초고성능컴퓨터 활용 교육

전장 유전체 연관분석 (이론)

(주)바이오투, 최준경

2023-07-19

목 차

NGS를 이용한 연구 (요약)

최신 NGS 장비의 능력

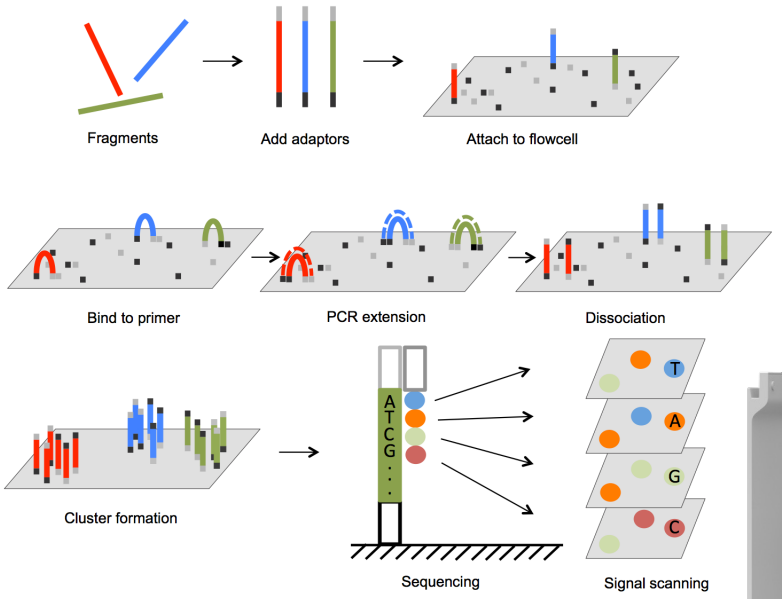


Table 1: NovaSeq 6000 System flow cell specifications

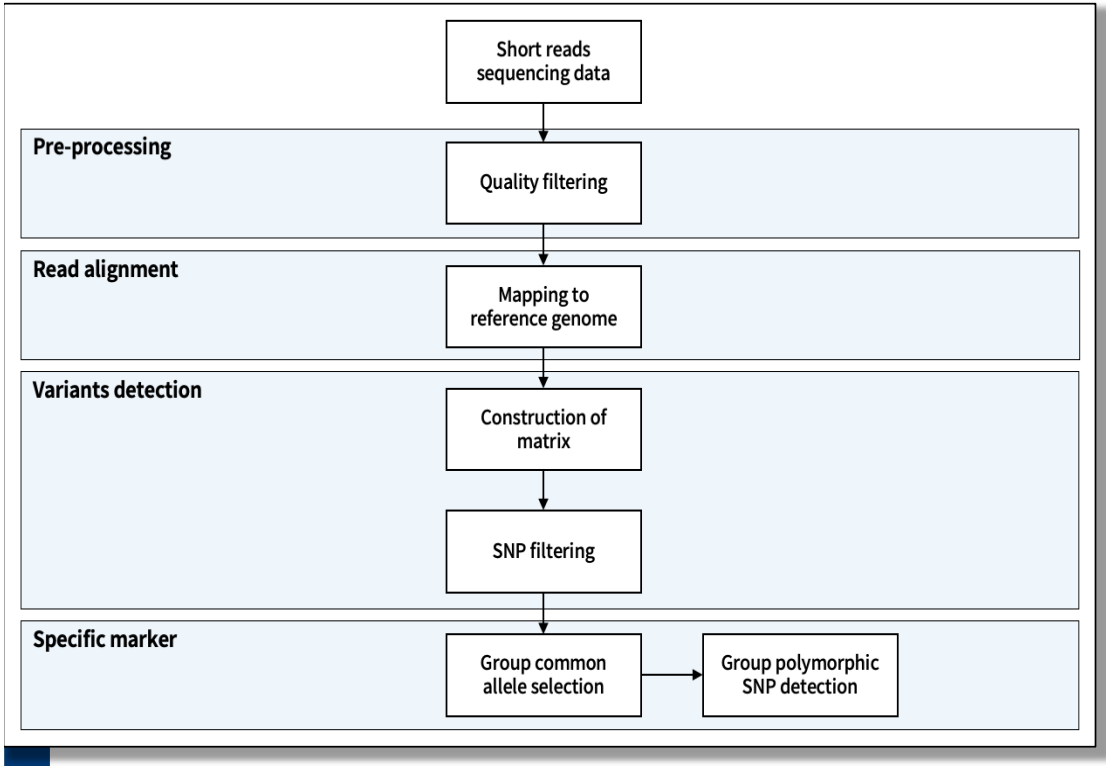
Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
1 × 35 bp	N/A			280-350 Gb
2 × 50 bp	65-80 Gb		Gb	N/A
2 × 100 bp	134-167 Gb	Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 × 250 bp	325-400 Gb	N/A	N/A	N/A
Single reads CPF	0.65-0.8B	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

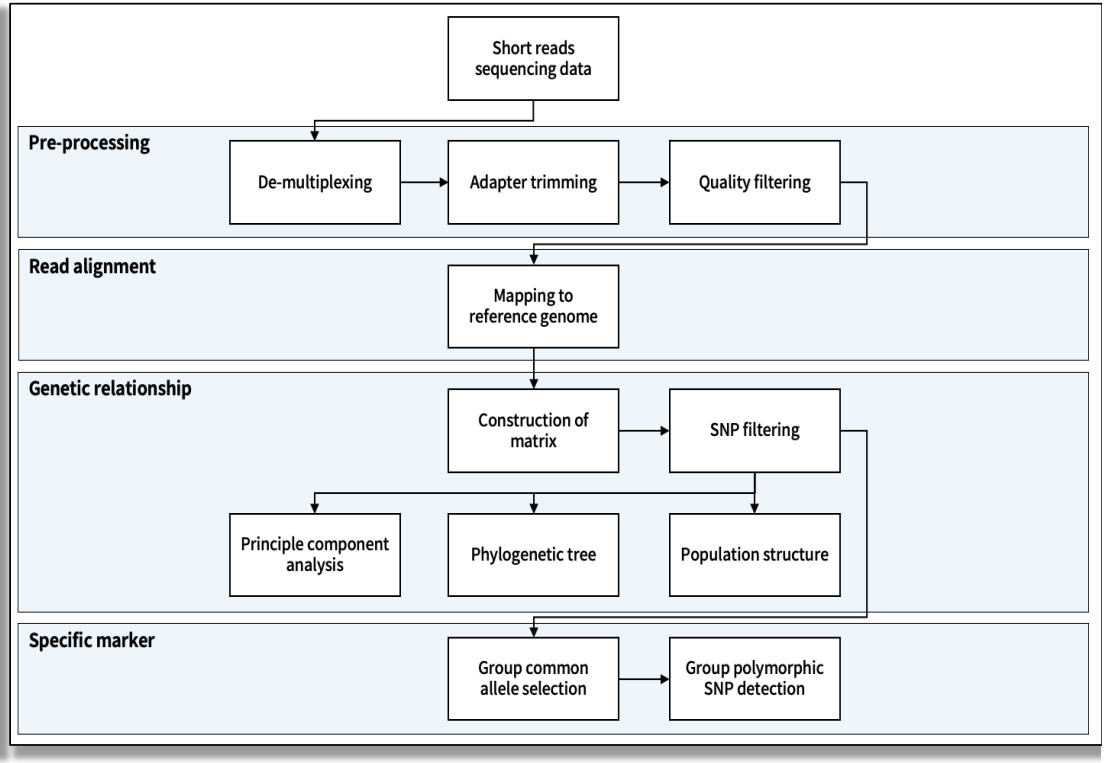
- https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html

NGS 분석 파이프라인

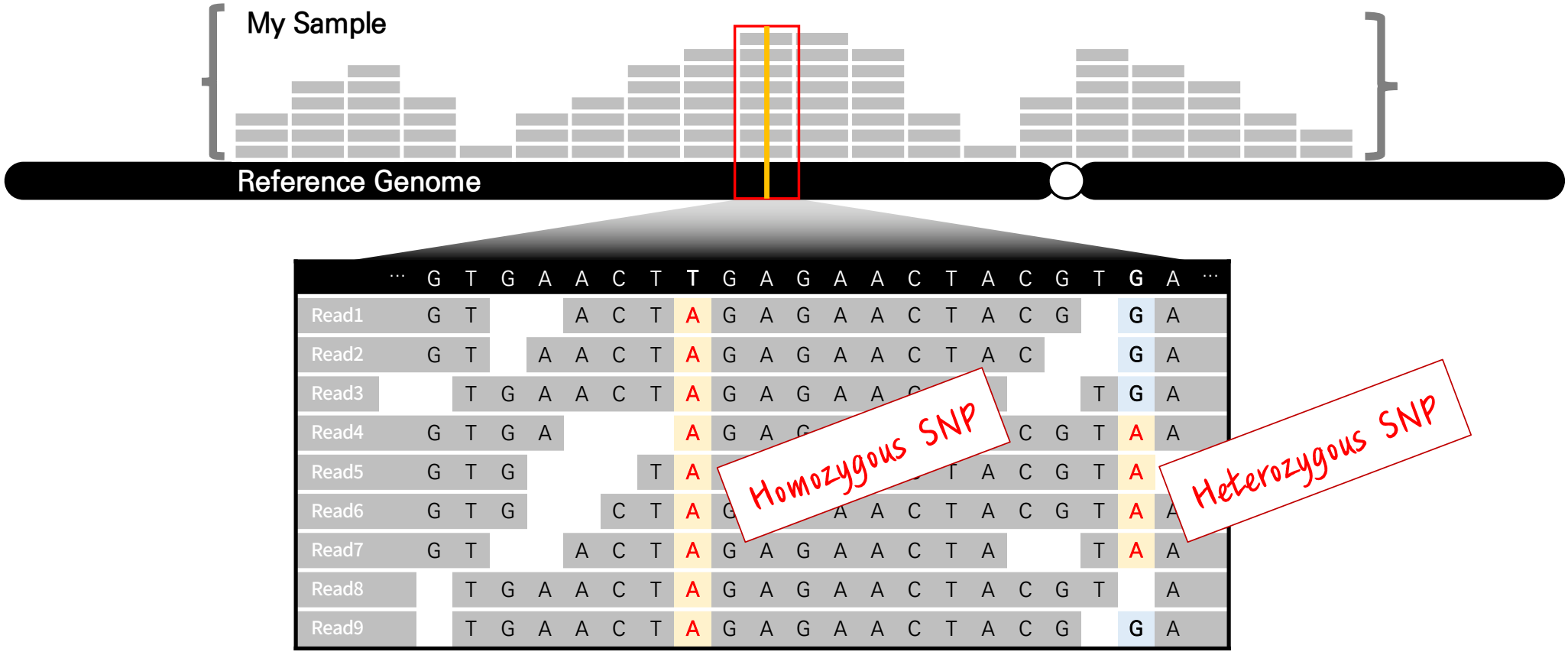
❖ WGS (Whole Genome Sequencing) 파이프라인



❖ GBS (Genotyping-By-Sequencing) 파이프라인

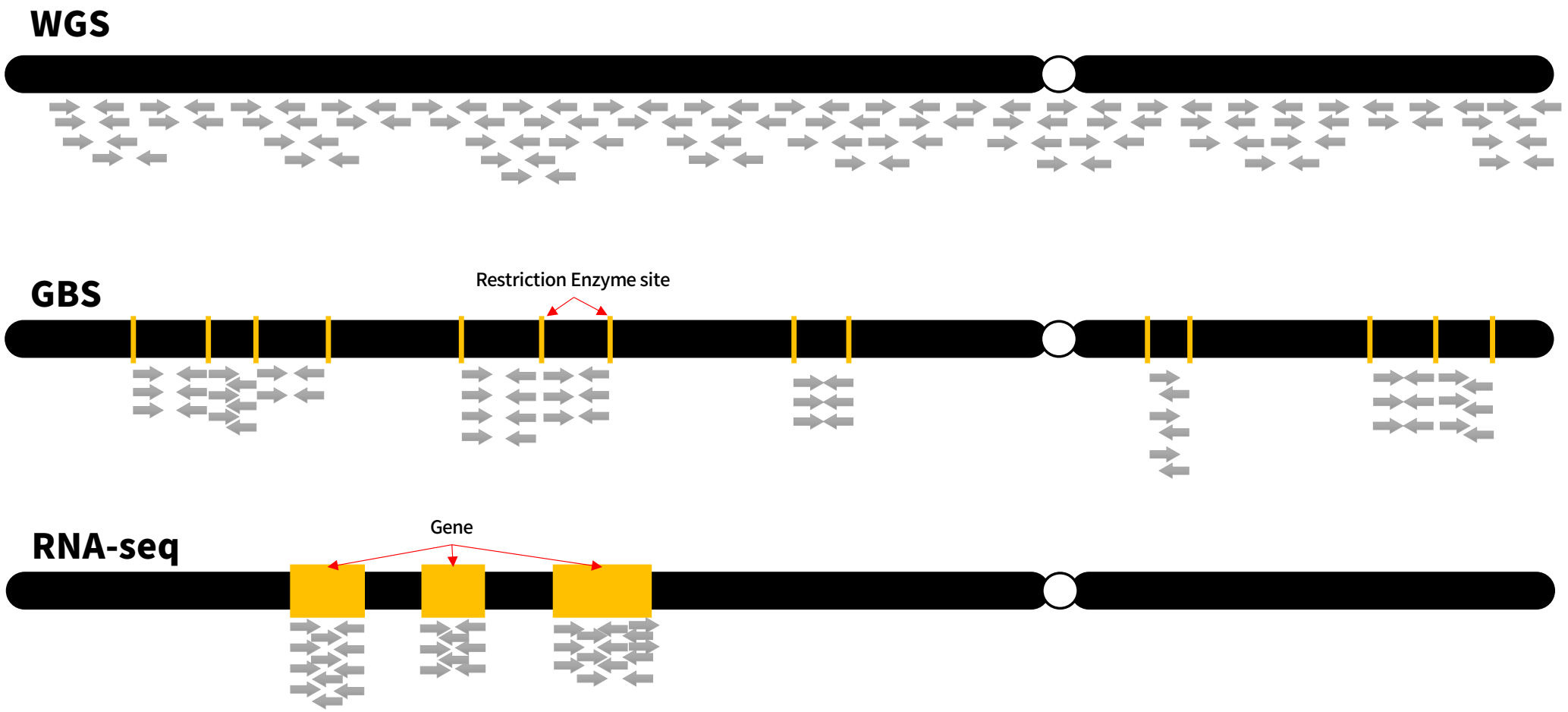


SNP 변이

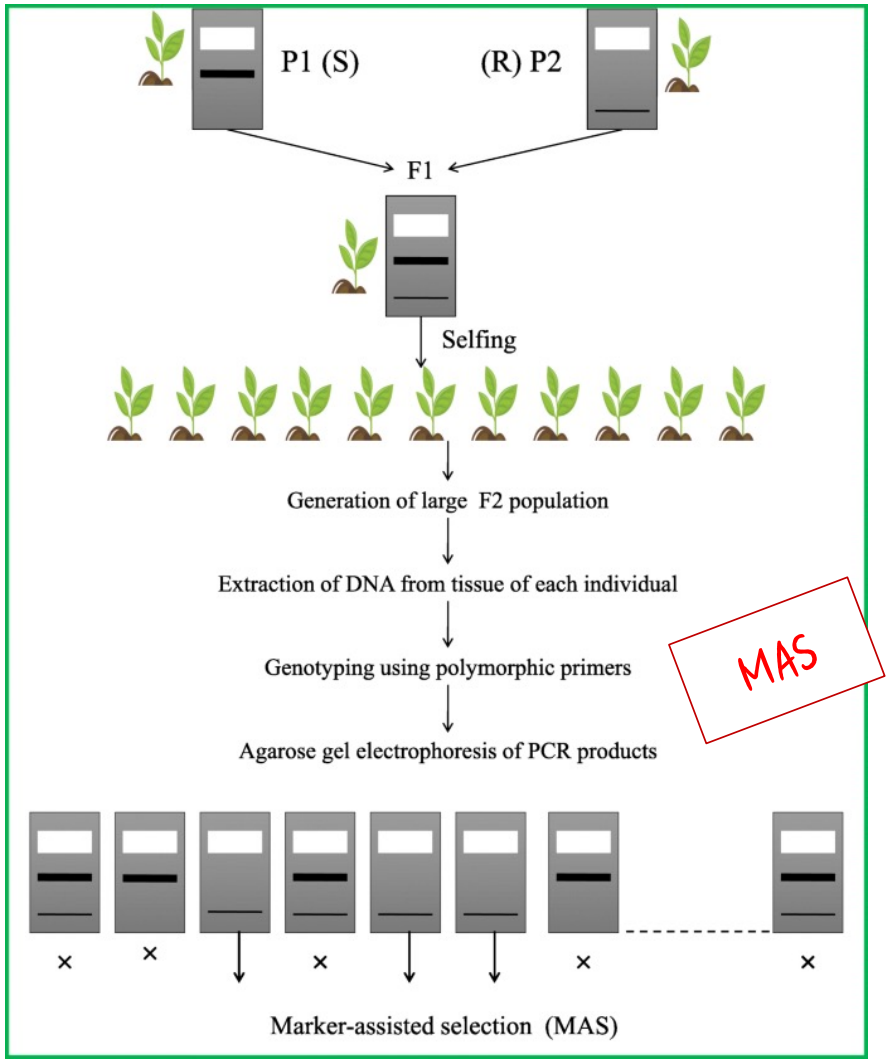


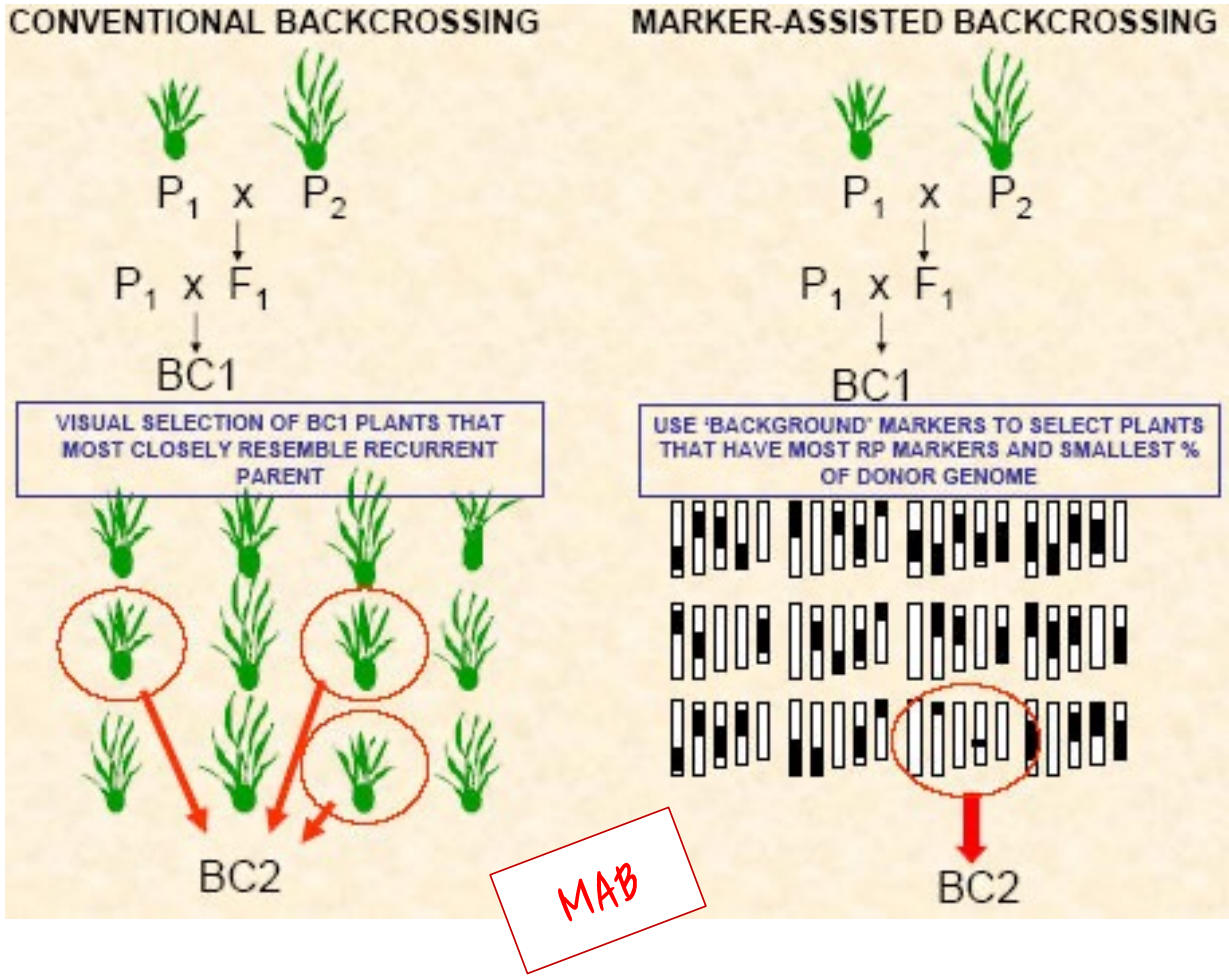
- Sachidanandam R et al. Nature. 2001;409(6822):928-933

WGS, GBS, RNA-seq의 sequencing 기법에 따른 차이



NGS로 대량 형질연관 SNP 획득 → 분자육종

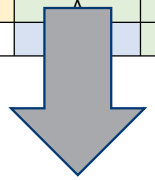




- https://jgeb.springeropen.com/articles/10.1186/s43141-021-00231-1
- http://www.knowledgebank.irri.org/ricebreedingcourse/Marker_assisted_breeding.htm

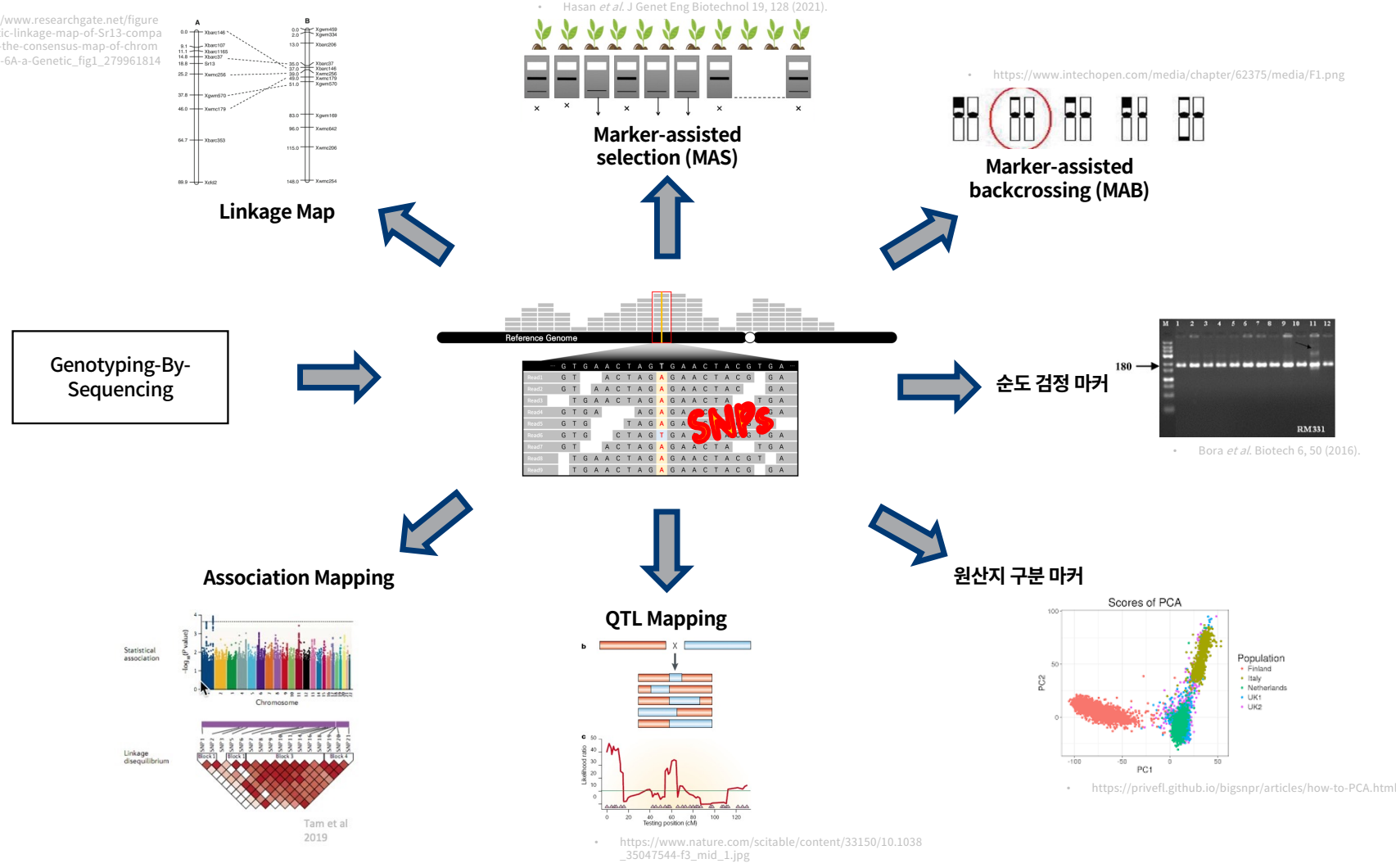
Filtering을 통한 최종 산물

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	G	G	A	-



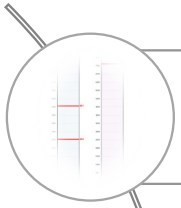
CHR	scaffold0777	scaffold0718	scaffold0776	scaffold0788
POS	175224	586757	85221	601667
Sample1	T	G	C	-
Sample2	T	G	A	G
Sample3	T	G	A	A
Sample4	T	A	A	G
Sample5	G	A	C	G
Sample6	T	A	A	A
Sample7	G	-	A	-

변이를 통해 할 수 있는 것

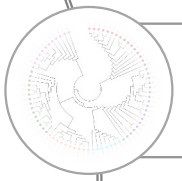

https://www.researchgate.net/figure/Genetic-linkage-map-of-Sr13-compared-to-the-consensus-map-of-chromosome-6A-a-Genetic_fig1_279961814
Linkage Map
Hasan et al. J Genet Eng Biotechnol 19, 128 (2021).
Marker-assisted selection (MAS)
https://www.intechopen.com/media/chapter/62375/media/F1.png
Marker-assisted backcrossing (MAB)
Genotyping-By-Sequencing
Reference Genome
SNPs
순도 검정 마커
180
RM331
Bora et al. Biotech 6, 50 (2016).
Association Mapping
Statistical association
Chromosome
Linkage disequilibrium
Tam et al 2019
QTL Mapping
Likelihood ratio
Testing position (cM)
https://www.nature.com/scitable/content/33150/10.1038_35047544-f3_mid_1.jpg
원산지 구분 마커
Scores of PCA
PC1
PC2
Population
Finland
Italy
Netherlands
UK1
UK2
https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

전장 유전체 변이를 이용한 다양한 응용 분석

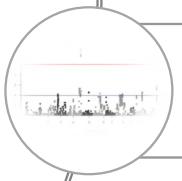
전장 유전체 변이를 이용한 응용 분석의 종류



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**

응용 1 – 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)

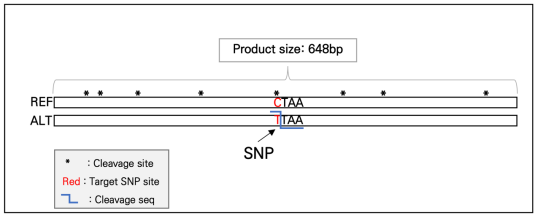
❖ **BARCODE SNP 예시**

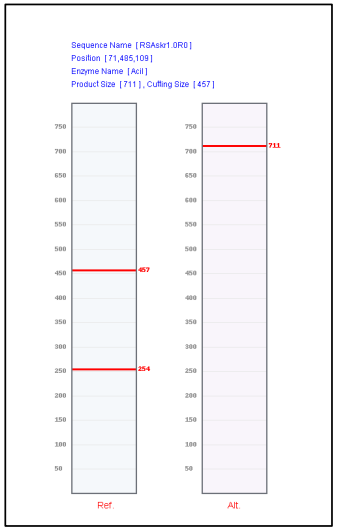
Barcode no.	1	2	3	4
Barcode-SNP type	abaabb	abaaba	ababbb	abbaaa
Barcode-SNP candidates	32	12	144	2
Sample1	a	a	a	a
Sample2	b	b	b	b
Sample3	a	a	a	b
Sample4	a	a	b	a
Sample5	b	b	b	a
Sample6	b	a	b	a

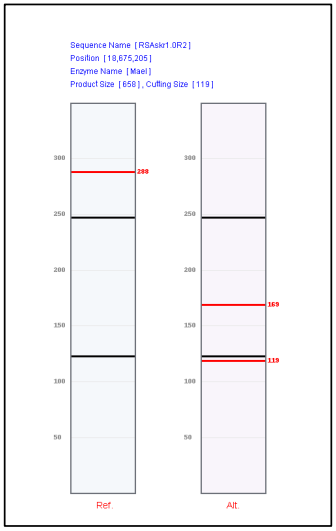
II. 개발 가능한 마커 종류

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

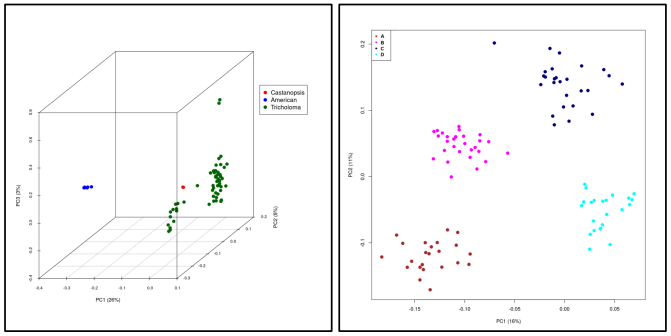
1. **주성분 분석 (PCA)**
2. **주좌표 분석 (PCoA)**
3. MDS plot
4. **계통수 분석 (Phylogenetic tree)**
5. **구조 분석 (STRUCTURE)**

유연관계 분석

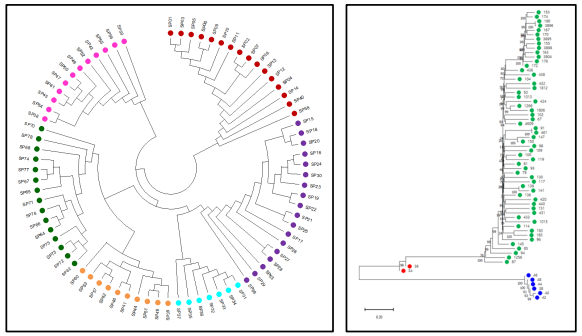
6. F_{ST} 분석
7. Nucleotide diversity 분석
8. Pairwise distance 분석
9. Heterozygosity 분석
10. AMOVA 분석

유전 다양성 분석

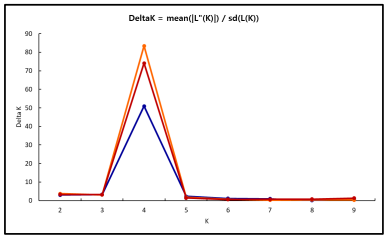
❖ 주성분 분석

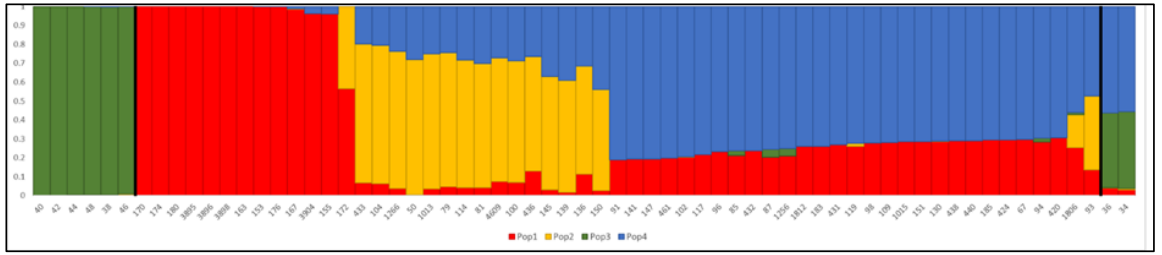


❖ 계통수 분석



❖ 구조 분석

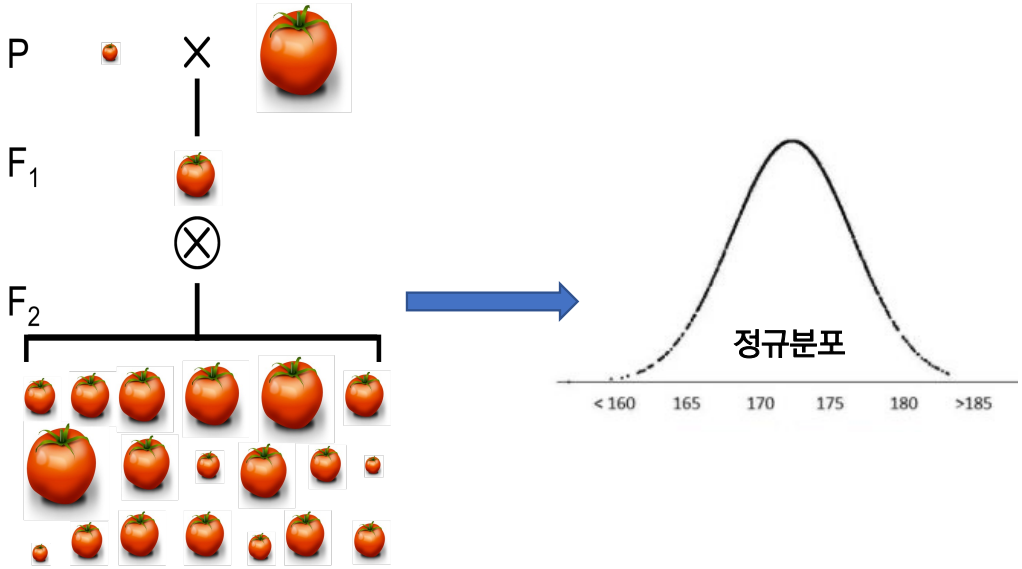




응용 3 – QTL mapping

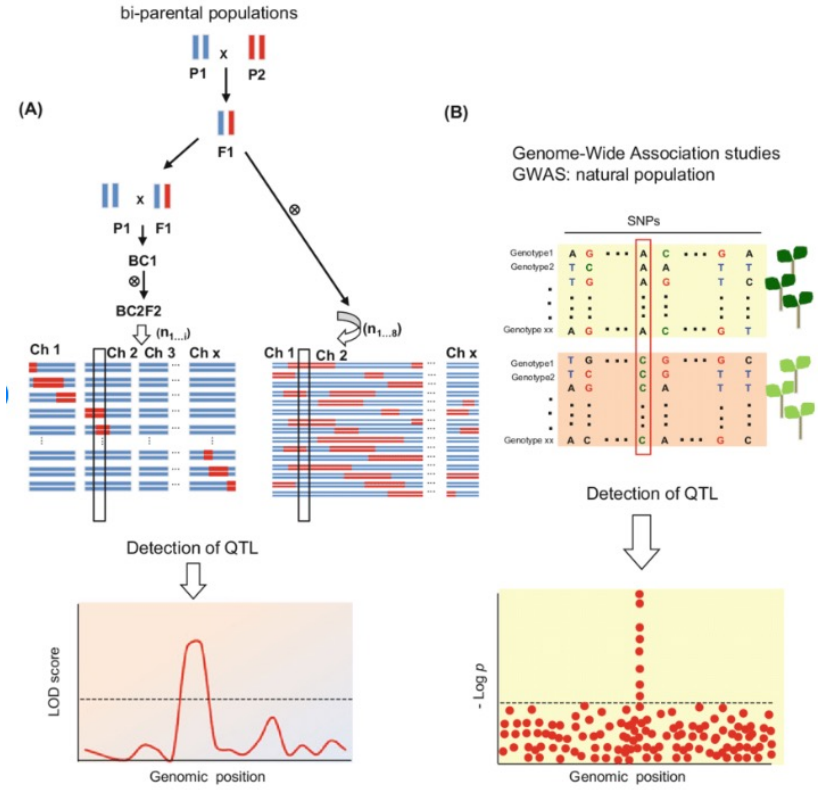
I. QTL mapping 분석이란?

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

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

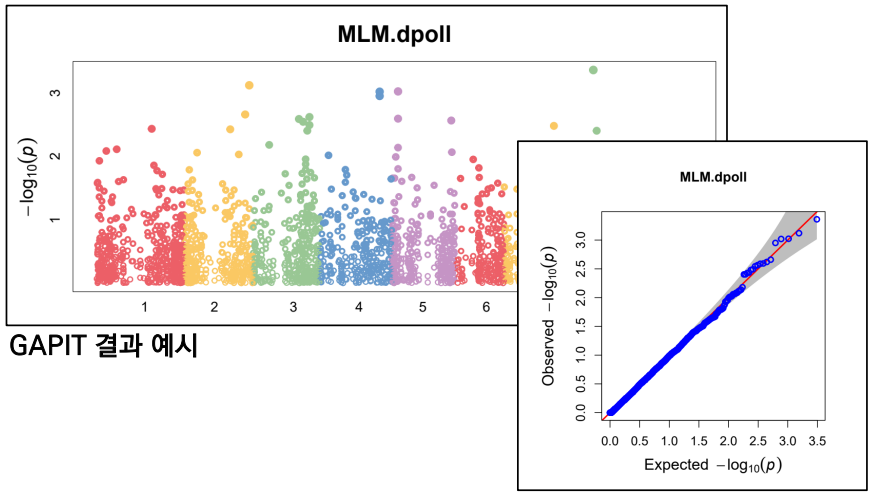
R package를 이용하여 분석하며, 기본 선형모델, 혼합 선형모델 등에 특화되어 분석할 수 있으며 간단한 명령어로 분석이 가능함.

b. TASSEL

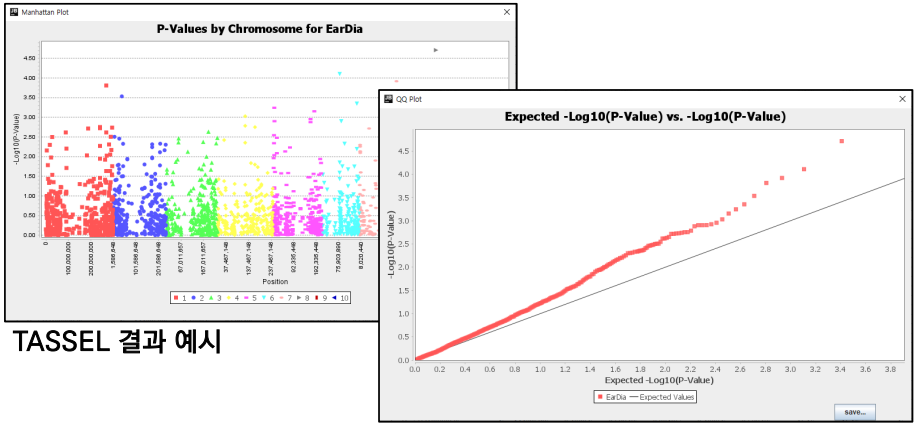
Windows 기반의 분석 프로그램으로 사용자의 편의성이 좋은 분석 프로그램으로 다양한 선형모델을 지원함.

c. PLINK

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



GAPIT 결과 예시

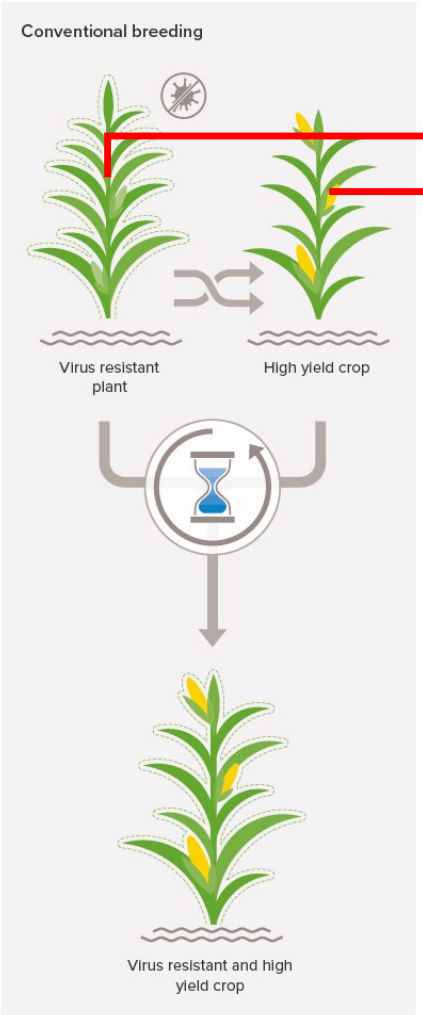


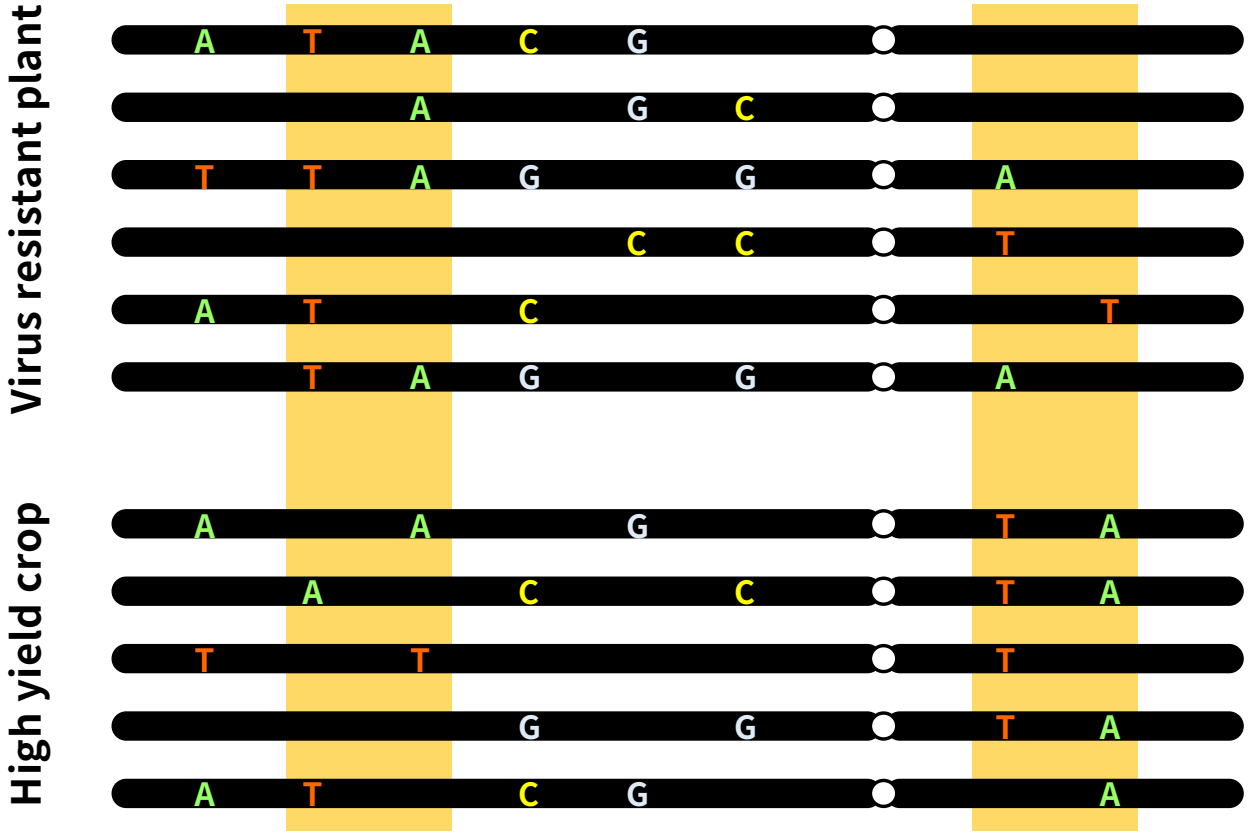
TASSEL 결과 예시

전장 유전체 연관 분석

(Genome-Wide Association Studies)

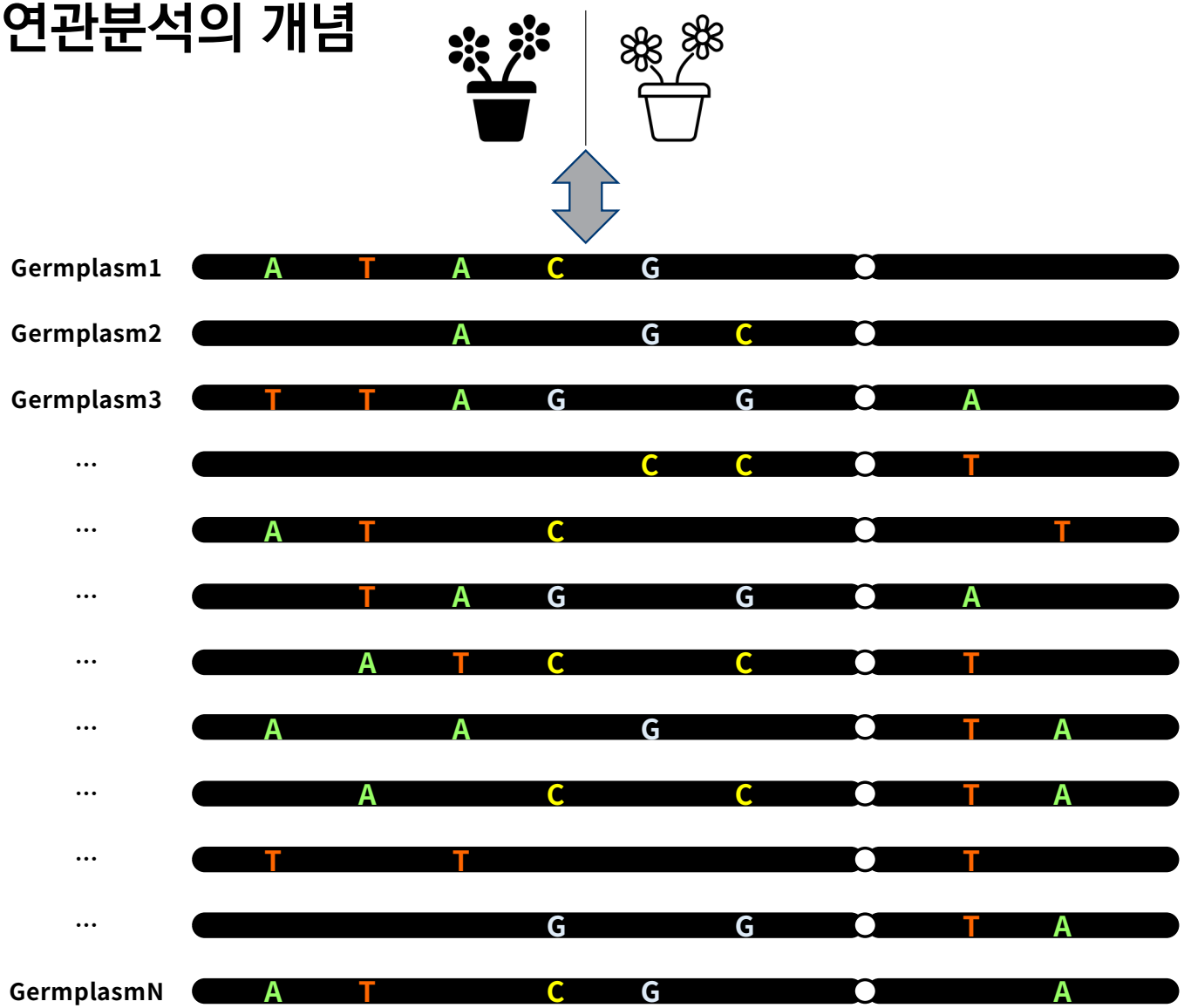
전장 유전체와 형질 연관영역의 분석 개념





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

전장 유전체 연관분석의 개념



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

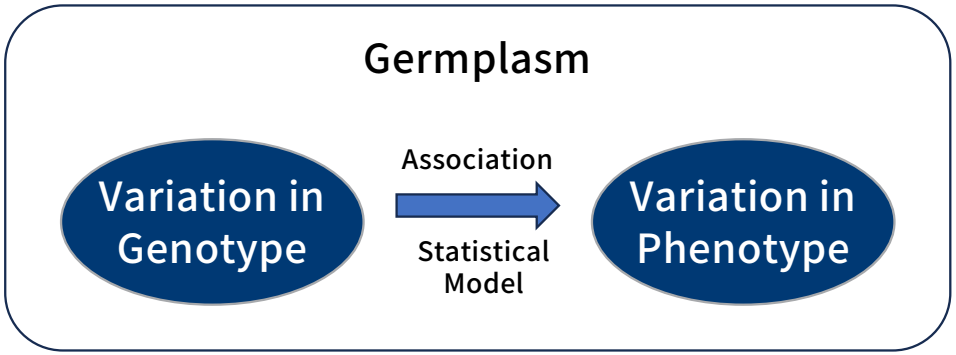
Myles *et al*. 2009

❖ GWAS의 목표:

유전형 variation과 표현형 variation 사이의
유의미한 연관성이 있는 유전적 마커 검색

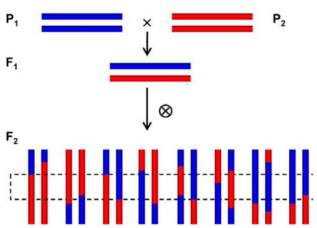
❖ GWAS의 key factors

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

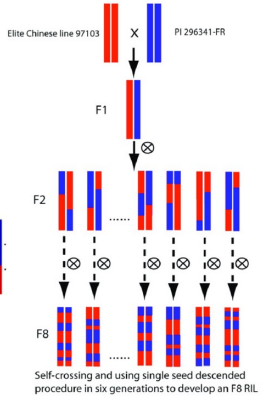


Linkage mapping과 Association Mapping의 비교

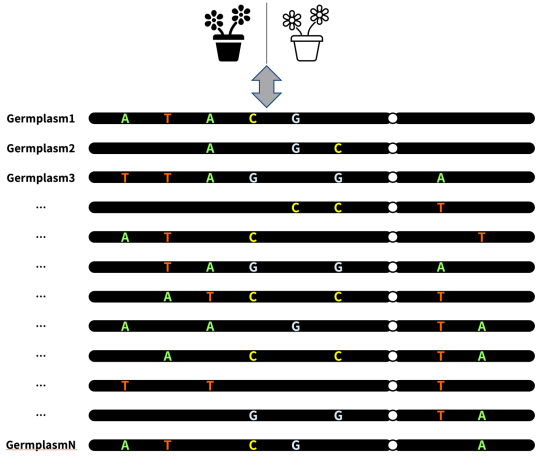
	Linkage Mapping	Association Mapping
장점	• 비교적 작은 집단 사용 • 낮은 density의 genetic marker • Fine-Mapping을 통해 QTL 분석 가능	• 높은 allele diversity • 교배 과정이 필요 없음. → 시간 절약 • Pedigree를 알고 있는 경우, 더 높은 mapping 해상도를 얻을 수 있음.
단점	• 낮은 allele diversity • 교배 과정에서 시간이 오래 걸림 • Recombination event가 낮을 경우, mapping 해상도가 낮음.	• 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



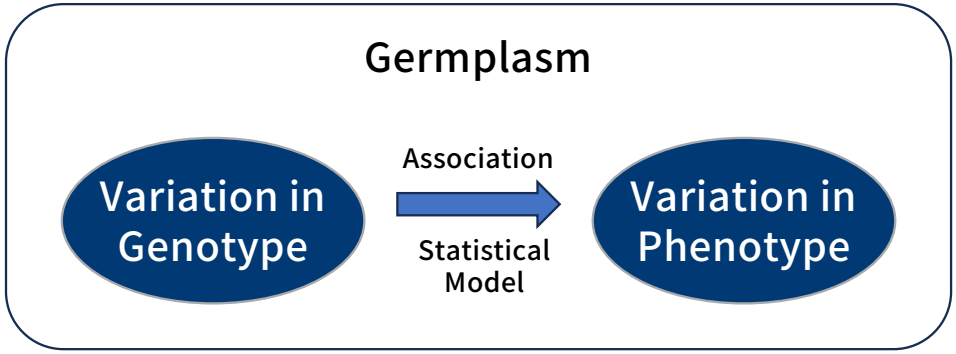
Zhu *et al.* (2008)



Ren *et al.* (2012)



GWAS 분석을 위해 고려할 사항



❖ GWAS의 key factors

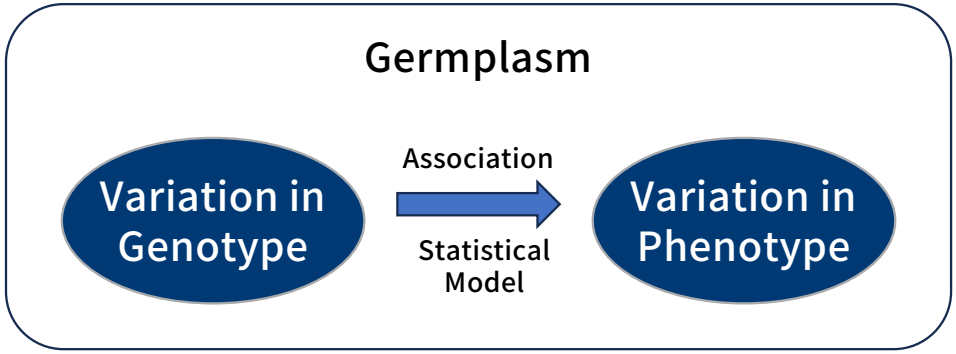
- Germplasm (population)
- Genetic markers
- Statistical models
- Phenotype

일반 선형 모형
GLM (General Linear Model)

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

- Y: 표현형 관측치 (observation)
- SNP: 유전자형 (fixed)
- $Q\ (or\ PCs)$: Population Structure (fixed)
- e: 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)

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

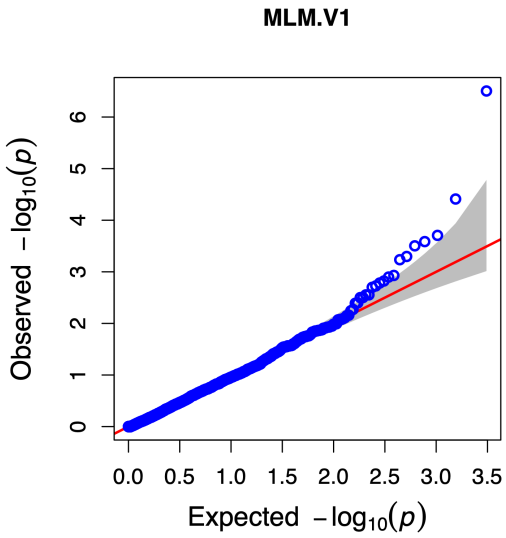
혼합 선형 모형
MLM (Mixed Linear Model)

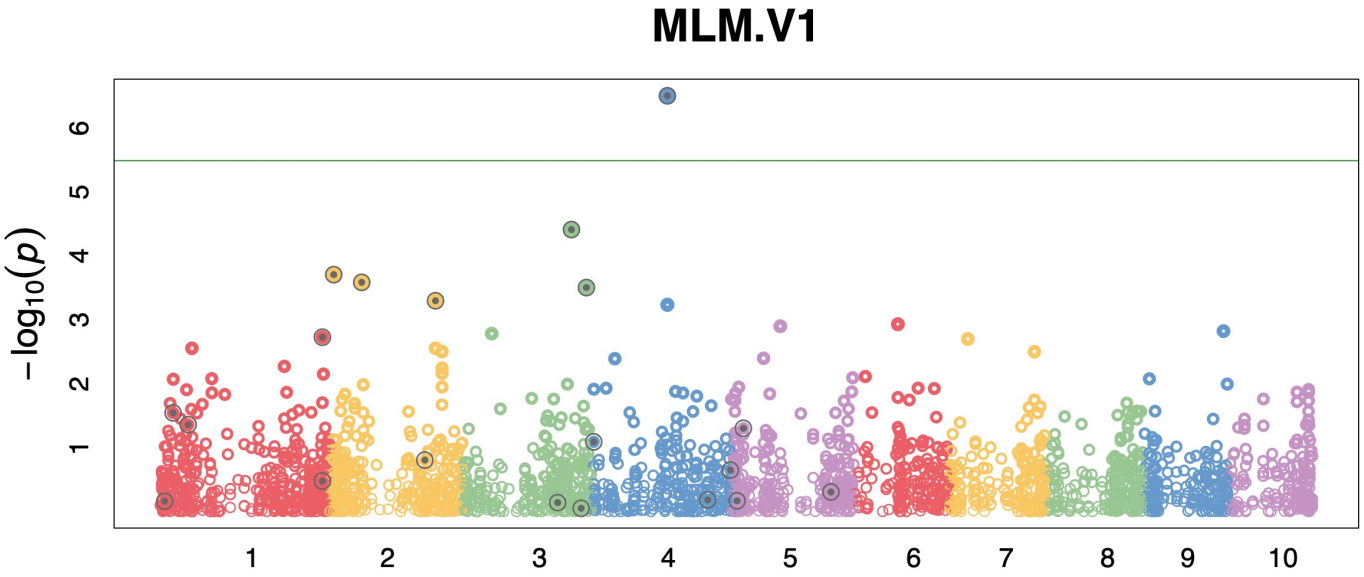
$$Y = SNP + Q\ (or\ PCs) + Kinship + e$$

(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)

GAPIT 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)

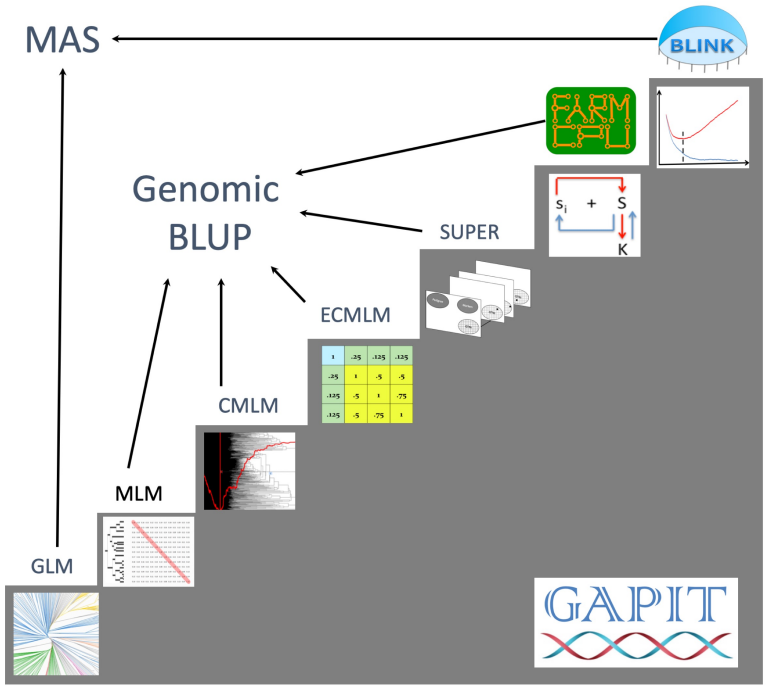
Tassel 5 Mac OS
Tassel 5 Windows 64 Bit
Tassel 5 Windows 32 Bit
Tassel 5 UNIX

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")
```

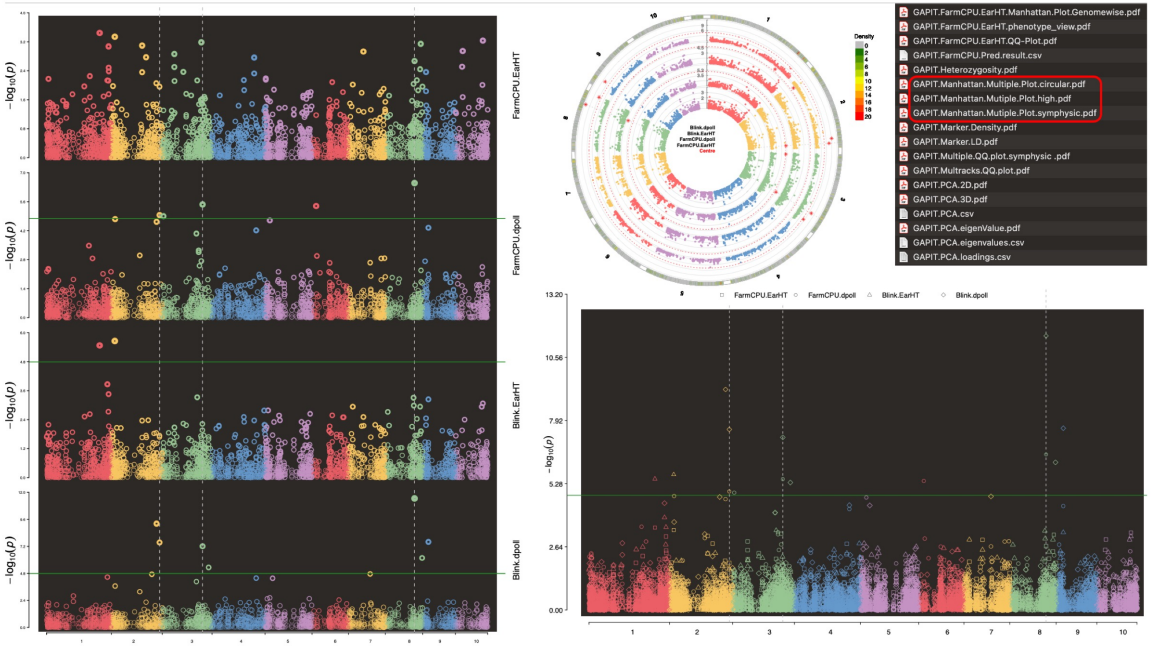
Or GitHub:

```
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)
```



강의 요약

- ❑ NGS의 발전으로 인하여 sequencing 데이터 생산의 가격은 낮아지고, 속도는 빨라짐. 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