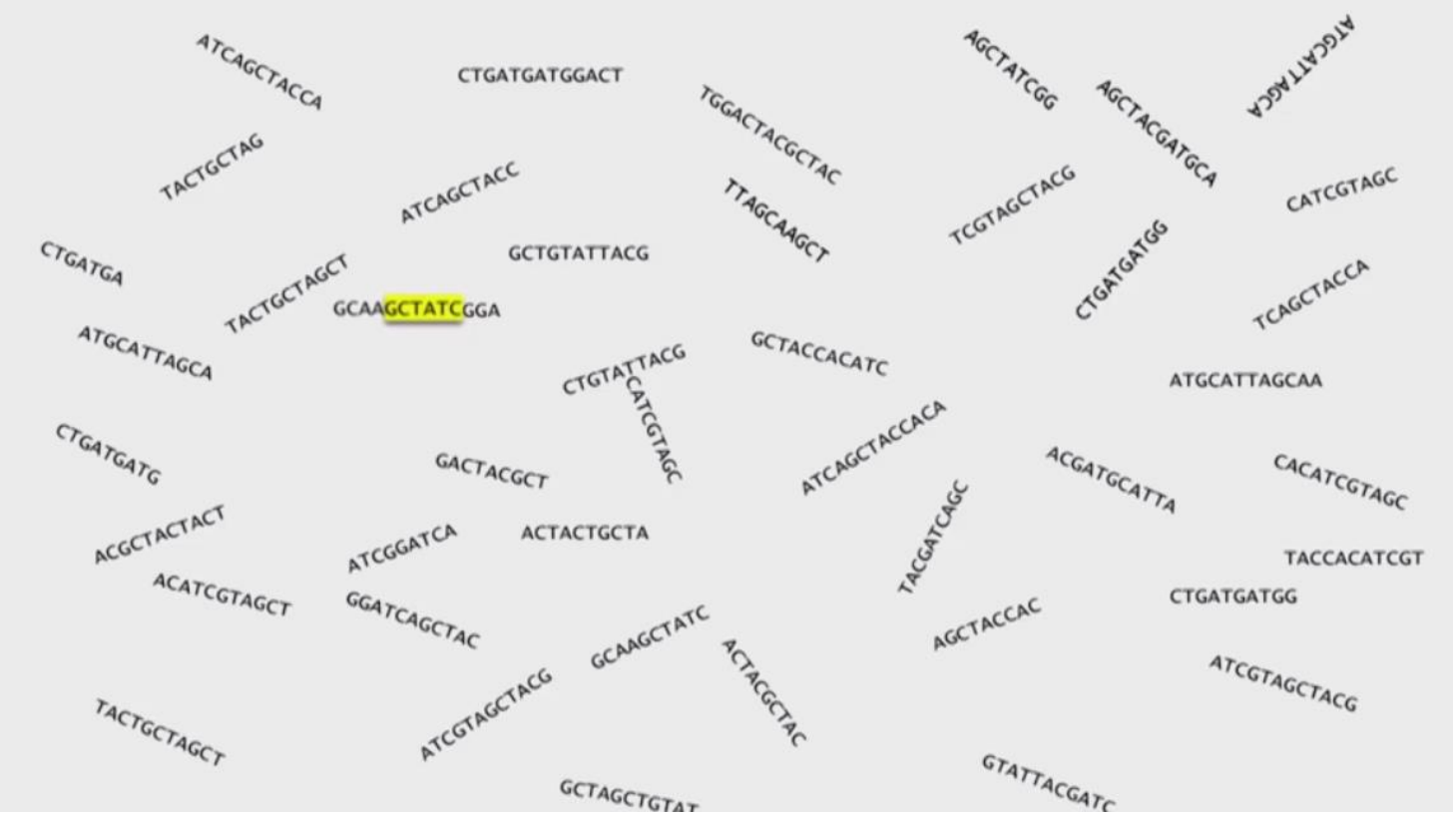
# Sequencing and newspaper explosions: 4

- Solve it as an overlap puzzle

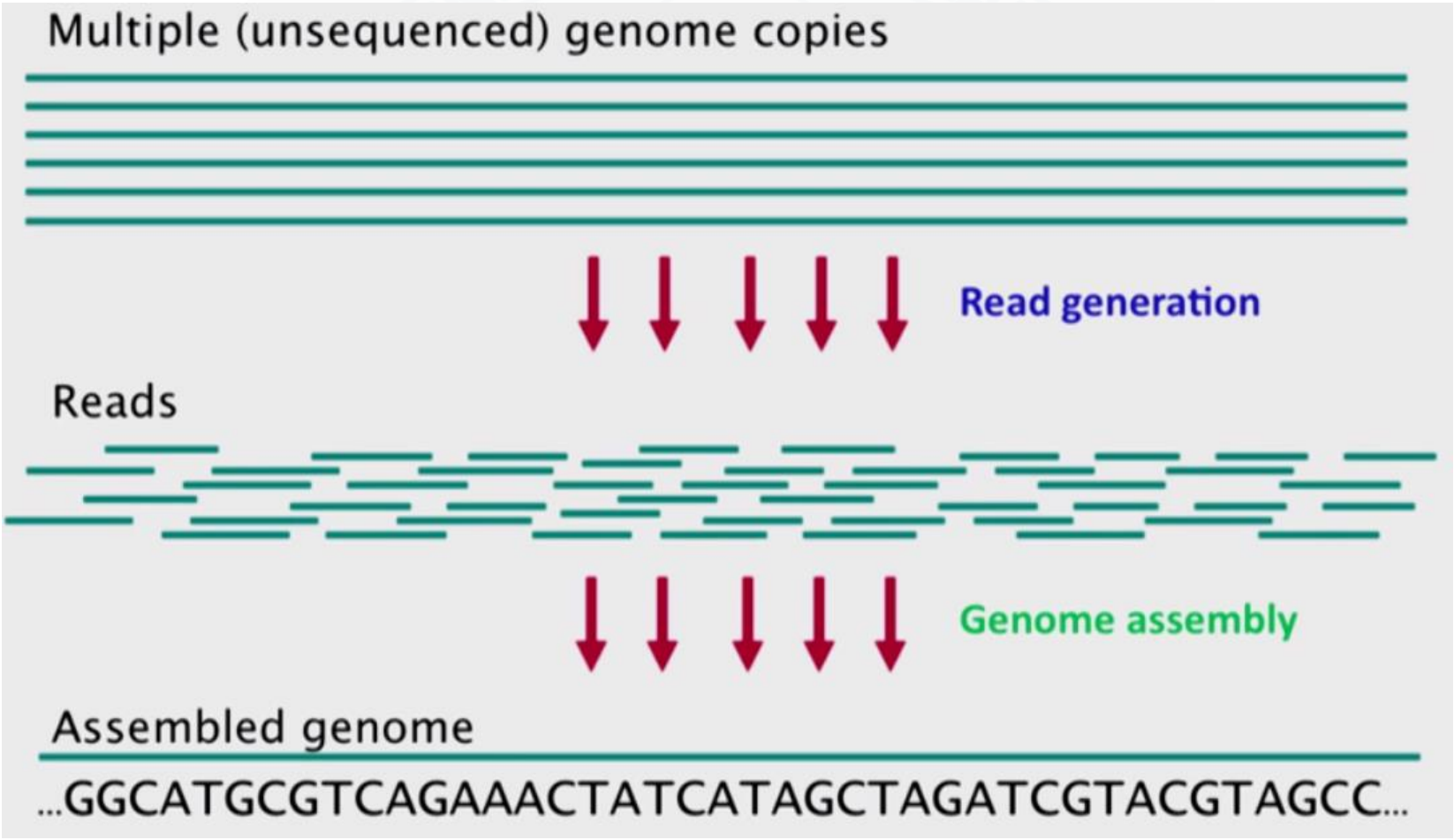


# Sequencing and newspaper explosions: 5

- Reconcile the pieces



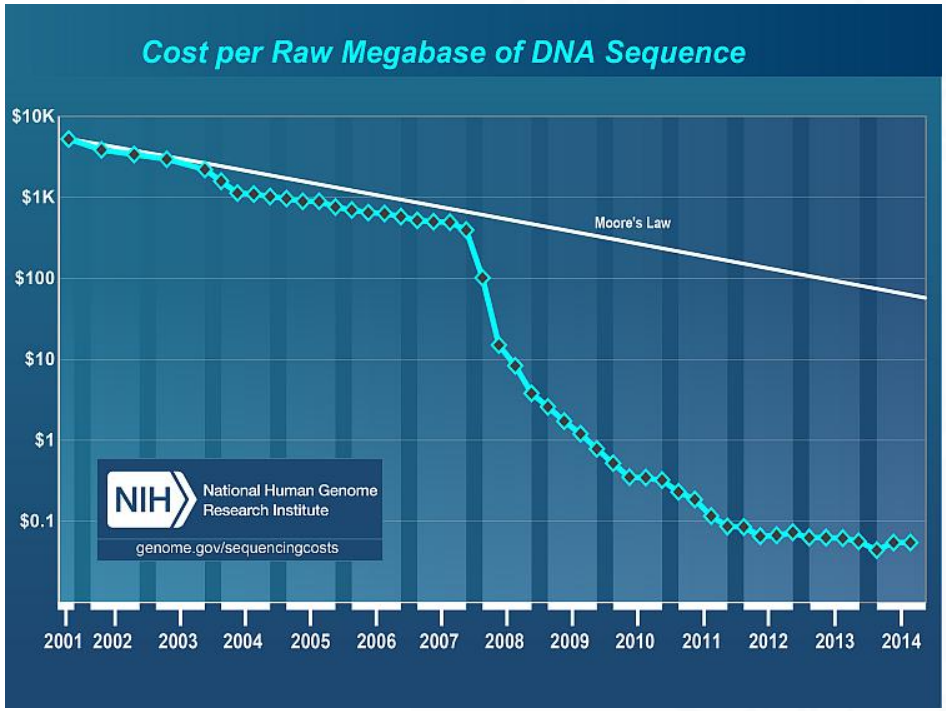
# Sequencing as a computational problem

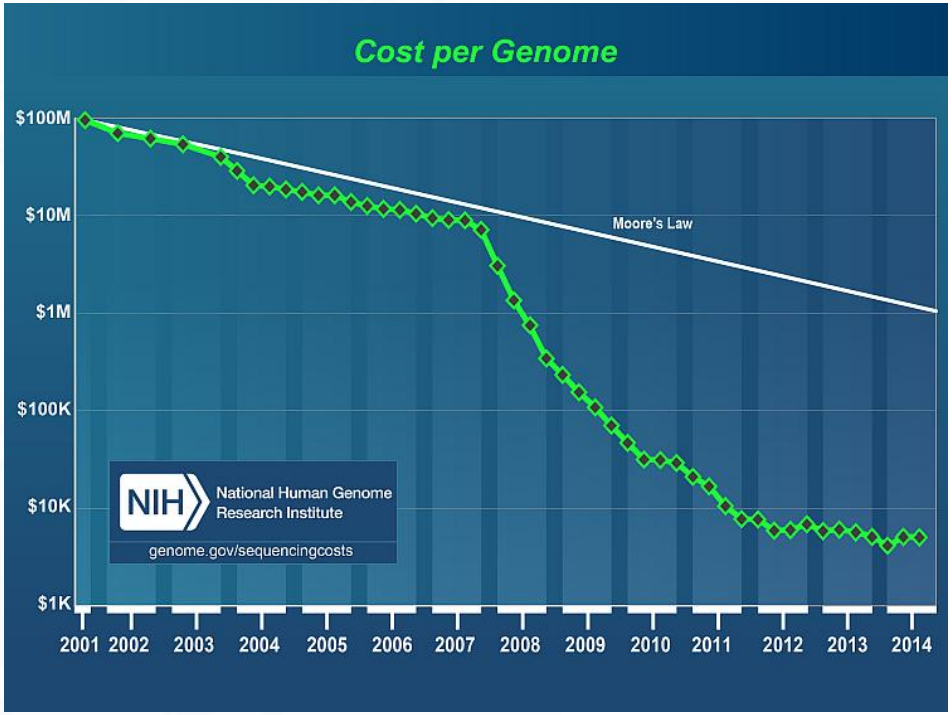


# Review of Sequencing Technologies

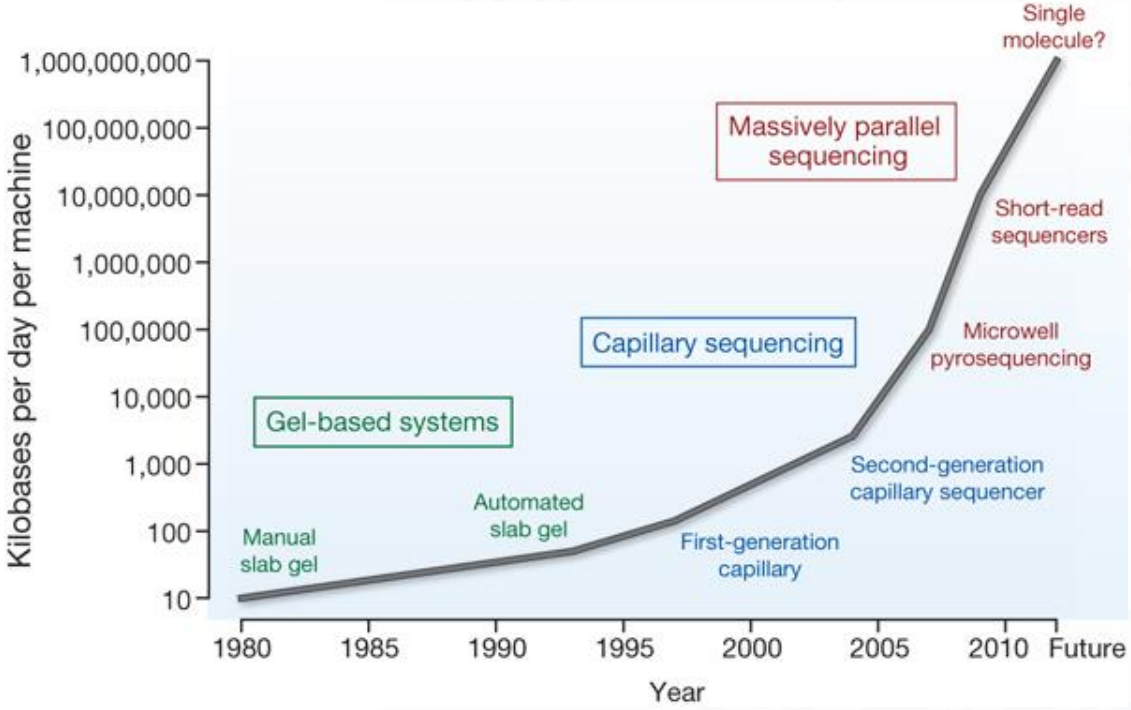
- Following concepts will be useful in comparing them
  - Reads
    - Read length
  - Depth
    - How many reads a certain location gets
  - Coverage
    - Are all parts covered?
  - Throughput
    - Reads per run
  - Accuracy / Quality
    - Read Errors
  - Run time and Cost

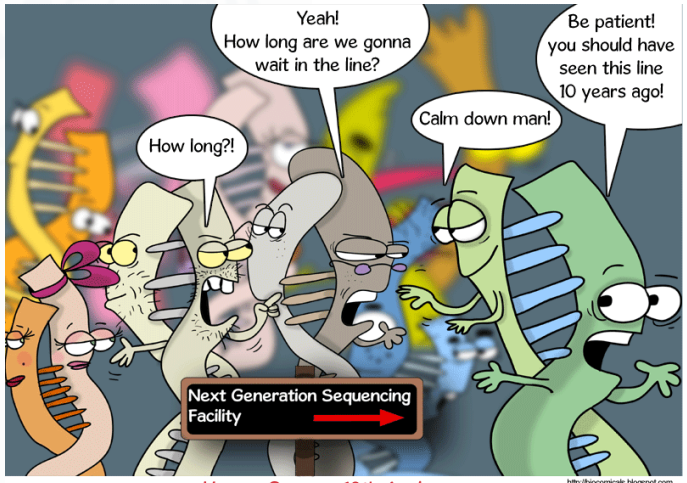
# Cost


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

# Throughput / Time



Human Genome 10th Anniversary

# Review of Sequencing Technologies

- Sanger Sequencing
- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
- Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton
- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION

HiSeq 2000


**First Generation**

**2nd Generation**
(Next Generation Sequencing, NGS)
(Deep Sequencing)
(High-throughput sequencing)
Amplified Single Molecule Sequencing
Most widely used right now

**3rd Generation**
(Next Next Generation Sequencing)
Single molecule sequencing

# Sanger Sequencing Video


http://www.dnalc.org/view/15479-Sanger-method-of-DNA-sequencing-3D-animation-with-narration.html

# Illumina Sequencing Video


www.youtube.com/watch?v=**l99aKKHcxC4**
http://www.ytpak.com/?component=video&task=view&id=l99aKKHcxC4

# SMRT Sequencing Video

http://www.ytpak.com/?component=video&task=view&id=_B_cUZ8hSYU
www.youtube.com/watch?v=**_B_cUZ8hSYU**

# Oxford Nanopore Sequencing Video

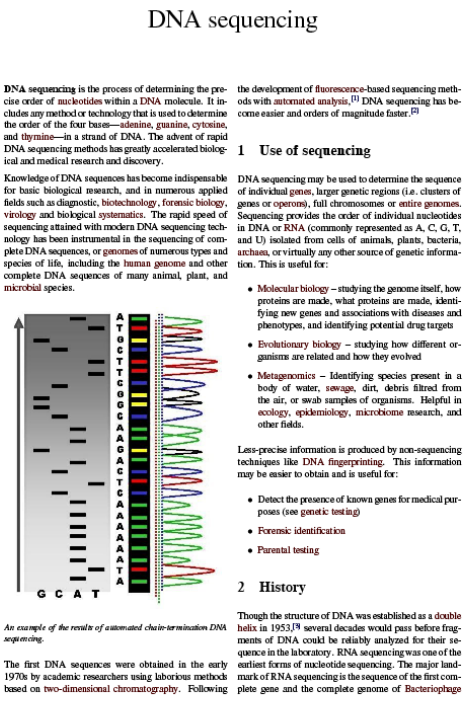
http://www.ytpak.com/?component=video&task=view&id=3UHw22hBpAk
www.youtube.com/watch?v=**3UHw22hBpAk**

# Sequencing Technologies Overview

- Practical issues
  - Read Errors
  - Produce FASTQ files which are essentially FASTA files but with read quality scores
- Uses
  - Whole Genome Sequencing
  - Genome wide association studies
  - SNP calling
  - Transcripts
- Types
  - Assembly (de novo)
  - Mapping

# Further reading

- http://en.wikipedia.org/wiki/DNA_sequencing



DNA sequencing

**DNA sequencing** is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as diagnostic, biotechnology, forensic biology, virology and biological systematics. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species.




*An example of the results of automated chain-termination DNA sequencing.*

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of fluorescence-based sequencing methods with automated analysis,[1] DNA sequencing has become easier and orders of magnitude faster.[2]

## 1 Use of sequencing

DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e. clusters of genes or operons), full chromosomes or entire genomes. Sequencing provides the order of individual nucleotides in DNA or RNA (commonly represented as A, C, G, T, and U) isolated from cells of animals, plants, bacteria, archaea, or virtually any other source of genetic information. This is useful for:

- Molecular biology – studying the genome itself, how proteins are made, what proteins are made, identifying new genes and associations with diseases and phenotypes, and identifying potential drug targets
- Evolutionary biology – studying how different organisms are related and how they evolved
- Metagenomics – Identifying species present in a body of water, sewage, dirt, debris filtred from the air, or swab samples of organisms. Helpful in ecology, epidemiology, microbiome research, and other fields.

Less-precise information is produced by non-sequencing techniques like DNA fingerprinting. This information may be easier to obtain and is useful for:

- Detect the presence of known genes for medical purposes (see genetic testing)
- Forensic identification
- Parental testing

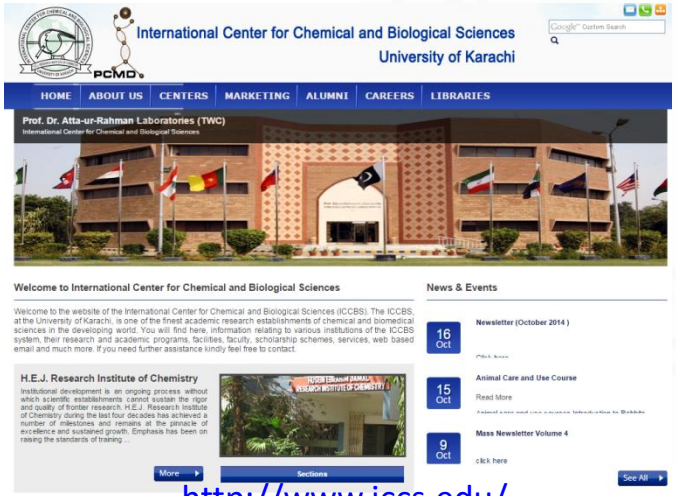
## 2 History

Though the structure of DNA was established as a double helix in 1953,[3] several decades would pass before fragments of DNA could be reliably analyzed for their sequence in the laboratory. RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage

1

# Sequencing in Pakistan

- ICCBS
  - HEJ Inst. Of Chemistry
    - NMR Spectroscopy
    - Mass Spectrometry
  - Dr. Panjwani Center for molecular medicine and drug Research (PCMD)
    - Jamil-ur-Rahman Center for Genome Research
      - DNA Sequencing facilities




http://www.iccs.edu/


http://www.dawn.com/news/640711/genome-mapping-of-first-pakistani-completed

Dr. Ata ur Rahman's genome was sequenced at a cost of $40K in 10 months. Pakistan already had sequences genomes of date palms, mangoes, etc.

# Sequencing in Pakistan

- IBGE
- PIEAS Affiliated: NIBGE / NIAB …



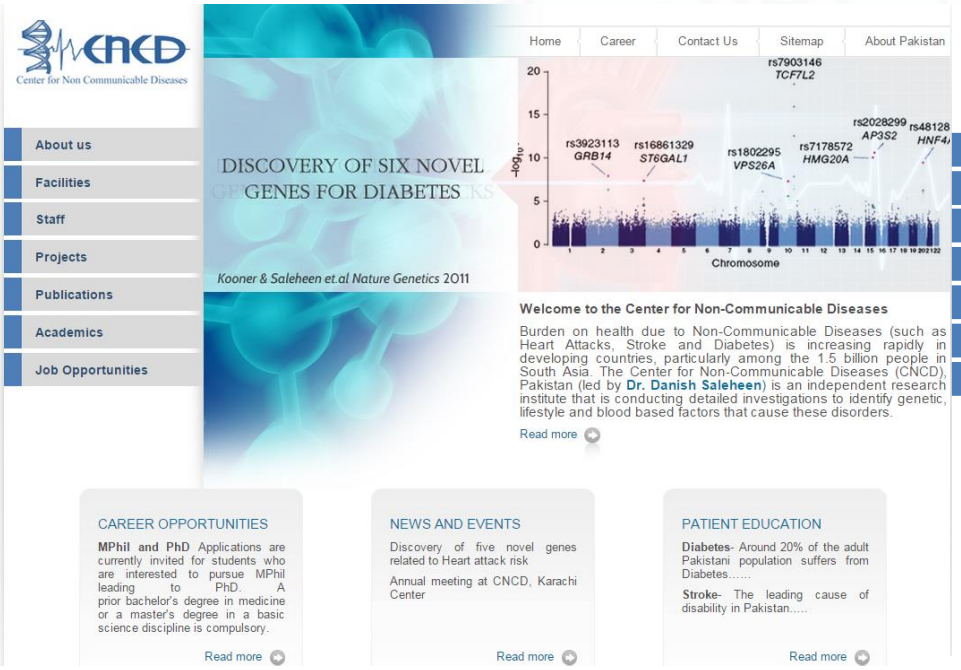
http://ibge.edu.pk/ http://nibge.org/

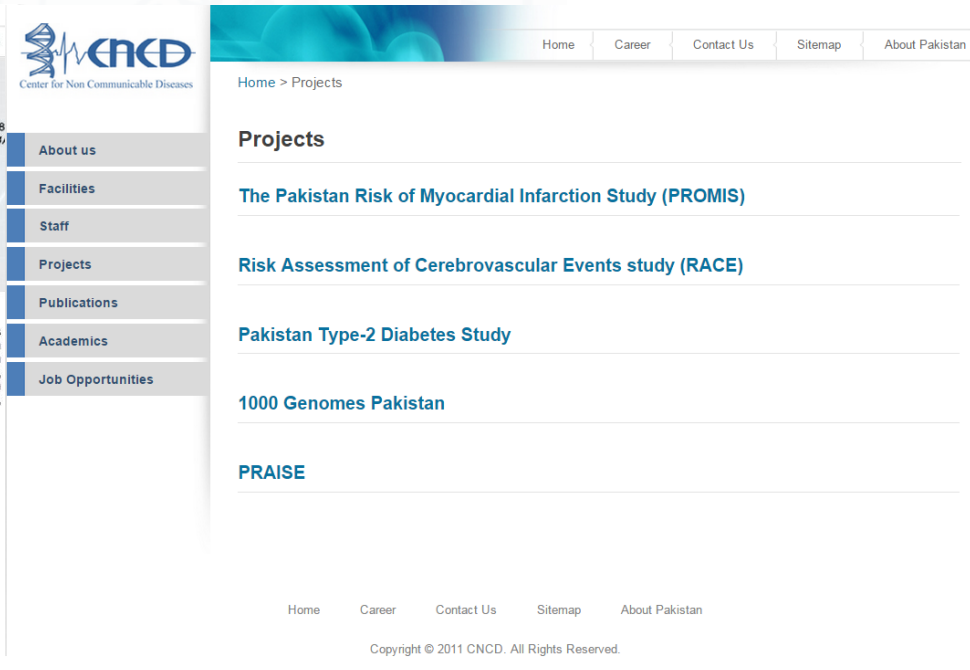
# Sequencing in Pakistan

- DNA Pakistan (http://dnapakistan.com/)
  - DNA Matching & Pairing
  - Center for Applied Molecular Biology (CAMB), University of the Punjab, Lahore
    - DNA sequencing, DNA genotyping and DNA synthesis

# Sequencing in Pakistan

- Center for noncommunicable diseases
  - http://www.cncdpk.com/
  - whole-genome sequencing, genome-wide association studies (GWAS), genetic markers …



# Sequencing in Pakistan

- Further reading
- Pakistan's silent revolution, by Suhail Yusuf Nov. 26, 2013
  - http://www.dawn.com/news/1058496
  - http://www.dawn.com/news/100343/paec-s-services-in-agriculture
  - http://www.dawn.com/authors/1083/suhail-yusuf/

# End of Lecture