초고성능컴퓨터 활용 교육

유전체 변이 분석 (이론)

(주)바이오투, 최준경

2023-07-18

목 차

I. NGS를 이용한 연구 방향

- NGS 발전에 따른 변화
- 최신 NGS 장비의 능력 및 분석 파이프라인 (예시)
- 유전체 상에 존재하는 다양한 변이(Variant) 및 SNP 변이
- NGS로 대량 형질연관 SNP 획득 → 분자육종

II. NGS를 이용하는 다양한 Sequencing 방법

- 분자마커 발굴을 위한 다양한 NGS 기법
- Sequencing 기법에 따른 차이
- 변이를 통해 할 수 있는 것

III. NGS를 이용한 유전체 변이 분석 과정

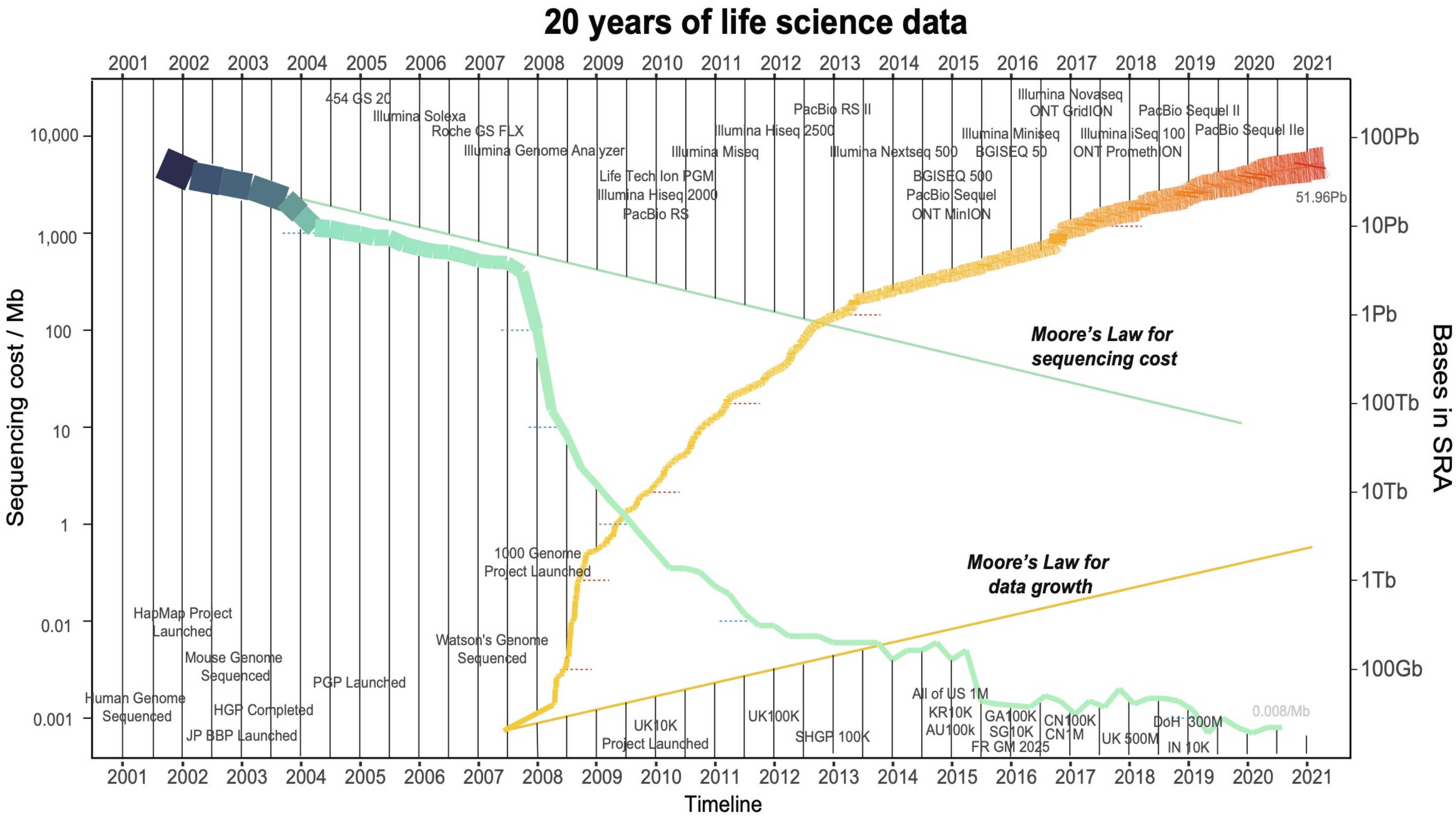
- NGS 분석 파이프라인
- Pre-processing
- Read Alignment
- Variant Detection
- Variant Filtration

Ⅳ. 강의 요약

- 강의 요약

NGS를 이용한 연구 방향

NGS 발전에 따른 변화



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

최신 NGS 장비의 능력

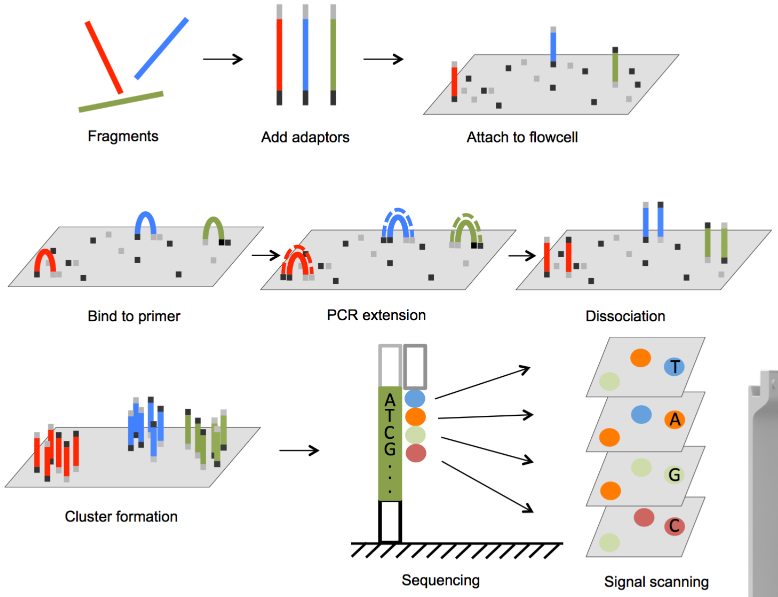
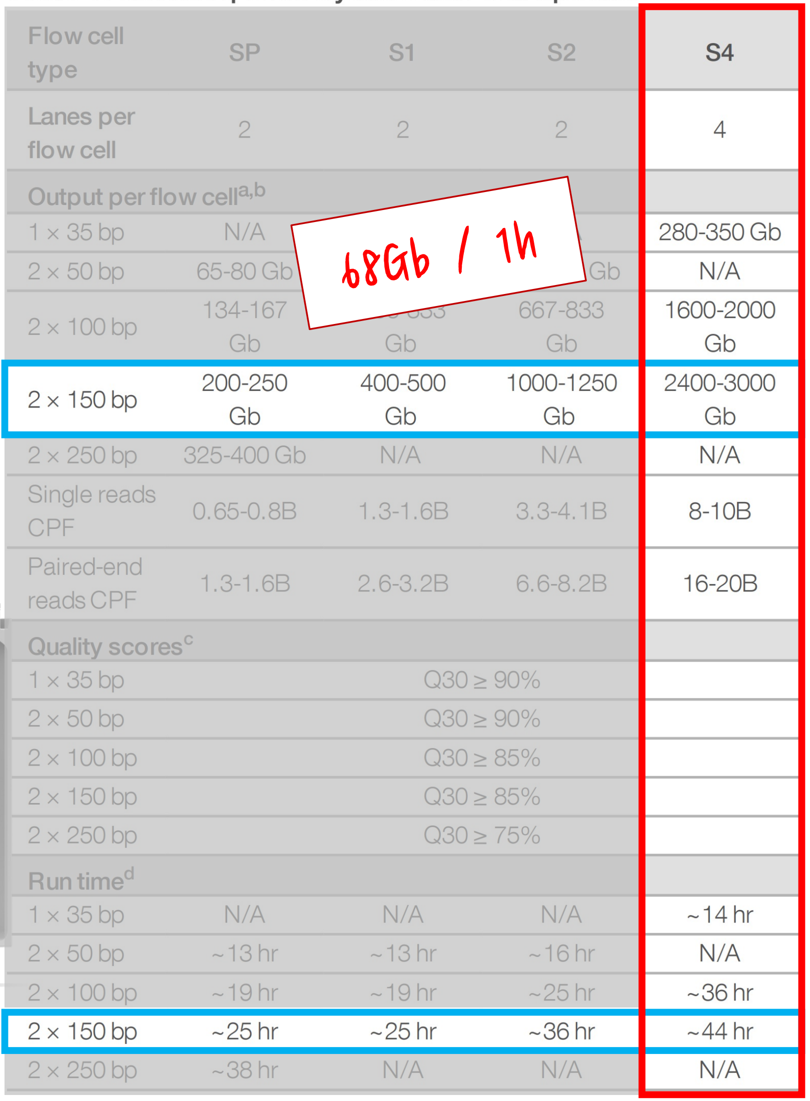


Table 1: NovaSeq 6000 System flow cell specifications

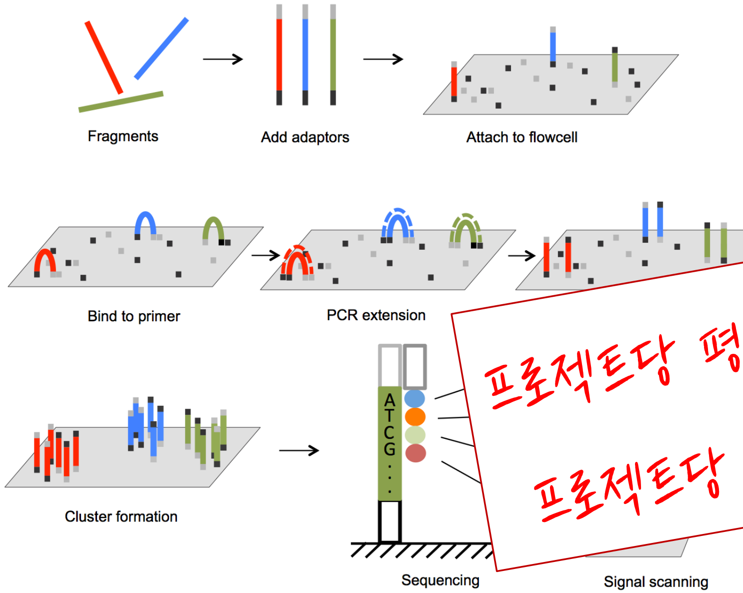


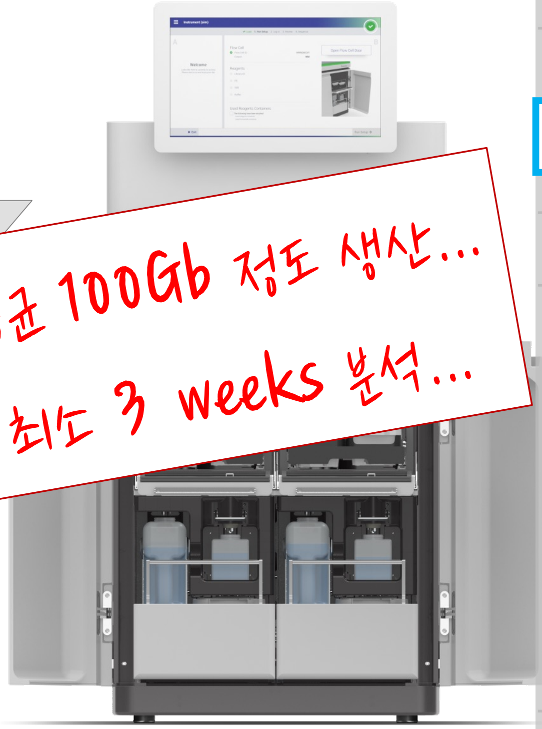
Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
1 × 35 bp	N/A	[illegible]	[illegible]	280-350 Gb
2 × 50 bp	65-80 Gb	[illegible]	[illegible] Gb	N/A
2 × 100 bp	134-167 Gb	[illegible]-833 Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000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

최신 NGS 장비의 능력





프로젝트당 평균 100Gb 정도 생산...

프로젝트당 최소 3 weeks 분석...

68Gb / 1h

하루에 벼 250 품종(30X 기준)을 한 번에 읽어 버릴 수 있다.

Oryza sativa genome size = 389 Mb

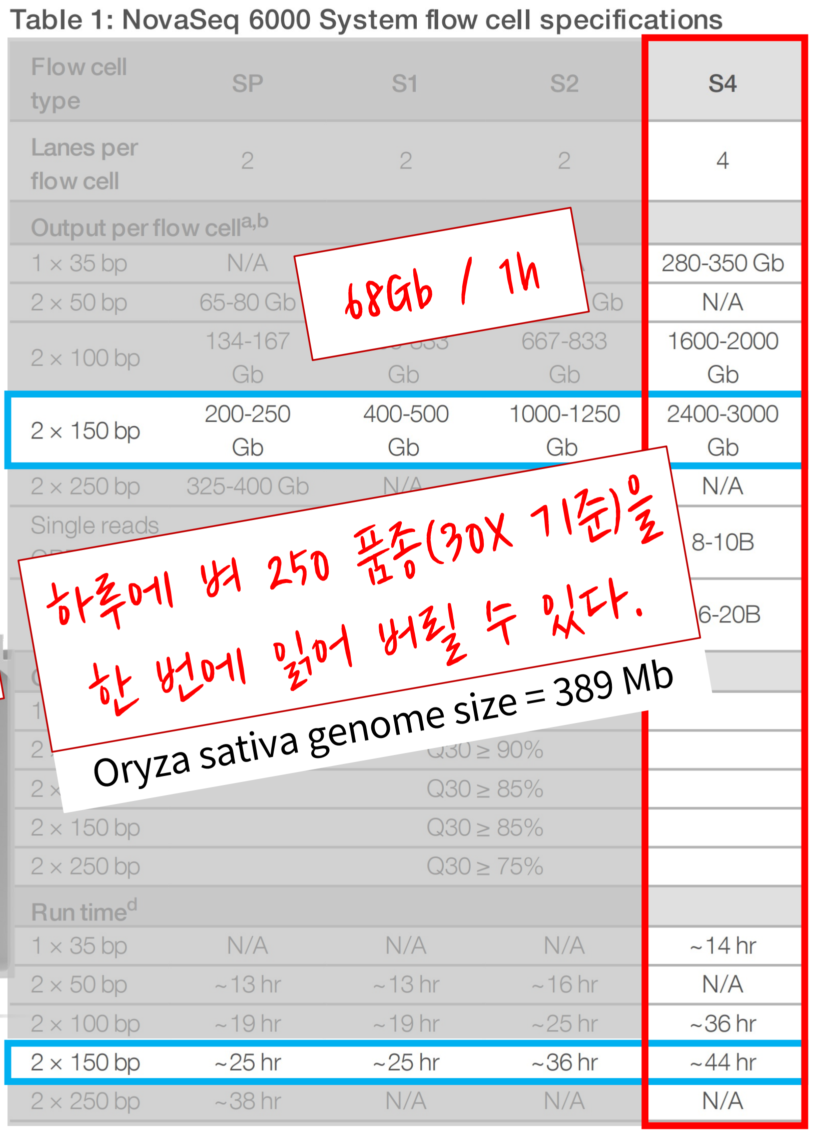


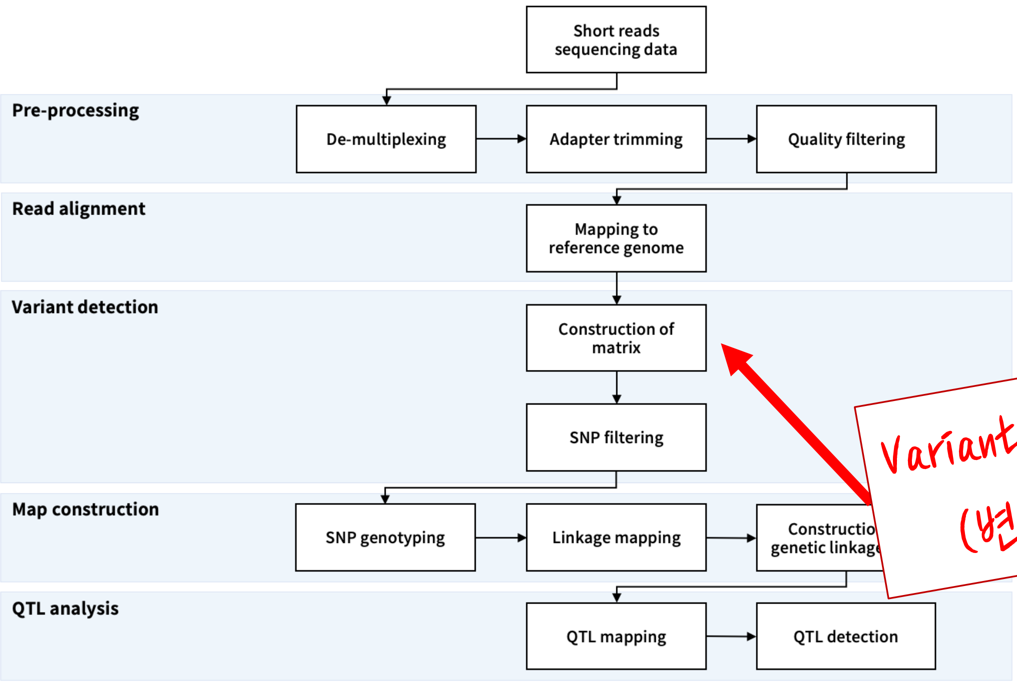
Table 1: NovaSeq 6000 System flow cell specifications

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	[illegible]	[illegible]	280-350 Gb
2 × 50 bp	65-80 Gb	[illegible]	[illegible] Gb	N/A
2 × 100 bp	134-167 Gb	[illegible]-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	[illegible]	N/A
Single reads	[illegible]	[illegible]	[illegible]	8-10B
[illegible]	[illegible]	[illegible]	[illegible]	6-20B
Q[illegible]				
1[illegible]	[illegible]			
2 [illegible]		Q30 ≥ 90%		
2 × [illegible]		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

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

NGS 데이터의 분석 파이프라인 (예시)

❖ Pipeline for QTL-mapping



Short reads sequencing data

Pre-processing: De-multiplexing → Adapter trimming → Quality filtering

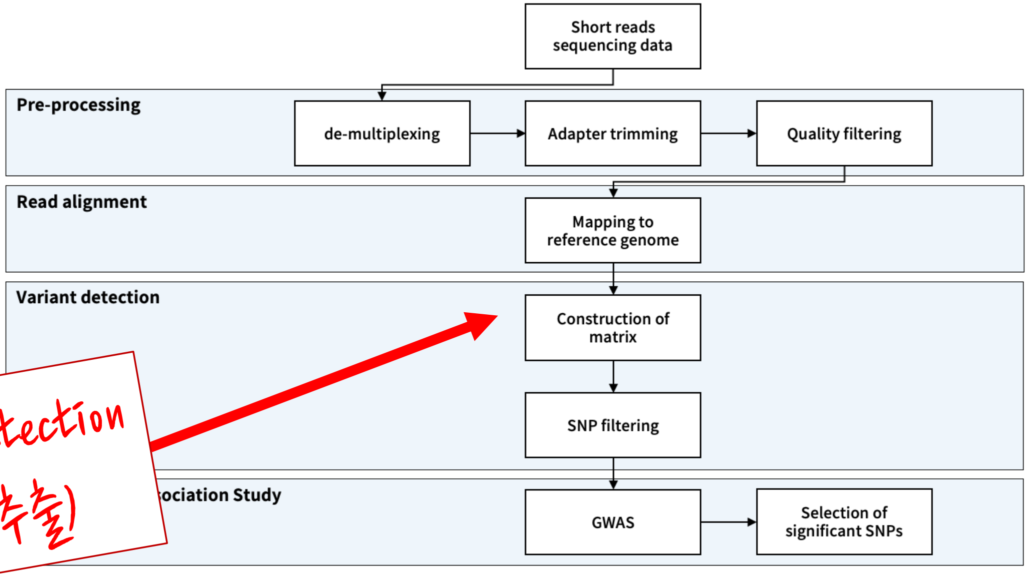
Read alignment: Mapping to reference genome

Variant detection: Construction of matrix → SNP filtering

Map construction: SNP genotyping → Linkage mapping → Constructio genetic linkage

QTL analysis: QTL mapping → QTL detection

❖ Pipeline for GWAS



Short reads sequencing data

Pre-processing: de-multiplexing → Adapter trimming → Quality filtering

Read alignment: Mapping to reference genome

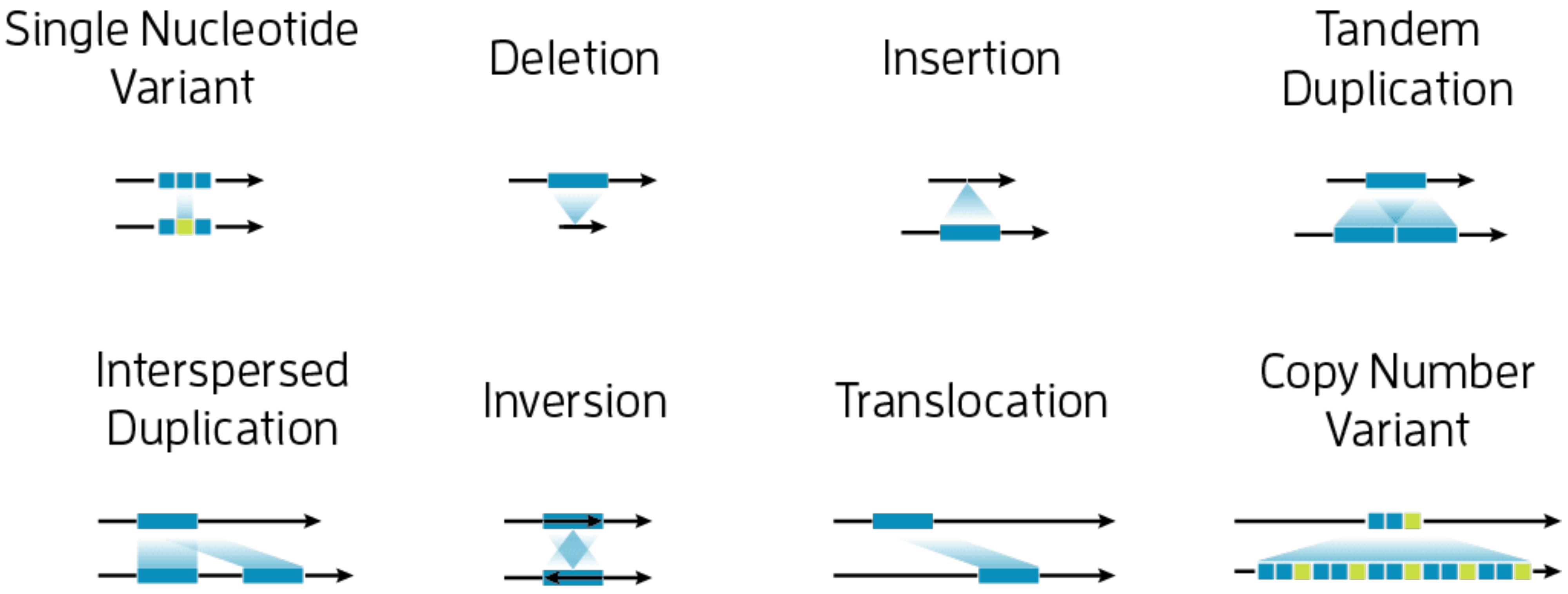
Variant detection: Construction of matrix → SNP filtering

ociation Study: GWAS → Selection of significant SNPs

Variant Detection

(변이 추출)

유전체 상에 존재하는 다양한 변이(Variant)



Single Nucleotide Variant

Deletion

Insertion

Tandem Duplication

Interspersed Duplication

Inversion

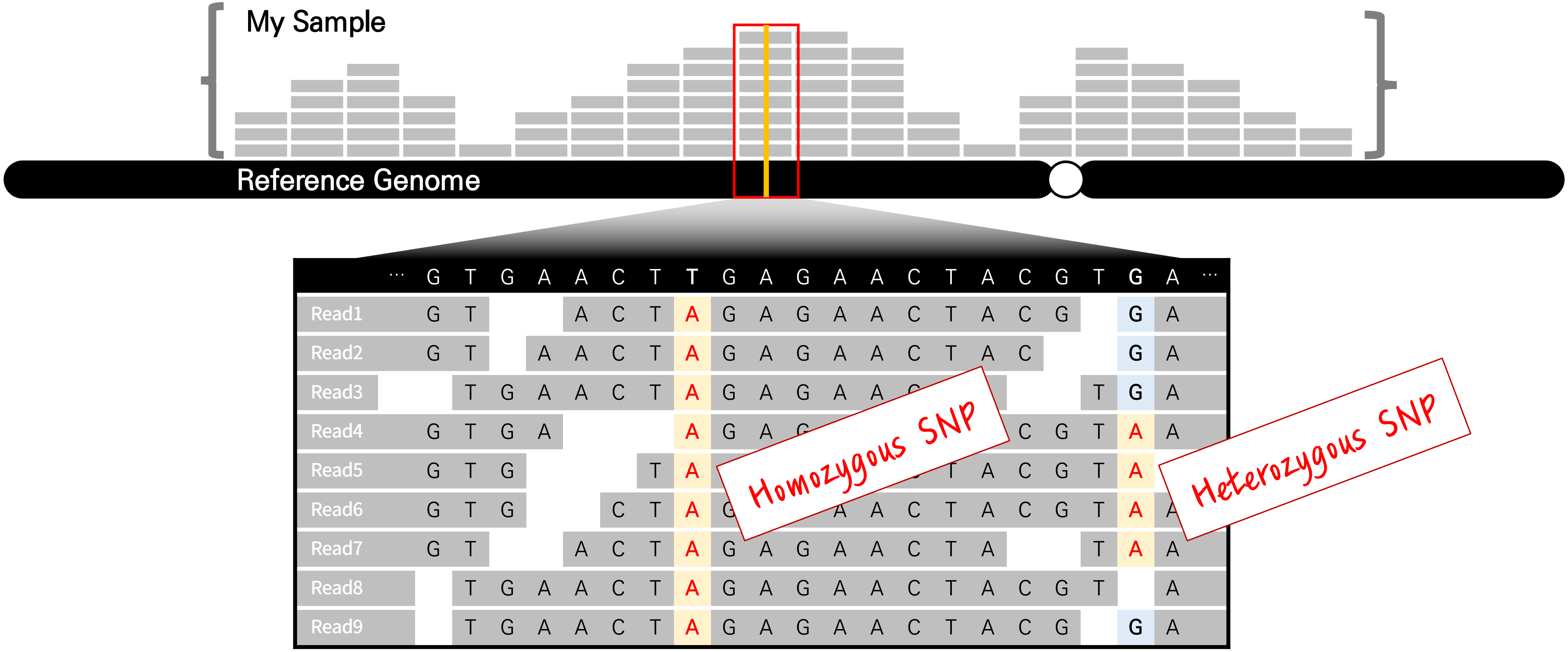
Translocation

Copy Number Variant

Types of Variants

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

SNP 변이

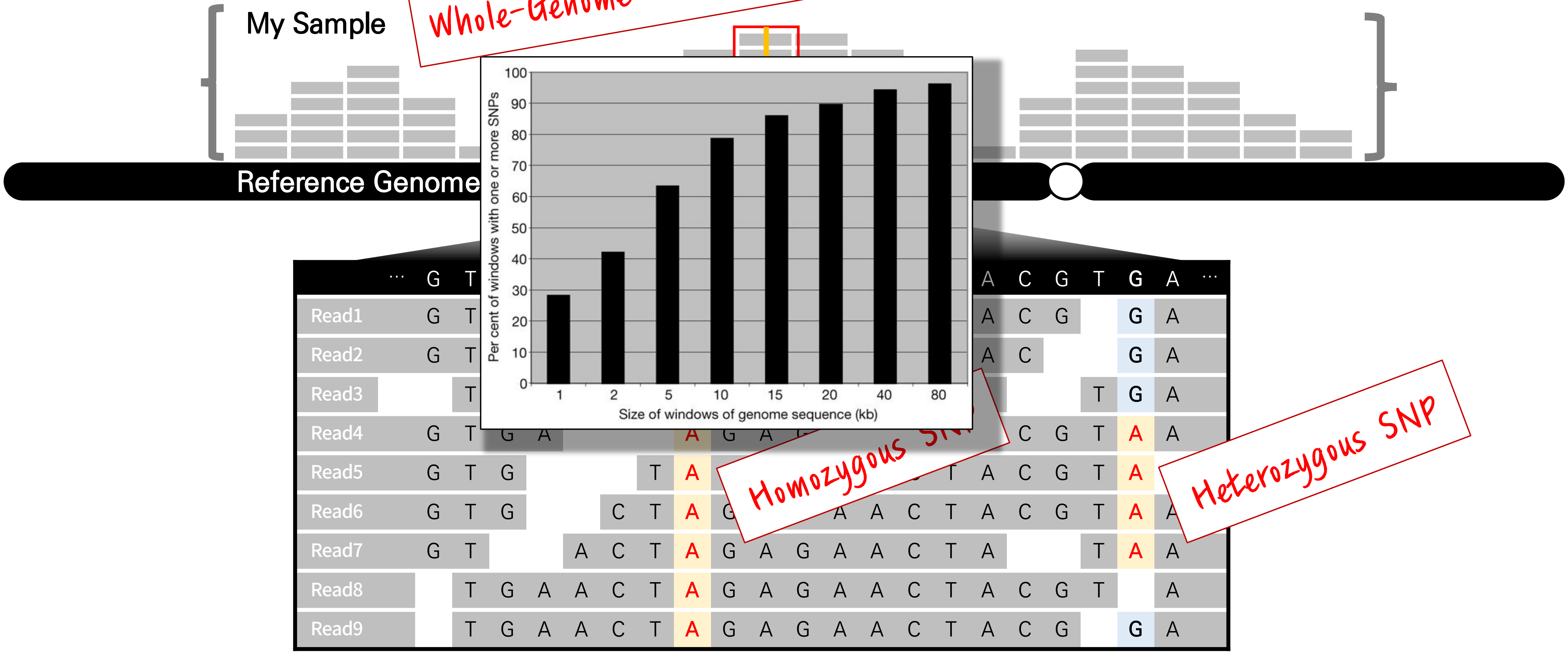


	…	G	T	G	A	A	C	T	**T**	G	A	G	A	A	C	T	A	C	G	T	**G**	A
Read1		G	T			A	C	T	A	G	A	G	A	A	C	T	A	C	G		**G**	A
Read2		G	T		A	A	C	T	A	G	A	G	A	A	C	T	A	C			**G**	A
Read3			T	G	A	A	C	T	A	G	A	G	A	A	C					T	**G**	A
Read4		G	T	G	A				A	G	A	G						C	G	T	A	A
Read5		G	T	G				T	A							T	A	C	G	T	A	
Read6		G	T	G			C	T	A	G			A	A	C	T	A	C	G	T	A	A
Read7		G	T			A	C	T	A	G	A	G	A	A	C	T	A			T	A	A
Read8			T	G	A	A	C	T	A	G	A	G	A	A	C	T	A	C	G	T		A
Read9			T	G	A	A	C	T	A	G	A	G	A	A	C	T	A	C	G		**G**	A

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

SNP 변이

Whole-Genome에 SNP가 고르게 분포되어 있음.



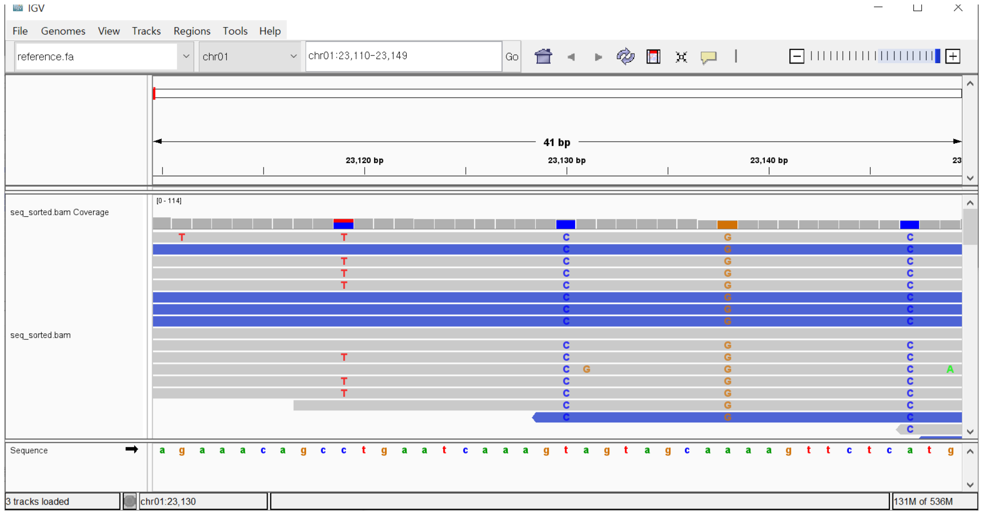
Homozygous SN

Heterozygous SNP

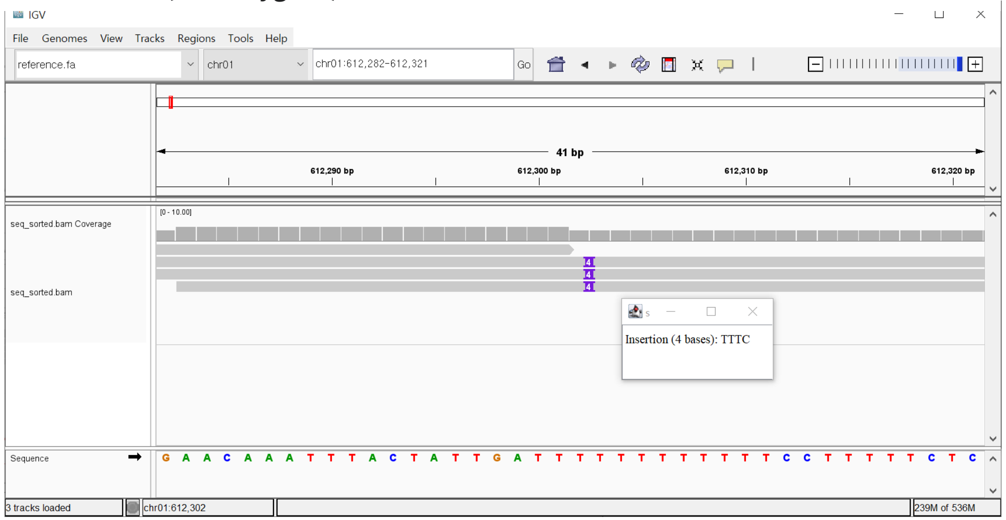
- Sachidanandam R et al. Nature. 2001;409(6822):928-933

IGV 프로그램을 통한 변이 관측

❖ SNP (Homozygous & Heterozygous)

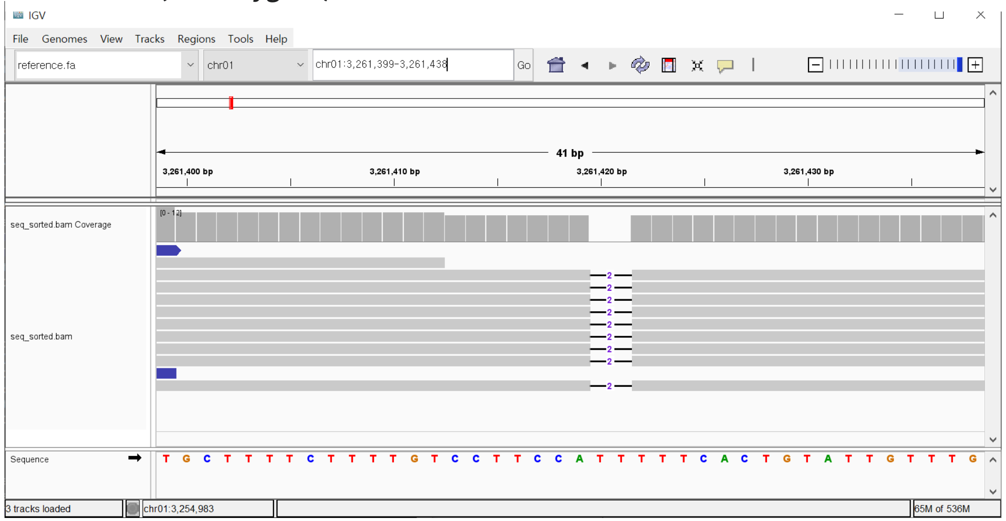


❖ Insertion (Homozygous)

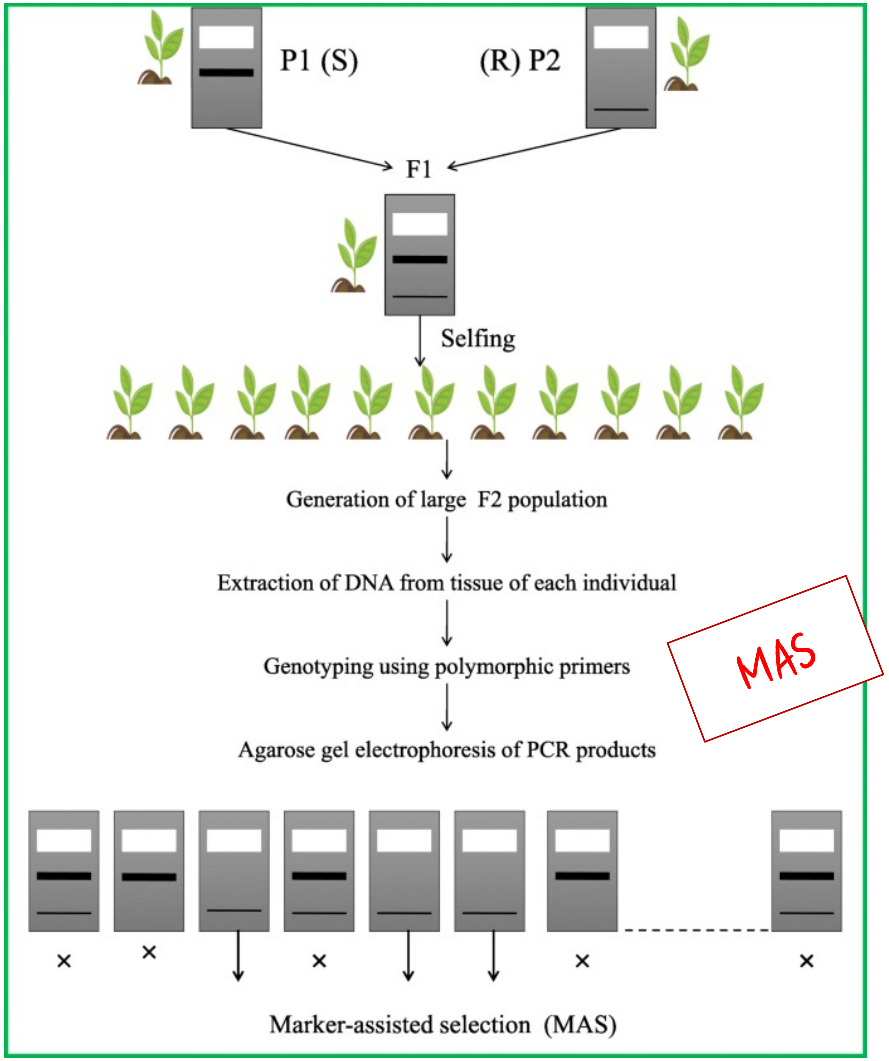


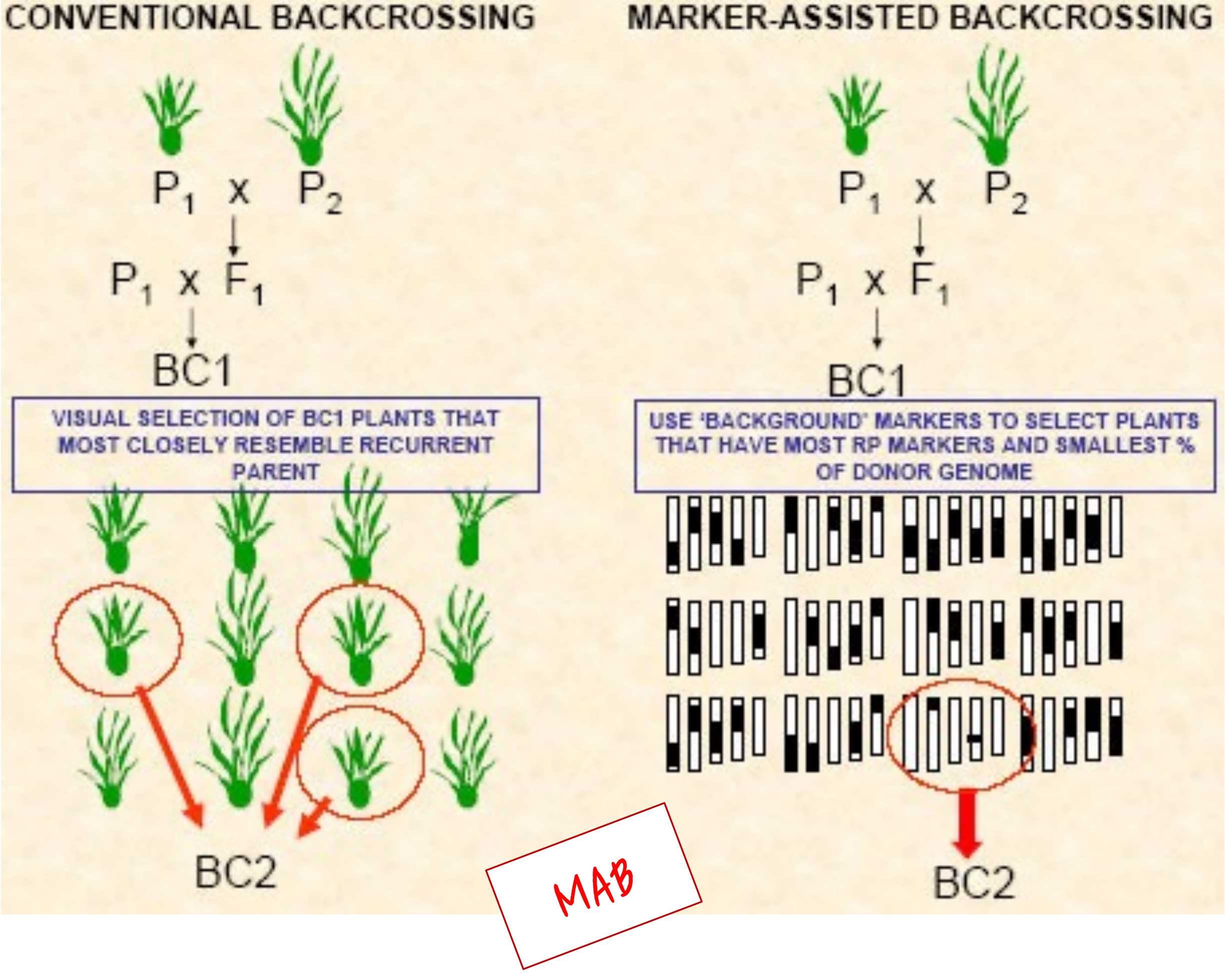
Insertion (4 bases): TTTC

❖ Deletion (Homozygous)



NGS로 대량 형질연관 SNP 획득 → 분자육종


P1 (S)
(R) P2
F1
Selfing
Generation of large F2 population
Extraction of DNA from tissue of each individual
Genotyping using polymorphic primers
Agarose gel electrophoresis of PCR products
×
×
×
×
×
Marker-assisted selection (MAS)
MAS

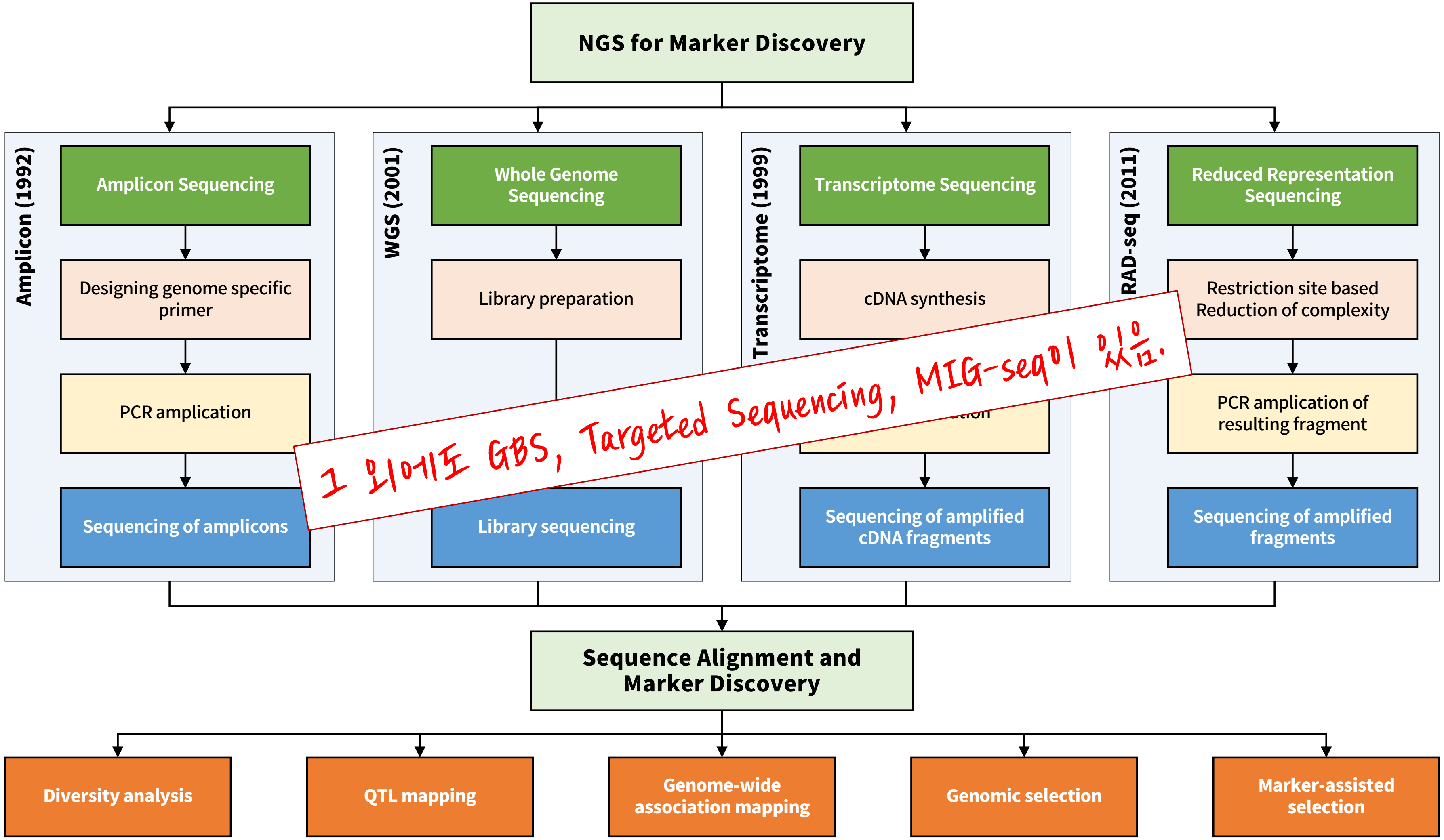

CONVENTIONAL BACKCROSSING
P1 x P2
P1 x F1
BC1
VISUAL SELECTION OF BC1 PLANTS THAT MOST CLOSELY RESEMBLE RECURRENT PARENT
BC2
MARKER-ASSISTED BACKCROSSING
P1 x P2
P1 x F1
BC1
USE 'BACKGROUND' MARKERS TO SELECT PLANTS THAT HAVE MOST RP MARKERS AND SMALLEST % OF DONOR GENOME
BC2
MAB

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

NGS를 이용하는 다양한 Sequencing 방법

분자마커 발굴을 위한 다양한 NGS 기법

BIOTO (주)바이오투



NGS for Marker Discovery

Amplicon (1992)
- Amplicon Sequencing
- Designing genome specific primer
- PCR amplication
- Sequencing of amplicons

WGS (2001)
- Whole Genome Sequencing
- Library preparation
- Library sequencing

Transcriptome (1999)
- Transcriptome Sequencing
- cDNA synthesis
- ation
- Sequencing of amplified cDNA fragments

RAD-seq (2011)
- Reduced Representation Sequencing
- Restriction site based Reduction of complexity
- PCR amplication of resulting fragment
- Sequencing of amplified fragments

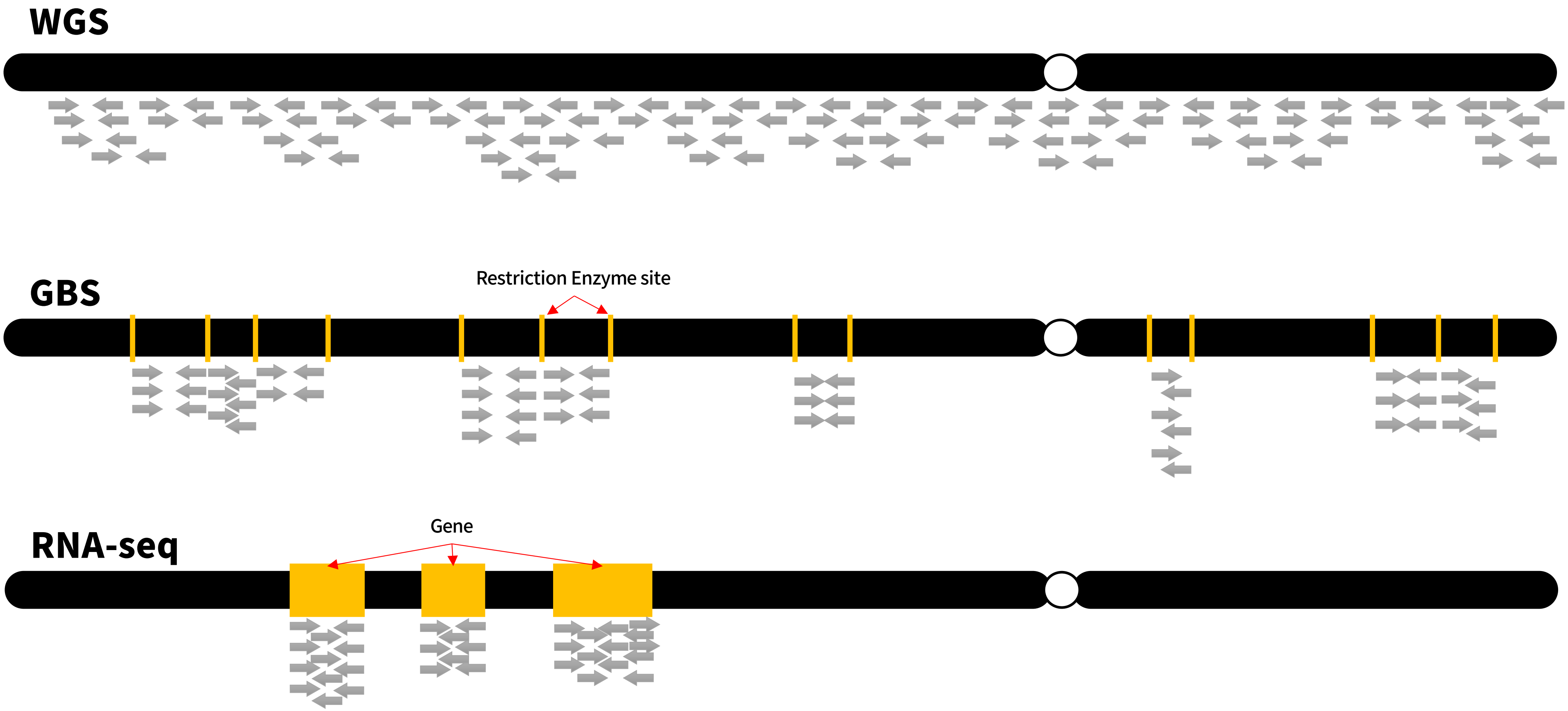
Sequence Alignment and Marker Discovery

- Diversity analysis
- QTL mapping
- Genome-wide association mapping
- Genomic selection
- Marker-assisted selection

그 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)

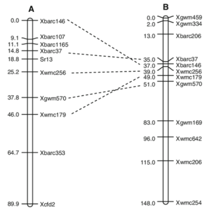


Sequencing 기법에 따른 차이 (2/2)

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

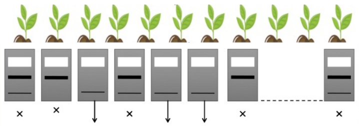
변이를 통해 할 수 있는 것

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



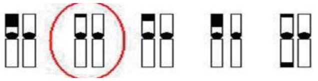
Linkage Map

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



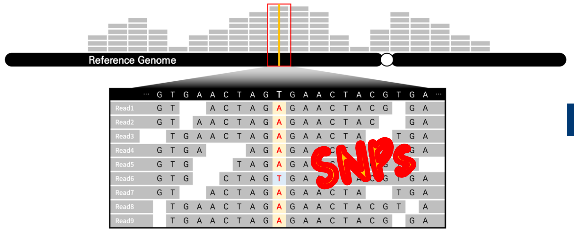
Marker-assisted selection (MAS)

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

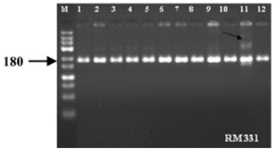


Marker-assisted backcrossing (MAB)

Next-Generation Sequencing

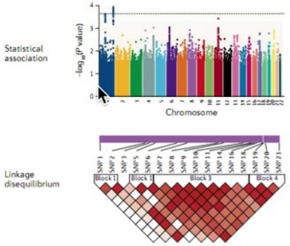


순도 검정 마커



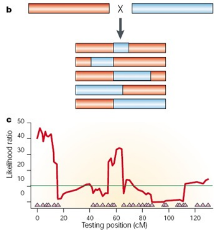
- Bora *et al.* Biotech 6, 50 (2016).

Association Mapping



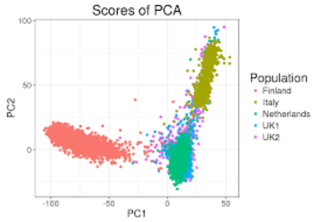
Tam et al 2019

QTL Mapping



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

원산지 구분 마커

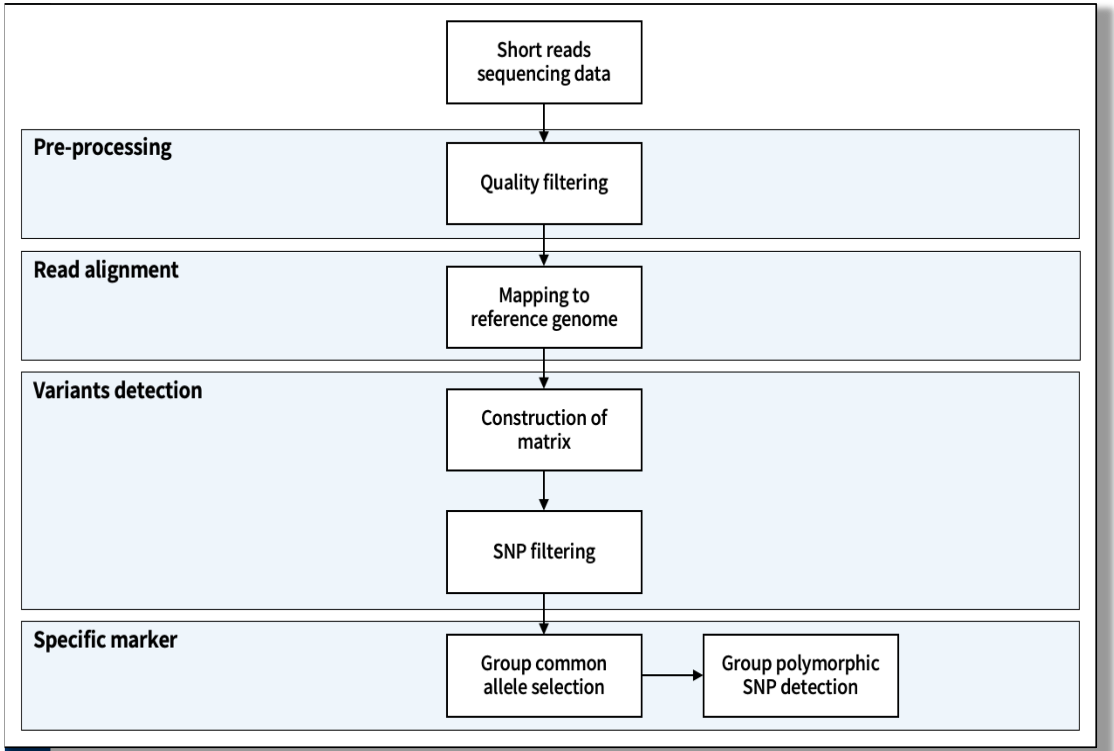


- https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

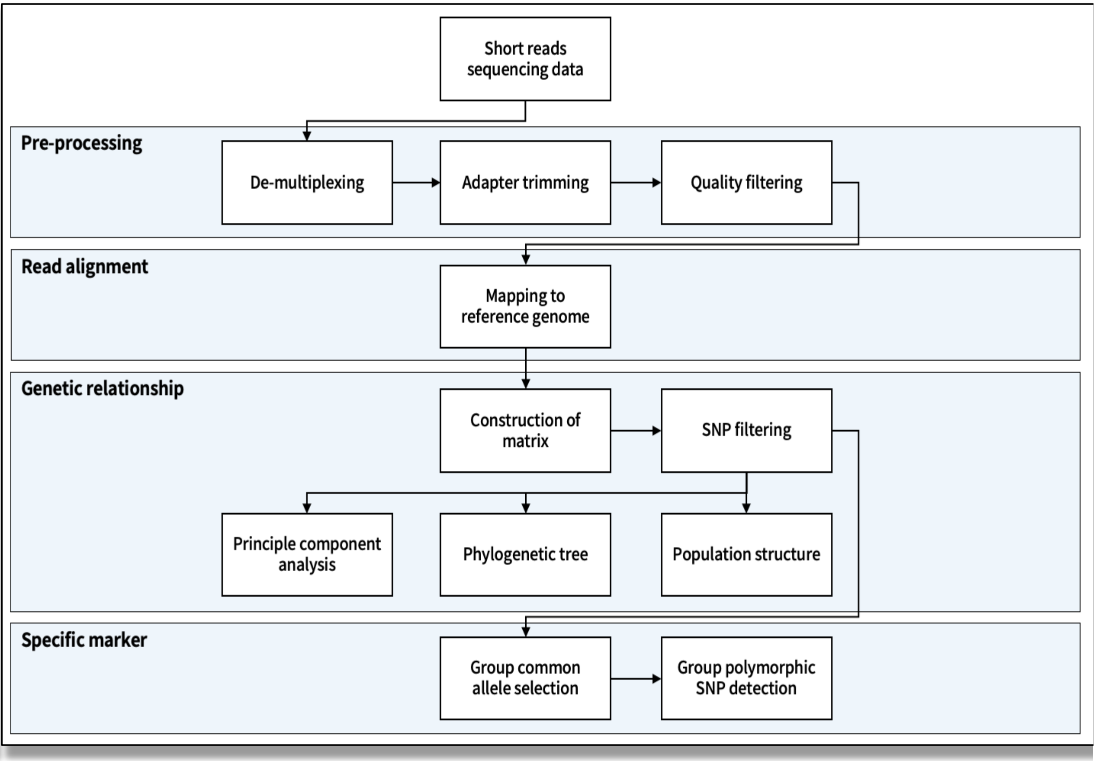
NGS를 이용한 유전체 변이 분석 과정

NGS 분석 파이프라인

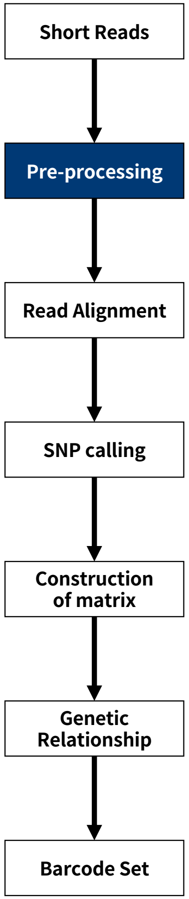
❖ WGS (Whole Genome Sequencing) 파이프라인


Short reads sequencing data
Pre-processing
Quality filtering
Read alignment
Mapping to reference genome
Variants detection
Construction of matrix
SNP filtering
Specific marker
Group common allele selection
Group polymorphic SNP detection

❖ GBS (Genotyping-By-Sequencing) 파이프라인



Pre-processing



❑ 시퀀싱 데이터의 전처리 과정

- Removal of technical sequences
- Quality and length filtering

❑ 많이 사용되는 프로그램

1. Trimmomatic

   A flexible trimmer for Illumina sequence data

2. FASTQC

   A quality control tool for high throughput sequence data.

3. FASTX-Toolkit

   The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

4. SolexaQA

   SolexaQA calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data.

Trimmomatic 실행

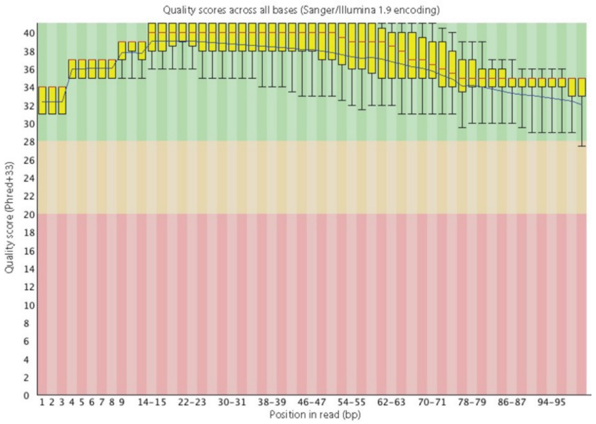
❑ Trimmomatic 프로그램 옵션

- 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)

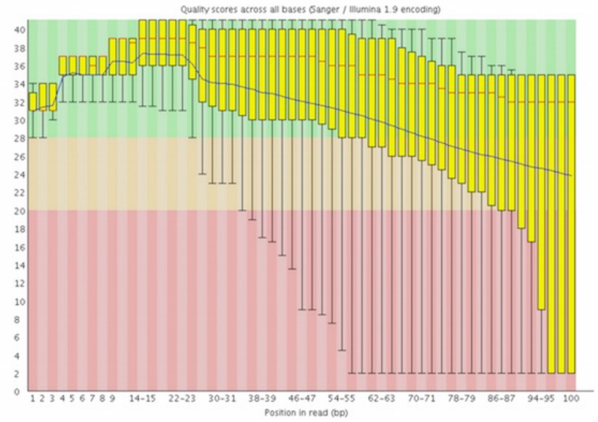
```
## Sequencing Data 디렉토리로 이동

cd /home/edu_01/1.rawdata


## 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
```

Pre-processing 결과 예시

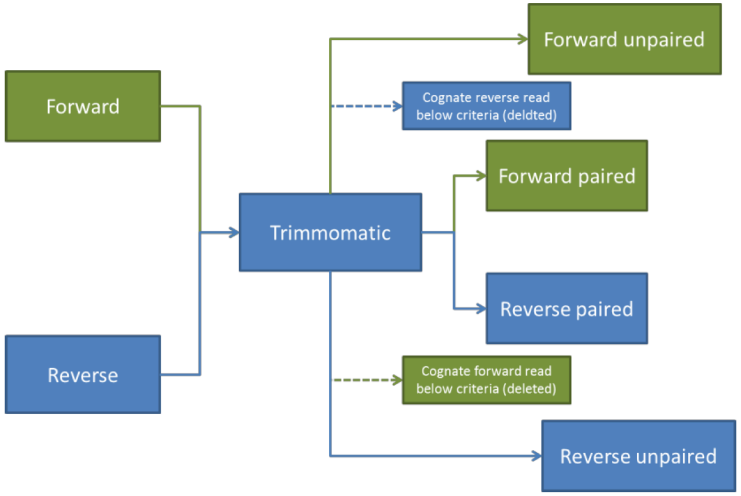
❖ FASTQC



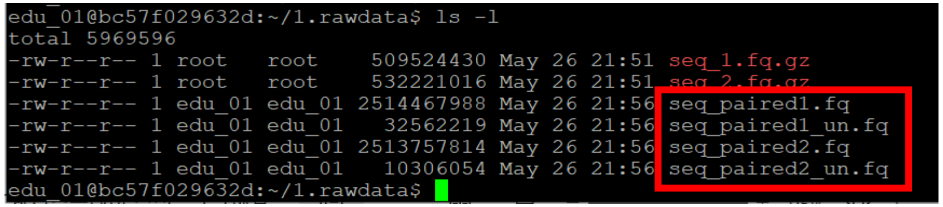
Per base sequence quality



❖ Flow of reads in Trimmomatic Paired End mode


Forward
Reverse
Trimmomatic
Forward unpaired
Cognate reverse read below criteria (deldted)
Forward paired
Reverse paired
Cognate forward read below criteria (deleted)
Reverse unpaired

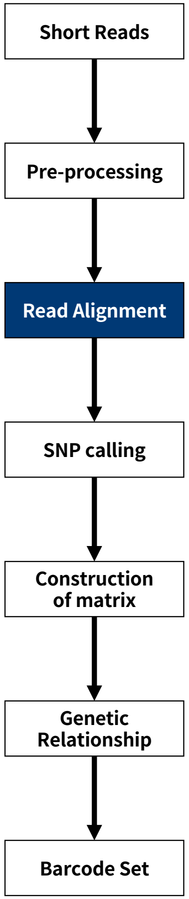
❖ Trimmomatic 결과 예시



```
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

Read Alignment



Short Reads

Pre-processing

Read Alignment

SNP calling

Construction of matrix

Genetic Relationship

Barcode Set

❑ 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

1. BWA (Burrows-Wheeler Aligner)

   BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.

2. Bowtie2

   Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

3. 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.

4. RUM, STAR, TopHat2, …

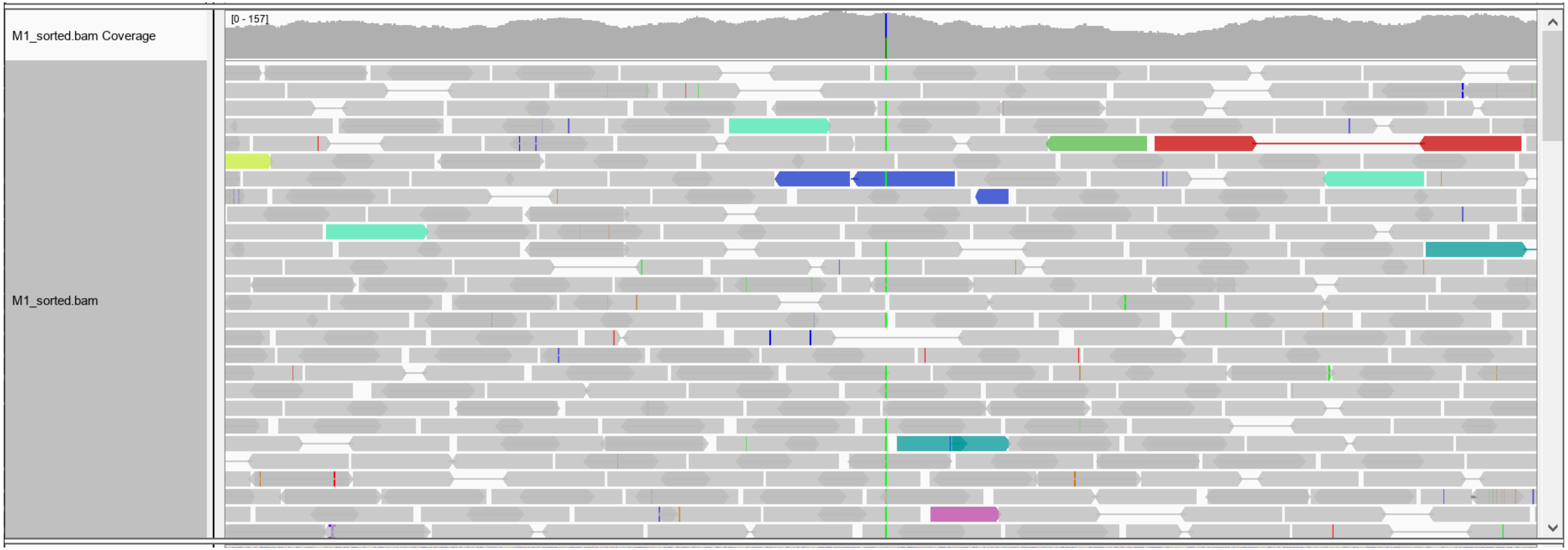
Alignment Tools 간의 비교

Tool	Alignment Reference	Description
Bowtie2	Transcriptome + Genome	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

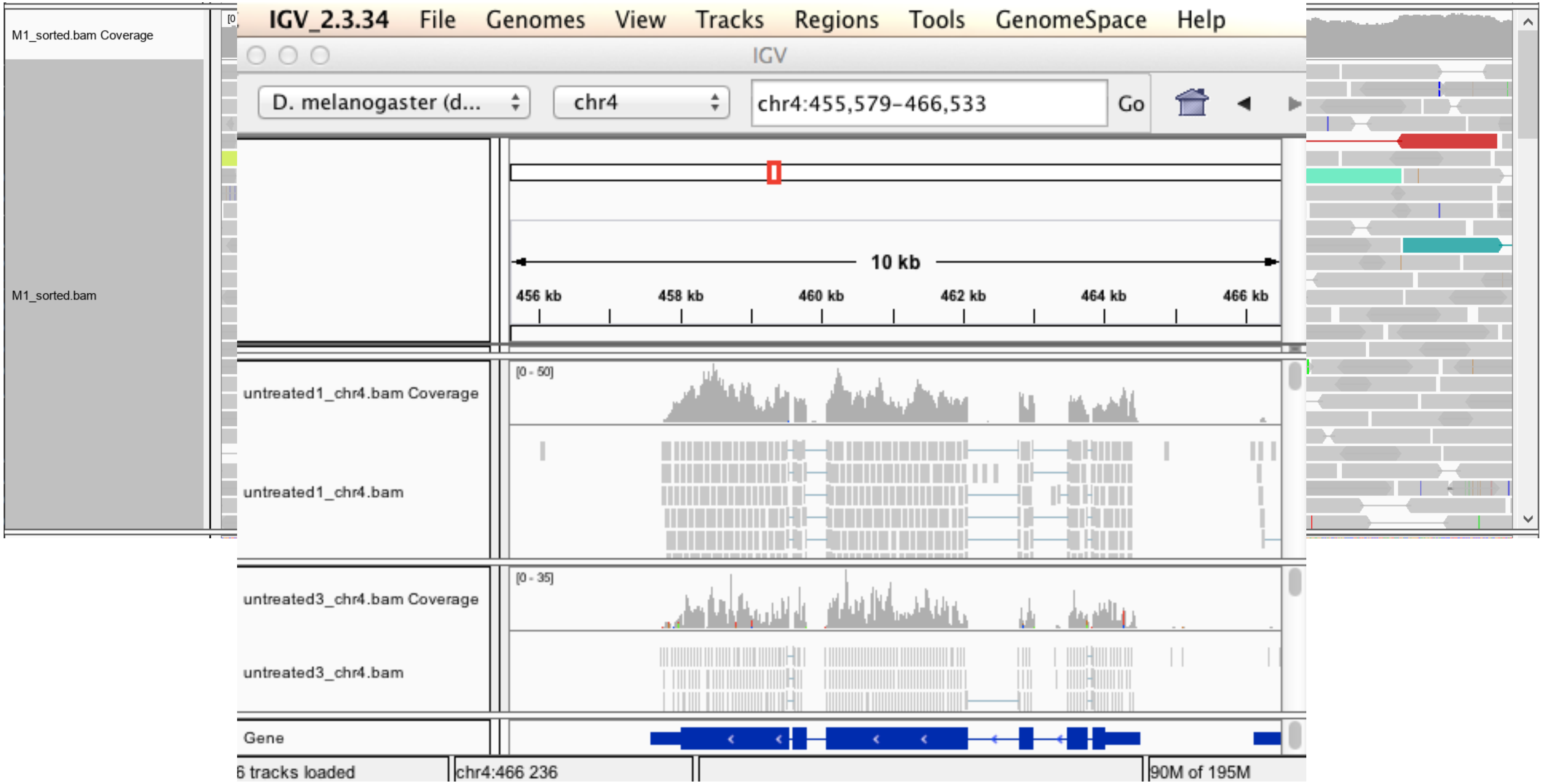
Read Alignment 결과 예시

Read alignment – IGV 결과

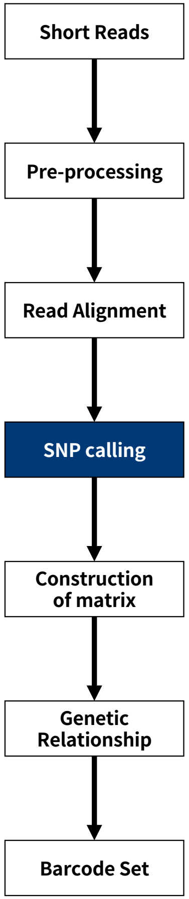


Read Alignment 결과 예시

Read alignment – IGV 결과



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

1. SAMTools

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

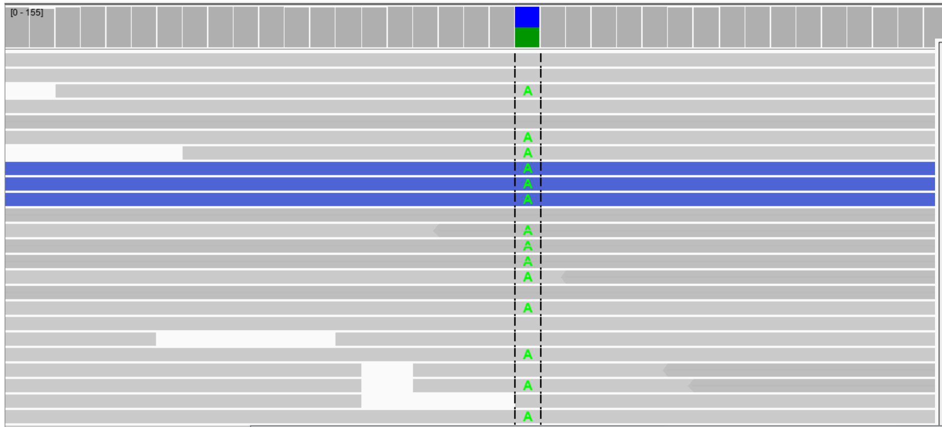
   Find variants

2. GATK

   A genomic analysis toolkit focused on variant discovery

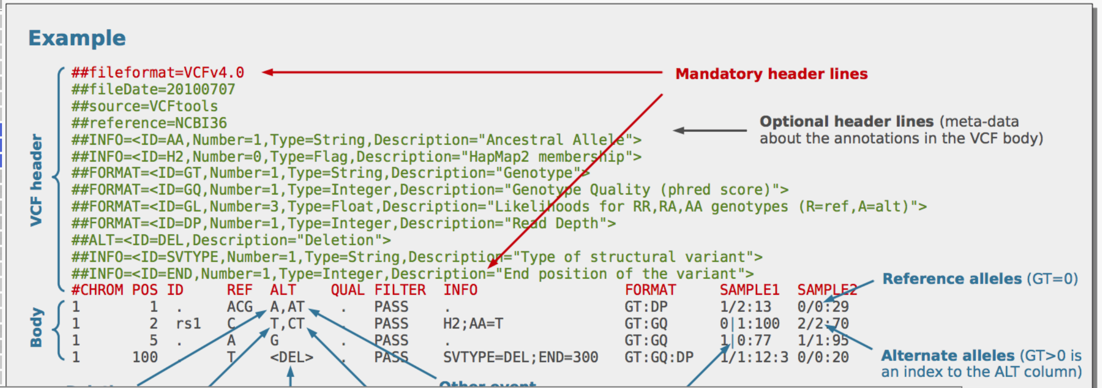
Variant Detection 결과 예시

Variant Detection - 결과



IGV

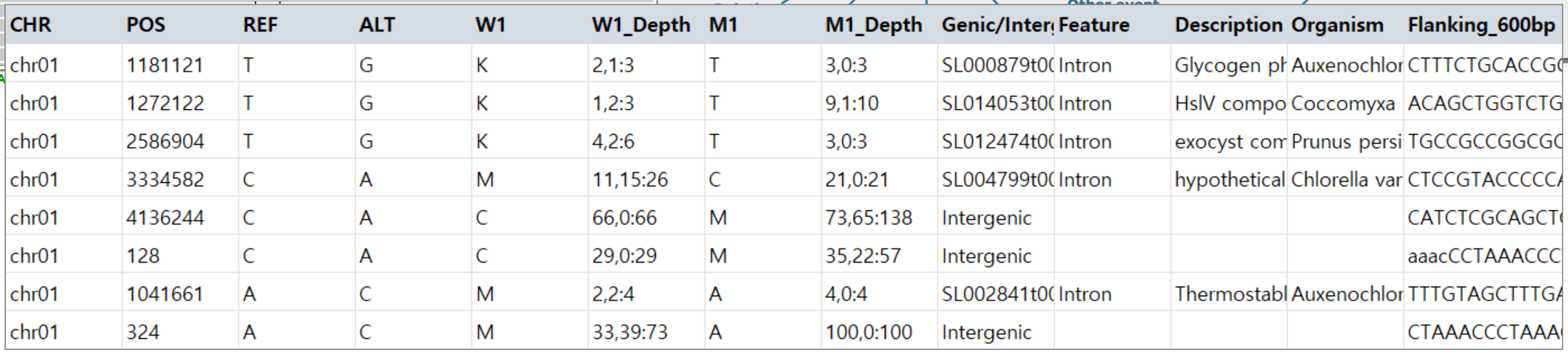
VCF (Variant Call Format)



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                FORMAT     SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT      .  PASS    .                   GT:DP      1/2:13    0/0:29
1        2 rs1    C    T,CT      .  PASS    H2;AA=T             GT:GQ      0|1:100   2/2:70
1        5 .      A    G         .  PASS    .                   GT:GQ      1|0:77    1/1:95
1      100 .      T    <DEL>     .  PASS    SVTYPE=DEL;END=300  GT:GQ:DP   1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Other event

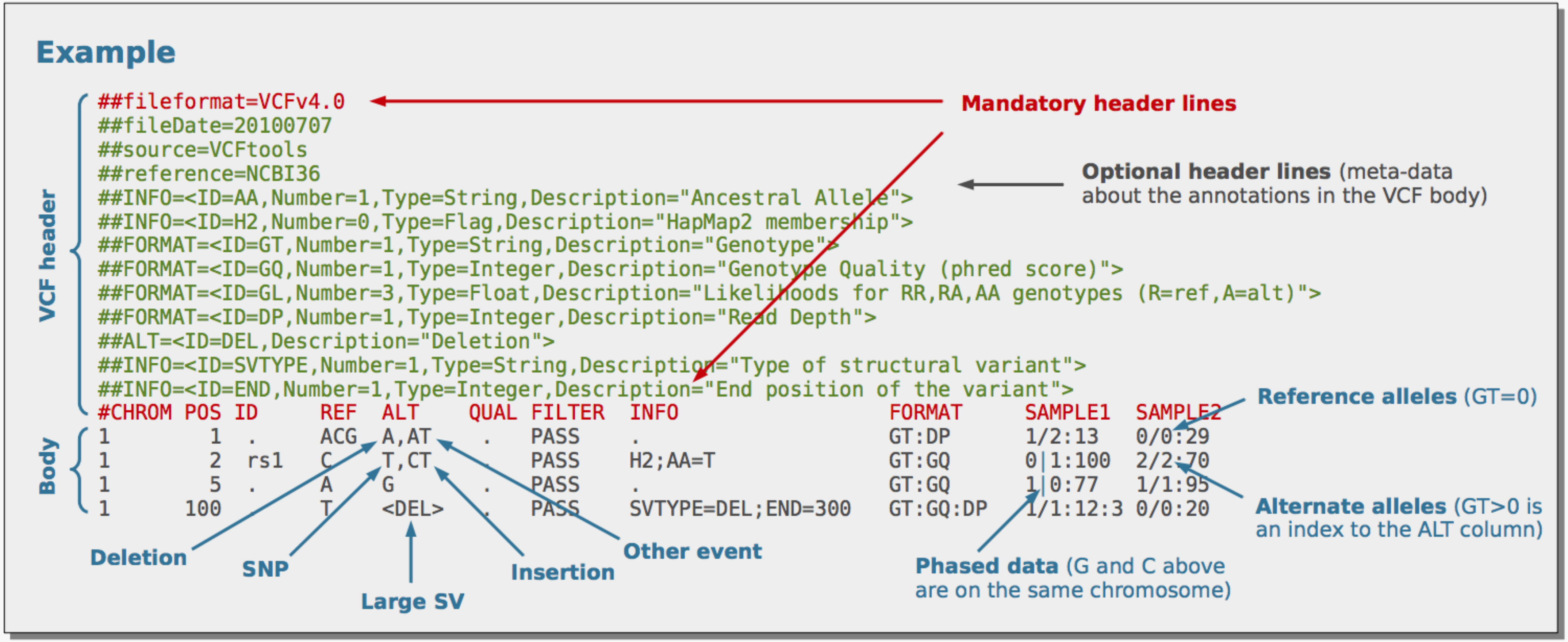


CHR	POS	REF	ALT	W1	W1_Depth	M1	M1_Depth	Genic/Interg	Feature	Description	Organism	Flanking_600bp
chr01	1181121	T	G	K	2,1:3	T	3,0:3	SL000879t0(	Intron	Glycogen pł	Auxenochlor	CTTTCTGCACCG(
chr01	1272122	T	G	K	1,2:3	T	9,1:10	SL014053t0(	Intron	HslV compo	Coccomyxa	ACAGCTGGTCTG
chr01	2586904	T	G	K	4,2:6	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

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

VCF (Variant Call Format) 결과 형태



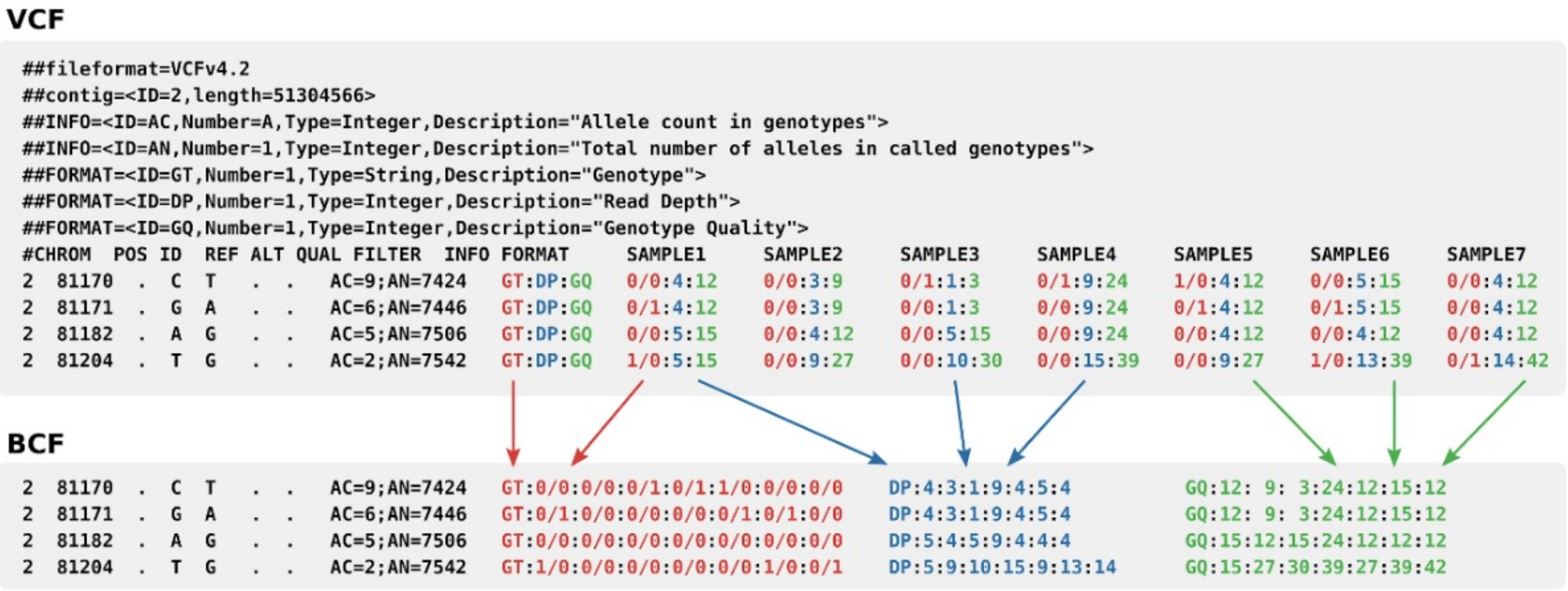
Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID    REF  ALT    QUAL FILTER INFO                FORMAT     SAMPLE1    SAMPLE2
1        1 .     ACG  A,AT   .    PASS   .                   GT:DP      1/2:13     0/0:29
1        2 rs1   C    T,CT   .    PASS   H2;AA=T             GT:GQ      0|1:100    2/2:70
1        5 .     A    G      .    PASS   .                   GT:GQ      1|0:77     1/1:95
1      100 .     T    <DEL>  .    PASS   SVTYPE=DEL;END=300  GT:GQ:DP   1/1:12:3   0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

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

VCF와 BCF의 차이



VCF

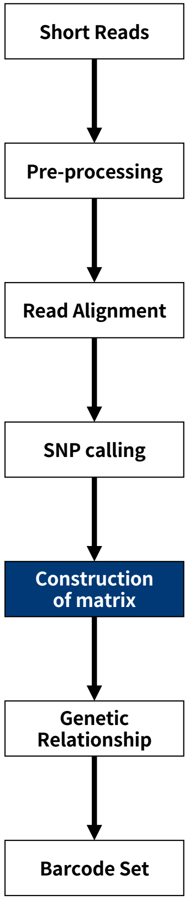
```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read 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:1/0:0/0:0/0:0/0:0/0:1/0:0/1  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



Short Reads

Pre-processing

Read Alignment

SNP calling

Construction of matrix

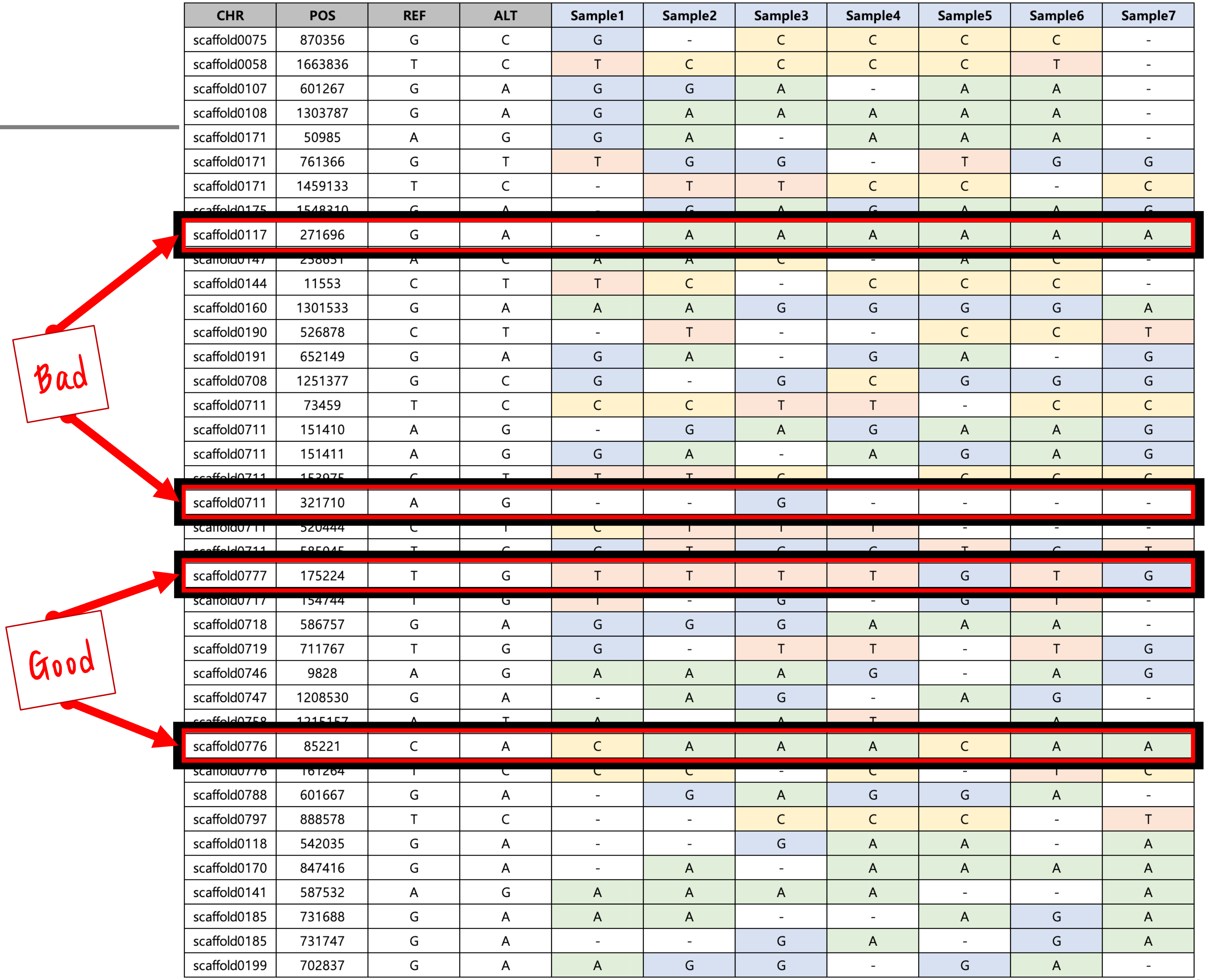
Genetic Relationship

Barcode Set

❑ 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.

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	238651	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	320444	C	T	C	T	T	T	-	-	-
scaffold0711	585045	T	C	C	T	C	C	T	C	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	-

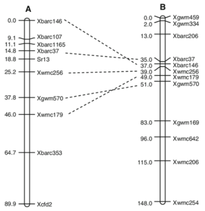
Filtering을 통한 최종 산물

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	-	[illegible]	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	-

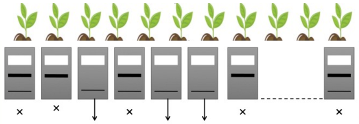
변이를 통해 할 수 있는 것

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



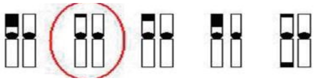
Linkage Map

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



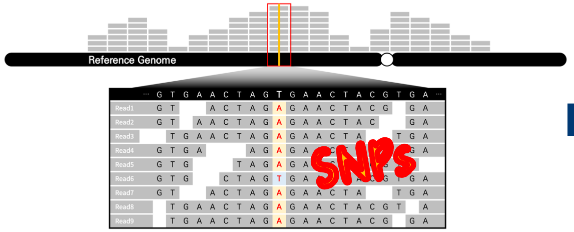
Marker-assisted selection (MAS)

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

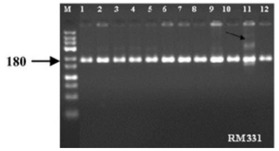


Marker-assisted backcrossing (MAB)

Next-Generation Sequencing

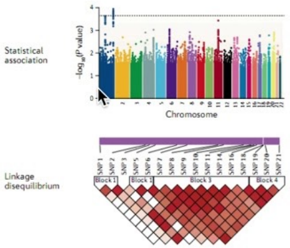

Reference Genome
··· G T G A A C T A G T G A A C T A C G T G A ···
Read1 G T A C T A G A G A A C T A C G G A
Read2 G T A A C T A G A G A A C T A C G A
Read3 T G A A C T A G A G A A C T A T G A
Read4 G T G A A G A G A G A
Read5 G T G T A G A G A
Read6 G T G C T A G T G A A C G T G A
Read7 G T A C T A G A G A A C T A T G A
Read8 T G A A C T A G A G A A C T A C G T A
Read9 T G A A C T A G A G A A C T A C G G A
SNPs

순도 검정 마커



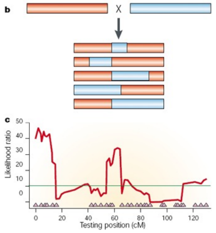
- Bora *et al.* Biotech 6, 50 (2016).

Association Mapping



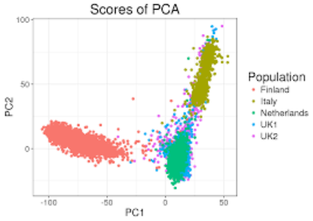
Tam et al 2019

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

강의 요약

- ❑ 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.

Q & A

강의를 경청해 주셔서 감사합니다.

🏢 대전시 유성구 테크노2로 187, B동 412호

✉ bi@bioto.co.kr

☎ 042-710-0077

📠 070-7585-5344