2025 한국식물분류학회

Technical Workshop

Plant organelle genome assembly

2025년 8월 5일 화요일

Workshop approach

1. eu.galaxy + galaxy.sungshin
2. rstudio.sungshin + Winscp
3. terminal + Winscp + Windows editor
4. terminal + Linux editor

Made by OpenAI ChatGPT

First things to do

- Create a user account at two kinds of websites for the workshop class:
  - usegalaxy.eu
  - usegalaxy.org
  - positCloud
- Assess users’ computer accessibility

galaxy: bioinformatics platform

- Two public galaxy websites: create your accounts
  - usegalaxy.org
  - https://usegalaxy.eu/
- Polap galaxy
  - galaxy.sungshin.ac.kr

positCloud

- https://posit.cloud/
  - create your account
- http://113.198.12.168:8787/
- ID: tw2025

About positCloud

1. Posit cloud - signup & login
2. Posit cloud - RStudio 환경
3. Posit cloud -- linux terminal

Made by OpenAI ChatGPT

About galaxy

1. galaxy - account creation
2. galaxy - account activation
3. galaxy - log-in
4. galaxy - tools & history
5. Posit cloud - signup & login

Made by OpenAI ChatGPT

YouTube demonstration

1. Bandage - select contigs
2. Bandage - extract a plastid sequence
3. Polap - seed contig selection using Bandage - Case: mtDNA of Vigna radiata
4. polap x galaxy 01 whole-genome assembly
5. polap x galaxy 02 organelle genome assembly
6. Organelle Genome Assembly of *Carex pseudochinensis* Using Galaxy Web Interface

Made by OpenAI ChatGPT

유전체 프로젝트

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 2019 Nov 17;18:9-19.

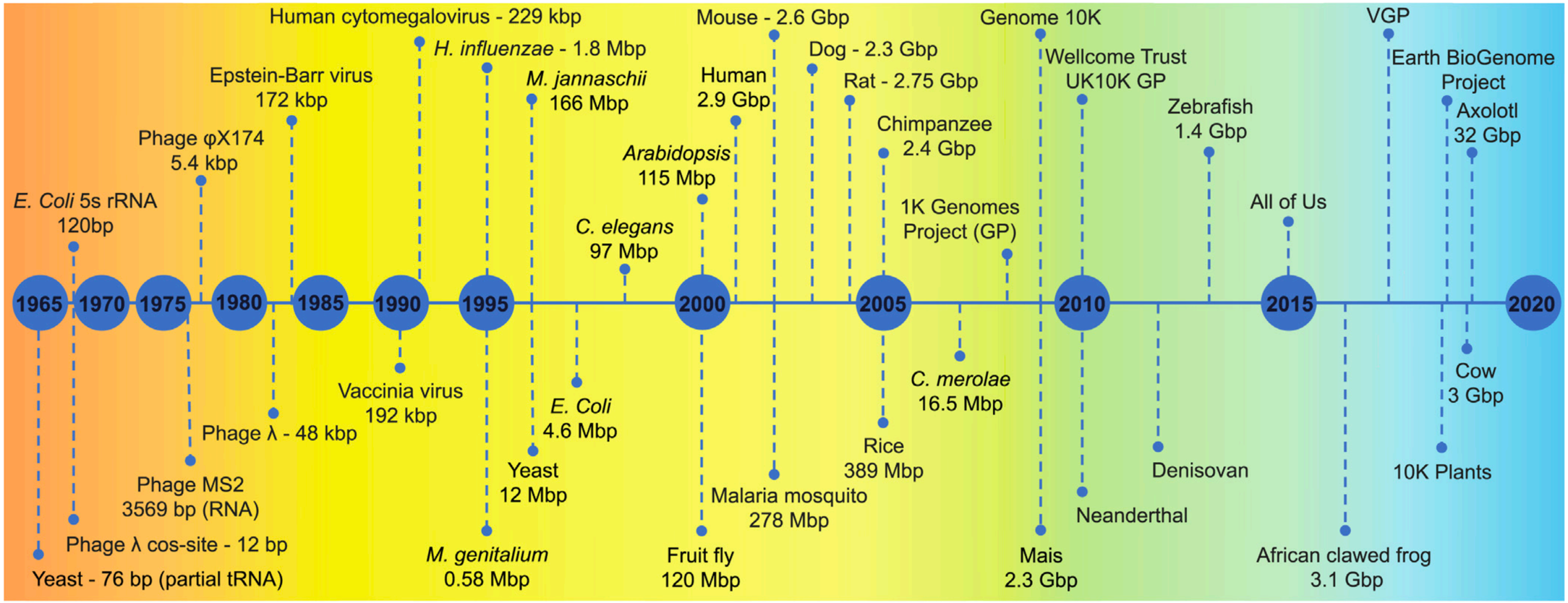


Fig. 1. Milestones in genome assembly. Timeline illustrating many of the major genome assembly achievements ranging from the beginning of the sequencing era to the large-scale genome projects currently ongoing. Each genome or genome project (GP) is placed under a color-coded background according to the sequencing approach adopted. Light red: early sequencing methods, Yellow: Sanger-based shotgun sequencing, Green: NGS, Light blue: TGS.

Sequencing technology

- DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule.
- Key for understanding genetic information and its applications in medicine, biology, and forensics.
- Three generations of sequencing technology
  - Sanger Sequencing
  - Next-Generation Sequencing (NGS)
  - Third-Generation Sequencing

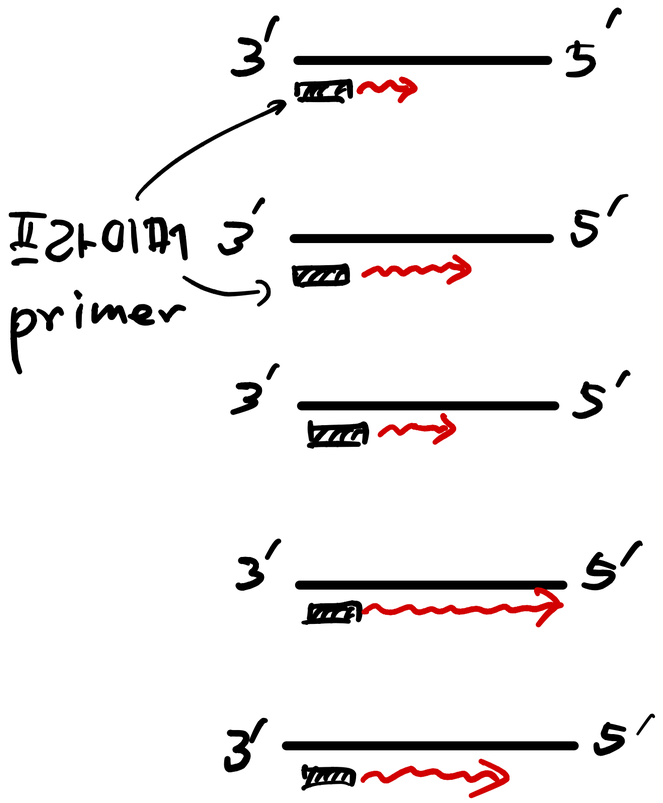
Made by OpenAI ChatGPT

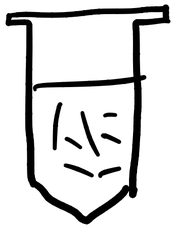
Sanger Sequencing

- Developed in the 1970s, considered the first generation of sequencing.
- Uses chain-terminating dideoxynucleotides.
- Pros: High accuracy, still used for small-scale projects.
- Cons: Slow, expensive for large genomes.
- Applications: Still used in clinical diagnostics and small-genome sequencing.

Made by OpenAI ChatGPT

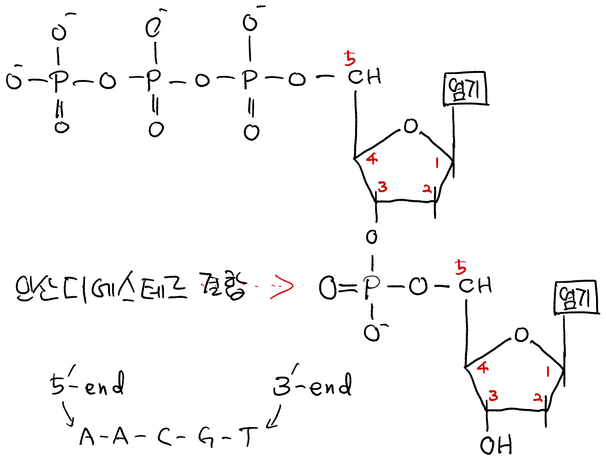
사슬종결 염기 서열 결정법: 생어 시퀀싱





\+ dNTP

\+ ddNTP



International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001)

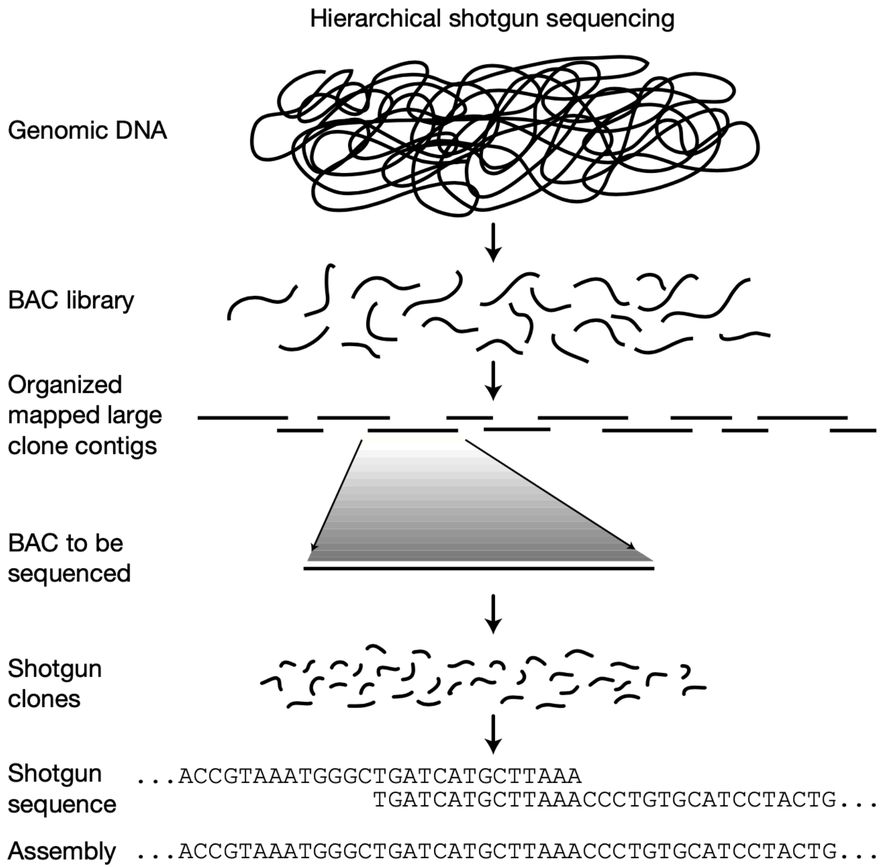


Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.

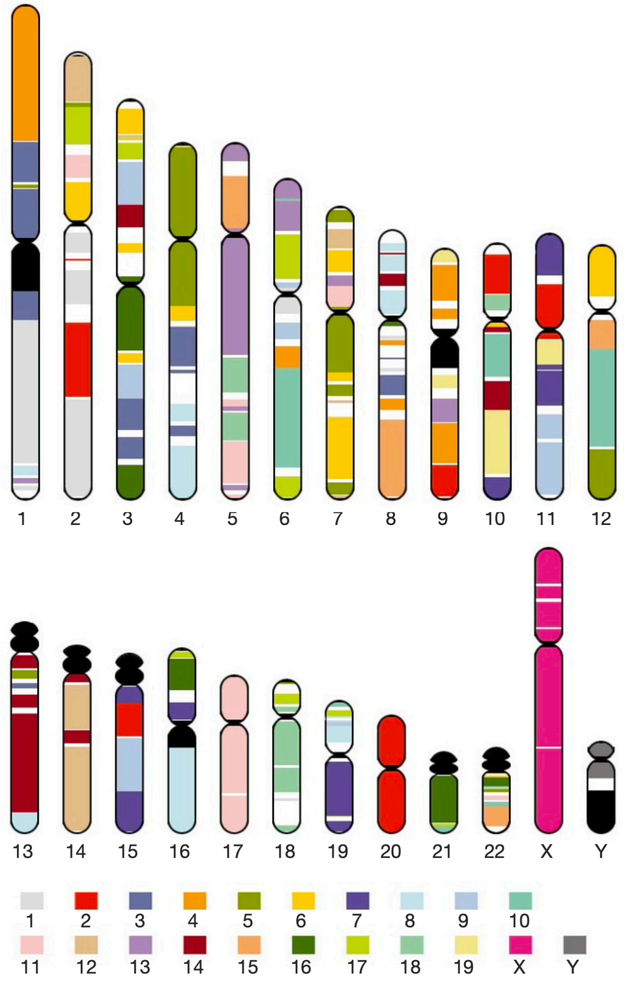
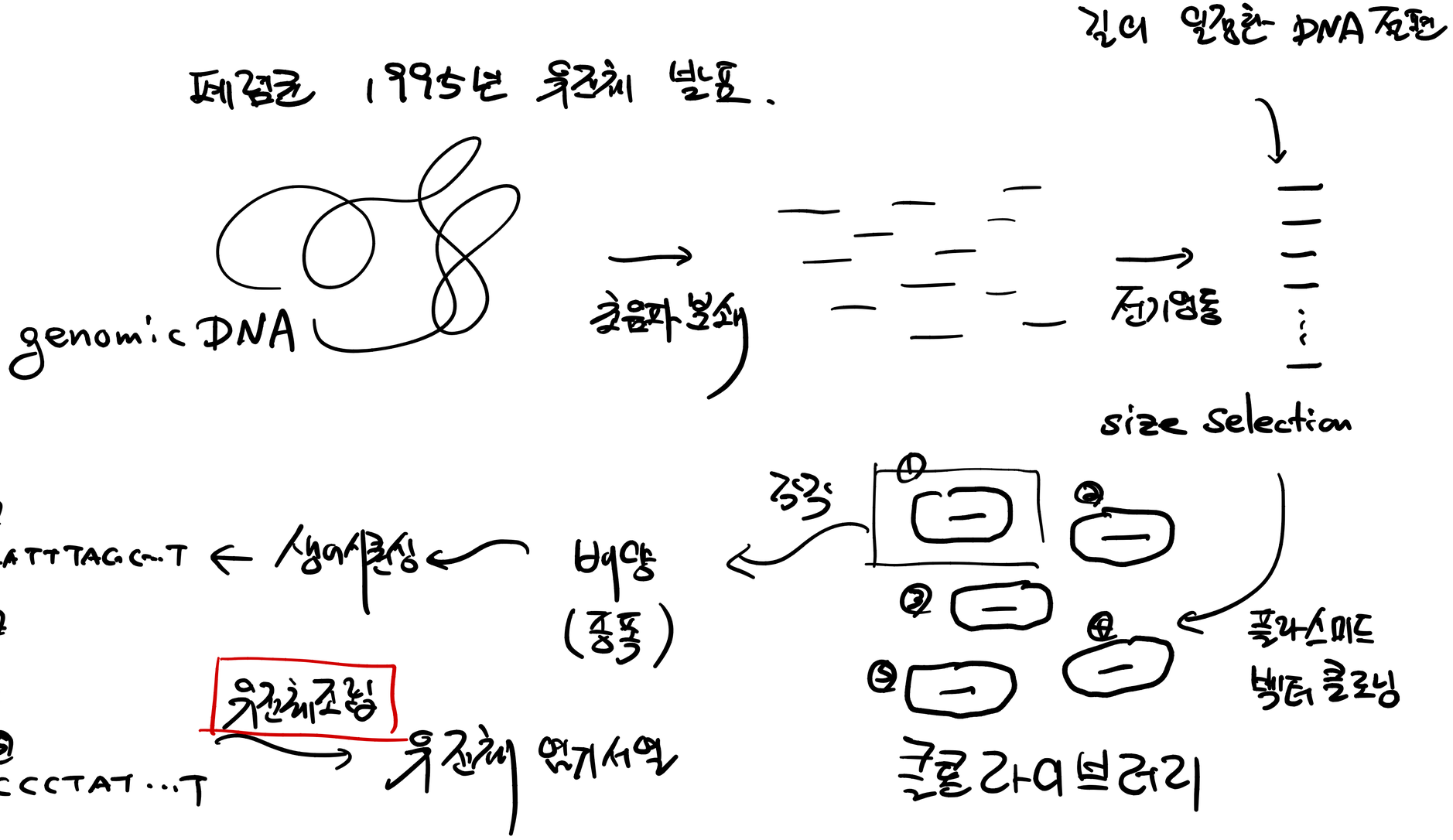


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

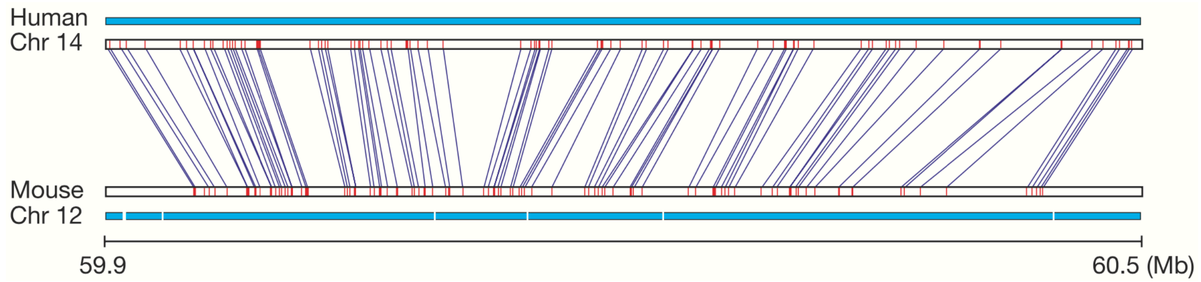
샷건 시퀀싱


최초의 완전한 DNA 염기서열
1995년 유전체 발표.
genomic DNA
초음파 분쇄
전기영동
size selection
① AATTTAGC~T
②
⑤ CCCTAT…T
생어시퀀싱
버퍼
(증폭)
유전체조립
유전체 염기서열
정상
플라스미드
벡터 클로닝
클론라이브러리

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium*



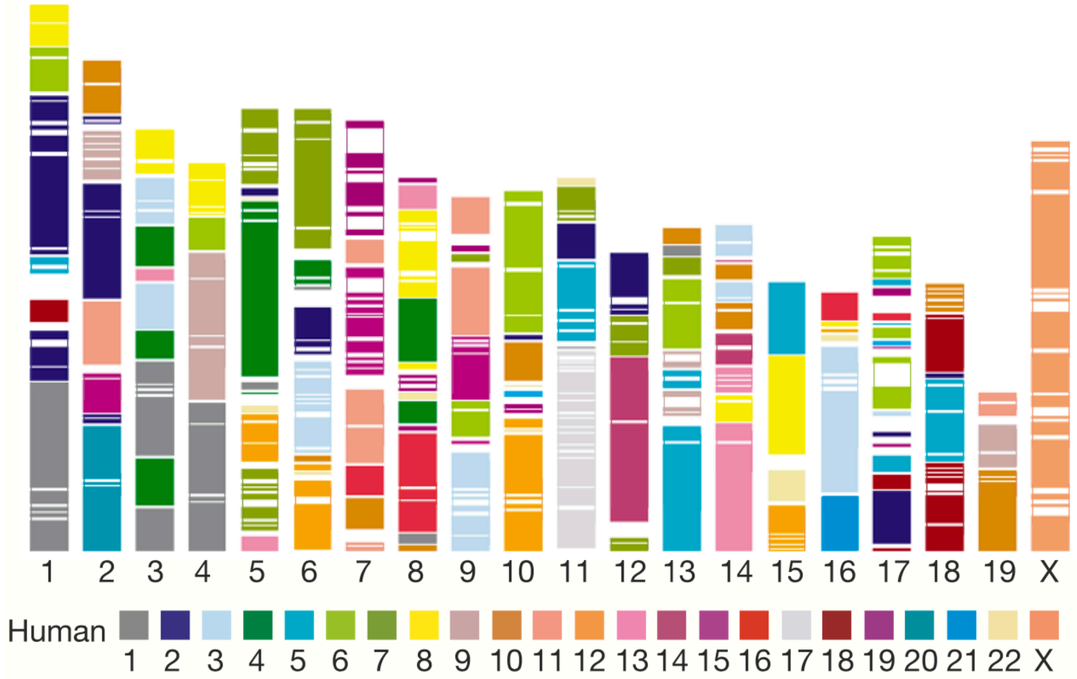


Figure 3 Segments and blocks >300 kb in size with conserved synteny in human are superimposed on the mouse genome. Each colour corresponds to a particular human chromosome. The 342 segments are separated from each other by thin, white lines within the 217 blocks of consistent colour.

Next-Generation Sequencing (NGS)

- Massively parallel sequencing, faster and cheaper than Sanger.
- Examples: Illumina, Ion Torrent.
- Pros: High throughput, cost-effective for large-scale sequencing.
- Cons: Requires more computational resources.
- Applications: Whole-genome sequencing, RNA-Seq, personalized medicine.

Made by OpenAI ChatGPT

유전체 프로젝트들

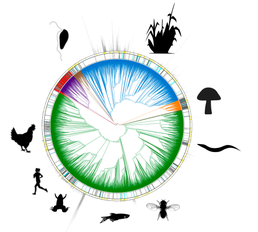


Fig. 1. Current status of the sequencing of life. Open Tree of Life (opentreeoflife.org) synthesis of phylogeny for all of life with resolution to the genus level, and showing phylogenetic information for Archaea (red), Bacteria (purple), Fungi (orange), Plantae (blue), Protista (brown), and Animalia (green). The current state of genomic

- **Earth BioGenome Project**: Sequencing life for the future of life. PNAS. April 23, 2018.115 (17) 4325-4333

- Genome 10K Community of Scientists, **Genome 10K**: A Proposal to Obtain Whole-Genome Sequence for **10000 Vertebrate** Species, Journal of Heredity, Volume 100, Issue 6, November-December 2009, Pages 659–674,

- Twyford, A.D. The road to **10,000 plant** genomes. Nature Plants 4, 312–313 (2018).

- Guo-Dong Wang, Greger Larson, Jeffrey M Kidd, Bridgett M vonHoldt, Elaine A Ostrander, Ya-Ping Zhang, **Dog10K**: the International Consortium of Canine Genome Sequencing, National Science Review, Volume 6, Issue 4, July 2019, Pages 611–613

Third-Generation Sequencing

- Examples: Pacific Biosciences (PacBio), Oxford Nanopore Technologies.
- Features: Real-time sequencing, can sequence long fragments.
- Pros: Long reads, real-time data generation, fewer assembly challenges.
- Cons: Higher error rates (although improving), expensive hardware.
- Applications: Structural variation detection

Made by OpenAI ChatGPT

더 많은 사람들의 유전체 정보 수집 프로젝트

UK10K Project

The UK10K project enabled researchers in the UK and beyond to better understand the link between low-frequency and rare genetic changes, and human disease caused by harmful changes to the proteins the body makes.

2010년 영국

https://www.uk10k.org/

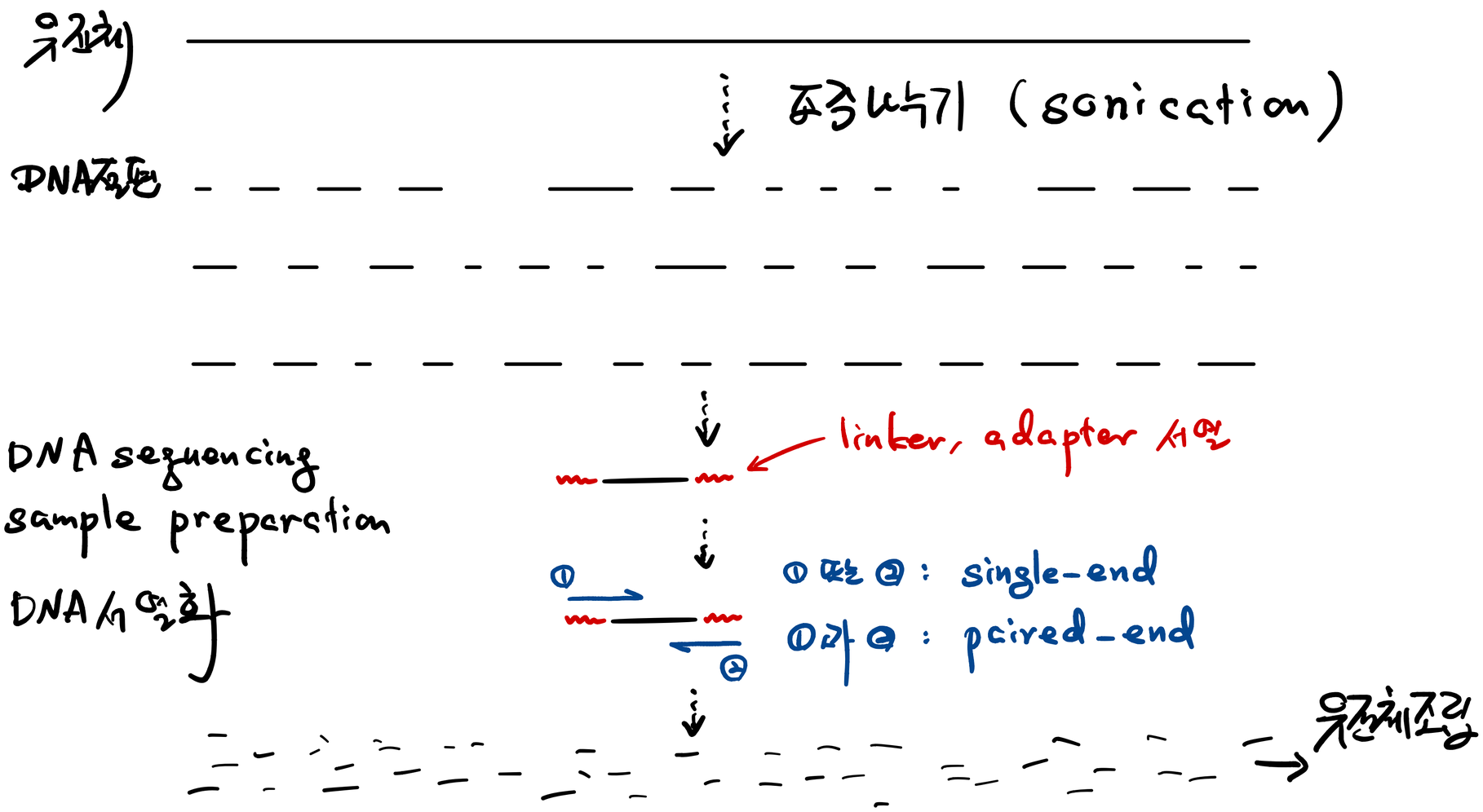
2015년 미국

https://www.researchallofus.org/

서열의 길이

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

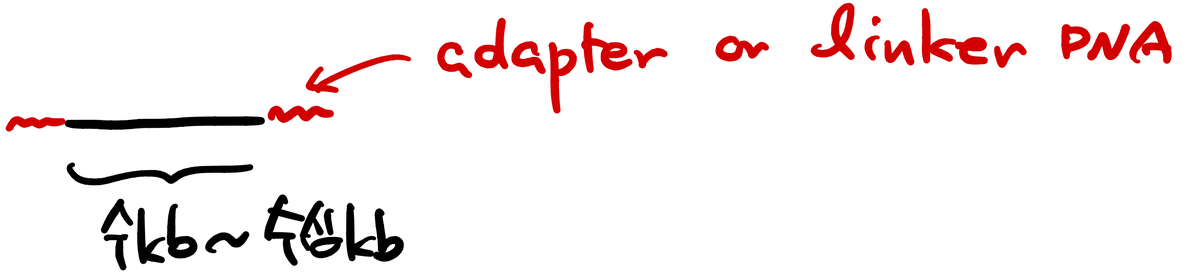
샷건 방법 (shotgun)

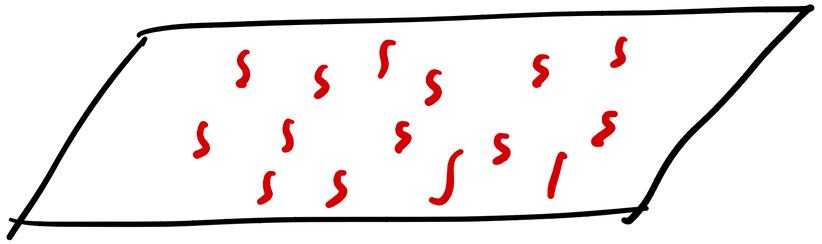


염기서열결정 라이브러리 제작

① DNA 절편 제작 : 초음파 분쇄 sonication
무작위로 DNA를 짧은 조각을 만든다.

② DNA 어댑터 연결.





Short-Read and Long-Read Sequencing

- Short-read sequencing and long-read sequencing are two widely used approaches in genomics, each with distinct advantages and challenges.
- Both methods are essential for various genomic studies, but they differ in terms of read length, accuracy, cost, and applications.
- Applications:
  - Short-Read Sequencing: Used for large-scale projects like whole-genome resequencing, RNA-seq, and SNP detection.
  - Long-Read Sequencing: Ideal for assembling de novo genomes, structural variant analysis, and resolving complex regions in genomes.

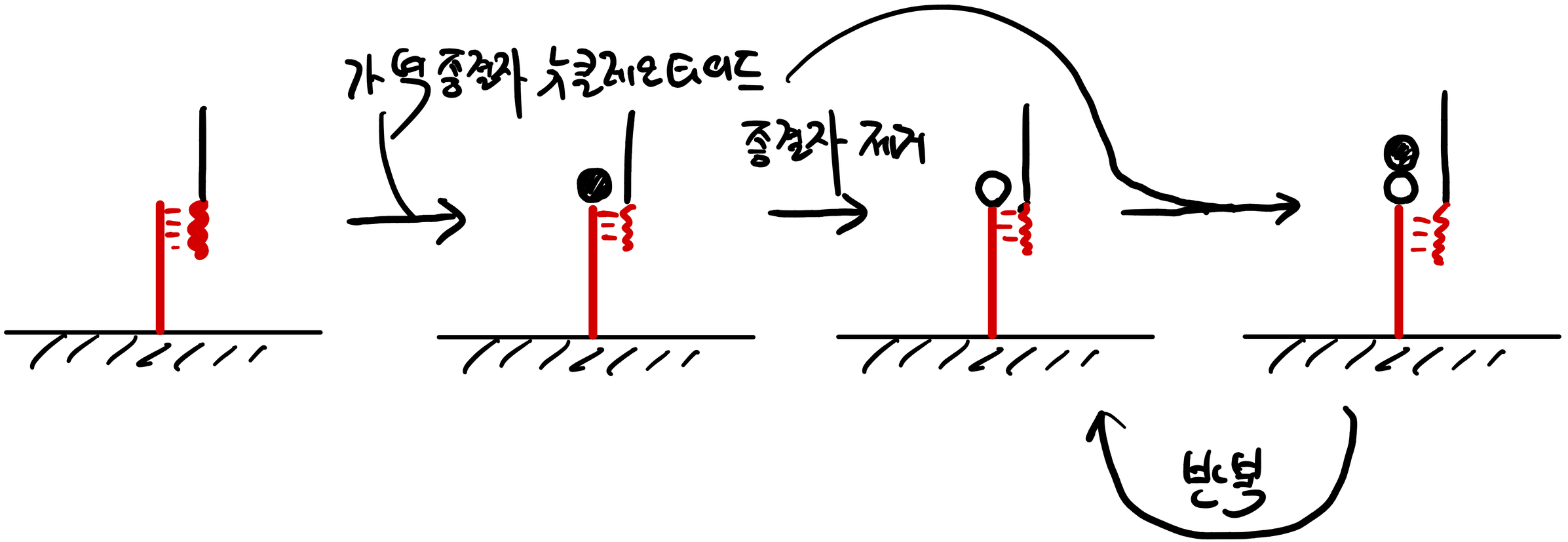
Made by OpenAI ChatGPT

Short-Read Sequencing

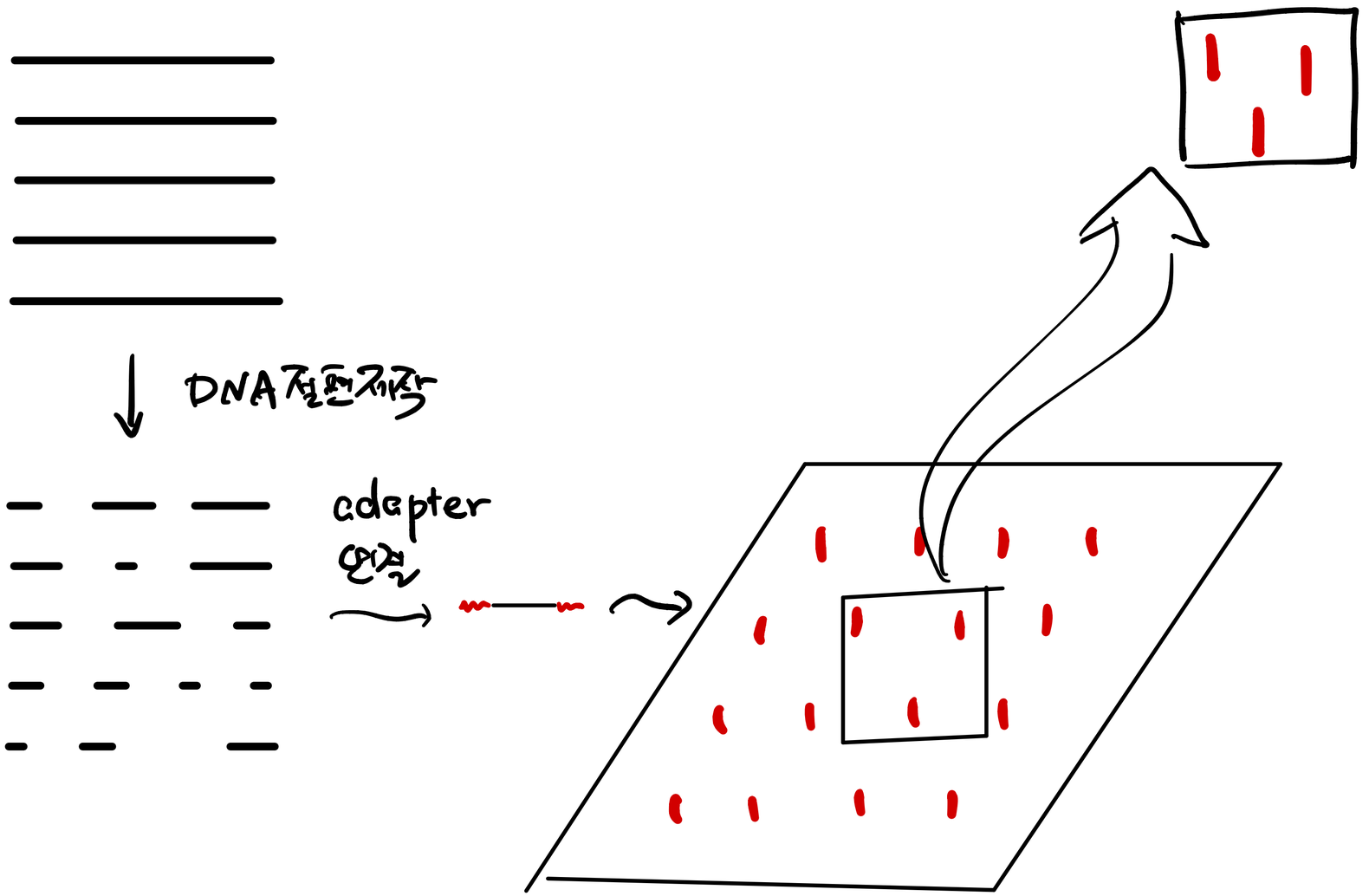
- Reads typically range from 50 to 300 base pairs (bp) in length.
  - Examples: Illumina platforms (most common).
- Advantages:
  - High throughput: Generates millions to billions of reads in a single run.
  - Lower cost: Economical for large-scale sequencing projects.
  - High accuracy: Reliable for detecting small variants like single nucleotide polymorphisms (SNPs).
- Challenges:
  - Assembly difficulty: Difficult to assemble repetitive regions or highly complex genomes due to short read lengths.
  - Limited to small variants: May struggle with detecting structural variants like large insertions, deletions, or rearrangements.

Made by OpenAI ChatGPT

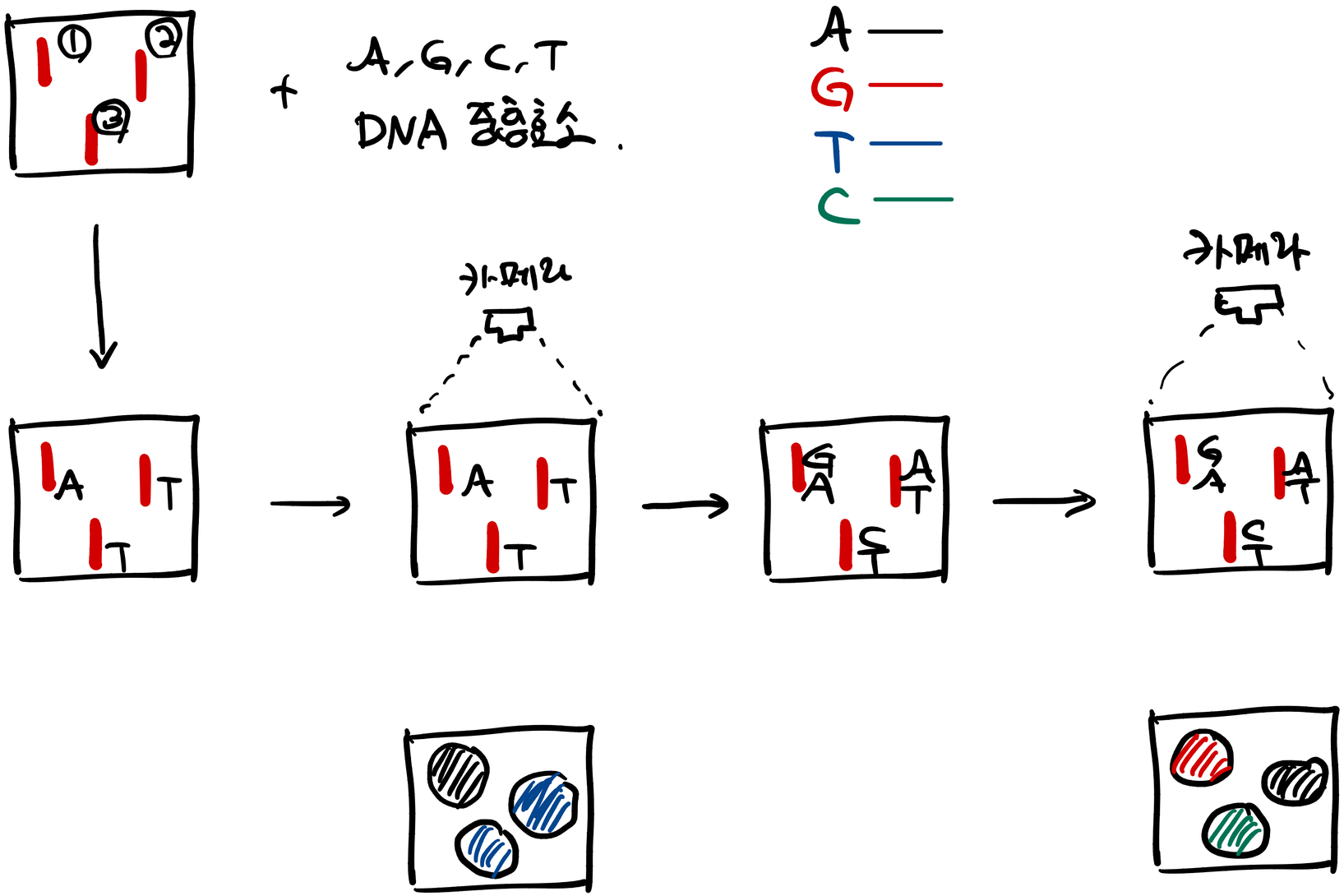
일루미나 시퀀싱 (1)



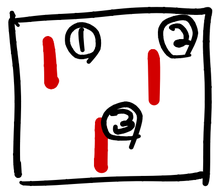
일루미나 시퀀싱 (2)

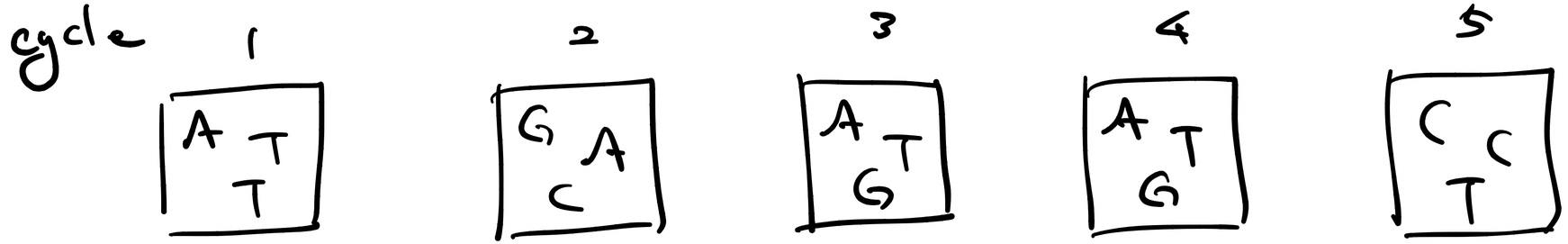


일루미나 시퀀싱 (3)



일루미나 시퀀싱 (4)





① A G A A C

② T A T T C

③ T C G G T

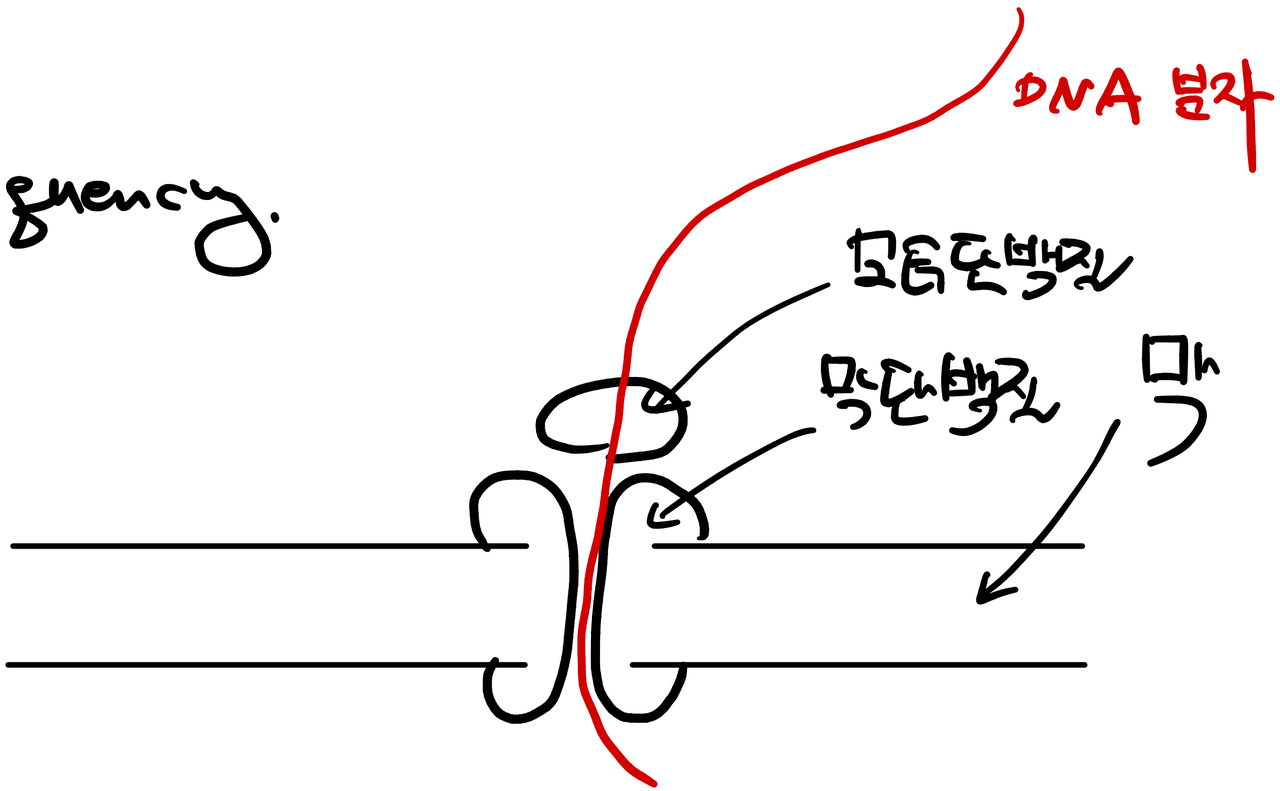
Long-Read Sequencing

- Reads typically range from thousands to millions of base pairs in length.
  - Examples: Oxford Nanopore and PacBio platforms.
- Advantages:
  - Better assembly: Longer reads simplify genome assembly, especially in repetitive regions.
  - Structural variant detection: Capable of detecting larger structural variations (insertions, deletions, rearrangements).
  - Single-molecule sequencing: Requires less DNA amplification, reducing bias.
- Challenges:
  - Higher cost: More expensive compared to short-read sequencing.
  - Lower accuracy: Typically less accurate on a per-base level, though error rates are improving with new technologies.
  - Lower throughput: Fewer reads generated per run compared to short-read technologies.

Made by OpenAI ChatGPT

3세대 염기서열 결정법

- SMART (single-molecule real-time sequencing)
  by PacBio
- nanopore sequencing



Summary of the short- and long-read sequencing

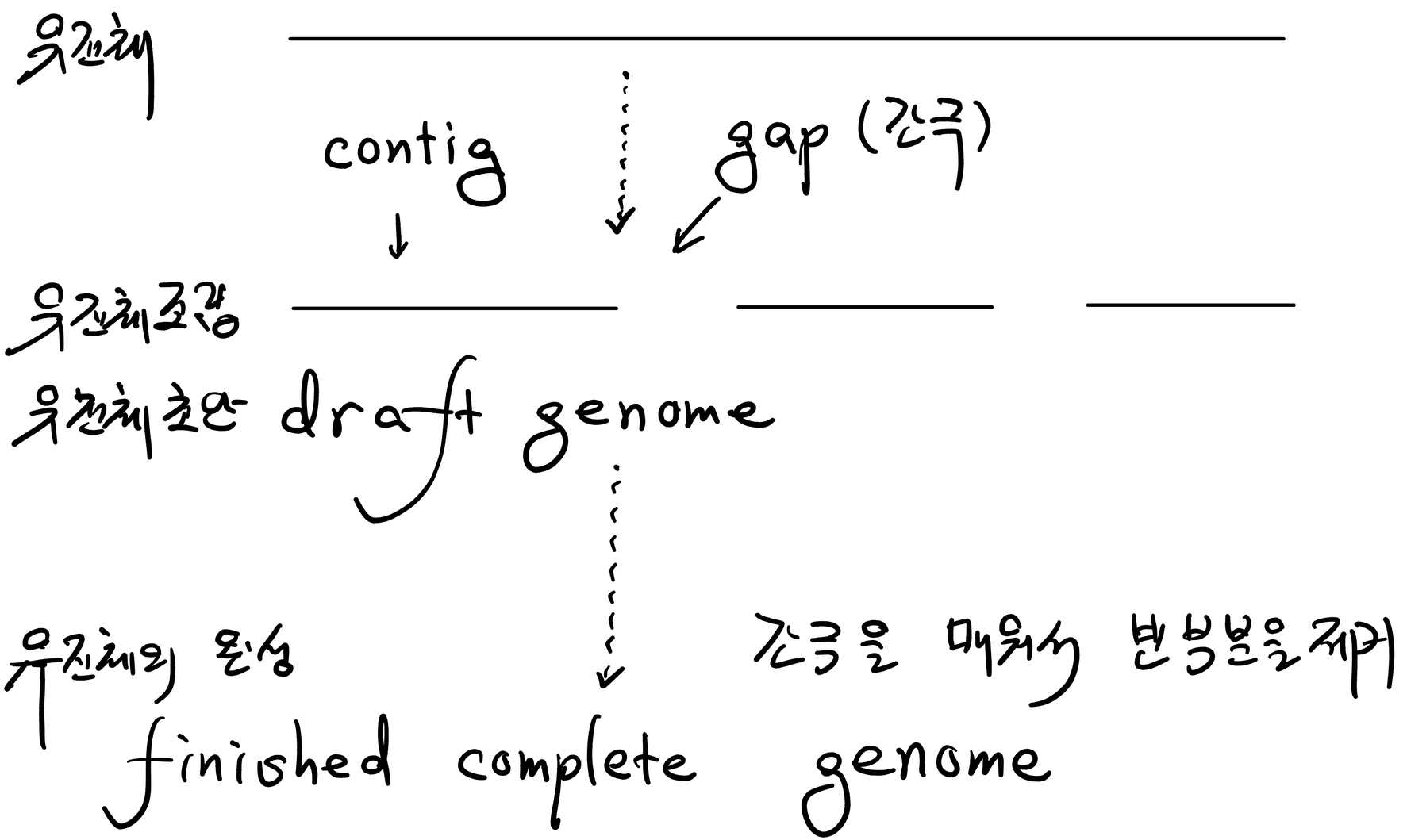
Feature	Short-Read Sequencing	Long-Read Sequencing
Read Length	50-300 bp	Thousands to millions of bp
Throughput	High	Lower
Cost	Lower	Higher
Accuracy	Higher accuracy per base	Lower accuracy per base
Assembly	Difficult with repetitive regions	Easier, especially for complex genomes
Applications	Variant calling, small genomes	Structural variant detection, complex genomes
Feature	Short-Read Sequencing	Long-Read Sequencing

Made by ChatGPT

유전체 완성의 어려움

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

유전체 완성의 어려움

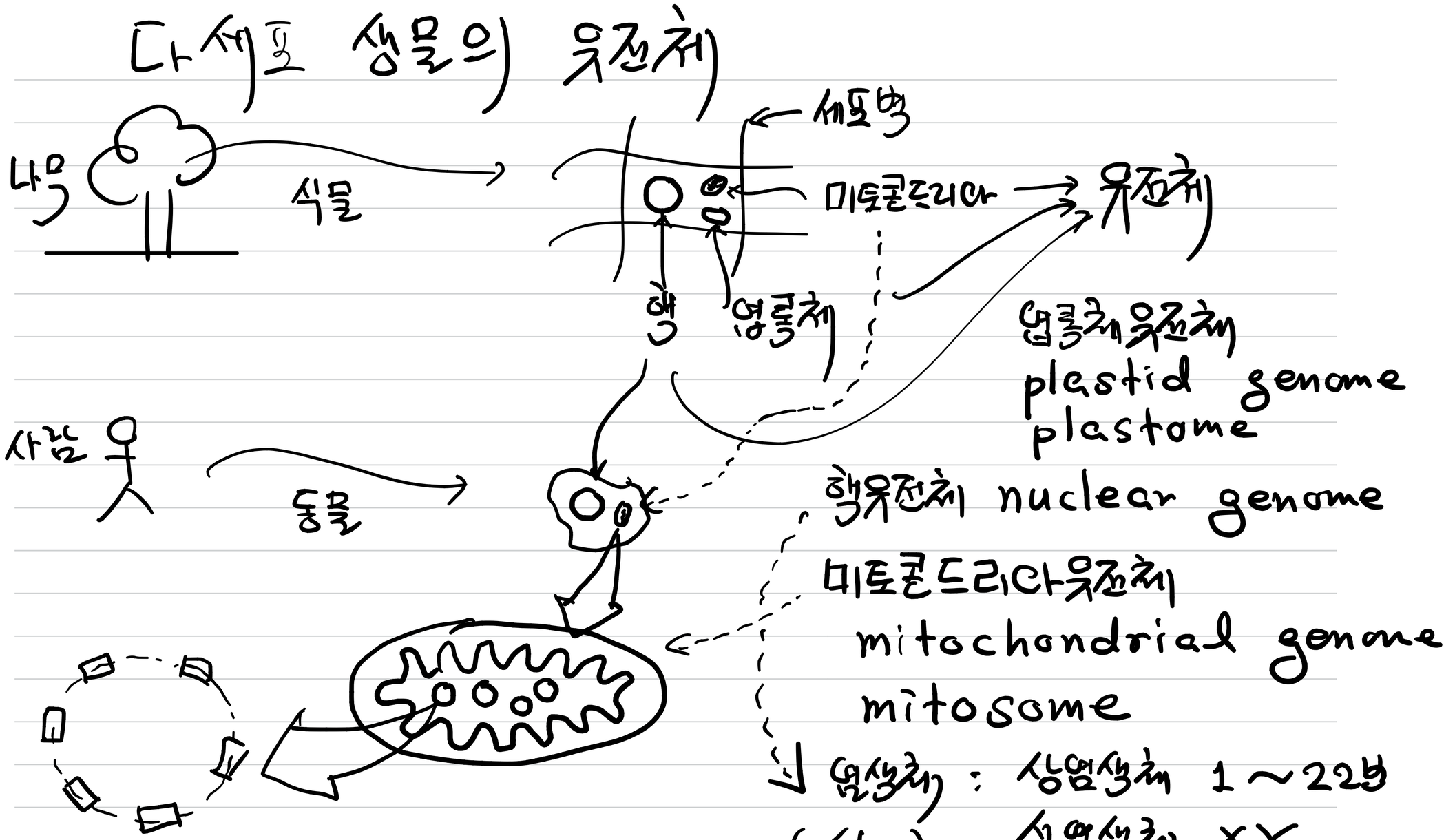

유전체
contig
gap (간극)
유전체조립
유전체초안 draft genome
유전체의 완성
finished complete genome
간극을 메워서 분포보을제거

Contigs and Gaps

- Contigs (continuous sequences) are formed by aligning and merging overlapping reads.
- Gaps are filled by further sequencing or computational methods.

Made by OpenAI ChatGPT

다세포 생물의 유전체


나무
식물
세포벽
미토콘드리아
유전체
핵
엽록체
엽록체유전체
plastid genome
plastome
사람
동물
핵유전체 nuclear genome
미토콘드리아유전체
mitochondrial genome
mitosome
염색체 : 상염색체 1~22번
(46개) 성염색체 X,Y

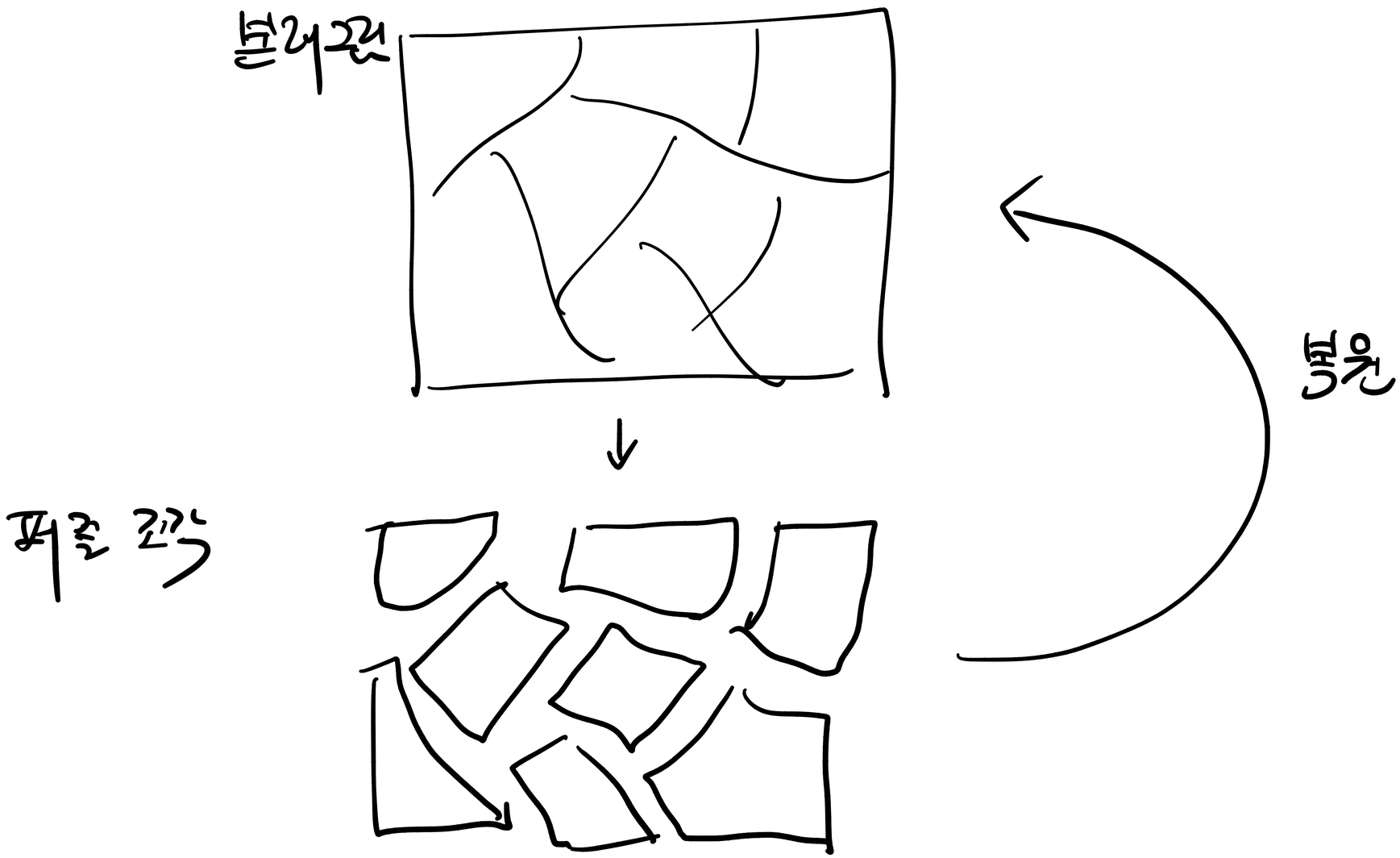
유전체 조립을 완성하는데 어려움

- 기술적 어려움: DNA 서열화 기계의 한계점
- 생물학적 어려움: 유전체의 반복 서열
- 진핵생물이 원핵생물의 유전체 조립보다 어려운 이유
  - 유전체가 더 크다.
  - 반복 서열이 더 많다.
  - DNA 클로닝이 어려운 염색체 부분이 있다.

유전체 조립 알고리즘

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

유전체 조립



TATGGGGTGC

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

k-mer와 그래프

TATGGGGTGC

TAT

ATG

TGG

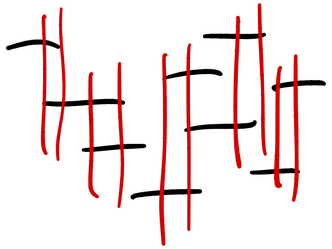
GGG

GGG

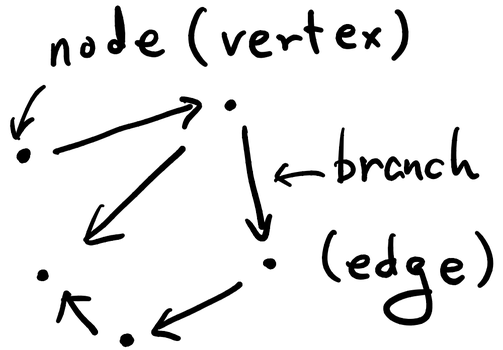
GGT

GTG

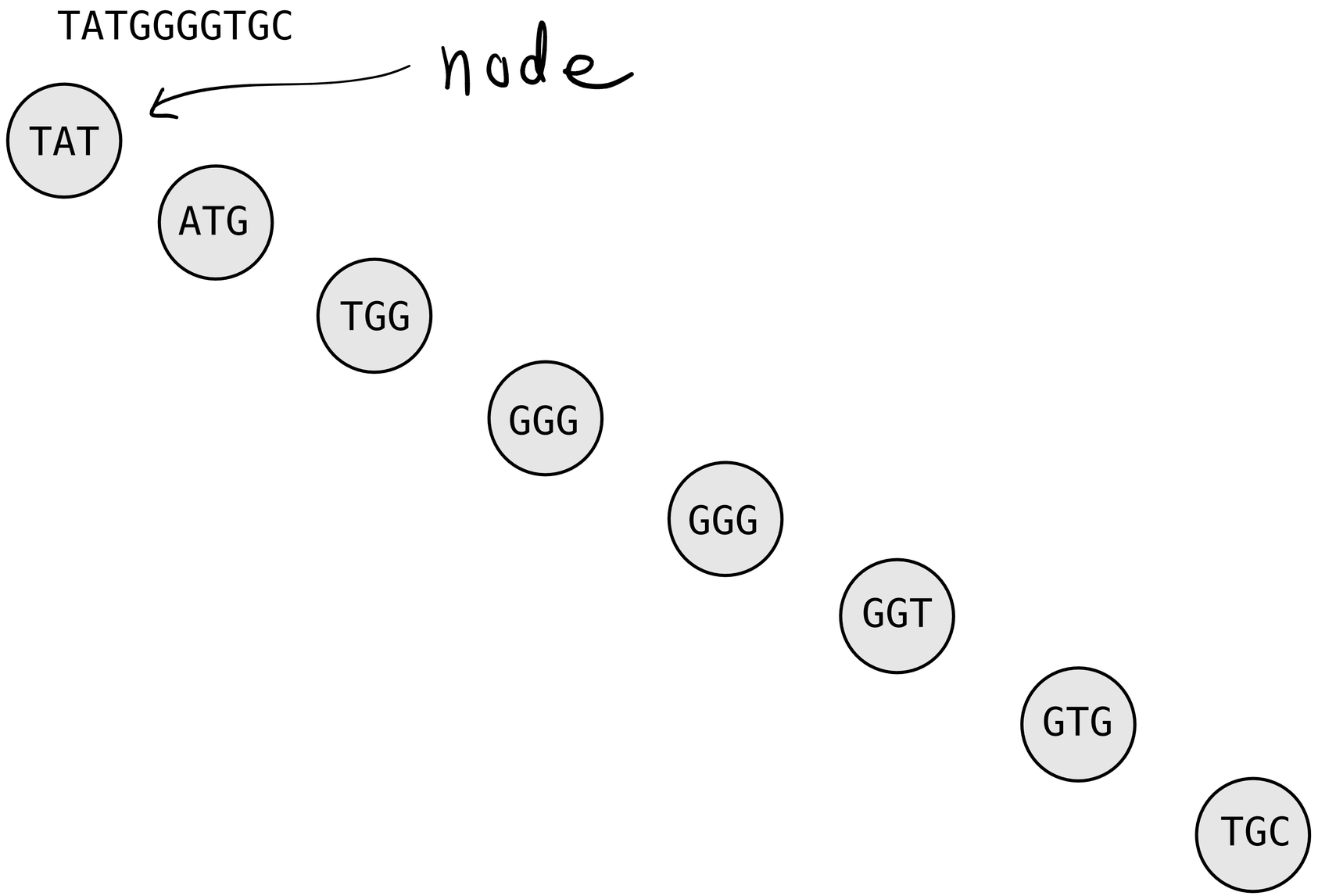
TGC



그래프에 대하여

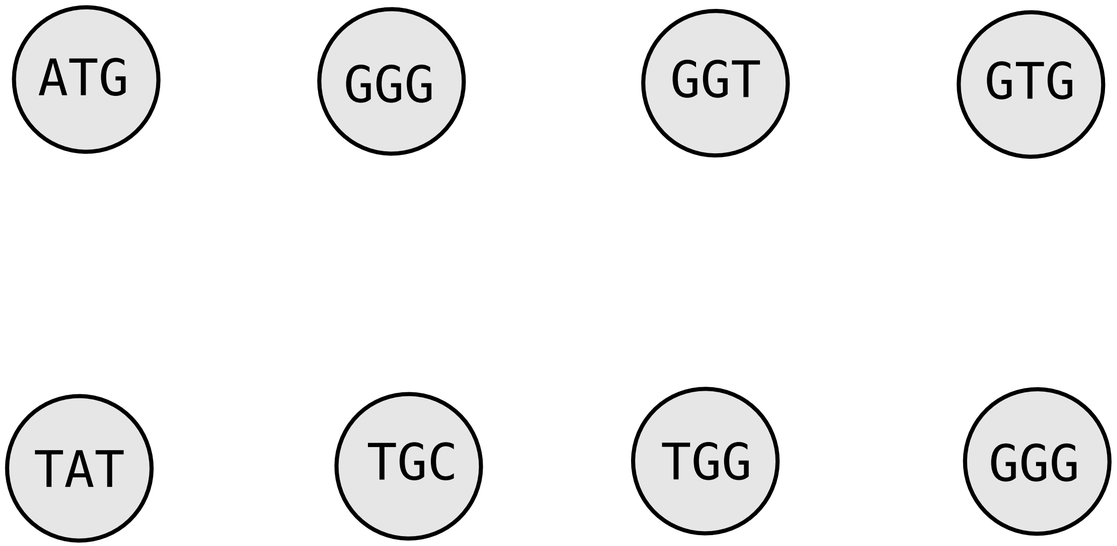


서열과 노드



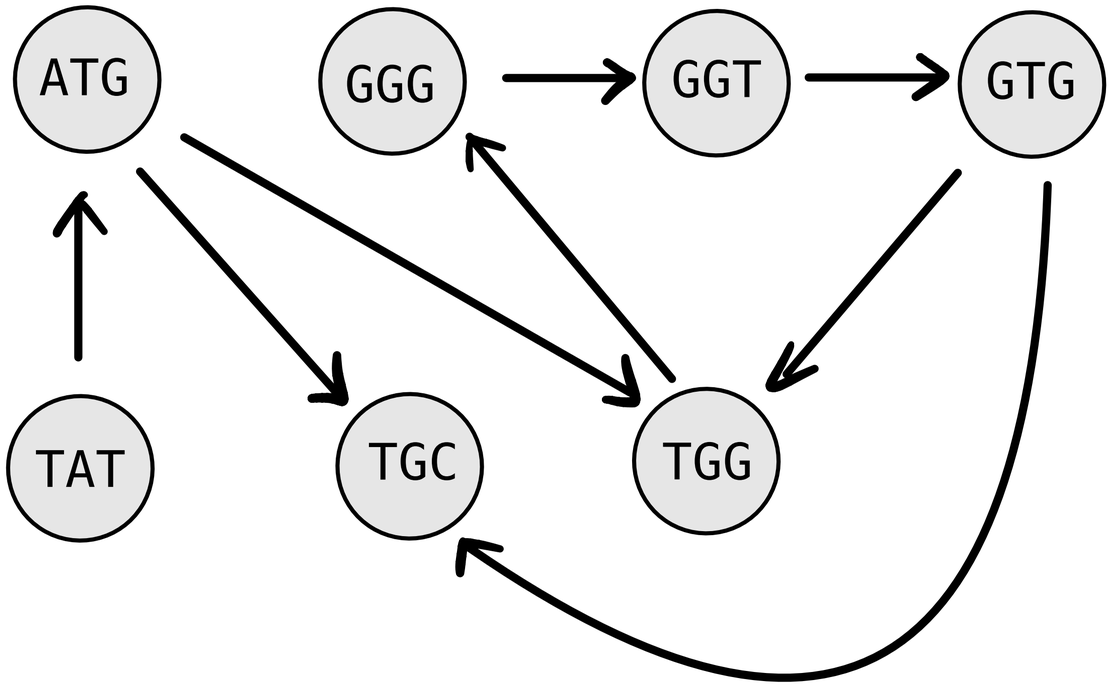
서열의 k-mer 사이의 관계

TATGGGGTGC

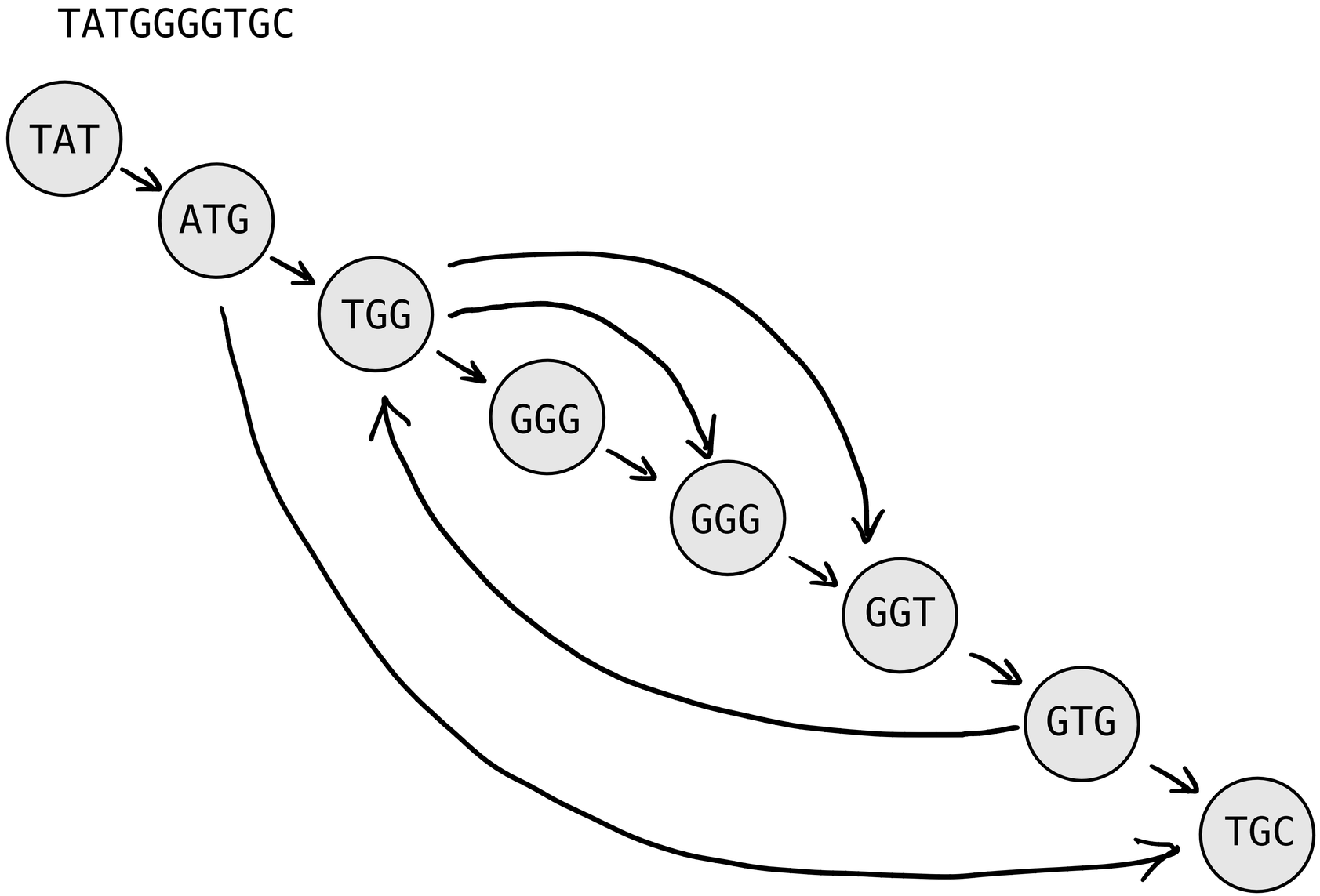


k-mer 사이의 연결선 (edge)

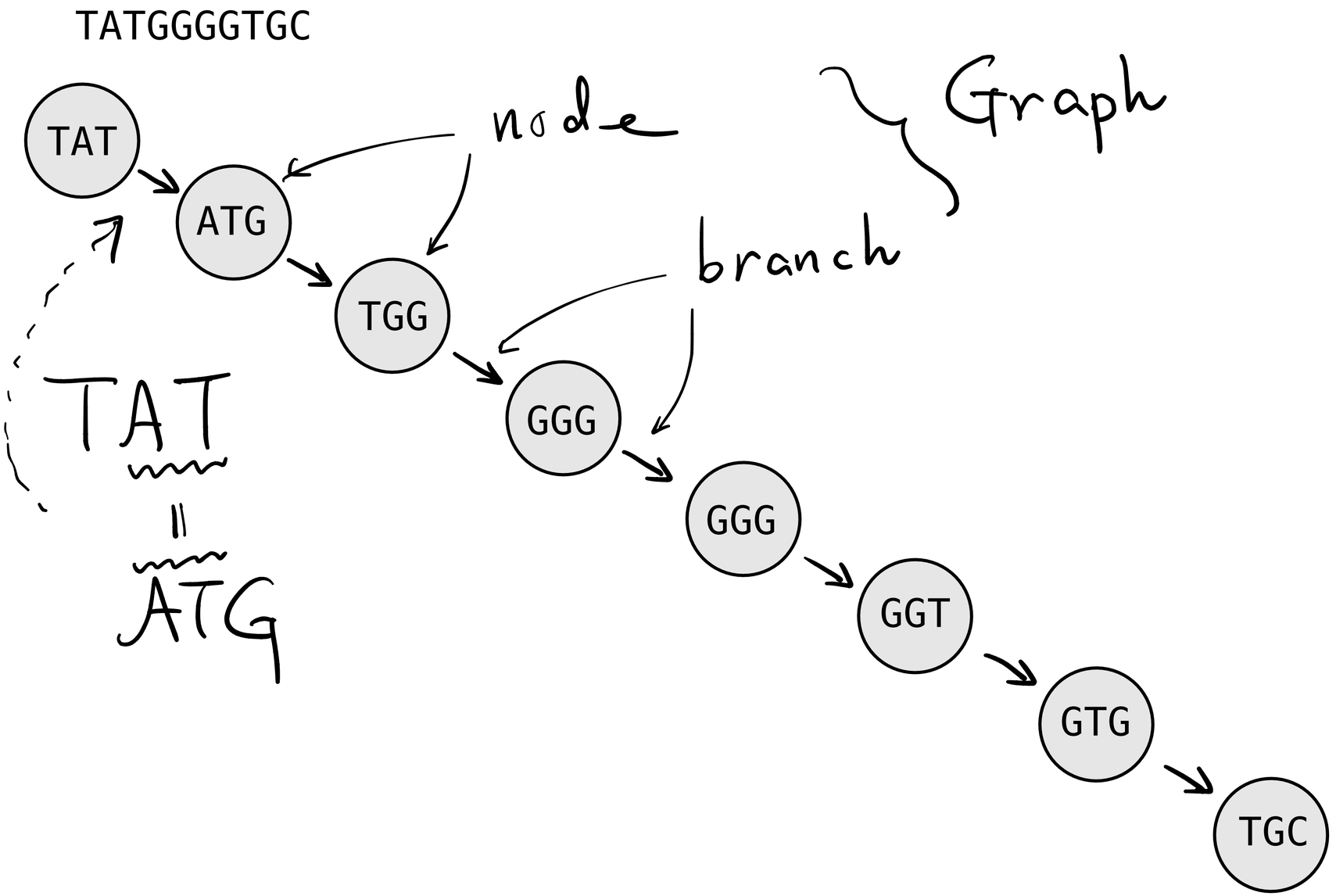
TATGGGGTGC



그래프 화살표 (edge)

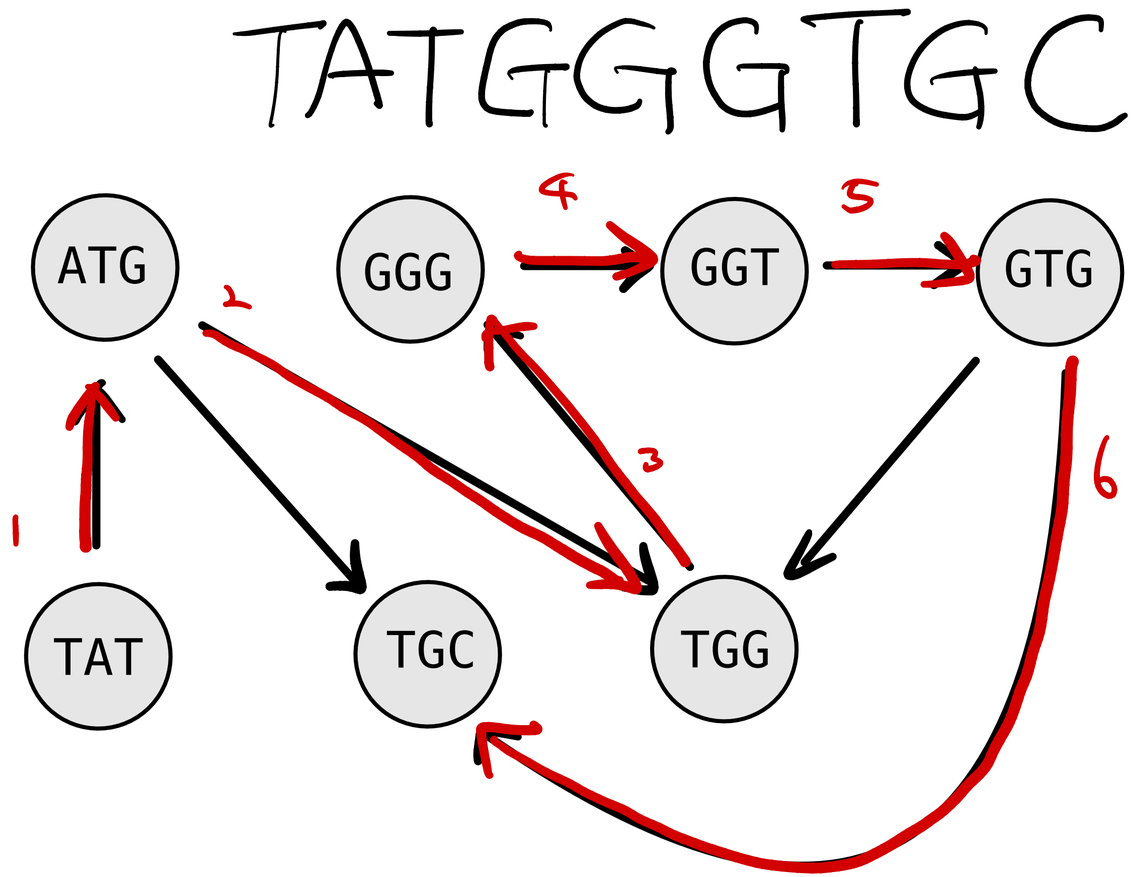


서열과 그래프


TATGGGGTGC
TAT
ATG
TGG
GGG
GGG
GGT
GTG
TGC
node
branch
Graph
TAT
=
ATG

유전체 염기 서열 조립은 미로에서 길 찾기와 유사

TATGGGGTGC

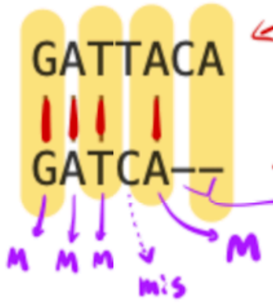


서열 정렬

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

서열 정렬 (1)

For example alignment of GATTACA and GATCA could look like:



gap { M: 4개, mis: 1개, gap: 2개 } → 정렬에 점수를 부여한다.

옆단위 3가지 종류 - match, mismatch, gap

For example, we could have arranged our example words like this

```
GATTACA
|||  ||
GAT--CA
```

{ M: 5개, mis: 0개, gap: 2개

or we could have shown them like this:

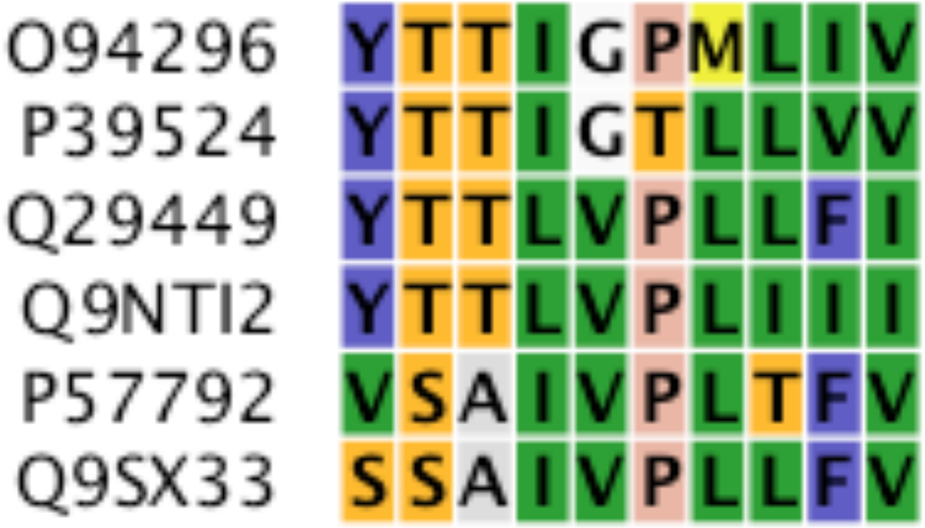
```
GATTACA
||  |  ||
GA-T-CA
```

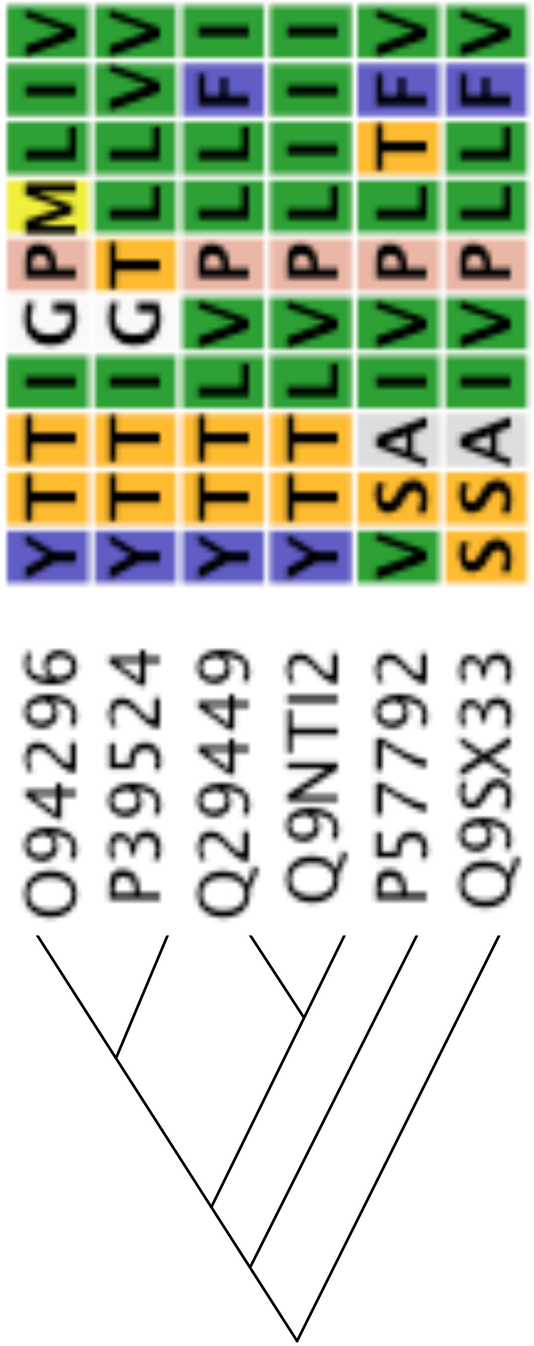
{ M: 5, mis: 0, gap: 2

서열정렬: 무수히 많은 정렬들 중에 가장 적당한 것을 선택하는 행위.

서열 정렬 (2)

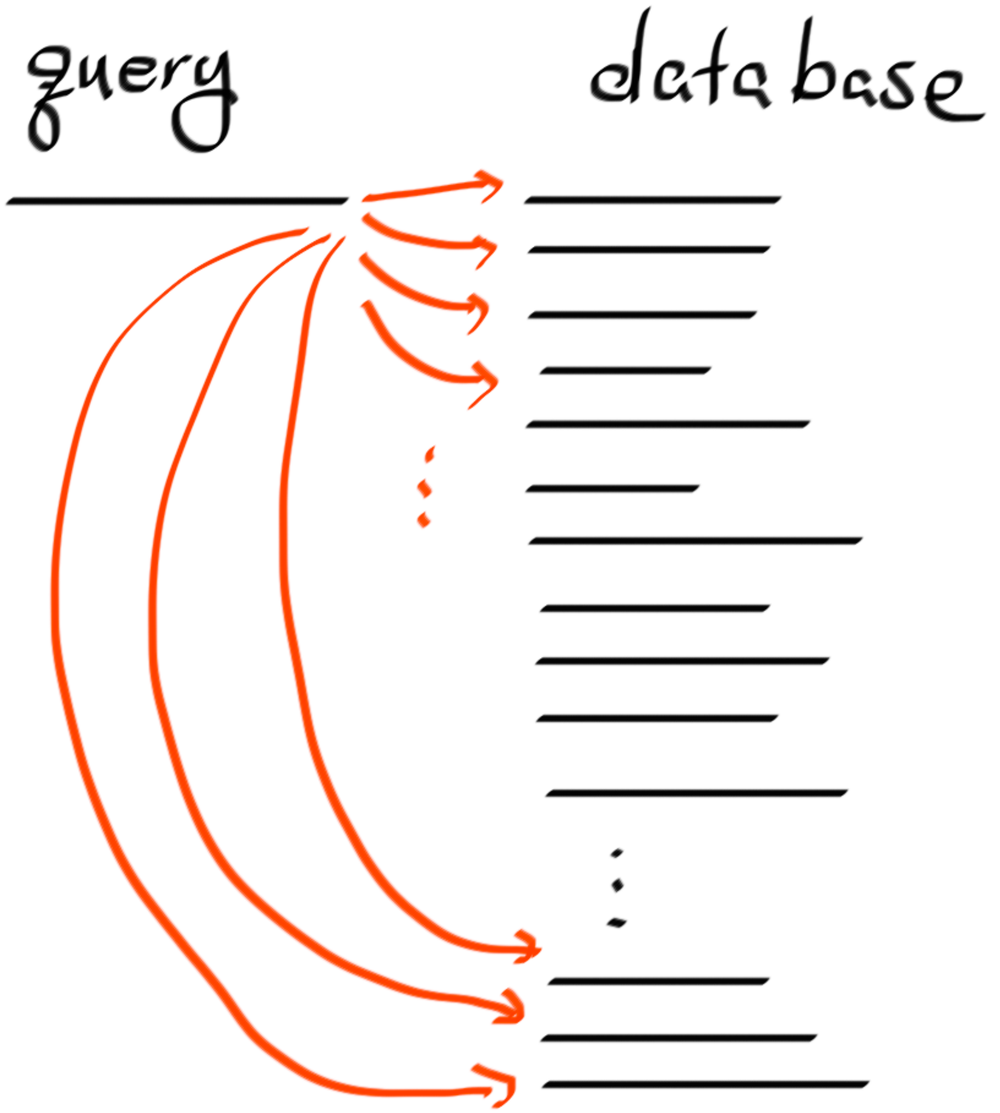
비슷한 부분의 염기 또는 아미노산 문자를 정렬한다.
서열 정렬은 진화적인 관계를 내포한다.
각 열을 사이트라고 하며 각 열은 진화적으로 상동 관계이다.





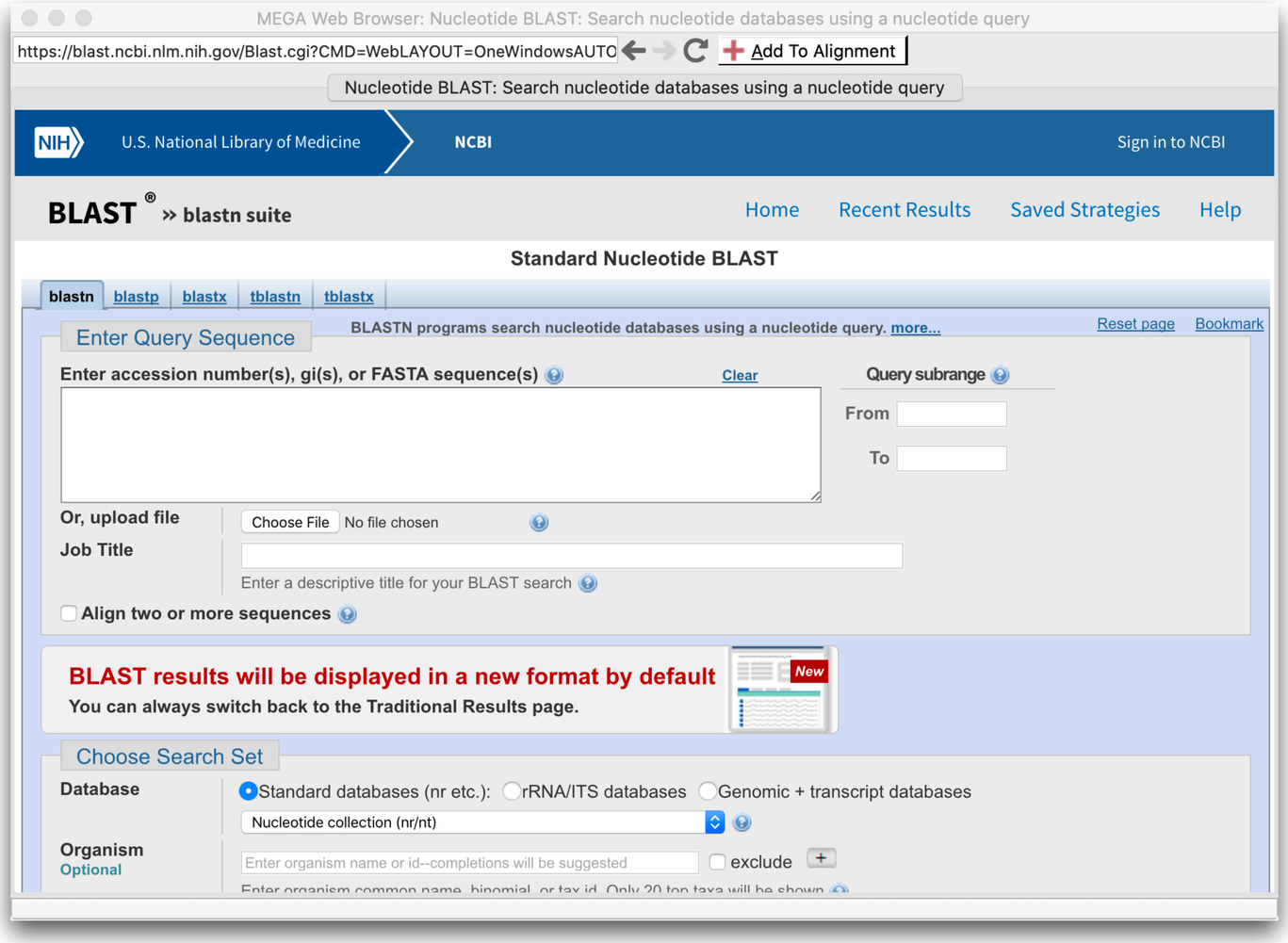
서열 검색 (1)

- Query (검색어): We have a biological sequence of DNA or amino acid sequences.
- Database (데이터베이스): We want to know the degrees of relatedness of such a sequence to a collection of other sequences, often called a sequence database.



서열 검색 (2)

- BLAST - Basic Local Alignment Search Tool



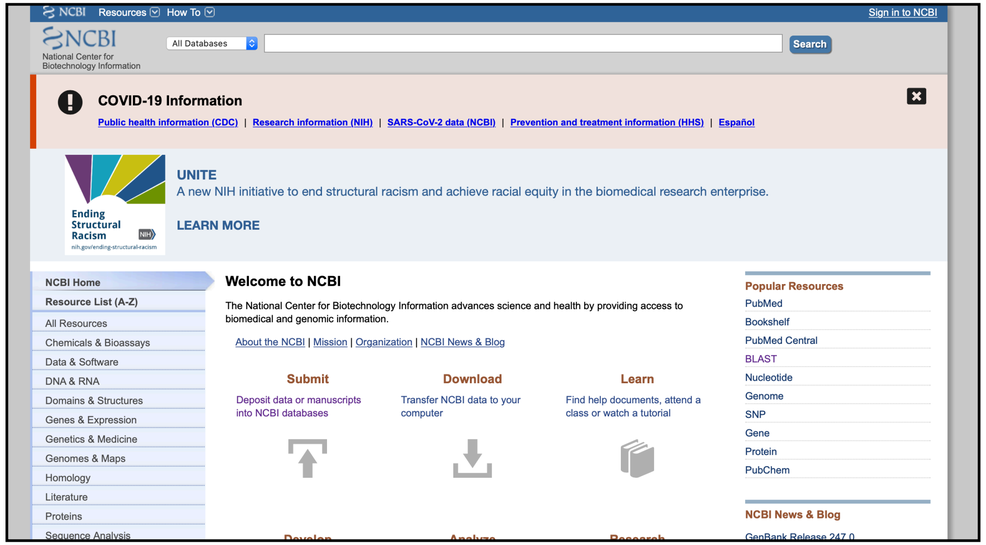
생물학 정보 데이터베이스

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

생물정보학 데이터

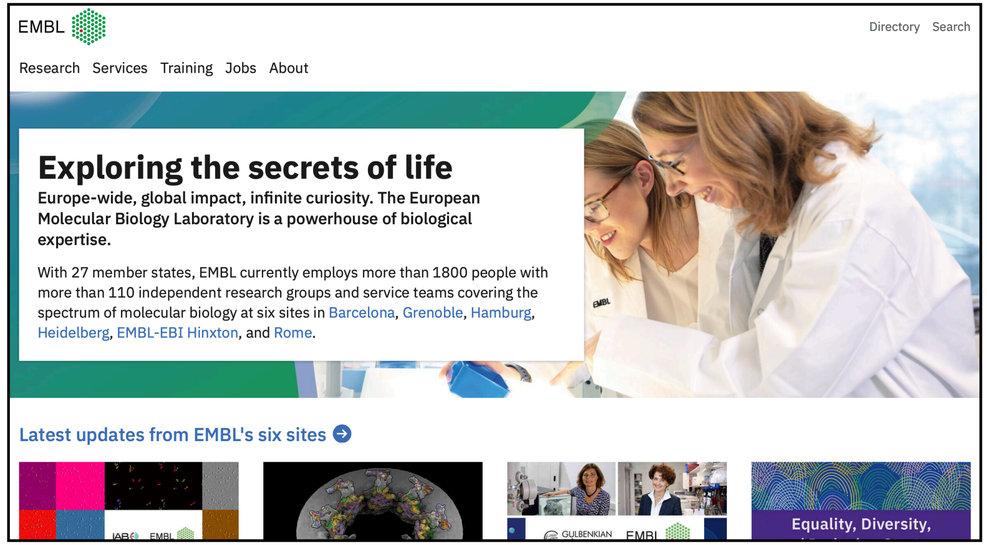
- 생물정보학 주요 웹사이트
- 미국 NCBI의 주요 데이터베이스
- 단백질 데이터베이스
- 대표적 모델 생물 데이터베이스
- 생물정보학 컴퓨터 데이터 파일

생물정보학 주요 웹사이트



NCBI: National Center for Biotechnology Information

미국 국립보건원 국립도서관 산하 기관

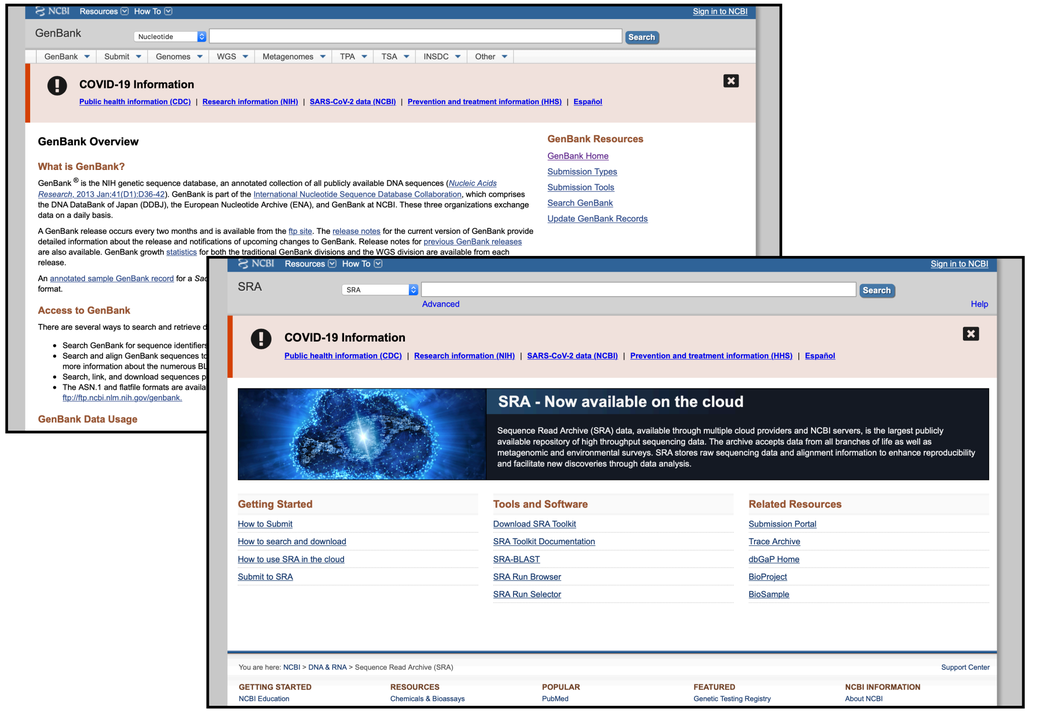


EMBL: European Molecular Biology Laboratory

유럽 분자생물학 연구소

미국 NCBI의 주요 데이터베이스

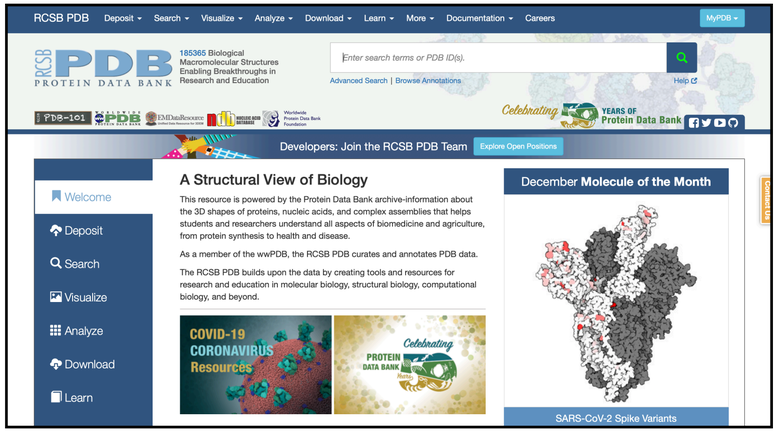
- 유전체: 유전체 서열 및 부가 정보를 저장하고 배포한다.
- 유전자: 유전자 정보를 저장하고 배포한다.
- 염기서열: 염기서열 정보를 저장하고 배포한다.
- 단백질서열: 단백질 서열 정보를 저장하고 배포한다.
- SNP: 다형성 정보를 저장하고 배포한다.
- 분류: 생물의 분류 정보를 저장하고 배포한다.



- Sequence Read Archive (SRA): SRA 데이터베이스는 대용량 서열 데이터를 저장하고 배포한다.
- PubMed: 생물학 및 의학 관련 도서와 논문 정보를 저장하고 배포한다.

단백질 데이터베이스

- PDB (Protein Data Bank): 단백질 3차 구조
- UniProt (Universal Protein Resource): 단백질 서열과 기능



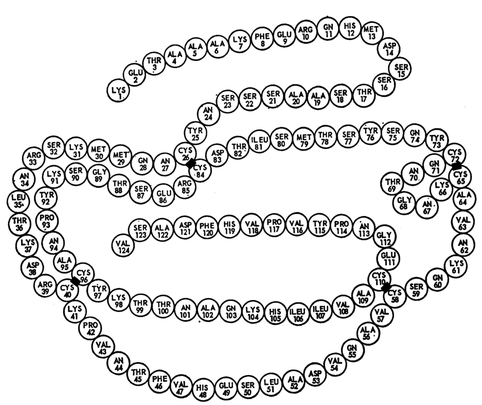


Fig. 1. The amino acid sequence of bovine pancreatic ribonuclease (*50*).

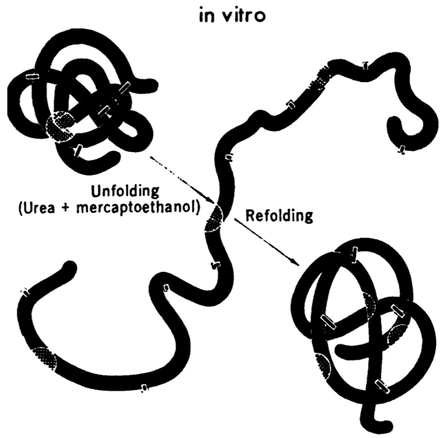
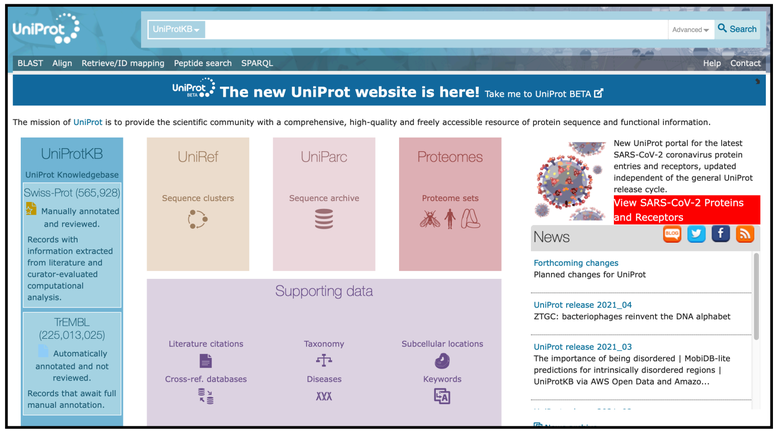


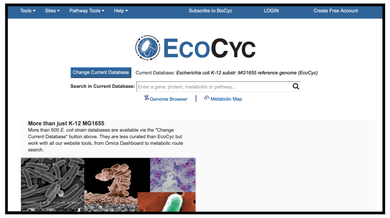
Fig. 2. Schematic representation of the reductive denaturation, in 8*M* urea solution containing 2-mercaptoethanol, of a disulfide–cross-linked protein. The conversion of the extended, denatured form to a randomly cross-linked, "scrambled" set of isomers is depicted at the lower right.

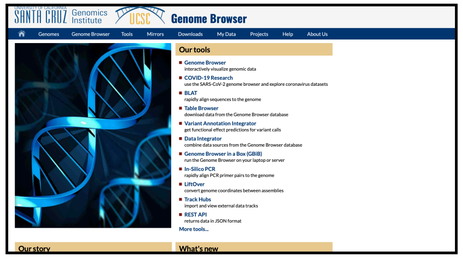


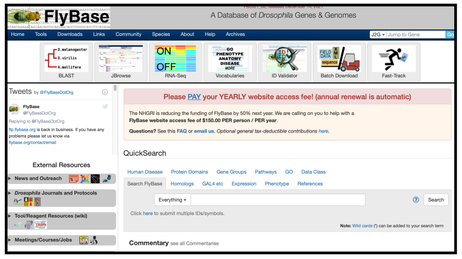
Anfinsen, C. B. Principles that govern the folding of protein chains. Science 161, 223–230 (1973).

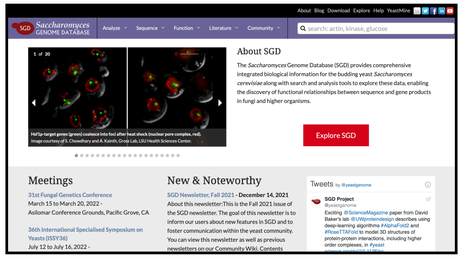
대표적 모델 생물 데이터베이스

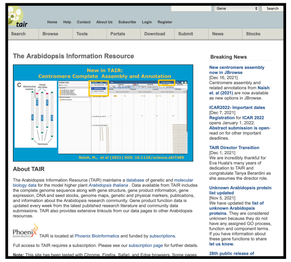
1. UCSC Genome Browser: 인간유전체를 비롯한 모델 생물의 유전체
2. FlyBase: 초파리
3. WormBase: 꼬마선충
4. Saccharomyces Genome Database(SGD): 효모
5. The Arabidopsis Information Resource(TAIR): 애기장대
6. EcoCyc (Encyclopedia of E. coli Genes and Metabolic Pathways): 대장균











서열 데이터 파일

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

Sequence data file formats (서열 데이터 파일 형식)

- 생물학 데이터는 대체로 텍스트 형태
- 데이터를 작성하는 약속을 데이터 형식
- Genbank file format
- FASTA file format
- FASTQ file format

Staphylococcus phage MSA6, complete genome

GenBank: JX080304.2

FASTA Graphics

```
LOCUS       JX080304              148243 bp    DNA     linear   PHG 31-MAR-2014
DEFINITION  Staphylococcus phage MSA6, complete genome.
ACCESSION   JX080304
VERSION     JX080304.2
KEYWORDS    .
SOURCE      Staphylococcus phage MSA6
  ORGANISM  Staphylococcus phage MSA6
            Viruses; Duplodnaviria; Heunggongvirae; Uroviricota;
            Caudoviricetes; Herelleviridae; Twortvirinae; Kayvirus.
REFERENCE   1  (bases 1 to 148243)
  AUTHORS   Lobocka,M., Hejnowicz,M.S., Dabrowski,K., Gozdek,A., Kosakowski,J.,
            Witkowska,M., Ulatowska,M.I., Weber-Dabrowska,B., Kwiatek,M.,
            Parasion,S., Gawor,J., Kosowska,H. and Glowacka,A.
  TITLE     Genomics of Staphylococcal Twort-like Phages - Potential
            Therapeutics of the Post-Antibiotic Era
  JOURNAL   Adv. Virus Res. 83, 143-216 (2012)
   PUBMED   22748811
```

```
>JX080304.2 Staphylococcus phage MSA6, complete genome
GGAATTCTTTTACCTCTCTCACTCAGCCTATTACTTATTACCGACTTCCCTAACTACTTATTCTATAGTT
ATAATATTCATTTATTATACAATACTTAAACTATAGTATTCTACTGTTAATCTATGCTGAAGCGGTCTTA
ATCTATGGTTATTATATAATAATCTTATATAATGGTACATTAATCTAGTATATTACATTAGAATCATTCT
AATCTAGGATTTTAATCTTTAGACCCTAGGAAAAGTGGTACTAAAATATAAAACCCTATAGGTATGGGAT
TCTTATTTTTAAAATTACTAAAAAGTATTAGGTTTTCCCTAGGGCAAAGT
```

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Sequence quality scores: FASTQ format

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI

|                          |    |        |

0.........................26...31.......40
```

$Q = -10 \log P$

Q	P
10	1/10
20	1/100
30	1/1000
40	1/10000

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

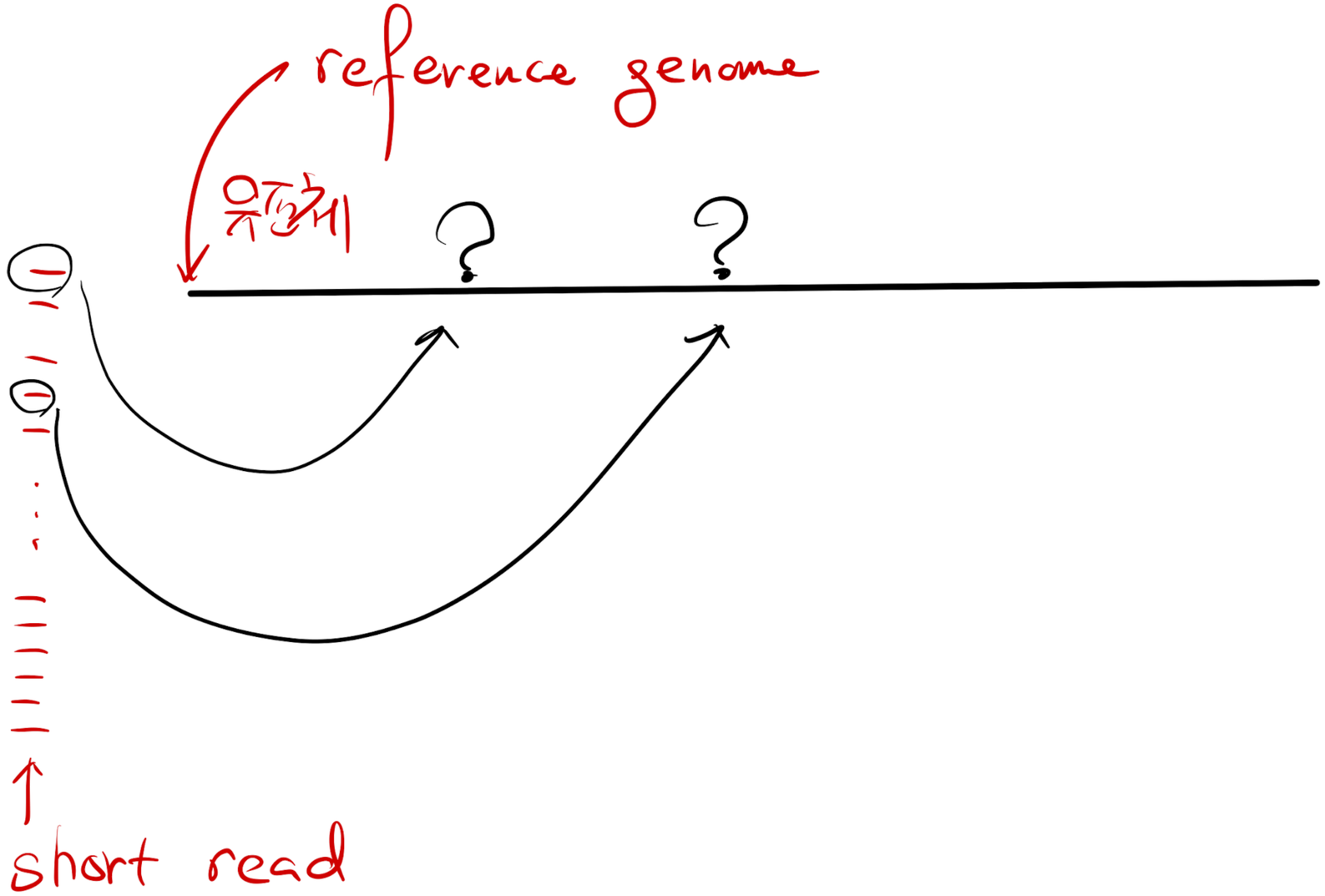
GFF file format: 유전체 주석 genome annotation

- 1열: 유전 정보가 포함되어 있는 염기 서열의 이름
- 2열: 유전 정보를 생성한 기관이나 생성시 사용한 방법
- 3열: 유전 정보의 종류
- 4열: 유전 정보의 1열의 염기 서열에서 시작 위치
- 5열: 유전 정보의 1열의 염기 서열에서 종료 위치
- 6열: 점수 또는 '.' 유전 정보와 관련된 수
- 7열: 유전 정보의 strand 정보 ('+' 또는 '-')
- 8일: CDS의 프레임 정보 (0, 1, 2)또는 '.'
- 9열: 나머지 정보들로 특별한 형식은 없다.

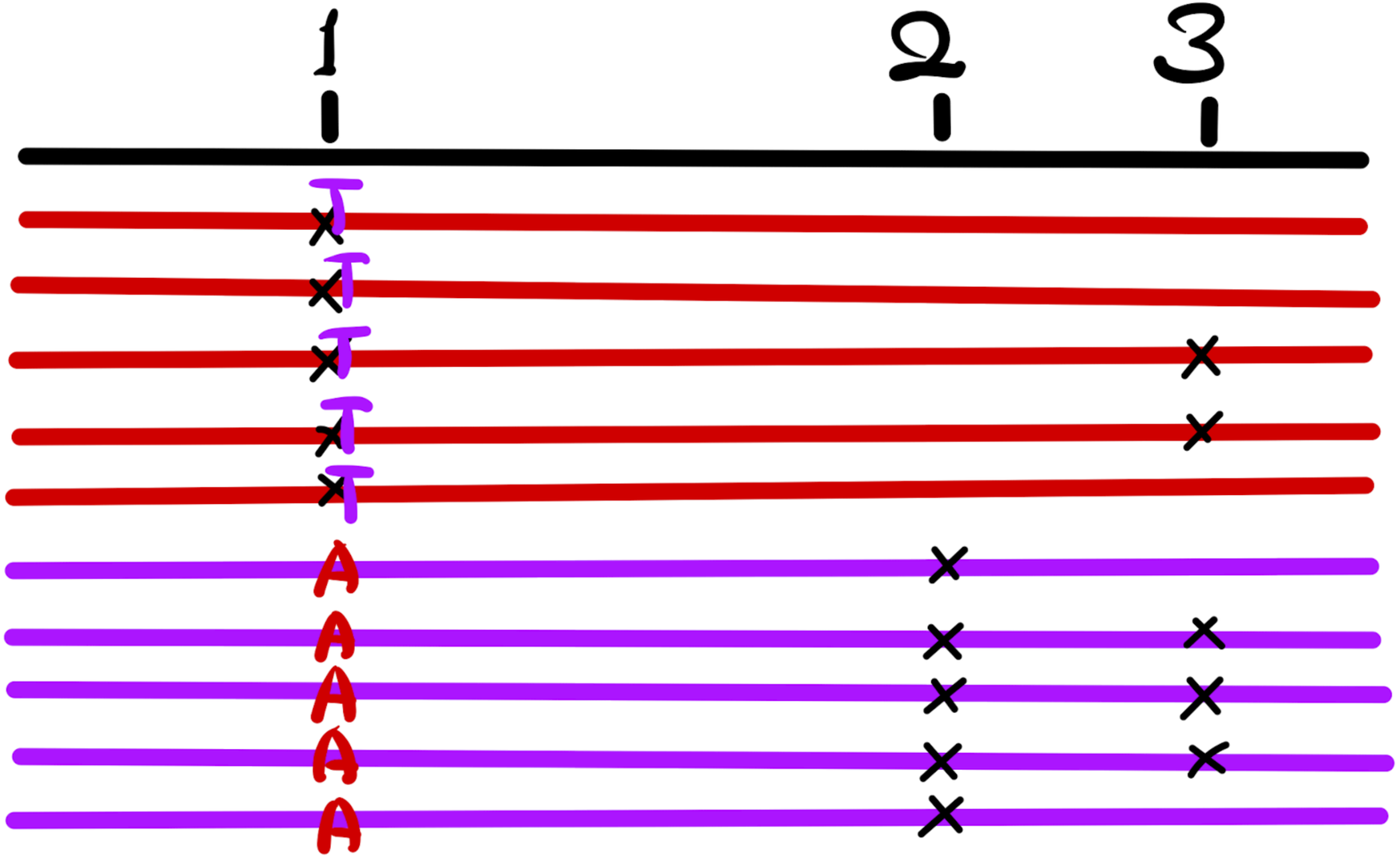
```
X       Ensembl   Repeat     2419108   2419128   42       .
X       Ensembl   Repeat     2419108   2419410   2502     -
X       Ensembl   Repeat     2419108   2419128   0        .
X       Ensembl   Pred.      2416676   2418760   450.19   -
X       Ensembl   Variation  2413425   2413425   .        +
X       Ensembl   Variation  2413805   2413805   .        +
```

http://m.ensembl.org/info/website/upload/gff.html

SAM/BAM file format - 유전체 맵핑



VCF format - 종의 여러 개체 유전체 변이



다양한 생물정보학 컴퓨터 파일 형식

- BED format
- bedGraph format
- CRAM format
- GTF format
- WIG format
- .2bit format

https://genome.ucsc.edu/FAQ/FAQformat.html

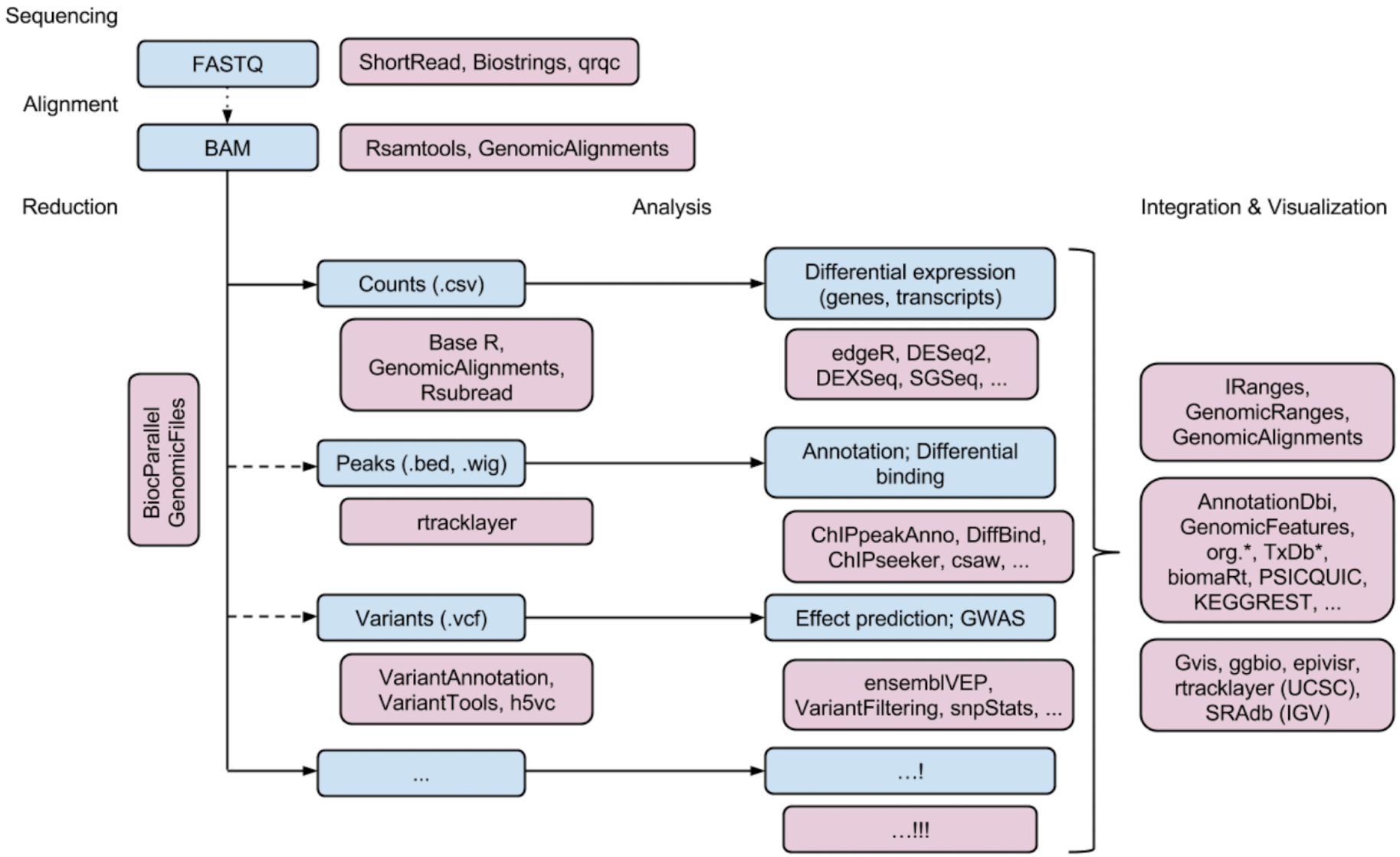
생물정보학 컴퓨터 도구

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

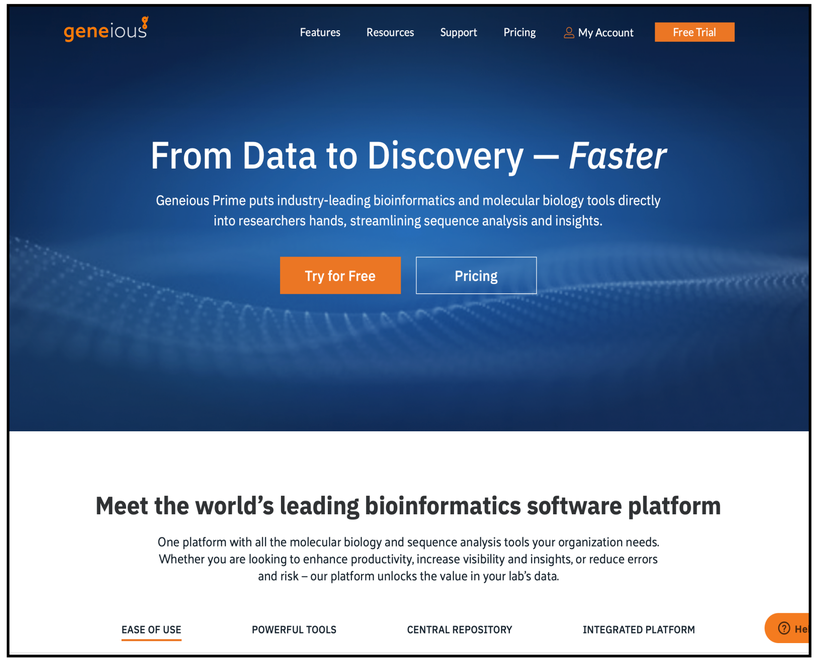
생물정보학 컴퓨터 도구

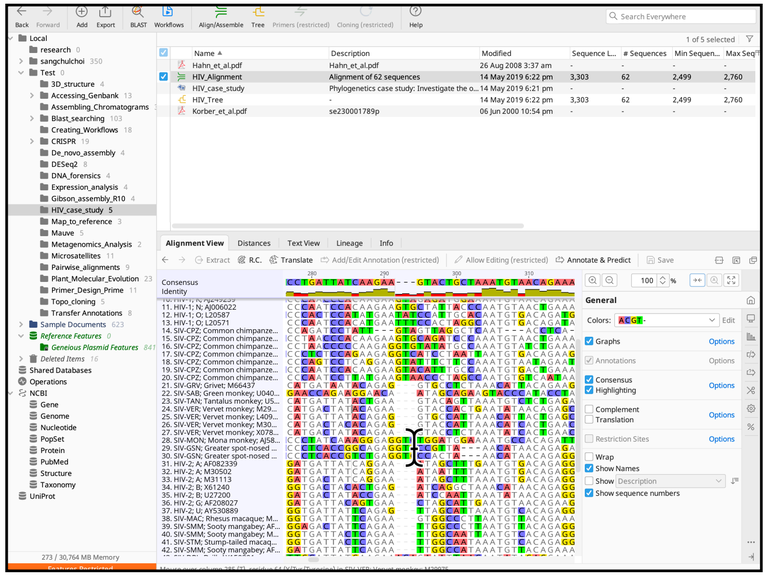
- R and RStudio
- Geneious
- Galaxy
- MegaX

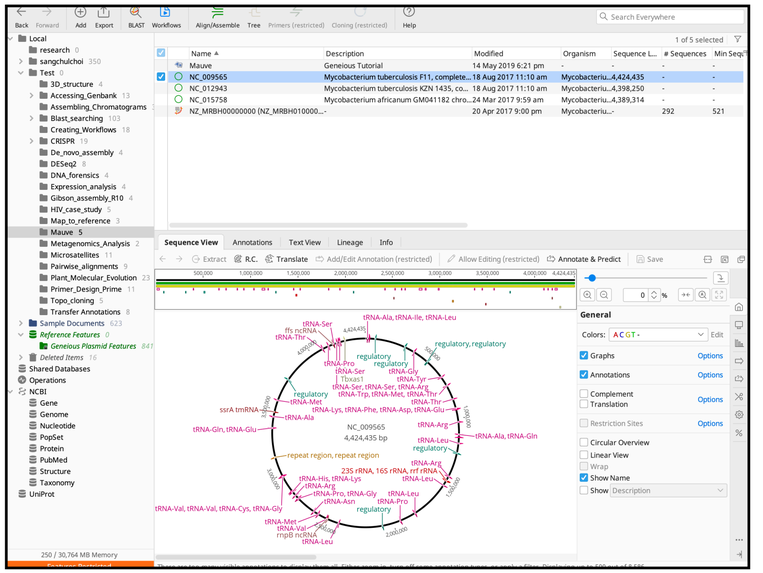
통계 프로그램 R and RStudio: bioconductor

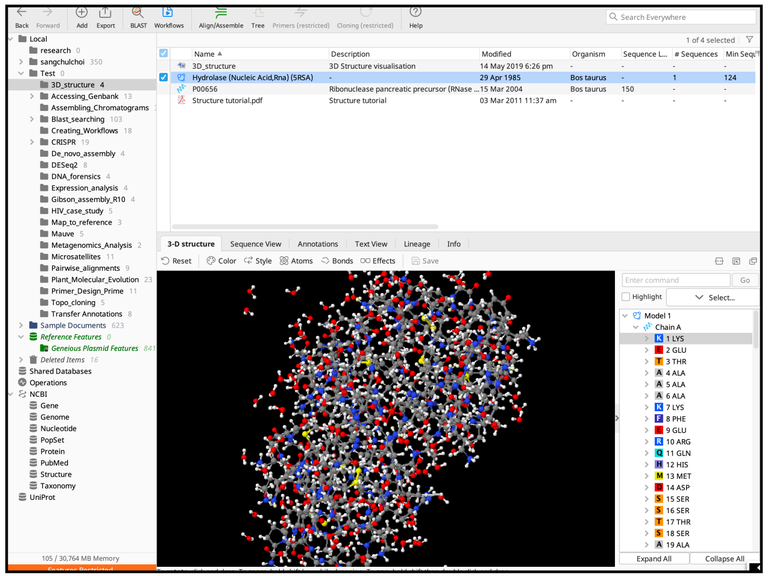


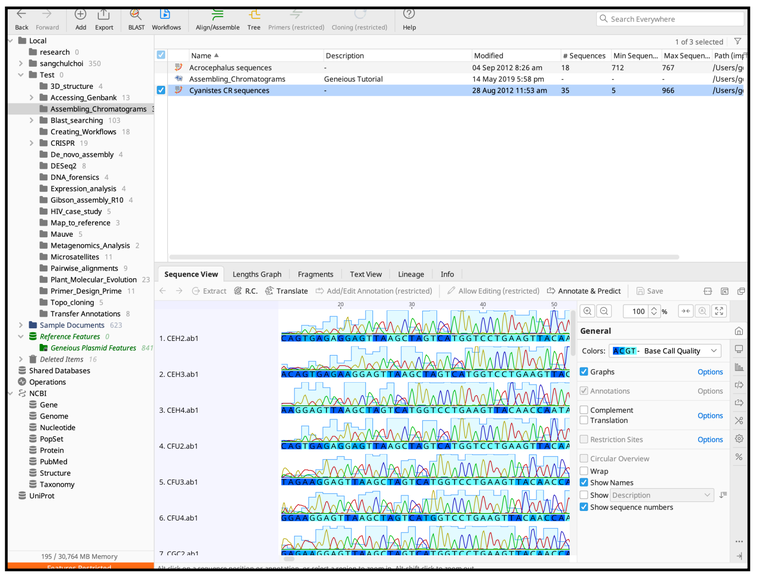
Geneious


geneious
Features
Resources
Support
Pricing
My Account
Free Trial
From Data to Discovery — Faster
Geneious Prime puts industry-leading bioinformatics and molecular biology tools directly into researchers hands, streamlining sequence analysis and insights.
Try for Free
Pricing
Meet the world's leading bioinformatics software platform
One platform with all the molecular biology and sequence analysis tools your organization needs. Whether you are looking to enhance productivity, increase visibility and insights, or reduce errors and risk – our platform unlocks the value in your lab's data.
EASE OF USE
POWERFUL TOOLS
CENTRAL REPOSITORY
INTEGRATED PLATFORM



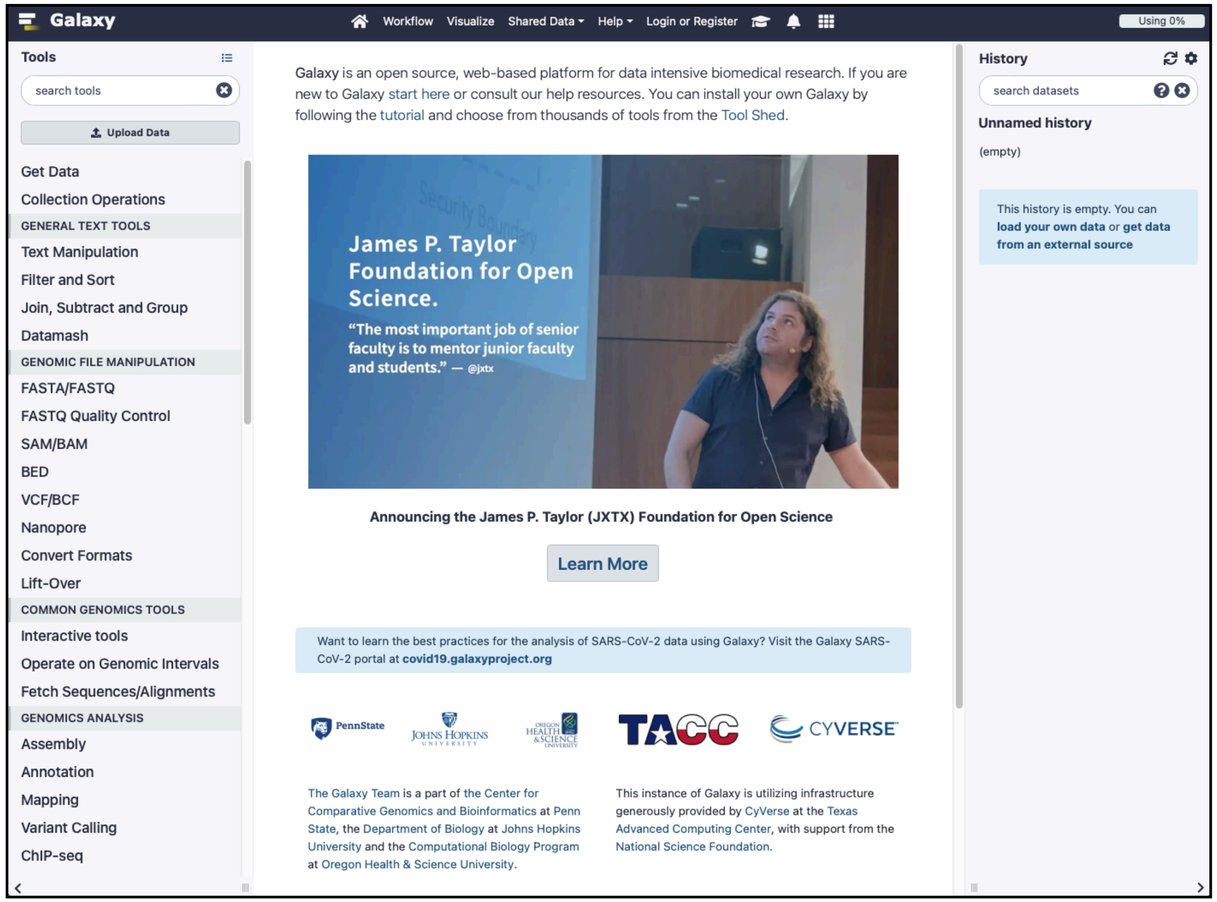




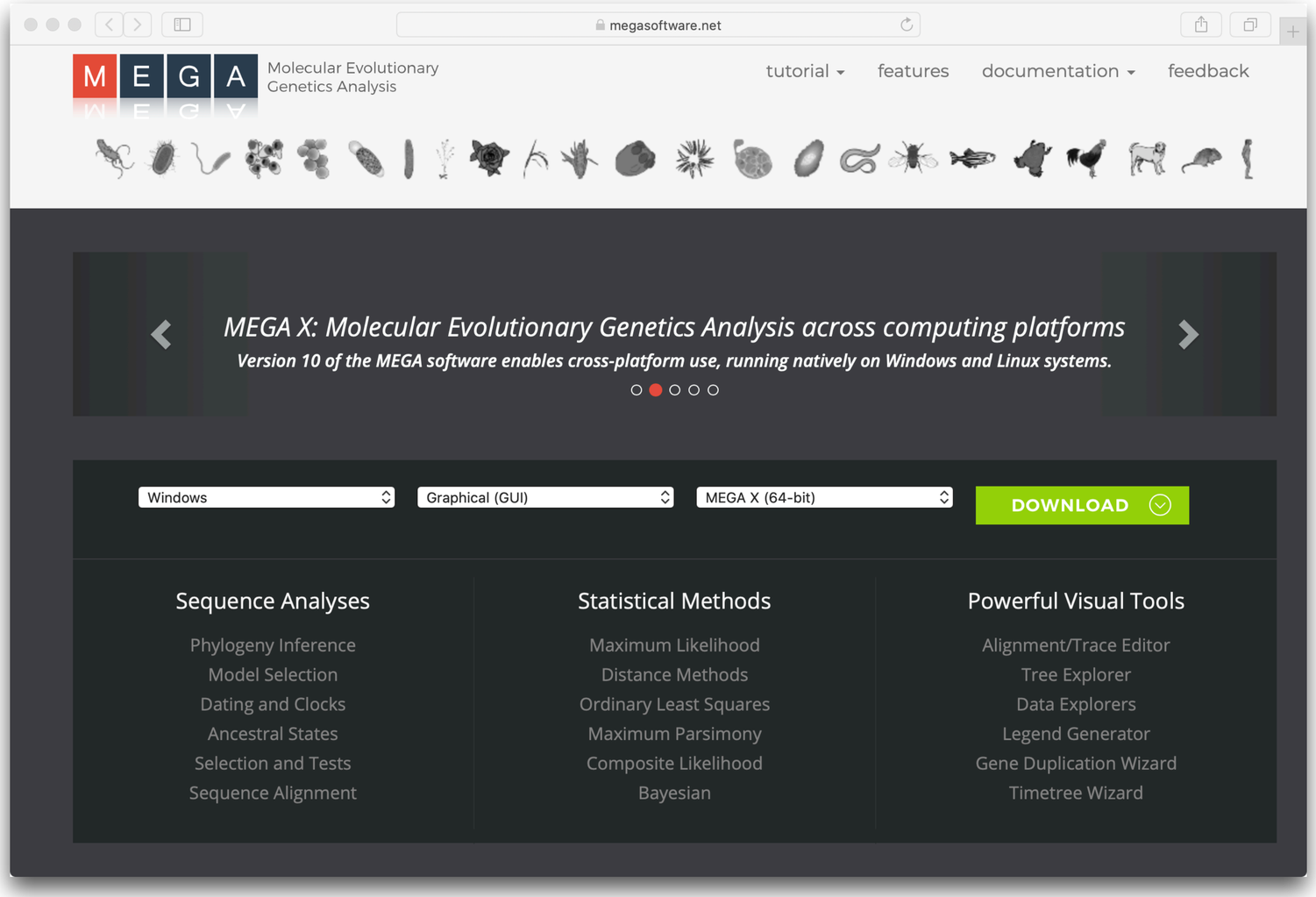


Galaxy

- 웹기반 생물정보학 도구의 사용자 인터페이스
- 대용량 데이터 처리
- 오픈소스



MegaX


megasoftware.net
M E G A
Molecular Evolutionary
Genetics Analysis
tutorial
features
documentation
feedback
MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms
Version 10 of the MEGA software enables cross-platform use, running natively on Windows and Linux systems.
Windows
Graphical (GUI)
MEGA X (64-bit)
DOWNLOAD
Sequence Analyses
Phylogeny Inference
Model Selection
Dating and Clocks
Ancestral States
Selection and Tests
Sequence Alignment
Statistical Methods
Maximum Likelihood
Distance Methods
Ordinary Least Squares
Maximum Parsimony
Composite Likelihood
Bayesian
Powerful Visual Tools
Alignment/Trace Editor
Tree Explorer
Data Explorers
Legend Generator
Gene Duplication Wizard
Timetree Wizard

END OF SLIDES

Access for free at https://openstax.org/books/biology-2e/pages/1-introduction