Основи секвенування нового покоління (NGS)

Анна Яручик

Експерт Бюро ВООЗ в Україні

World Health Organization

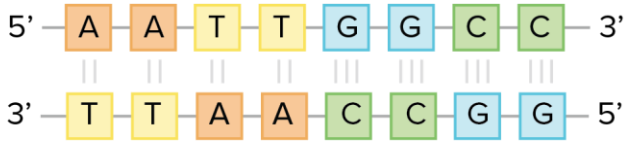
European Region

План:

1. Терміни.
2. Огляд основних технологій NGS.
3. Аналіз даних NGS.
4. Бази даних.
5. Значення NGS та геномного нагляду в охороні здоров’я.

World Health Organization
European Region

Терміни:



- **Принцип комплементарності:** чітка відповідність нуклеотидів між двома ланцюгами ДНК, які за допомогою водневих зв'язків формують парні комплекси: аденін в одному ланцюзі протилежний тиміну (або урацилу в РНК) в іншому ланцюзі, а цитозин в одному ланцюзі протилежний гуаніну в іншому.
- **Секвенування нового покоління (Next Generation Sequencing, NGS, масивне паралельне секвенування) -** це високопродуктивна технологія секвенування нуклеїнових кислот, яка дозволяє швидко та ефективно визначити послідовність багатьох молекул ДНК/РНК одночасно в одному об’ємі біохімічної реакції.
- **Зчитування (Read):** послідовність одного ДНК-фрагмента, яка секвенується.
- **Бібліотека (Library):** колекція ДНК-фрагментів з приєднаними адаптерами, підготовлена до секвенування.
- **Адаптери:** олігонуклеотиди з відомими послідовностями, які лігуються з ДНК-фрагментами для зв’язування їх з матрицею в інструменті NGS.
- **Лігування:** це з'єднання двох фрагментів нуклеїнової кислоти ферментом лігазою.
- **Індекс (баркод):** коротка послідовність олігонуклеотидів, прикріплених до зразка для його ідентифікації при змішуванні з іншими зразками. Сам індекс також секвенується, таким чином діючи як «штрих-код» для ідентифікації зразка.

World Health Organization
European Region

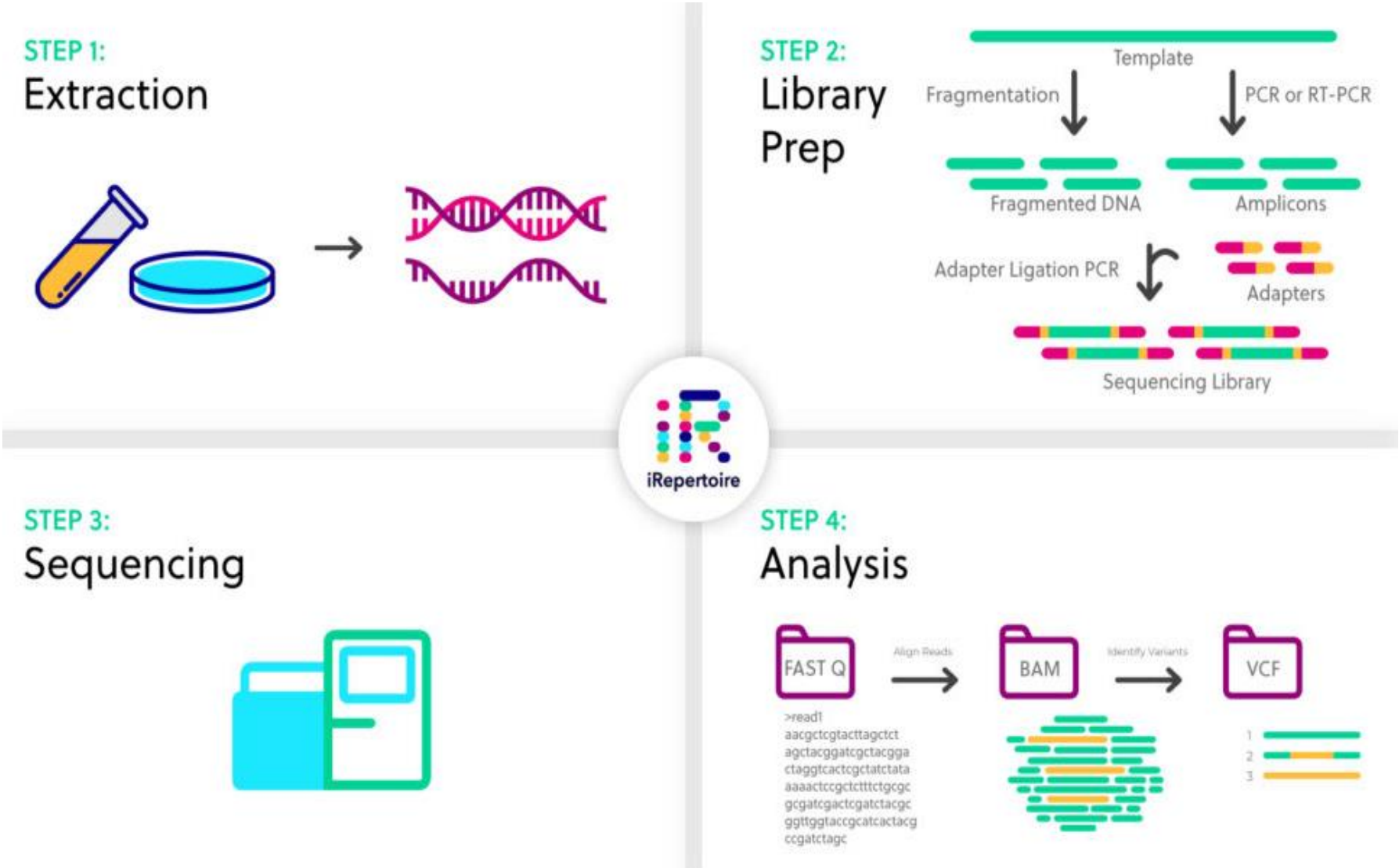
Підходи секвенування

- **Таргетне секвенування:** секвенування лише обраних ділянок геному за допомогою праймерів або зондів, які специфічно зв'язуються з цільовими ділянками геному. Це дозволяє зменшити обсяг даних, отриманих в процесі секвенування, зменшуючи час і витрати на аналіз даних.
- **Секвенування методом дробовика (Shotgun sequencing):** стратегія секвенування, яка полягає в розбитті випадковим чином геному на велику кількість коротких фрагментів, що секвенуються окремо та потім збираються в одну послідовність. Метод використовується для дослідження нової або невідомої ДНК, а також для секвенування метагеномів - зразків з довкілля, що містять суміш багатьох організмів, таких як грунт, вода, кишкові мікроорганізми тощо.
- **Повногеномне секвенування (Whole Genome Sequencing):** секвенування повного геному цільового виду.
- **Повноекзомне секвенування (Whole Exome Sequencing):** секвенування лише екзонів (ділянок, що кодують білки) цільового виду.

World Health Organization
European Region

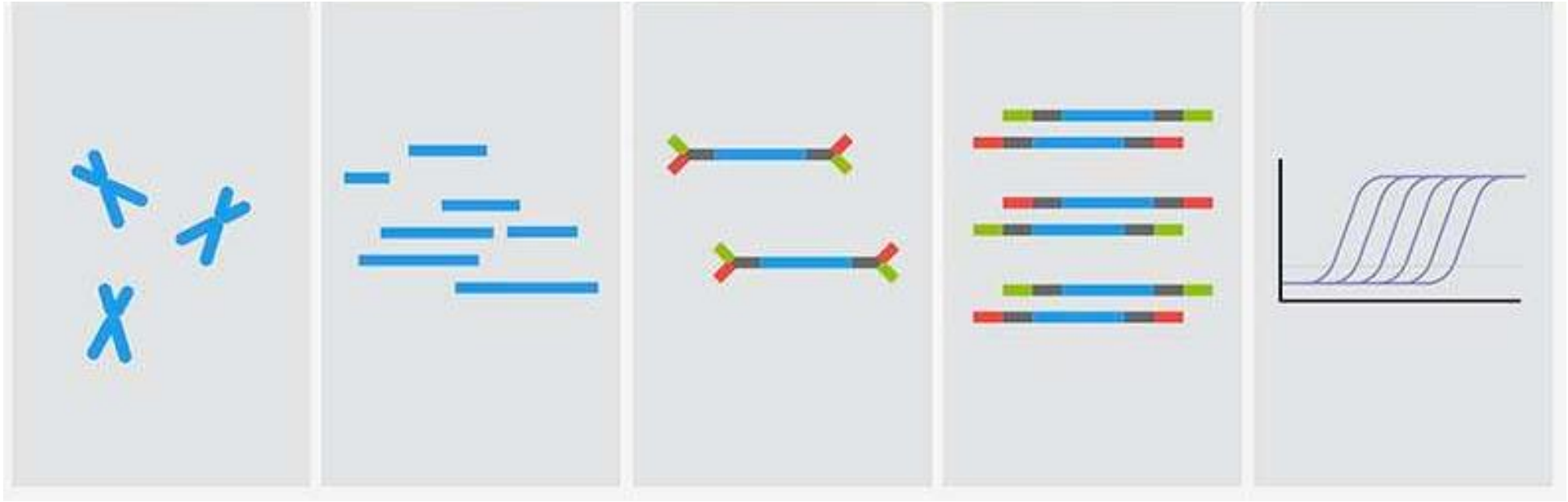
Основні етапи робочого процесу NGS

1. **Екстракція ДНК/РНК зі зразків,**
   оцінка кількості та якості.
   *** Зворотня транскрипція для РНК**
2. **Підготовка бібліотеки**
3. **Секвенування**
4. **Аналіз даних**



World Health Organization
European Region

Підготовка бібліотеки



Утворення фрагментів ДНК

Приєднання адаптерів та індексів

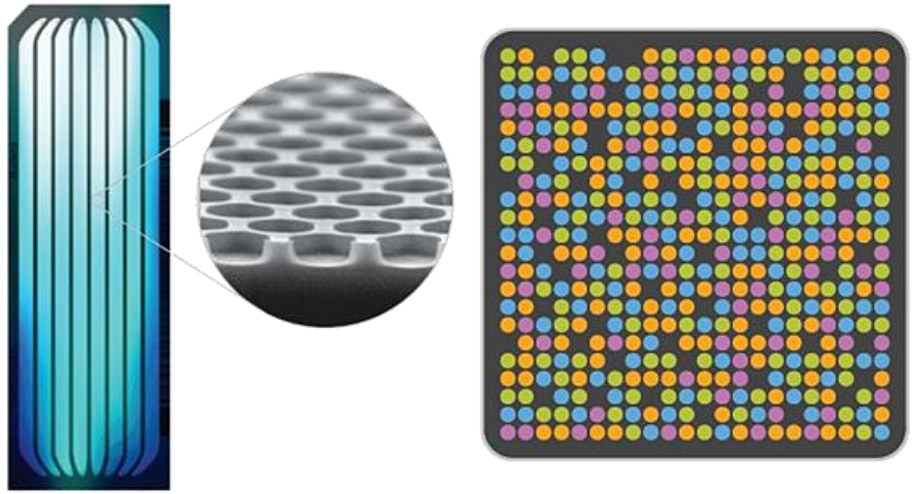
Ампліфікація та відбір фрагментів потрібної довжини

Оцінка кількості бібліотеки

World Health Organization
European Region

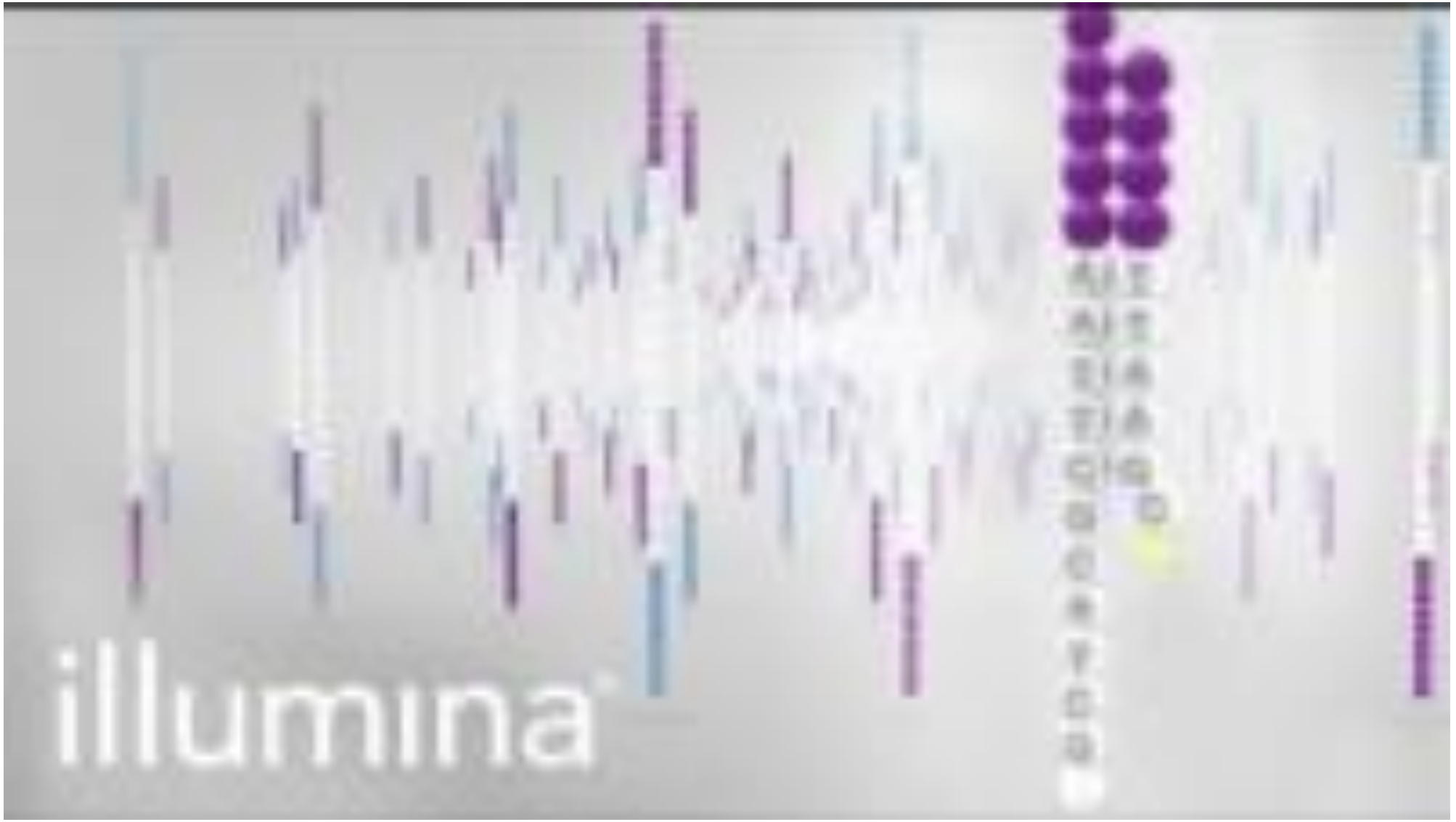
Секвенування Illumina

- Секвенування шляхом синтезу (SBS- Sequencing by synthesis)
- Технологія SBS використовує чотири флуоресцентно мічені нуклеотиди
- Генерування кластерів шляхом місткової ампліфікації
- Паралельне секвенування мільйонів ДНК-фрагментів на проточній лунці (flow cell)
- Короткі зчитування (до 300bp)
- Висока точність та низький рівень помилок



World Health Organization
European Region

Секвенування Illumina

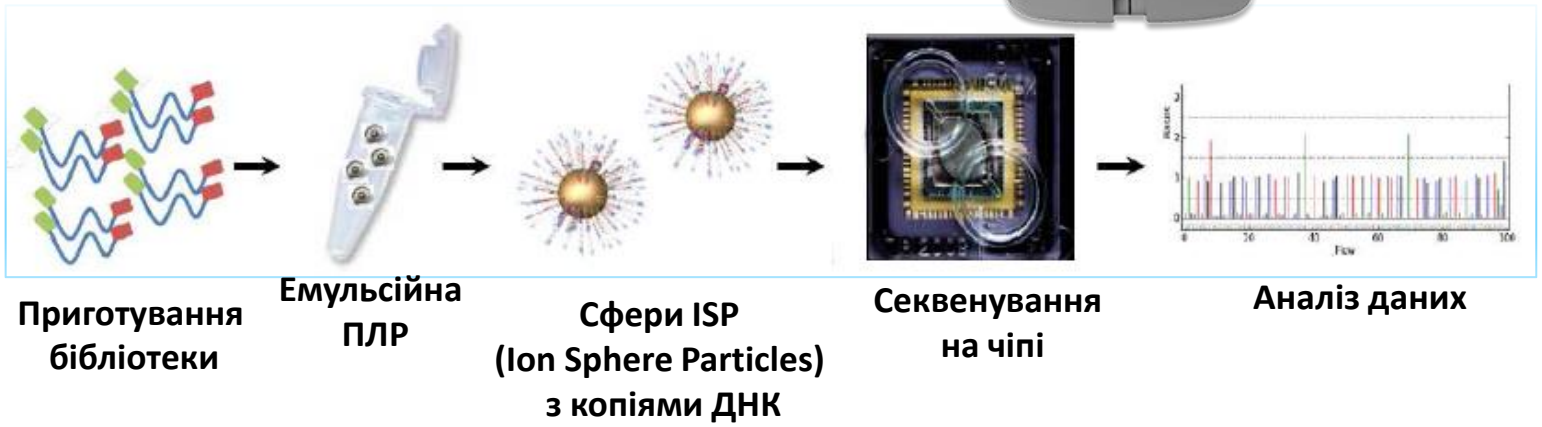


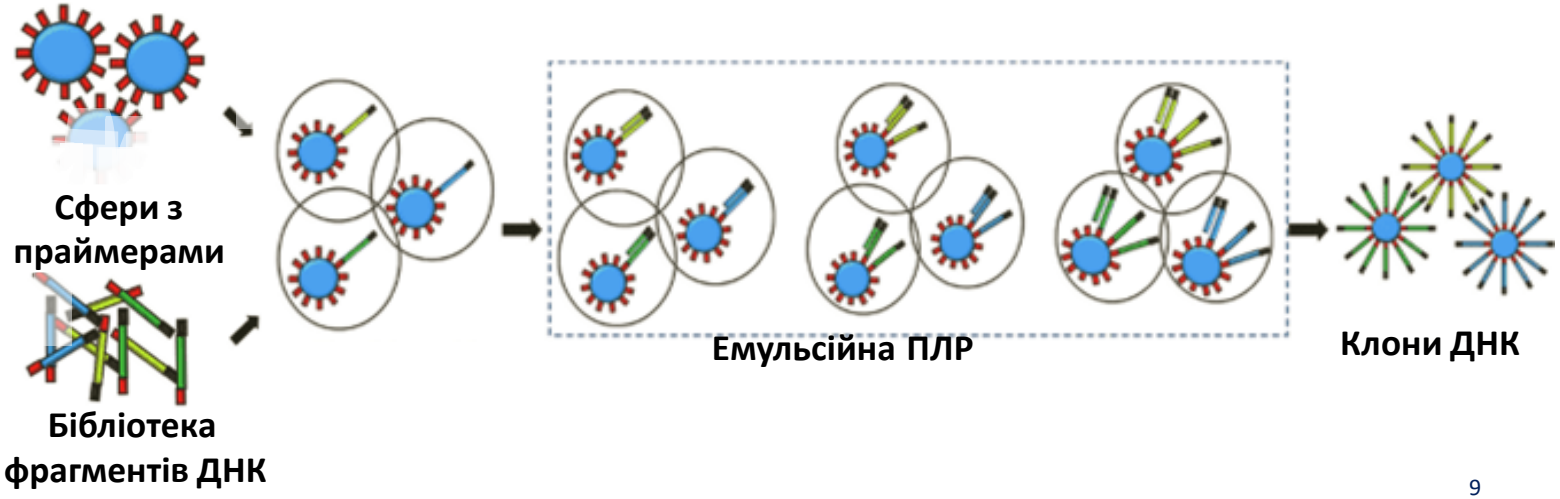
Відео:
https://www.youtube.com/watch?v=fCd6B5HRaZ8

World Health Organization
European Region

Секвенування Ion Torrent

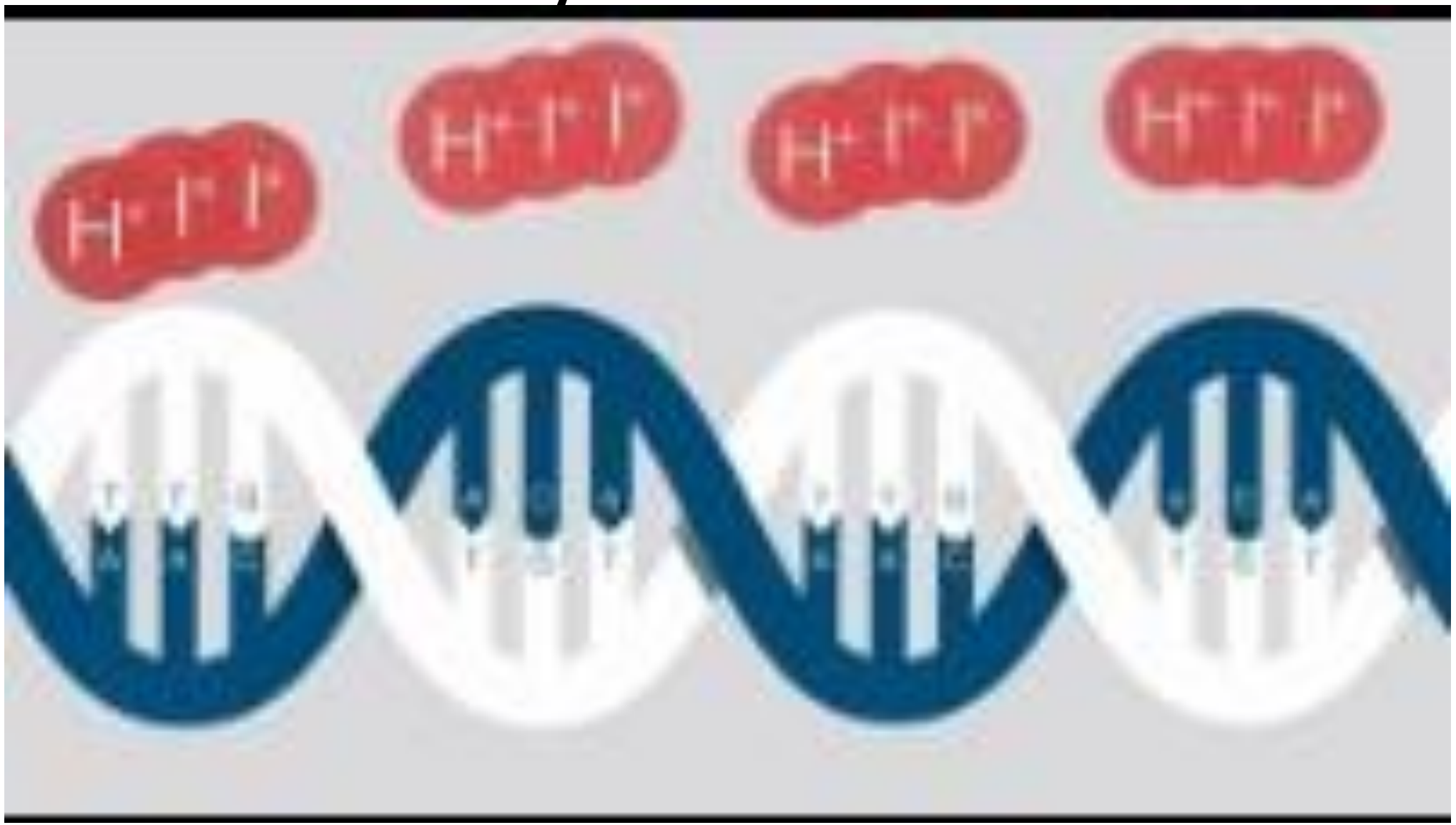
- Напівпровідникове секвенування на основі іонного потоку
(ion semiconductor sequencing)
- Засноване на реєстрації йонів водню, виділених в середовище у ході синтезу полімеразою ДНК-фрагментів > зміна pH
- Клональна ампліфікація шляхом емульсійної ПЛР (ePCR):
бібліотеки приєднуються до іонних сфер та шляхом ПЛР утворюється багато копій на поверхні сфери
- Короткі зчитування (200bp-600bp)
- Висока точність та низький рівень помилок





World Health Organization
European Region

Секвенування Ion Torrent

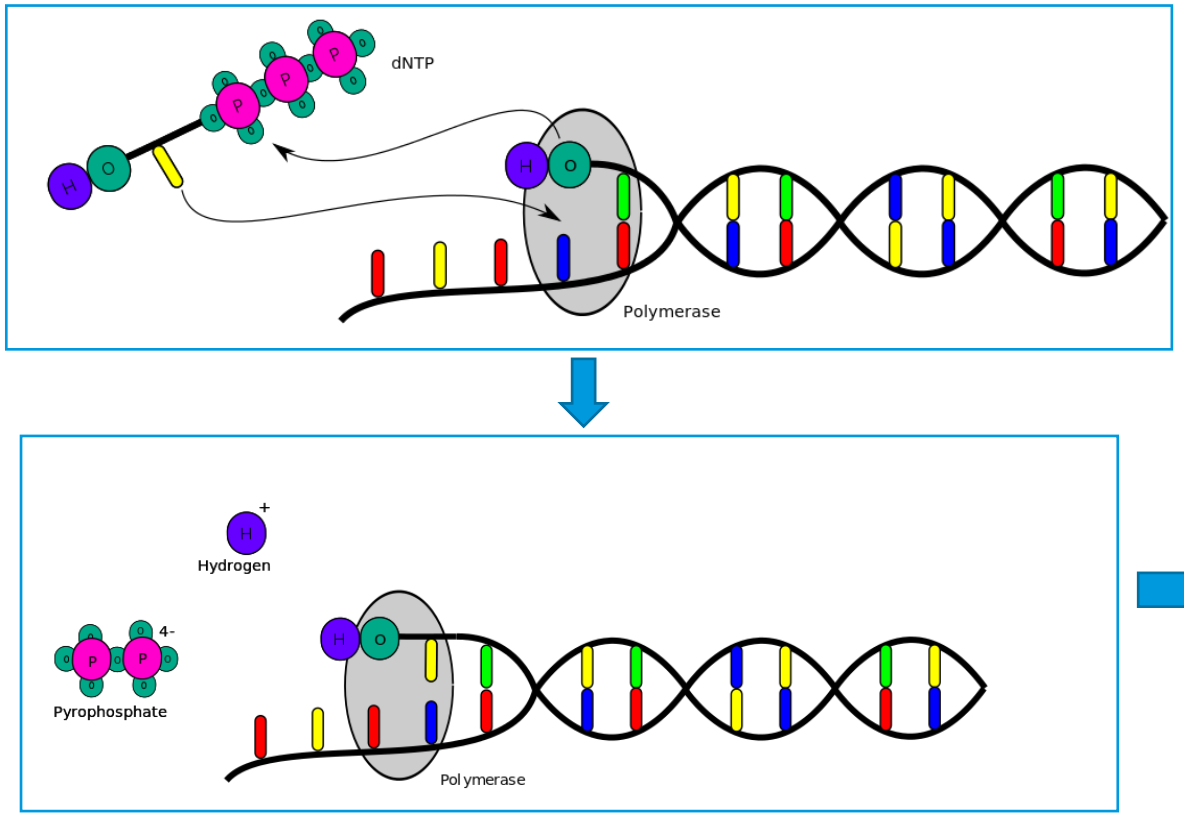


Відео:
https://youtu.be/zBPKj0mMcDg?t=25

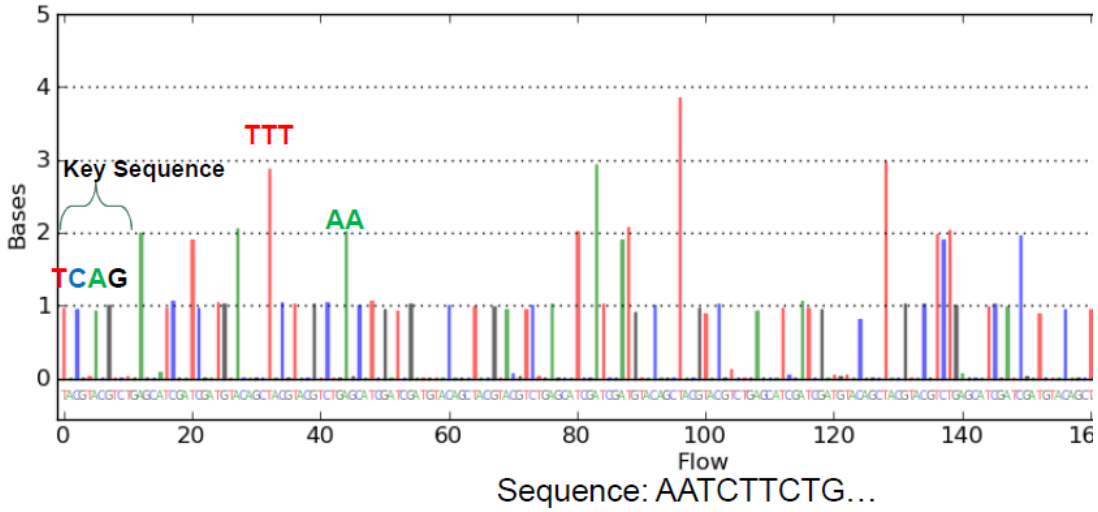
World Health Organization
European Region

Секвенування Ion Torrent

Реакція приєднання полімеразою нуклеотиду до ланцюга ДНК, внаслідок якої вивільняється водень і пірофосфат



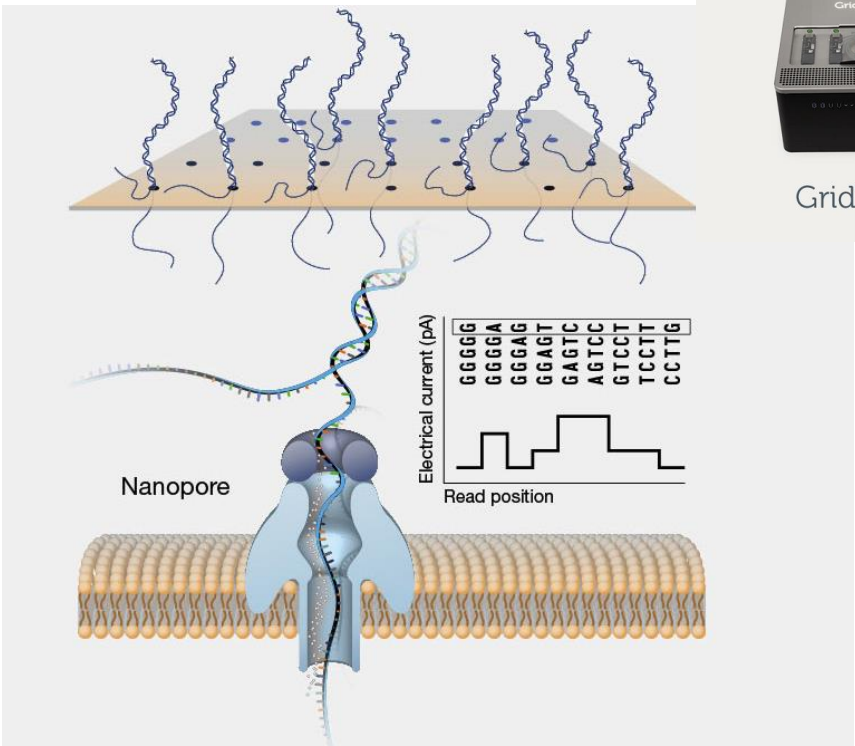
Детекція нуклеотидів в послідовності ДНК



World Health Organization
European Region

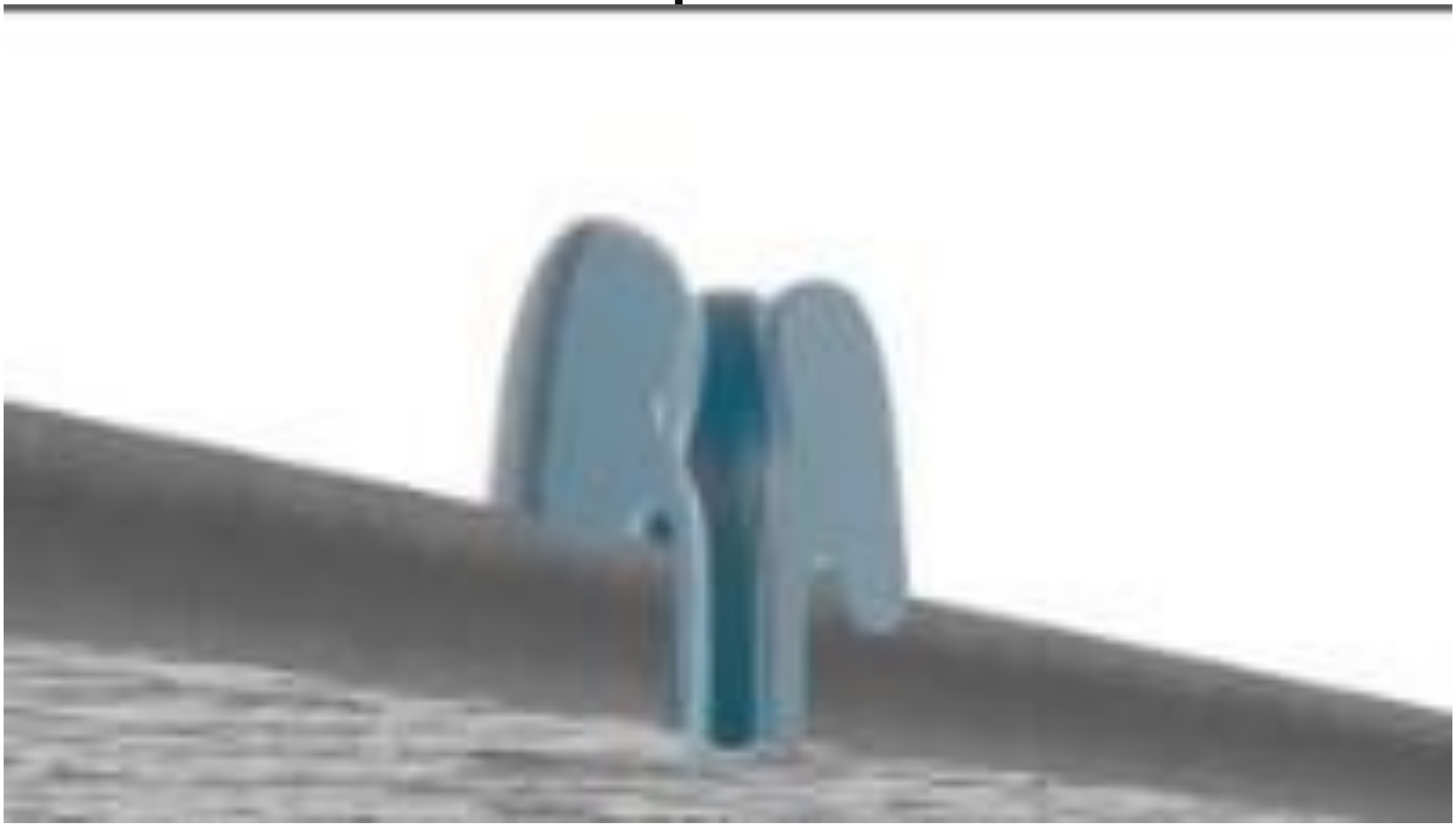
Oxford Nanopore

- **Базується на використанні нанопор для вимірювання зміни струму, що протікає через пору під час проходження ДНК-фрагментів, що дає змогу визначити послідовність нуклеотидів.**
- **Можливість секвенувати одну молекулу ДНК або РНК без необхідності ампліфікації.**
- **Довгі зчитування (тисячі bp), що дає змогу вирішувати задачі, пов'язані з дослідженням структури та функцій геномів.**
- **Метод дозволяє проводити секвенування в режимі реального часу.**
- **Однією з особливостей є портативність та можливість використання в польових умовах.**



World Health Organization
European Region

Oxford Nanopore



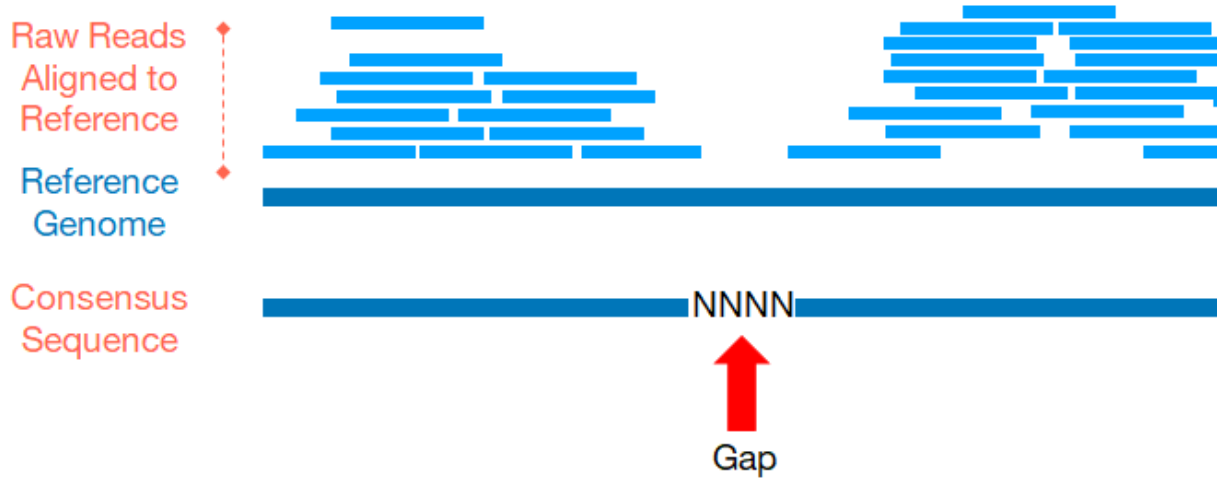
Відео:

https://www.youtube.com/watch?v=CGWZvHIi3i0

World Health Organization
European Region

Аналіз даних NGS

- **Аналіз даних NGS** - це процес обробки та інтерпретації великих обсягів даних, отриманих з NGS-платформ, з використанням біоінформатичних інструментів з метою отримання консенсусної послідовності ДНК.
- **Консенсусна послідовність** - це узагальнена послідовність досліджуваної ДНК або РНК, яка представляє найбільш поширену основу в кожній позиції.
- **Референсна послідовність геному** - формально визнана офіційна послідовність відомого геному, зібрана вченими як репрезентативний приклад геному окремого виду, використовується як стандарт для порівняння досліджуваної ДНК цього виду.



World Health Organization
European Region

Основні етапи аналізу даних

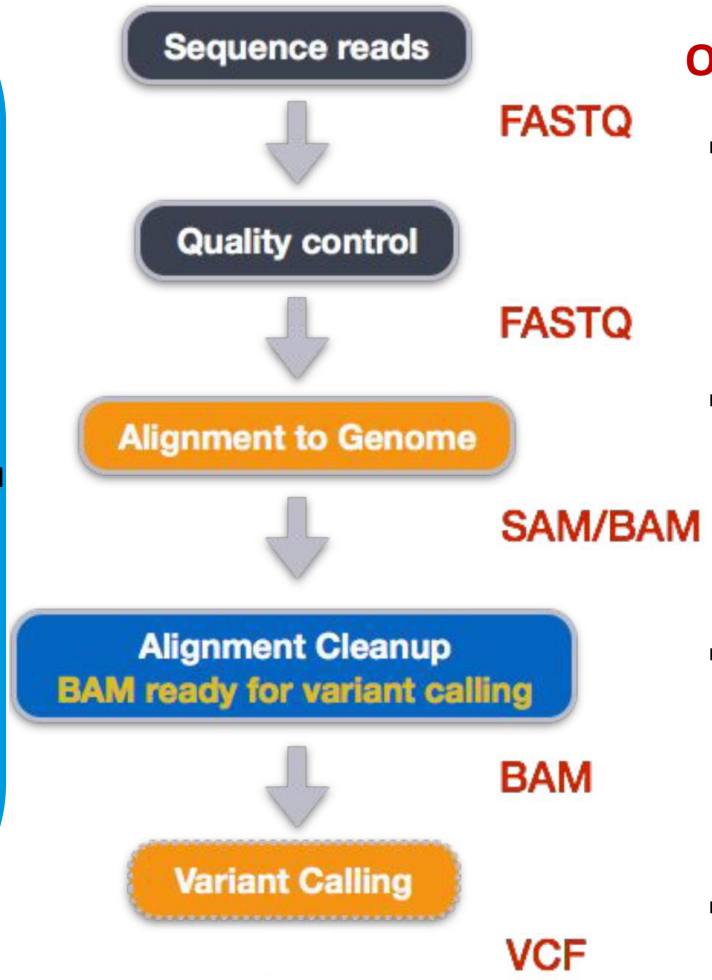
1. **Визначення послідовності сирих даних (Basecalling)** - перетворення реєстрованих сигналів з секвенатора в послідовність нуклеотидів (наприклад, в форматі FASTQ).

2. Оцінка якості даних (Quality control) - перевірка якості секвенування та обробки зчитувань, для визначення наявності помилок та можливих проблем, які потребують додаткової обробки.

3. Вирівнювання (Mapping/Alignment) –вирівнювання усіх зчитувань та на референсний геном (якщо такий використовується) та збірка консенсусної послідовності.

4. **Визначення варіантів (Variant calling)** - визначення відмінностей між секвенованою послідовністю та референсною послідовністю.

5.Інтерпретація результатів.



Основні формати файлів даних

- **FASTQ file:** зберігає послідовності зчитувань, які отримані під час секвенування, та показники якості визначення кожного нуклеотиду.
- **SAM (Sequence Alignment/Map)** – стандартний текстовий формат файлу для зберігання результатів вирівнювання (Alignment) послідовностей.
- **BAM (Binary Alignment/Map)** - бінарний аналог формату SAM, який забезпечує швидше зберігання та обробку великих об'ємів даних.
- **VCF (Variant Call Format)** - формат файлу для зберігання інформації про варіанти (SNPs, InDels) геному порівняно з референсом.

World Health Organization
European Region

Сирі дані: FASTQ file

1. **унікальний ідентифікатор для кожного зчитування**
2. **рядок послідовності ДНК у вигляді тексту**
3. **рядок заголовка (+)**
4. **рядок символів, які кодують показник якості для кожної основи.**

```
@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG
AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAAGAAATG
+
AAAFFJJJJJFJJ<J<FJJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<
@K00359:71:HJJL7BBXX:3:1101:2240:1508 1:N:0:ATCACG
GTAAAGGATGCGTAGGGATGGGAGGGCGATGAGGACTAGGATGATGGCGG
+
AAFFFJJJJJJJF<J7JJFJJJJJJFFFJFJJJJJJJJJJJJJJJJJJJ
@K00359:71:HJJL7BBXX:3:1101:2402:1508 1:N:0:ATCACG
GTCGACCATGTGGGCAGAACCTTGATGTTGGATTCCAGCAGGACCTGTCC
+
AAFFFJJJJJJJJ<JJJJJJJJJ<JFJJJJJJJJJJJJJJJFJJJJJJJ
@K00359:71:HJJL7BBXX:3:1101:2463:1508 1:N:0:ATCACG
ATGTGGTGTATGCATCGGGGTAGTCCGAGTAACGTCGGGGCATTCCGGAT
+
AAAFFFFJJJJJJJJJJJJJFJJJJJJJJJJJJJJJFJ7JJJJJJJJJJ
```

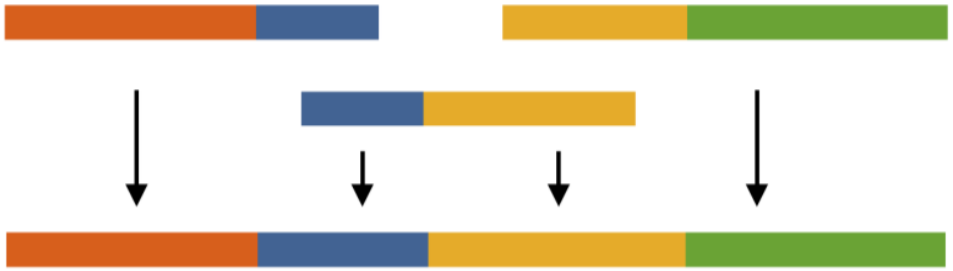
World Health Organization
European Region

Вирівнювання послідовностей (Alignment)

- **Вирівнювання на відому референсну послідовність**

- **De novo збірка геному:**

без попереднього знання послідовності шляхом виявлення взаємного перекриття зчитувань

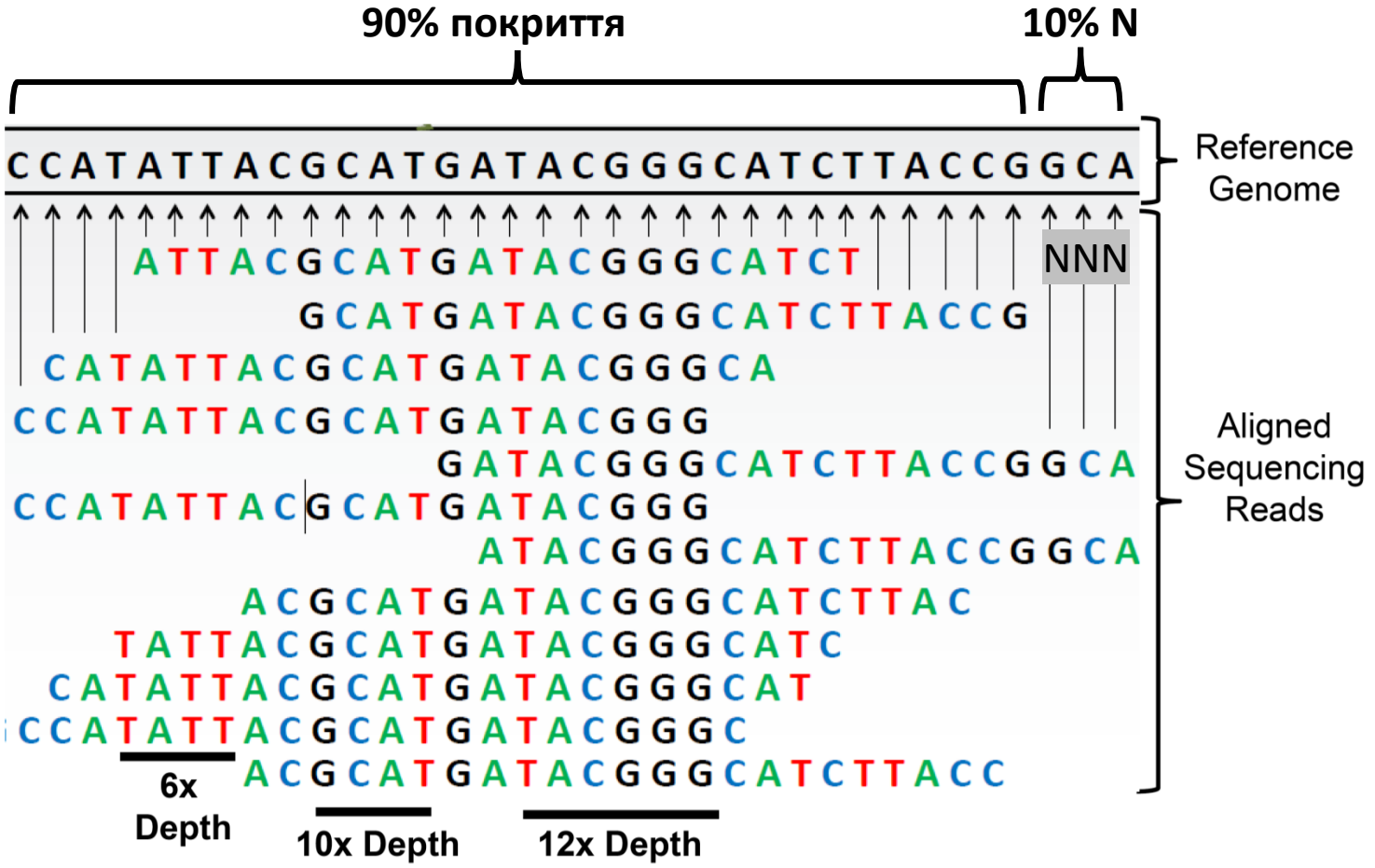


World Health Organization
European Region

Кількісні показники секвенування:

- **Довжина зчитування (ріда):** довжина секвенованого фрагмента в бібліотеці
- **Глибина покриття (Depth)** - кількість зчитувань вирівняних на окрему позицію геному (12x - 12-кратне покриття) Чим більша глибина зчитування, тим вища впевненість у визначеному нуклеотиді.
- **Середнє покриття (Coverage):** середня кількість вирівняних зчитувань, які перекривають усі позиції на цільовому геномі.
- **Загальне покриття послідовності** - % геному, яка має достатню кількість рідів для аналізу
- **% N** - Відсоток основ у кожній позиції без визначеного нуклеотиду

Приклад: консенсусний геном охоплює 90%, тобто ми ідентифікували послідовність для 90% геному, але маємо 10% N з відсутніми даними.



World Health Organization
European Region

Типи мутацій (варіантів)

Хромосомні

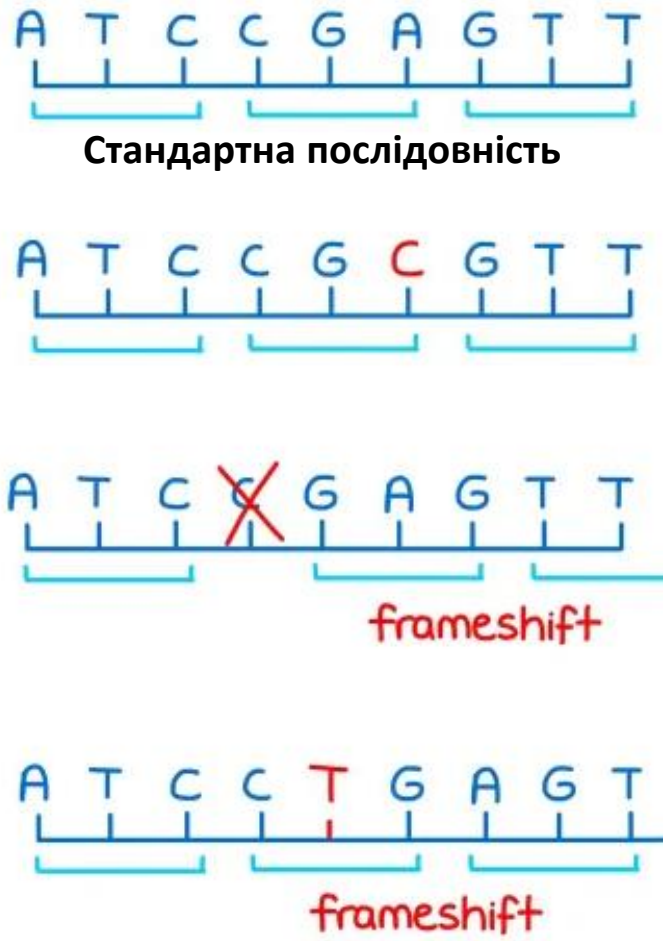
Точкові

Заміна: заміна одного нуклеотиду на інший

Делеція: втрата нуклеотиду

Інсерція: вставка додаткового нуклеотиду

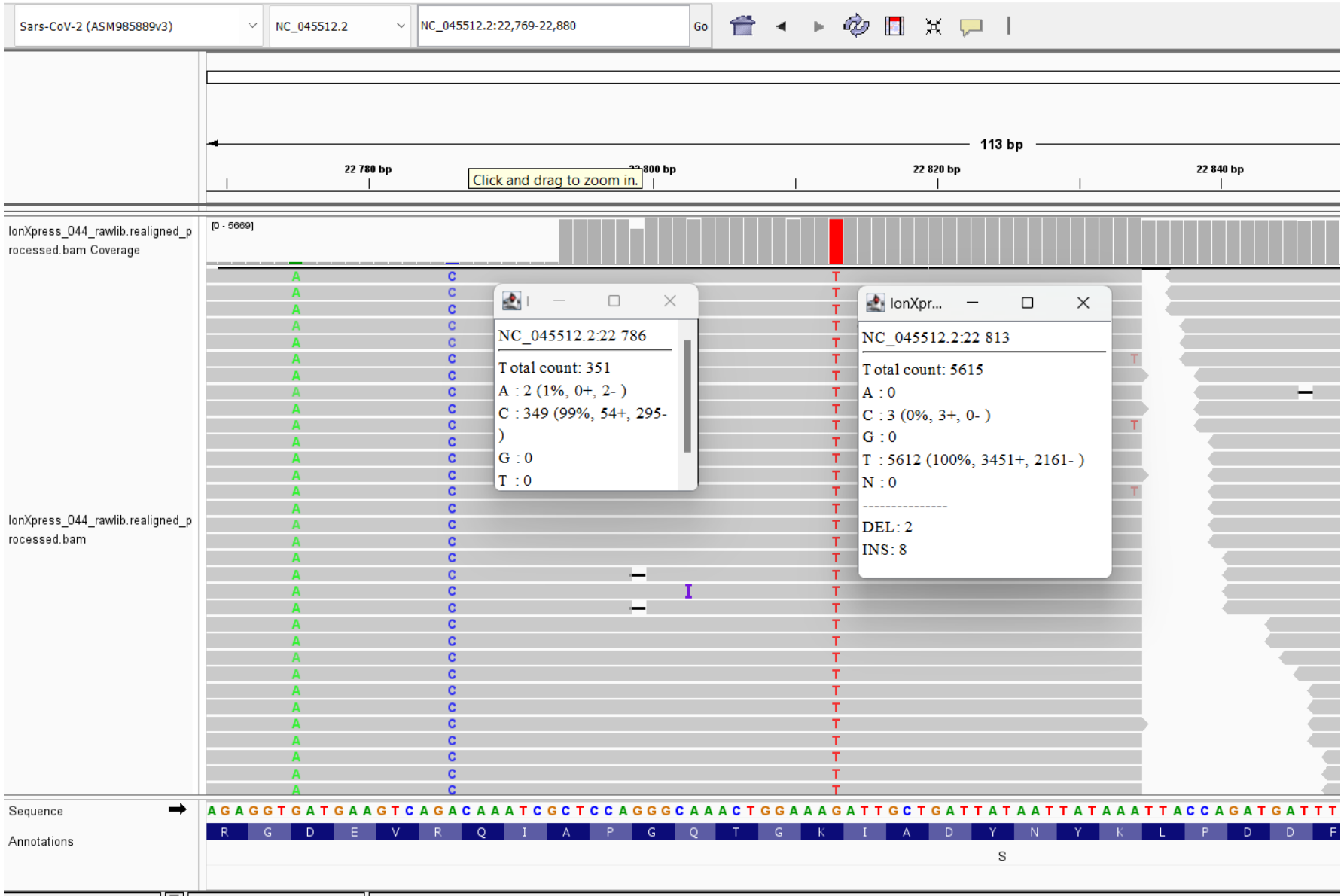
Зсув рамки зчитування (Frameshift): коли додається або видаляється нуклеотид, змінюється вся послідовність кодонів, що йде за мутацією.



World Health Organization
European Region

IGV

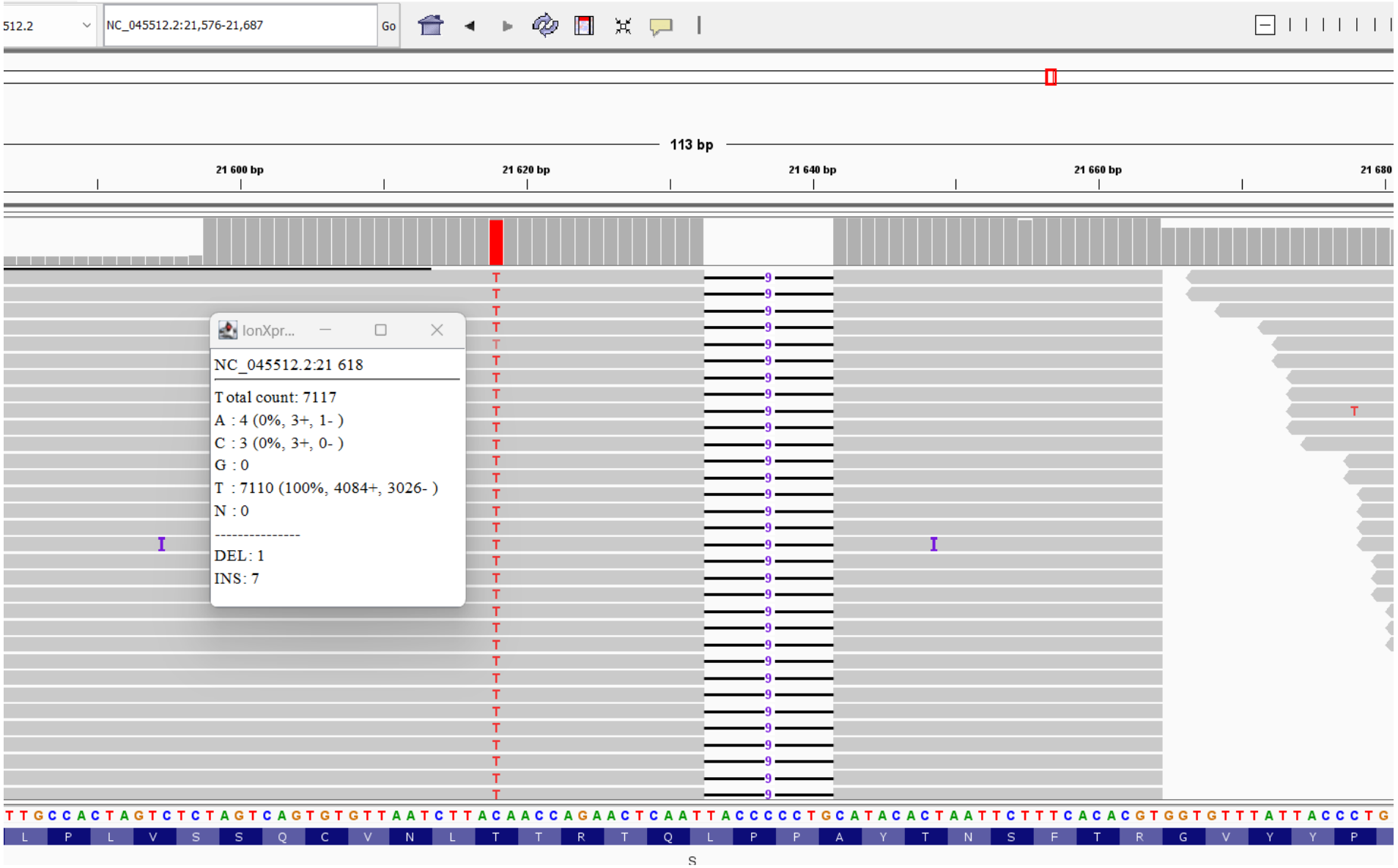
(Integrative Genomics Viewer) – Інтегрований геномний переглядач



World Health Organization
European Region

IGV

(Integrative Genomics Viewer) – Інтегрований геномний переглядач



World Health Organization

European Region

Формат FASTA

- **простий текстовий формат для файлів послідовностей ДНК і білків.**
- **Кожній послідовності передує рядок заголовка, який починається з «>» і за яким йде назва.**
- **Будь-який інший текст після першого рядка вважається частиною послідовності.**
- **Один файл може містити багато послідовностей**

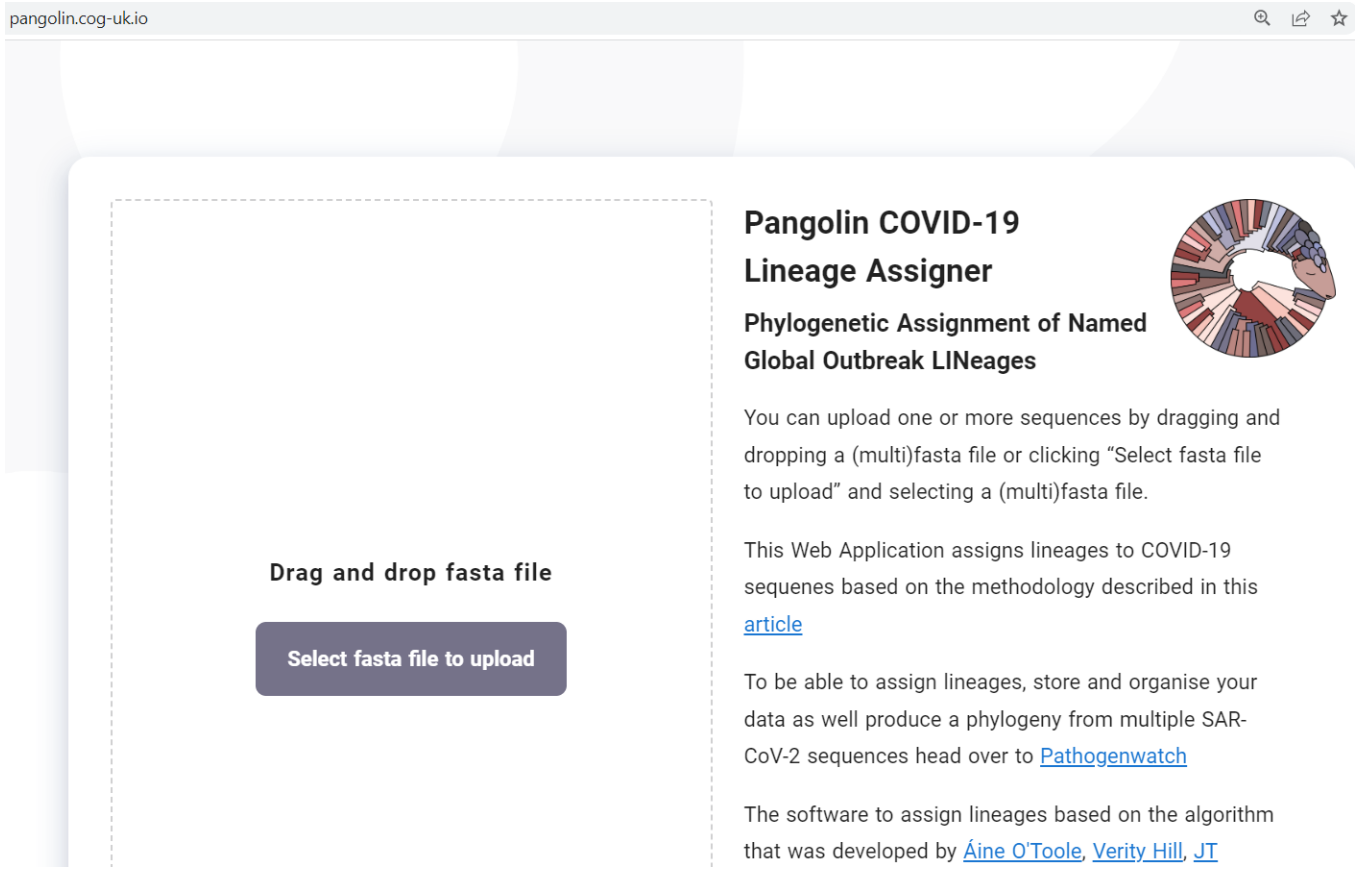
```
>hCoV-19/Ukraine/74373/2023
TTGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTA
TAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCG
CAGCANNNNNNNNNNNNNNNNNNNNNNNNNNNNAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAG
TTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAG
GATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCAT
CTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGT
CCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTG
GATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAA
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGC
GTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTG
TGCTGCCGTGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGAC
CAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACT
GAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATGCAACC
ATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTG
AAGAAGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCAC
GCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTG
TCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGGTTGTAACCA
GTTCCGAAGGTCTTAATGACAACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGT
NNNNNNNNNNNNNNNNNNNNNNNNNNNGCTTCCACAAGTGCTTTTGTGGAAACTGTGAAAGGTTTGGATT
GTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAGCTAAAAAAGGTGCCTGGAATATTGGTGAACA
TTTATGCATTTGCATCAGAGGCTGCTCGTGTTGTACGATCAATTTTCTCCCGCACTCTTGAAACTGCTCAA
GAAGGCCGCTATAACAATACTAGATGGAATTTCACAGTATTCACTGAGACTCATTGATGCTATGATGTTCA
```

World Health Organization
European Region

Pangolin – визначення варіантів SARS-CoV-2

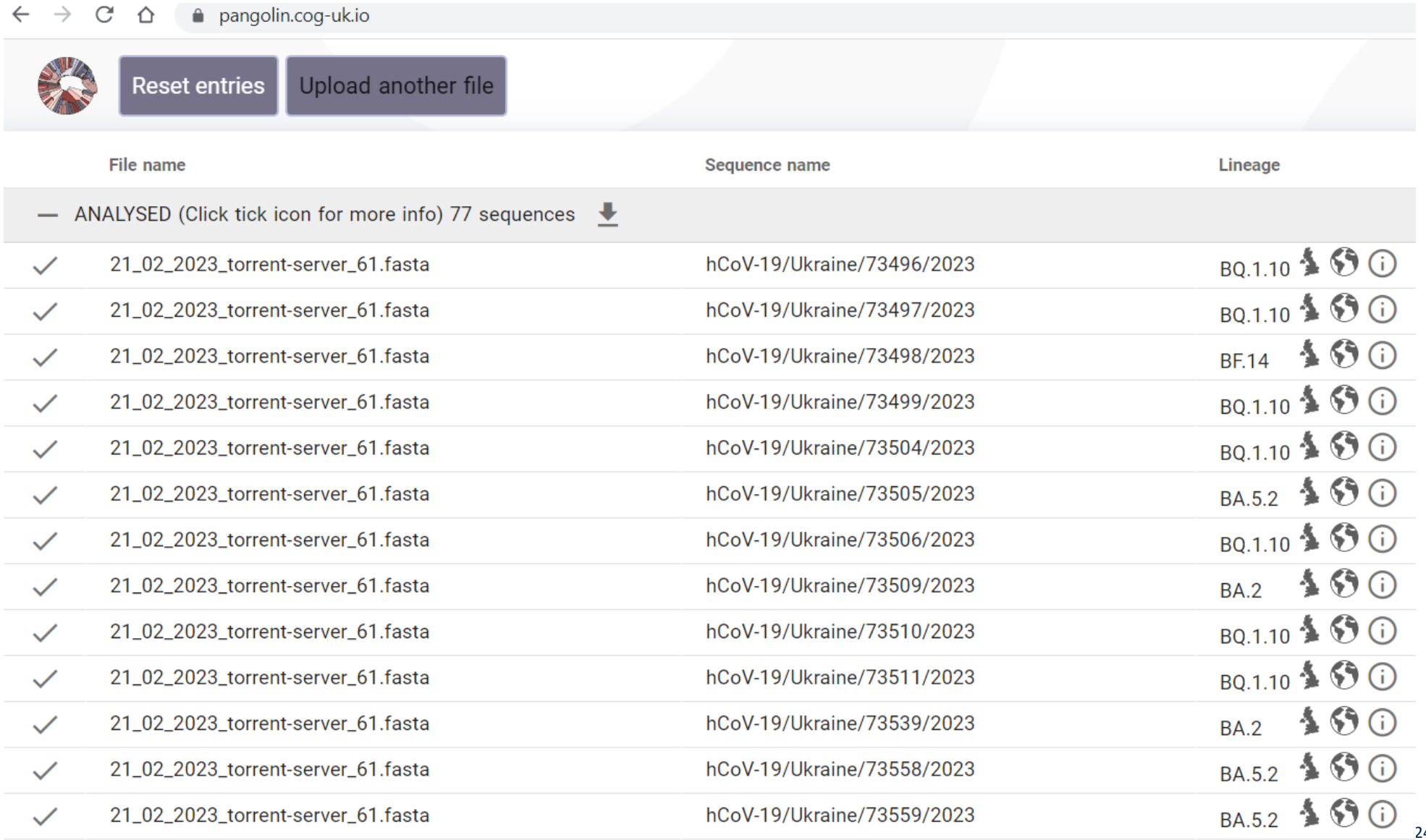
Варіанти ДНК ≠ варіанти SARS-CoV-2

- **Варіант ДНК/РНК (мутація) - це відмінність в генетичній послідовності в порівнянні з референсною послідовністю.**
- **Варіант SARS-CoV-2 - це вірусний геном, який відрізняється від інших вірусів однією або кількома мутаціями. Група вірусів з подібною послідовністю та загальними характеристиками (вірулентність, інфекційність тощо)**



World Health Organization
European Region

Pangolin – визначення варіантів SARS-CoV-2



File name	Sequence name	Lineage
— ANALYSED (Click tick icon for more info) 77 sequences		
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73496/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73497/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73498/2023	BF.14
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73499/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73504/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73505/2023	BA.5.2
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73506/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73509/2023	BA.2
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73510/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73511/2023	BQ.1.10
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73539/2023	BA.2
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73558/2023	BA.5.2
✓ 21_02_2023_torrent-server_61.fasta	hCoV-19/Ukraine/73559/2023	BA.5.2

World Health Organization
European Region

Бази даних

1. **Послідовність геному**
2. **Метадані** - набір даних, який дає інформацію про первинні дані

   -додаткові дані про зразок, з якого була отримана послідовність ДНК/РНК:

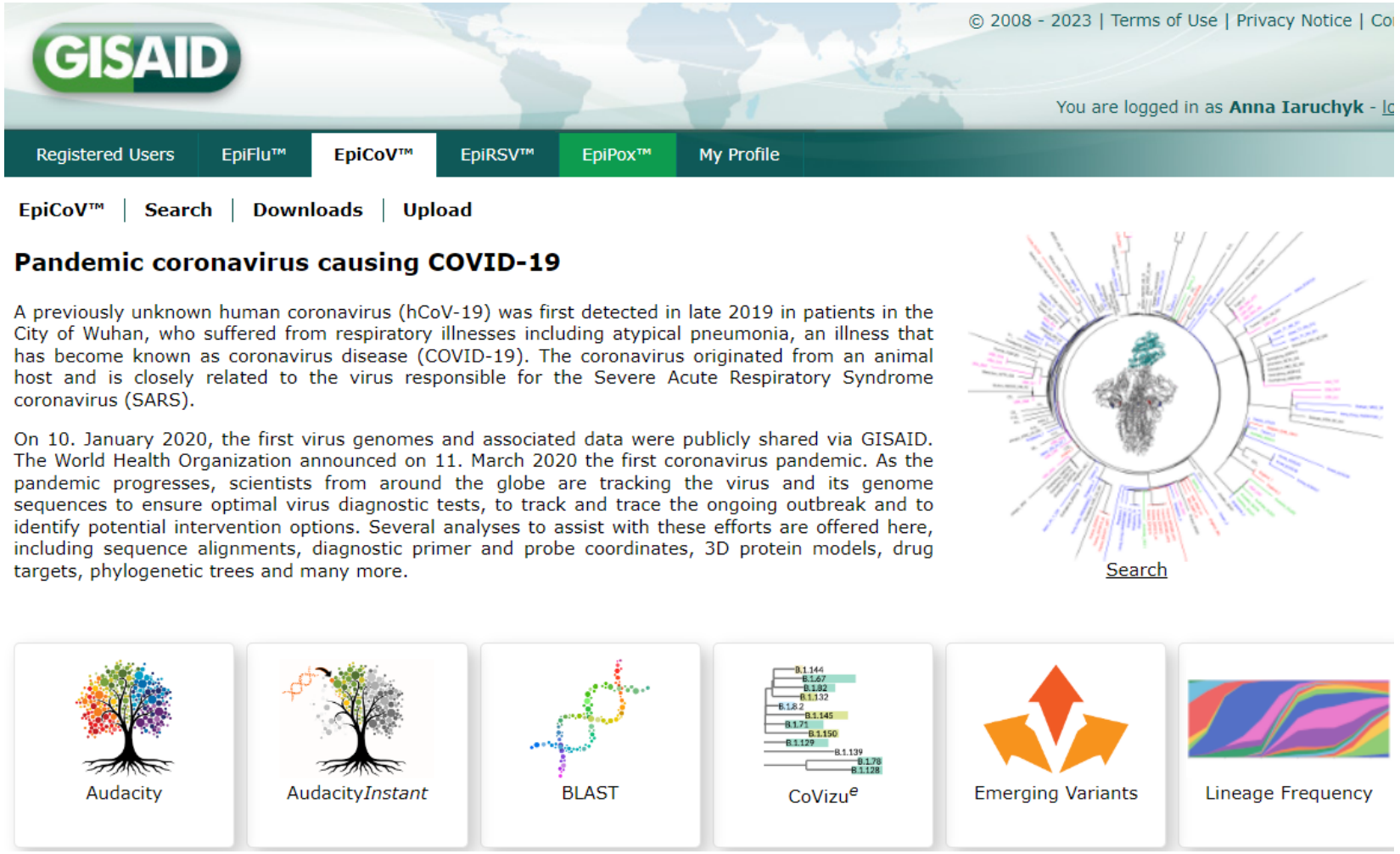
- Дата забору
- Географічні дані
- Організм
- Вид біоматеріалу
- Стать і вік пацієнта
- Статус вакцинації пацієнта
- інші дані

! Метадані повинні бути стандартизовані між усіма лабораторіями

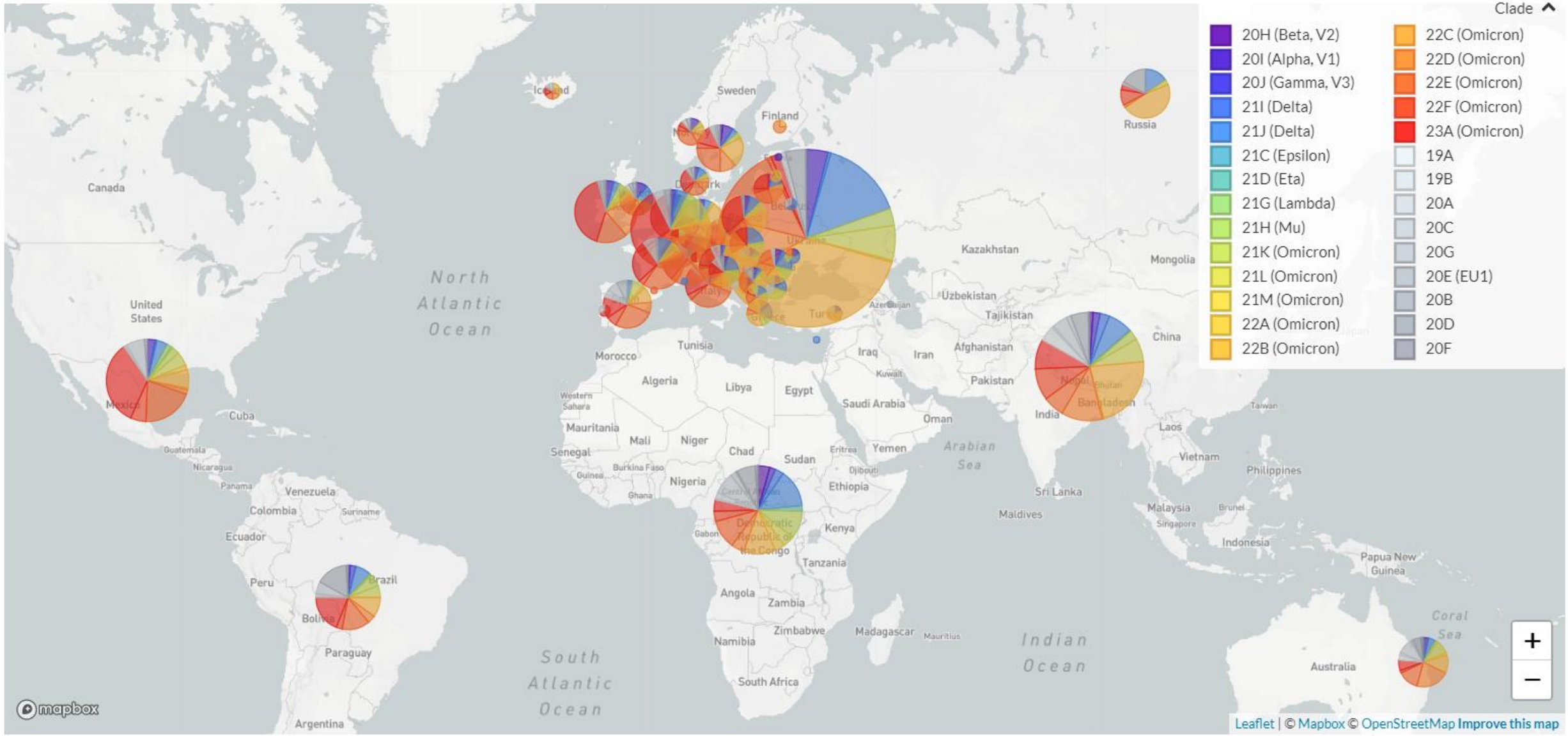
World Health Organization
European Region

База даних GISAID

- це глобальна науково-дослідна ініціатива та основне джерело геномних даних і метаданих усіх вірусів грипу, респіраторно-синцитіального вірусу (RSV) і SARS-CoV-2



World Health Organization
European Region

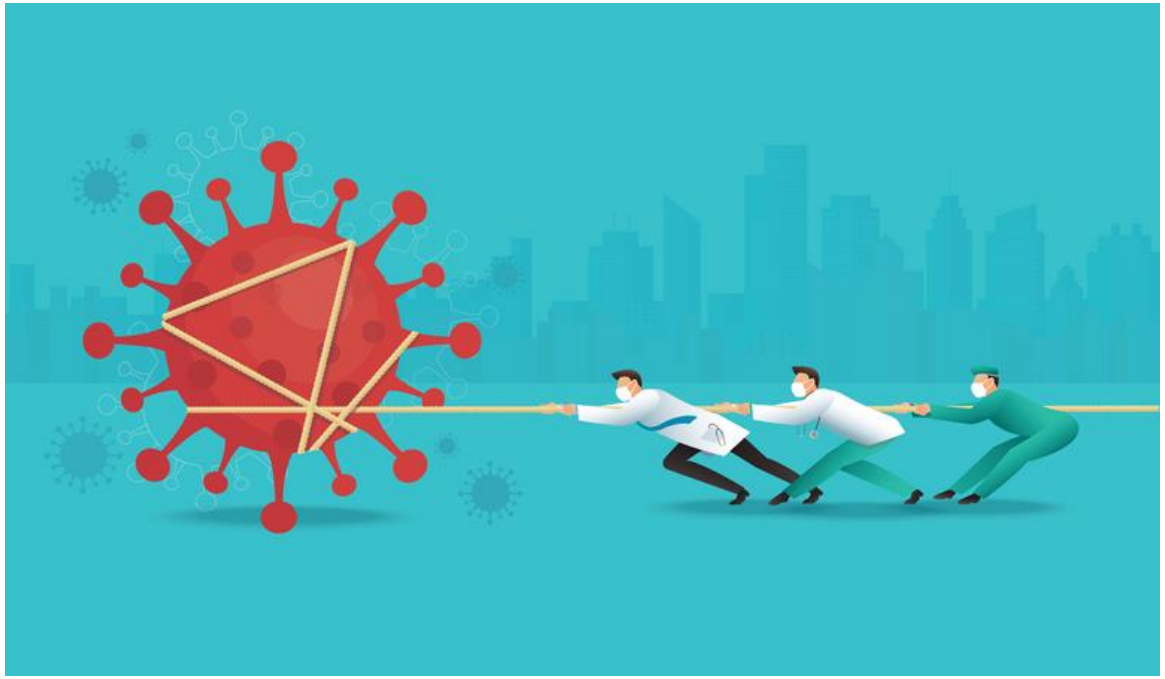


World Health Organization
European Region

https://next.nextstrain.org

Значення NGS та геномного нагляду в охороні здоров’я

- Дослідження походження інфекційних захворювань
- Виявлення спалахів/епідемій/пандемій інфекційних захворювань та управління ними
- Дослідження резистентності мікроорганізмів
- Розуміння джерел/резервуарів
- Виявлення нових патогенів
- Підтримка розробки діагностики, ліків і вакцин, а також інформування про реакцію на спалах



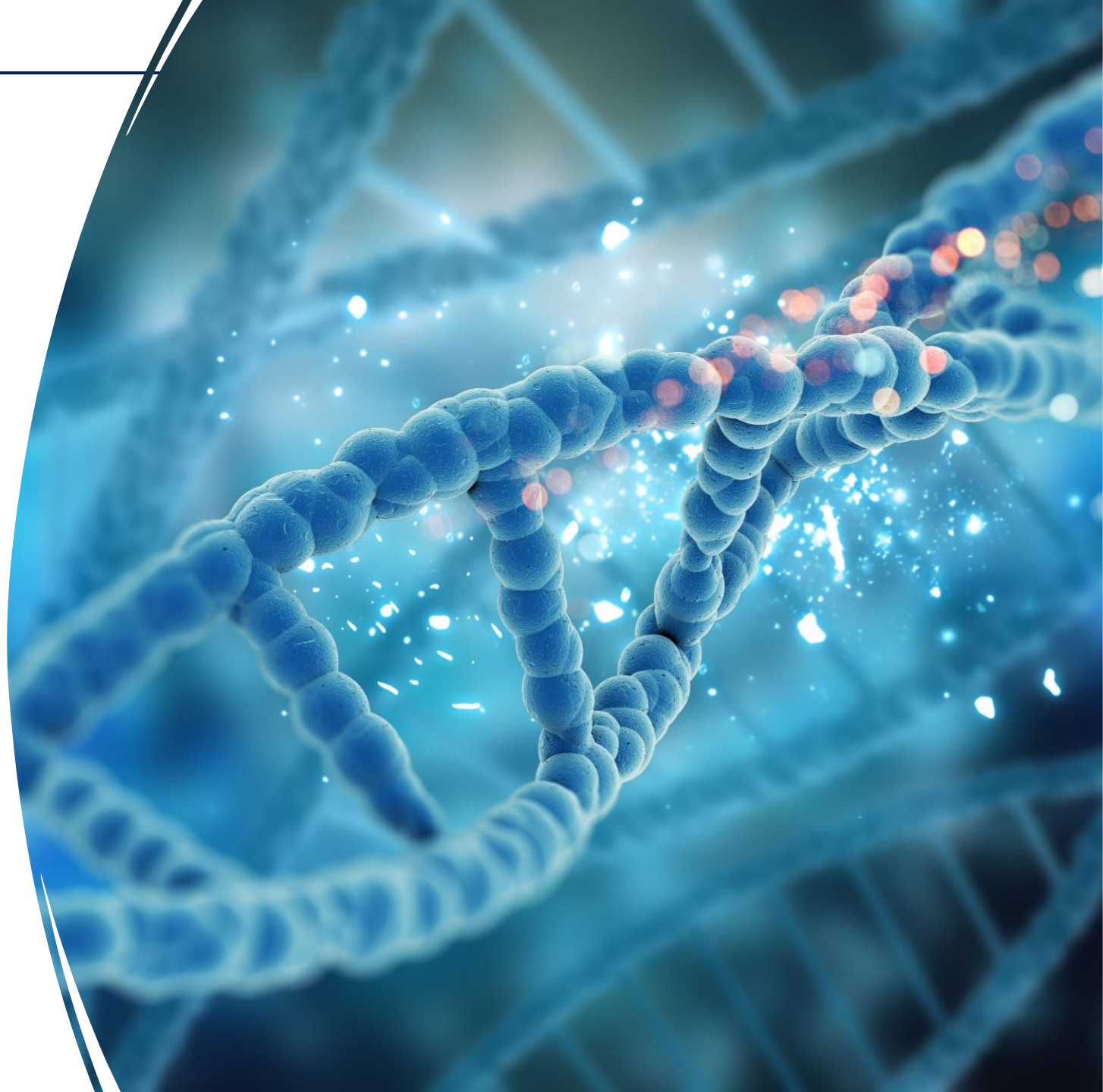
World Health Organization
European Region

Дякую за увагу!

Анна Яручик

Експерт Бюро ВООЗ в Україні

iaruchyka@who.int



World Health Organization

European Region