생물정보학과 디지털육종

(주)바이오투

2023-08-16

목 차

from Bioinformatics To biology

(주)바이오투는 최적화된 바이오 데이터 분석/처리 기술을 바탕으로,

고객이 **바이오 데이터의 진정한 의미와 가치**를 발견하고,

인사이트를 확보할 수 있도록 도움을 드립니다.

이를 통해 최적의 **형질전환 분자마커** 및 **진단용 분자마커**를 개발하고 키트화된 제품을 준비하고 있습니다.

(주)바이오투

(주)바이오투는 최적화된 바이오 데이터 분석/처리 기술을 바탕으로,

고객이 **바이오 데이터의 진정한 의미와 가치**를 발견하고,

인사이트를 확보할 수 있도록 도움을 드립니다.

이를 통해 최적의 **형질전환 분자마커** 및 **진단용 분자마커**를 개발하고 키트화된 제품을 준비하고 있습니다.

회사 일반현황 및 연혁

회 사 명	주식회사 바이오투	대표자	최준경, 이봉우
사업분야	▪ 생물정보 분석 , 생물정보 시스템 개발 ▪ 실험 분석, 분자마커 키트 개발		
주 소	▪ 대전광역시 유성구 테크노2로 187, B동 412호 나		
전화번호	▪ 전화: 042 -710 - 0077 / FAX: 070 - 7585 - 5344		
회사설립년도	▪ 2021년 08월		
해당부문 종사기간	▪ 2021년 08월 ~ 2023년 6월 (1년 10개월)		

주요 연혁 HISTORY

2023~	**신규 국가 표준에 따른 NABIC DB 재구축 및 서비스 기술 개발(07월~)** 관엽식물 가해 뿌리썩이선충 및 미소해충 신속 분류·동정 기술 개발 수주(01월) 국외 활엽수 목재 수종식별 바코드 자동화시스템 개발 연구 과제 수주(01월)
~ 2023	물봉선속 GBS 분석 2SET 유전자원분석 (01월) RNA-SEQ을 통한 사과 2품종 측지관련 유 전자 발현분석 및 SNP분석(02월) 분자마커 개발을 위한 염색체 Consensus (02월)
~2022	클로렐라 WGS 데이터를 이용한 SNP, INDEL 분석외 2건 (12월) 플라스틱 산화 곰팡이 탄소원 조건별(11월) 무 종자생산 연관 유전집단 GBS 분석 (10월)
~	온도변화에 따른 국화 꽃색 발현 관련 (10월) 수박 GBS SNP matrix 분석 (10월) 마우스 종양 관련 RNA 4샘플 3반복 DEG분석 (10월)
~	GBS 방법에 의한 남생이 유연관계 분석 (11월) (원산지검정과) 벼(쌀) DB 구축을 위한 유전체 정보 분석 계약(10월) 농작물 주요 병해충,잡초 유전,표본자원 관리 및 유지 프로그램 구축(7월) 농생명 연구데이터 분석 공유를 위한 최적 인프라 체계 연구(6월)
2022~	기름나물속 추가 분류군 엽록체 유전체 분석 (3월) RNASeq 통한 사상성진균 유전자 발현 연구,개발(3월) 2022년 우포따오기 개체별 유전자 특성 분석 연구(3월)

(주)바이오투

(주)바이오투는 최적화된 바이오 데이터 분석/처리 기술을 바탕으로,

고객이 **바이오 데이터의 진정한 의미와 가치**를 발견하고,

인사이트를 확보할 수 있도록 도움을 드립니다.

이를 통해 최적의 **형질전환 분자마커** 및 **진단용 분자마커**를 개발하고 키트화된 제품을 준비하고 있습니다.

주요사업 내용

- 게놈분석서비스
- 전사체분석서비스
- SNP 마커 개발

- 바이오 S/W 개발
- 바이오 포털 DB 개발
- 유전체 브라우저 및 오믹스 브라우저

- Real-Time PCR 기반 실험
- 정밀진단 키트 제작
- 등온 PCR 기반 현장진단 키트 제작

- 대용량 실험장비를 통한 유전자형 분석
- 실험장비 및 실험기자재 검증
- 유전자형 기반 빅 데이터 축적

우리의 주제

생물정보학과 디지털육종

생물정보학으로 뭘 하나요?

육종에 디지털을 어떻게 적용하나요?

생물정보학과 디지털육종을 어떻게 연결하나요?

강의의 목적

종자생명산업 맞춤형 인력양성 사업

생물정보학 분야에서 일하려면 뭘 준비해야 하나요?

디지털육종을 위해 할 수 있는 일이 뭔가요?

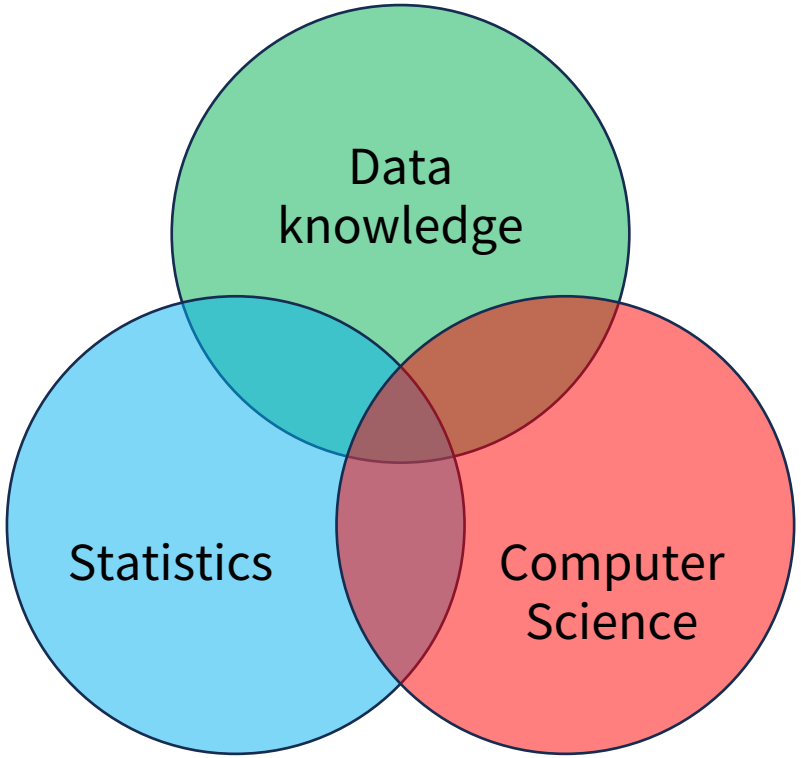
생물정보학과 디지털육종, 하려면 뭘 준비해야 하나요?

I. 생물정보학

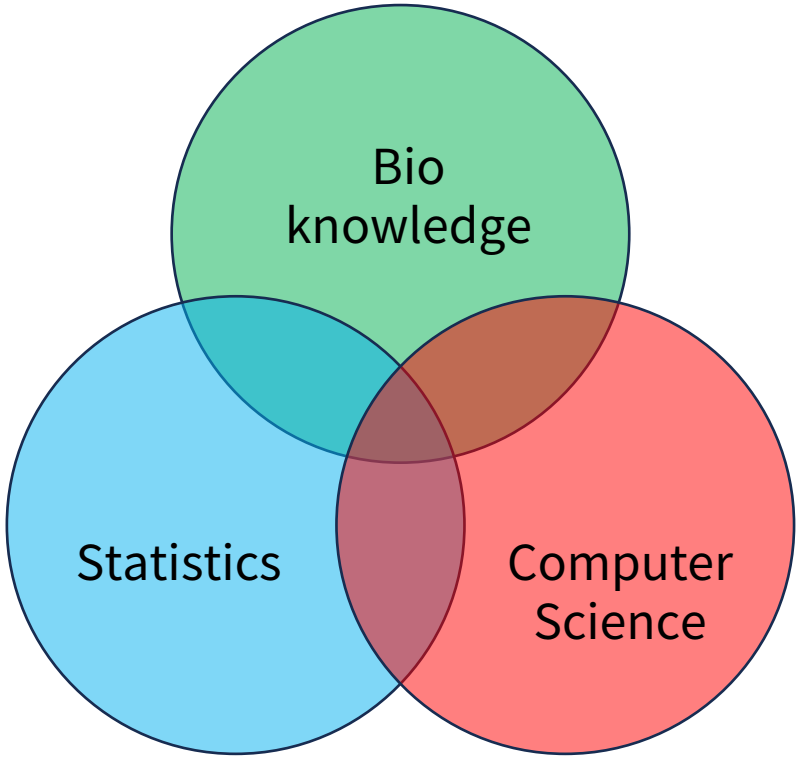
1. 생물정보학이란?

생물정보학이란?

컴퓨터를 이용하여 생명과학을 연구하는 모든 분야



Data Science



Bioinformatics

생물정보학이란?

컴퓨터를 이용하여 생명과학을 연구하는 모든 분야

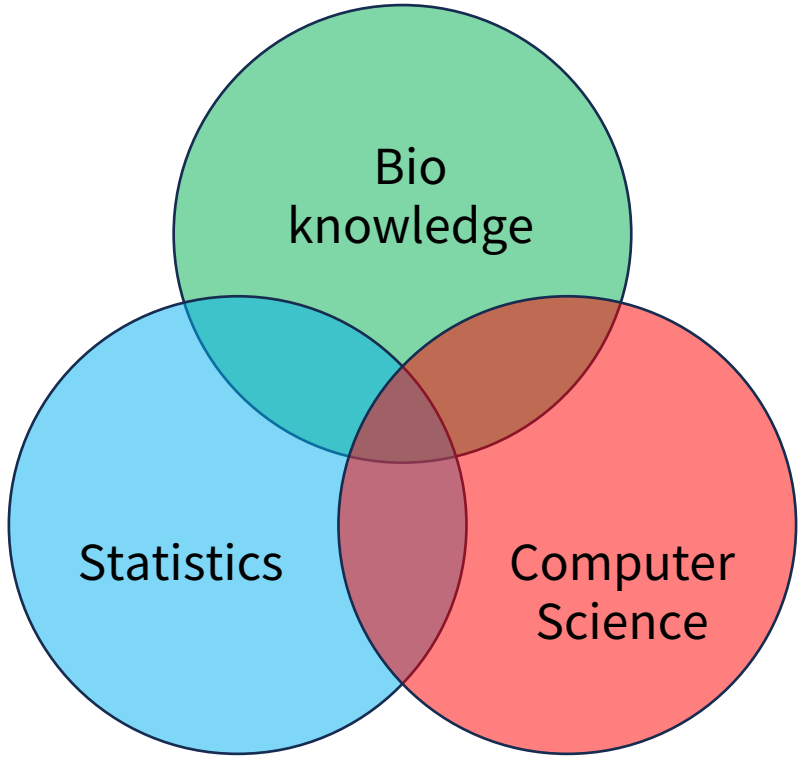
- 생물학적 데이터를 저장, 검색, 구성, 분석하는 데,
  IT를 사용하는 모든 분야

- 1970년 Paulien Hogeweg와 Ben Hesper가
  처음으로 용어 사용

- 처음 용어의 뜻:
  "생물 시스템 내에서의 정보 프로세스 연구"

- 현재는 좀 더 넓은 의미로 사용되어
  - 컴퓨터 구조 생물학,
  - 화학 생물학,
  - 시스템 생물학,
  - 데이터 통합,
  - 시스템 모델링

등을 포괄적으로 사용



Bioinformatics

생물정보학의 발달과정

최초의 생물정보학 논문

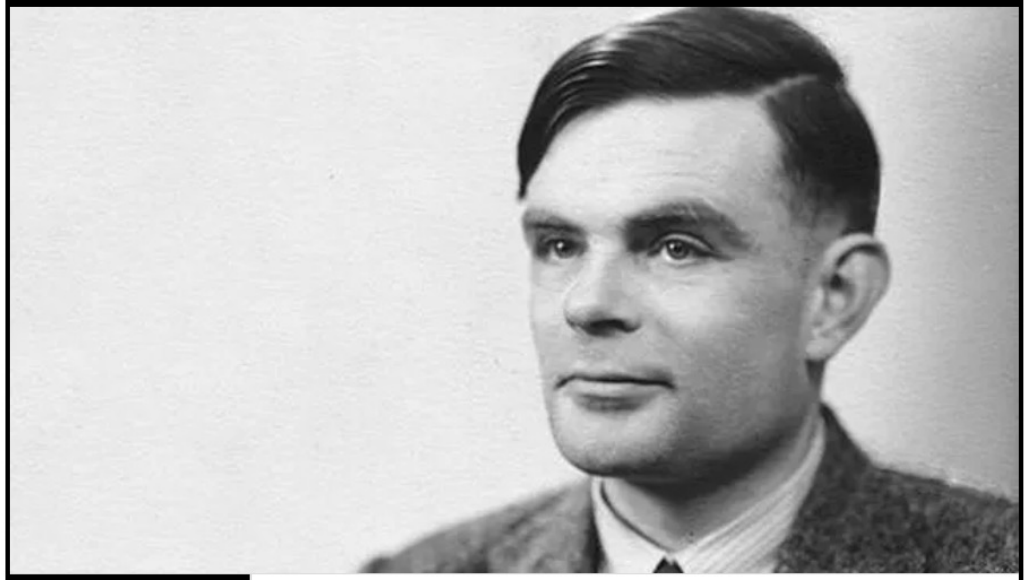
1952년

형태형성의 화학적인 토대
(The Chemical Basis of Morphogenesis)

- 생물의 발생에서 새로운 형태가 생겨나는 과정인 '형태형성'을 모의 실험
- 지금은 당연한 유전자에서 태아가 발생하는 부분을 이 시기에 귀납적으로 예측

- https://namu.wiki/w/앨런%20튜링

OBE FRS
앨런 튜링
Alan Turing



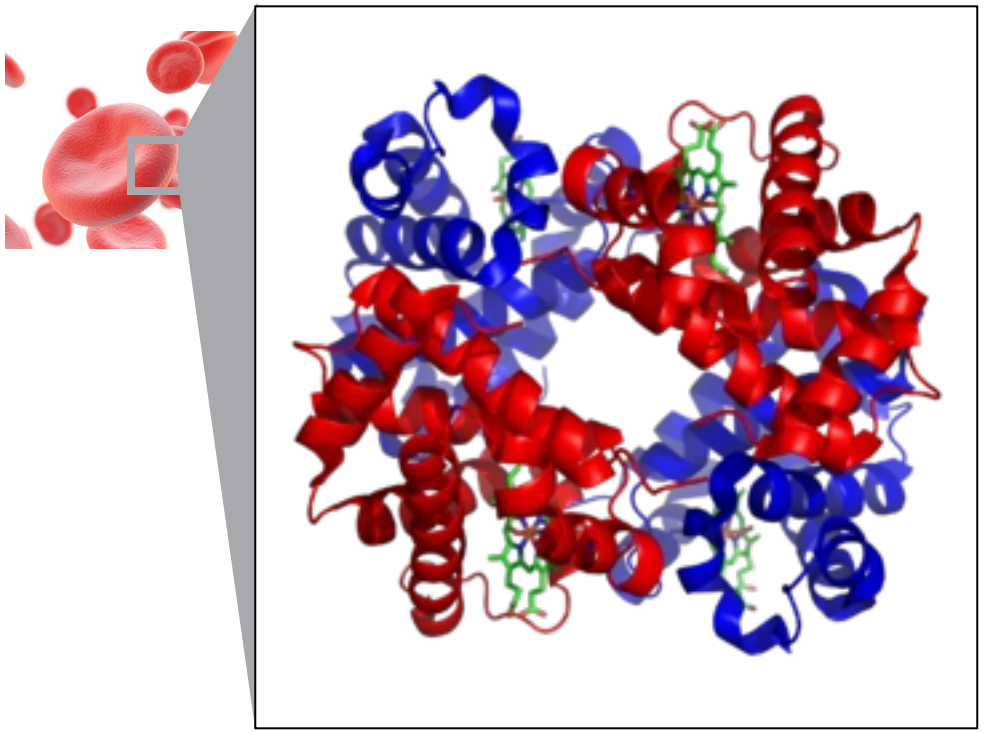
본명	앨런 매시슨 튜링 Alan Mathison Turing
출생	1912년 6월 23일 그레이트브리튼 아일랜드 연합왕국 런던 웨스트민스터 마이다 베일
사망	1954년 6월 7일 (향년 41세) 영국 잉글랜드 체셔 윔슬로
국적	영국
직업	컴퓨터과학자
분야	인지심리학, 수학, 논리학, 컴퓨터과학, 철학

생물정보학의 발달과정

영국 의학연구위원회(Medical Research Council; MRC)

컴퓨터를 이용한

헤모글로빈의 3차원 구조 예측



Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr **UAA Stop** **UAG Stop**	UGU Cys UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile **AUG Met**	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

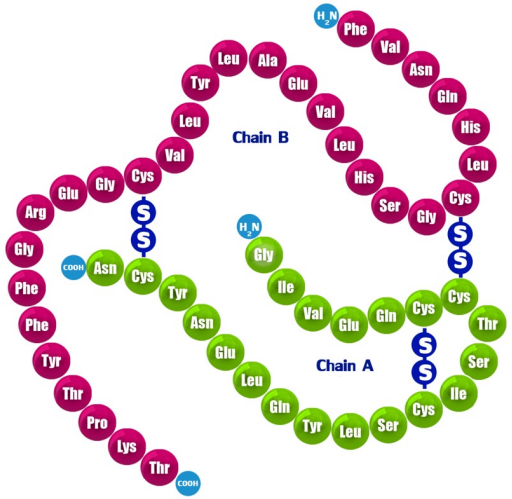
유전 부호 (Genetic Codon) 발견

생물정보학의 발달과정

소(Cow)에서

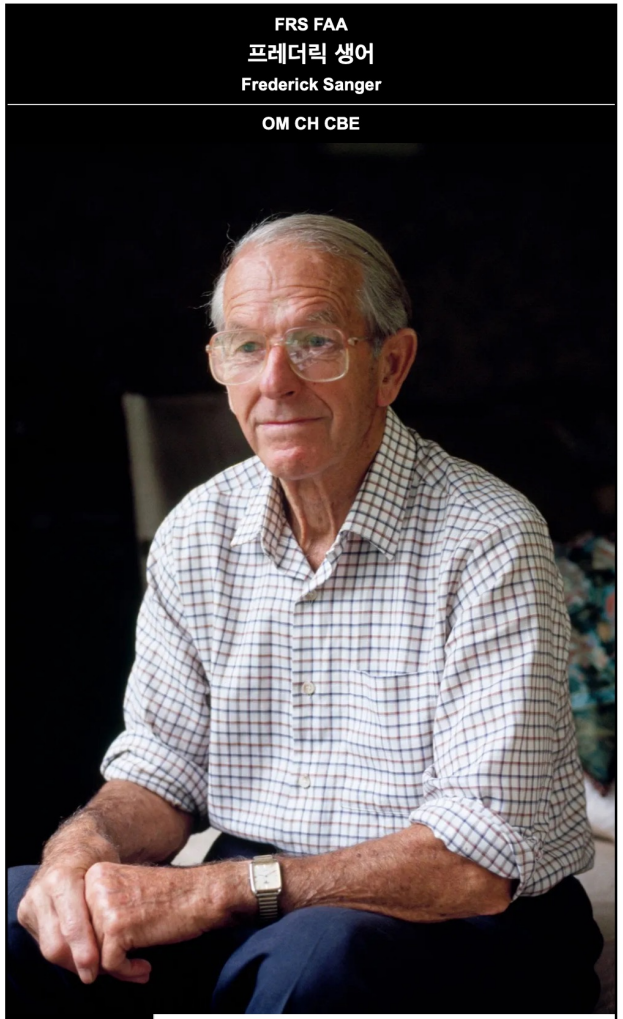
- 인슐린 B (1951년, 30개 amino acid 서열)와
- 인슐린 A (1952년, 21개 amino acid 서열)

완벽하게 분석



- 1975년 앨런 콜슨(Alan Coulson)과 함께 첫번째 DNA 해독법인 '음양법'(Plus and Minus technique)을 만들어 박테리오파지 φX174의 유전체 서열을 밝혀내어 최초로 DNA 유전체를 해독했으나 한계점이 많았다.

- 1977년 생어 연구팀은 사슬종결법(Chain-termination method),
  일명 **생어해독법** 혹은 **생어시퀀싱**을 개발하여 획기적으로 DNA 서열을 해독할 수 있게 만들었다.



출생	1918년 8월 13일 잉글랜드 글로스터셔 렌드콤
사망	2013년 11월 19일 (향년 95세) 잉글랜드 케임브리지셔 케임브리지
국적	영국
직업	생화학자
학력	케임브리지 대학교 세인트 존스 컬리지 (생화학 / M.Sc.)[1] 케임브리지 대학교 세인트 존스 컬리지 (생화학 / Ph.D.)[2]
주요 업적	아미노산 서열 해독 염기서열 해독법 개발

- https://namu.wiki/w/프레더릭%20생어

생물정보학의 발달과정

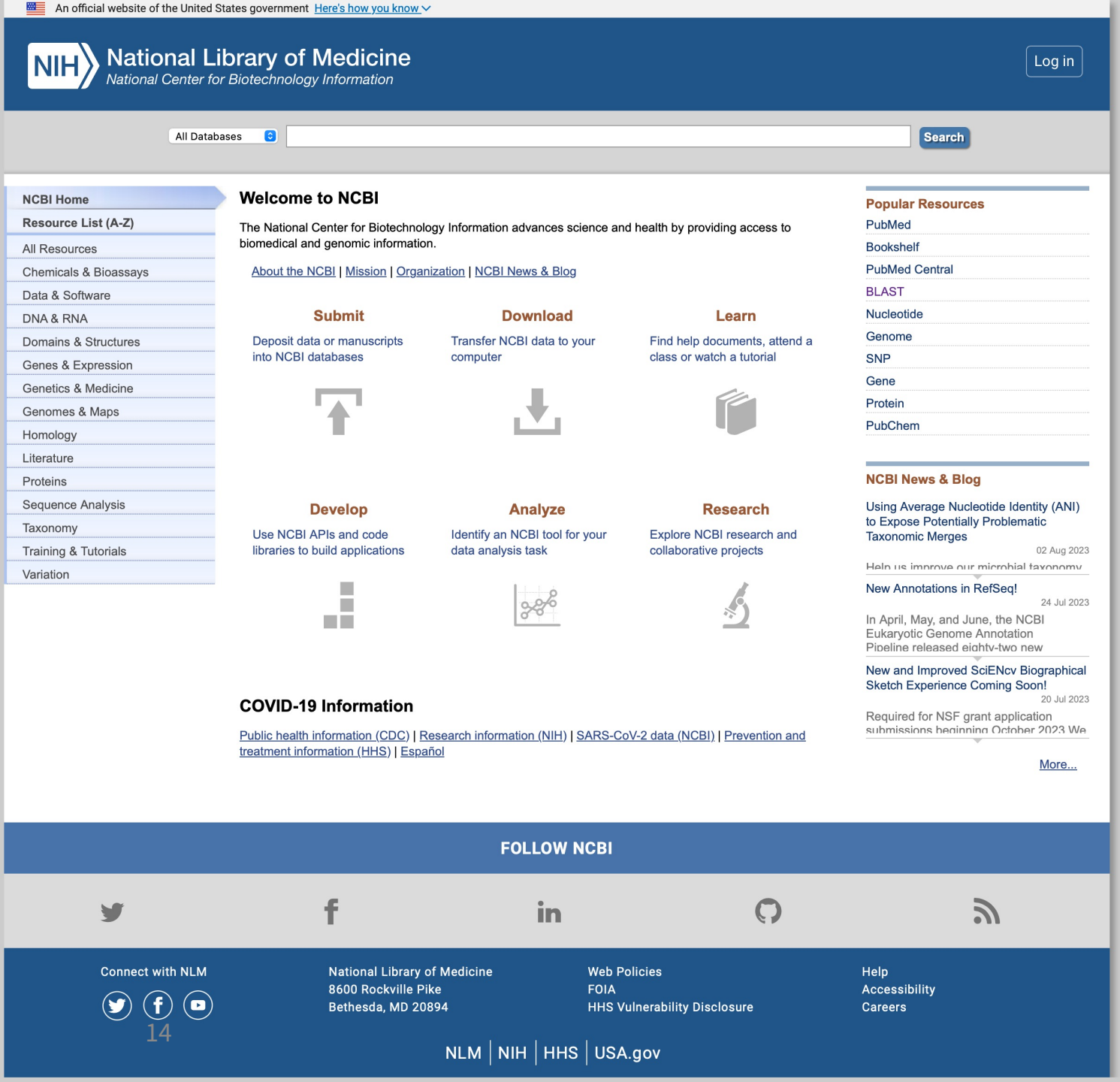
미국 국립생물공학정보센터
(National Center for Biotechnology Information, **NCBI**)

1988년 설립

- PubMed: 생명과학 및 의학 논문 DB
- GenBank: 유전체 서열 DB
- SRA: 시퀀싱 데이터 Archive용 DB

각종 생명공학 정보들을 담고 있으며,

이 모든 정보들은 Entrez 검색엔진으로 온라인으로 열람



- https://ko.wikipedia.org/wiki/미국_국립생물공학정보센터

Human Genome Project

2000년 완료 발표

J. Craig Venter (left)

Bill Clinton (center)

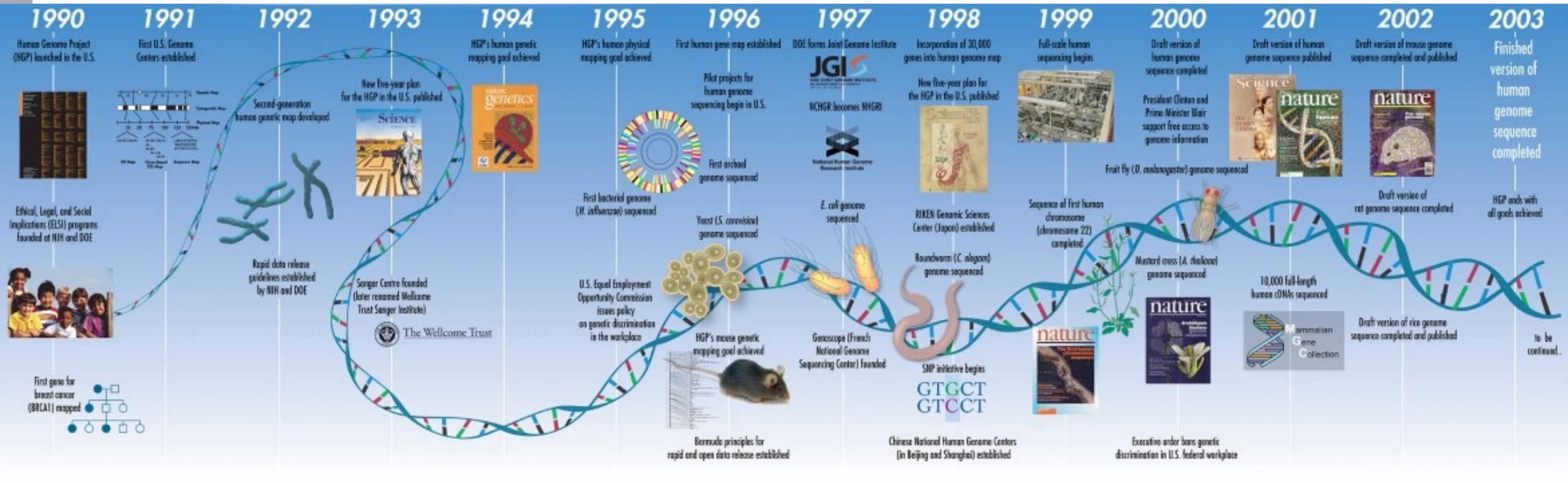
Francis S. Collins (right)

- https://www.genengnews.com/insights/the-human-genome-project-in-2020-hindsight/

생물정보학의 발달과정

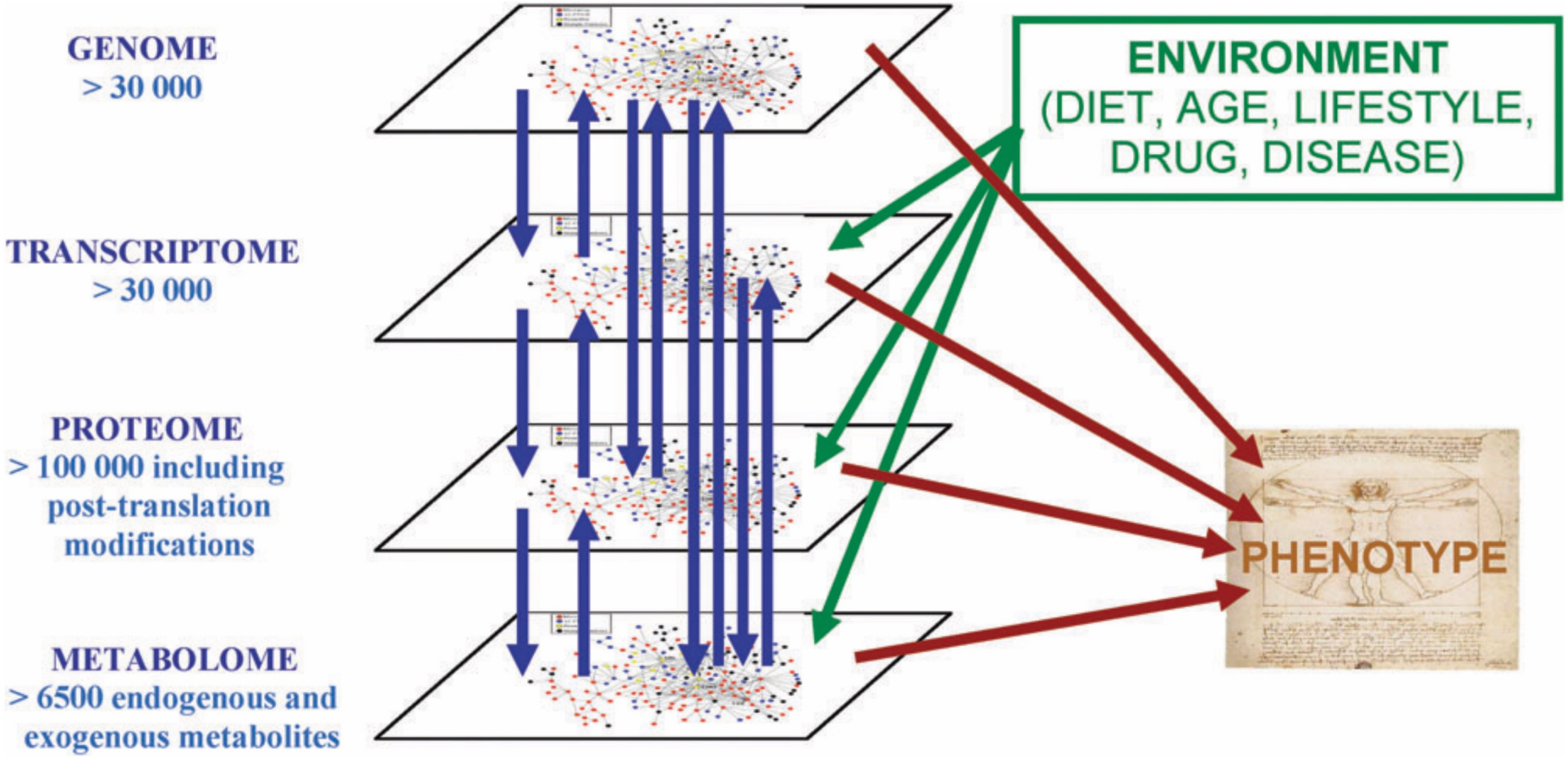
2003년 Human Genome Project 종료와 더불어

대용량 유전체 데이터 생성의 시대 도래



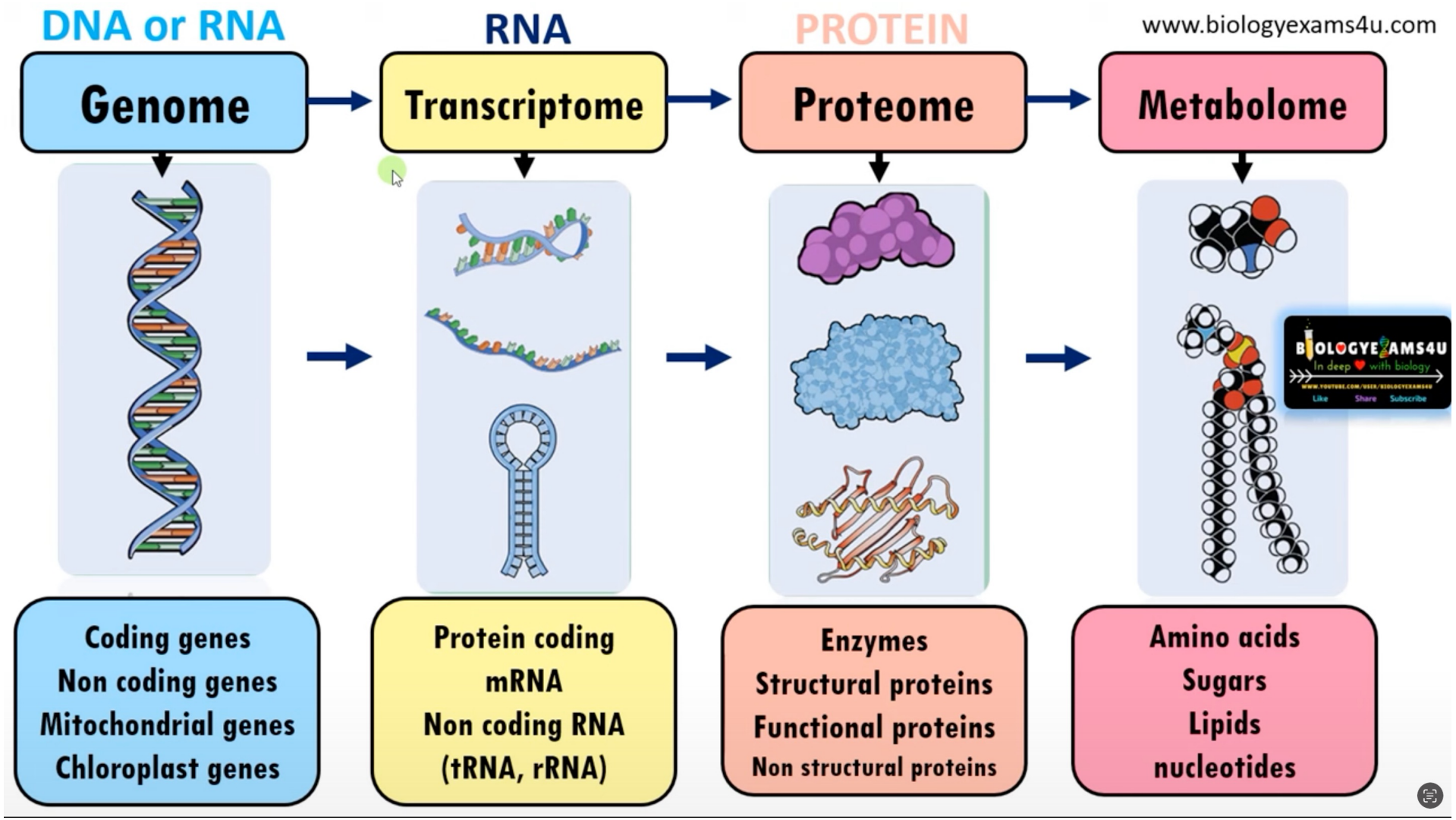
- https://www.stressmarq.com/june-26-2000-dna-sequence-released-by-human-genome-project/?v=3e8d115eb4b3

The complex interactions in biological systems



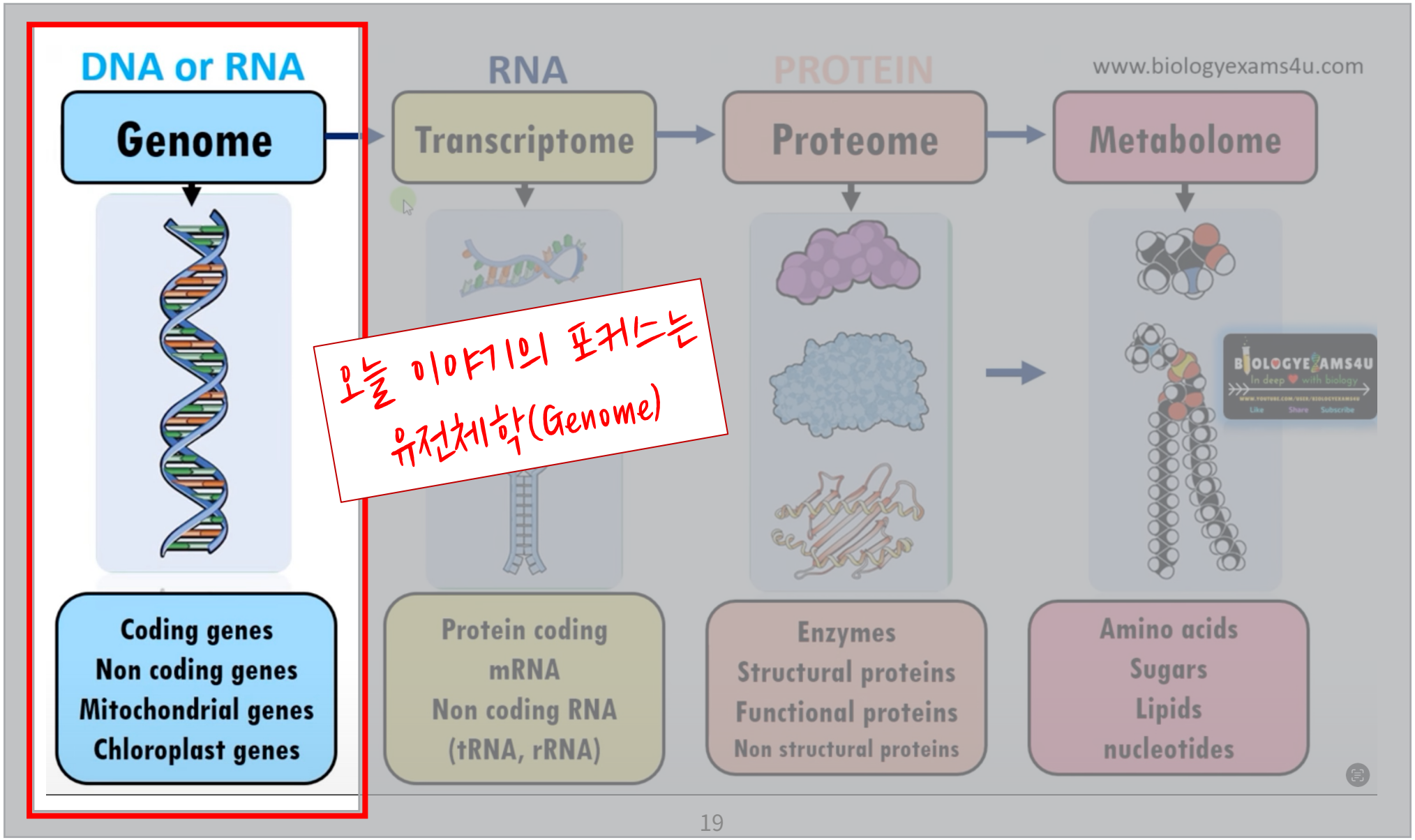
- Dunn WB *et al.* Chem Soc Rev. 2011

Bioinformatics의 범위



- https://www.youtube.com/watch?v=FJ5iN-v0MFM

Bioinformatics의 범위

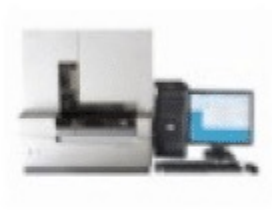


- https://www.youtube.com/watch?v=FJ5iN-v0MFM

I. 생물정보학

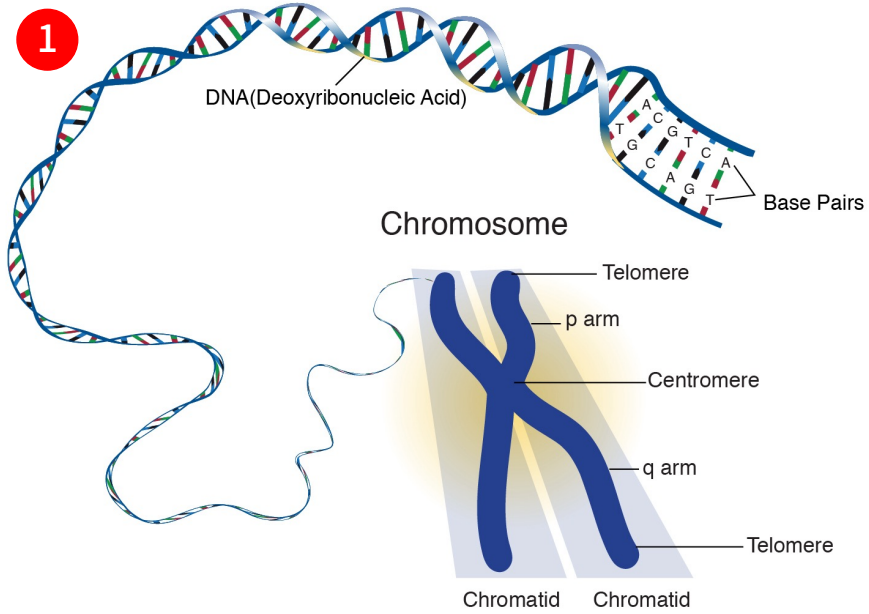
2. Next-Generation Sequencing (NGS)

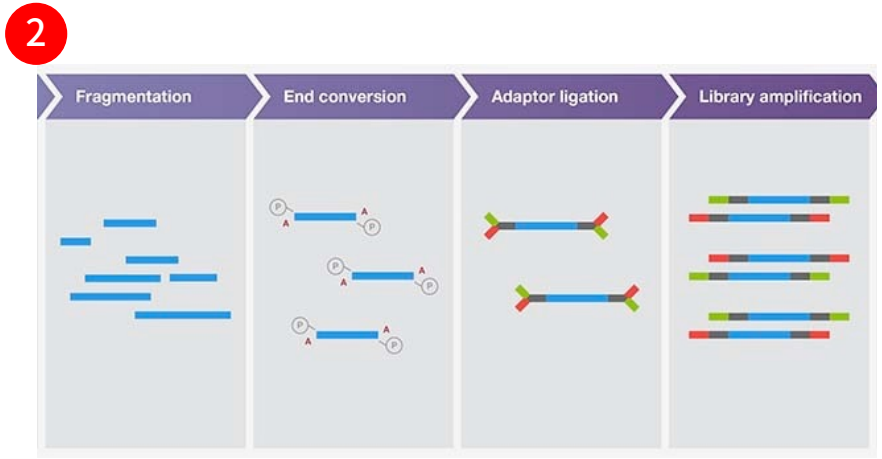
Next-Generation Sequencing (NGS)의 발달

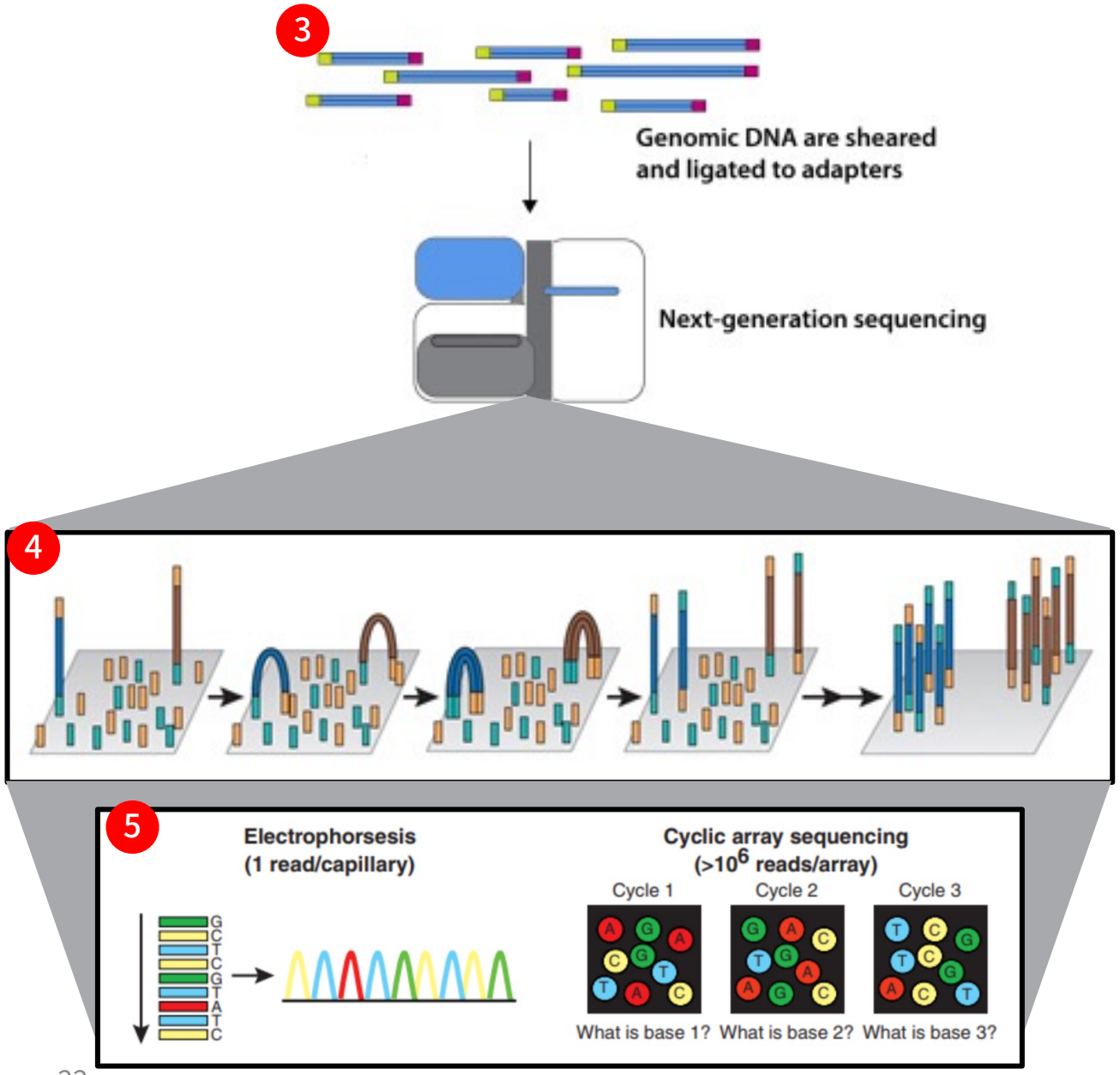
First generation	Second generation (next generation sequencing)	Third generation
  	  	 
Sanger sequencing Maxam and Gilbert Sanger chain termination	454, Solexa, Ion Torrent, Illumina	PacBio Oxford Nanopore
Infer nucleotide identity using dNTPs, then visualize with electrophoresis	High throughput from the parallelization of sequencing reactions	Sequence native DNA in real time with single-molecule resolution
500–1,000 bp fragments	~50–500 bp fragments	Tens of kb fragments, on average
Short-read sequencing		Long-read sequencing

- https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

NGS 원리

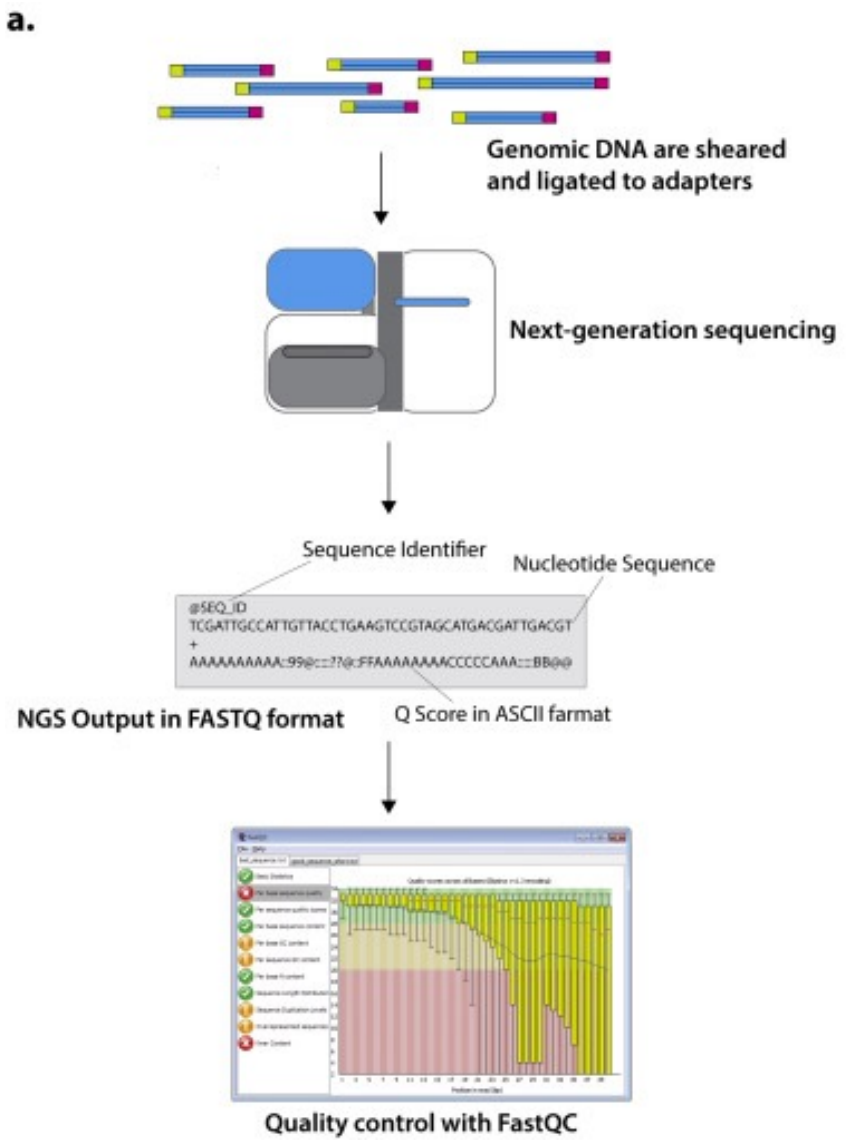






- Ang MY *et al.* Clin Chim Acta. 2019;498:38-46

NGS 원리



- Ang MY *et al.* Clin Chim Acta. 2019;498:38-46

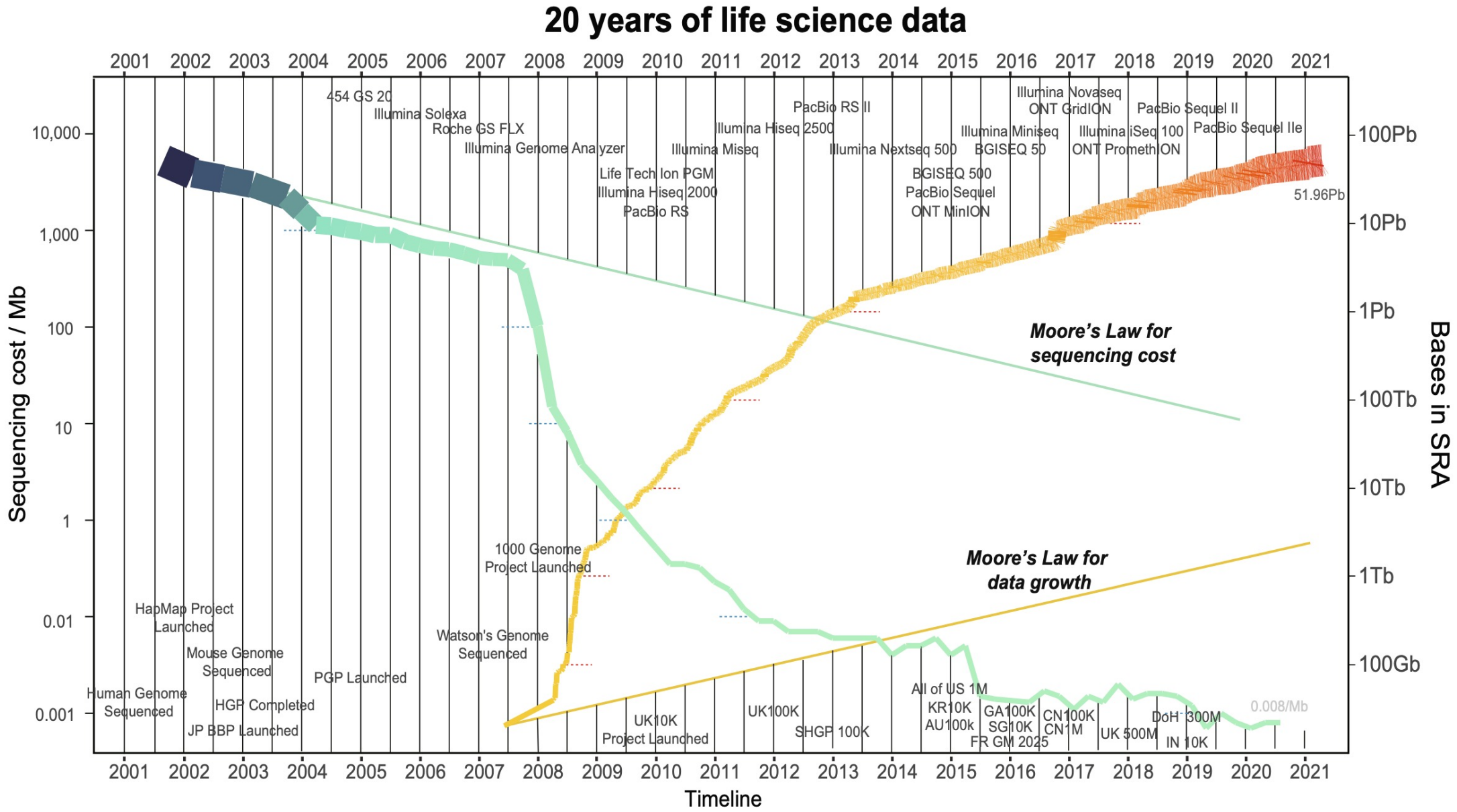
NGS 장비의 종류




	HiSeq X Ten	HiSeq 2500				NextSeq 500		MiSeq
		Hiseq v4	Truseq v3	Rapid v2	Rapid	High-output	Mid-output	v3
Read Length (bp)	2x150	2x125	2x100	2x250	2x150	2x150	2x150	2x300
Run Time	3일	6일	11일	60시간	40시간	29시간	26시간	~65시간
Total Output	1.8 Tb	1 Tb	~600 Gb	~300 Gb	~110 Gb	~100 Gb	~36 Gb	~ 5.5 Gb
단일 read 갯수	60억	40억	30억	12억	6억	4억	1.3억	2천5백만
Quality Score	> 75% above Q30	> 80% above Q30	> 80% above Q30	> 75% above Q30	> 75% above Q30	> 75% above Q30	> 75% above Q30	> 70% above Q30

- https://m.ibric.org/miniboard/read.php?id=2278&Board=report

NGS의 파급력



- Jiang *et al.* Ccf Transactions High Perform Comput 3, 344–352 (2021).

최신 NGS 장비의 능력

Table 1: NovaSeq 6000 System flow cell specifications

Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
1 × 35 bp	N/A	N/A	N/A	280-350 Gb
2 × 50 bp	65-80 Gb	134-167 Gb	333-417 Gb	N/A
2 × 100 bp	134-167 Gb	266-333 Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 × 250 bp	325-400 Gb	N/A	N/A	N/A
Single reads CPF	0.65-0.8B	1.3-1.6B	3.3-4.1B	8-10B
Paired-end reads CPF	1.3-1.6B	2.6-3.2B	6.6-8.2B	16-20B
Quality scores[c]				
1 × 35 bp	Q30 ≥ 90%			
2 × 50 bp	Q30 ≥ 90%			
2 × 100 bp	Q30 ≥ 85%			
2 × 150 bp	Q30 ≥ 85%			
2 × 250 bp	Q30 ≥ 75%			
Run time[d]				
1 × 35 bp	N/A	N/A	N/A	~14 hr
2 × 50 bp	~13 hr	~13 hr	~16 hr	N/A
2 × 100 bp	~19 hr	~19 hr	~25 hr	~36 hr
2 × 150 bp	~25 hr	~25 hr	~36 hr	~44 hr
2 × 250 bp	~38 hr	N/A	N/A	N/A

최신 NGS 장비의 능력

Table 1: NovaSeq 6000 System flow cell specifications

Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
1 × 35 bp	N/A			280-350 Gb
2 × 50 bp	65-80 Gb		Gb	N/A
2 × 100 bp	134-167 Gb	Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 × 250 bp	325-400 Gb	N/A	N/A	N/A
Single reads CPF	0.65-0.8B	1.3-1.6B	3.3-4.1B	8-10B
Paired-end reads CPF	1.3-1.6B	2.6-3.2B	6.6-8.2B	16-20B
Quality scores[c]				
1 × 35 bp		Q30 ≥ 90%		
2 × 50 bp		Q30 ≥ 90%		
2 × 100 bp		Q30 ≥ 85%		
2 × 150 bp		Q30 ≥ 85%		
2 × 250 bp		Q30 ≥ 75%		
Run time[d]				
1 × 35 bp	N/A	N/A	N/A	~14 hr
2 × 50 bp	~13 hr	~13 hr	~16 hr	N/A
2 × 100 bp	~19 hr	~19 hr	~25 hr	~36 hr
2 × 150 bp	~25 hr	~25 hr	~36 hr	~44 hr
2 × 250 bp	~38 hr	N/A	N/A	N/A

68Gb / 1h

최신 NGS 장비의 능력

프로젝트당 평균 100Gb 정도 생산...

프로젝트당 최소 3 weeks 분석...

Table 1: NovaSeq 6000 System flow cell specifications

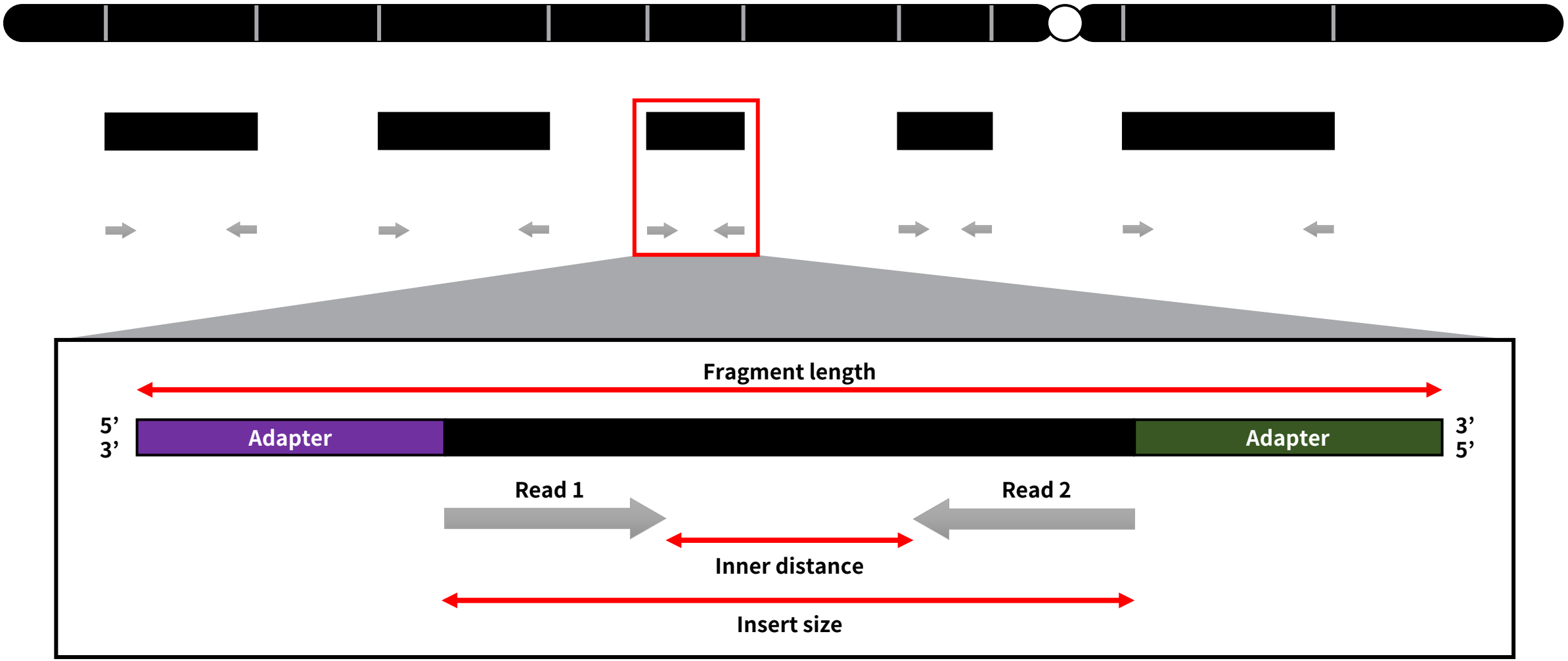
Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
1 × 35 bp	N/A			280-350 Gb
2 × 50 bp	65-80 Gb		Gb	N/A
2 × 100 bp	134-167 Gb	Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 × 250 bp	325-400 Gb	N/A	N/A	N/A
Single reads CPF	0.65-0.8B			8-10B
				16-20B
			Q30 ≥ 85%	
2 ×			Q30 ≥ 85%	
2 × 250 bp			Q30 ≥ 75%	
Run time[d]				
1 × 35 bp	N/A	N/A	N/A	~14 hr
2 × 50 bp	~13 hr	~13 hr	~16 hr	N/A
2 × 100 bp	~19 hr	~19 hr	~25 hr	~36 hr
2 × 150 bp	~25 hr	~25 hr	~36 hr	~44 hr
2 × 250 bp	~38 hr	N/A	N/A	N/A

68Gb / 1h

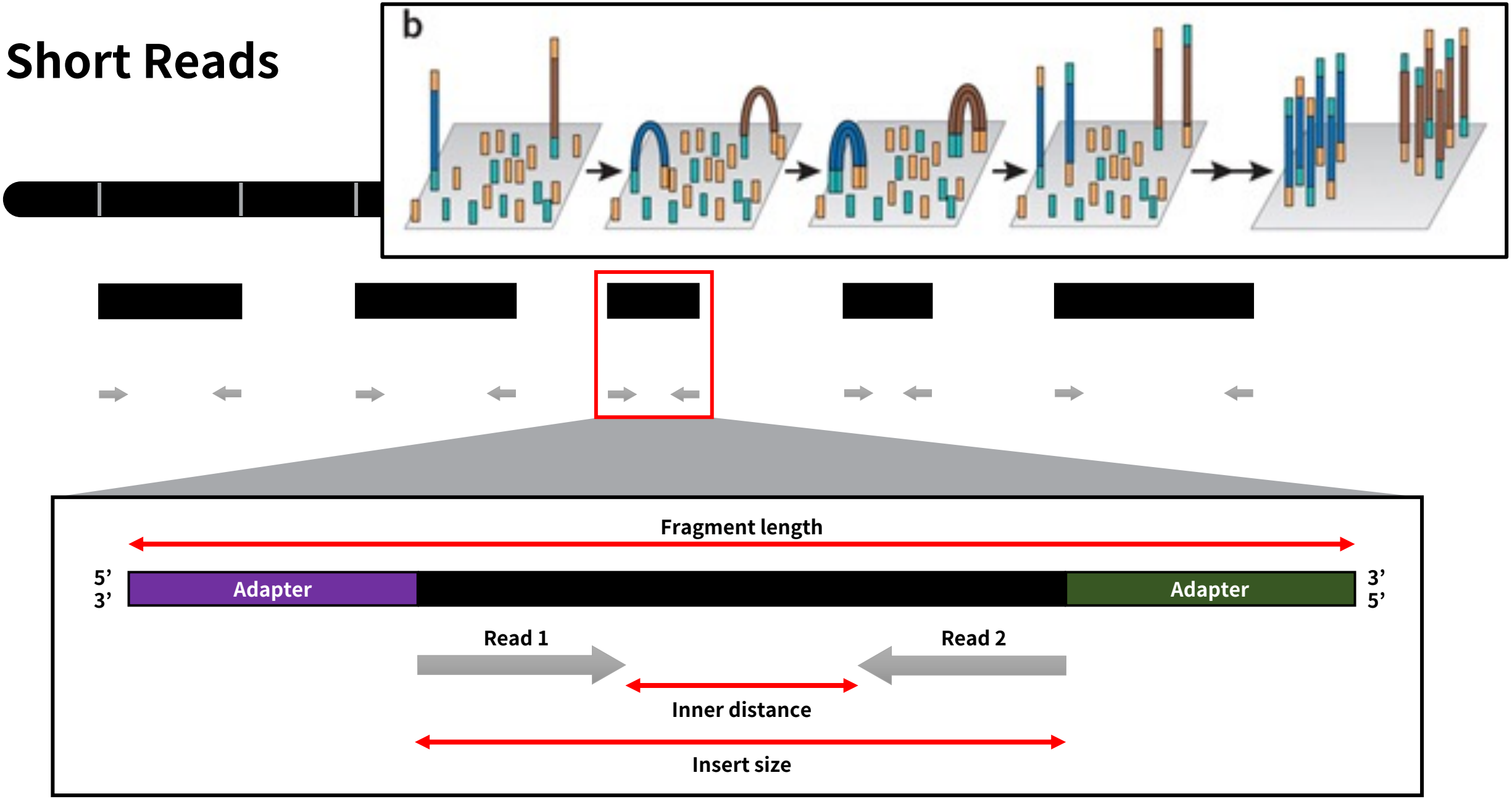
이들에 비 250 품종(30X 기준)을 한 번에 읽어 버릴 수 있다.

Oryza sativa genome size = 389 Mb

Short Reads


Fragment length
5’
3’
Adapter
Adapter
3’
5’
Read 1
Read 2
Inner distance
Insert size

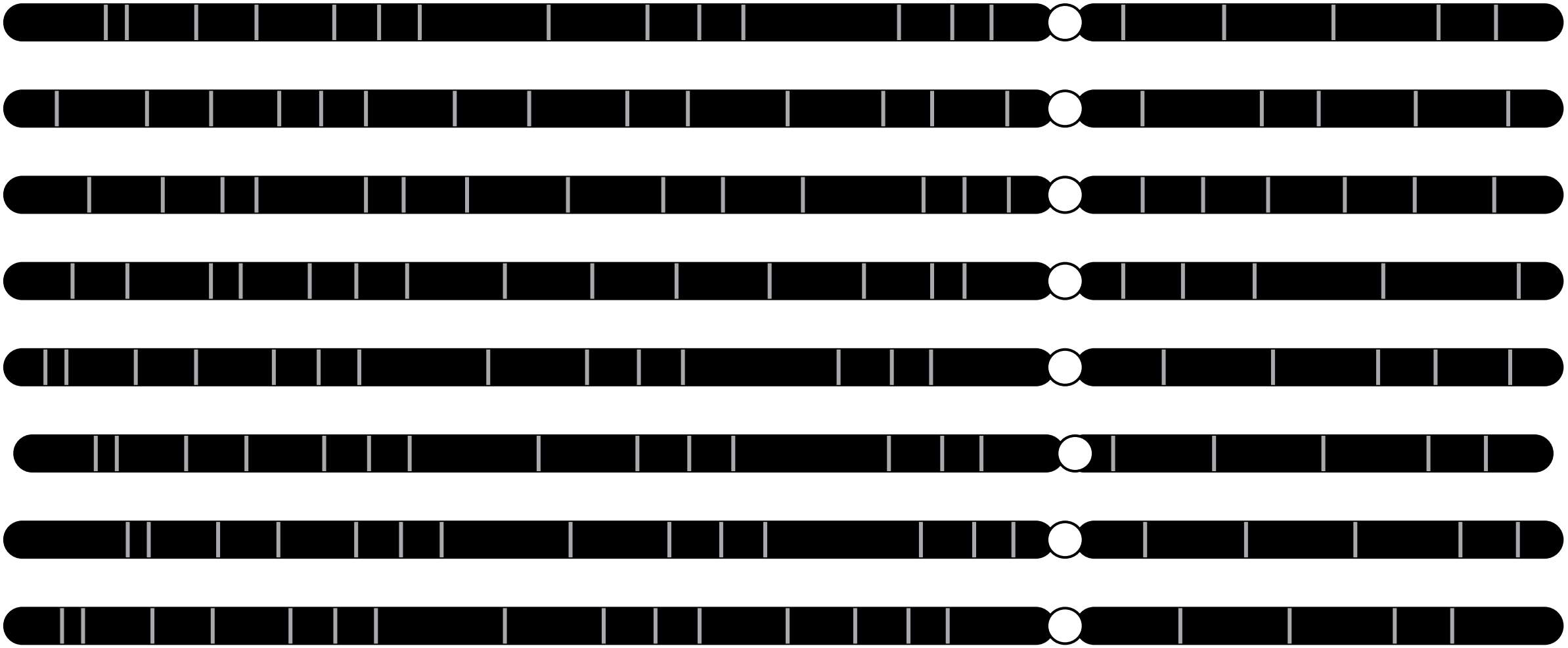
Short Reads


b
Fragment length
5’
3’
Adapter
Adapter
3’
5’
Read 1
Read 2
Inner distance
Insert size

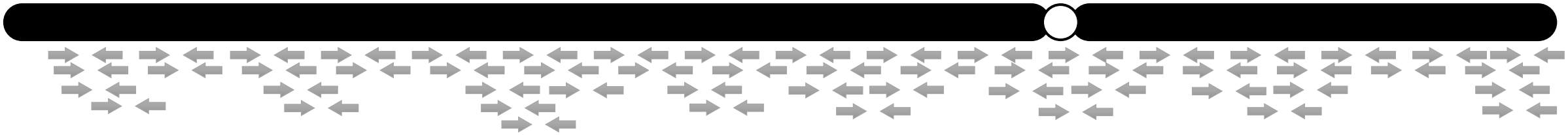
Short Reads

Short Reads

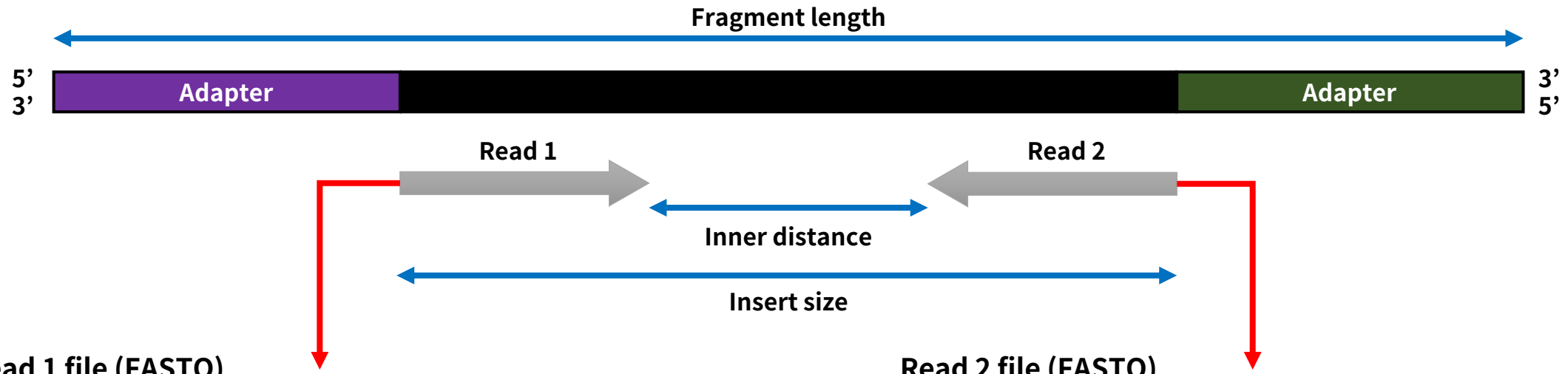
Short Reads



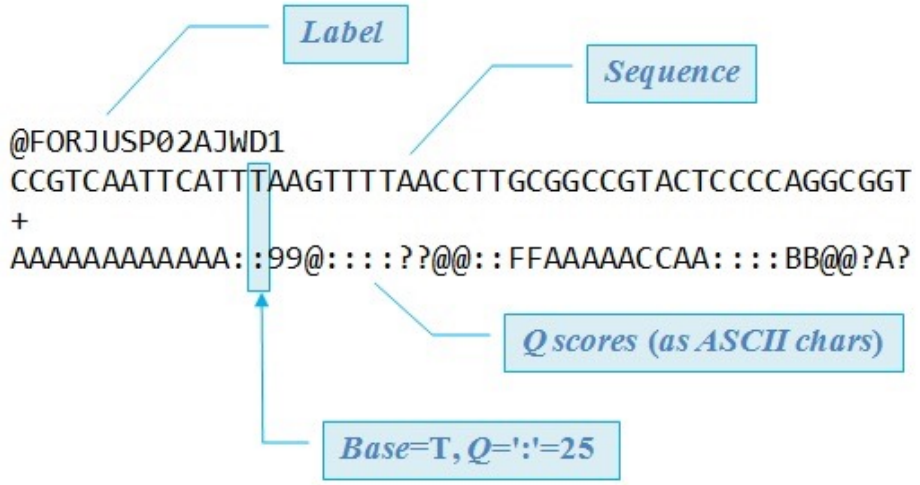
Short Reads



FASTQ



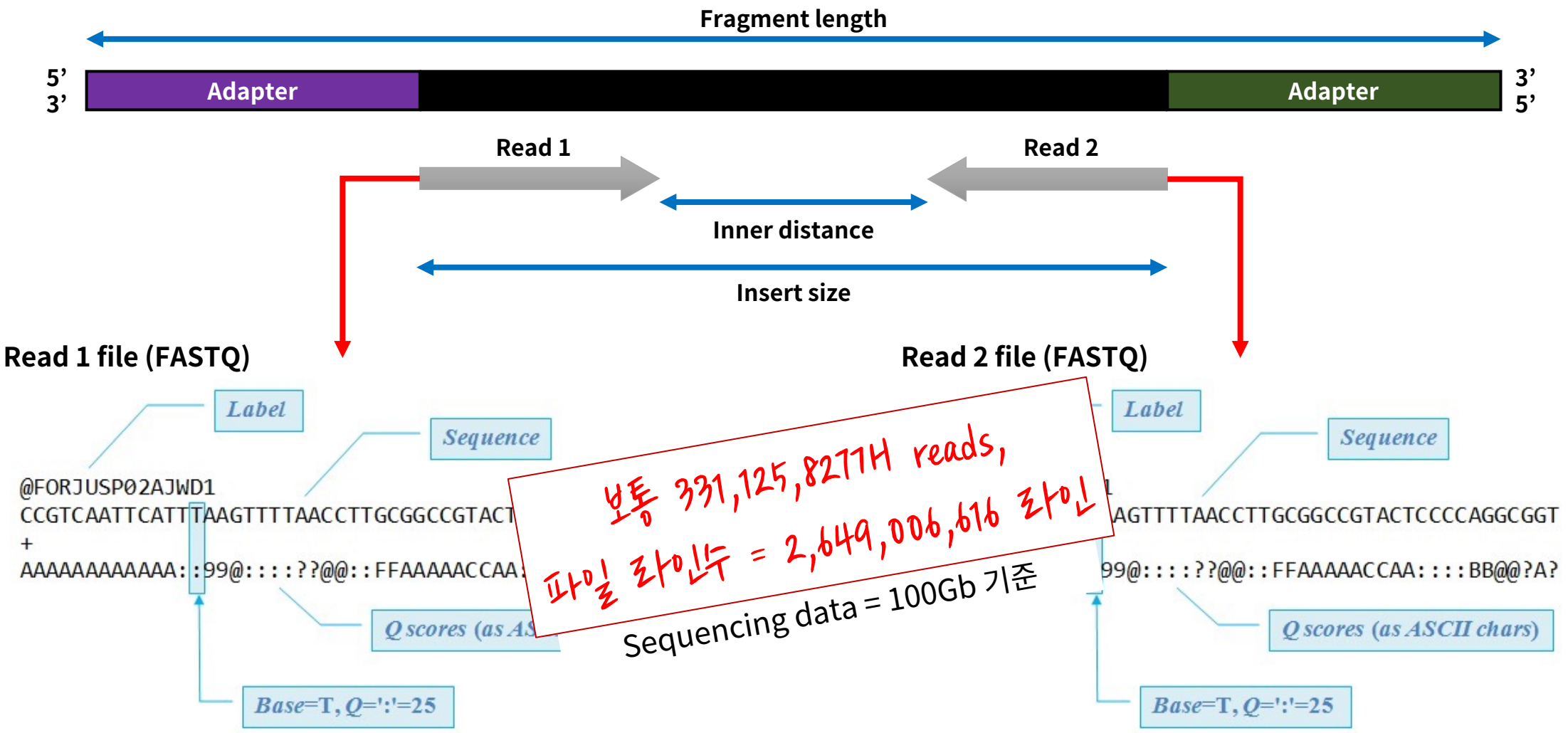
Read 1 file (FASTQ)



Read 2 file (FASTQ)

- https://www.drive5.com/usearch/manual/fastq_files.html

FASTQ



보통 331,125,827개 reads,
파일 라인수 = 2,649,006,616 라인

Sequencing data = 100Gb 기준

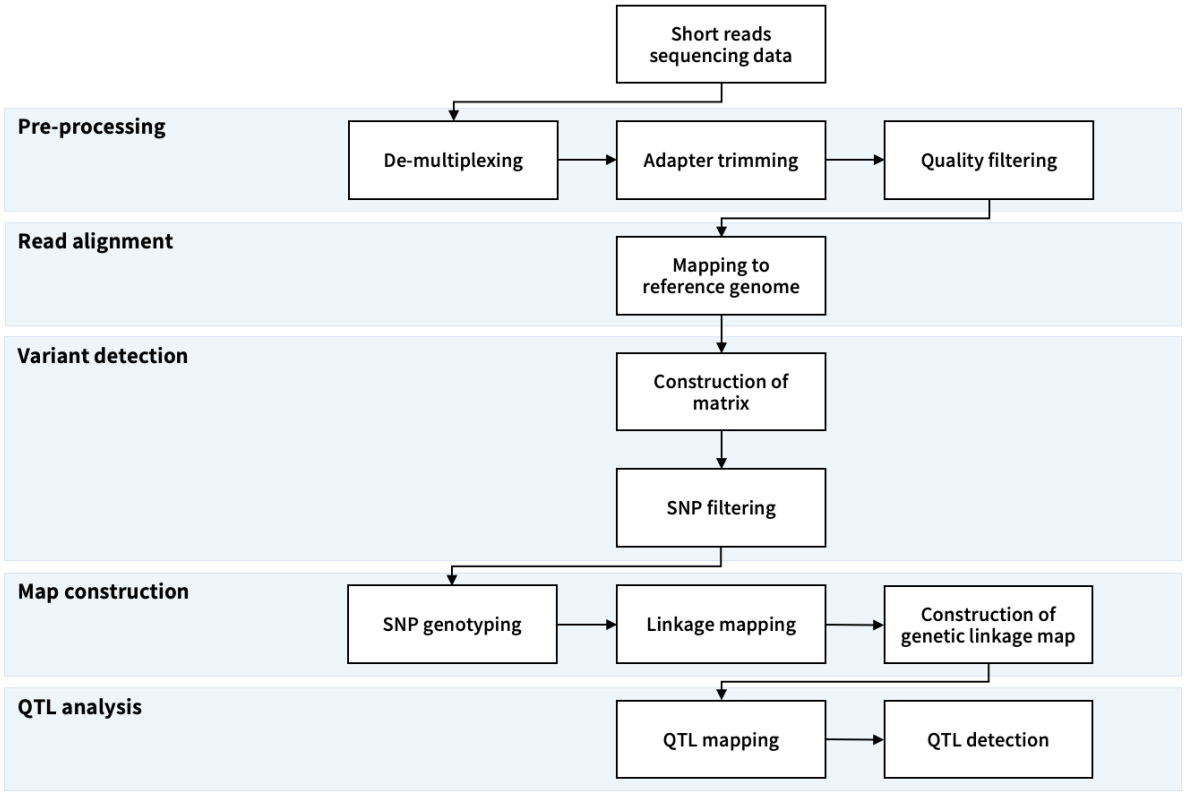
- https://www.drive5.com/usearch/manual/fastq_files.html

I. 생물정보학

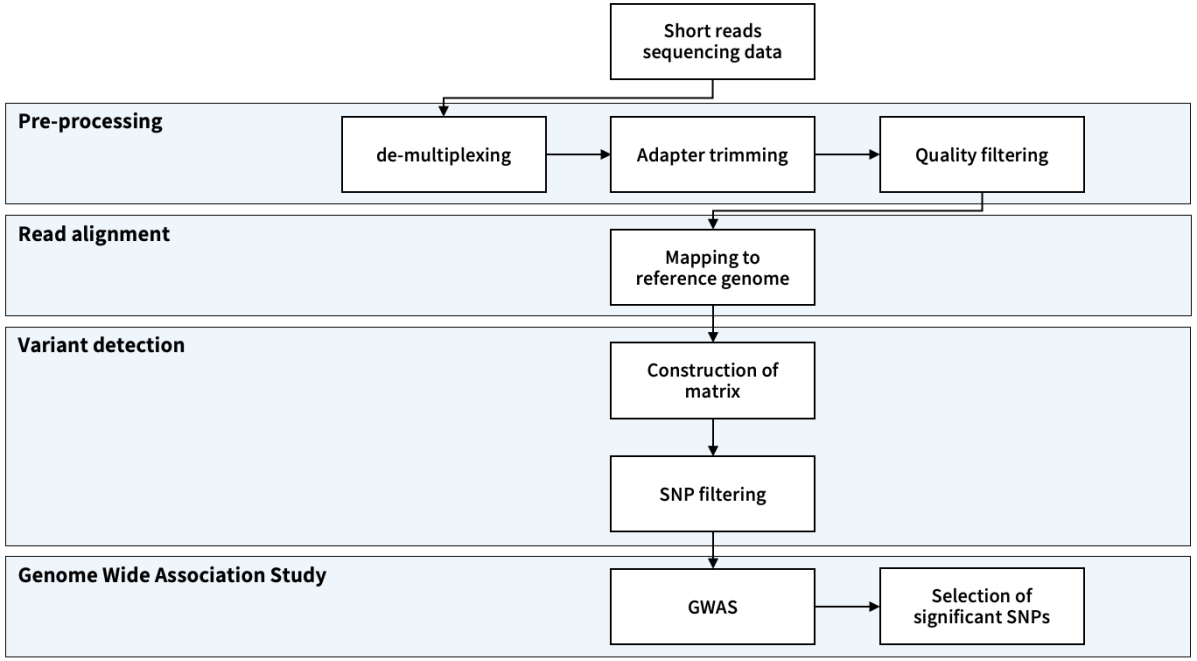
3. NGS 데이터로 하는 일

NGS 데이터 분석 파이프라인 (예시)

❖ Pipeline for QTL-mapping

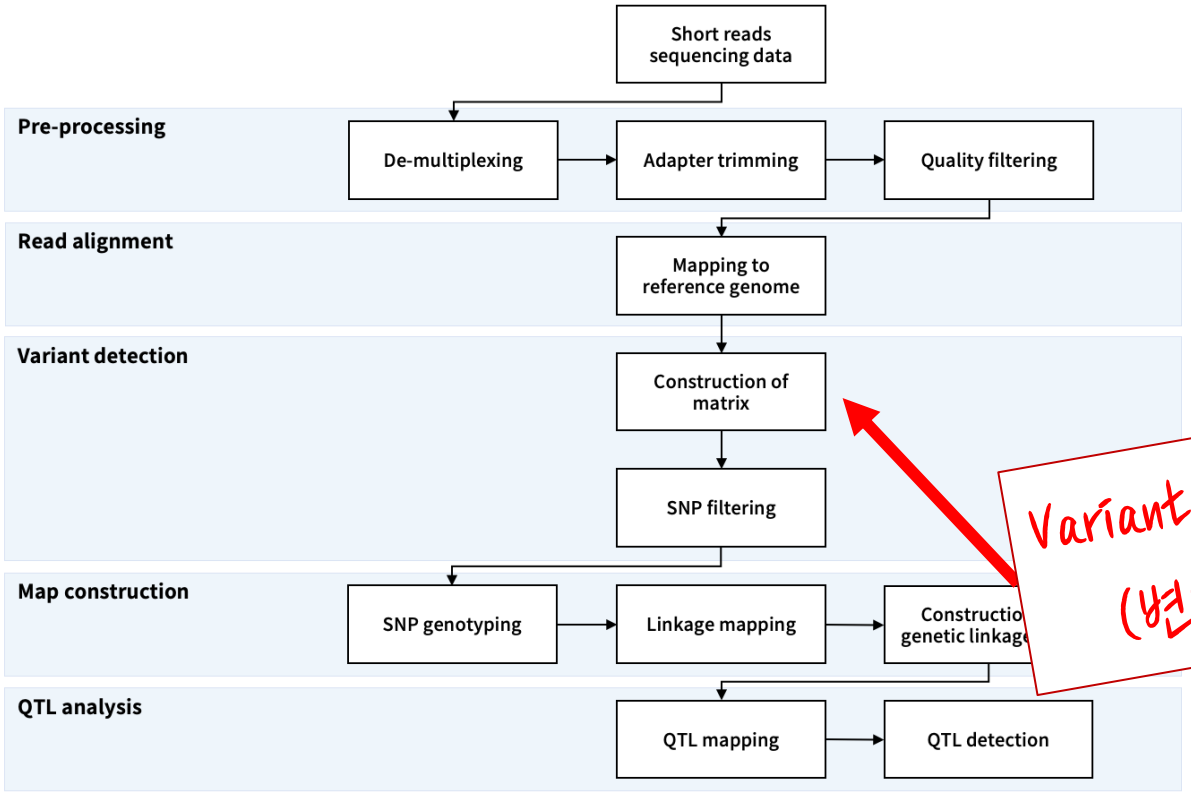


❖ Pipeline for GWAS

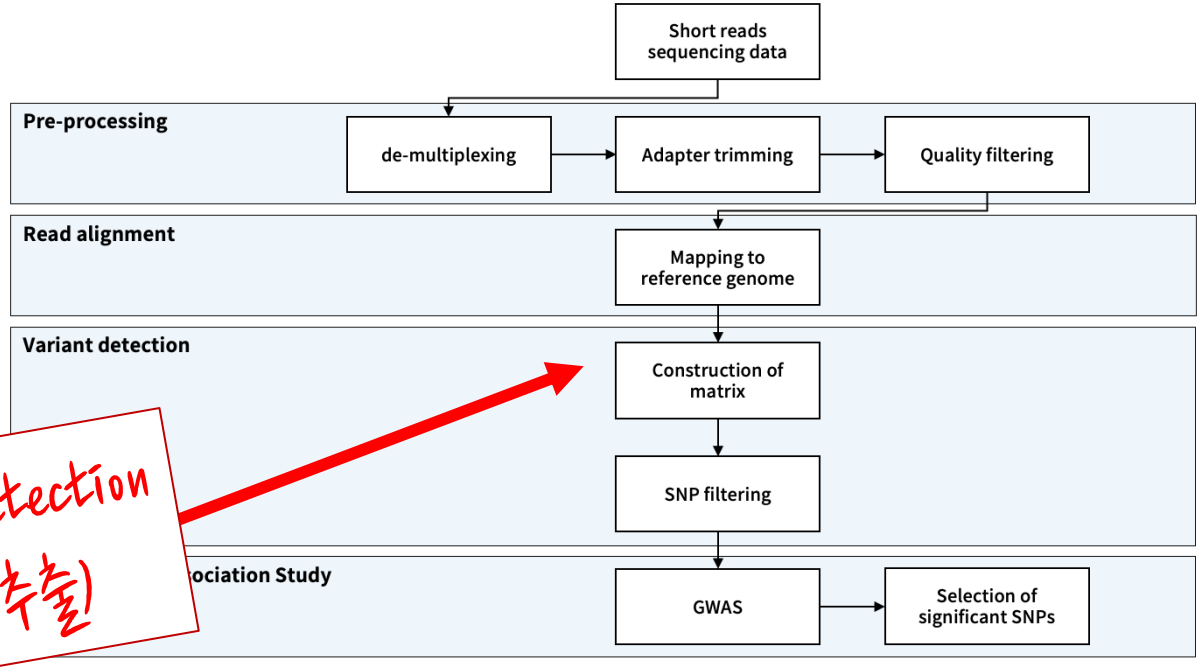


NGS 데이터 분석 파이프라인 (예시)

❖ Pipeline for QTL-mapping

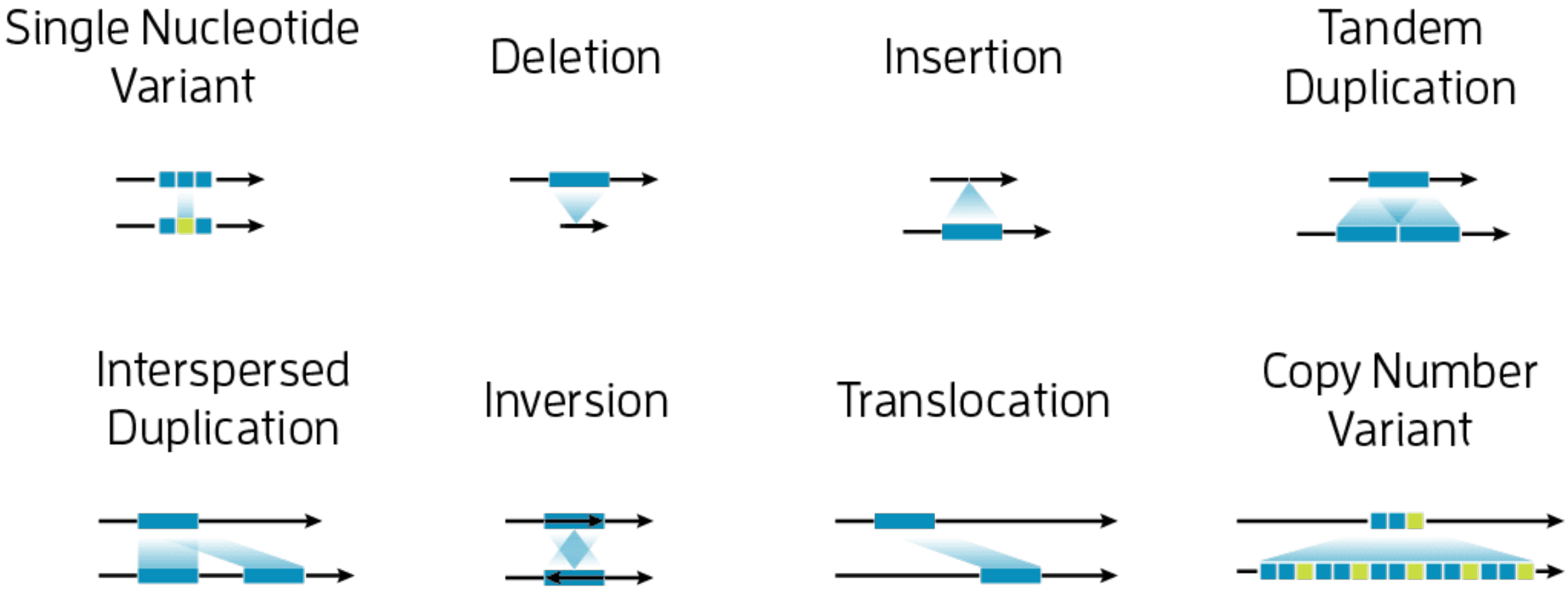


❖ Pipeline for GWAS



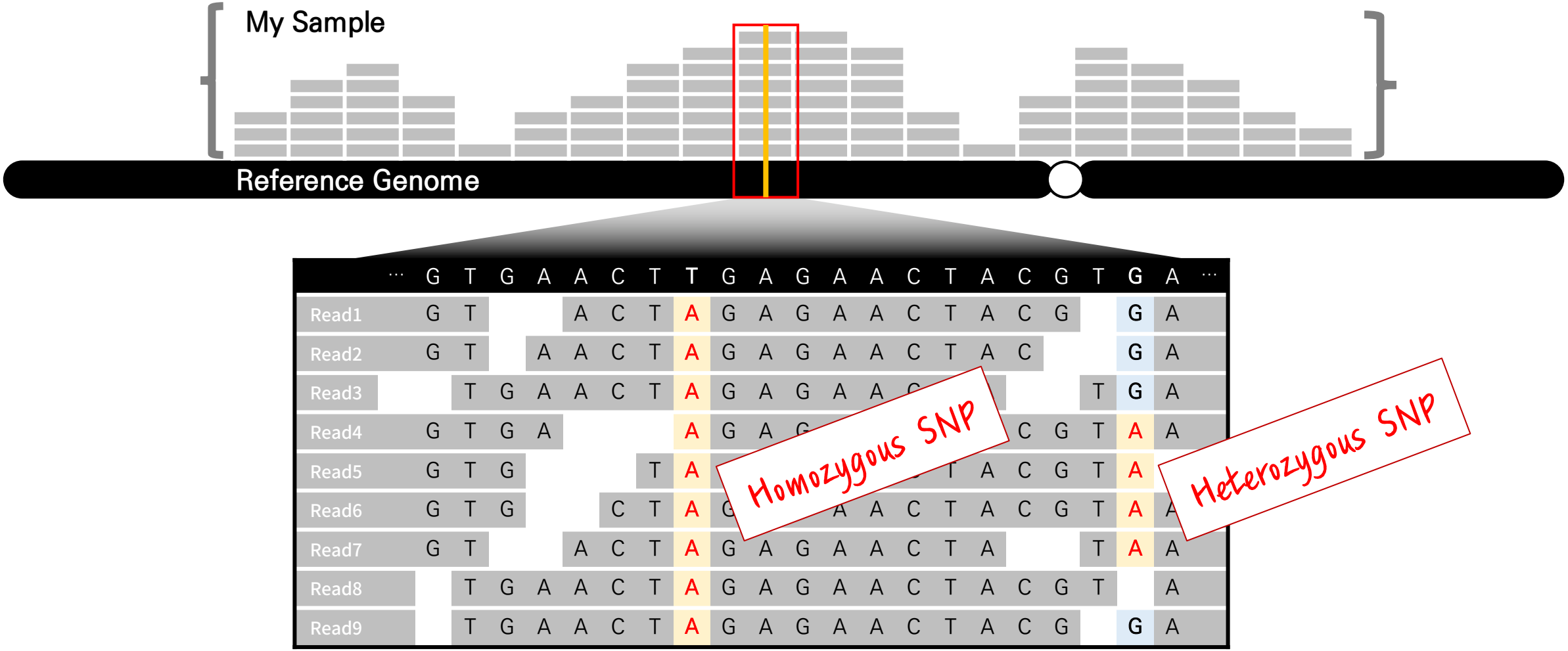
Variant Detection
(변이 추출)

유전체 상에 존재하는 다양한 변이(Variant)



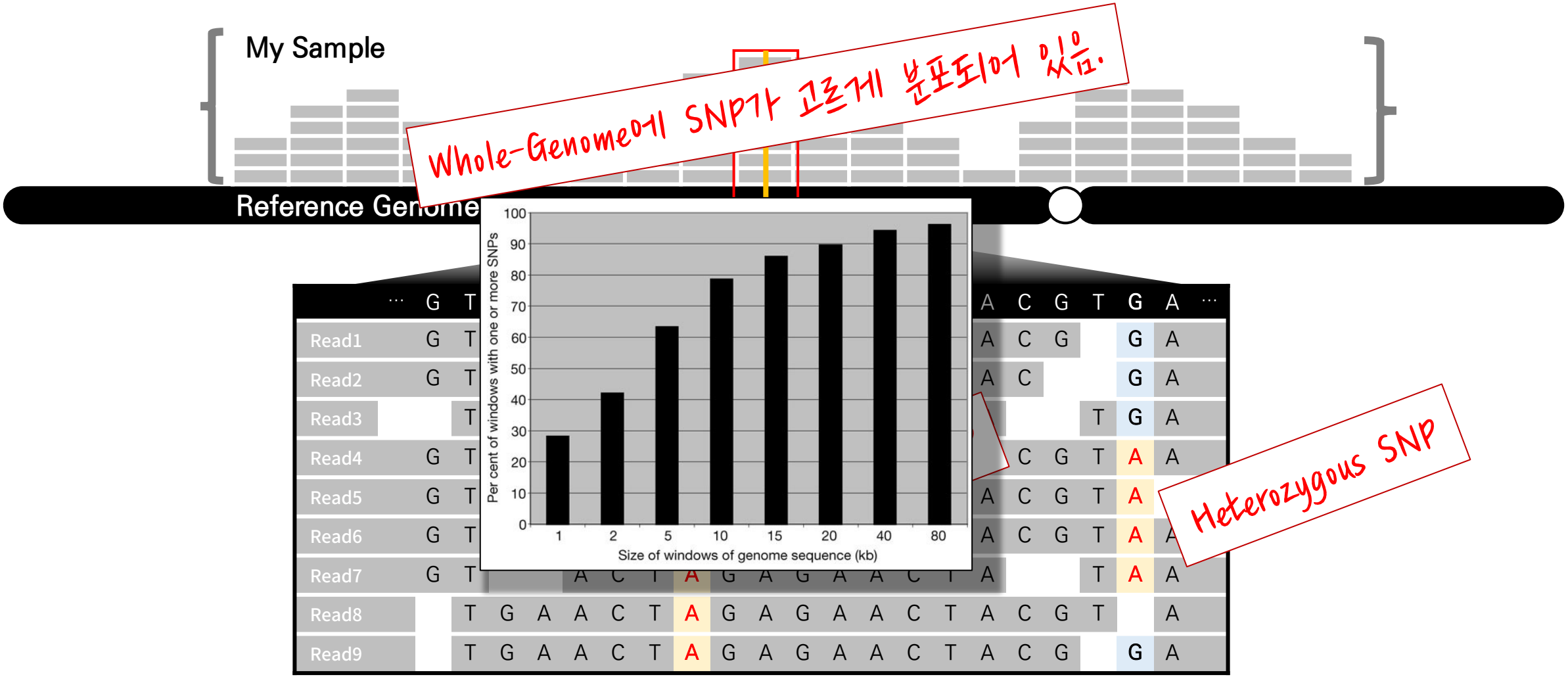
- https://www.pacb.com/?page_id=9004

SNP 변이



- Sachidanandam R et al. Nature. 2001;409(6822):928-933

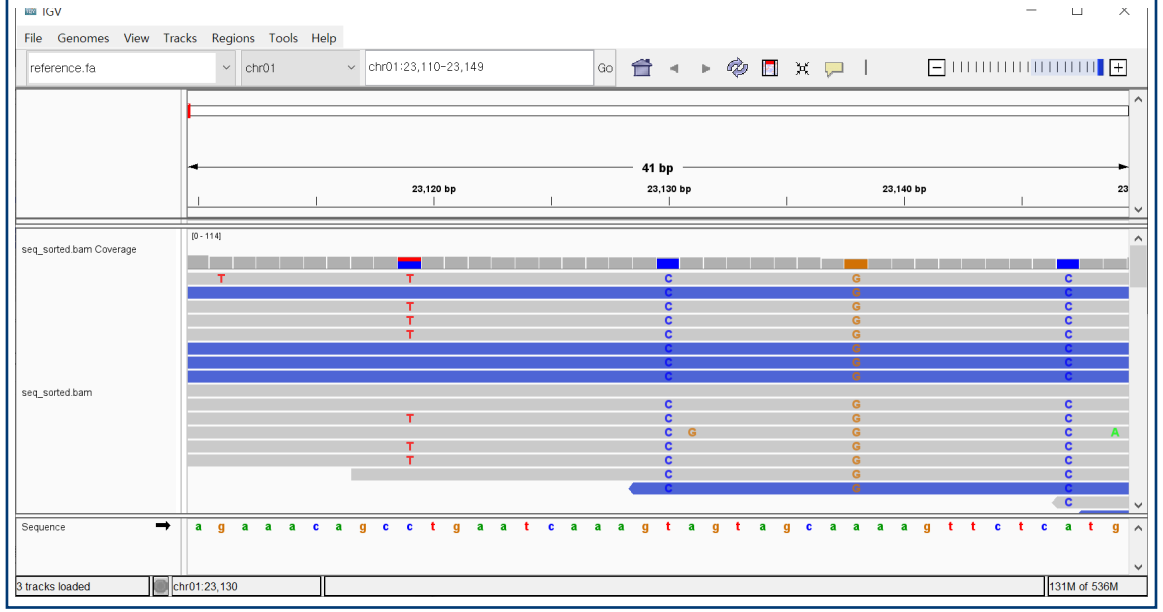
SNP 변이



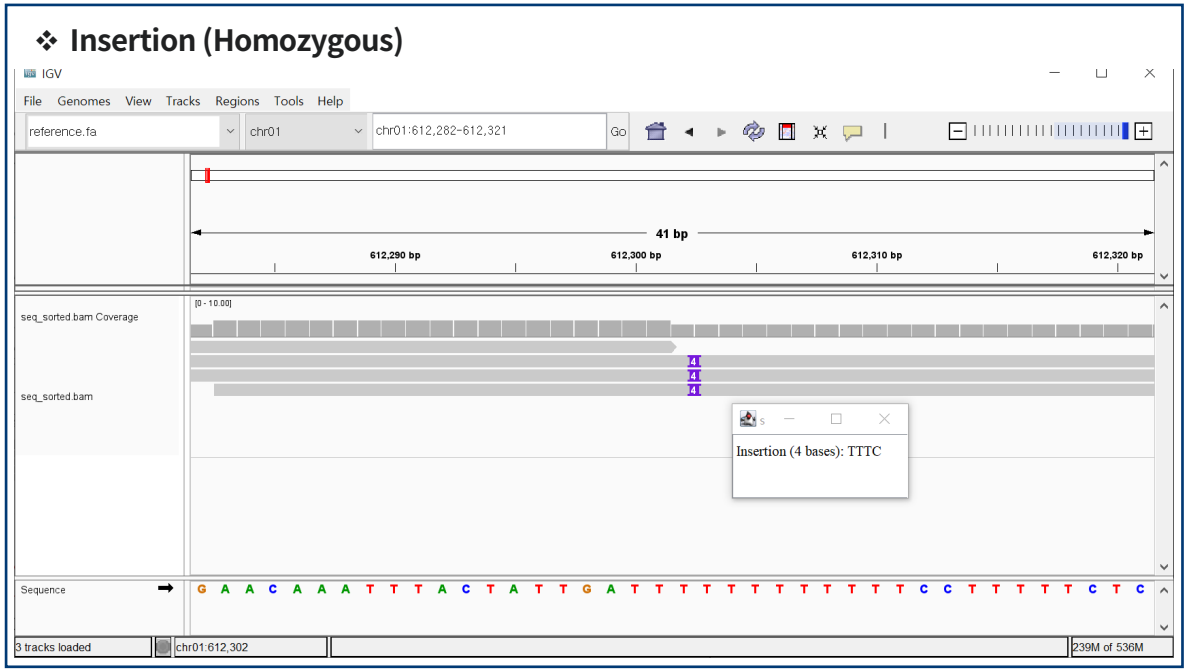
- Sachidanandam R et al. Nature. 2001;409(6822):928-933

IGV 프로그램을 통한 변이 관찰

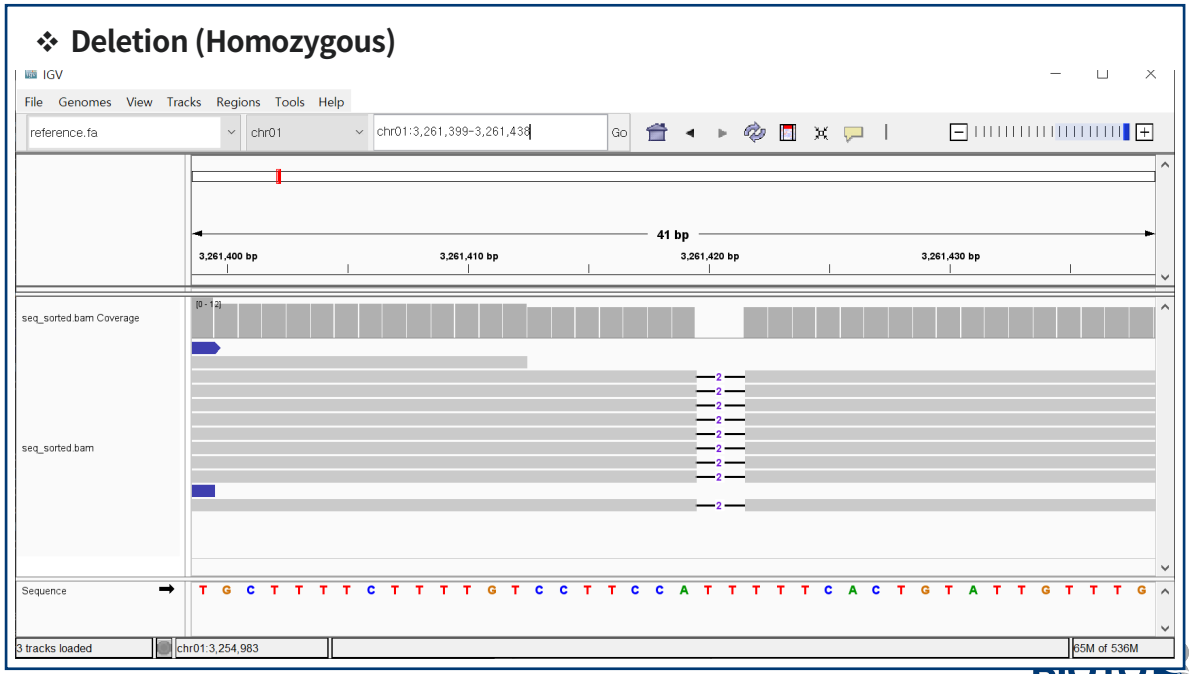
❖ SNP (Homozygous & Heterozygous)



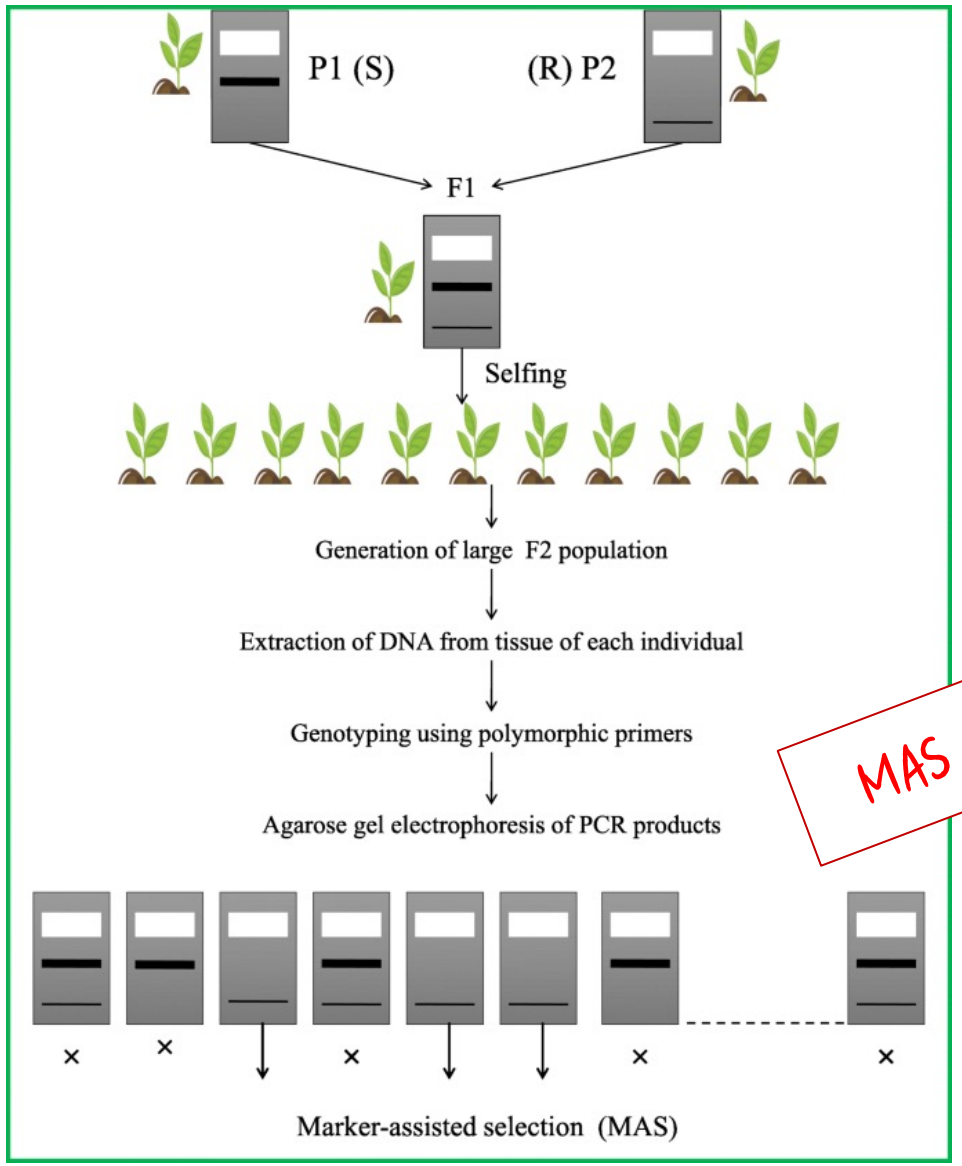
❖ Insertion (Homozygous)



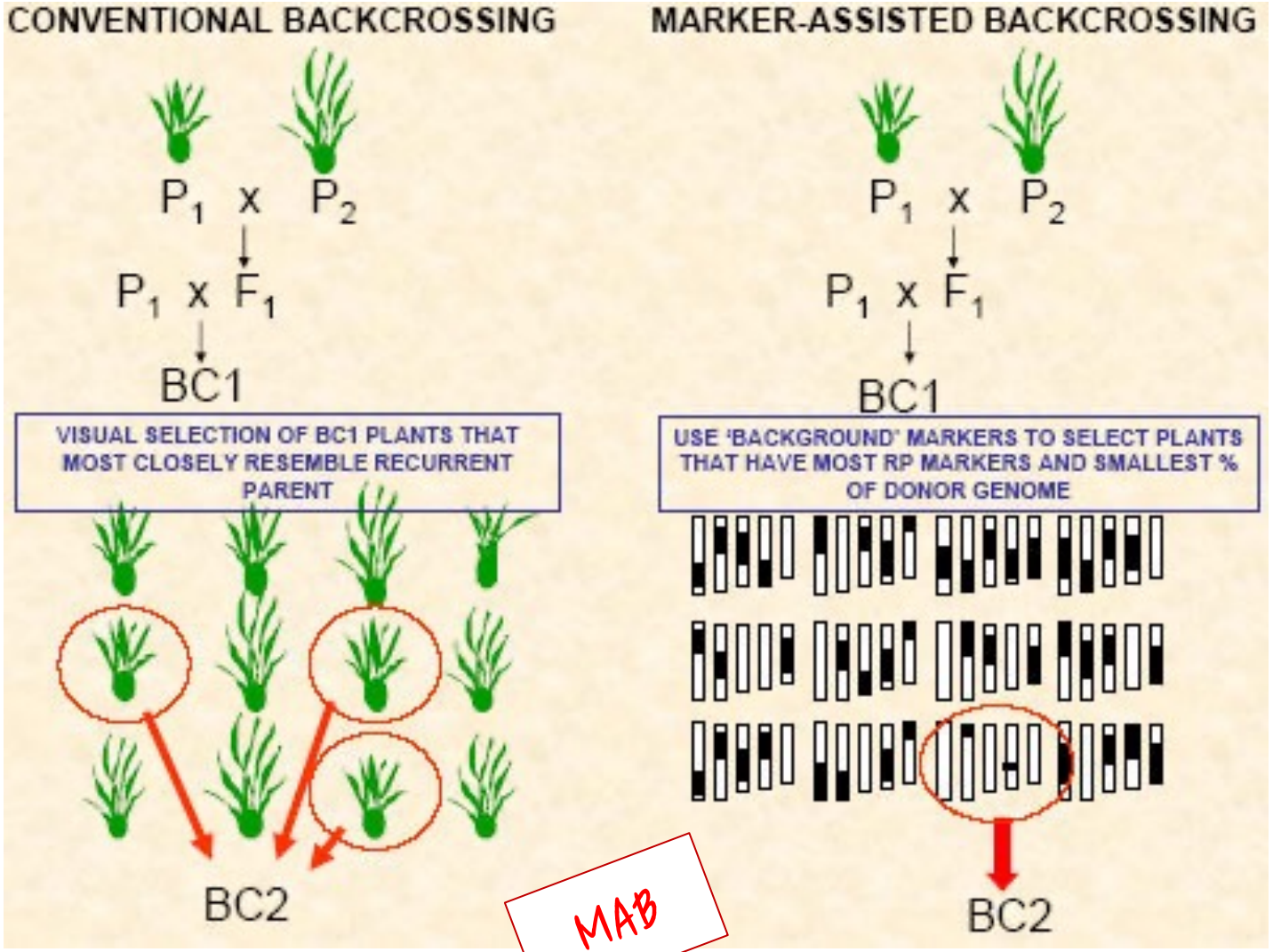
❖ Deletion (Homozygous)



NGS로 대량 형질연관 SNP 획득 → 분자육종



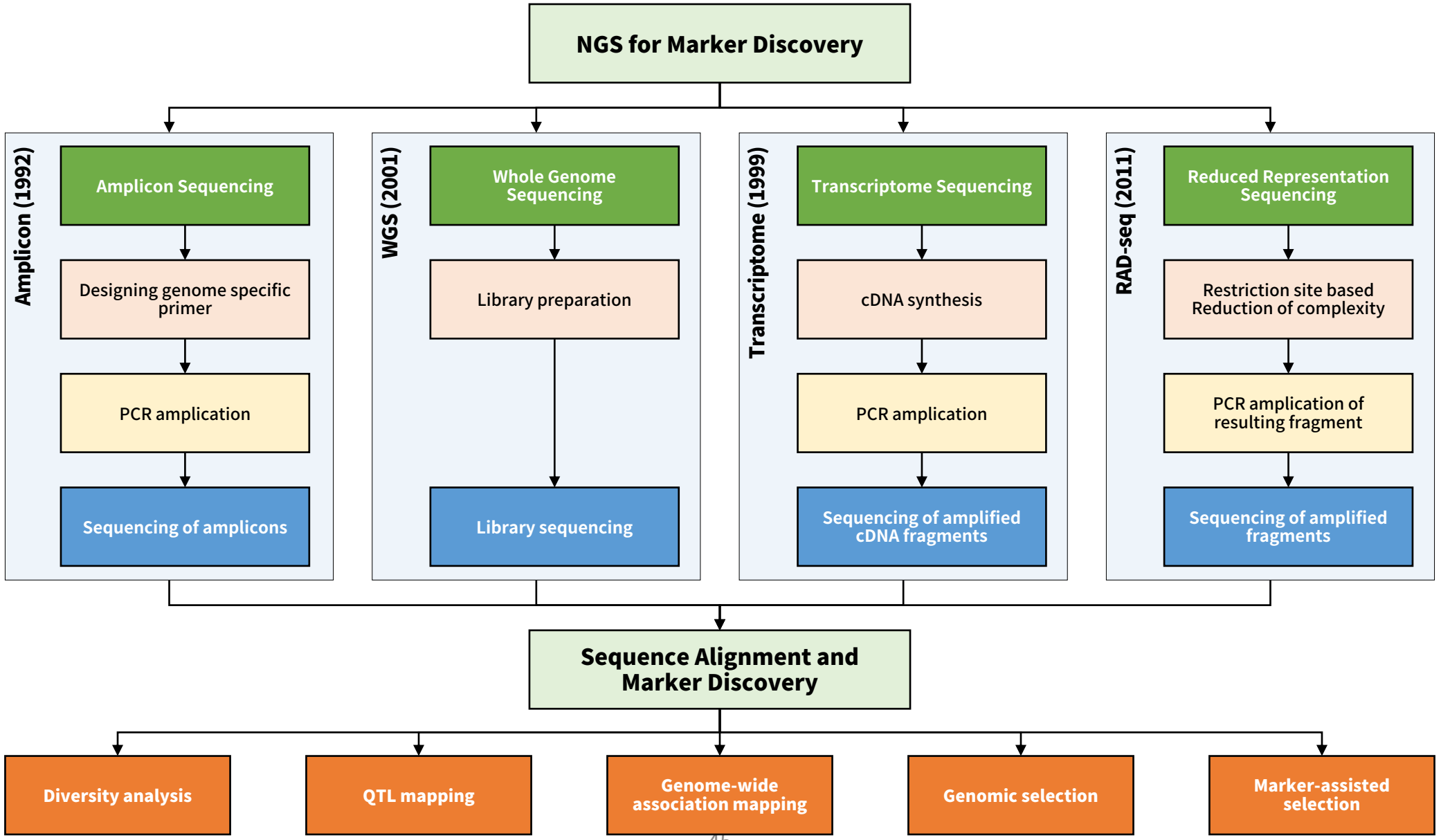
MAS



MAB

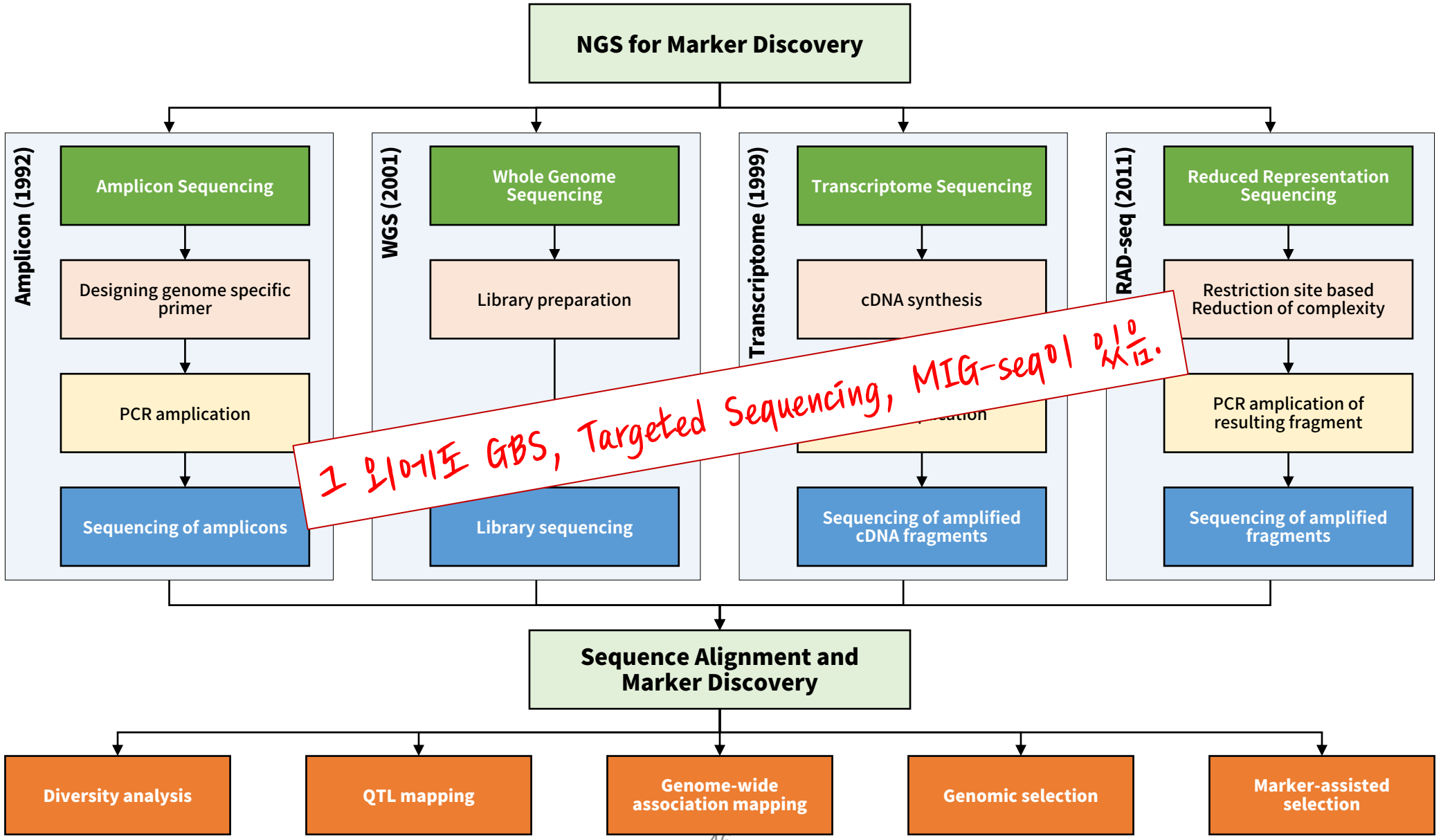
- https://jgeb.springeropen.com/articles/10.1186/s43141-021-00231-1
- http://www.knowledgebank.irri.org/ricebreedingcourse/Marker_assisted_breeding.htm

분자마커 발굴을 위한 다양한 NGS 기법



- https://www.researchgate.net/figure/Illustration-of-reduced-representation-sequencing-amplicon-sequencing-and-transcriptome_fig1_289685635 인용하여 수정

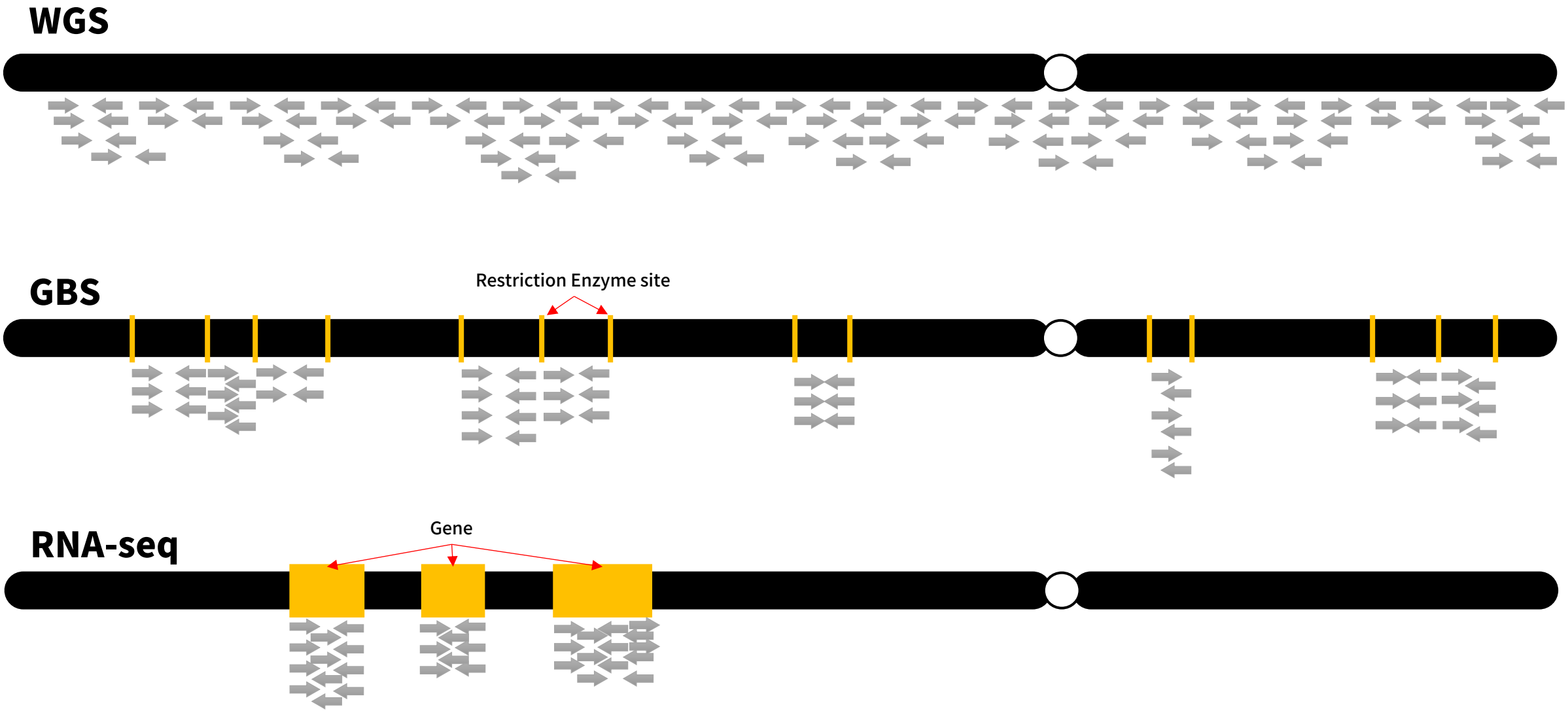
분자마커 발굴을 위한 다양한 NGS 기법



1 외에도 GBS, Targeted Sequencing, MIG-seq이 있음.

- https://www.researchgate.net/figure/Illustration-of-reduced-representation-sequencing-amplicon-sequencing-and-transcriptome_fig1_289685635 인용하여 수정

Sequencing 기법에 따른 차이 (1/2)

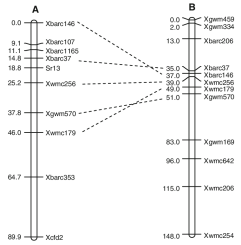


Sequencing 기법에 따른 차이 (2/2)

비교 사항	WGS	GBS	RNA-seq
분석 영역	유전체 모든 영역	Restriction Site 인근 영역	유전자 coding 영역
추천 시퀀싱 양	genome 기준 10X ~ 30X / 샘플	1Gbp / 샘플	2Gbp ~ 5Gbp / 샘플
유전체 크기	보통	적절	-
대량 샘플	샘플당 sequencing 양 증가	대량 샘플 OK	샘플당 sequencing 양 증가
변이 수	~ 수 십만 개	~ 수 천 개	~ 수 천 개
비용	80만원/샘플	16만원/샘플	80만원/샘플
발현 계산	X	X	O

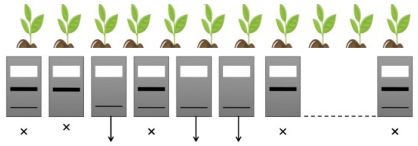
변이를 통해 할 수 있는 것

- https://www.researchgate.net/figure/Genetic-linkage-map-of-Sr13-compared-to-the-consensus-map-of-chromosome-6A-a-Genetic_fig1_279961814



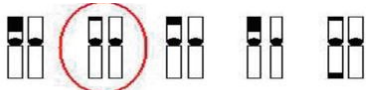
Linkage Map

- Hasan *et al.* J Genet Eng Biotechnol 19, 128 (2021).



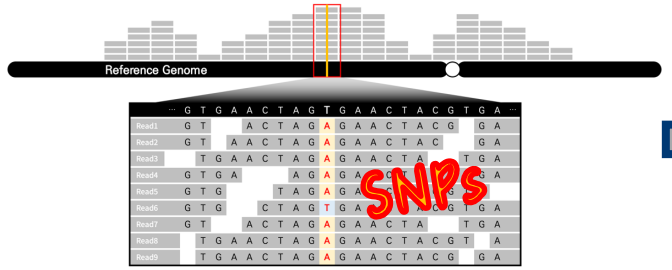
Marker-assisted selection (MAS)

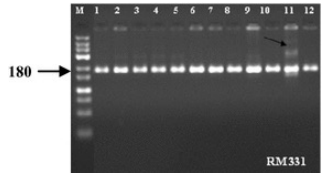
- https://www.intechopen.com/media/chapter/62375/media/F1.png



Marker-assisted backcrossing (MAB)

Next-Generation Sequencing

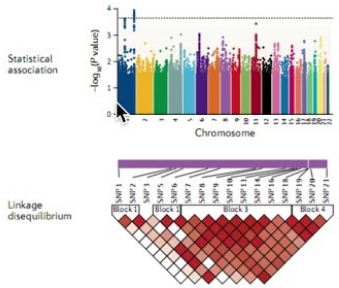




순도 검정 마커

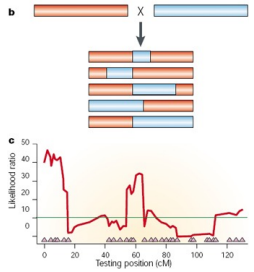
- Bora *et al.* Biotech 6, 50 (2016).

Association Mapping



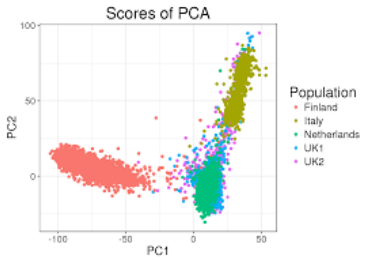
Tam et al 2019

QTL Mapping



- https://www.nature.com/scitable/content/33150/10.1038_35047544-f3_mid_1.jpg

원산지 구분 마커

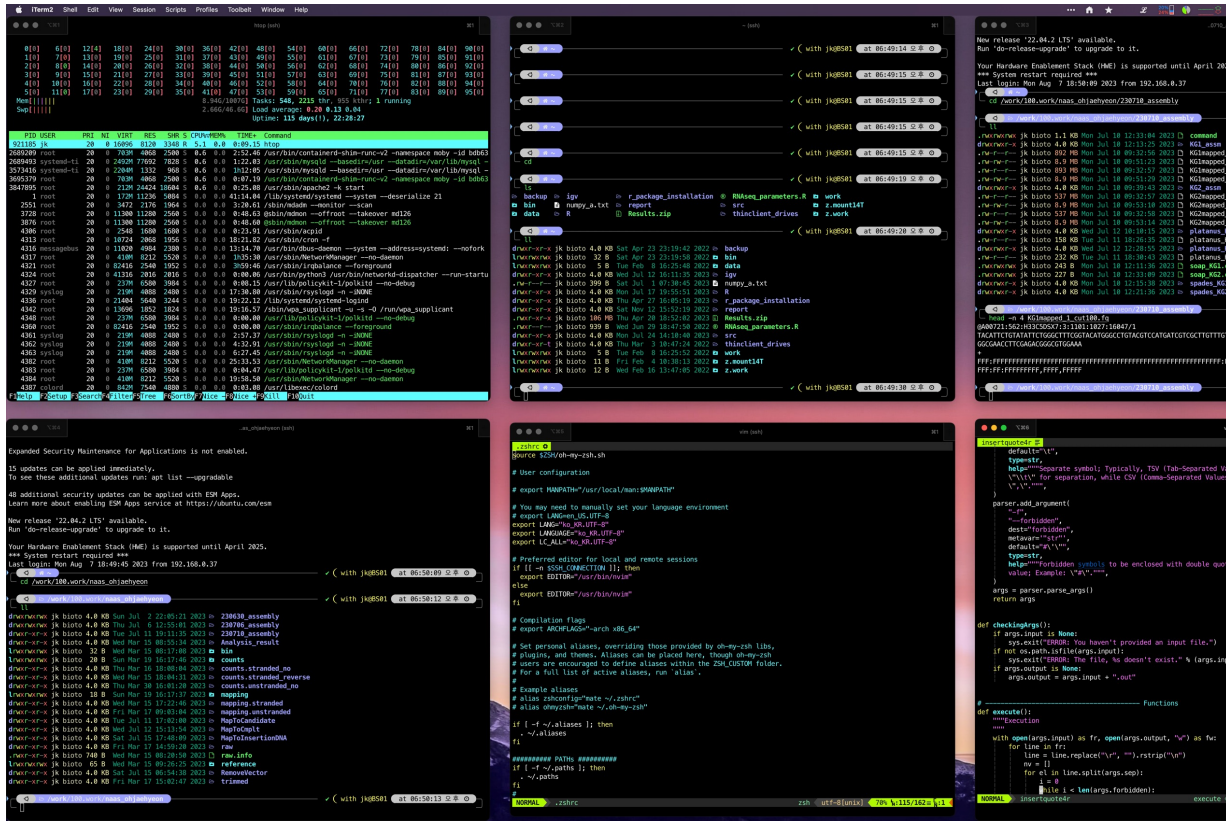


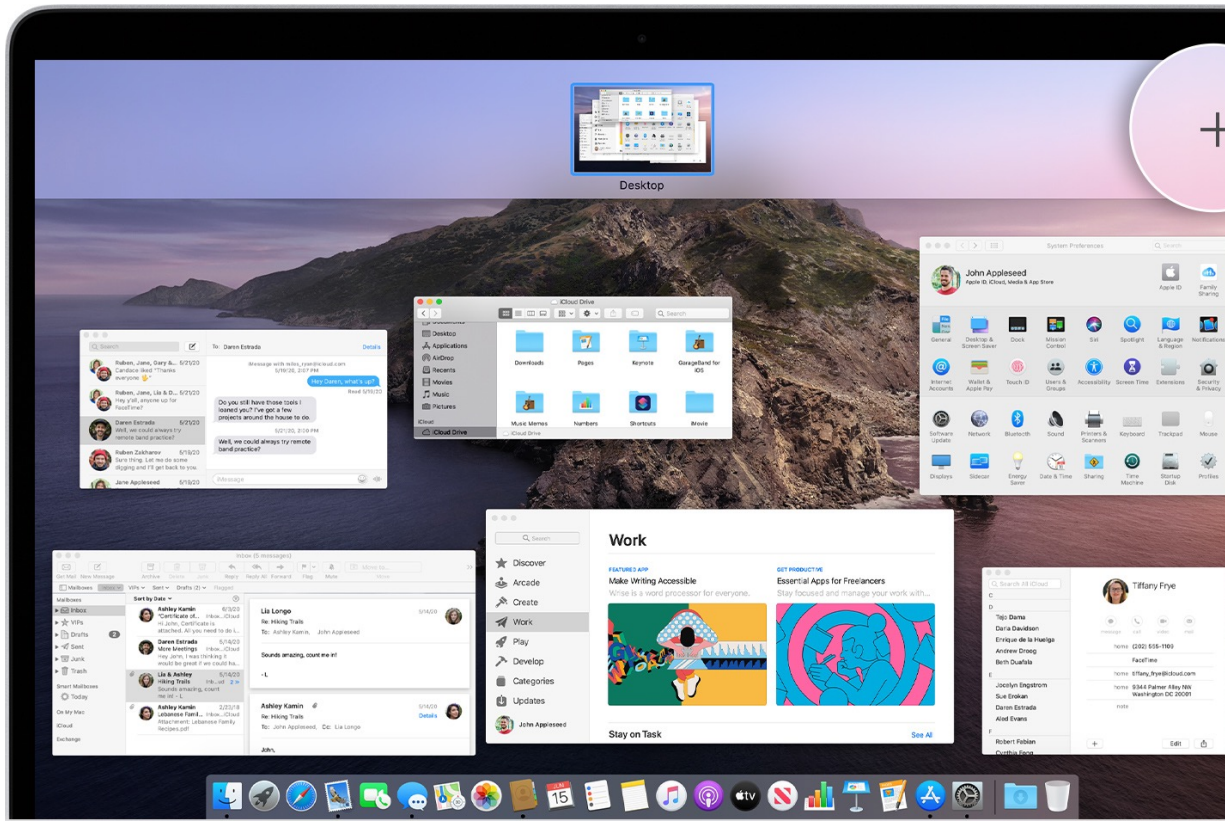
- https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

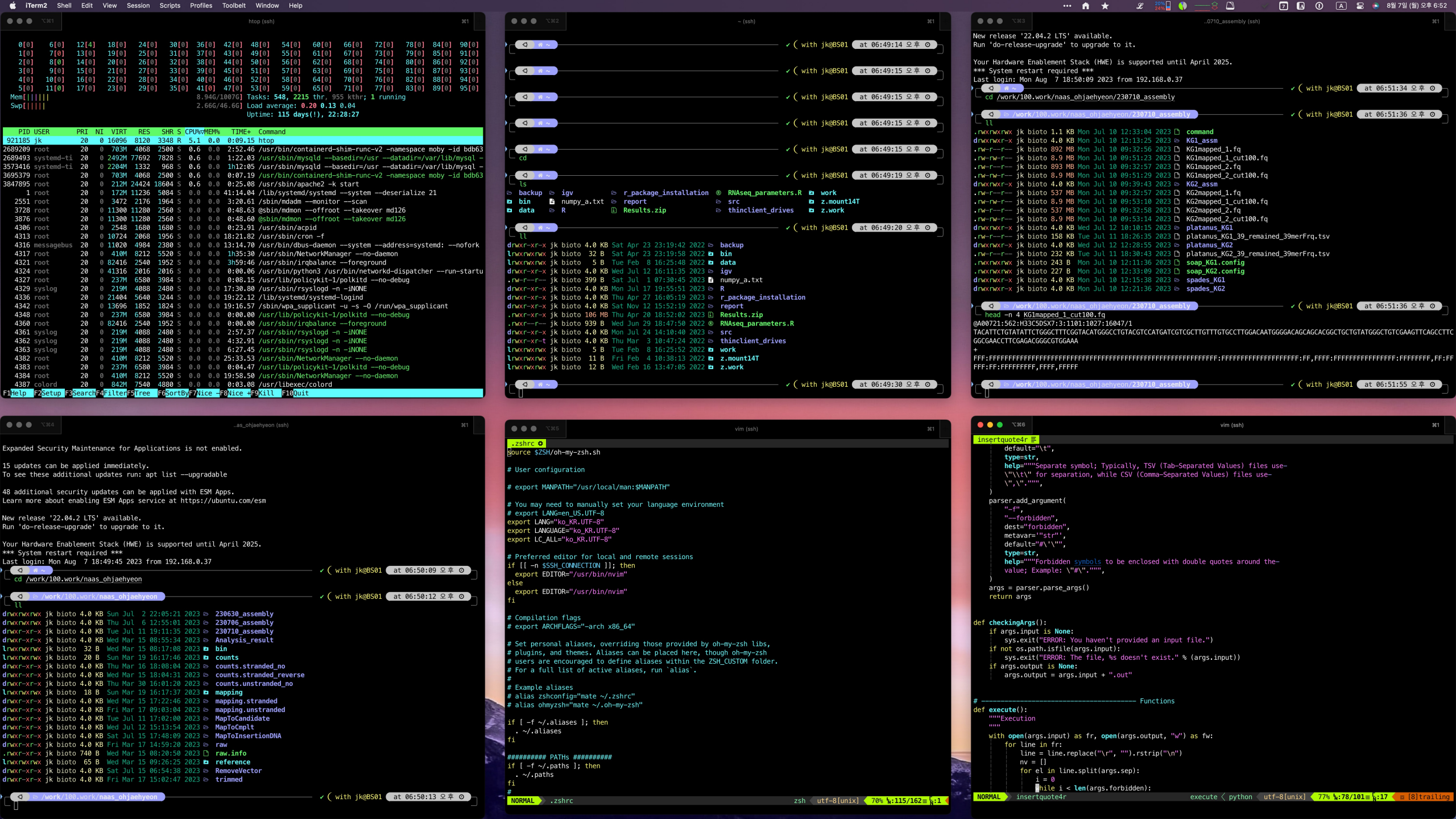
I. 생물정보학

4. NGS 데이터를 어떻게 분석하는가?

CLI vs. GUI




Desktop
Work
Make Writing Accessible
Essential Apps for Freelancers
Stay on Task


iTerm2 Shell Edit View Session Scripts Profiles Toolbelt Window Help
8월 7일 (월) 오후 6:52
htop (ssh)
Tasks: 548, 2215 thr, 955 kthr; 1 running
Load average: 0.20 0.13 0.04
Uptime: 115 days(!), 22:28:27
F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9Kill F10Quit
~ (ssh)
with jk@BS01 at 06:49:14 오후
with jk@BS01 at 06:49:15 오후
with jk@BS01 at 06:49:19 오후
with jk@BS01 at 06:49:20 오후
with jk@BS01 at 06:49:30 오후
..0710_assembly (ssh)
New release '22.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
Your Hardware Enablement Stack (HWE) is supported until April 2025.
*** System restart required ***
Last login: Mon Aug 7 18:50:09 2023 from 192.168.0.37
cd /work/100.work/naas_ohjaehyeon/230710_assembly
with jk@BS01 at 06:51:34 오후
with jk@BS01 at 06:51:36 오후
with jk@BS01 at 06:51:55 오후
..as_ohjaehyeon (ssh)
Expanded Security Maintenance for Applications is not enabled.
15 updates can be applied immediately.
To see these additional updates run: apt list --upgradable
48 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm
New release '22.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
Your Hardware Enablement Stack (HWE) is supported until April 2025.
*** System restart required ***
Last login: Mon Aug 7 18:49:45 2023 from 192.168.0.37
cd /work/100.work/naas_ohjaehyeon
with jk@BS01 at 06:50:09 오후
with jk@BS01 at 06:50:12 오후
with jk@BS01 at 06:50:13 오후
vim (ssh)
.zshrc
source $ZSH/oh-my-zsh.sh
NORMAL .zshrc zsh utf-8[unix] 70% ㏑:115/162≡ ㏇:1
insertquote4r
NORMAL insertquote4r execute 〈 python utf-8[unix] 77% ㏑:78/101≡ ㏇:17 ≡ [8]trailing

생물정보학자가 알아야 하는 것

데이터 분석 이론

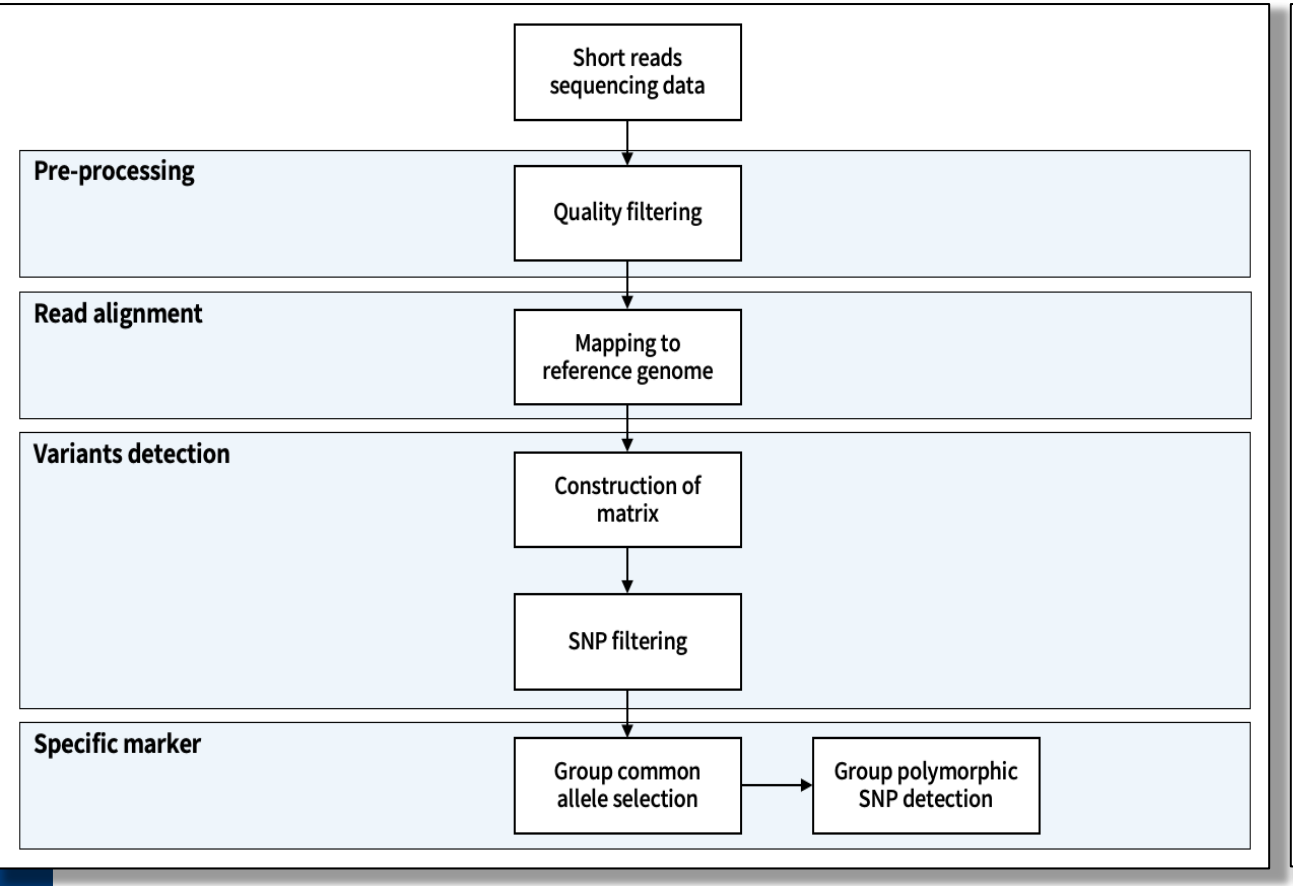
분석을 위한
Program 작동법

Linux 운영체제

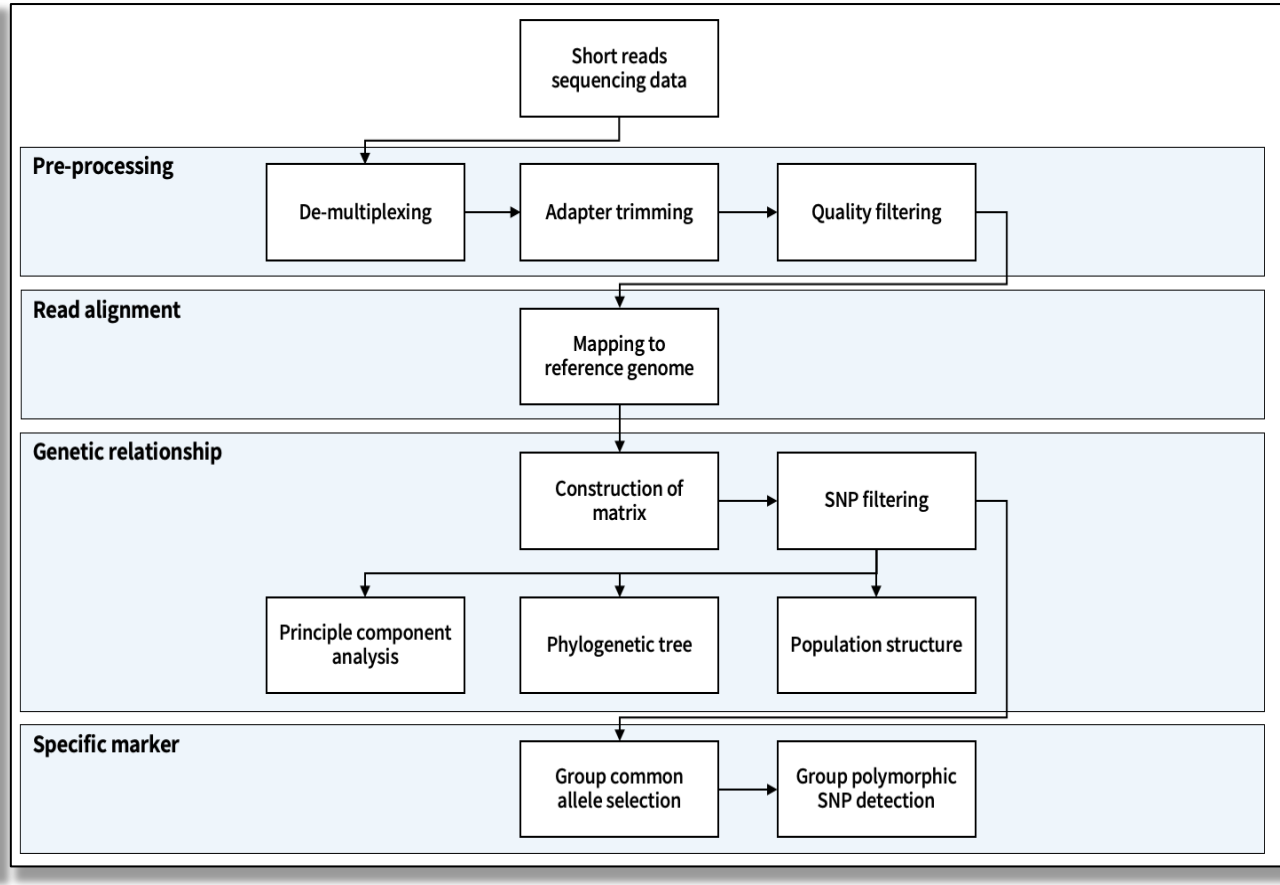
프로그래밍 언어
(Python, Java)

NGS 분석 파이프라인

❖ WGS (Whole Genome Sequencing) 파이프라인



❖ GBS (Genotyping-By-Sequencing) 파이프라인



Pre-processing

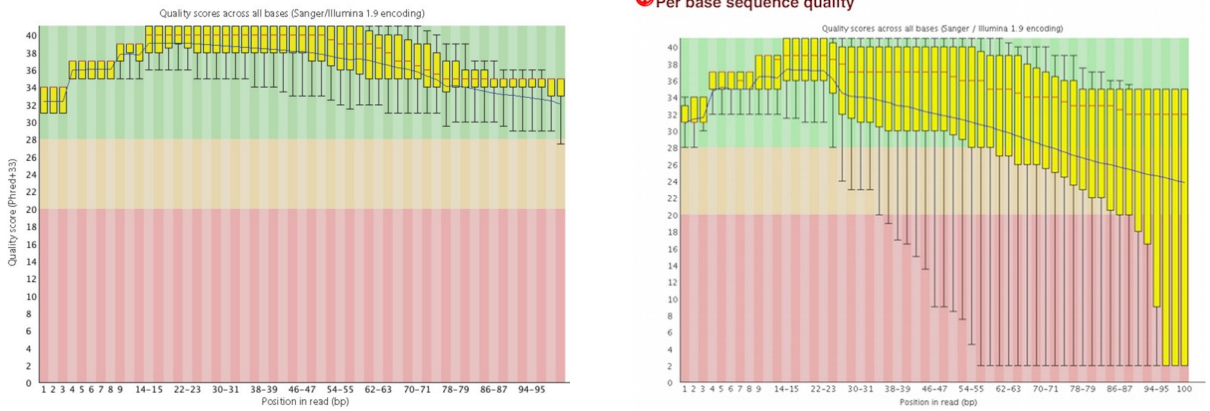
❑ 시퀀싱 데이터의 전처리 과정

- Removal of technical sequences
- Quality and length filtering

❑ 많이 사용되는 프로그램

- **Trimmomatic**
  A flexible trimmer for Illumina sequence data
- **FASTQC**
  A quality control tool for high throughput sequence data.
- **FASTX-Toolkit**
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- **SolexaQA**
  SolexaQA calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data.

❖ FASTQC



❖ Trimmomatic 결과 예시

```
edu_01@bc57f029632d:~/1.rawdata$ ls -l
total 5969596
-rw-r--r-- 1 root   root    509524430 May 26 21:51 seq_1.fq.gz
-rw-r--r-- 1 root   root    532221016 May 26 21:51 seq_2.fq.gz
-rw-r--r-- 1 edu_01 edu_01 2514467988 May 26 21:56 seq_paired1.fq
-rw-r--r-- 1 edu_01 edu_01   32562219 May 26 21:56 seq_paired1_un.fq
-rw-r--r-- 1 edu_01 edu_01 2513757814 May 26 21:56 seq_paired2.fq
-rw-r--r-- 1 edu_01 edu_01   10306054 May 26 21:56 seq_paired2_un.fq
edu_01@bc57f029632d:~/1.rawdata$
```

- https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html
- http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

Trimmomatic 실행

❑ Trimmomatic 프로그램 옵션

- Phred33
- Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
- Remove leading low quality or N bases (LEADING:3)
- Remove trailing low quality or N bases (TRAILING:3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

Trimmomatic 수행

java -jar /home/Trimmomatic-0.39/trimmomatic-0.39.jar PE -threads 10 -phred33 seq_1.fq.gz seq_2.fq.gz seq_paired1.fq seq_paired1_un.fq seq_paired2.fq seq_paired2_un.fq ILLUMINACLIP:/home/Trimmomatic-0.39/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Read Alignment

❑ Read Alignment / Read Mapping 과정

- Read alignment (mapping)는 sequencing reads들을 표준유전체 서열과 비교하여 reads의 염기서열과 일치하는 위치를 표준유전체 서열에서 찾는 과정

❑ 많이 사용되는 프로그램

- **BWA (Burrows-Wheeler Aligner)**
  BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.
- **Bowtie2**
  Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.
- **HISAT2**
  The HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome.
- **RUM, STAR, TopHat2, …**

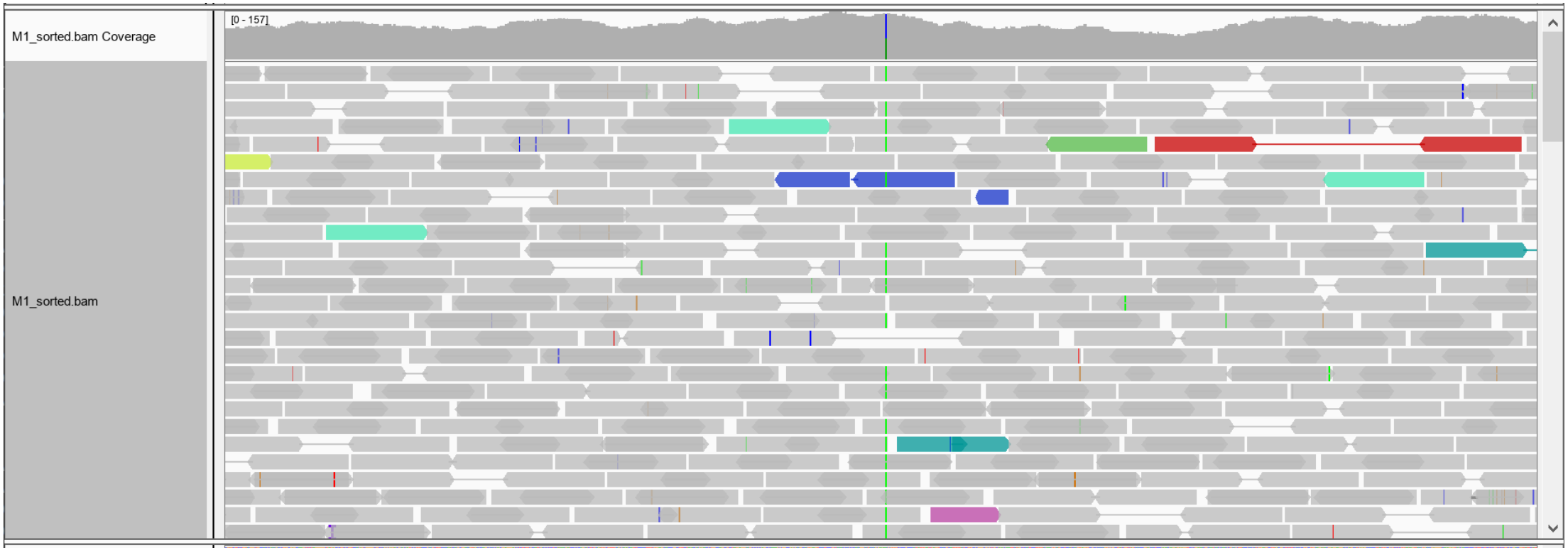
Alignment Tools 간의 비교

Tool	Alignment Reference	Description
Bowtie2	Transcriptome	Bowtie2 aligns reads by combining full-text minute index and hardware-accelerated dynamic programming to produce sensitive and accurate alignments (Langmead and Salzberg, 2012).
Bwa	Genome	Bwa aligns short DNA sequences against a reference genome by constructing a suffix array and applying Burrows-Wheeler transformation that matches the sequences using a backward search (Li et al., 2013).
HiSat2	Genome	HISAT aligns reads using an indexing scheme based on Burrows-Wheeler transform and the Ferragina-Mangini index (Kim et al., 2015).
RUM	Genome + Transcriptome	RUM is an alignment and feature quantification pipeline developed specifically for Illumina RNAseq data. RUM uses Bowtie algorithm for alignment (Grant et al., 2015).
STAR	Genome or Transcriptome	STAR aligns raw reads by using a seed - extension search based on uncompressed suffix arrays and detects splice junctions.
TopHat2	Genome	TopHat2 has the ability to identify novel splice sites and mapping directly to known transcripts that produces sensitive and accurate alignments (Kim et al., 2013).

- https://www.elucidata.io/blog/bulk-rna-sequencing-a-comparison-of-the-most-popular-tools-and-pipelines

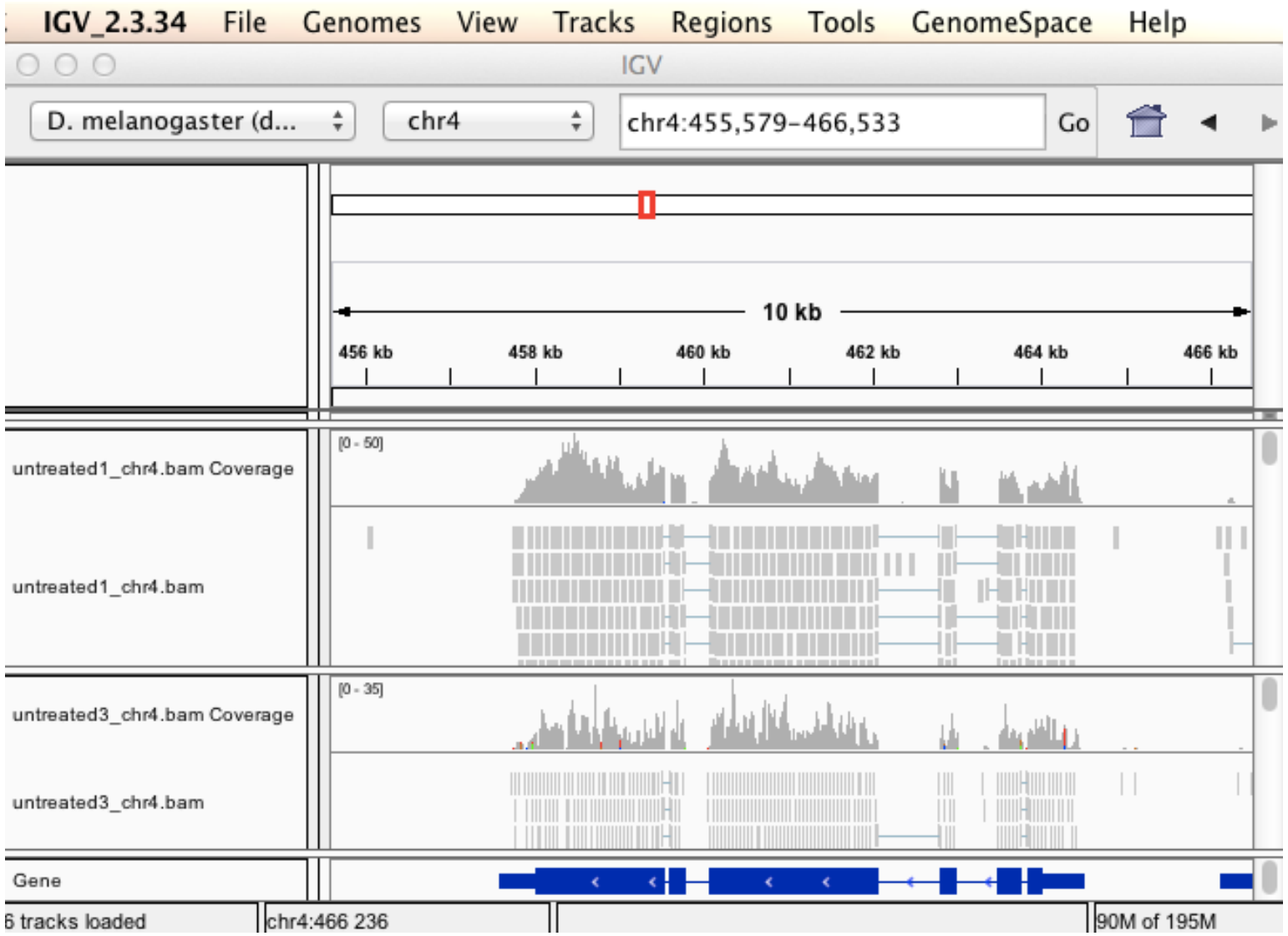
Read Alignment 결과 예시

Read alignment – IGV 결과 예시 (Genome)



Read Alignment 결과 예시

Read alignment – IGV 결과 예시 (RNA)



Variant Detection

❑ Variant Detection 과정

- Read alignment (mapping) 산물을 이용하여 시퀀싱 샘플과 표준유전체 서열과의 차이 (SNP, In/Del 등)를 찾는 과정

❑ 많이 사용되는 프로그램

- **SAMTools**

  Provides various utilities for manipulating alignments in the SAM/BAM format.

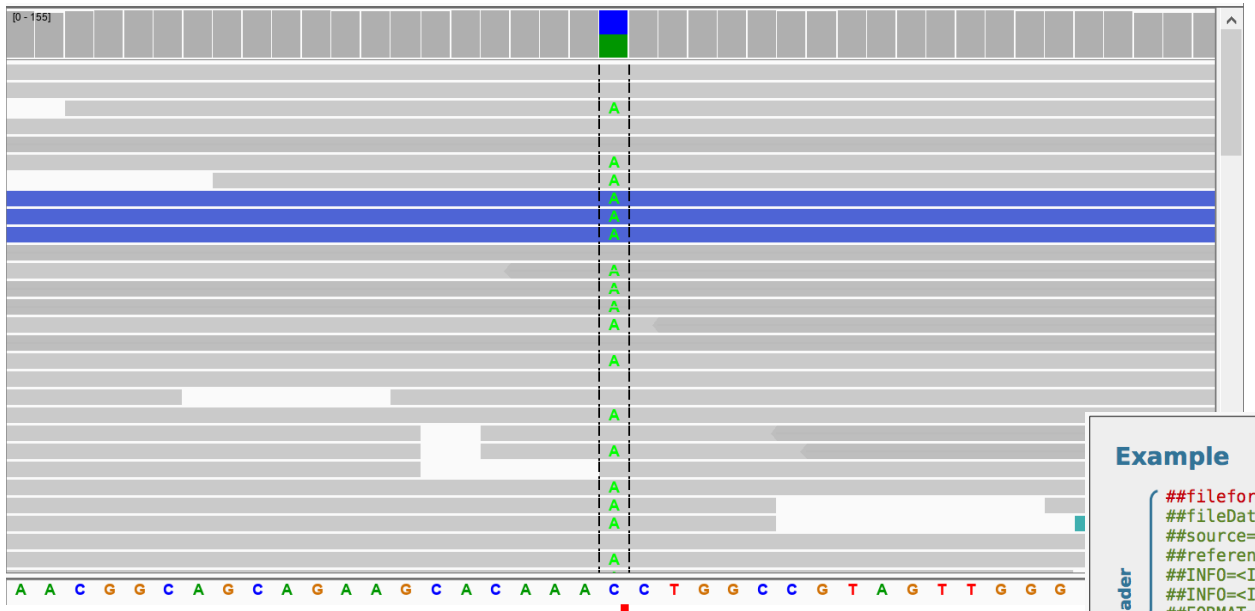
  Find variants

- **GATK**

  A genomic analysis toolkit focused on variant discovery

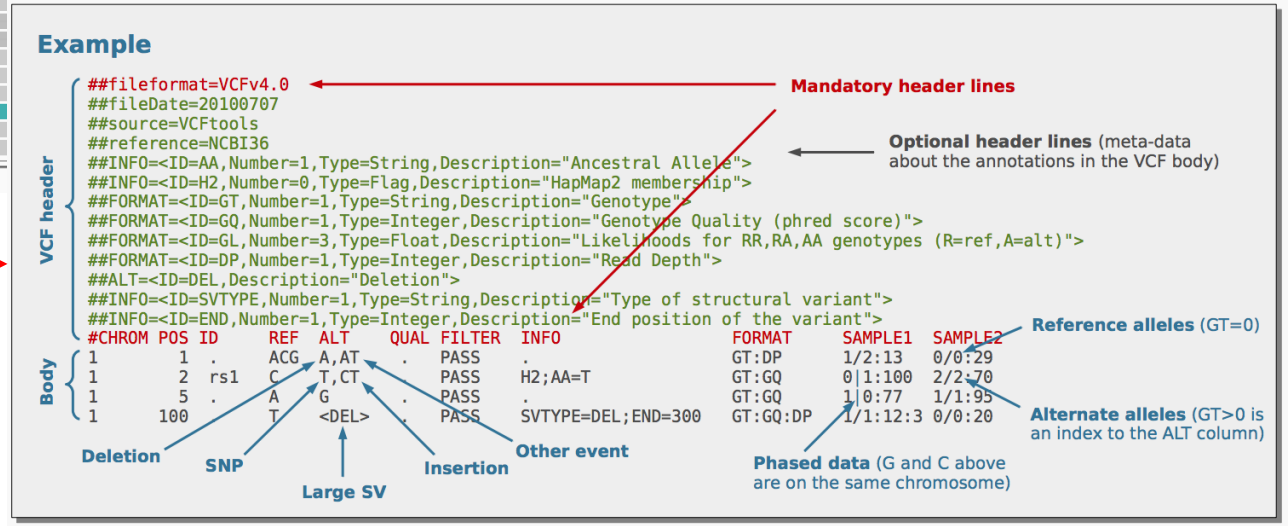
Variant Detection 결과 예시

Variant Detection - 결과



IGV

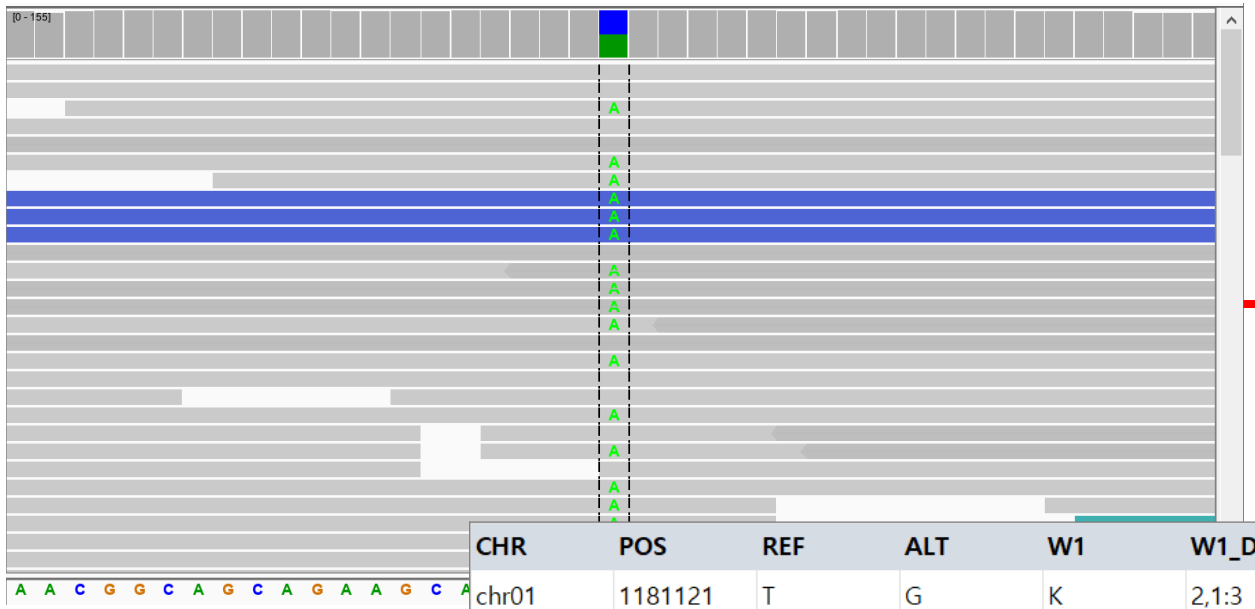
VCF (Variant Call Format)



- https://davetang.github.io/learning_vcf_file/

Variant Detection 결과 예시

Variant Detection - 결과

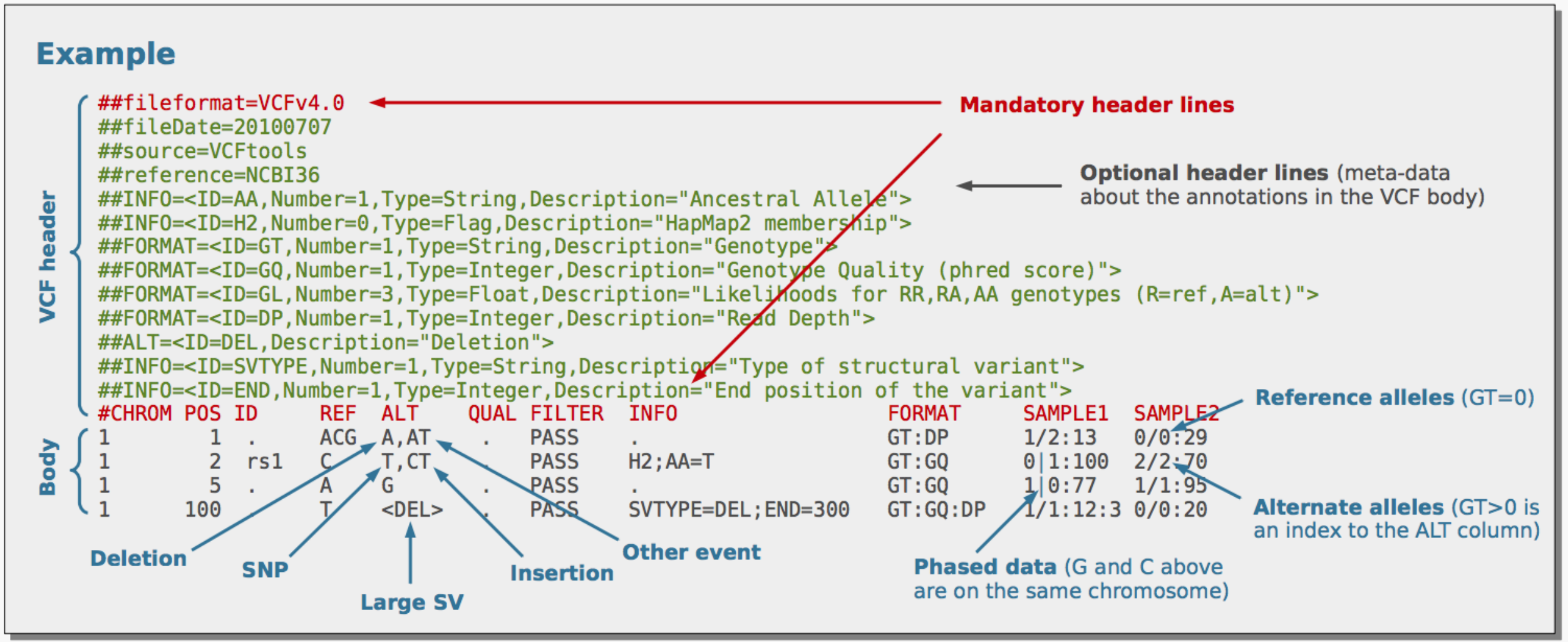


IGV

CHR	POS	REF	ALT	W1	W1_Depth	M1	M1_Depth	Genic/Inter	Feature	Description	Organism	Flanking_600bp
chr01	1181121	T	G	K	2,1:3	T	3,0:3	SL000879t0(	Intron	Glycogen ph	Auxenochlor	CTTTCTGCACCG(
chr01	1272122	T	G	K	1,2:3	T	9,1:10	SL014053t0(	Intron	HslV compo	Coccomyxa	ACAGCTGGTCTG
chr01	2586904	T	G	K	4,2:6	T	3,0:3	SL012474t0(	Intron	exocyst com	Prunus persi	TGCCGCCGGCG(
chr01	3334582	C	A	M	11,15:26	C	21,0:21	SL004799t0(	Intron	hypothetical	Chlorella var	CTCCGTACCCCC
chr01	4136244	C	A	C	66,0:66	M	73,65:138	Intergenic				CATCTCGCAGCT(
chr01	128	C	A	C	29,0:29	M	35,22:57	Intergenic				aaacCCTAAACCC
chr01	1041661	A	C	M	2,2:4	A	4,0:4	SL002841t0(	Intron	Thermostabl	Auxenochlor	TTTGTAGCTTTG
chr01	324	A	C	M	33,39:73	A	100,0:100	Intergenic				CTAAACCCTAAA(

SNP matrix

VCF (Variant Call Format) 결과 형태



- https://davetang.github.io/learning_vcf_file/

VCF와 BCF의 차이

VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:DP:GQ   0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:DP:GQ   0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:DP:GQ   0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:DP:GQ   1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

BCF

```
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0   DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0   DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0   DP:5:4:5:9:4:4:4        GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:1/0:0/0:0/0:0/0:0/0:0/1:0/1   DP:5:9:10:15:9:13:14    GQ:15:27:30:39:27:39:42
```

https://en.wikipedia.org/wiki/Variant_Call_Format

Variant Filtration

❑ Variant Filtration 과정에 사용되는 기준

- SNPs Low quality
- Number of alleles: It is possible to filter out the non-biallelic or the monomorphic SNPs.
- High Coverage: It can be false positives due to repetitive regions.
- Missing genotypes
- Minor Allele Frequency (MAF)
- Observed Heterozygosity
- By genome localization: exon, UTR, etc,.
- Amino-acid change: We can select the SNPs with large impacts in the coded proteins.
- Linkage Disequilibrium: If we have genotype a segregant population it could be useful to filter out the SNPs that are not in linkage disequilibrium with their closest SNPs.

I. 생물정보학

5. Marker Filtration (필요한 마커를 골라내는 방법)

Filtering 예시

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0075	870356	G	C	G	-	C	C	C	C	-
scaffold0058	1663836	T	C	T	C	C	C	C	T	-
scaffold0107	601267	G	A	G	G	A	-	A	A	-
scaffold0108	1303787	G	A	G	A	A	A	A	A	-
scaffold0171	50985	A	G	G	A	-	A	A	A	-
scaffold0171	761366	G	T	T	G	G	-	T	G	G
scaffold0171	1459133	T	C	-	T	T	C	C	-	C
scaffold0175	1548310	G	A	-	G	A	G	A	A	G
scaffold0117	271696	G	A	-	A	A	A	A	A	A
scaffold0147	258651	A	C	A	A	C	-	A	C	-
scaffold0144	11553	C	T	T	C	-	C	C	C	-
scaffold0160	1301533	G	A	A	A	G	G	G	G	A
scaffold0190	526878	C	T	-	T	-	-	C	C	T
scaffold0191	652149	G	A	G	A	-	G	A	-	G
scaffold0708	1251377	G	C	G	-	G	C	G	G	G
scaffold0711	73459	T	C	C	C	T	T	-	C	C
scaffold0711	151410	A	G	-	G	A	G	A	A	G
scaffold0711	151411	A	G	G	A	-	A	G	A	G
scaffold0711	153975	C	T	T	T	C	-	C	C	C
scaffold0711	321710	A	G	-	-	G	-	-	-	-
scaffold0711	520444	C	T	C	T	T	T	-	-	-
scaffold0711	585045	T	G	G	T	G	G	T	G	T
scaffold0777	175224	T	G	T	T	T	T	G	T	G
scaffold0717	154744	T	G	T	-	G	-	G	T	-
scaffold0718	586757	G	A	G	G	G	A	A	A	-
scaffold0719	711767	T	G	G	-	T	T	-	T	G
scaffold0746	9828	A	G	A	A	A	G	-	A	G
scaffold0747	1208530	G	A	-	A	G	-	A	G	-
scaffold0758	1215157	A	T	A	-	A	T	-	A	-
scaffold0776	85221	C	A	C	A	A	A	C	A	A
scaffold0776	161264	T	C	C	C	-	C	-	T	C
scaffold0788	601667	G	A	-	G	A	G	G	A	-
scaffold0797	888578	T	C	-	-	C	C	C	-	T
scaffold0118	542035	G	A	-	-	G	A	A	-	A
scaffold0170	847416	G	A	-	A	-	A	A	A	A
scaffold0141	587532	A	G	A	A	A	A	-	-	A
scaffold0185	731688	G	A	A	A	-	-	A	G	A
scaffold0185	731747	G	A	-	-	G	A	-	G	A
scaffold0199	702837	G	A	A	G	G	-	G	A	-

Filtering 예시

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0075	870356	G	C	G	-	C	C	C	C	-
scaffold0058	1663836	T	C	T	C	C	C	C	T	-
scaffold0107	601267	G	A	G	G	A	-	A	A	-
scaffold0108	1303787	G	A	G	A	A	A	A	A	-
scaffold0171	50985	A	G	G	A	-	A	A	A	-
scaffold0171	761366	G	T	T	G	G	-	T	G	G
scaffold0171	1459133	T	C	-	T	T	C	C	-	C
scaffold0175	1548310	G	A	-	G	A	G	A	A	G
scaffold0117	271696	G	A	-	A	A	A	A	A	A
scaffold0147	258651	A	C	A	A	C	-	A	C	-
scaffold0144	11553	C	T	T	C	-	C	C	C	-
scaffold0160	1301533	G	A	A	A	G	G	G	G	A
scaffold0190	526878	C	T	-	T	-	-	C	C	T
scaffold0191	652149	G	A	G	A	-	G	A	-	G
scaffold0708	1251377	G	C	G	-	G	C	G	G	G
scaffold0711	73459	T	C	C	C	T	T	-	C	C
scaffold0711	151410	A	G	-	G	A	G	A	A	G
scaffold0711	151411	A	G	G	A	-	A	G	A	G
scaffold0711	153975	C	T	T	T	C	-	C	C	C
scaffold0711	321710	A	G	-	-	G	-	-	-	-
scaffold0711	520444	C	T	C	T	T	T	-	-	-
scaffold0711	585045	T	G	G	T	G	G	T	G	T
scaffold0777	175224	T	G	T	T	T	T	G	T	G
scaffold0717	154744	T	G	T	-	G	-	G	T	-
scaffold0718	586757	G	A	G	G	G	A	A	A	-
scaffold0719	711767	T	G	G	-	T	T	-	T	G
scaffold0746	9828	A	G	A	A	A	G	-	A	G
scaffold0747	1208530	G	A	-	A	G	-	A	G	-
scaffold0758	1215157	A	T	A	-	A	T	-	A	-
scaffold0776	85221	C	A	C	A	A	A	C	A	A
scaffold0776	161264	T	C	C	C	-	C	-	T	C
scaffold0788	601667	G	A	-	G	A	G	G	A	-
scaffold0797	888578	T	C	-	-	C	C	C	-	T
scaffold0118	542035	G	A	-	-	G	A	A	-	A
scaffold0170	847416	G	A	-	A	-	A	A	A	A
scaffold0141	587532	A	G	A	A	A	A	-	-	A
scaffold0185	731688	G	A	A	A	-	-	A	G	A
scaffold0185	731747	G	A	-	-	G	A	-	G	A
scaffold0199	702837	G	A	A	G	G	-	G	A	-

Filtering 예시

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0075	870356	G	C	G	-	C	C	C	C	-
scaffold0058	1663836	T	C	T	C	C	C	C	T	-
scaffold0107	601267	G	A	G	G	A	-	A	A	-
scaffold0108	1303787	G	A	G	A	A	A	A	A	-
scaffold0171	50985	A	G	G	A	-	A	A	A	-
scaffold0171	761366	G	T	T	G	G	-	T	G	G
scaffold0171	1459133	T	C	-	T	T	C	C	-	C
scaffold0175	1548310	G	A	-	G	A	G	A	A	G
scaffold0117	271696	G	A	-	A	A	A	A	A	A
scaffold0147	258651	A	C	A	A	C	-	A	C	-
scaffold0144	11553	C	T	T	C	-	C	C	C	-
scaffold0160	1301533	G	A	A	A	G	G	G	G	A
scaffold0190	526878	C	T	-	T	-	-	C	C	T
scaffold0191	652149	G	A	G	A	-	G	A	-	G
scaffold0708	1251377	G	C	G	-	G	C	G	G	G
scaffold0711	73459	T	C	C	C	T	T	-	C	C
scaffold0711	151410	A	G	-	G	A	G	A	A	G
scaffold0711	151411	A	G	G	A	-	A	G	A	G
scaffold0711	153975	C	T	T	T	C	-	C	C	C
scaffold0711	321710	A	G	-	-	G	-	-	-	-
scaffold0711	520444	C	T	C	T	T	T	-	-	-
[illegible]										
scaffold0777	175224	T	G	T	T	T	T	G	T	G
[illegible]										
scaffold0718	586757	G	A	G	G	G	A	A	A	-
scaffold0719	711767	T	G	G	-	T	T	-	T	G
scaffold0746	9828	A	G	A	A	A	G	-	A	G
scaffold0747	1208530	G	A	-	A	G	-	A	G	-
[illegible]										
scaffold0776	85221	C	A	C	A	A	A	C	A	A
[illegible]										
scaffold0788	601667	G	A	-	G	A	G	G	A	-
scaffold0797	888578	T	C	-	-	C	C	C	-	T
scaffold0118	542035	G	A	-	-	G	A	A	-	A
scaffold0170	847416	G	A	-	A	-	A	A	A	A
scaffold0141	587532	A	G	A	A	A	A	-	-	A
scaffold0185	731688	G	A	A	A	-	-	A	G	A
scaffold0185	731747	G	A	-	-	G	A	-	G	A
scaffold0199	702837	G	A	A	G	G	-	G	A	-

Filtering을 통한 최종 산물

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0777	175224	T	G	T	T	T	T	G	T	G
[redacted]										

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
[redacted]										
scaffold0776	85221	C	A	C	A	A	A	C	A	A
[redacted]										

Filtering을 통한 최종 산물

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0777	175224	T	G	T	T	T	T	G	T	G
scaffold0776	85221	C	A	C	A	A	A	C	A	A

Filtering을 통한 최종 산물

CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0777	175224	T	G	T	T	T	T	G	T	G
scaffold0776	85221	C	A	C	A	A	A	C	A	A

Filtering을 통한 최종 산물

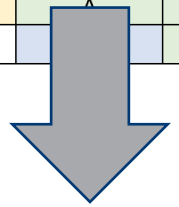
CHR	POS	REF	ALT	Sample2	Sample3	Sample4	Sample6
scaffold0777	175224	T	G	T	T	T	T
scaffold0718	586757	G	A	G	G	A	A
scaffold0776	85221	C	A	A	A	A	A

Filtering을 통한 최종 산물

CHR	POS	REF	ALT		Sample2	Sample3	Sample4		Sample6	
scaffold0777	175224	T	G		T	T	T		T	
scaffold0718	586757	G	A		G	G	A		A	
scaffold0776	85221	C	A		A	A	A		A	
scaffold0788	601667	G	A		G	A	G		A	

Filtering을 통한 최종 산물

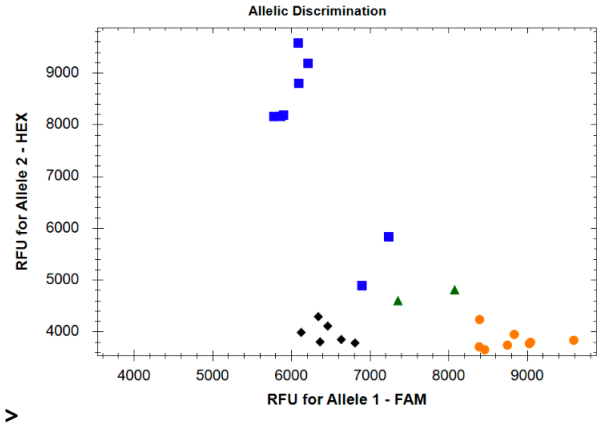
CHR	POS	REF	ALT	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
scaffold0777	175224	T	G	T	T	T	T	G	T	G
scaffold0718	586757	G	A	G	G	G	A	A	A	-
scaffold0776	85221	C	A	C	A	A	A	C	A	A
scaffold0788	601667	G	A	-	A	A	G	G	A	-



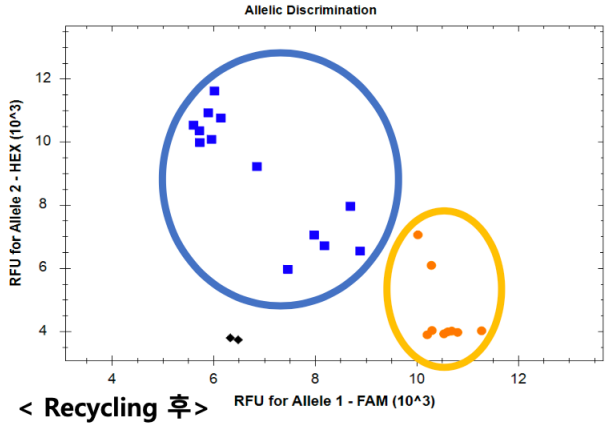
CHR	scaffold0777	scaffold0718	scaffold0776	scaffold0788
POS	175224	586757	85221	601667
Sample1	T	G	C	-
Sample2	T	G	A	G
Sample3	T	G	A	A
Sample4	T	A	A	G
Sample5	G	A	C	G
Sample6	T	A	A	A
Sample7	G	-	A	-

실험을 통한 검증 (KASP 예시)

Assay_name	Allele1	Allele2	Primer_AlleleX	Primer_AlleleY	Primer_Common
4_16052654	C	A	AAGTCTTTCGTGCCATTTAGCAACC	AAAGTCTTTCGTGCCATTTAGCAACA	CTACCTTACATCAATAACCTCCTCTCTTT



< Recycling 전>



< Recycling 후>

◆ = No call
X = Undetermined
● = Allele 1
■ = Allele 2
▲ = Hetero

sample	Allele type	sample	Allele type	sample	Allele type	sample	Allele type
A1	1	B1	He	C1	1		
A2	2	B2	1	C2	1	A2	1
A3	1	B3	No	C3	No		
A4	1	B4	1	C4	2		
A5	He	B5	2	C5	No		
A6	2	B6	2	C6	No		
A7	2	B7	2	C7	2		

sample	Allele type	sample	Allele type	sample	Allele type	sample	Allele type
A1	1	B1	1	C1	1		
A2	2	B2	1	C2	1	A2	1
A3	1	B3	2	C3	2		
A4	1	B4	1	C4	2		
A5	2	B5	2	C5	2		
A6	2	B6	2	C6	2		
A7	2	B7	2	C7	2		

Summary

- NGS의 발전으로 인하여 sequencing 데이터 생산의 가격은 낮아지고, 속도는 빨라짐.
- 유전체 상에 존재하는 다양한 변이 정보를 NGS 데이터로 분석할 수 있음.
- NGS로 얻은 변이 정보를 이용하여 형질연관마커, MAS, MAB, 순도검정, 원산지 구분, QTL-mapping, GWAS와 같은 다양한 분석이 가능
- 마커 개발을 위한 NGS 기법에는 WGS, RNA-seq, GBS와 같은 다양한 기법이 존재
- NGS 데이터를 분석하기 위해 Pre-processing, Read Alignment, Variants Detection, 마커 후보군 개발의 순으로 분석이 진행됨.

II. 디지털육종

1. 디지털육종이란?

육종

❑ 정의

- 인간이 원하는 형태로 작물을 개선(진화 또는 변형)
- 농작물이나 가축을 개량하여 경제(실용) 가치가 더 높은 새로운 품종을 개발하고 증식하여 보급하는 기술
- 예) 수량 증대, 품질 향상, 내재해성, 내병성, 맛, 향기(풍미), 모양, 사육 환경 등

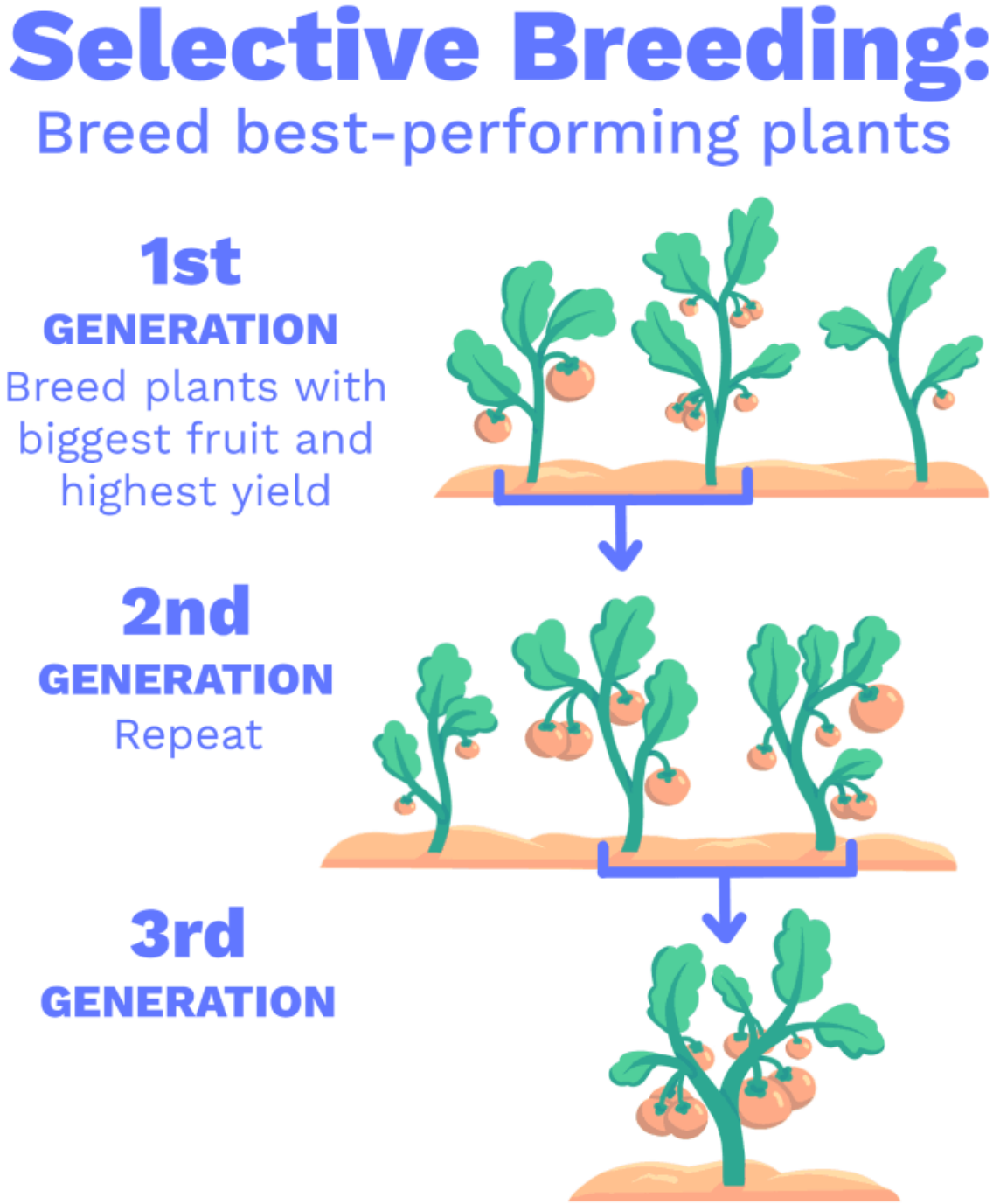
❑ 대상

- 농경을 시작한 이래로 산업적으로 유용한 형질(표현형)을 가진 모든 생물체
- 동식물을 망라하고 인간에게 유용한 경제 형질을 가진 모든 분야에서 육종 진행
- 예) 마블링이 우수한 1등급 '한우', 매운맛의 강자 '청양고추', 수확량의 제왕 '통일벼', 밀을 대체할 벼 품종 '가루미' 등

❑ 문제점

- 기존의 전통적인 분리육종만으로는 (세대진전을 위해) 수년 ~ 수십 년 이상을 필요
- 현대 육종방법에서는 최첨단 과학기술을 사용하여 기간 단축을 위한 기술이 요구됨.

관습육종

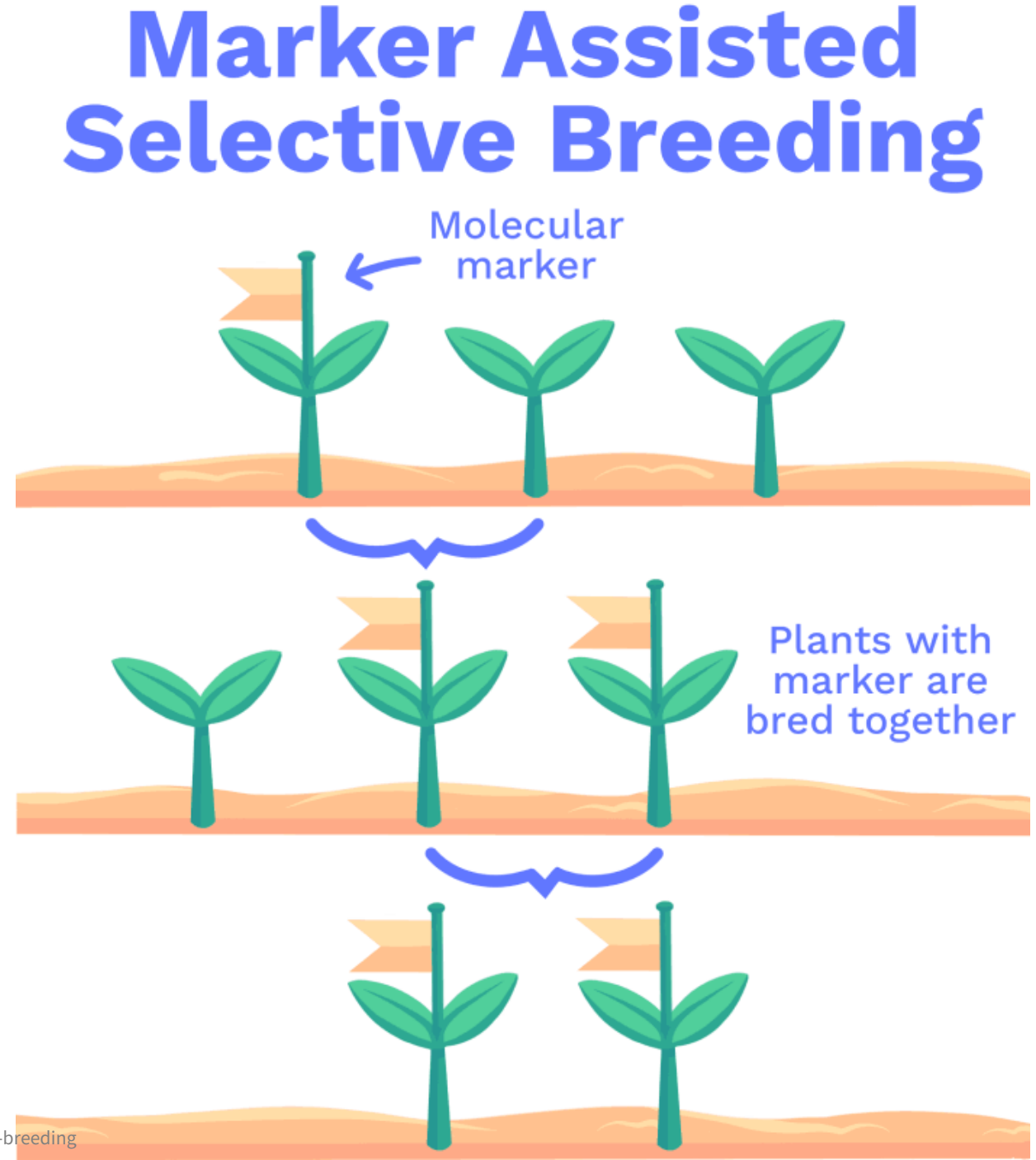


- https://climatescience.org/advanced-food-selective-breeding

관습육종

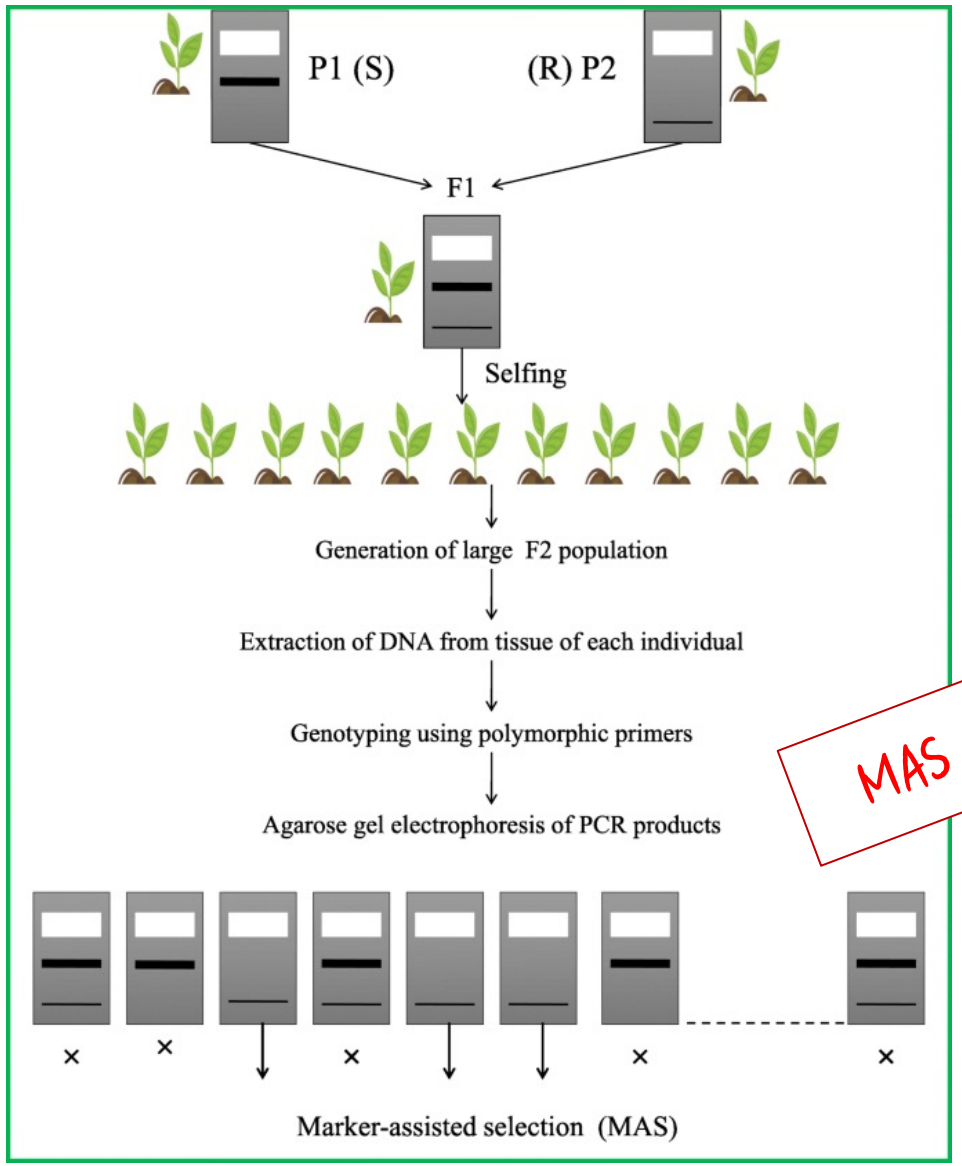
- https://climatescience.org/advanced-food-selective-breeding

분자육종

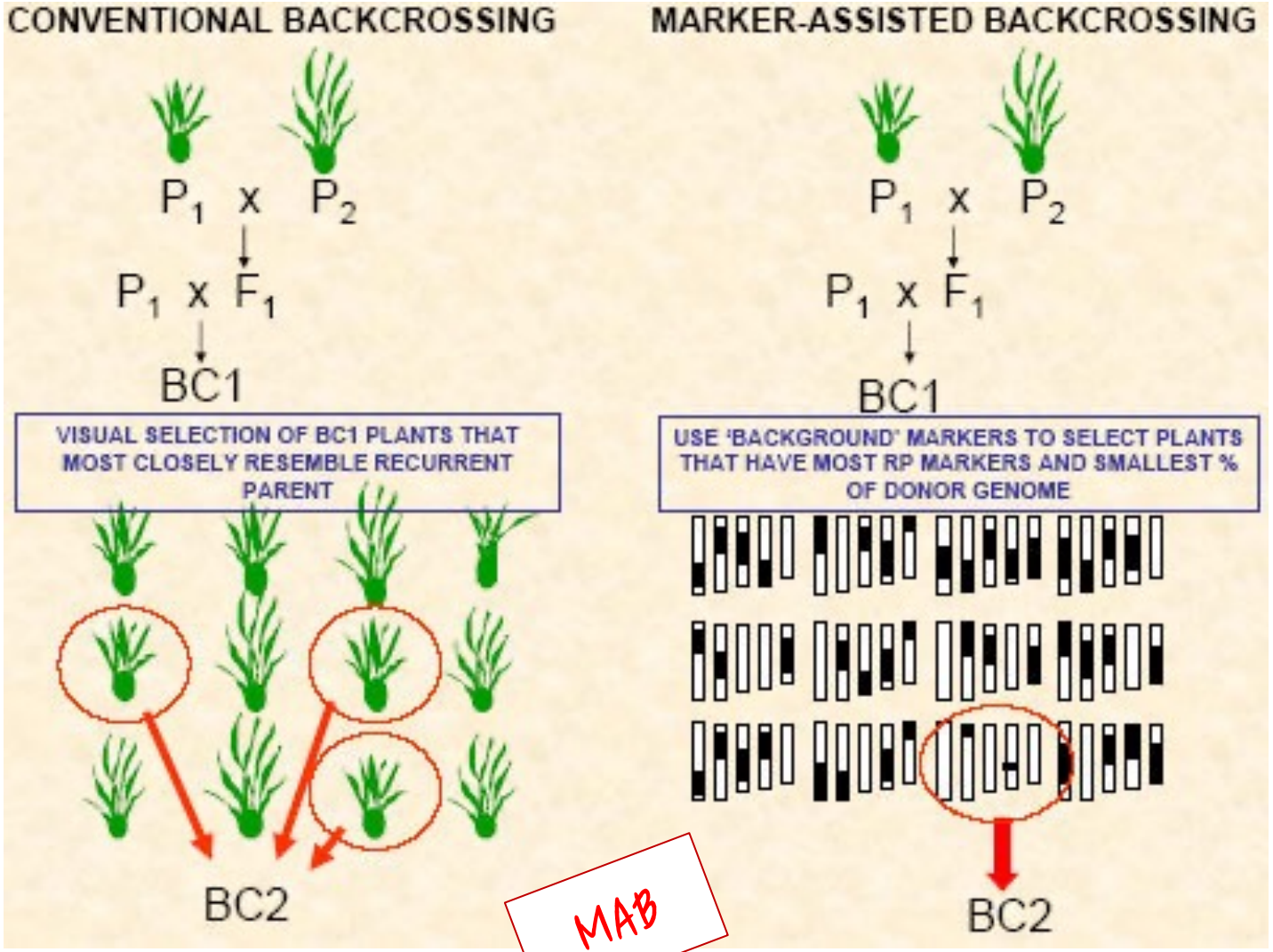


- https://climatescience.org/advanced-food-selective-breeding

분자마커의 효용성은 선발을 빨리 할 수 있다는 것



MAS



MAB

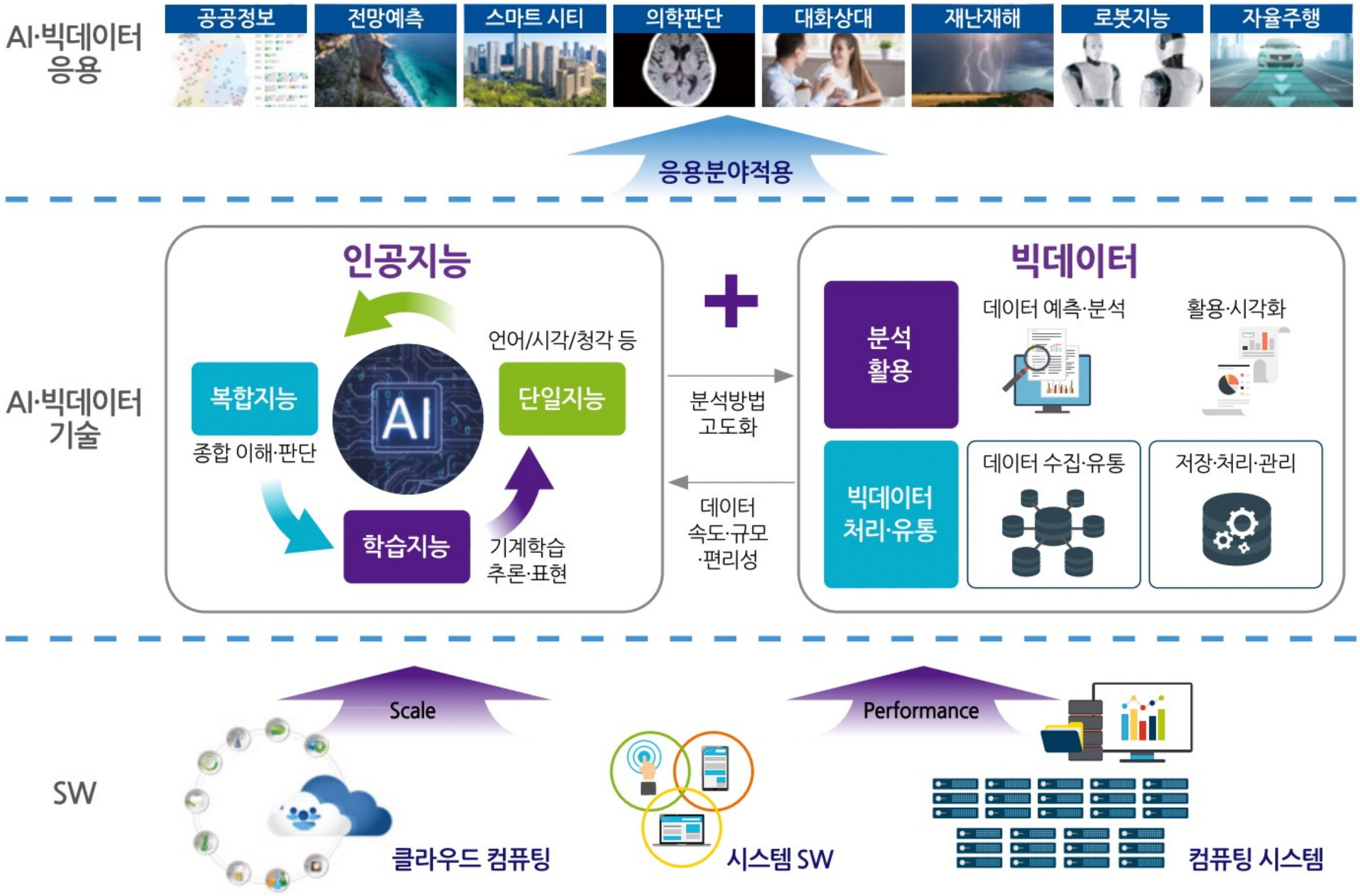
- https://jgeb.springeropen.com/articles/10.1186/s43141-021-00231-1
- http://www.knowledgebank.irri.org/ricebreedingcourse/Marker_assisted_breeding.htm

인공지능 (AI) 시대

인공지능을 활용한 육종

↓

디지털 육종



디지털육종을 위한 필수 요소 3가지

❑ 개체(샘플)

- 현재 관찰이 가능한 모든 개체를 대상으로 함.
- 야생종부터 돌연변이 종까지 제한이 없으며, 동일한 생장 조건이 필요하지도 않음.
- 예) 밤나무: 전국의 수집 가능한 모든 밤나무가 대상

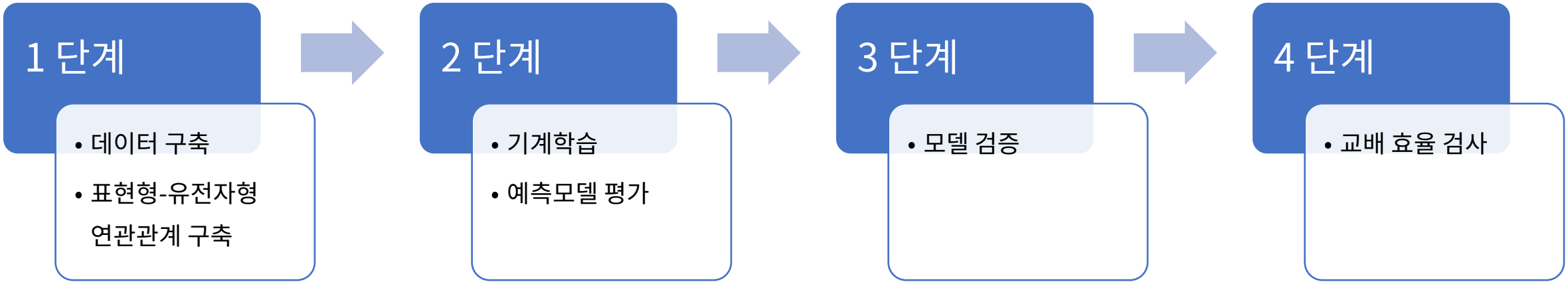
❑ 표현형

- 현재 확보된 개체들에서 관찰되는 모든 것을 대상으로 함.
- 예) 밤나무: 알곡의 크기, 지역, 수확 시기, 나무의 크기, 밤송이의 수 등 현재 시점에서 관찰할 수 있는 모든 것
- 이를 좀 더 효율적으로 수집하고자 디지털화된 장비를 이용한 표현형 수집 방법이 부상
- 온실에 카메라를 설치하고 주기적으로 사진을 촬영하여 영상 분석을 통해 표현형을 처리하는 AI 기술이 접목된 스마트 팜

❑ 유전형

- NGS 기술과 생물정보의 발달로 누구나 쉽게 얻을 수 있음.

디지털육종의 과정


1 단계
• 데이터 구축
• 표현형-유전자형 연관관계 구축
2 단계
• 기계학습
• 예측모델 평가
3 단계
• 모델 검증
4 단계
• 교배 효율 검사

디지털육종의 과정 (1단계)

❑ 데이터베이스 구축 단계

- 각 재배 및 사육 단계에서 육안 또는 영상, ICT 장비를 이용해 데이터를 체계적으로 수집하는 단계
- 데이터를 수치화하고 객관화시켜 신뢰할 수 있는 표현체 빅데이터를 구축해야 함.

❑ 지식 정보 그래프 구축 단계

- 수집 데이터를 '노드'-'엣지'의 그래프로 구현하는 단계
- 데이터 관계를 명확하게 정의하여 표현형, 기능, 유전형의 관계를 체계적으로 구축하는 단계

❑ 연관 관계 분석 단계

- 각 수집 요소에 대하여 알고리즘을 적용하는 단계
- 표현형, 환경정보, 유전형의 연관 관계에 중요도를 부여 → 특정 표현형을 선발하기 위한 유전형 우선순위 정보 제공

❑ 맞춤형 분석 단계

- 데이터베이스를 재구성하거나 탐색을 통해 자신과 가진 데이터를 비교하거나 분석할 수 있는 플랫폼을 제공하는 단계
- 특정 표현형에 대한 집단 비교 분석 및 AI 기술 도입을 통해 최적의 교배 지침을 제공하는 단계

디지털육종의 과정 (2단계)

❑ 기계학습

- 표현형에 따른 집단의 구분력을 보이는 변이만을 유전형 데이터로 활용하여 표현형-유전형 기계학습을 수행
- 수집된 개체의 75%를 학습 데이터로, 나머지 25%는 학습된 기계학습의 정확도 평가를 위해 사용

- 예) 밤 알곡의 크기를 예측하기 위한 기계학습
  - 1단계에서 선별된 변이 마커가 30개라면, 개체별 30개의 유전형 정보에 따라 측정된 알곡의 크기를 표현형 정보로 학습
  - 이후 구축된 예측모델을 이용해 남겨 두었던 25% 개체의 유전형 정보를 넣고 해당 표현형이 예측될 가능성이 얼마나 될지 확률치 계산

❑ 예측모델 평가

- True Positive, False Negative로 계산되는 Specificity와 Sensitivity 이용
- 만약 만족스럽지 못하면, 기계학습 알고리즘 변경 or 학습 데이터 변경

디지털육종의 과정 (3 ~ 4단계)

❑ (3단계) 예측모델 검증

- 구축된 기계학습모델로 다른 개체에 적용
- 해당 표현형을 예측하는 데 사용되는 마커는 1단계에서 30개로 선별되었고, 이후 미지의 시료에 대해 30개 마커 유전형만을 타입핑하여 표현형 예측

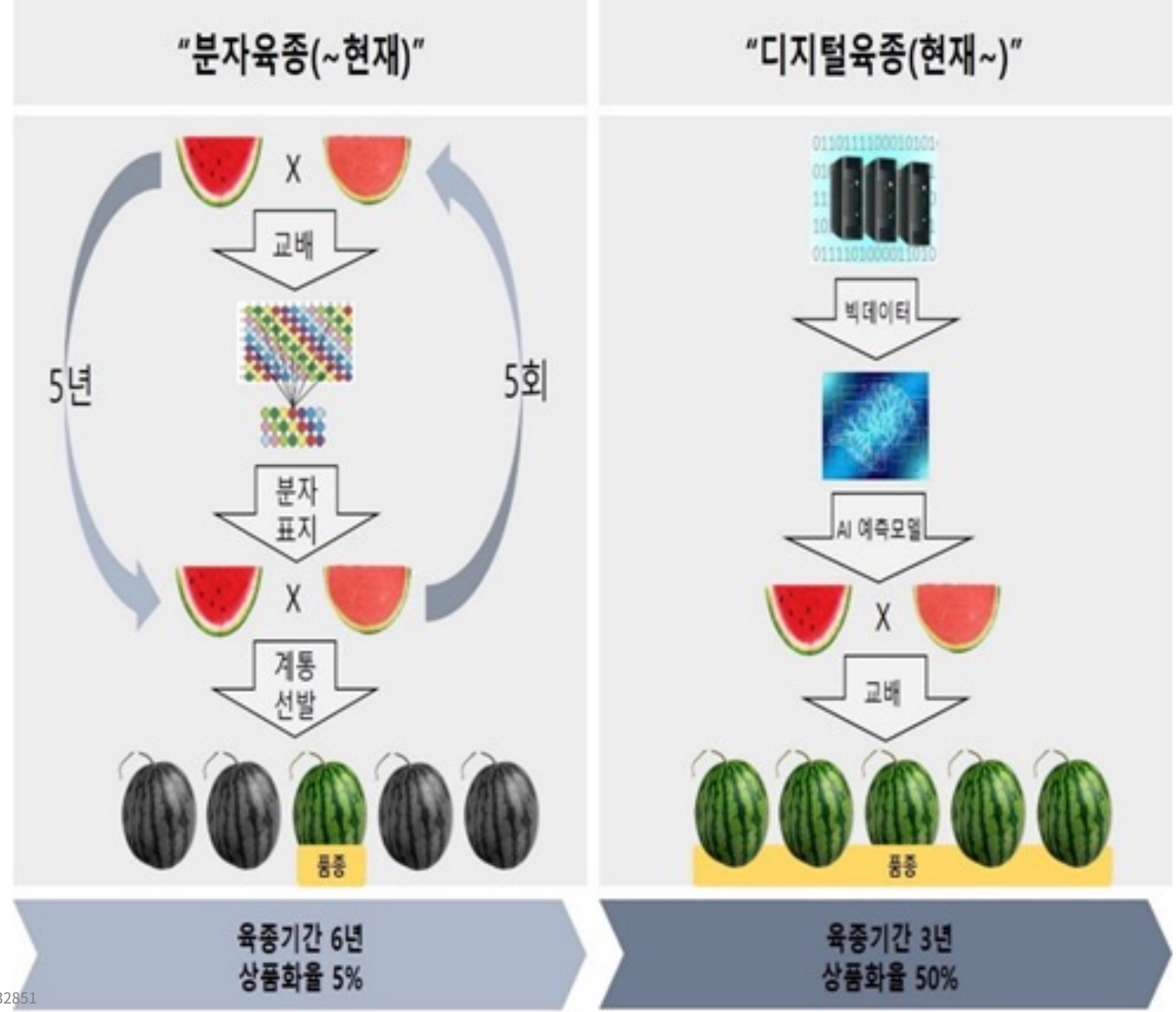
❑ (4단계) 시뮬레이션을 통한 교배 효율 검사

- 최적의 교배 지침을 위한 F1 세대의 표현형 예측 시뮬레이션 진행
- 부·모가 될 개체의 유전형을 기반으로 F1 세대에서 나타날 수 있는 유전형을 무작위 방식 구성

- 예) F1 세대의 개체 수는 2,000개체 이상, 유전형은 해당 표현형을 예측하는 마커 수 (예: 30개 유전형)를 가진 임의 생성
  - 이후 2,000 개체의 유전형을 이용해 구축된 기계학습으로 표현형을 예측
  - F1 세대에서 해당 표현형을 가질 수 있는 평균 개체 수가 어느 정도 되는지 수치화

- 이러한 방식으로 F1 세대에서 해당 표현형을 가질 수 있는 개체수가 많은 순서로 교배 조합을 시뮬레이션

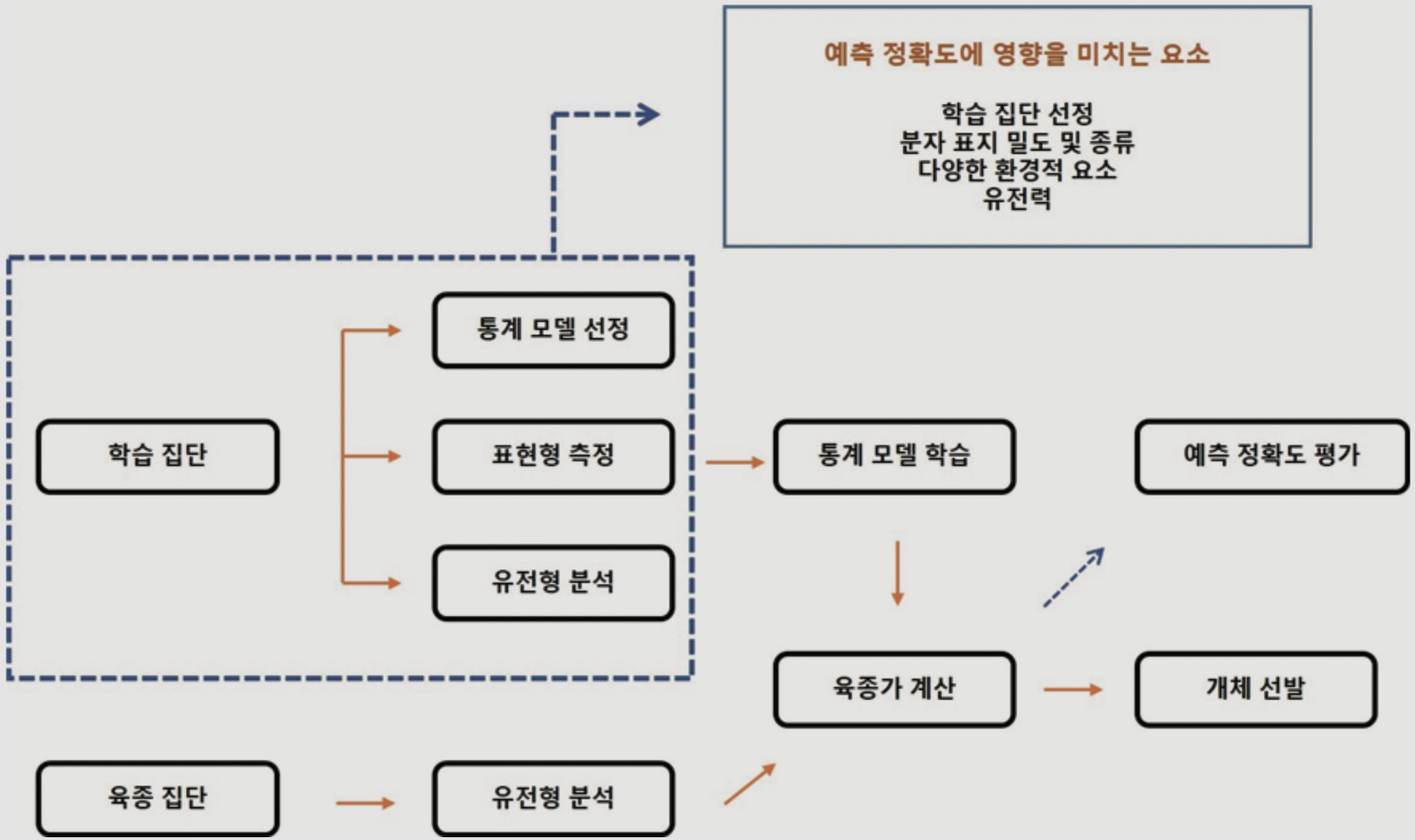
분자육종과 디지털육종 비교



- http://www.newsam.co.kr/mobile/article.html?no=32851

예측 육종의 기술 중 하나인 유전체 기반 선발법

[그림 3] 예측 육종의 기술 중 하나인 유전체 기반 선발법



- BIO ECONOMY BRIEF. “식물 디지털 육종 기술의 발전과 미래의 종자산업”. Nov. 2022. Issue 170.

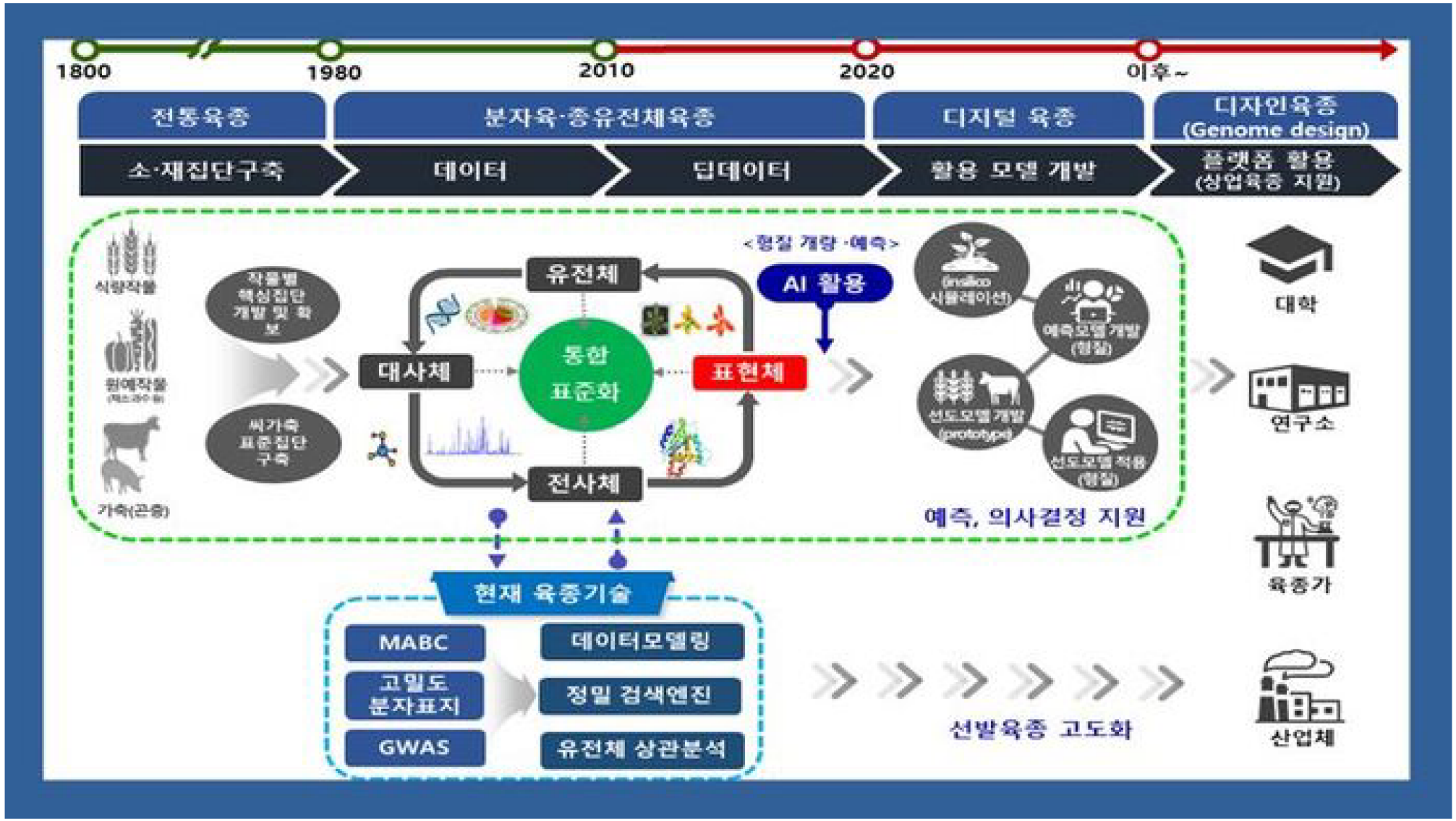
디지털 육종의 단계별 기술 정의 및 자율주행 기술 단계별 분류와의 비교

[표 1] 디지털 육종의 단계별 기술 정의 및 자율주행 기술 단계별 분류와의 비교

단계	기술정의	육종통계모델	육종법	자율주행 기술정의 (SAE, 미국자동차공학회, 2016)
0	디지털기술 적용 없음	없음	전통육종	운전자 항시운행
			비자동화	
1	분자표지 개발을 위한 분자수준의 데이터 이용	QTL (composite interval model)	분자육종	시스템이 차간거리 조향등 보조
			육종가(운전자) 적극 개입	
2	분자표지 개발을 위한 빅데이터 이용	GWAS (GLM, MLM, FarmCPU, etc.)	분자육종	특정 조건에서 시스템이 보조 주행
			부분 자동화	
3	양적형질에 대한 유전체기반 육종가 예측	GS (BLUP, LASSO, Bayesian, Machine learning, etc.)	예측육종	특정 조건에서 자율주행, 위험시 운전자 개입
			조건부 자동화	
4	환경요소를 고려한 표현형의 예측	Phenotype prediction (ML, DL)	예측육종	운전자 개입 불필요
			고도 자동화	
5	인공지능을 활용한 식물 육종의 완전 자동화	Automated breeding design of all processes (DL)	인공지능육종	운전자 불필요
			완전 자동화	

- BIO ECONOMY BRIEF. “식물 디지털 육종 기술의 발전과 미래의 종자산업”. Nov. 2022. Issue 170.

디지털육종의 생명공학 기술 변천

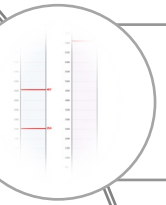


- http://www.kfdn.co.kr/60796

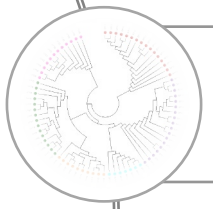
II. 디지털육종

2. 전장 유전체 변이를 이용한 다양한 응용 분석

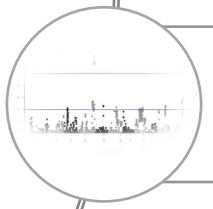
전장 유전체 변이를 이용한 응용 분석의 종류



형질, 특성에 차이가 나는 개체/집단의 정보를 탐색하는 **마커 개발**



개체 간의 관계 및 구조를 분석하는 **유연관계 분석**



교배/육성 집단과 형질 간의 양적유전자 좌를 탐색하는 **QTL 분석**

유전자원과 형질 간의 연관성을 분석하는 **GWAS 분석**

응용1 – 마커 개발

I. 마커 개발이 필요한 이유?

1. 품종 또는 집단의 구분이 필요한 경우
2. 원산지 판별이 필요한 경우
3. 특정 형질과 연관된 마커가 필요한 경우
4. 집단 육성을 위해 마커가 필요한 경우 (MAS, MAB)

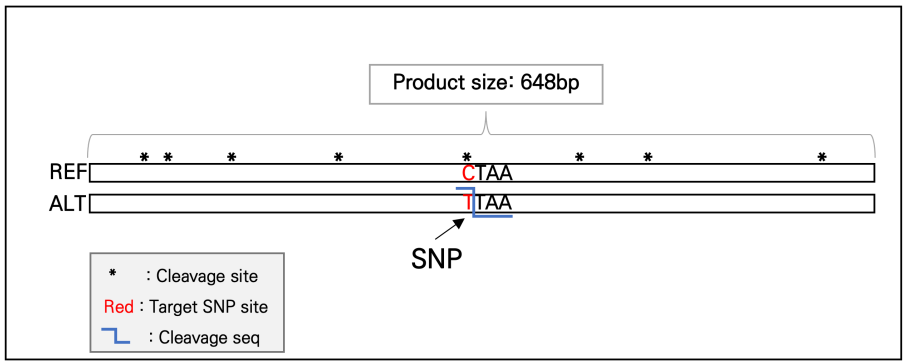
❖ **BARCODE SNP 예시**

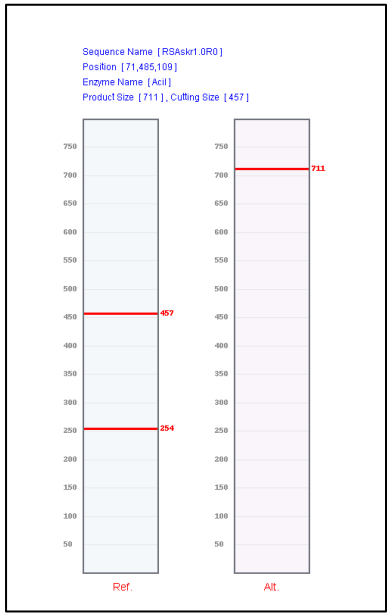
Barcode no.	1	2	3	4
Barcode-SNP type	abaabb	abaaba	ababbb	abbaaa
Barcode-SNP candidates	32	12	144	2
Sample1	a	a	a	a
Sample2	b	b	b	b
Sample3	a	a	a	b
Sample4	a	a	b	a
Sample5	b	b	b	a
Sample6	b	a	b	a

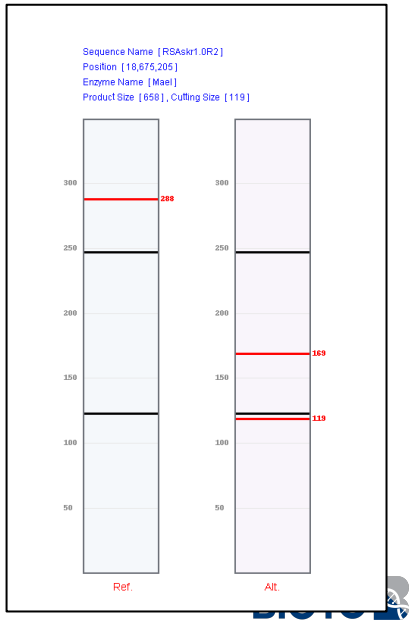
II. 개발 가능한 마커 종류

1. SNP 기반의 마커 – HRM, KASP 등
2. SNP, In/Del 기반의 마커 – CAPS, SCAR, gel base PCR 등
3. SSR 기반의 마커

❖ **CAPS 예시**







응용 2 - 유전 분석 (Genetic Analysis)

I. 유전 분석이란?

SNP 또는 In/Del과 같은 다양한 변이 정보를 이용하여 개체 또는 집단의 다양성, 다형성, 진화, 집단의 특성이나 구조 등을 분석함.

II. 유전 분석의 종류

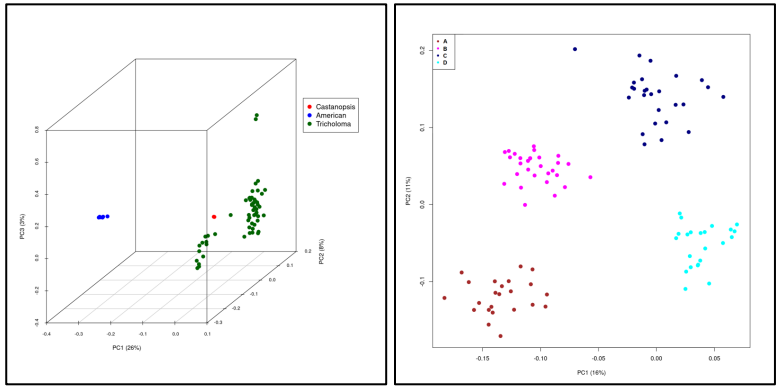
유연관계 분석

1. **주성분 분석 (PCA)**
2. **주좌표 분석 (PCoA)**
3. MDS plot
4. **계통수 분석 (Phylogenetic tree)**
5. **구조 분석 (STRUCTURE)**

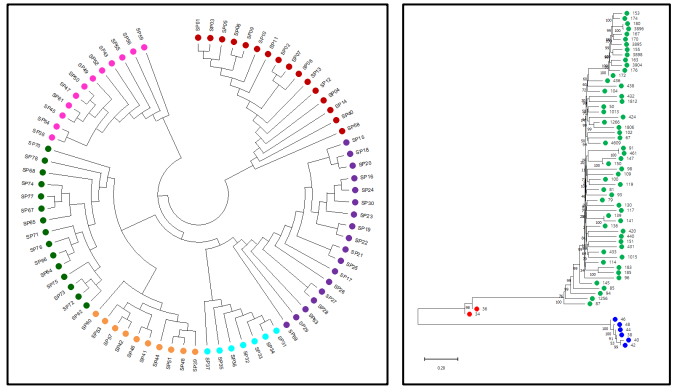
유전 다양성 분석

6. F_{ST} 분석
7. Nucleotide diversity 분석
8. Pairwise distance 분석
9. Heterozygosity 분석
10. AMOVA 분석

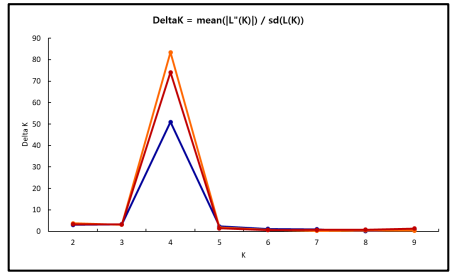
❖ 주성분 분석

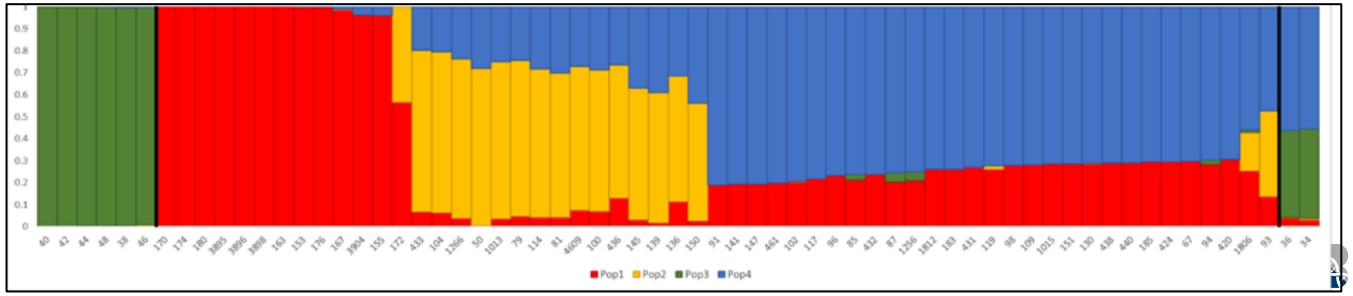


❖ 계통수 분석



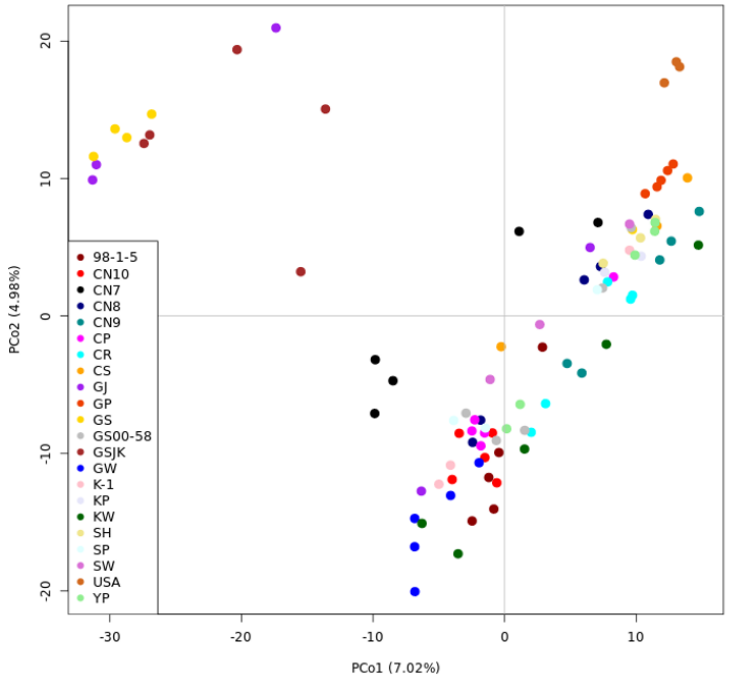
❖ 구조 분석



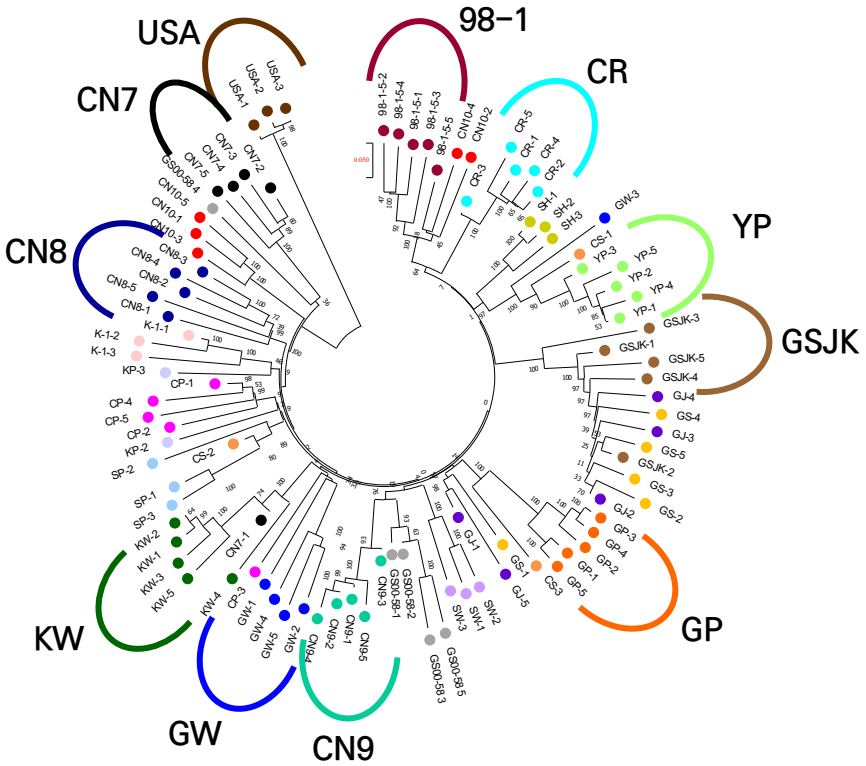


유연관계 사례 (인삼)

❖ PCoA 분석 결과



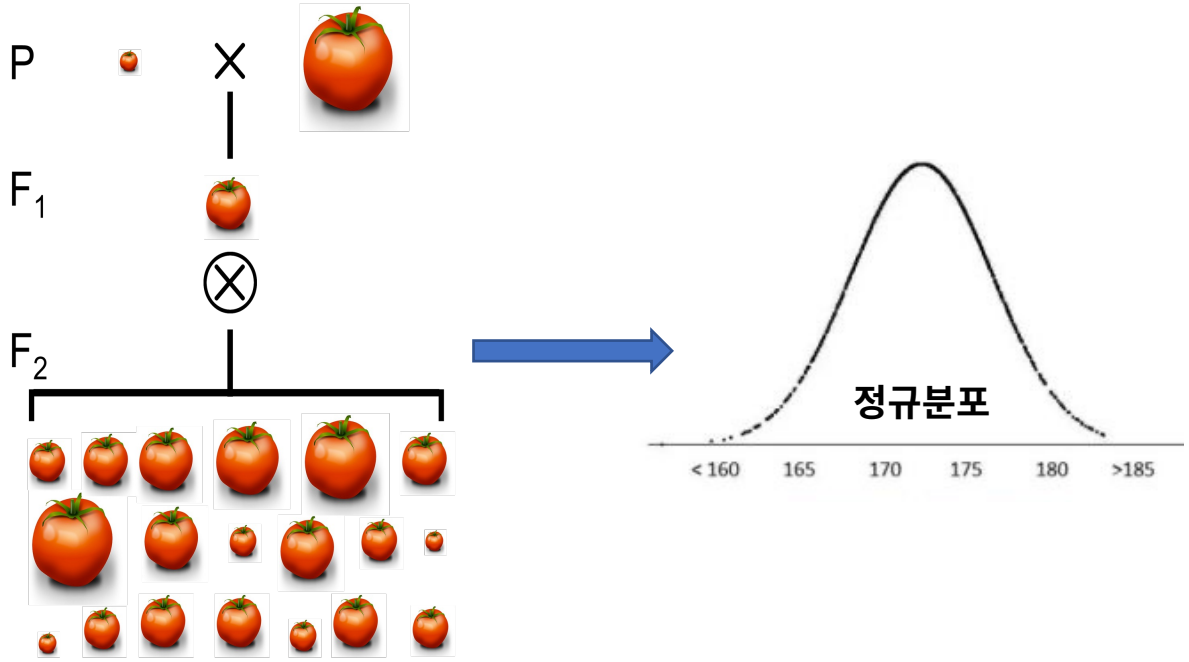
❖ Phylogenetic Tree 분석 결과



응용 3 – QTL mapping

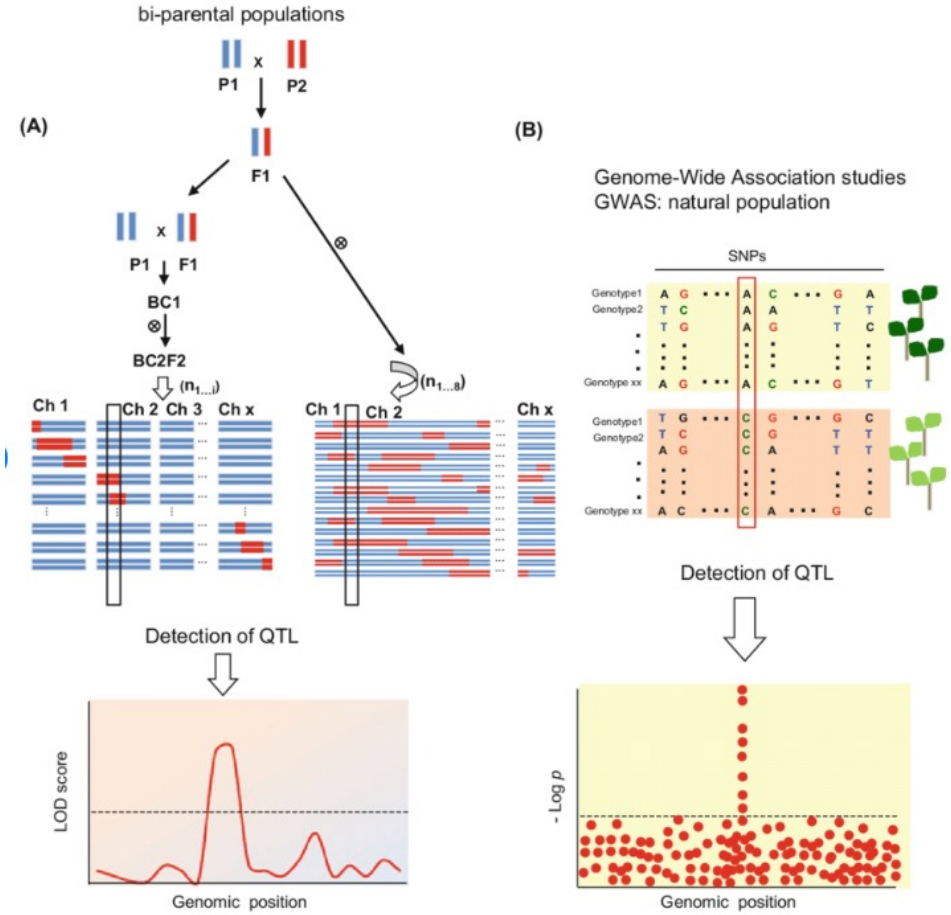
I. QTL mapping 분석이란?

1. 유전적 특성이나 형질을 조절하는 유전자 위치를 찾기 위한 분석 방법
2. 주로 **자손집단 데이터**를 이용하여 분석함.
3. 형질은 대부분 양적 형질을 이용함. 하지만 질적형질도 분석 가능.



출처: https://www.pngwing.com/en/free-png-nrwdn

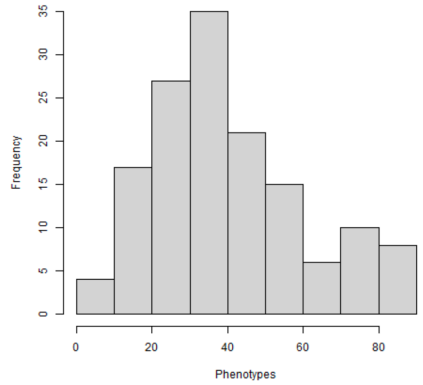
QTL-mapping vs GWAS

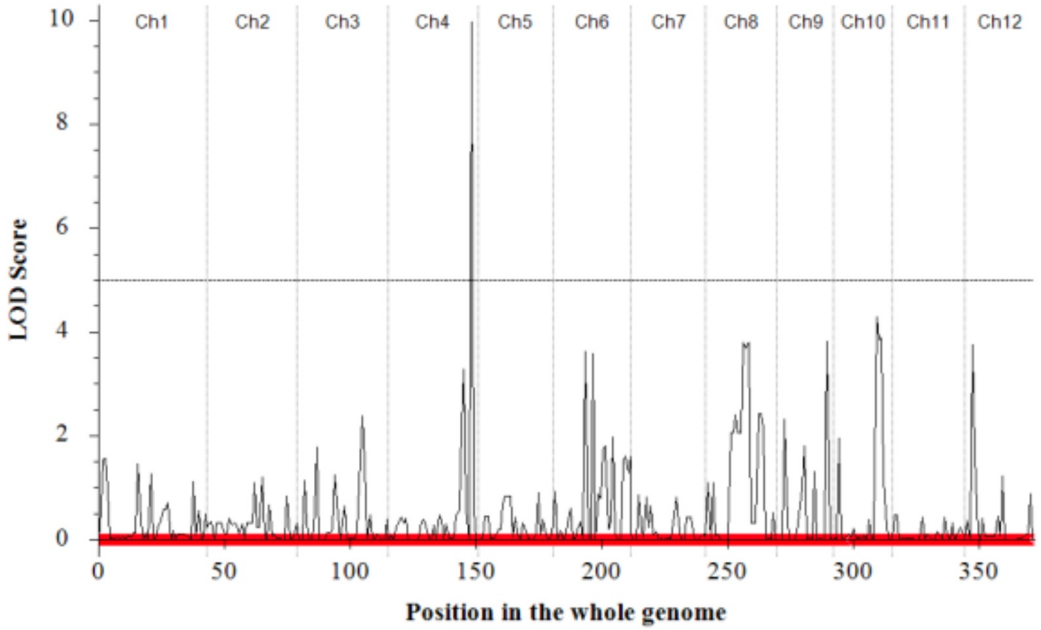


출처:Alseekh et al. 2018

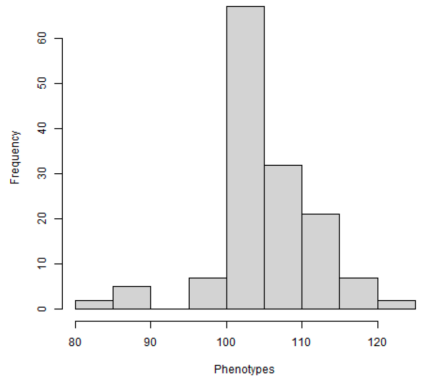
QTL-mapping 사례 (화본과)

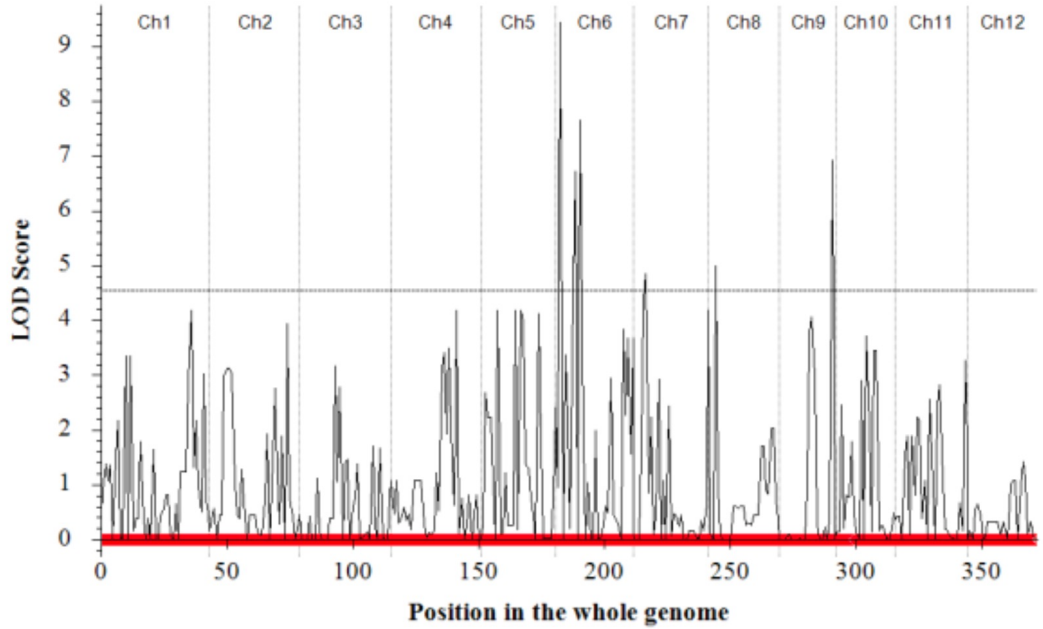
❖ Fusarium 저항성





❖ 출수기





응용 4 – GWAS 분석

I. GWAS (Genome-wide Association Study) 분석이란?

전체 게놈(유전체)에 대한 포괄적인 연구를 통해 유전적 변이와 특정 표현형 (질병 또는 형질) 간의 관계를 조사하는 유전학적 분석 방법. GWAS는 수많은 개인의 유전체 데이터와 표현형 데이터를 조합하여 특정 유전적 변이가 특정 표현형과 관련이 있는지 식별하여 분석을 함.

II. 어떤 프로그램으로 분석할 수 있나?

a. GAPIT

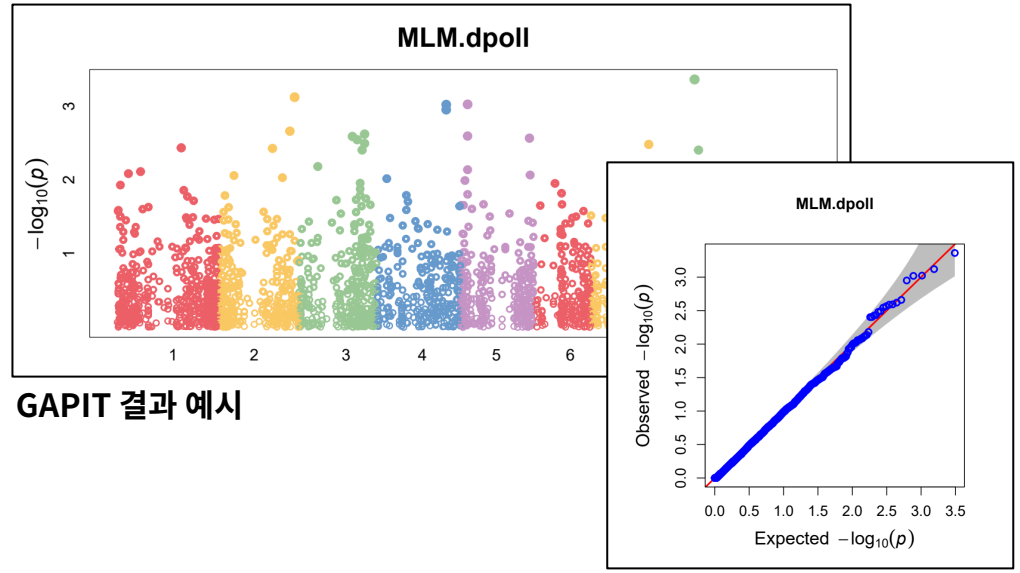
R package를 이용하여 분석하며, 기본 선형모델, 혼합 선형모델 등에 특화되어 분석할 수 있으며 간단한 명령어로 분석이 가능함.

b. TASSEL

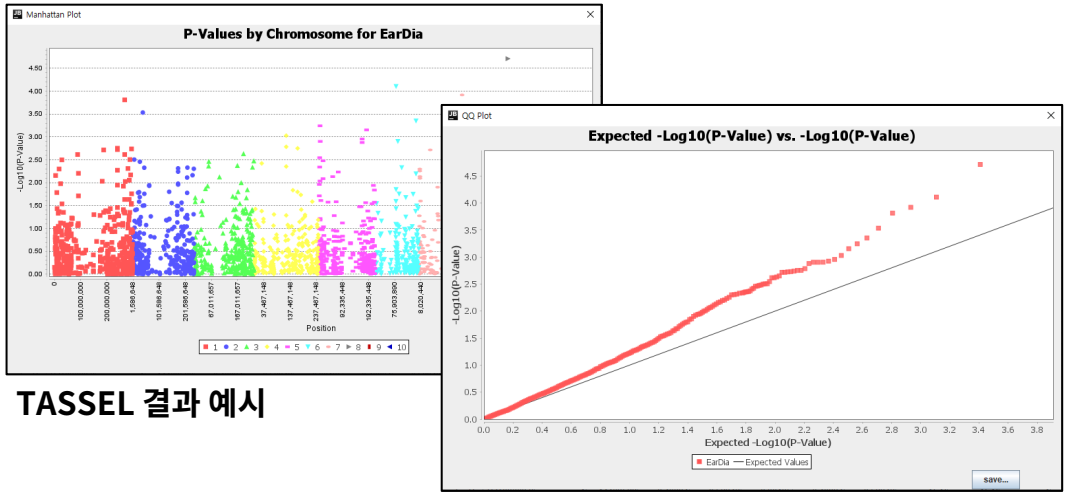
Windows 기반의 분석 프로그램으로 사용자의 편의성이 좋은 분석 프로그램으로 다양한 선형모델을 지원함.

c. PLINK

리눅스 기반의 분석 프로그램으로 대규모 GWAS 연구에 적합함. 단, 사용자별 옵션값에 따라 결과의 차이가 크게 나타날 수 있음.



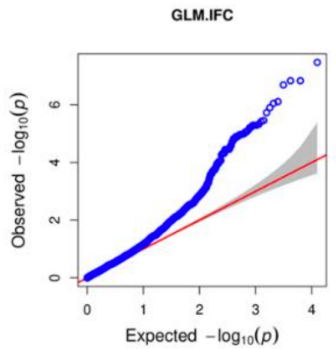
GAPIT 결과 예시

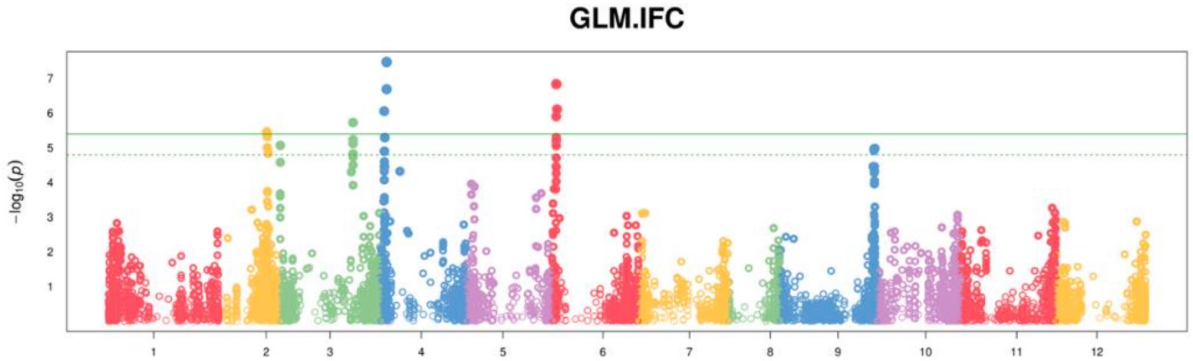


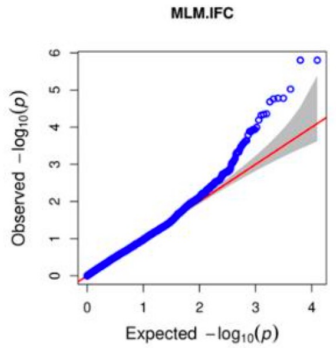
TASSEL 결과 예시

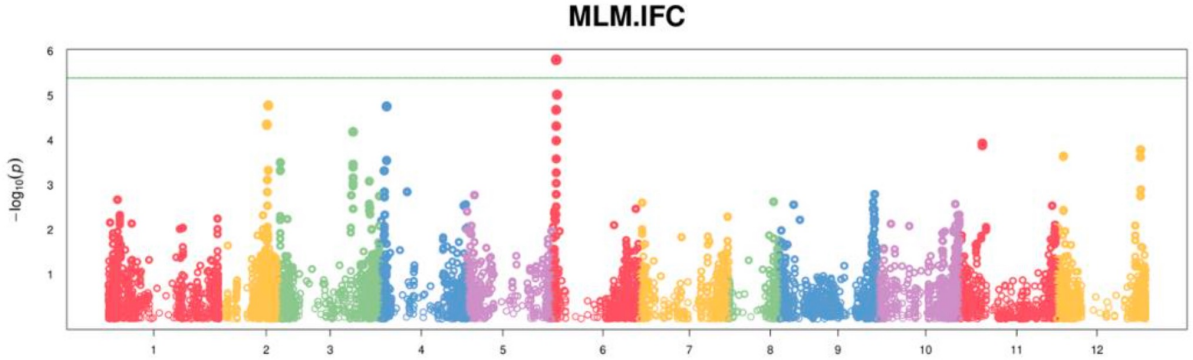
GWAS 사례 (가지과)

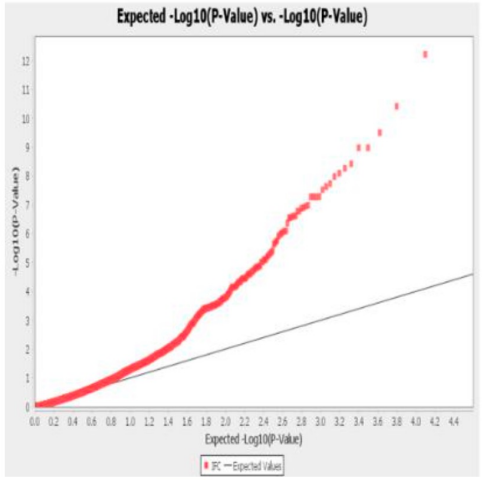
❖ 과색

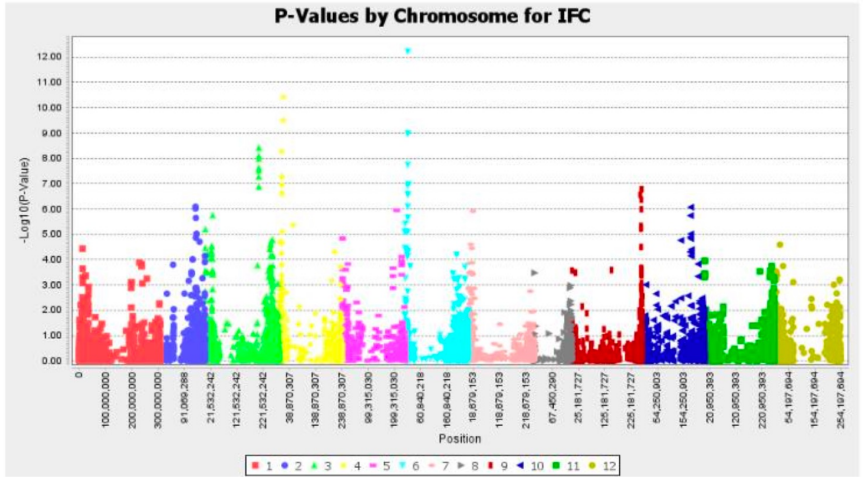








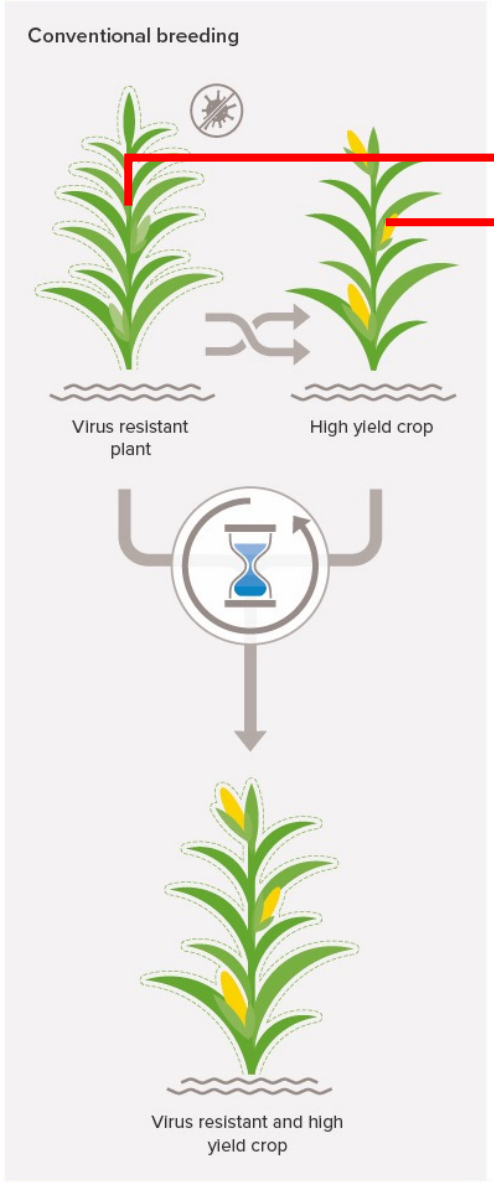


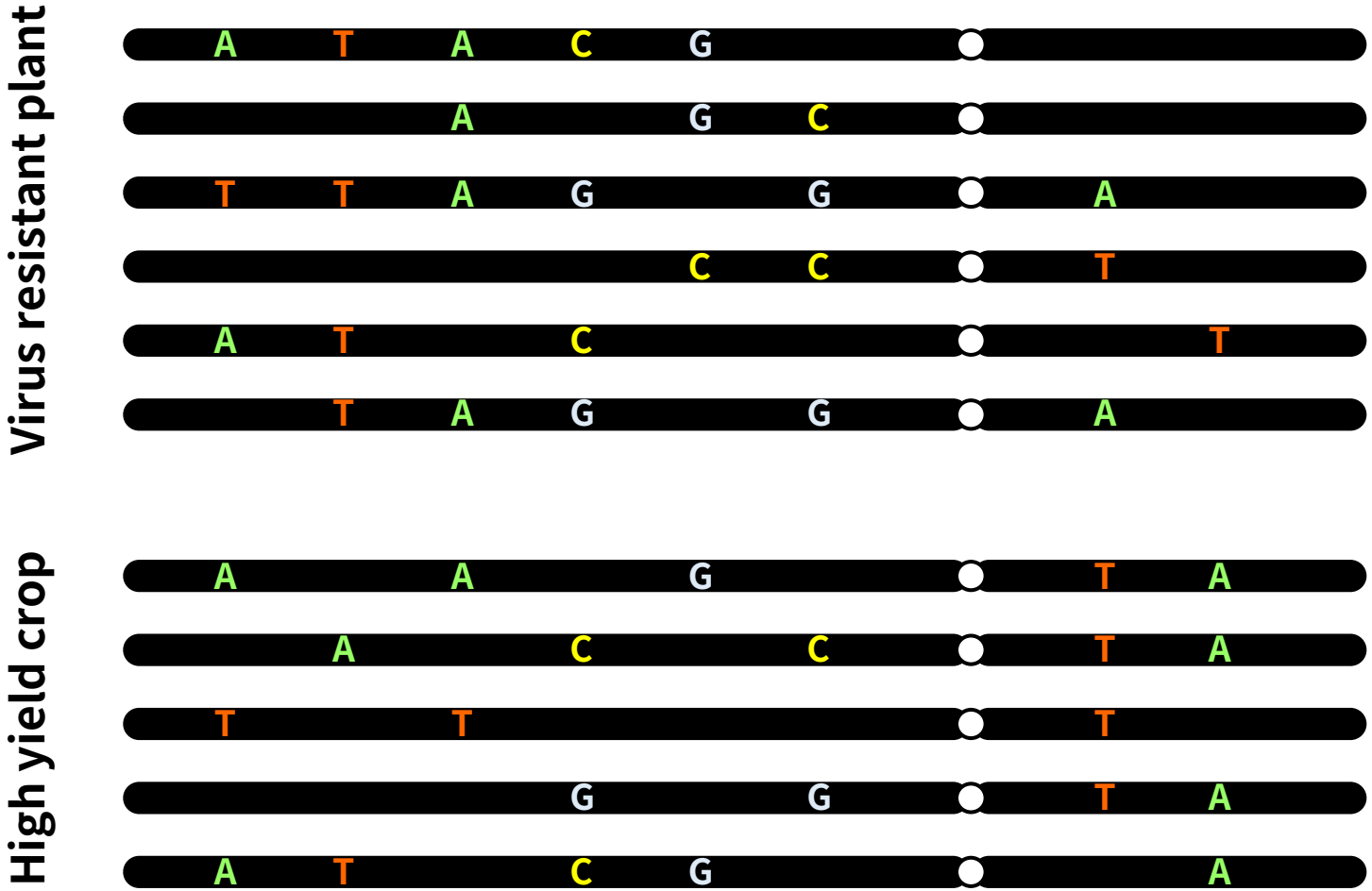


II. 디지털육종

3. 유전체-표현체 연관 분석의 기본 개념

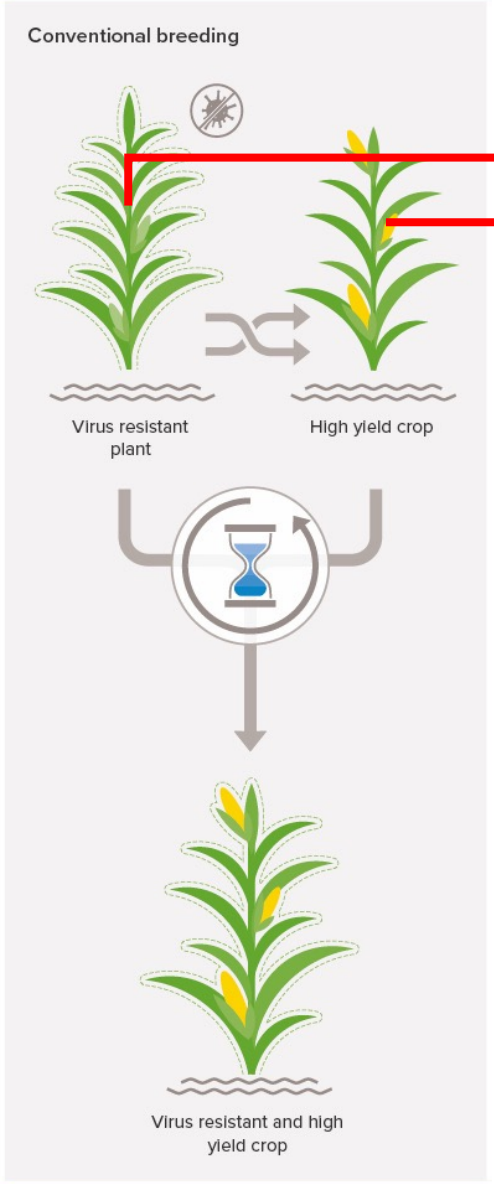
전장 유전체와 형질 연관영역의 분석 개념

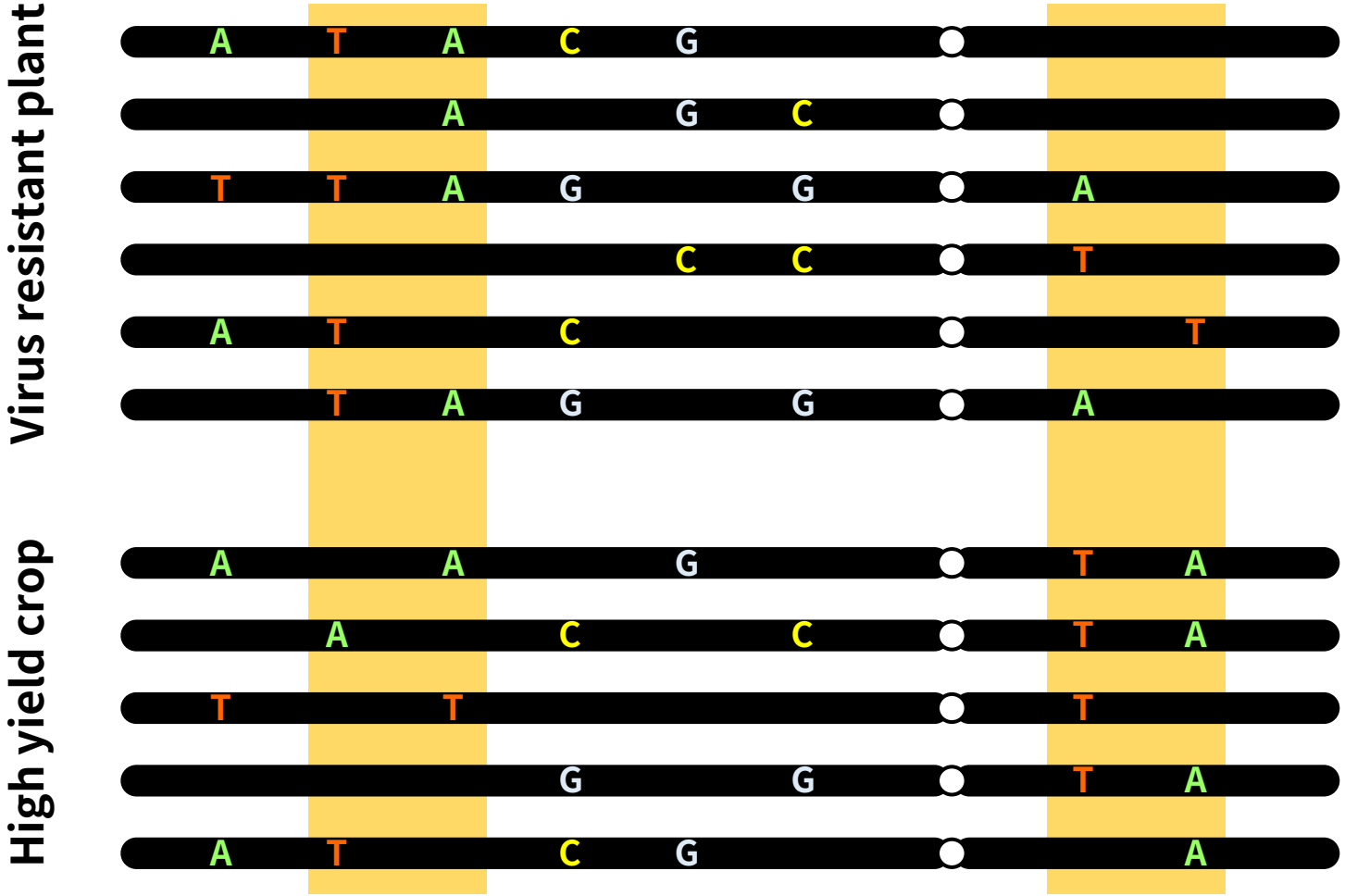




- https://royalsociety.org/topics-policy/projects/gm-plants/how-does-gm-differ-from-conventional-plant-breeding/

전장 유전체와 형질 연관영역의 분석 개념

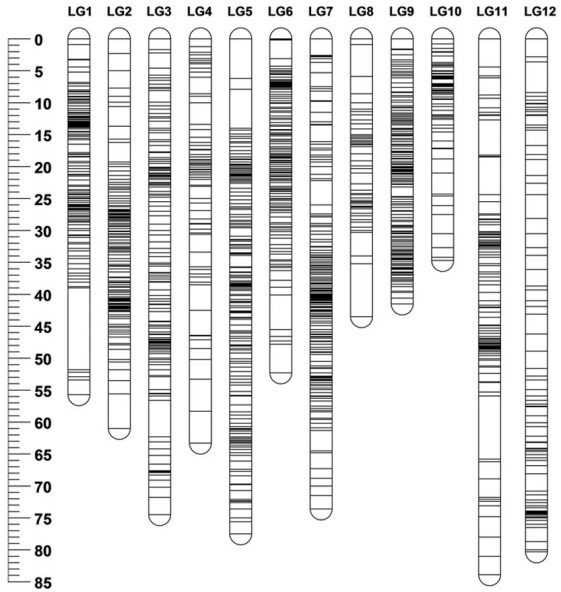




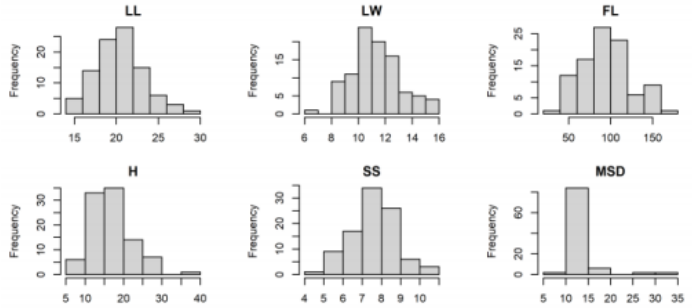
- https://royalsociety.org/topics-policy/projects/gm-plants/how-does-gm-differ-from-conventional-plant-breeding/

QTL-mapping의 분석 과정

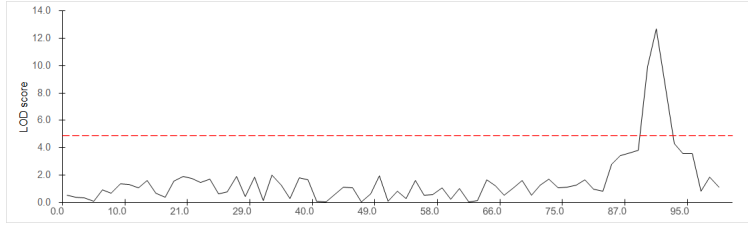
1. 집단의 변이 정보를 이용하여 유전자 연관지도 작성



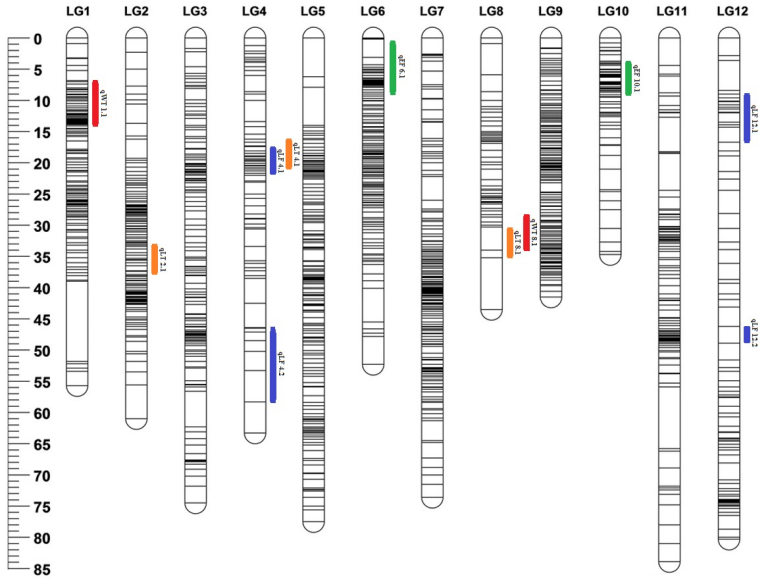
2. 연관지도 작성에 사용된 개체의 형질 데이터



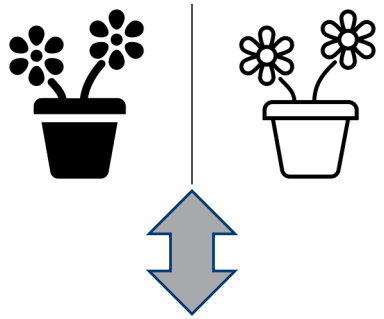
3. 1과 2의 데이터를 이용하여 QTL mapping 분석을 진행

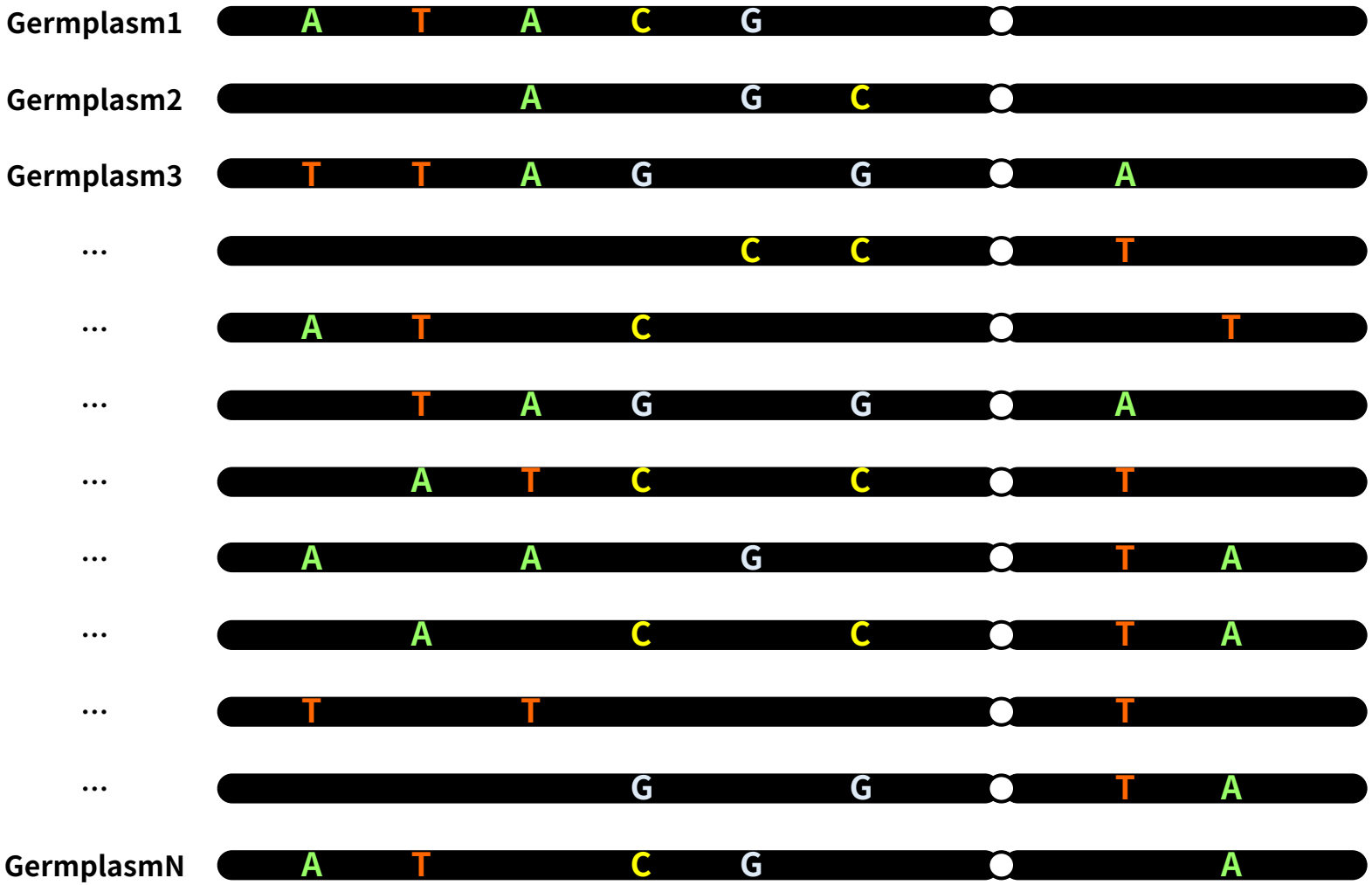


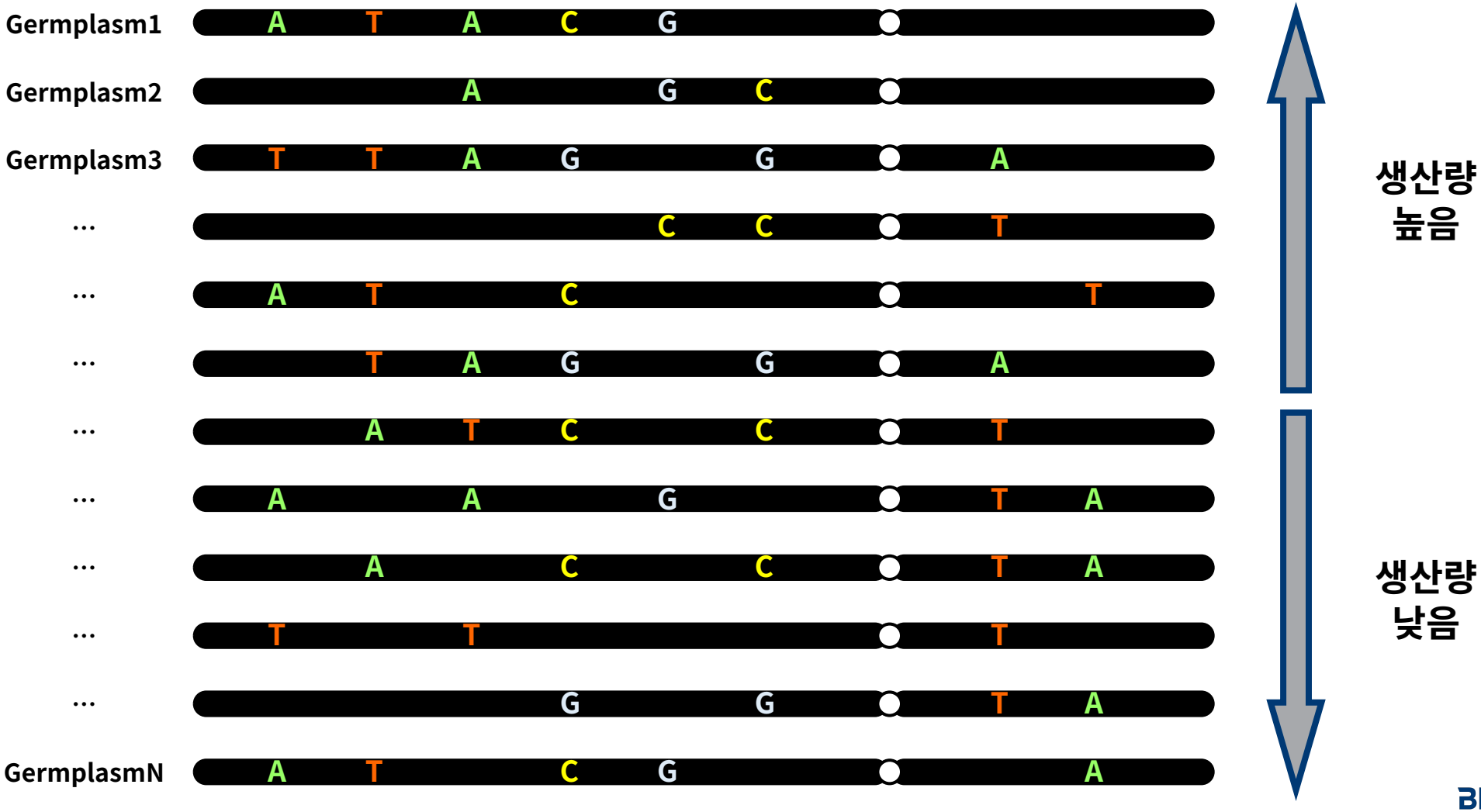
4. QTL mapping 결과에서 확인된 QTL 영역을 연관지도에 정리



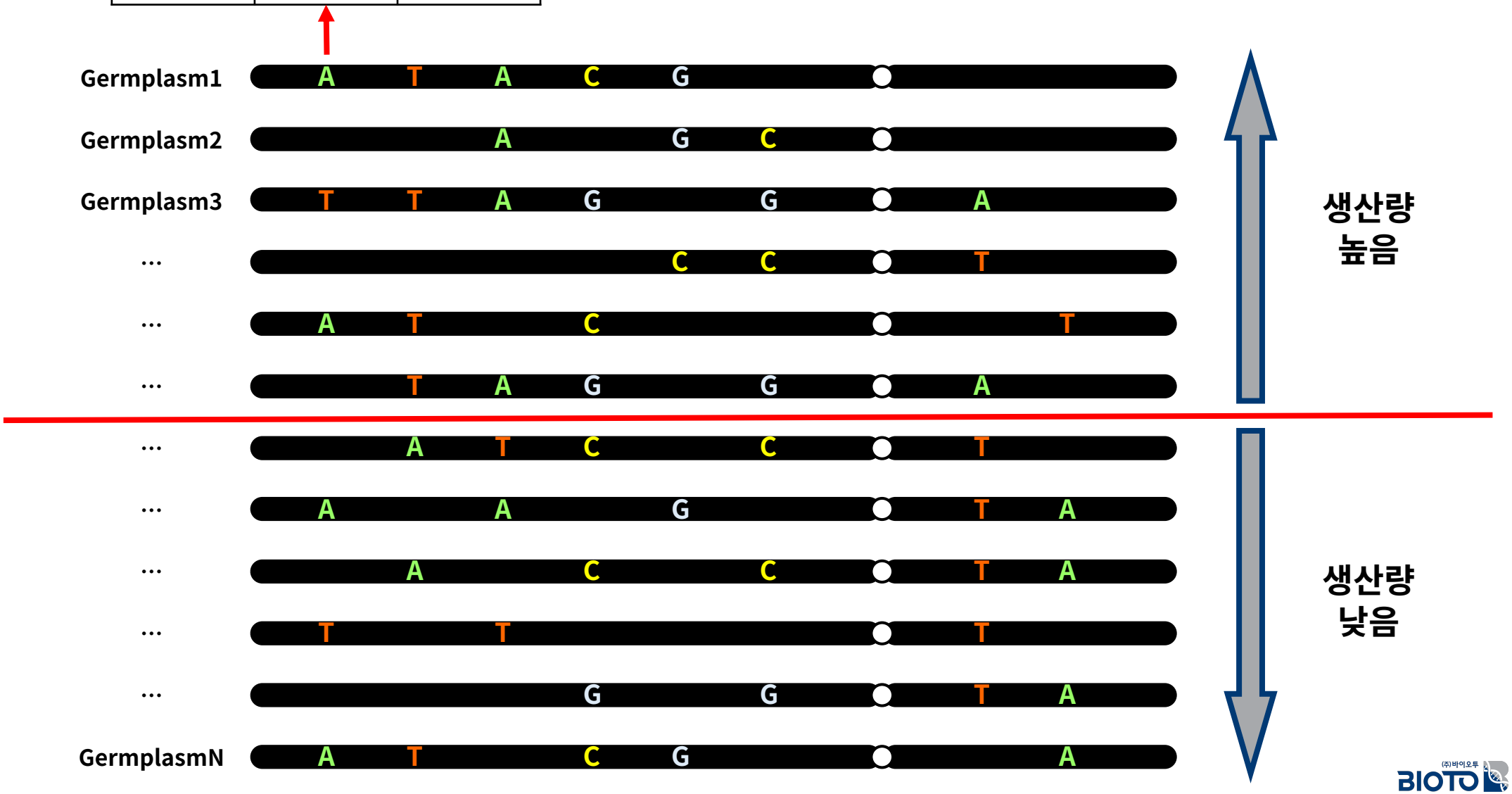
형질 연관영역을 결정하는 법



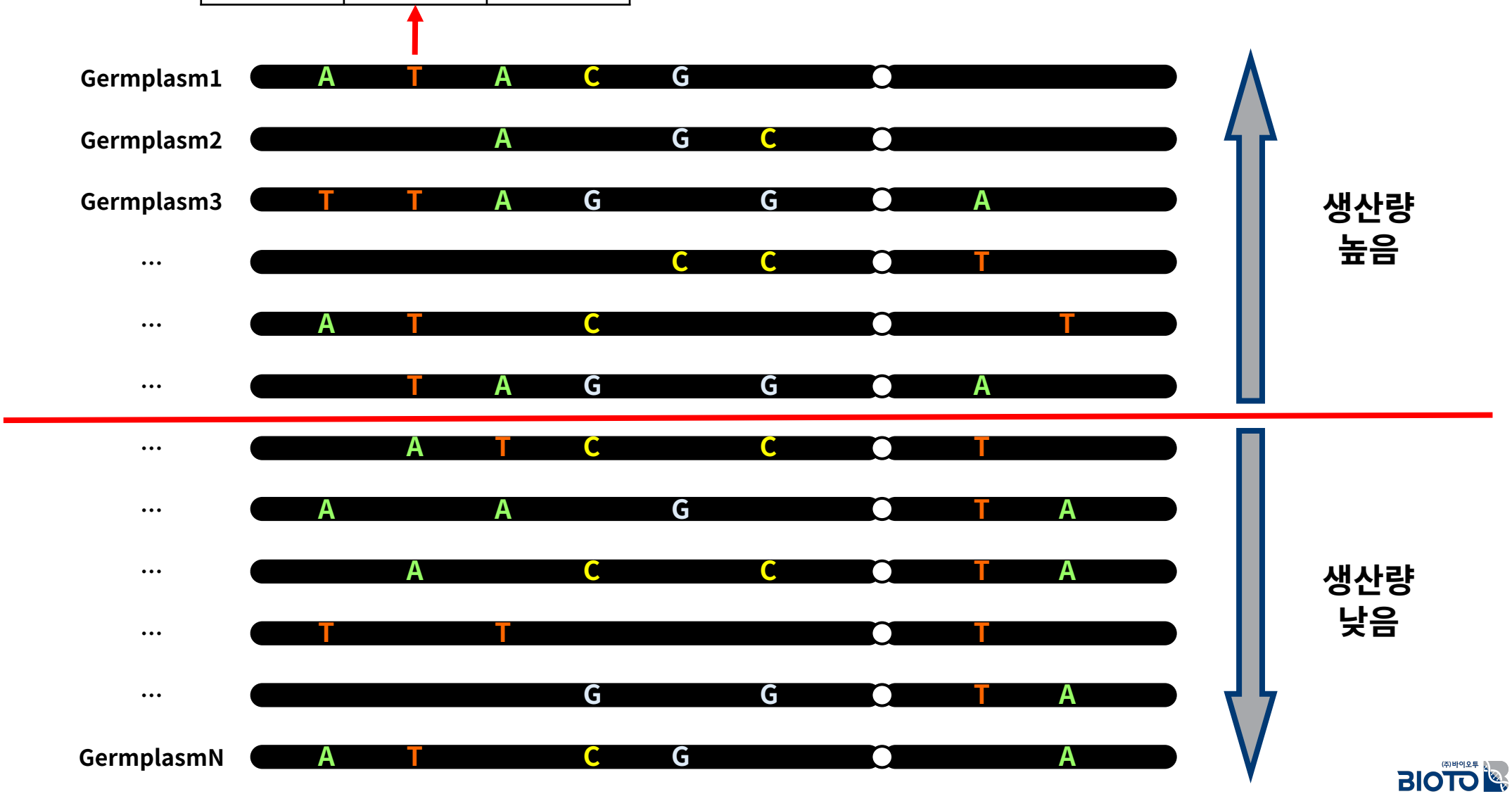



Germplasm1
A T A C G
Germplasm2
A G C
Germplasm3
T T A G G A
... C C T
... A T C T
... T A G G A
... A T C C T
... A A G T A
... A C C T A
... T T T
... G G T A
GermplasmN
A T C G A
생산량
높음
생산량
낮음

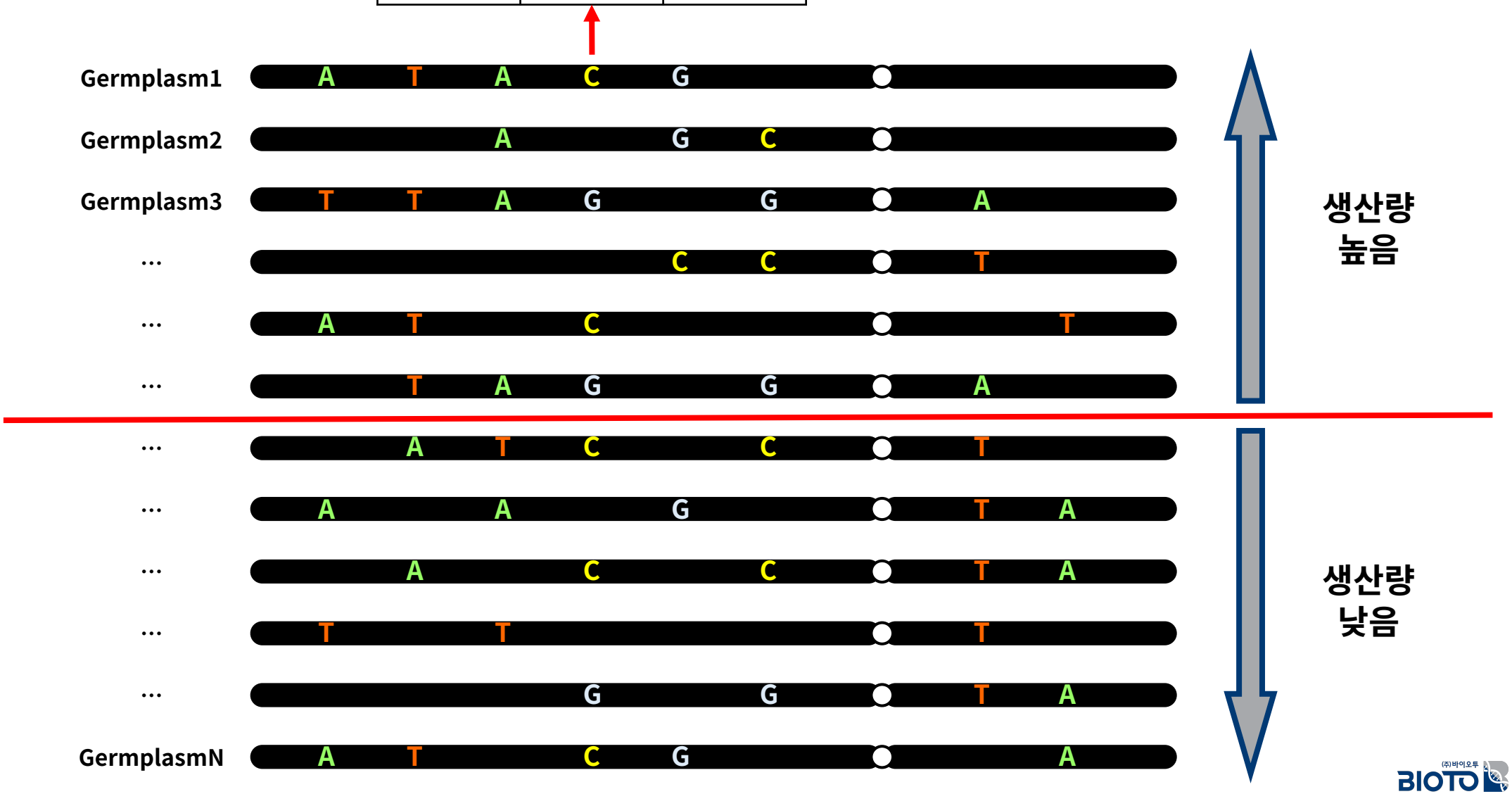
	생산량 높음	생산량 낮음
Ref. (A)	2	2
Alt. (T)	1	1



	생산량 높음	생산량 낮음
Ref. (A)	0	2
Alt. (T)	4	1



	생산량 높음	생산량 낮음
Ref. (C)	2	3
Alt. (G)	2	1



❖ **Marker 1**

	생산량 높음	생산량 낮음
Ref. (A)	2	2
Alt. (T)	1	1

❖ **Marker 2**

	생산량 높음	생산량 낮음
Ref. (A)	0	2
Alt. (T)	4	1

❖ **Marker 3**

	생산량 높음	생산량 낮음
Ref. (C)	2	3
Alt. (G)	2	1

Fisher's Exact Test

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

❖ **Marker 1**

	생산량 높음	생산량 낮음
Ref. (A)	2	2
Alt. (T)	1	1

- p-value = 1.0

❖ **Marker 2**

	생산량 높음	생산량 낮음
Ref. (A)	0	2
Alt. (T)	4	1

- p-value = 0.1429

❖ **Marker 3**

	생산량 높음	생산량 낮음
Ref. (C)	2	3
Alt. (G)	2	1

- p-value = 1.0

❖ **Example**

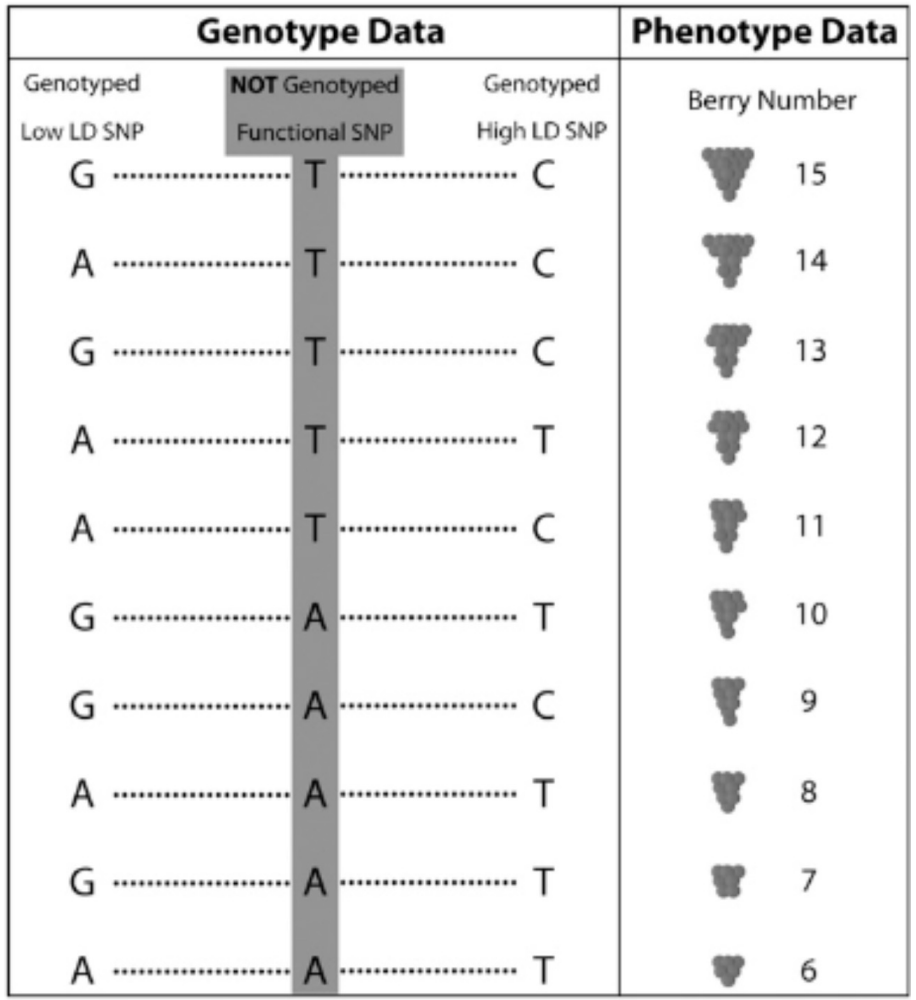
	생산량 높음	생산량 낮음
Ref. (A)	0	8
Alt. (T)	4	1

- p-value = 0.006993

II. 디지털육종

4. 전장 유전체 연관 분석
(Genome-Wide Association Studies)

표현형 정보와 유전자형 연결성 계산



- Myles *et al.* 2009

표현형 정보와 유전자형 연결성 계산

Genotype Data			Phenotype Data
Genotyped Low LD SNP	**NOT** Genotyped Functional SNP	Genotyped High LD SNP	Berry Number
G	T	C	15
A	T	C	14
G	T	C	13
A	T	T	12
A	T	C	11
G	A	T	10
G	A	C	9
A	A	T	8
G	A	T	7
A	A	T	6

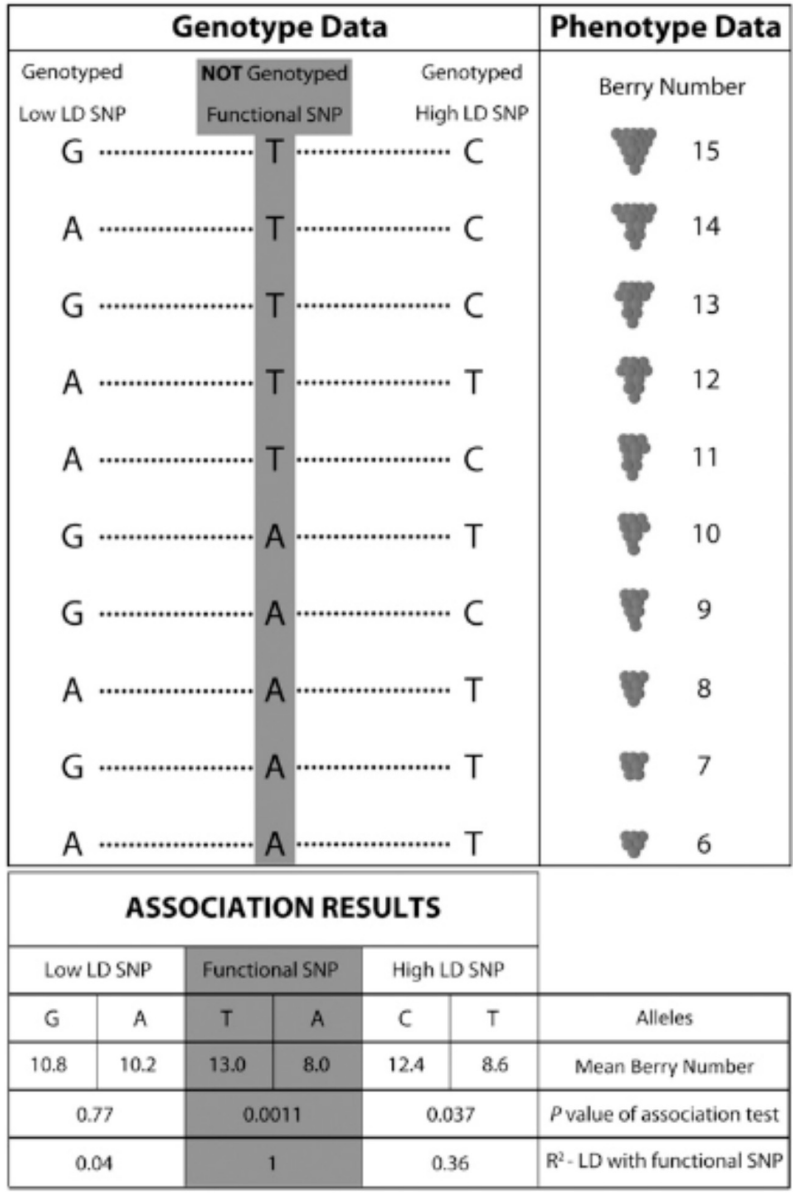
ASSOCIATION RESULTS

Low LD SNP		Functional SNP		High LD SNP		
G	A	T	A	C	T	Alleles
10.8	10.2	13.0	8.0	12.4	8.6	Mean Berry Number
0.77		0.0011		0.037		*P* value of association test
0.04		1		0.36		R^2 - LD with functional SNP

Myles *et al*. 2009

- Myles *et al*. 2009

Genome-Wide Association Studies (GWAS)



Genotype Data						Phenotype Data
Genotyped Low LD SNP		NOT Genotyped Functional SNP		Genotyped High LD SNP		Berry Number
G		T		C		15
A		T		C		14
G		T		C		13
A		T		T		12
A		T		C		11
G		A		T		10
G		A		C		9
A		A		T		8
G		A		T		7
A		A		T		6

ASSOCIATION RESULTS

Low LD SNP		Functional SNP		High LD SNP		
G	A	T	A	C	T	Alleles
10.8	10.2	13.0	8.0	12.4	8.6	Mean Berry Number
0.77		0.0011		0.037		*P* value of association test
0.04		1		0.36		R^2 - LD with functional SNP

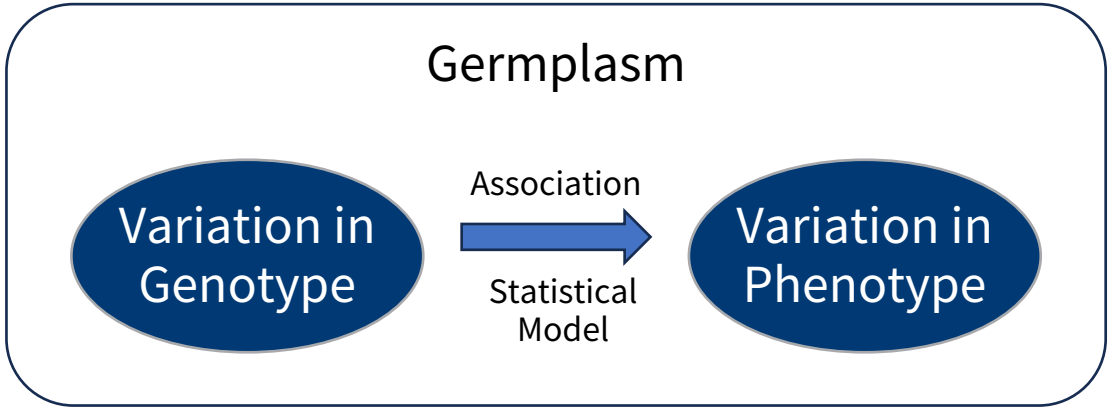
Myles *et al*. 2009

❖ GWAS의 목표:

유전형 variation과 표현형 variation 사이의
유의미한 연관성이 있는 유전적 마커 검색

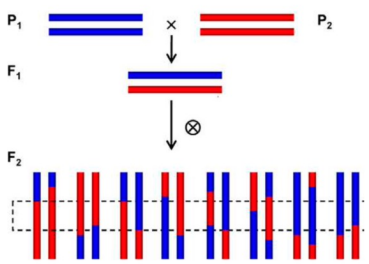
❖ GWAS의 key factors

- Germplasm (population)
- Genetic markers
- Statistical models
- Phenotype

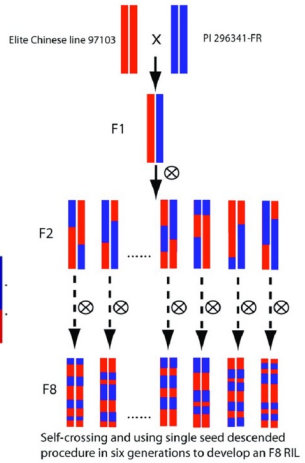


Linkage mapping과 Association Mapping의 비교

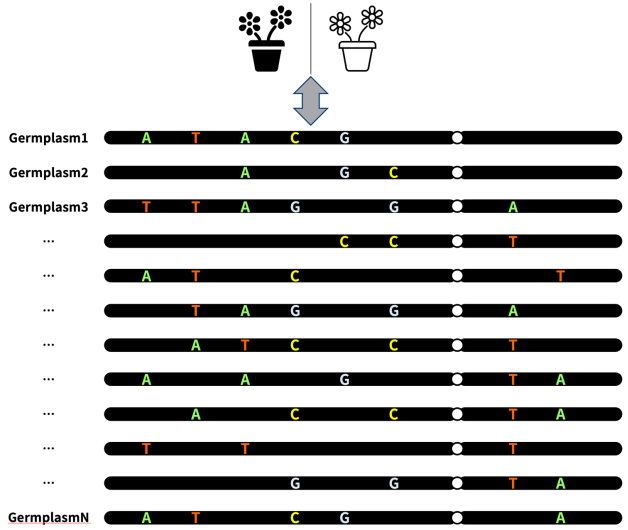
	Linkage Mapping	Association Mapping
장점	• 비교적 작은 집단 사용 • 낮은 density의 genetic marker • Fine-Mapping을 통해 QTL 분석 가능	• 높은 allele diversity • 교배 과정이 필요 없음. → 시간 절약 • Pedigree를 알고 있는 경우, 더 높은 mapping 해상도를 얻을 수 있음.
단점	• 낮은 allele diversity • 교배 과정에서 시간이 오래 걸림 • Recombination event가 낮을 경우, mapping 해상도가 낮음.	• phenotype간의 관계도에 의해 복잡성이 증가함. • QTL을 찾기 위해서는 매우 큰 집단의 사용 필요 • 높은 density의 genetic marker 요구 • 후보 유전자를 찾기 위해서는 후속 validation 필요



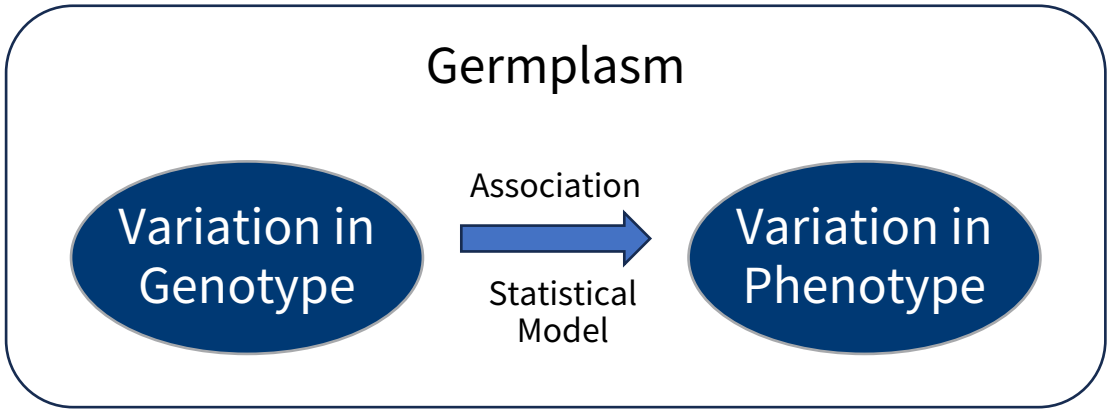
Zhu *et al.* (2008)



Ren *et al.* (2012)



GWAS 분석을 위해 고려할 사항



❖ GWAS의 key factors

- Germplasm (population)
- Genetic markers
- Statistical models
- Phenotype

일반 선형 모형
GLM (General Linear Model)

$$Y = SNP + Q\ (or\ PCs) + e$$

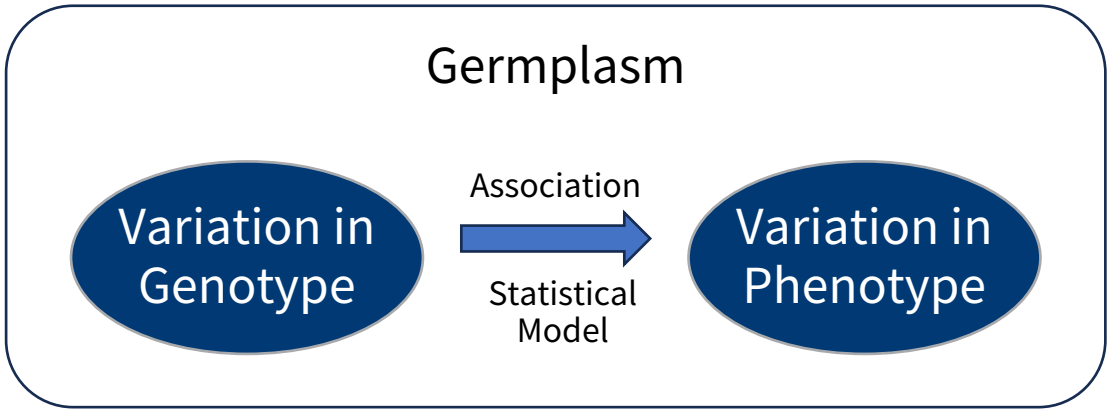
표현형 관측치 (observation)

유전자형 (fixed)

Population Structure (fixed)

오차 (error)

GWAS 분석을 위해 고려할 사항



❖ GWAS의 key factors

- Germplasm (population)
- Genetic markers
- Statistical models
- Phenotype
  - Kinship
  - 환경 변수

일반 선형 모형
GLM (General Linear Model)

$$Y = SNP + Q\ (or\ PCs) + e$$

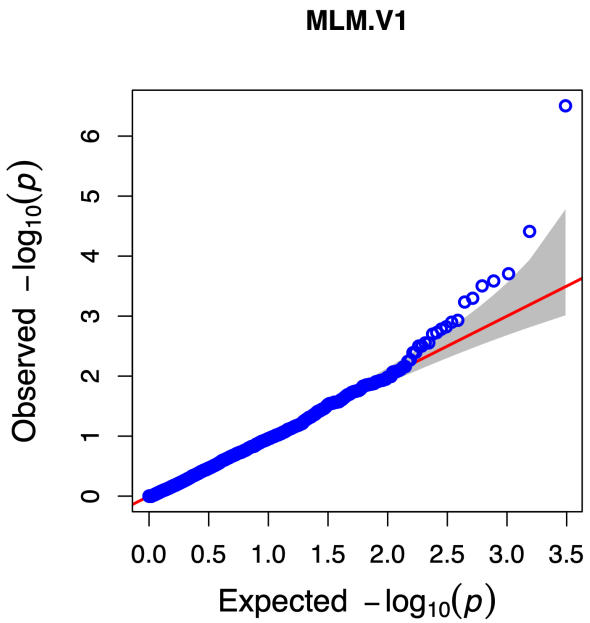
혼합 선형 모형
MLM (Mixed Linear Model)

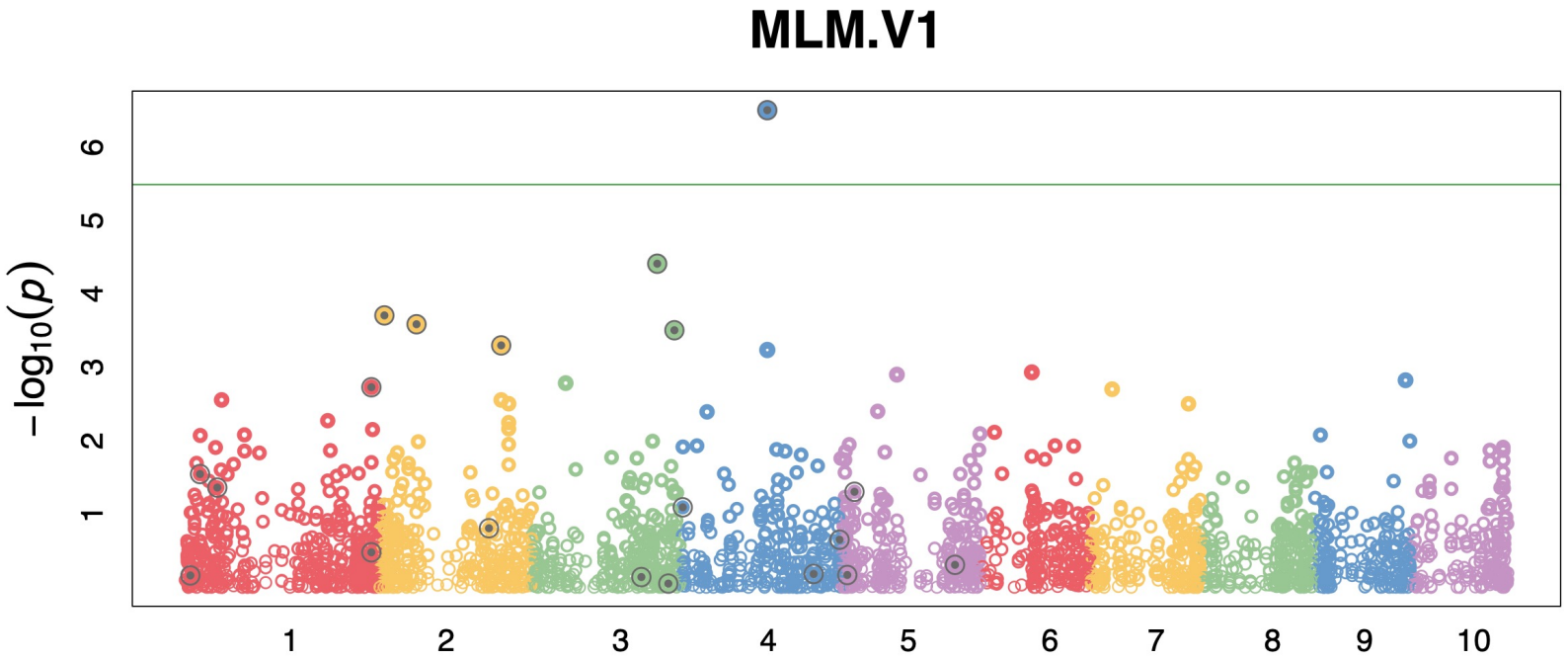
$$Y = SNP + Q\ (or\ PCs) + Kinship + e$$

(Yu *et al.* 2005, Nature Genetics)

- Y: 표현형 관측치 (observation)
- SNP: 유전자형 (fixed)
- $Q\ (or\ PCs)$: Population Structure (fixed)
- $Kinship$: Unequal Relatedness (Random)
- e: 오차 (error)

GWAS 주요 결과


MLM.V1
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0 1 2 3 4 5 6
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5


MLM.V1
$-\log_{10}(p)$
1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10

GWAS 분석을 위한 프로그램

GAPIT

Genomic Association and Prediction Integrated Tool

(Version 3)

Last updated on NOV 15, 2021

https://zzlab.net/GAPIT/

※ GAPIT 은 R 기반으로 R studio 로 실행

TASSEL - Trait Analysis by aSSociation, Evolution and Linkage

TASSEL is a software package used to evaluate traits associations, evolutionary patterns, and linkage disequilibrium. Strengths of this software include:

1. The opportunity for a number of new and powerful statistical approaches to association mapping such as a General Linear Model (GLM) and Mixed Linear Model (MLM). MLM is an implementation of the technique which our recently published Nature Genetics paper - Unified Mixed-Model Method for Association Mapping - which reduces Type I error in association mapping with complex pedigrees, families, founding effects and population structure.

2. An ability to handle a wide range of indels (insertion & deletions). Most software ignore this type of polymorphism; however, in some species (like maize), this is the most common type of polymorphism.

Read more at:
Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

TASSEL Version 5.0 *(Getting Started!)*
(Build: November 16, 2021 Requires: Java 1.8)

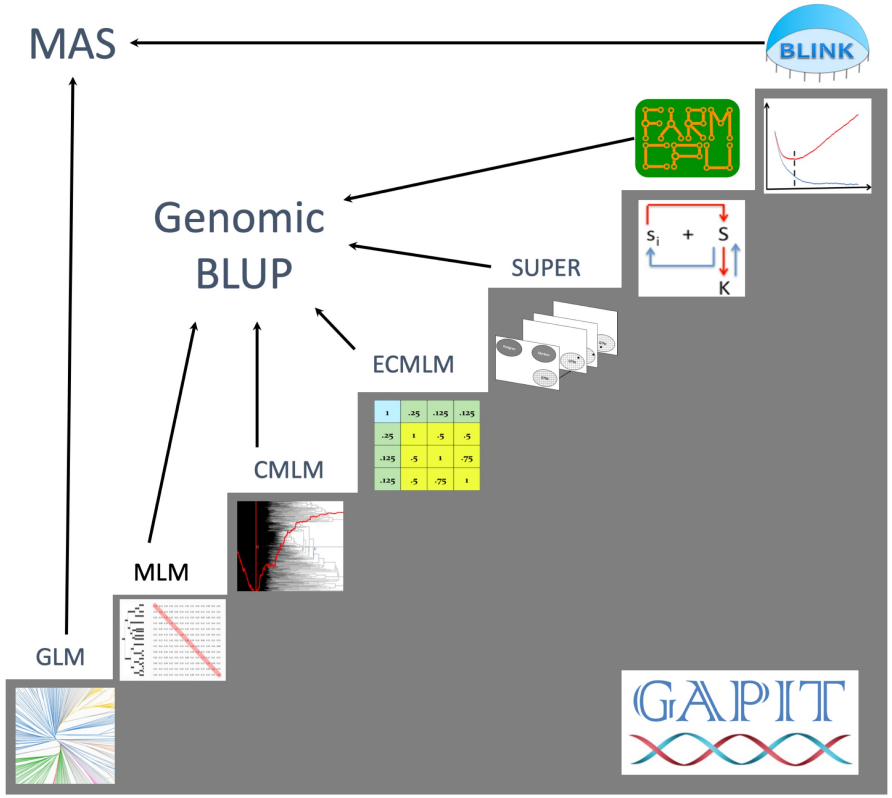
Tassel 5 Mac OS
Tassel 5 Windows 64 Bit
Tassel 5 Windows 32 Bit
Tassel 5 UNIX

Alignment Viewer

https://www.maizegenetics.net/tassel

※ TASSEL 프로그램 설치는 기본 default 값으로 진행

GAPIT 소개



Citation: Multiple statistical methods are implemented in GAPIT version 1, 2 and 3. Citations of GAPIT vary depending on methods and versions used in the analysis:

Method	Method paper	Version 1[1]	Version 2[2]	Version 3[3]
General Linear Model (GLM)	Price et al, 2006, *Nature Genetics*[4]	✓	✓	✓
Mixed Linear Model (MLM)	Yu et al, 2005, *Nature Genetics*[5]	✓	✓	✓
Compression MLM (CMLM)	Zhang et al, 2010, *Nature Genetics*[6]	✓	✓	✓
gBLUP	Zhang et al, 2007, *J. Anim. Science*[7]	✓	✓	✓
Enriched CMLM	Li et al, 2014, *BMC Biology*[8]		✓	✓
SUPER	Wang et al, 2014, *PLoS One*[9]		✓	✓
MLMM	Segura et al, 2012, *Nature Genetics*[10]			✓
FarmCPU	Liu et al, 2016, *PloS Genetics*[11]			✓
cBLUP and sBLUP	Wang et al, 2019, *Heredity*[12]			✓
BLINK	Huang et al, 2019, *GigaScience*[13]			✓

Note: These references are listed in section of Reference.

- https://zzlab.net/GAPIT/gapit_help_document.pdf

1.2 Getting Started

GAPIT is a package that is run in the R software environment, which can be freely downloaded from http://www.r-project.org or http://www.rstudio.com. There are two sources to install GAPIT package.

Zhiwu Zhang Lab website:

```
source("http://zzlab.net/GAPIT/GAPIT.library.R")
source("http://zzlab.net/GAPIT/gapit_functions.txt")
```

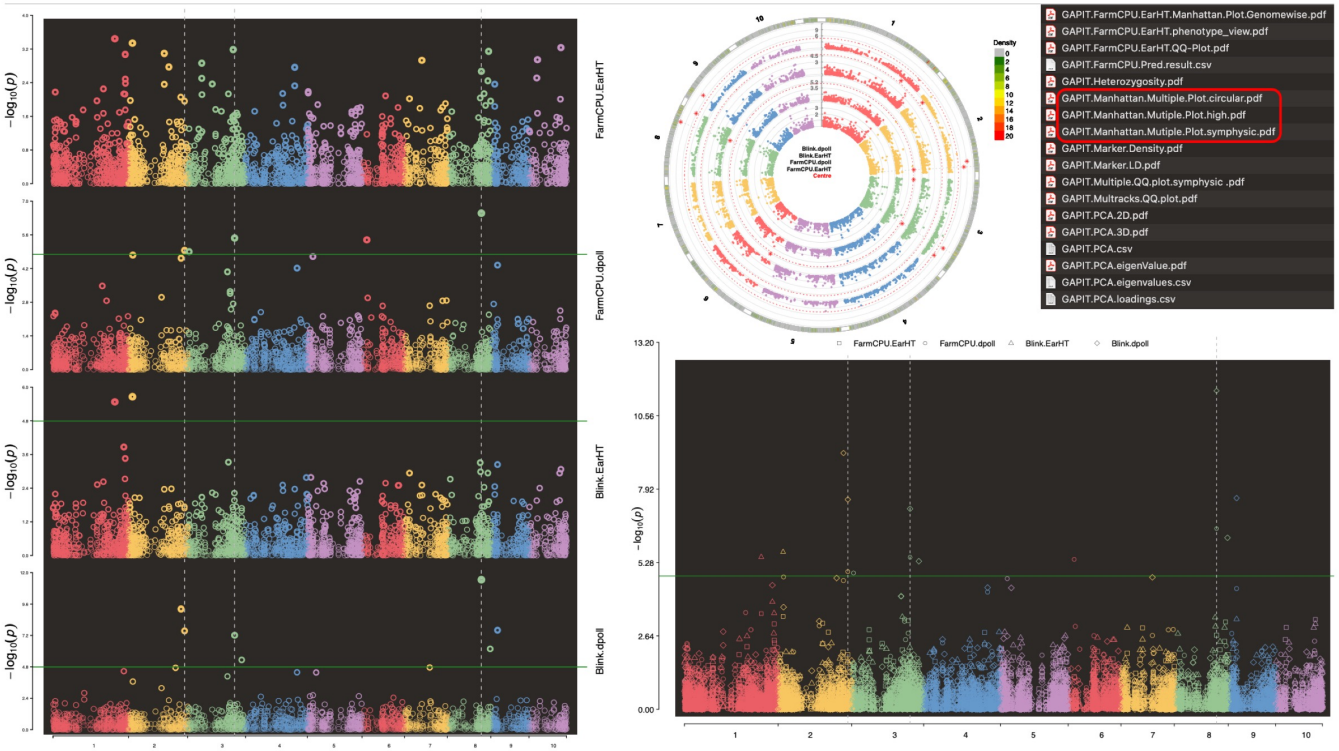
Or GitHub:

```
install.packages("devtools")
devtools::install_github("jiabowang/GAPIT3",force=TRUE)
library(GAPIT3)
```

The easiest way is to COPY/PASTE GAPIT tutorial script. Here are example code and outputs:

```
#Import data from Zhiwu Zhang Lab
myY  <- read.table("http://zzlab.net/GAPIT/data/mdp_traits.txt", head = TRUE)
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#GWAS
myGAPIT=GAPIT(
Y=myY[,c(1,2,3)], #fist column is ID
GD=myGD,
GM=myGM,
PCA.total=3,
model=c("FarmCPU", "Blink"),
Multiple_analysis=TRUE)
```



GWAS를 통해 이룩하고 싶은 것

SCIENTIFIC REPORTS

OPEN

A simulation-based breeding design that uses whole-genome prediction in tomato

Received: 15 October 2015
Accepted: 08 December 2015
Published: 20 January 2016

Eiji Yamamoto[1], Hiroshi Matsunaga[1], Akio Onogi[2], Hiromi Kajiya-Kanegae[2], Mai Minamikawa[2], Akinori Suzuki[2], Kenta Shirasawa[3], Hideki Hirakawa[3], Tsukasa Nunome[1], Hirotaka Yamaguchi[1], Koji Miyatake[1], Akio Ohyama[4], Hiroyoshi Iwata[2] & Hiroyuki Fukuoka[1]

Trait	Abbreviation	Trait category	h^2	Details
Percentage of fruit set (%)	PF	Yield	0.300	Percentage of flowers that reached fruit set in a plant
Total fruit weight (g/plant)	TFW	Yield	0.507	Total fruit weight per plant
Average fruit weight (g)	AFW	Yield	0.538	Average weight of all fruits from a plant
Percentage of marketable fruits (%)	PMF	Yield	0.401	Percentage of fruits of 100 g or more, without physiological disorders, in a plant
Total marketable fruit weight (g/plant)	TMFW	Yield	0.449	Total marketable fruit weight per plant
Average marketable fruit weight (g)	AMFW	Yield	0.469	Average weight of marketable fruits in a plant
Soluble solids content (°Brix)	SSC	Quality	0.600	Degree of Brix measured by saccharimeter (average of 4 marketable fruits per plant)
Pericarp colour	PCol	Quality	1.000	Colourless (pink tomato) and yellow (red tomato) pericarp counted as 0 and 1, respectively
Style scar	SS	Quality	0.492	Size of style scar on ripened fruit was scored based on the length of major axis. 0: <4 mm, 1: 4 ~ 10 mm, 2: >10 mm
Percentage of blossom-end rot fruits (%)	PBF	Physiological disorder of fruit	0.389	Percentage of blossom-end rot fruits in a plant
Percentage of irregular-shaped fruits (%)	PIF	Physiological disorder of fruit	0.478	Percentage of irregular-shaped fruits in a plant
Percentage of cracked fruits (%)	PCF	Physiological disorder of fruit	0.338	Percentage of cracked fruits in a plant
Percentage of small fruits (%)	PSF	Physiological disorder of fruit	0.372	Percentage of fruits less than 100 g in a plant
Leaf length (mm)	LL	Others	0.492	Length of a leaf under the first truss
Leaf width (mm)	LW	Others	0.464	Width of a leaf under the first truss
Stem width (mm)	SW	Others	0.377	Width of a stem at the position of the first truss
Height to the first truss (cm)	H1T	Others	0.370	Height of the first truss from ground
Number of flowers	NFlo	Others	0.382	Number of flowers after defloration (maximum number of flowers is 6 per truss)
Days to flowering	DTF	Others	0.106	Number of days from seeding to first flower
Number of leaves under the first truss	NL1T	Others	0.389	Number of true leaves under the first truss

Table 1. List of traits analysed in this study.

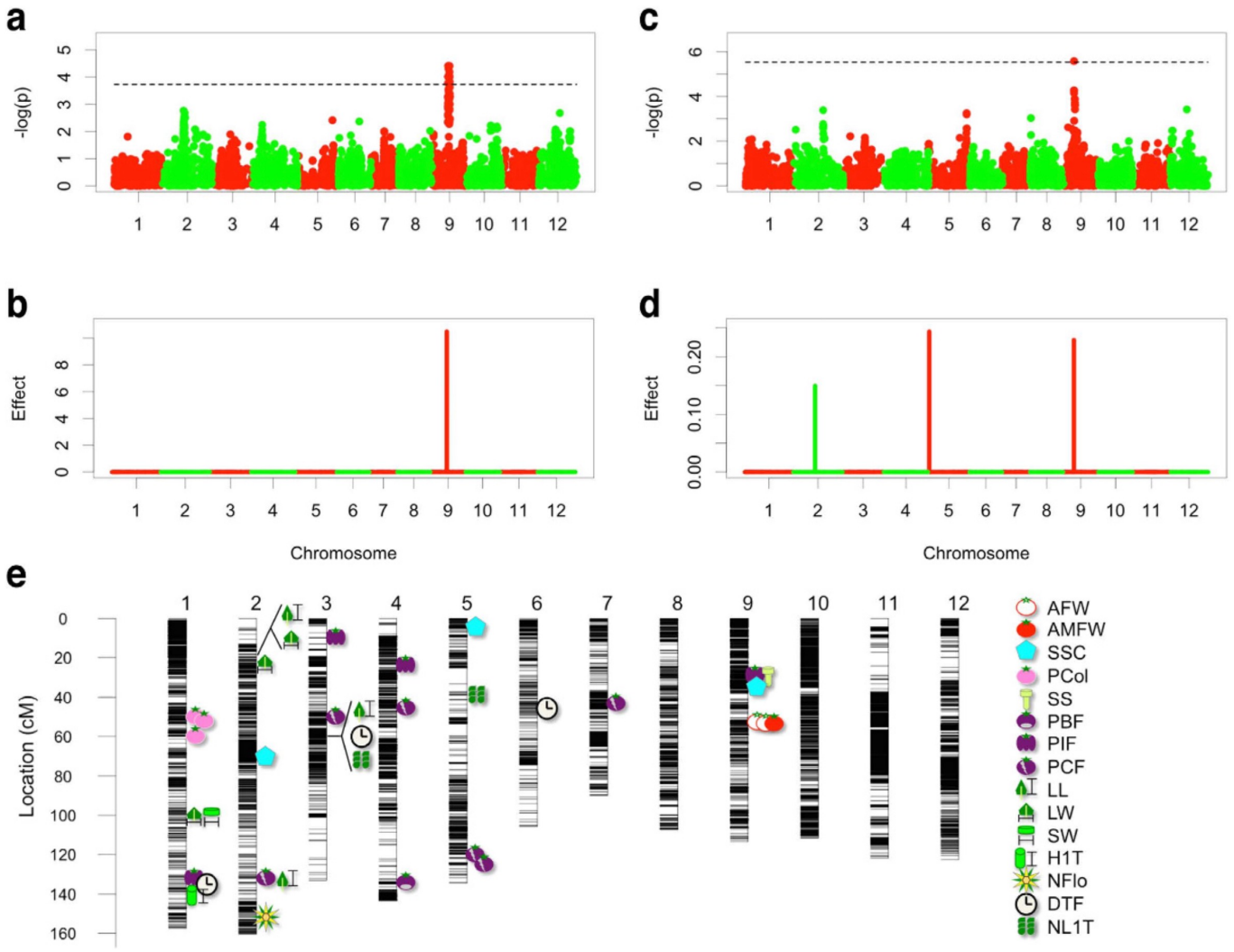
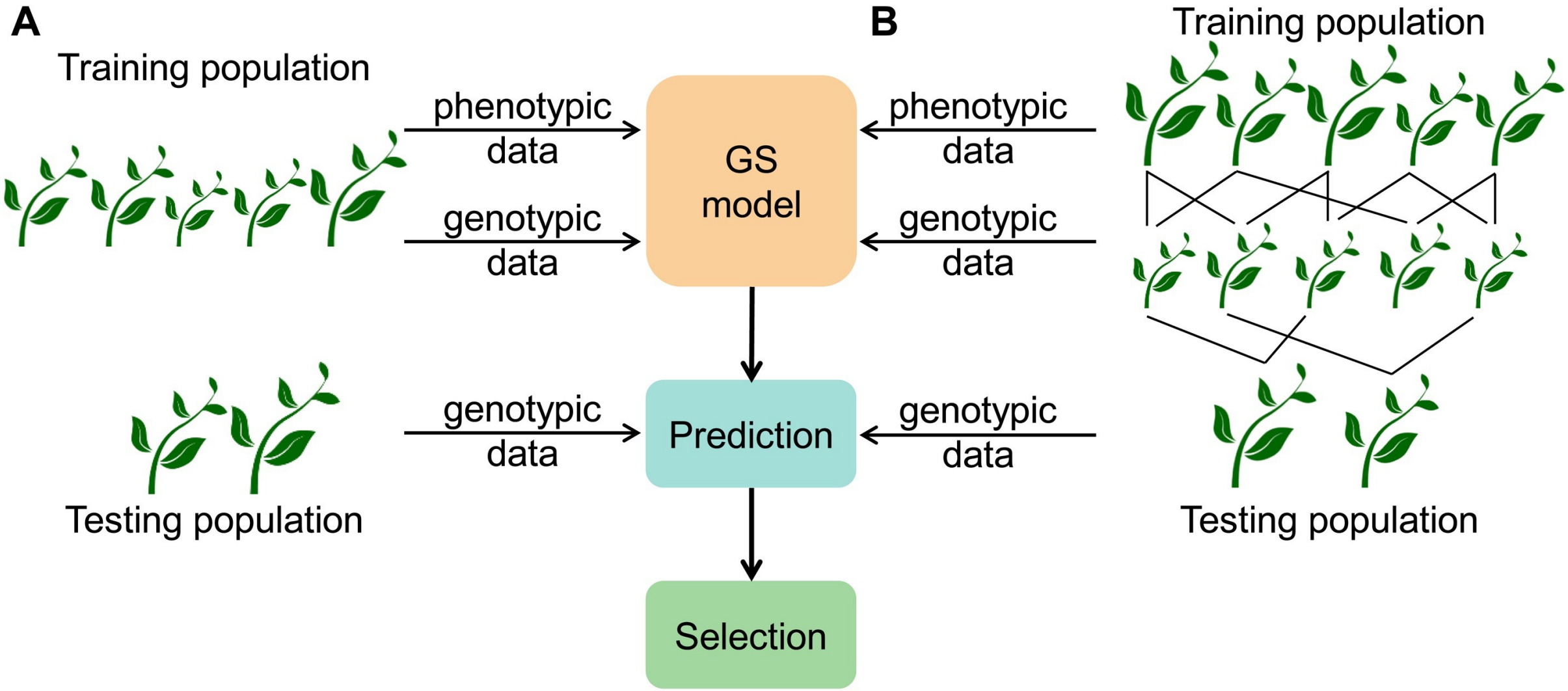


Figure 2. Summary of the genome-wide association study (GWAS) results. (**a,b**) GWAS results for average marketable fruit weight. (**c,d**) GWAS results for soluble solids content. (**a,c**) Manhattan plots for mixed linear model. The horizontal dashed lines indicate the threshold obtained from the 5% false discovery rate. (**b,d**) Posterior means of all marker effects for extended Bayesian Lasso. The values were obtained by using hyperparameter $\theta = 0.0001$. (**e**) Chromosomal distribution of significant associations detected by the GWAS. AFW, average fruit weight; AMFW, average marketable fruit weight; SSC, soluble solids content; PCol, pericarp colour; SS, style scar; PBF, percentage of blossom-end rot fruits; PIF, percentage of irregular-shaped fruits; PCF, percentage of cracked fruits; LL, leaf length; LW, leaf width; SW, stem width; H1T, height to the first truss; NFlo, number of flowers; DTF, days to flowering; NL1T, number of leaves under the first truss. See Table 1 for details.

Genomic Selection



- Tong H & Nikoloski Z. J Plant Physiol. 2021

강의 요약

- NGS로 얻은 변이 정보를 이용하여 형질연관마커, MAS, MAB, 순도검정, 원산지 구분, QTL-mapping, GWAS와 같은 다양한 분석이 가능

- 전장 유전체 상에 존재하는 변이 정보를 이용하여 다양한 분석을 수행할 수 있음.
  - 형질, 특성에 차이가 나는 개체/집단의 정보를 탐색하는 마커 개발
  - 개체 간의 관계 및 구조를 분석하는 유연관계 분석
  - 교배/육성 집단과 형질 간의 양적유전자 좌를 탐색하는 QTL 분석
  - 유전자원과 형질간의 연관성을 분석하는 GWAS 분석

- GWAS (Genome-Wide Association Study) 분석은 전장 유전체에 대한 대량의 유전적 변이와 표현형 간의 관계를 조사하는 유전학적 분석 방법으로 GAPIT, TASSEL, PLINK와 같은 프로그램으로 수행할 수 있음.

- GAPIT에는 GLM, MLM, CMLM, gBLUP, Enriched CMLLM, SUPER, MLMM, FarmCPU, cBLUP and sBLUP, BLINK와 같은 다양한 분석 방법론을 적용하여 분석에 사용할 수 있음.

Q & A

강의를 경청해 주셔서 감사합니다.

대전시 유성구 테크노2로 187, B동 412호

bi@bioto.co.kr

042-710-0077

070-7585-5344