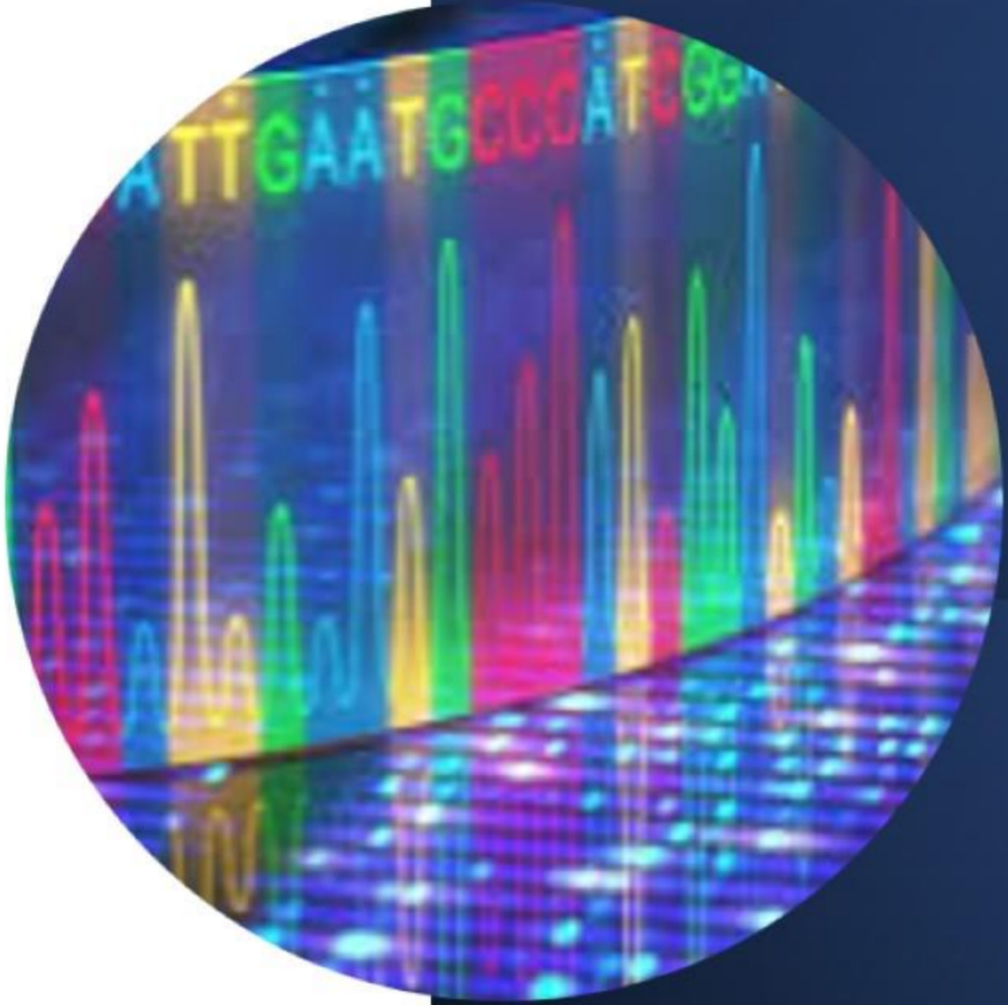
DNA sequencing

Lec 3 / Bioinformatics

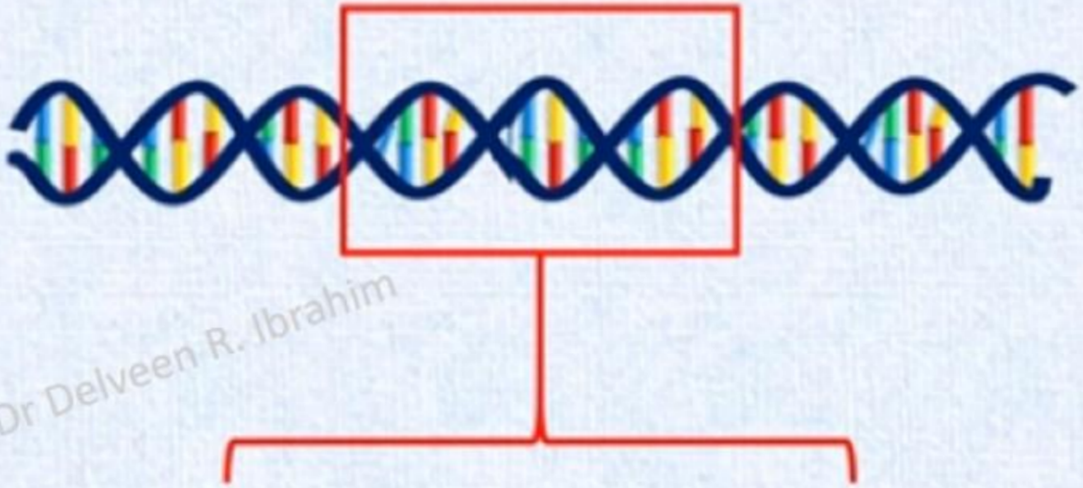
By

Dr Delveen R. Ibrahim



DNA sequencing

The technique by which the **precise order of nucleotides** in a DNA segment can be determined.

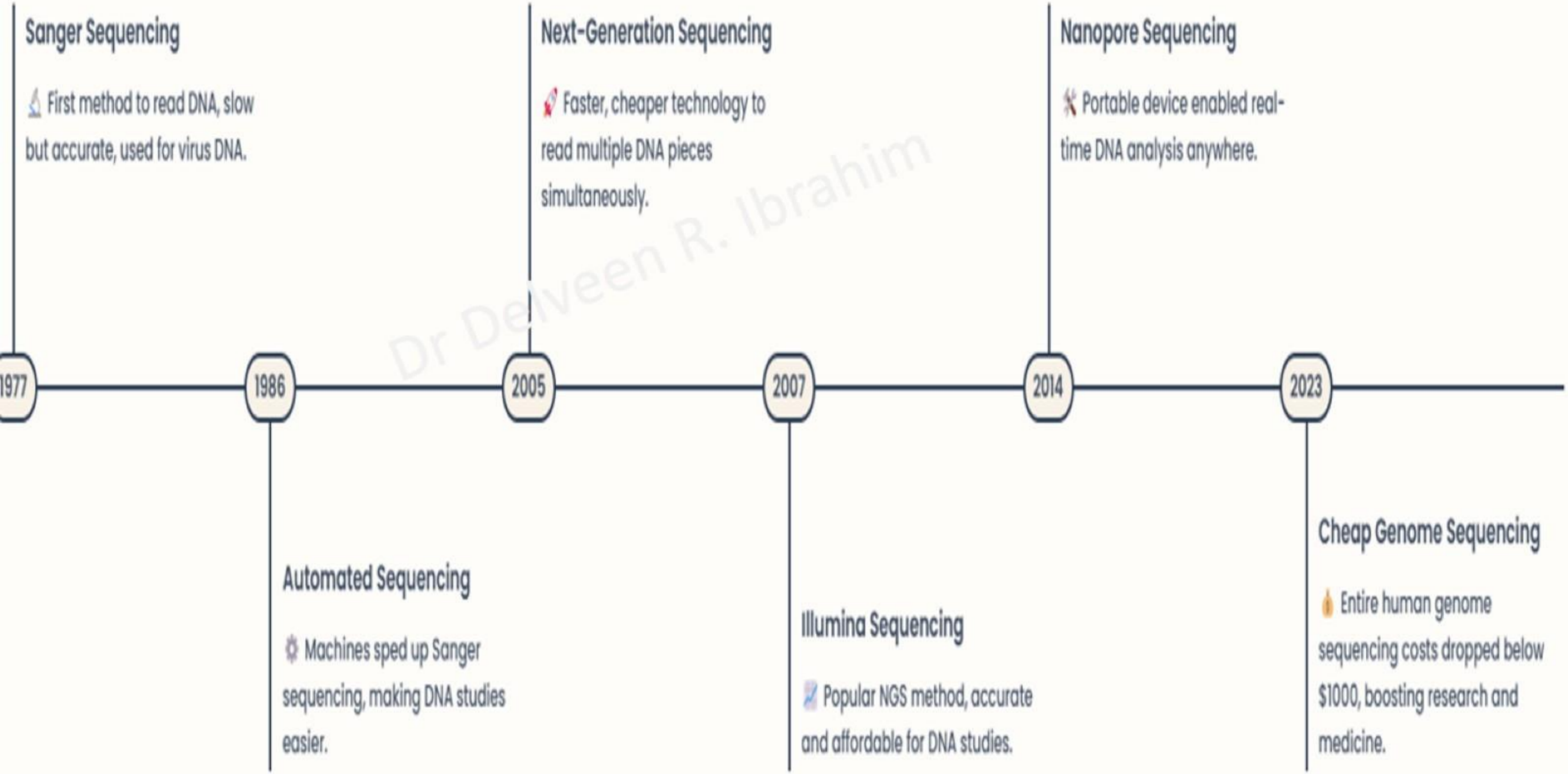


5’ ATGCACTTGATC **3’**
3’ TACGTGAACTAG **5’**

DNA sequencing:

- Simply, Sequencing DNA means determining the order of the four chemical building blocks called "bases" ,that make up the DNA molecule using laboratory technique .
- In the DNA double helix, the four chemical nucleotide or bases always bond with the same partner to form "base pairs.“
- These bases are Adenine, Cytosine, Thymine, Guanine.
- Adenine (A) always pairs with thymine (T); cytosine (C) always pairs with guanine (G).
- So, the role of DNA sequencing is to understand and interpret the genetics code to all biological life on earth as well as to understand and treat genetic diseases.

DNA Sequencing Developments



Sanger Sequencing

🔬 First method to read DNA, slow but accurate, used for virus DNA.

Automated Sequencing

⚙ Machines sped up Sanger sequencing, making DNA studies easier.

Next-Generation Sequencing

🚀 Faster, cheaper technology to read multiple DNA pieces simultaneously.

Illumina Sequencing

📈 Popular NGS method, accurate and affordable for DNA studies.

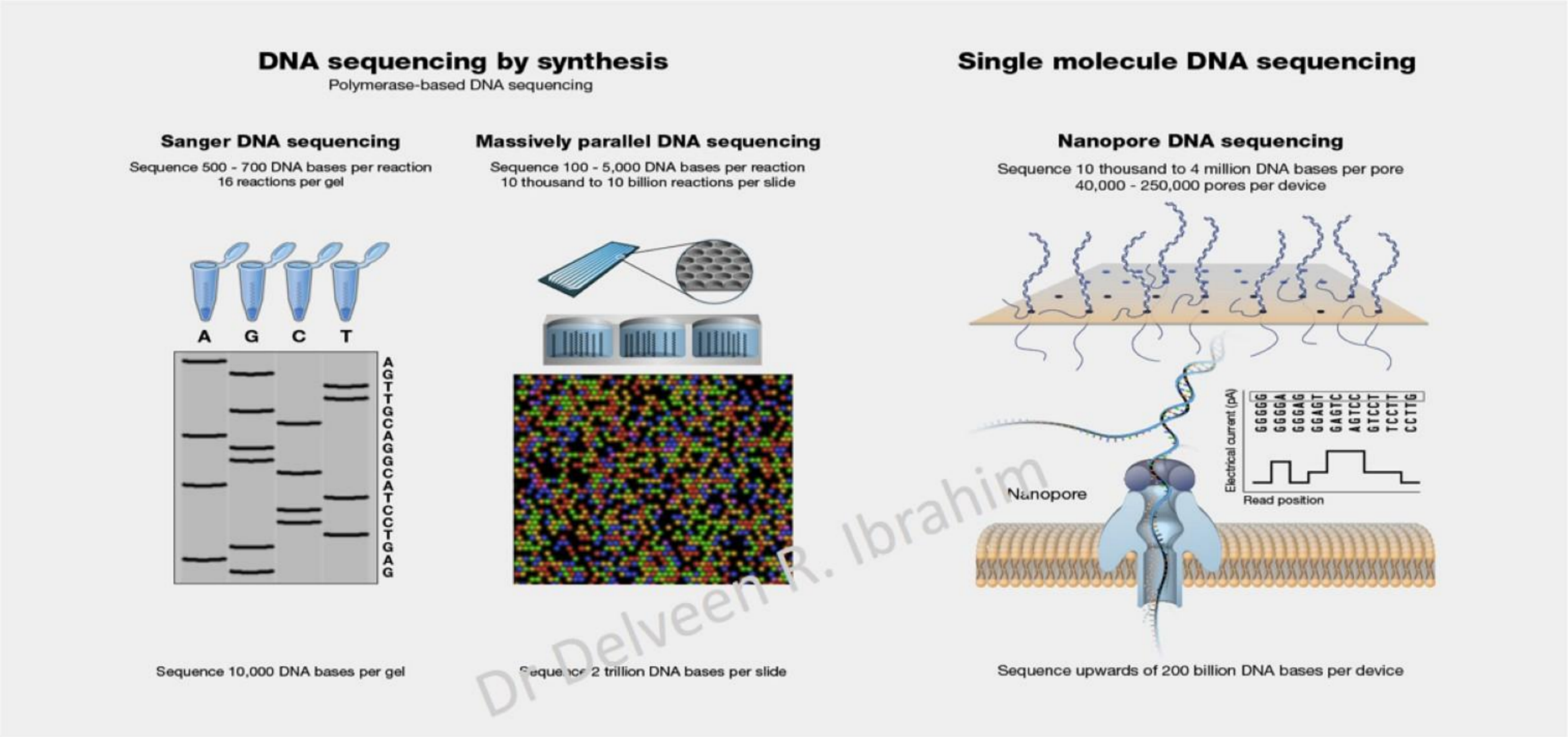
Nanopore Sequencing

🛠 Portable device enabled real-time DNA analysis anywhere.

Cheap Genome Sequencing

💰 Entire human genome sequencing costs dropped below $1000, boosting research and medicine.

DNA Sequencing methods: there are many methods which are using different mechanisms for sequencing



DNA sequencing by synthesis

Polymerase-based DNA sequencing

Sanger DNA sequencing

Sequence 500 - 700 DNA bases per reaction
16 reactions per gel

Sequence 10,000 DNA bases per gel

Massively parallel DNA sequencing

Sequence 100 - 5,000 DNA bases per reaction
10 thousand to 10 billion reactions per slide

Sequence 2 trillion DNA bases per slide

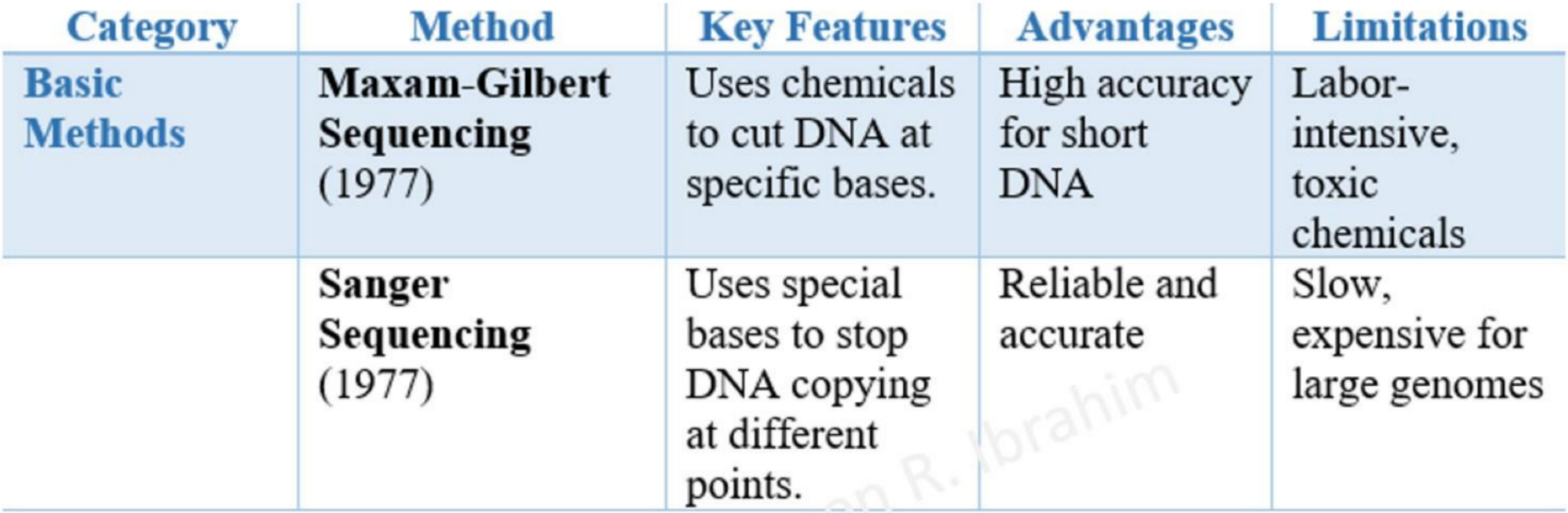
Single molecule DNA sequencing

Nanopore DNA sequencing

Sequence 10 thousand to 4 million DNA bases per pore
40,000 - 250,000 pores per device

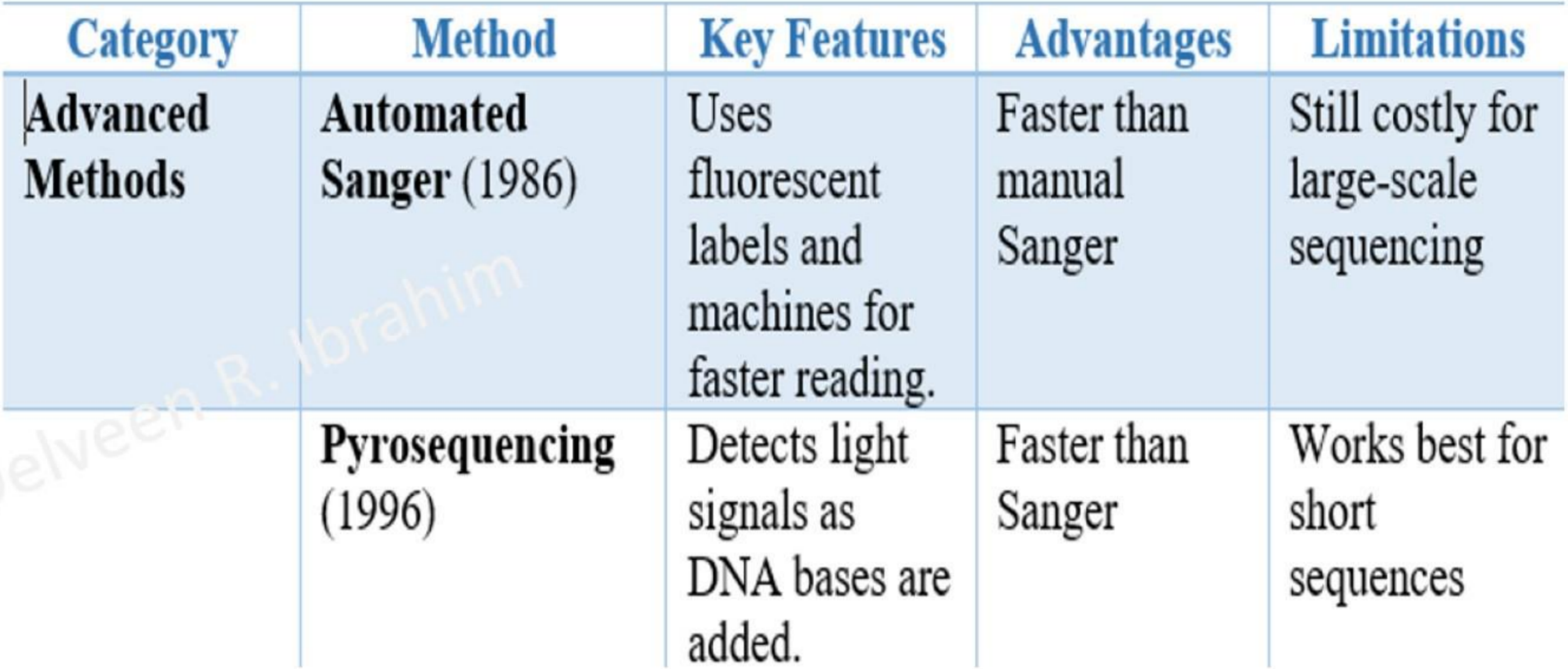
Sequence upwards of 200 billion DNA bases per device

1. Basic methods:



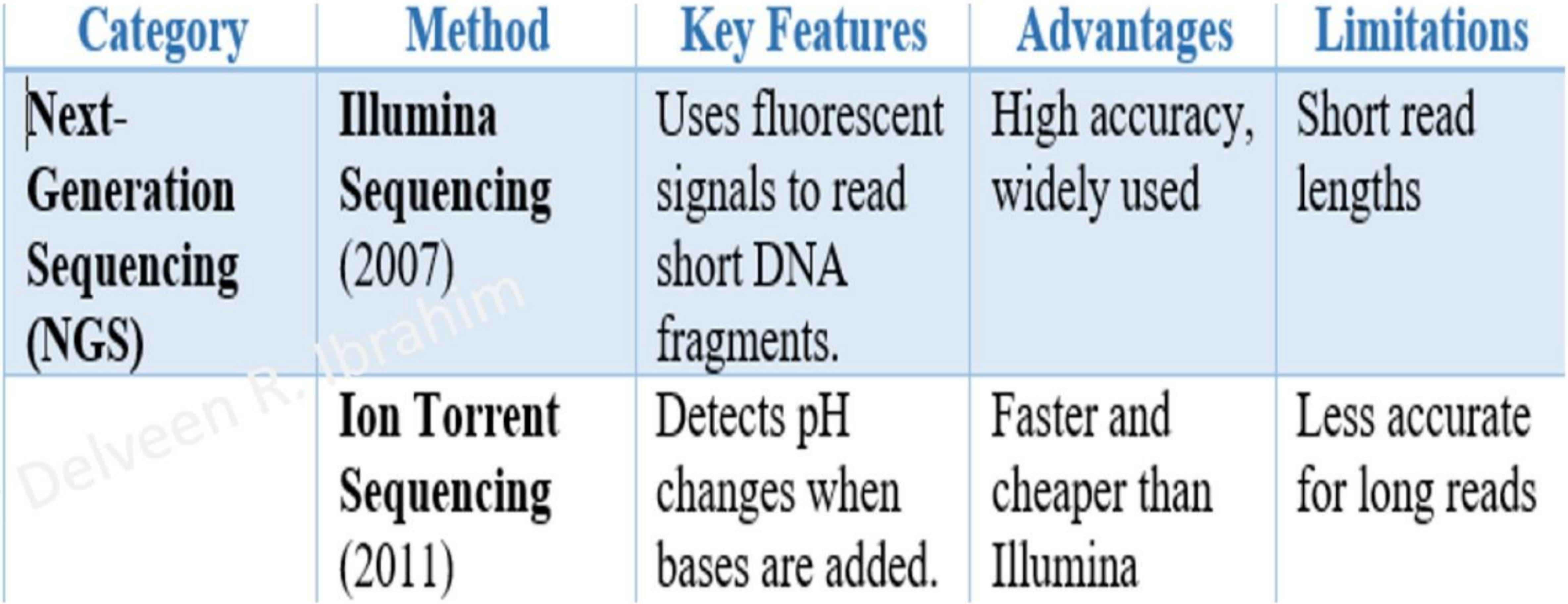
Category	Method	Key Features	Advantages	Limitations
Basic Methods	**Maxam-Gilbert Sequencing** (1977)	Uses chemicals to cut DNA at specific bases.	High accuracy for short DNA	Labor-intensive, toxic chemicals
	Sanger Sequencing (1977)	Uses special bases to stop DNA copying at different points.	Reliable and accurate	Slow, expensive for large genomes

2. Advanced methods:

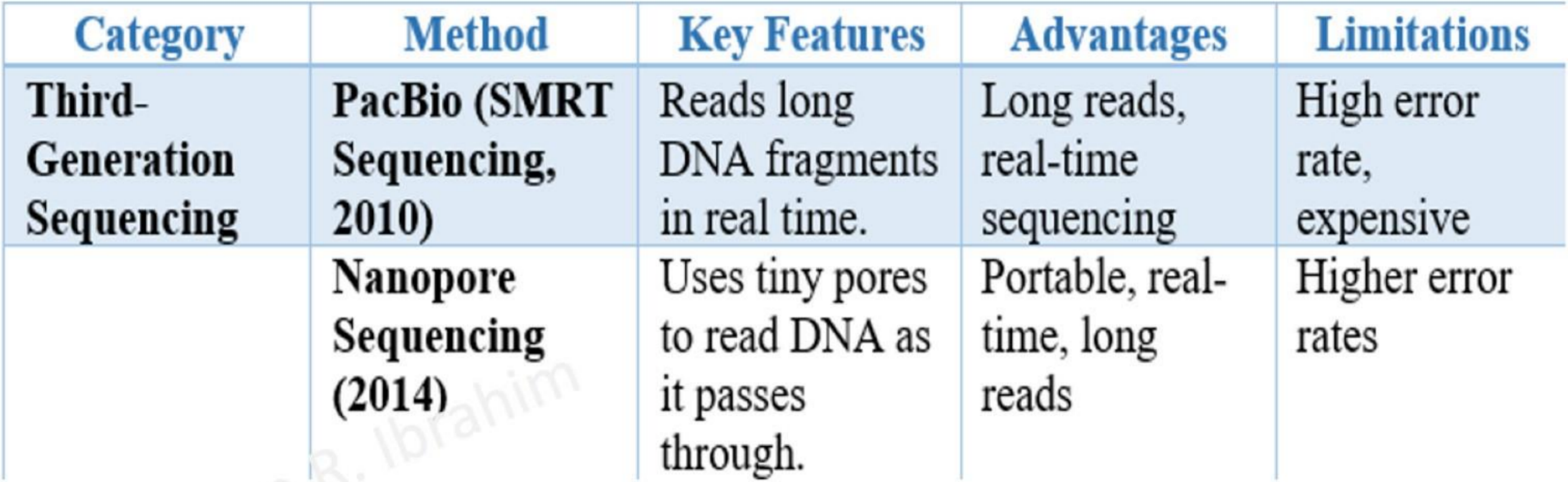


Category	Method	Key Features	Advantages	Limitations
Advanced Methods	**Automated Sanger** (1986)	Uses fluorescent labels and machines for faster reading.	Faster than manual Sanger	Still costly for large-scale sequencing
	Pyrosequencing (1996)	Detects light signals as DNA bases are added.	Faster than Sanger	Works best for short sequences

3. Next Generation Sequencing:



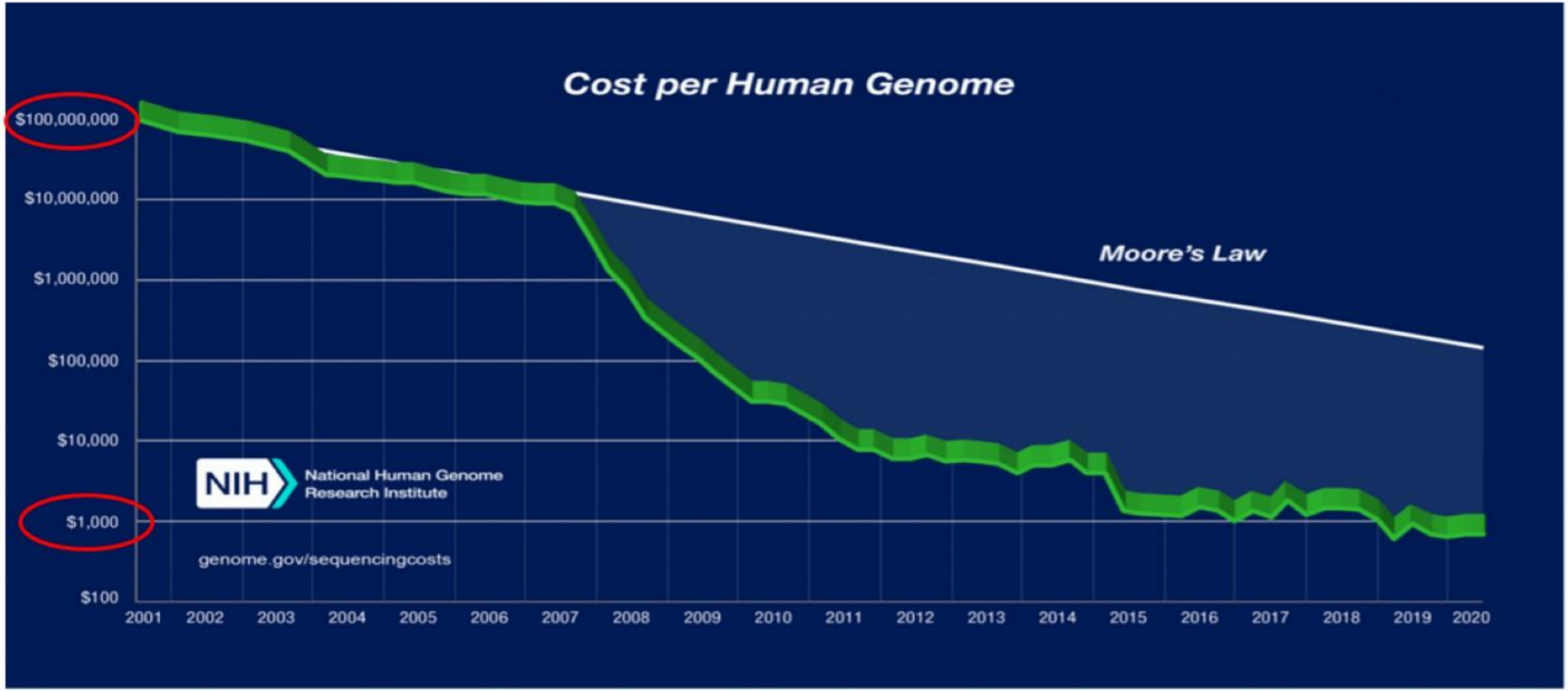
Category	Method	Key Features	Advantages	Limitations
Next-Generation Sequencing (NGS)	**Illumina Sequencing** (2007)	Uses fluorescent signals to read short DNA fragments.	High accuracy, widely used	Short read lengths
	Ion Torrent Sequencing (2011)	Detects pH changes when bases are added.	Faster and cheaper than Illumina	Less accurate for long reads



Category	Method	Key Features	Advantages	Limitations
Third-Generation Sequencing	**PacBio (SMRT Sequencing, 2010)**	Reads long DNA fragments in real time.	Long reads, real-time sequencing	High error rate, expensive
	Nanopore Sequencing (2014)	Uses tiny pores to read DNA as it passes through.	Portable, real-time, long reads	Higher error rates

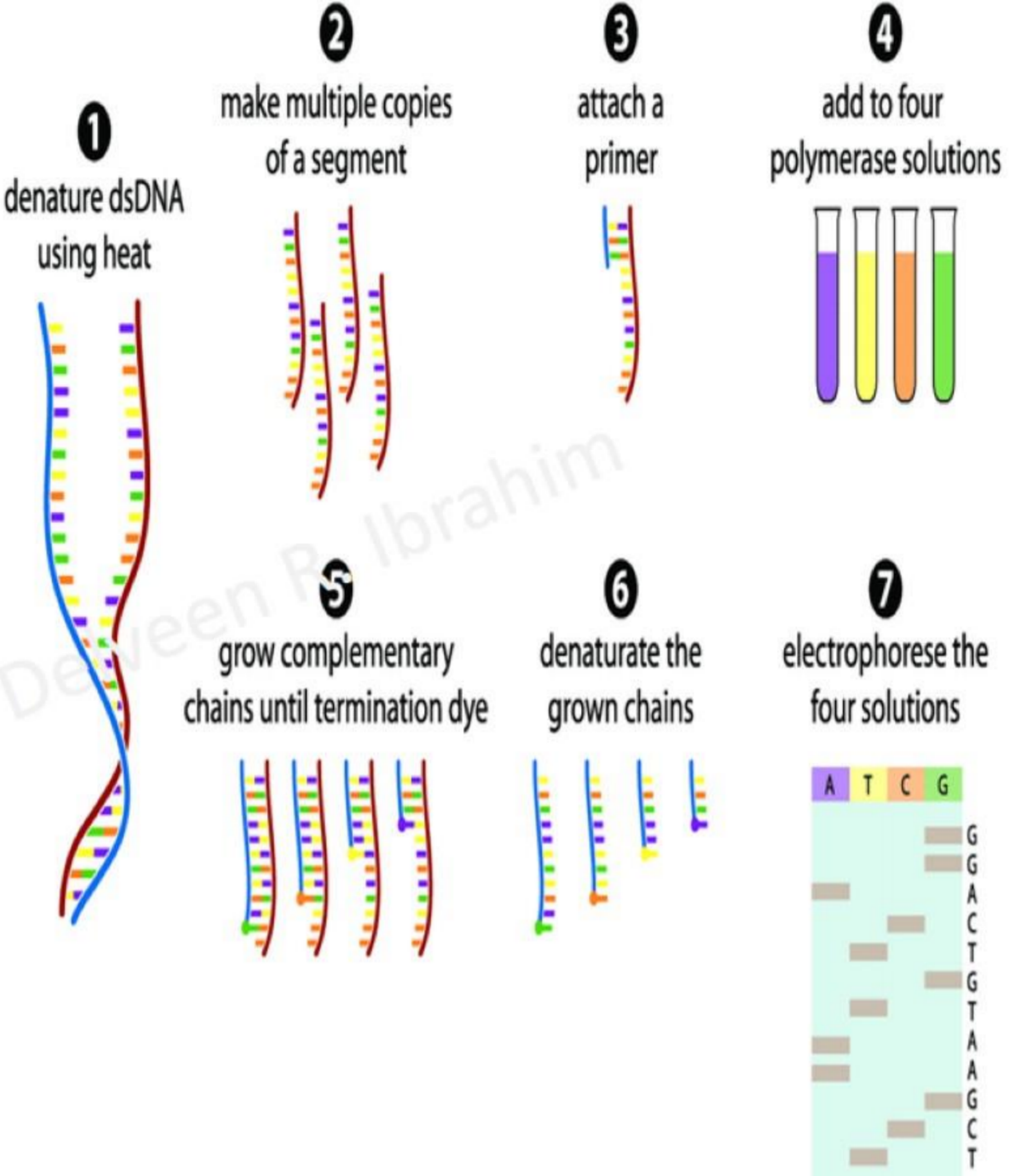
Note: Long reads and **short reads** refer to the length of DNA sequences that a sequencing method can read at a time

Due to the advanced methods used in DNA sequencing the cost also has been reduced tremendously which make it more affordable, see the figure below

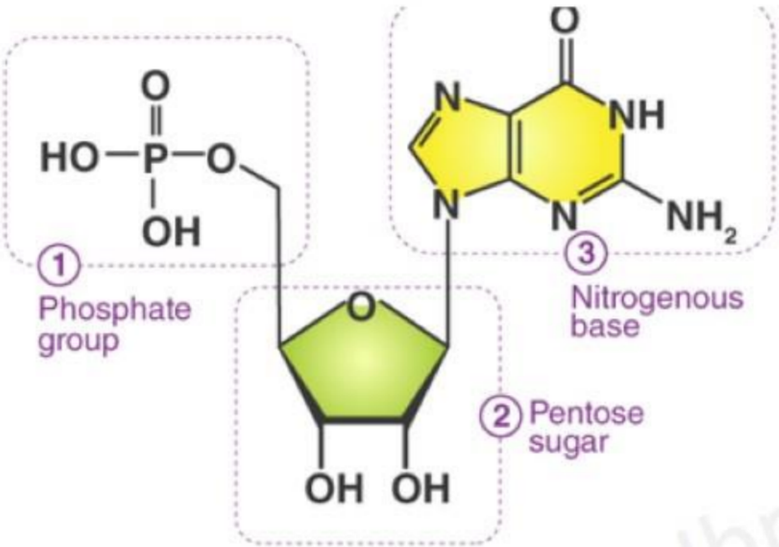


Sanger (chain termination method)

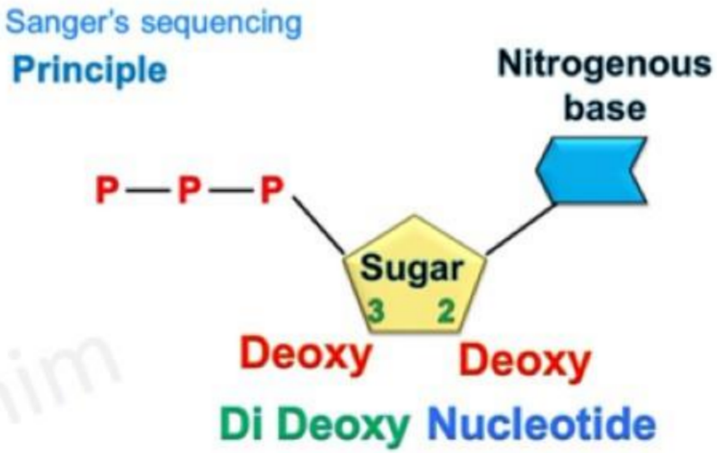
- This method include many steps starting with DNA extraction, purification , PCR, fragment separation , detection and data analysis. See the provided video in Moodle.
- But mainly this process depend on **DNA Polymerase** and **Dideoxynucleotides (ddNTPs):** DNA polymerase, the enzyme responsible for synthesizing new DNA strands, is then introduced along with a mixture of standard deoxynucleotides (dNTPs) and small amounts of modified nucleotides called dideoxynucleotides (ddNTPs). These ddNTPs lack a 3'-OH group, which is necessary for DNA strand elongation. So, after involving one of the ddNTPs, termination occur.


1 denature dsDNA using heat
2 make multiple copies of a segment
3 attach a primer
4 add to four polymerase solutions
5 grow complementary chains until termination dye
6 denaturate the grown chains
7 electrophorese the four solutions
A T C G
G
G
A
C
T
G
T
A
A
G
C
T

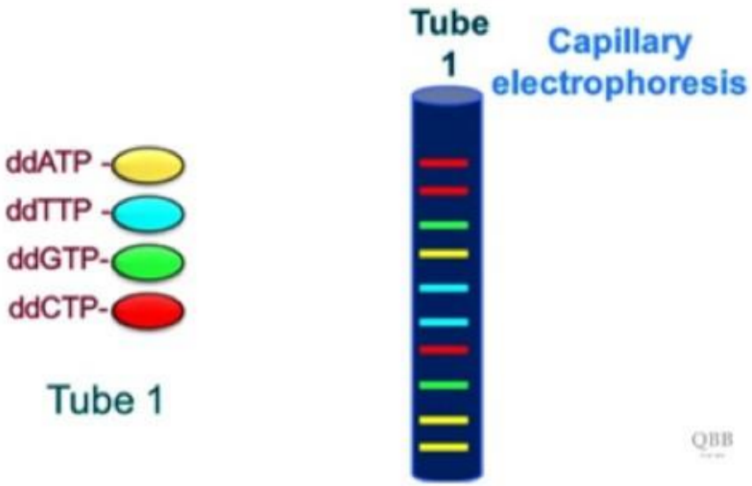
Nucleotide



Di Deoxy Nucleotide



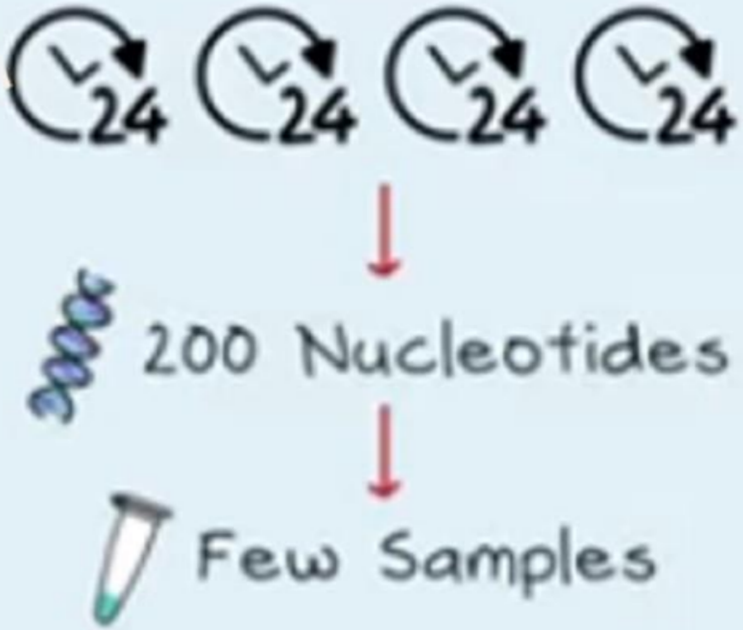
Fragment separation



Sanger Sequencing
Original Method

Labour Intensive

4 Days

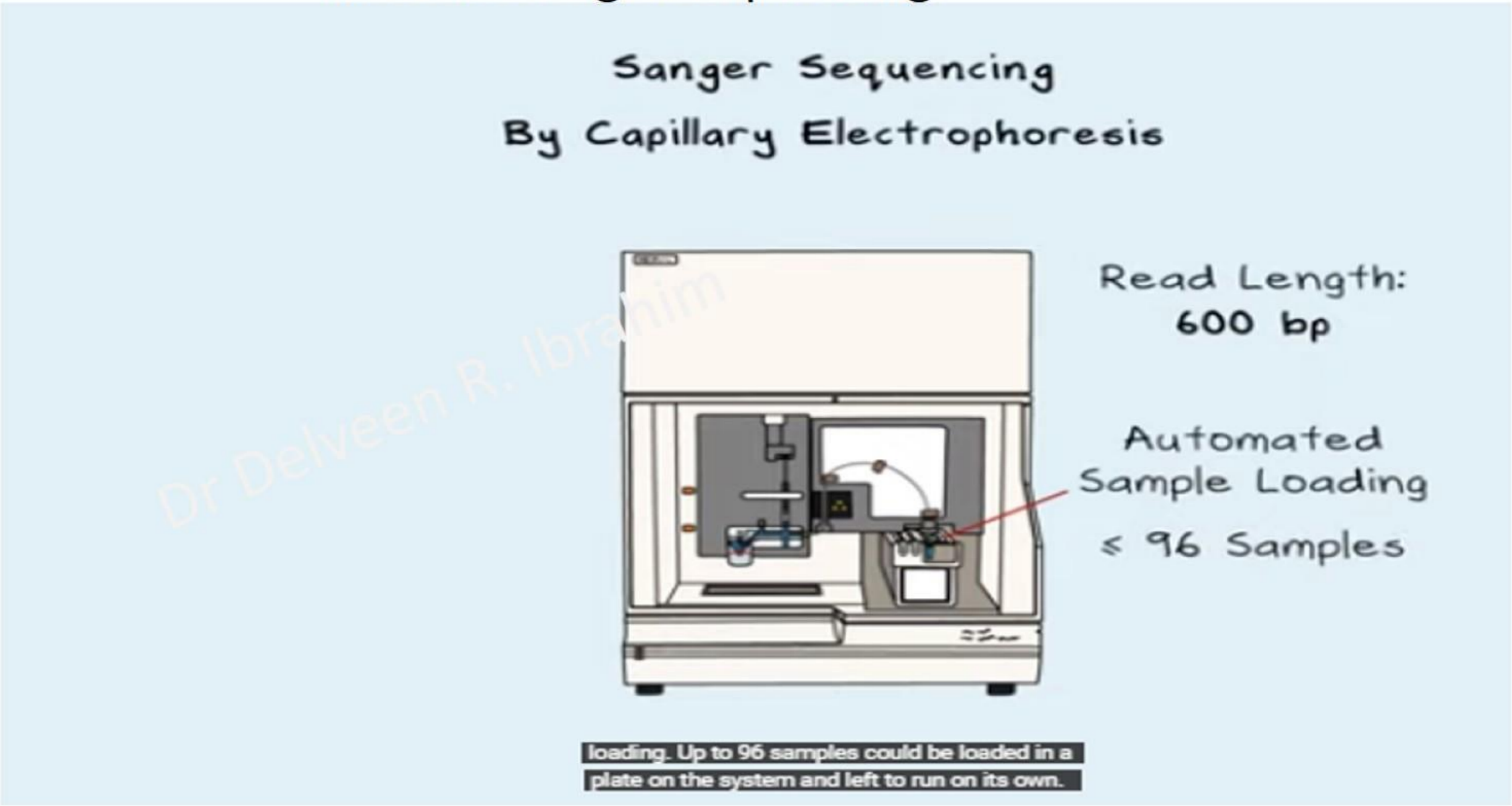


200 Nucleotides

Few Samples

!

Modern Sanger sequencing



Sanger Sequencing

By Capillary Electrophoresis

Read Length:
600 bp

Automated
Sample Loading

≤ 96 Samples

loading. Up to 96 samples could be loaded in a
plate on the system and left to run on its own.

Sanger Sequencing

Why Use Sanger Sequencing Today?

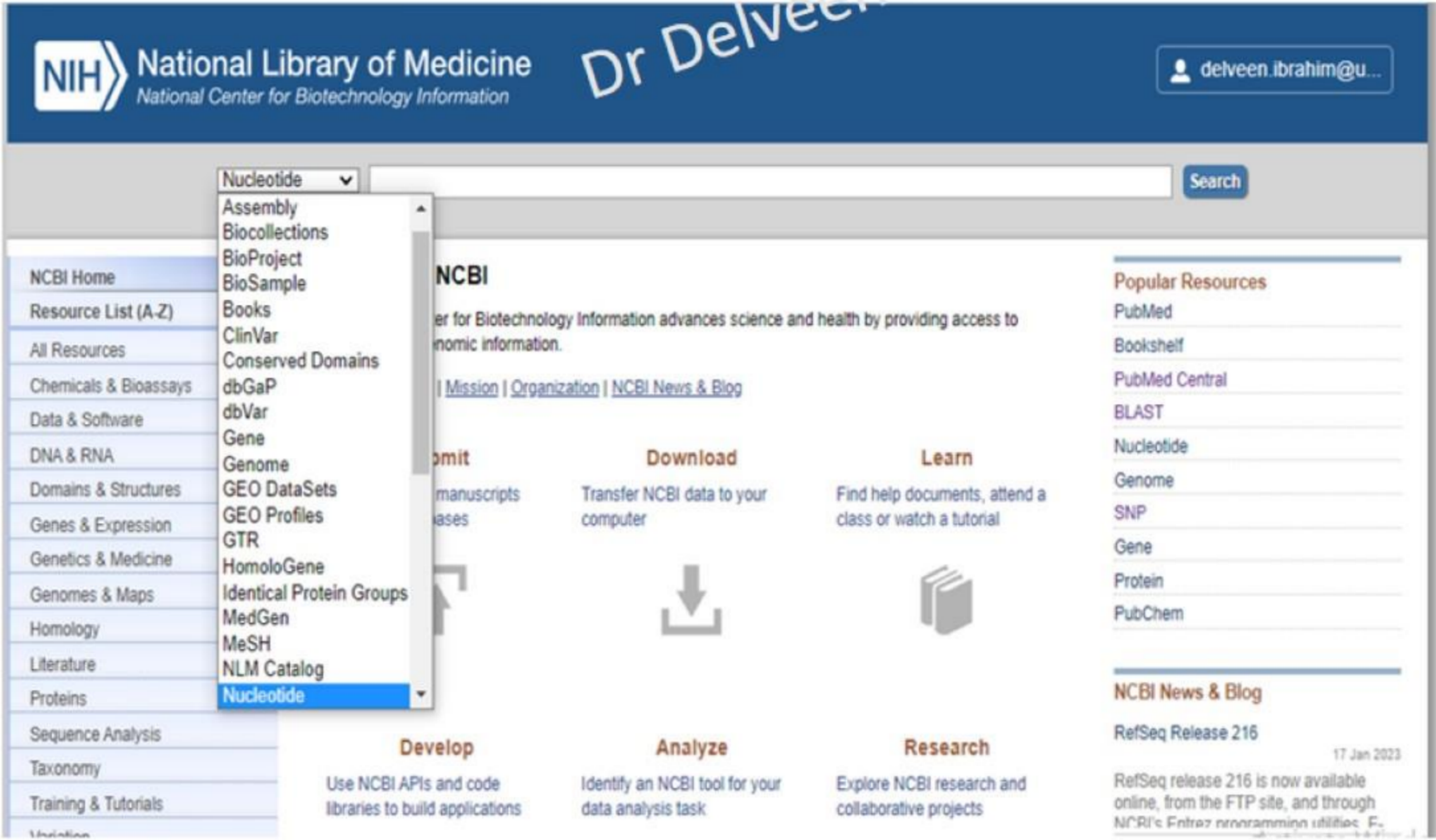
	Sanger Sequencing	NGS
Accuracy	99.9 %	99 - 99.9 %
Cost Effectiveness	< 20 Samples	> 20 Samples
Speed < 20 Samples	Fast	Slow
Speed > 20 Samples	Slow	Fast
Sensitivity	15 - 20 %	1 %
Sample Coverage	1 Read/Sample (300 - 850 bp)	Billions of Reads/Sample (Up to 16 Tb)

Retrieving nucleotide and protein sequence using NCBI

Practical / Lab 3

Finding Nucleotide sequence by using **accession number**, or **key words** such as the gene name and organism by searching the Nucleotide section in NCBI

On left upper side select **Nucleotide**. Type accession number , name of gene or gene symbol and organism name



...er for Biotechnology Information advances science and health by providing access to ...nomic information.

...manuscripts ...ases

Transfer NCBI data to your computer

Find help documents, attend a class or watch a tutorial

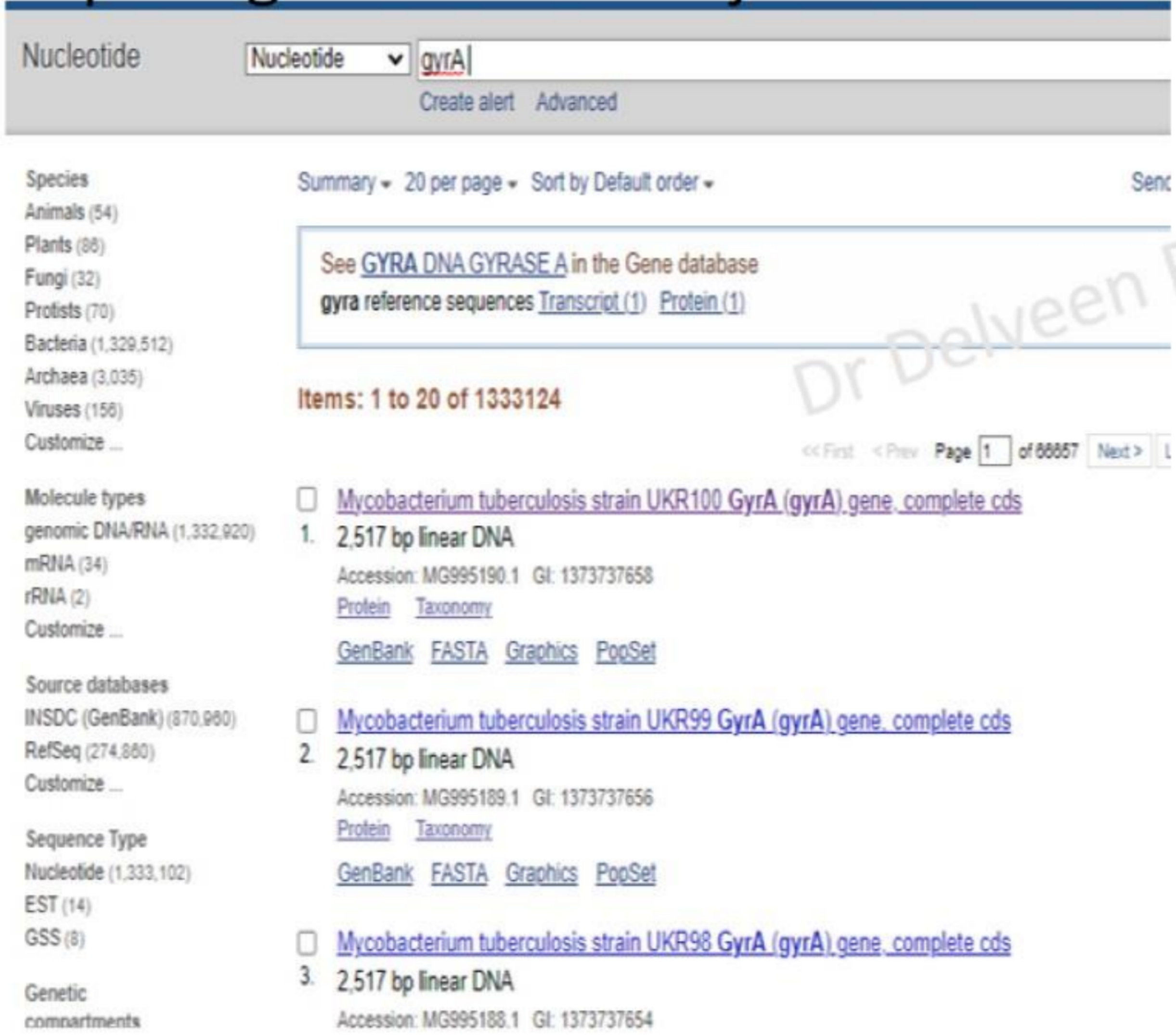
Use NCBI APIs and code libraries to build applications

Identify an NCBI tool for your data analysis task

Explore NCBI research and collaborative projects

RefSeq release 216 is now available online, from the FTP site, and through NCBI's Entrez programming utilities, E-

Type *gyrA* (gyrase subunit A) for example, you will get the window below. Click on the first option , you will get the GenBank and you can do copy the complete gene or the CDS just like before



Species
Animals (54)
Plants (88)
Fungi (32)
Protists (70)
Bacteria (1,329,512)
Archaea (3,035)
Viruses (156)
Customize ...

Molecule types
genomic DNA/RNA (1,332,920)
mRNA (34)
rRNA (2)
Customize ...

Source databases
INSDC (GenBank) (870,960)
RefSeq (274,860)
Customize ...

Sequence Type
Nucleotide (1,333,102)
EST (14)
GSS (8)

Genetic compartments

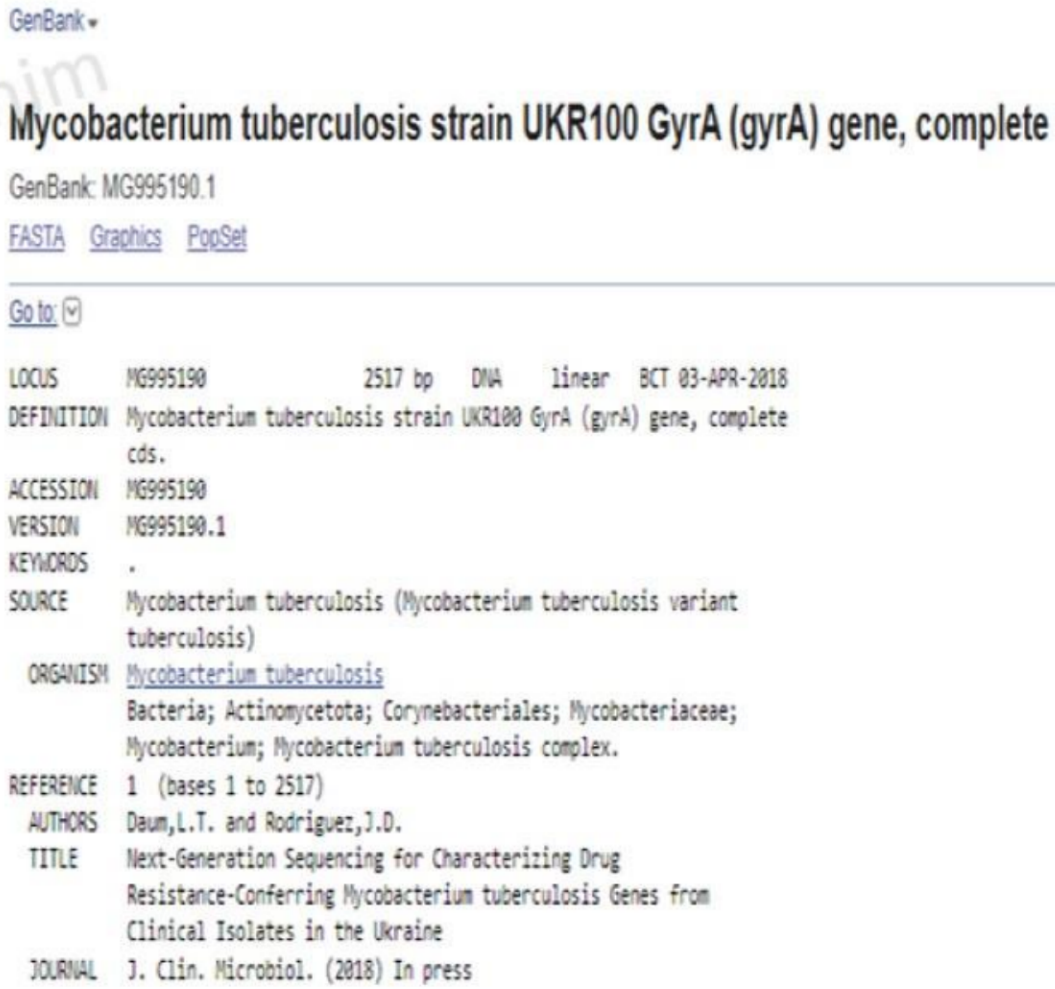
See GYRA DNA GYRASE A in the Gene database
gyra reference sequences Transcript (1) Protein (1)

Items: 1 to 20 of 1333124

□ Mycobacterium tuberculosis strain UKR100 GyrA (gyrA) gene, complete cds
1. 2,517 bp linear DNA
Accession: MG995190.1 GI: 1373737658
Protein Taxonomy
GenBank FASTA Graphics PopSet

□ Mycobacterium tuberculosis strain UKR99 GyrA (gyrA) gene, complete cds
2. 2,517 bp linear DNA
Accession: MG995189.1 GI: 1373737656
Protein Taxonomy
GenBank FASTA Graphics PopSet

□ Mycobacterium tuberculosis strain UKR98 GyrA (gyrA) gene, complete cds
3. 2,517 bp linear DNA
Accession: MG995188.1 GI: 1373737654



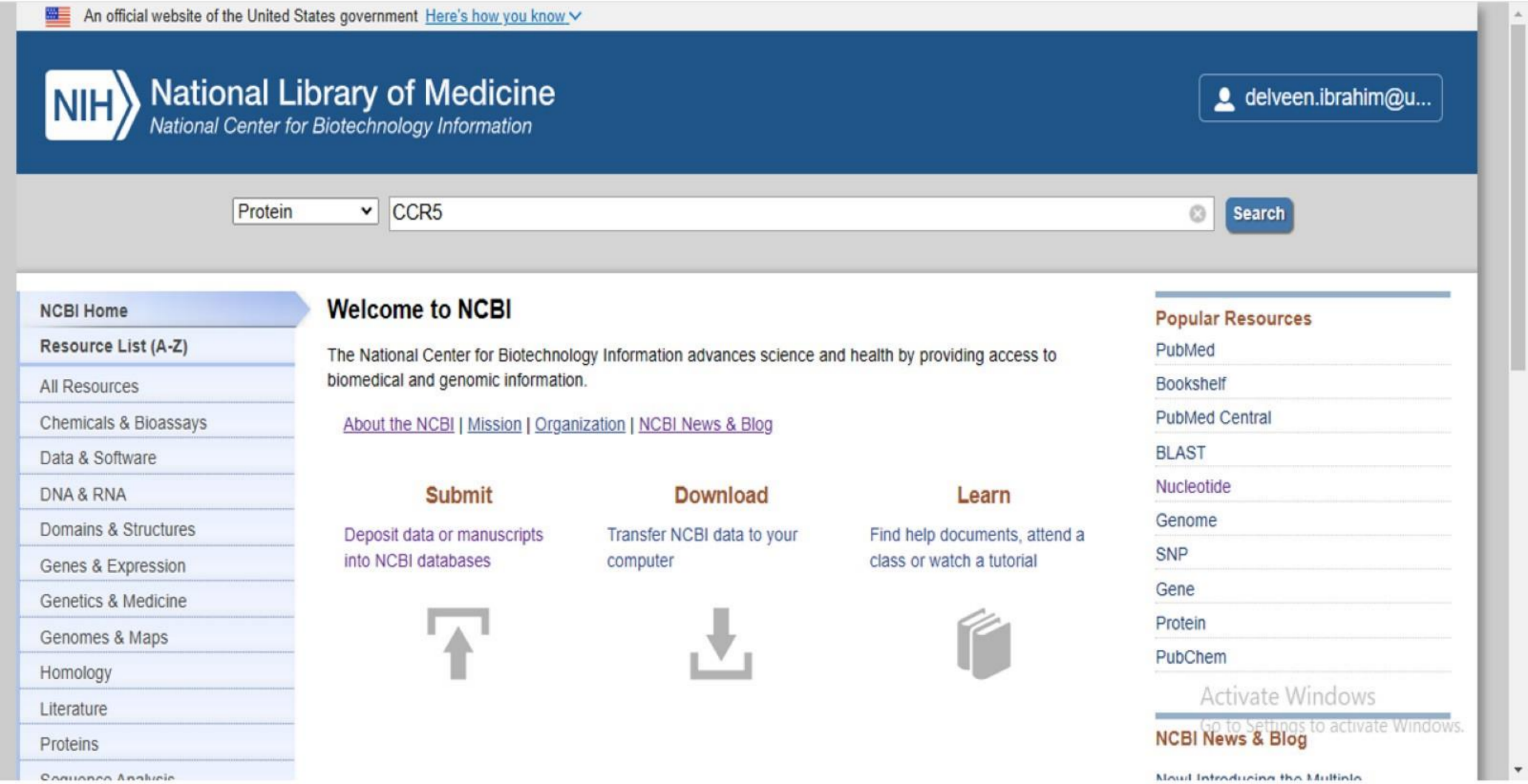
Mycobacterium tuberculosis strain UKR100 GyrA (gyrA) gene, complete

GenBank: MG995190.1

```
LOCUS       MG995190                2517 bp    DNA     linear   BCT 03-APR-2018
DEFINITION  Mycobacterium tuberculosis strain UKR100 GyrA (gyrA) gene, complete
            cds.
ACCESSION   MG995190
VERSION     MG995190.1
KEYWORDS    .
SOURCE      Mycobacterium tuberculosis (Mycobacterium tuberculosis variant
            tuberculosis)
  ORGANISM  Mycobacterium tuberculosis
            Bacteria; Actinomycetota; Corynebacteriales; Mycobacteriaceae;
            Mycobacterium; Mycobacterium tuberculosis complex.
REFERENCE   1  (bases 1 to 2517)
  AUTHORS   Daum,L.T. and Rodriguez,J.D.
  TITLE     Next-Generation Sequencing for Characterizing Drug
            Resistance-Conferring Mycobacterium tuberculosis Genes from
            Clinical Isolates in the Ukraine
  JOURNAL   J. Clin. Microbiol. (2018) In press
```

What are the prefix means in accession numbers?

- **NM_**: Refers to mRNA sequences (messenger RNA).
- **NP_**: Refers to protein sequences (translated products of mRNA).
- **NG_**: Refers to incomplete genomic regions (usually gene-specific loci).
- **NC** stands for **Non-redundant Curated** sequence



Protein: CCR5

Welcome to NCBI

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

Submit

Deposit data or manuscripts into NCBI databases

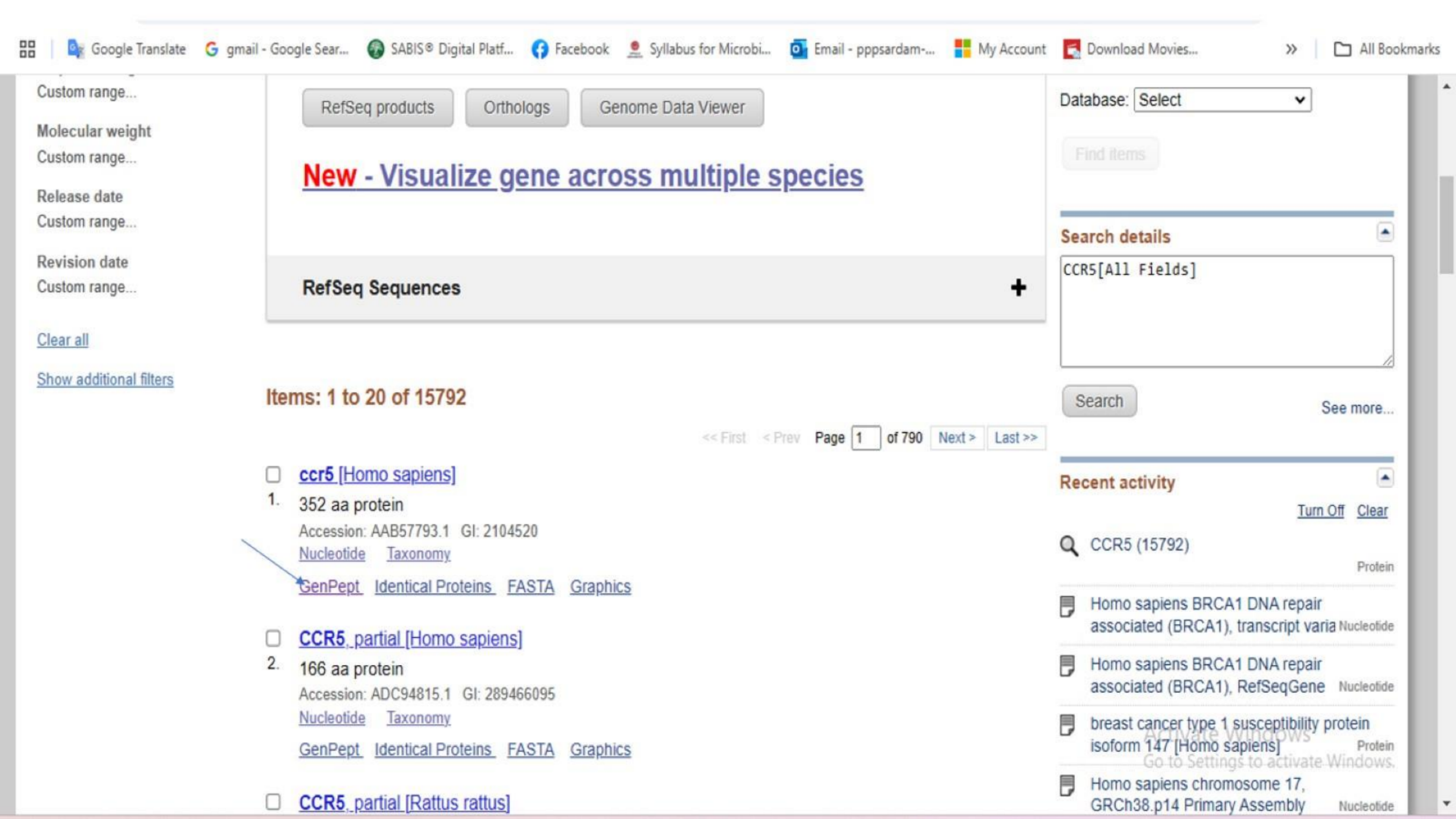
Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

New! Introducing the Multiple



New - Visualize gene across multiple species

RefSeq Sequences +

Search details

CCR5[All Fields]

Items: 1 to 20 of 15792

Page 1 of 790

1. ccr5 [Homo sapiens]
352 aa protein
Accession: AAB57793.1 GI: 2104520
Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

2. CCR5, partial [Homo sapiens]
166 aa protein
Accession: ADC94815.1 GI: 289466095
Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

CCR5, partial [Rattus rattus]

Recent activity

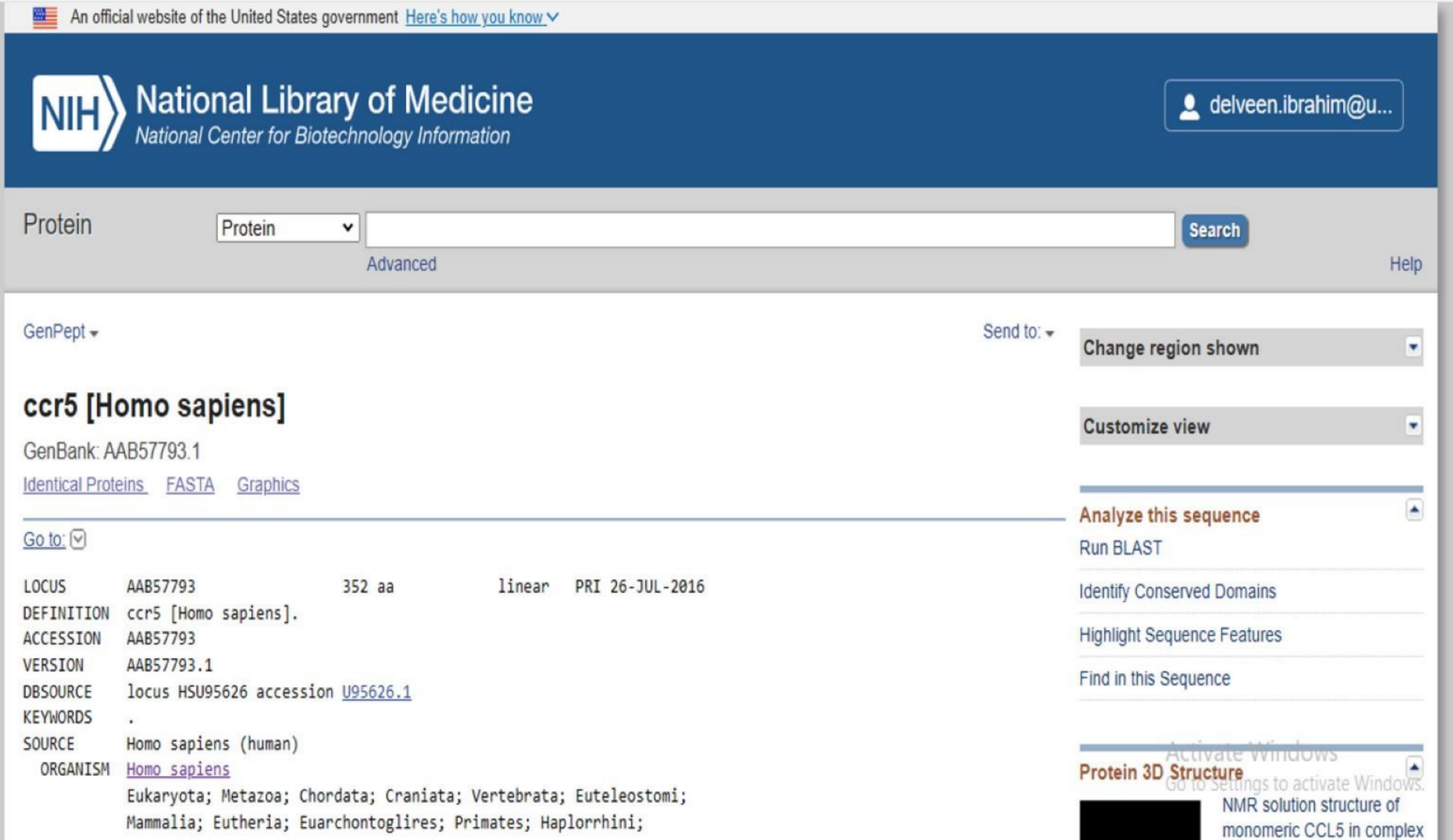
CCR5 (15792) Protein

Homo sapiens BRCA1 DNA repair associated (BRCA1), transcript varia Nucleotide

Homo sapiens BRCA1 DNA repair associated (BRCA1), RefSeqGene Nucleotide

breast cancer type 1 susceptibility protein isoform 147 [Homo sapiens] Protein

Homo sapiens chromosome 17, GRCh38.p14 Primary Assembly Nucleotide



ccr5 [Homo sapiens]

GenBank: AAB57793.1

```
LOCUS       AAB57793                 352 aa            linear   PRI 26-JUL-2016
DEFINITION  ccr5 [Homo sapiens].
ACCESSION   AAB57793
VERSION     AAB57793.1
DBSOURCE    locus HSU95626 accession U95626.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
```

Tasks :

You will be given different tasks in the lecture