분자마커 개발 이론

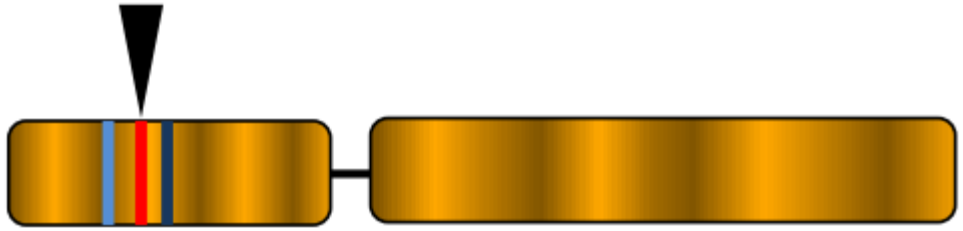
㈜제노믹베이스
이사 남 문
nammoon8406@genomicbase.co.kr
010-2607-5344

분자 마커는 ?

A **molecular marker** is a molecule contained within a sample taken from an organism (biological **markers**) or other matter. ... DNA, for example, is a **molecular marker** containing information about genetic disorders, genealogy and the evolutionary history of life.

Ref) WIKIPEDIA

- **Genetic loci** or **specific DNA sequences** that can be easily tracked and quantified in a population and may be associated with a particular gene or trait of interest.



- Individuals within a population of a sexually reproducing species will have some degree of heritable genomic variation caused by mutations, insertion/deletions (INDELS), in versions, duplications, and translocations.
- genetic loci that may be associated with a particular gene or trait of interest.

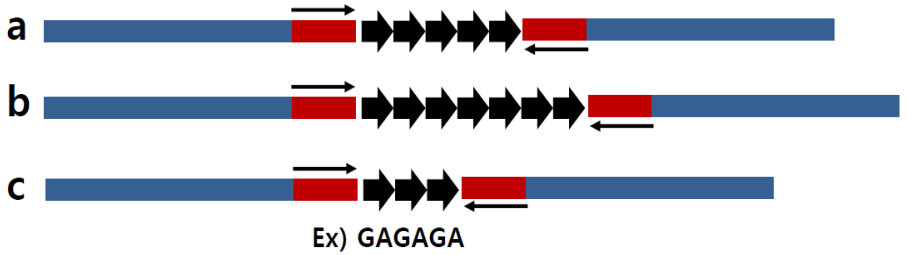
Ref) Hayward et. al. 2015 Methods Mol. Biol.

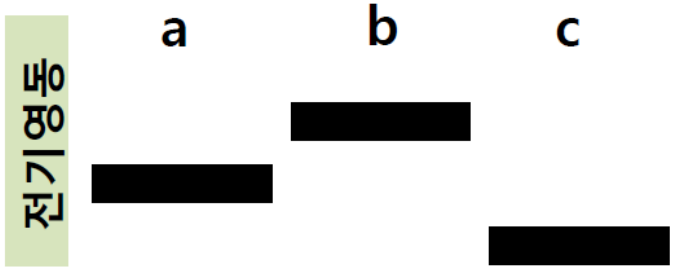
분자 마커는?_분자 마커의 종류

- **RFLP**: Restriction Fragment Length Polymorphism
- **SSLP**: Simple Sequence Length Polymorphism
- **AFLP**: Amplified Fragment Length Polymorphism
- **RAPD**: Random Amplification of Polymorphic DNA
- **SSR**: Simple Sequence Repeat
- **SNP**: Single Nucleotide Polymorphism
- **ISSR**: Inter Simple Sequence Repeat
- **IRAP**: Inter-Retrotransposon Amplified Polymorphism
- **CAPS**: Cleaved Amplified Polymorphism Sequence

최근 주로 사용되는 분자마커

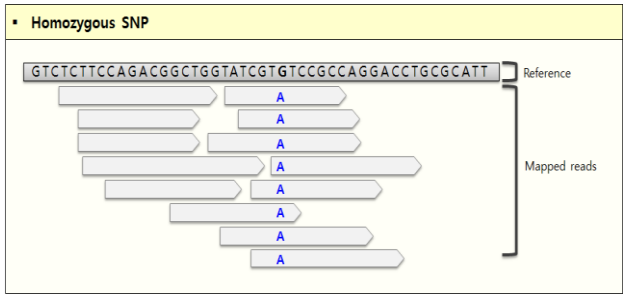
SSR(Simple Sequence Repeat) 마커

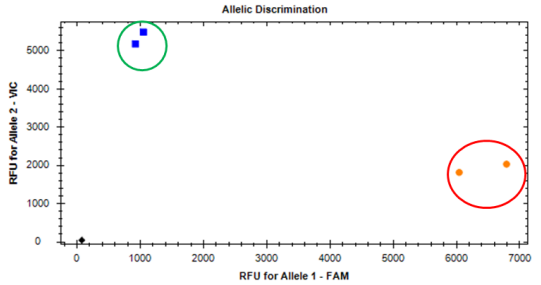


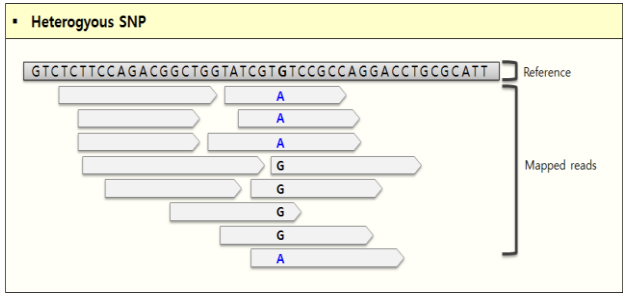


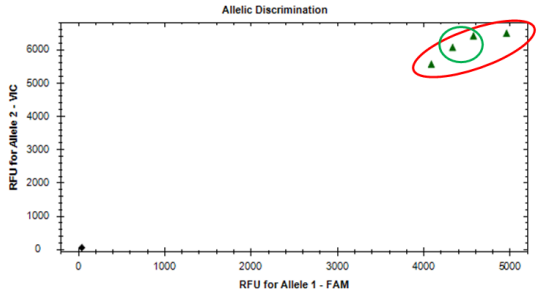
- PCR products를 전기영동 혹은 sanger sequencing을 통해 시퀀스 repeat의 수를 측정함

Homozygous/Heterozygous SNP 마커



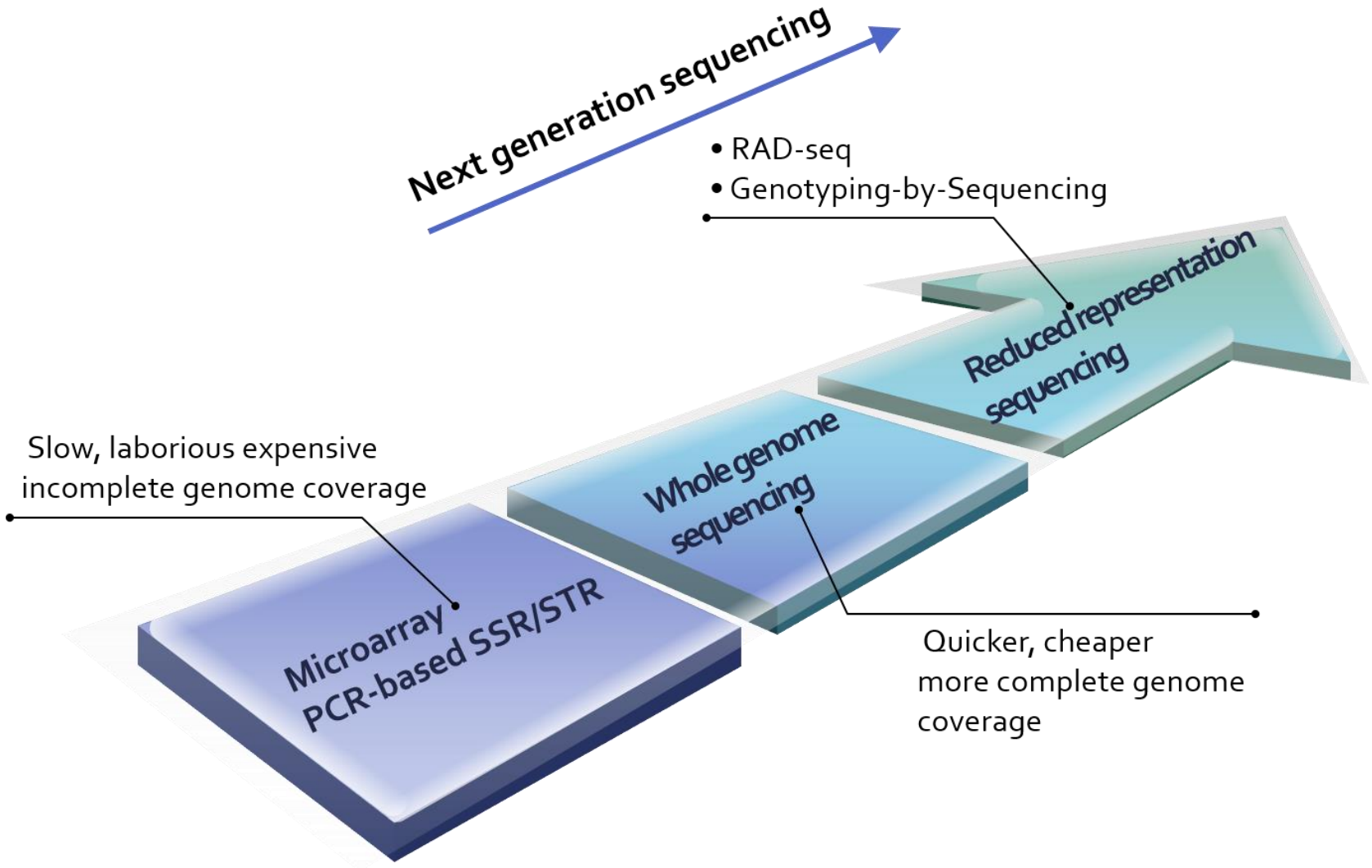






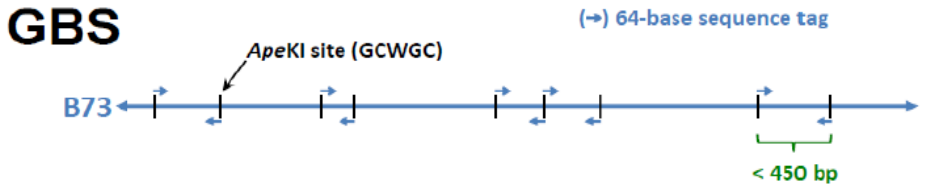
- Real-time PCR 장비를 이용하여 SNP genotyping을 통해 SNP를 측정함

분자마커를 어떻게 찾는가?

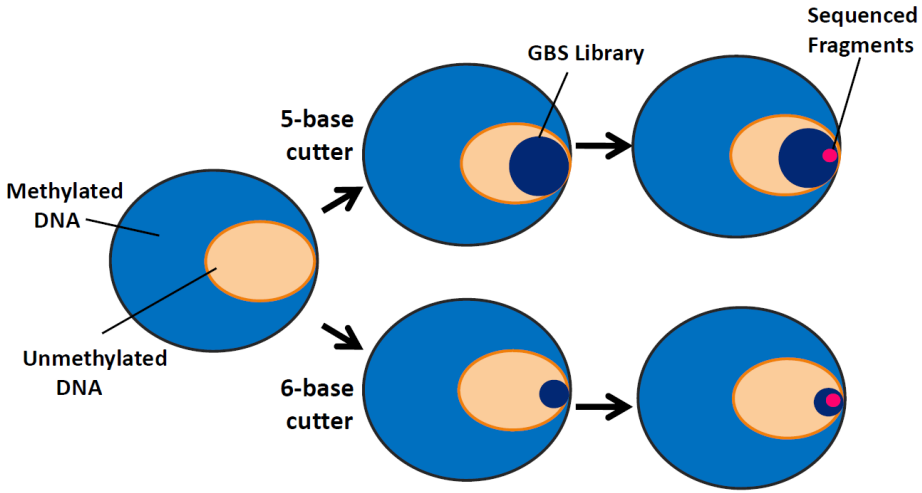


Reduced Genome Representation through GBS

(Genotyping-by-Sequencing)



(Whole Genome Shotgun Sequencing)



Sampling large genomes with methylation sensitive restriction enzymes


Giant Squid


Goose

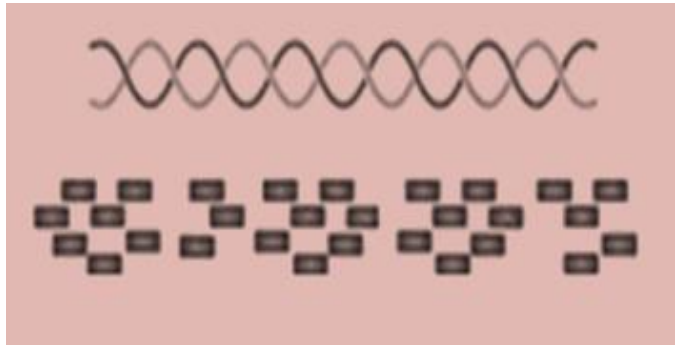

Deer Mouse


Shrub Willow

유전체 분석을 이용한 다양한 분자마커 선발 방법

NGS library

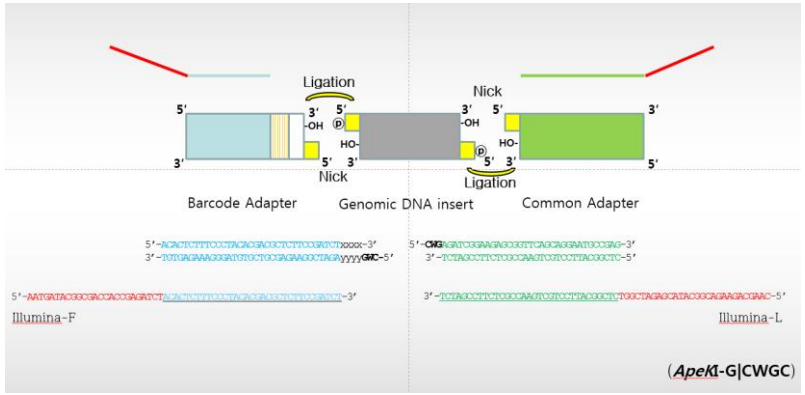
Whole genome sequencing (WGS)



- 게놈 전체 부위를 fragmentation 하여 NGS 시퀀싱 진행
- 게놈 size가 클 경우 시퀀싱 비용이 매우 높아짐
- 일반적으로 집단 보다는 개체가 중요할 경우 진행하여 reference genome 제작에 사용

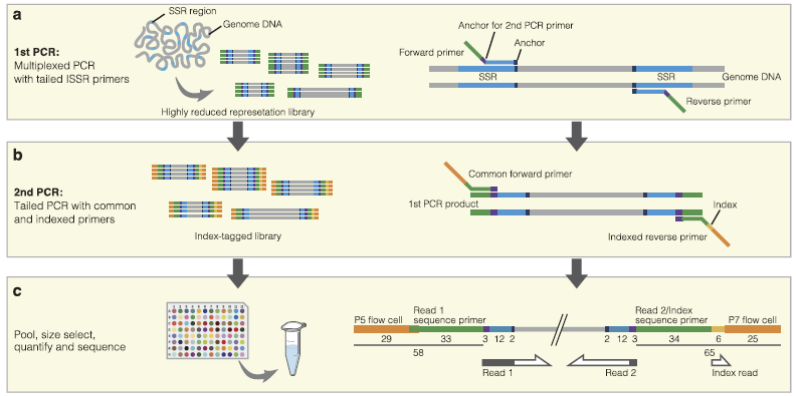
Reduced representation library (RRS)

Genotyping-by-Sequencing (GBS)



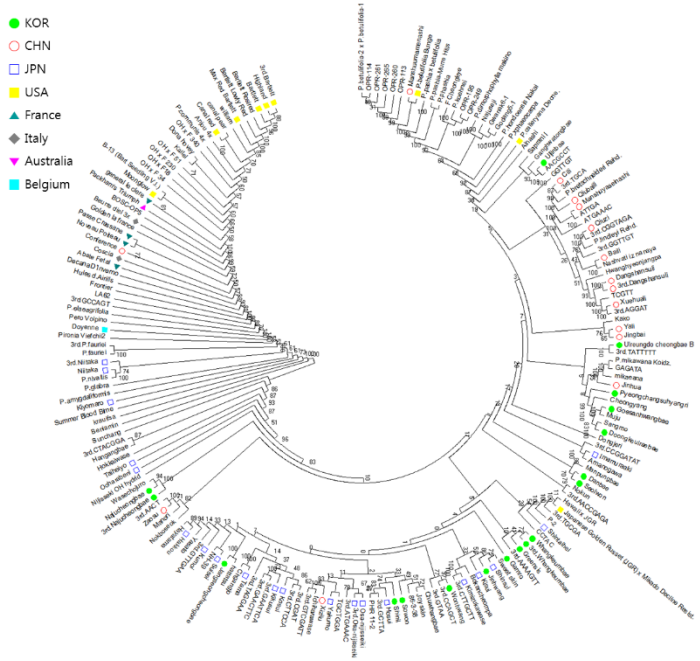
- 게놈의 전체가 아닌 일부분만 대표적으로 시퀀싱이 이루어짐
- 일반적으로 집단(약 96개체 이상) 내 유전적 다양성 분석에 사용
- 제한효소(RE)를 이용하여 라이브러리를 제작함

Multiplexed ISSR genotyping by sequencing (MIG-seq)

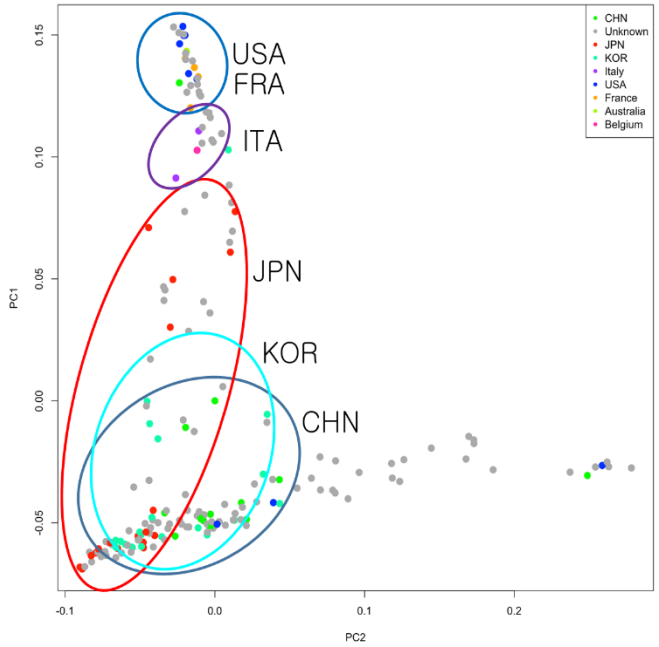


- 게놈의 전체가 아닌 일부분만 대표적으로 시퀀싱이 이루어짐
- 집단(약 96개체 이상) 내 유전적 다양성 분석에 사용하며 GBS library 제작이 어려운 경우 사용함
- Two PCR step을 사용하여 library 제작

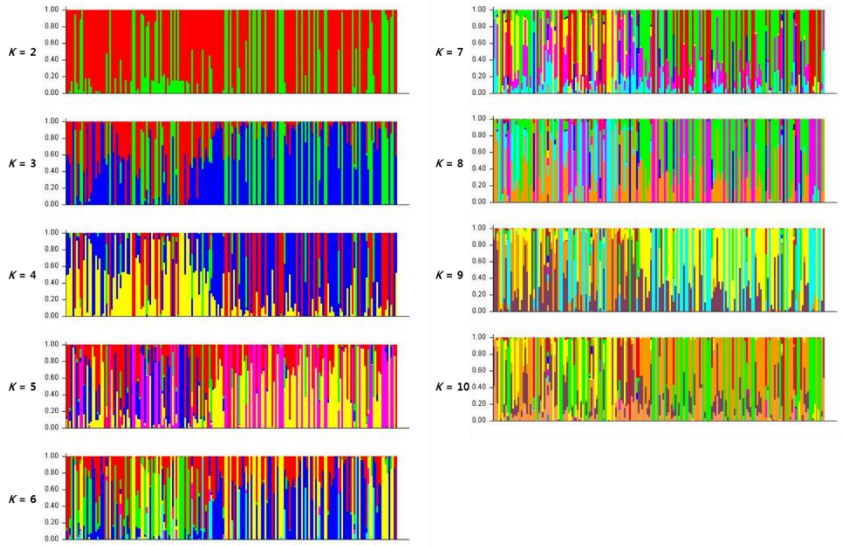
Genome-wide marker를 이용한 집단 분석



과수 유전자원 190개 샘플의 SNP 31,138개를 이용한 Phylogenetic tree



과수 유전자원 190개 샘플의 SNP 10,000개를 이용한 주성분 분석 (PCA)



과수 유전자원 190개 샘플의 SNP 10,000개를 이용한 population structure

Genotyping-by-Sequencing (GBS) 분석

GBS 분석 개요

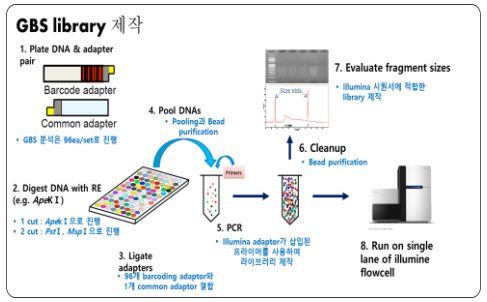
- **Genotyping-by-sequencing (GBS)** is a simple, highly-multiplexed system for constructing reduced representation libraries for use with next generation sequencing technologies such as the Illumina platforms.
- GBS enables the analysis of large numbers of single nucleotide polymorphisms (SNPs) in studies of genetic variation.
- GBS uses restriction enzymes to reduce genome complexity and to avoid the repetitive fraction of the genome.
- Key features of GBS compared to other approaches include:
  - Reduced Sample handling
  - Few PCR & purification steps
  - No DNA size fractionation
  - Efficient barcoding system
  - Simultaneous marker discovery and genotyping
  - Scales very well

GBS 분석을 이용한 집단간 유연관계 분석

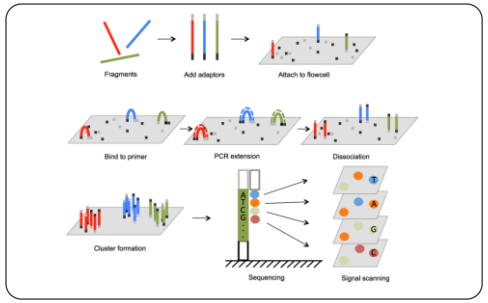
GBS 분석을 위한 시료 준비

- GBS 분석 서비스는 1세트 96샘플을 대상으로 수행
- gDNA는 약 50ng/㎕의 농도로 20-30 ㎕가 필요
- 분리된 gDNA는 agarose gel에서 QC 후 library 제작 진행

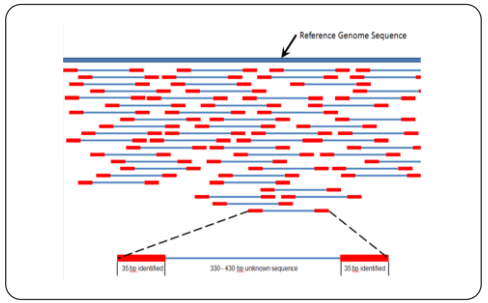
(Genotyping-by-Sequencing) GBS library 제작



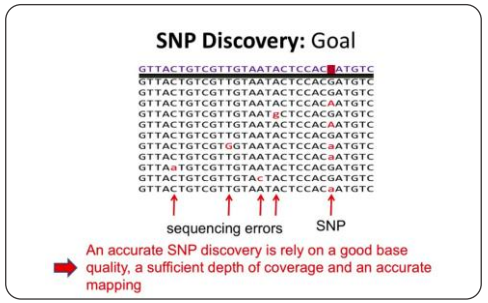
NGS short read sequencing



Read mapping

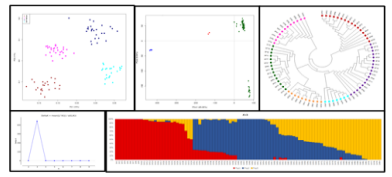


SNP calling

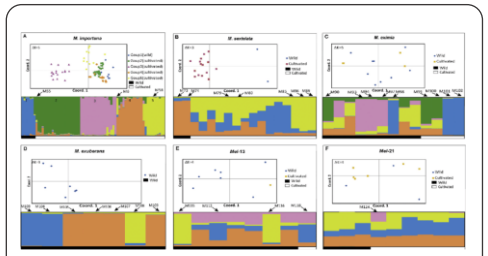


유연관계 분석 (Phylogenetic tree, PCA, Genetic Structure)

대량의 샘플에서 선발된 변이정보를 이용하여 집단/자원의 유연관계 분석 가능



유전다양성 분석 (GeneAlEx)

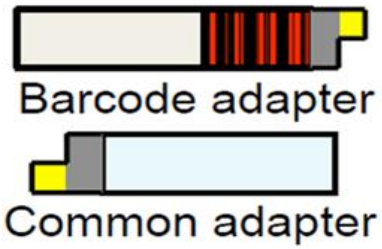


- GBS 분석은 1세트 96개의 시료를 대상으로 유연관계 분석 및 유전다양성 분석에 매우 효과적인 분석 tool 임
- 제한효소(1cut : *Ape*K I, 2cut : *Pst* I and *Msp* I)를 이용하여 gDNA를 절단 한 후 절단된 부위를 중심으로 NGS 시퀀싱을 진행함
- WGS(whole genome sequencing)에 비해 시퀀싱되는 영역을 줄임으로써 대량의 샘플을 효과적으로 분석 가능함
- GBS 분석을 통해 선발된 genome-wide SNP 마커를 이용하여 집단별 유연관계 분석, 유전다양성 분석, 집단별 구분마커 개발, GWAS 분석, QTL mapping 등 다양한 분석 수행이 가능함

GBS 라이브러리 구축 방식

GBS library 제작

1. Plate DNA & adapter pair



- GBS 분석은 96ea/set로 진행

2. Digest DNA with RE (e.g. *Ape*K I)

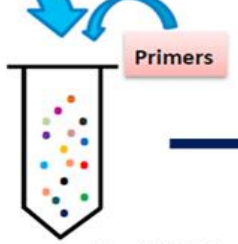
- 1 cut : *Ape*k I 으로 진행
- 2 cut : *Pst* I , *Msp* I 으로 진행

3. Ligate adapters

- 96개 barcoding adaptor와 1개 common adaptor 결합

4. Pool DNAs

- Pooling과 Bead purification

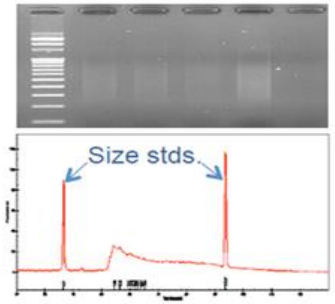


5. PCR

- Illumina adaptor가 삽입된 프라이머를 사용하여 라이브러리 제작

6. Cleanup

- Bead purification



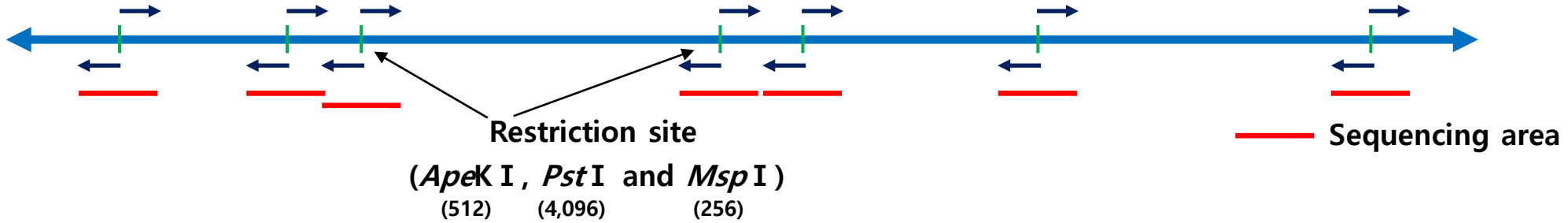
7. Evaluate fragment sizes

- Illumina 시퀀서에 적합한 library 제작

8. Run on single lane of illumine flowcell

GBS 라이브러리 구축 방식

Selective primer(AC) 사용



R.E. cut

*Ape*K I 5′...GCWGC...3′ / 3′...CGWCG...5′

NNNNNNG CWGCNNN.....NNNG CWGCNNNNN
NNNNNNCGWC GNNN.....NNNCGWC GNNNNN

*Ape*K I *Ape*K I

Adaptor ligation

Barcode adaptor Common adaptor

GATCTxxxx CWGCNN.....NNNG CWGAGATCG
CTAGAyyyyGWC GNN.....NNNCGWC TCTAGC

Ligation Ligation

GATCTxxxxCWGCNN.....NNNGCWGAGATCG
CTAGAyyyyGWCGNN.....NNNCGWCTCTAGC

*Ape*K I (X) *Ape*K I (X)

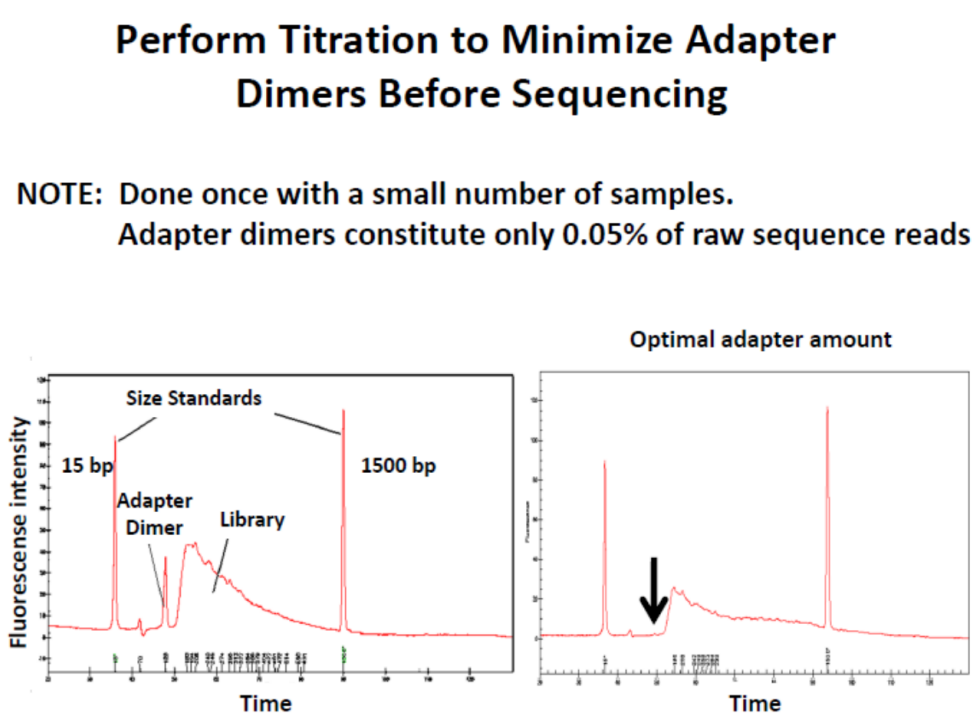
PCR amplification

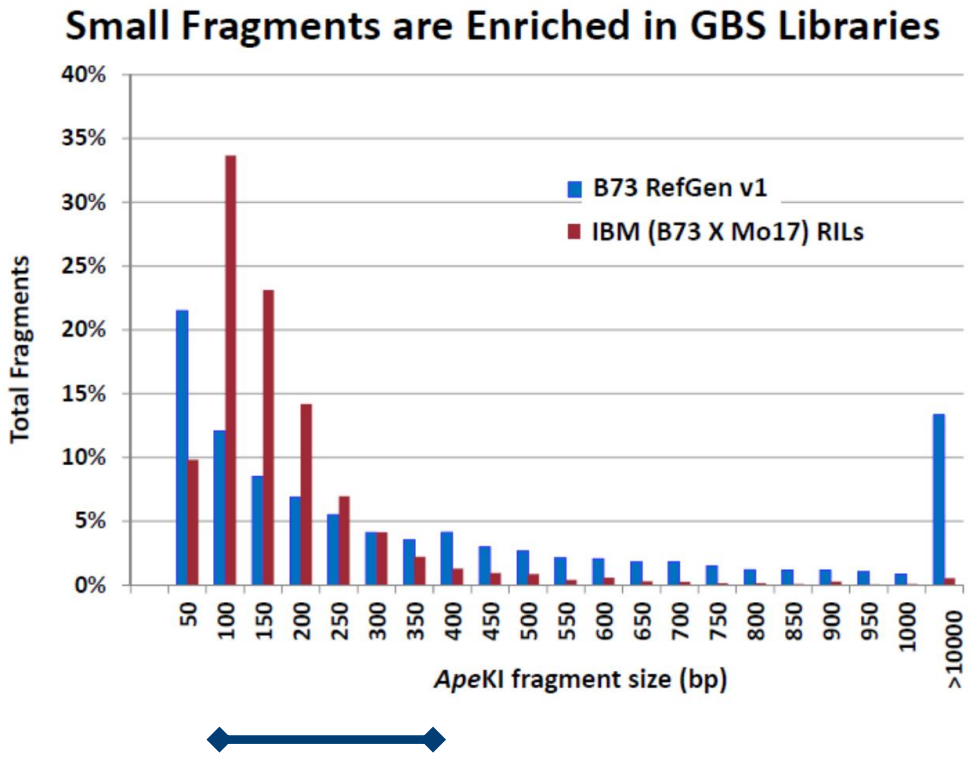
3′-YXCGACTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGCTAGAGCATACGGCAGAAGACGAAC-5′
3′-TCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGCTAGAGCATACGGCAGAAGACGAAC-5′

5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxCWGCNNN NNNGCWGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′
3′-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAyyyyGWCGNNN NNNCGWCTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC-5′

5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

GBS 라이브러리 구축

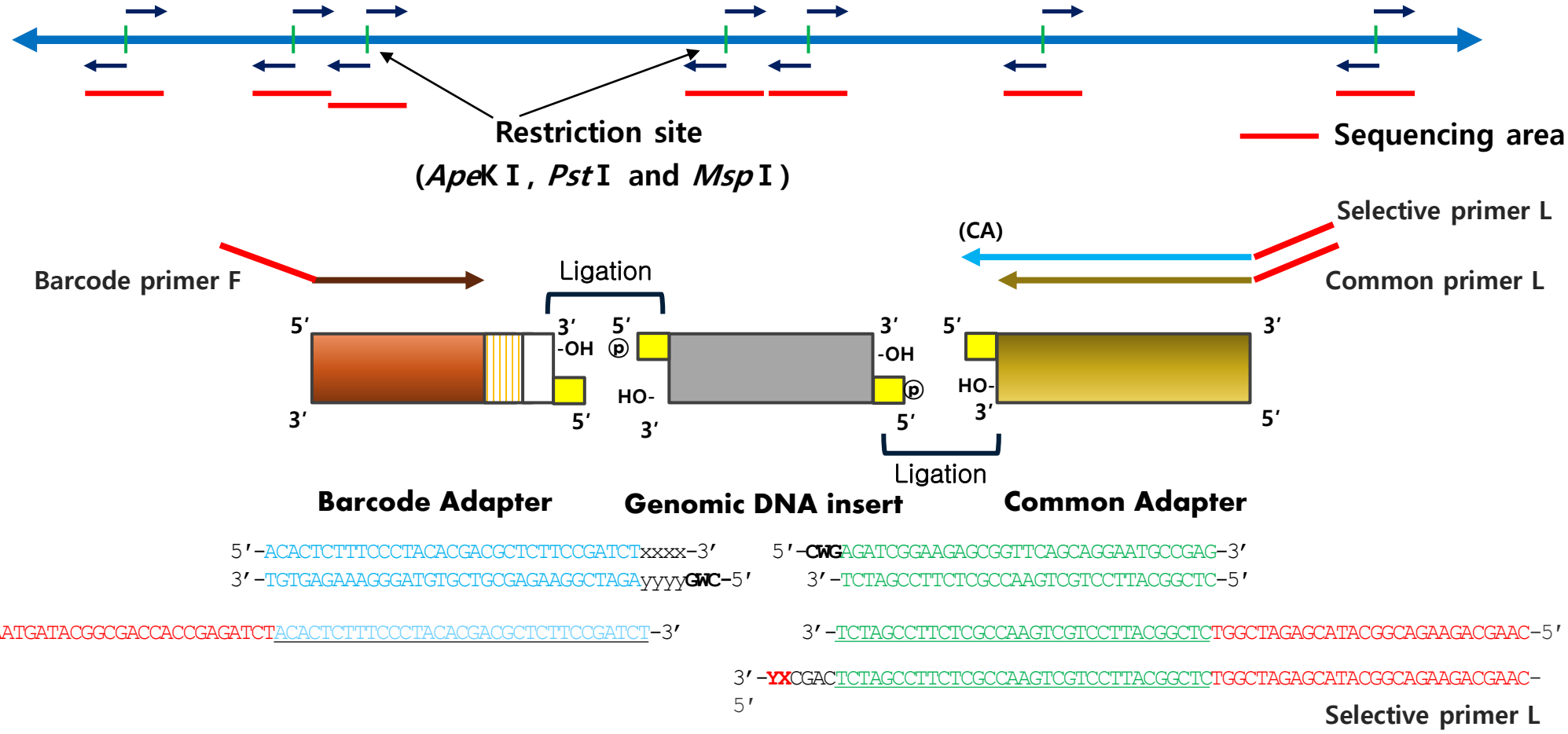




- GBS library는 약 170bp – 350bp 정도에서 주로 fragment가 만들어져야 시퀀싱 효율이 높아짐.
- PCR purificatio을 통해 adaptor dimer 등을 제거 할 수 있음.

GBS 라이브러리 구축 방식

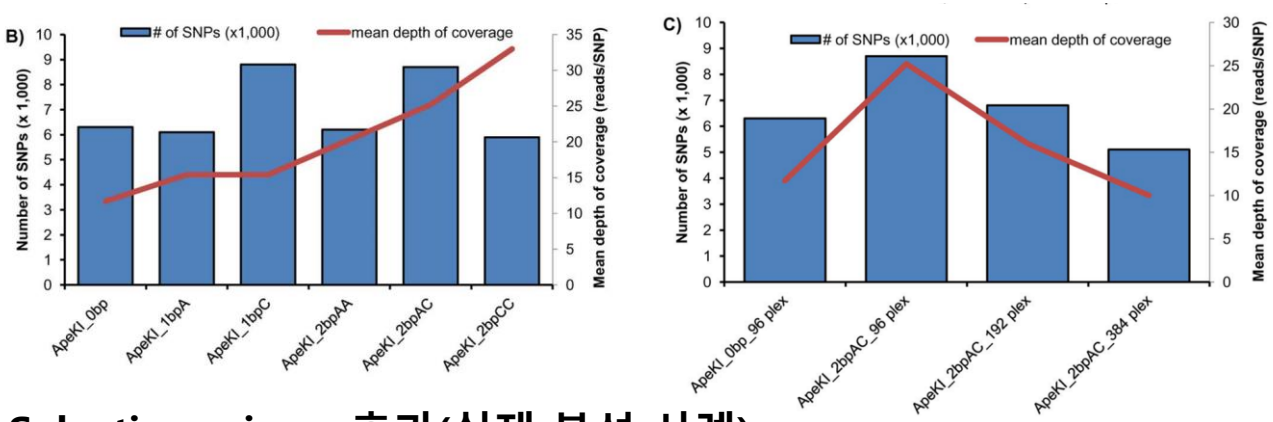
Selective primer(AC) 사용



- 제한효소를 이용하여 라이브러리를 제작하며, 바코딩 프라이머를 통해 96-192개 집단 시료를 대상으로 분자마커 선발이 가능.
- Selective primer 사용으로 GBS 분석의 단점인 missing 비율을 최소화 함으로써, SNP 마커 선발 효율을 높일 수 있음.

GBS 분석 사례

➢ **Selective primer 효과(논문에서 발췌)**



OPEN ACCESS Freely available online
PLOS ONE

An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping

Humira Sonah[1], Maxime Bastien[1], Elmer Iquira[1], Aurélie Tardivel[1], Gaétan Légaré[2], Brian Boyle[2], Éric Normandeau[3], Jérôme Laroche[4], Stéphane Larose[4], Martine Jean[1], François Belzile[1]*

1 Département de Phytologie and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada, 2 Plateforme d'analyses génomiques and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada, 3 Département de Biologie, and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada, 4 Plate-forme de bio-informatique and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada

➢ **Selective primer 효과(실제 분석 사례)**

		단계 별 SNP 좌 수	
No.	필터 단계	Normal GBS	Selective primer
1	Total SNP{ (95개체 통합 matrix)	468,050	**629,213**
2	Missing < 40%	289,964	**346,139**
3	Missing < 30%	250,372	**296,793**
4	MAF > 5%	56,204	**107,118**
5	Missing < 40% and MAF > 5%	7,895	**14,801**
6	Missing < 30% and MAF > 5%	5,489	**10,100**

➢ **실제 효과**

Read depth ↑

SNP 좌 수 ↑

Missing 좌 수 ↓

➢ GBS 라이브러리 제작에 selective primer를 사용할 경우 read mapping 영역의 수를 줄임으로써 read depth를 높이고, missing 좌수를 줄이고, 최종 선발되는 SNP 좌의 수를 높이는 효과가 있음.

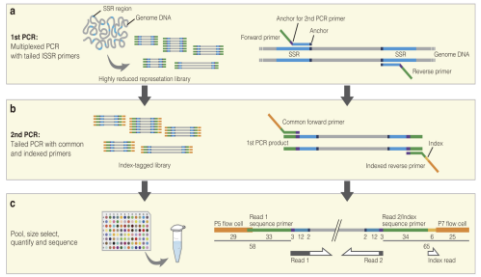
Multiplexed inter-simple sequence repeats genotyping by sequencing (MIG-seq analysis)

MIG-seq 분석을 이용한 집단간 유연관계 분석 서비스

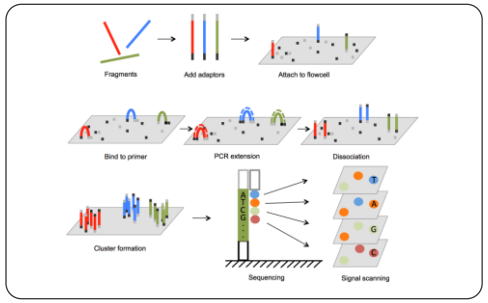
MIG-seq 분석을 위한 시료 준비

- GBS 분석 서비스는 1세트 96샘플을 대상으로 수행
- gDNA는 약 1ng/㎕ 이상의 농도로 10-20 ㎕가 필요하며, PCR이 진행될 수 있을 정도의 amount와 quality가 필요
- 분리된 gDNA를 이용하여 MIG-seq library를 제작한 후 agarose gel에서 QC를 진행 하여 제작 성공 여부를 판단함

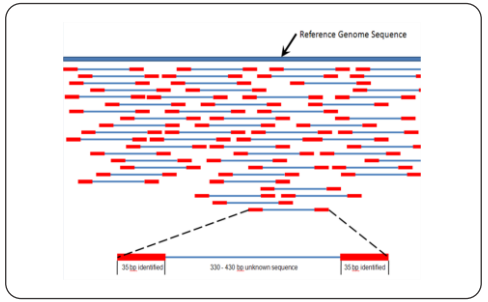
(Genotyping-by-Sequencing) GBS library 제작



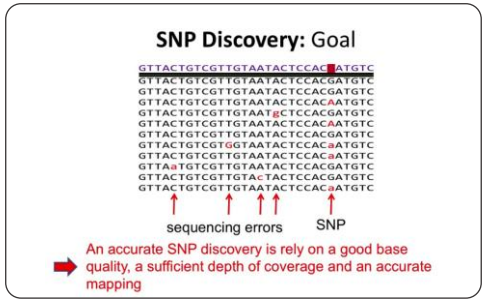
NGS short read sequencing



Read mapping

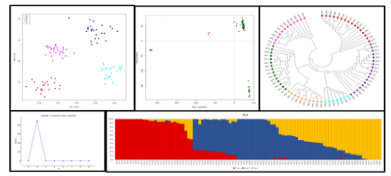


SNP calling

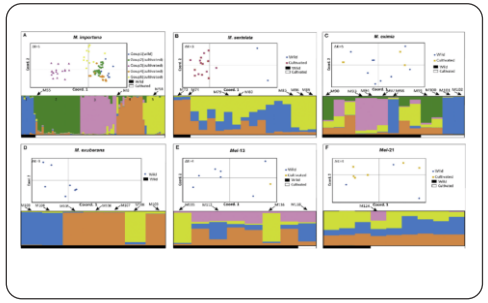


유연관계 분석 (Phylogenetic tree, PCA, Genetic Structure)

대량의 샘플에서 선발된 변이정보를 이용하여 집단/자원의 유연관계 분석 가능

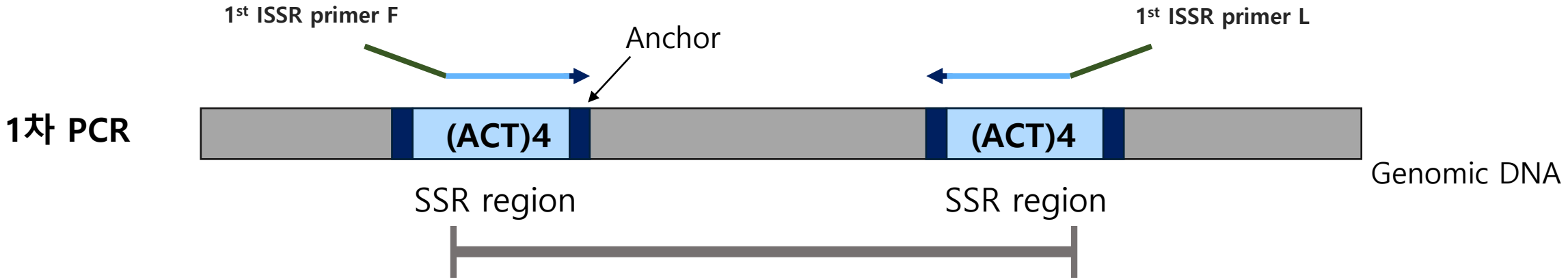


유전다양성 분석 (GeneAlEx)

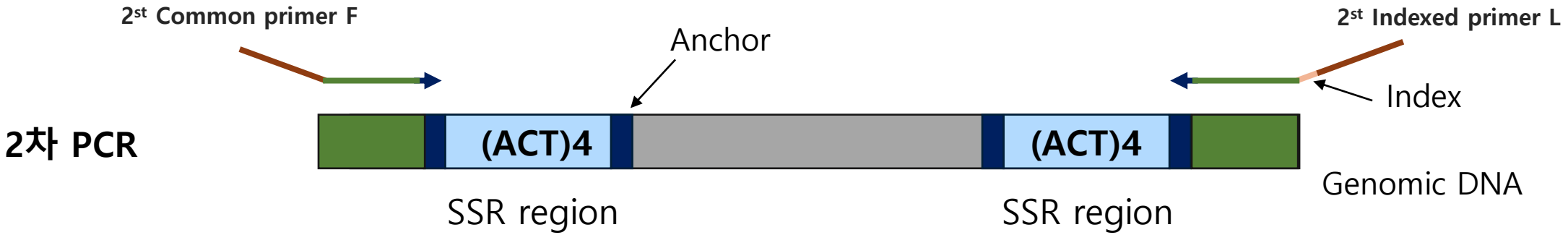


- MIG-seq 분석은 1세트 96개의 시료를 대상으로 유연관계 분석 및 유전다양성 분석에 매우 효과적임.
- MIG-seq 라이브러리는 multiplexed inter-simple sequence repeat (ISSR) primer를 이용하여 제작하며, 게놈 내 수천개의 genome-wide SNP를 선발하여 다양한 분석에 활용이 가능함.
- MIG-seq 분석은 GBS 분석에 비해 선발되는 후보 SNP 마커의 수는 부족하지만 gDNA의 문제로 GBS 분석이 어려운 시료를 대상으로 매우 좋은 대안이 됨 .
- MIG-seq 분석을 통해 선발된 genome-wide SNP 마커를 이용하여 집단별 유연관계 분석, 유전다양성 분석, 집단별 구분 마커 개발, linkage map 작성, GWAS 분석, QTL mapping 등 다양한 분석 수행이 가능함.

MIG-seq 라이브러리 구축 방식



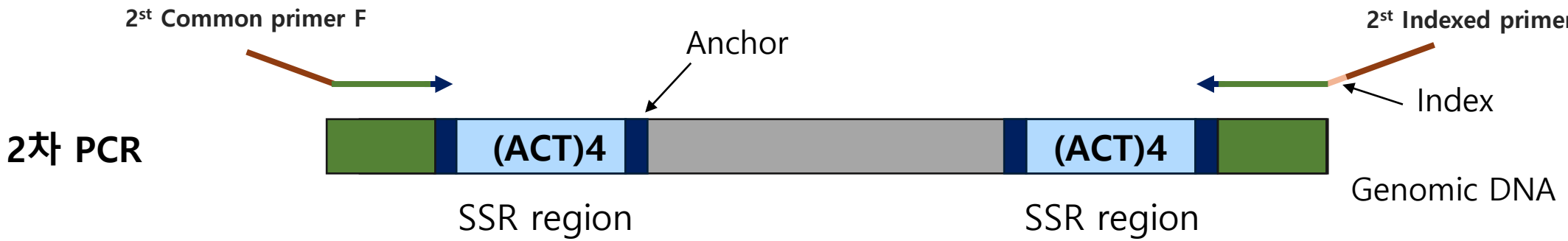
➢ 2개의 SSR motif 내의 영역을 시퀀싱 하여 polymorphic SNP 마커를 탐색함



➢ PCR 기반의 라이브러리 구축을 통해 소량의 gDNA 혹은 quality가 낮은 gDNA를 대상으로 유전체 분석을 수행할 수 있음.

➢ MIG-seq 분석은 GBS 분석에 비해 약 1/10 정도의 SNP 마커 선발 효율을 갖음.

MIG-seq 라이브러리 구축 방식

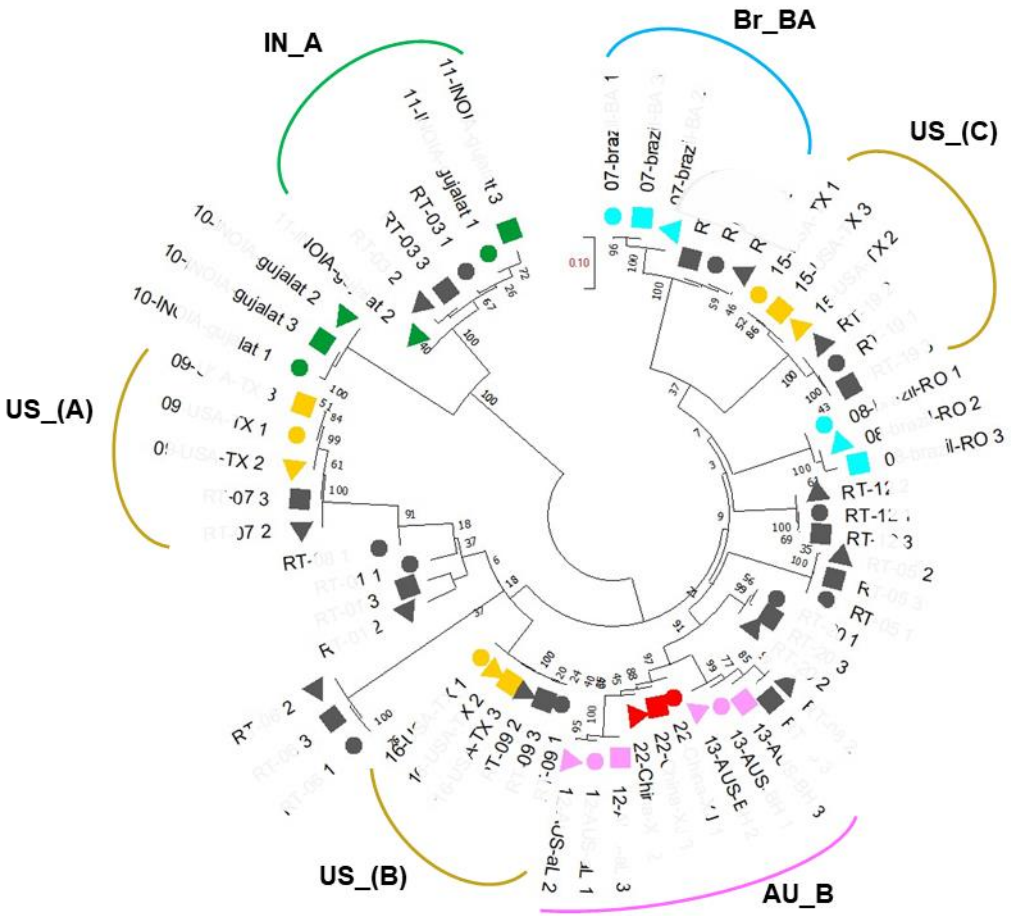


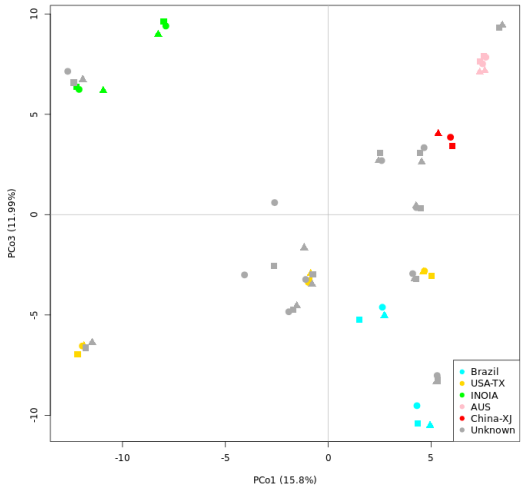
No.	Primer set name	Primer name	Sequence (5' - 3')
1	PS1_4-5	ACT4TG-f_4-5	CGCTCTTCCGATCTCTGACTACTACTNNTTG
2	PS1_4-5	CTA4TG-f_4-5	CGCTCTTCCGATCTCTGCTACTACTANNATG
3	PS1_4-5	TTG4AC-f_4-5	CGCTCTTCCGATCTCTGTTGTTGTTGNNGAC
4	PS1_4-5	GTT4CC-f_4-5	CGCTCTTCCGATCTCTGGTTGTTGTTNNTCC
5	PS1_4-5	GTT4TC-f_4-5	CGCTCTTCCGATCTCTGGTTGTTGTTNNTTC
6	PS1_4-5	GTG4AC-f_4-5	CGCTCTTCCGATCTCTGGTGGTGGTGNNGAC
7	PS1_4-5	GT6TC-f_4-5	CGCTCTTCCGATCTCTGGTGTGTGTGNNTTC
8	PS1_4-5	TG6AC-f_4-5	CGCTCTTCCGATCTCTGTGTGTGTGTNNGAC

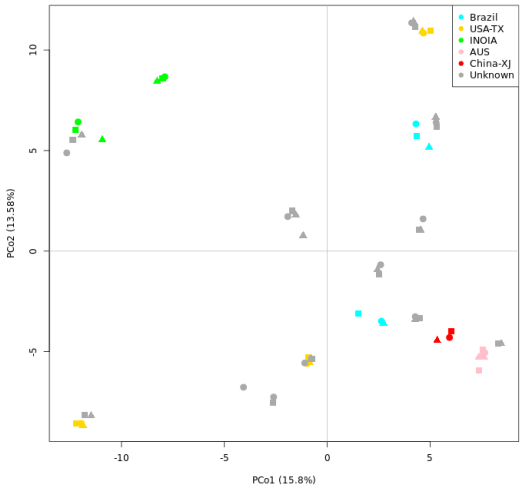
No.	Primer set name	Primer name	Sequence (5' - 3')
1	PS1_4-5	ACT4TG-r_4-5	TGCTCTTCCGATCTGACACTACTACTNNTTG
2	PS1_4-5	CTA4TG-r_4-5	TGCTCTTCCGATCTGACCTACTACTANNATG
3	PS1_4-5	TTG4AC-r_4-5	TGCTCTTCCGATCTGACTTGTTGTTGNNGAC
4	PS1_4-5	GTT4CC-r_4-5	TGCTCTTCCGATCTGACGTTGTTGTTNNTCC
5	PS1_4-5	GTT4TC-r_4-5	TGCTCTTCCGATCTGACGTTGTTGTTNNTTC
6	PS1_4-5	GTG4AC-r_4-5	TGCTCTTCCGATCTGACGTGGTGGTGNNGAC
7	PS1_4-5	GT6TC-r_4-5	TGCTCTTCCGATCTGACGTGTGTGTGNNTTC
8	PS1_4-5	TG6AC-r_4-5	TGCTCTTCCGATCTGACTGTGTGTGTNNGAC

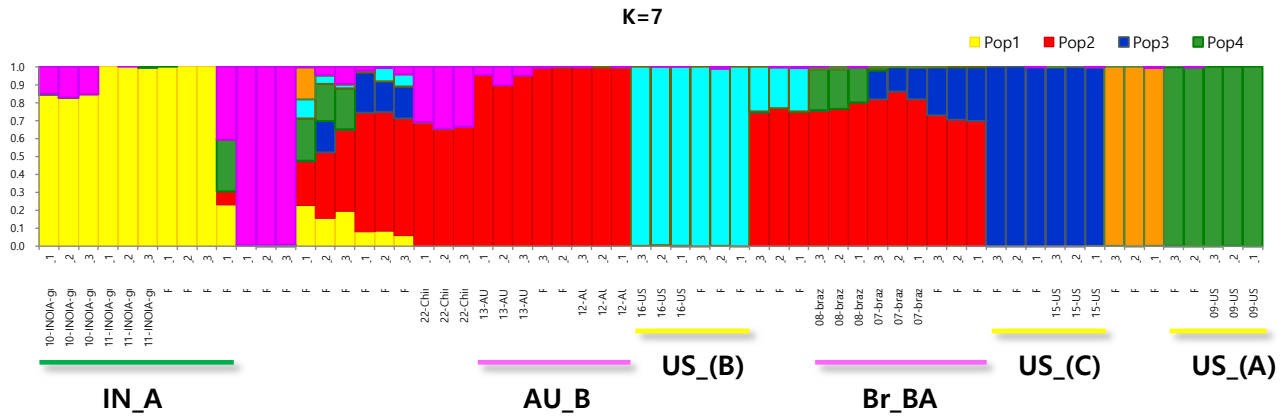
- SSR region을 대상으로 프라이머를 디자인 하며 다양한 SSR motif를 대상으로 PCR 증폭이 일어날 수 있도록 프라이머를 제작함
- 최근 연구에 따르면 SSR motif 앞에 NN(degenerate oligonucleotide)을 삽입하여 유전자 다형성을 탐색하는데 더 효율적인 것으로 확인함.

면화_MIG-seq 분석_원산지 구분 마커 선발(사례)









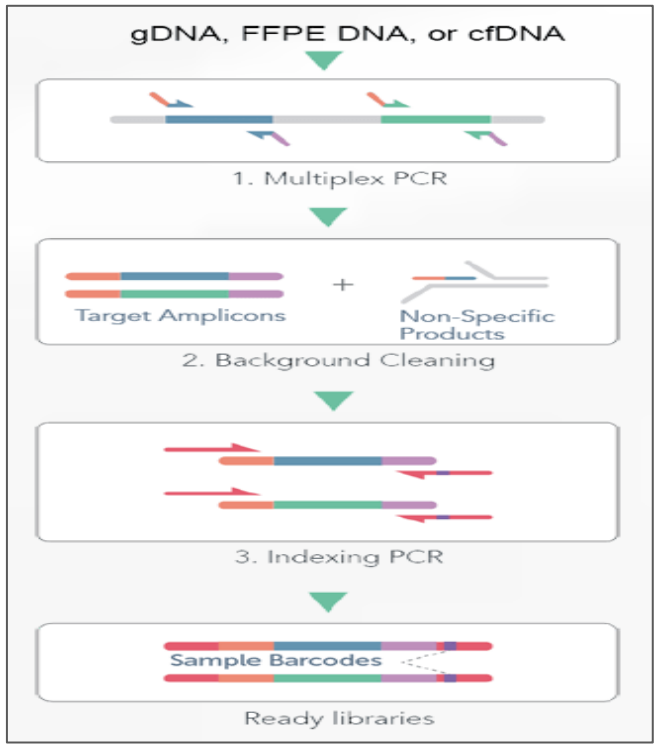
- 면화의 섬유 조직에서 미량의 gDNA를 분리하여 면화가 생산되는 원산지를 구분할 수 있는 마커를 선발함.
- 선발된 1,081개의 SNP 마커를 이용하여 원산지를 모르는 미지의 면사를 대상으로 원산지 확인함.

Targeted sequencing

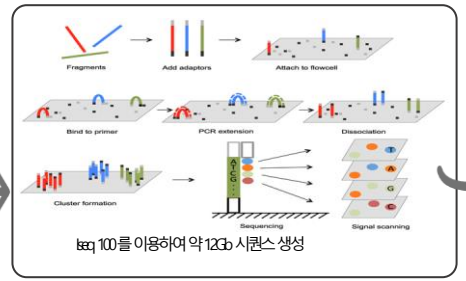
- Multiplex amplicon 방식
- Hybrid bait 방식

Targeted sequencing_Multiplex amplicon 방식

Multiplex amplicon 방식의 library 제작

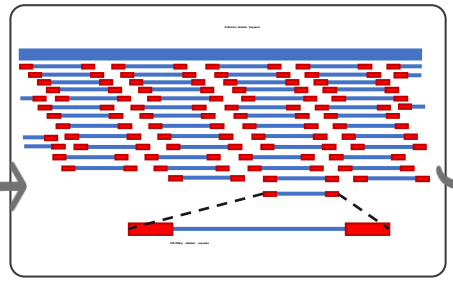


NGS sequencing

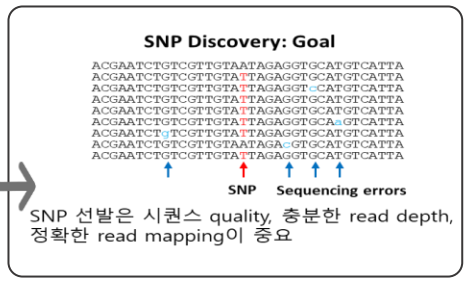


Illumina 100를 이용하여 약 120Gb 시퀀스 생성

Short read mapping



SNP calling



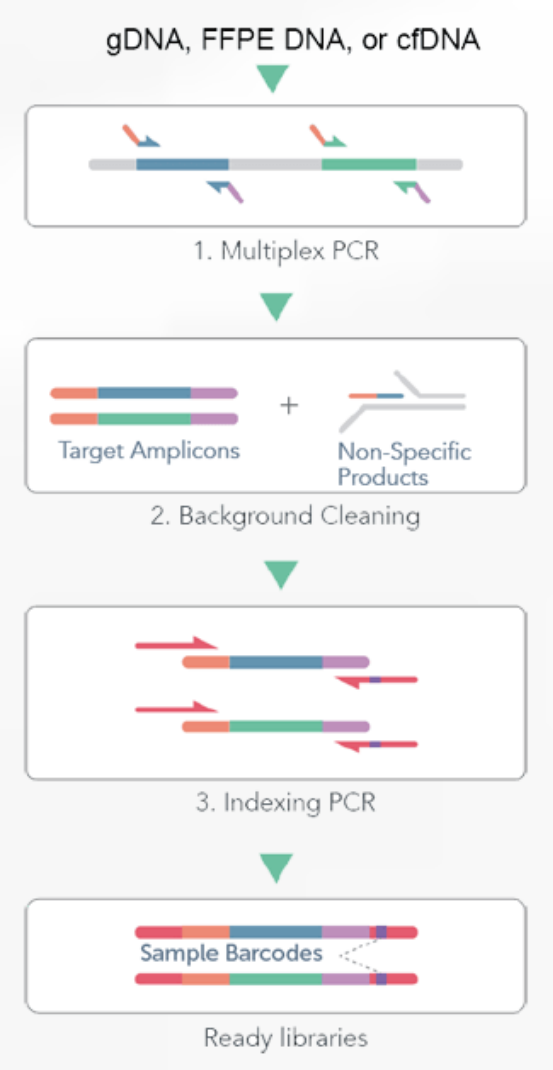
SNP 선발은 시퀀스 quality, 충분한 read depth, 정확한 read mapping이 중요

Applied analysis

- Multiplex amplicon 기반의 targeted 시퀀싱은 기 선발된 genome-wide SNP(약 300-400여개) 마커를 대상으로 multiplex PCR 방식으로 원하는 부위만을 시퀀싱 함.
- 집단 시료를 대상으로 원하는 SNP 스크리닝에 강점을 가지고 있음.
- 주요 응용 분석 : MABC genotyping, Genomic selection (GS), QTL screening, High throughput genotyping

- 두 방식 모두 96샘플/1세트로 분석을 수행하며, reference 게놈 서열내 타겟 영역의 염기서열 정보를 이용하여 사전에 **Amplicon primer 세트** 혹은 **Hybrid bait 세트**의 제작이 필요함.
- **Multiplex amplicon 방식** : 96샘플 기준으로 약 300 여개의 SNP 정보를 지노타입핑 및 분석하는데 효과적임.

Multiplex amplicon 방식의 library 제작

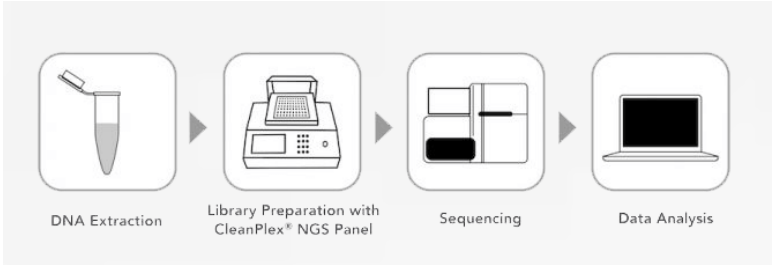


CleanPlex DNA Target Enrichment and Library Preparation Workflow

Step 1 PCR

Step 2 PCR

- The most advanced NGS amplicon sequencing technology for targeted DNA and RNA seq



Targeted DNA Sequencing Workflow

Dual-Indexed PCR Primers for Illumina

Primer Sequences

Each sample is indexed by a pair of Indexed PCR Primers for sequencing on Illumina platforms. XXXXXXXX denotes the index region of the primer. Index sequences are listed below.

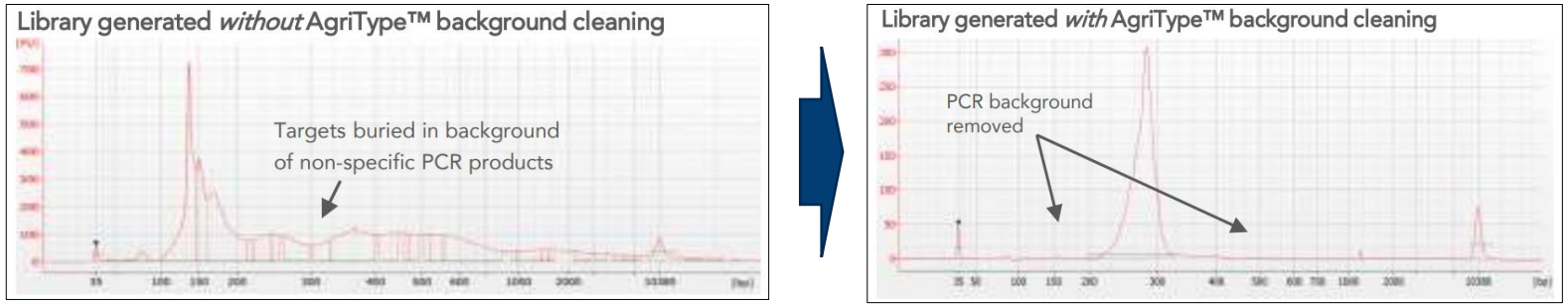
i5 Indexed Primer :
AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATC*T

i7 Indexed Primer :
CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T

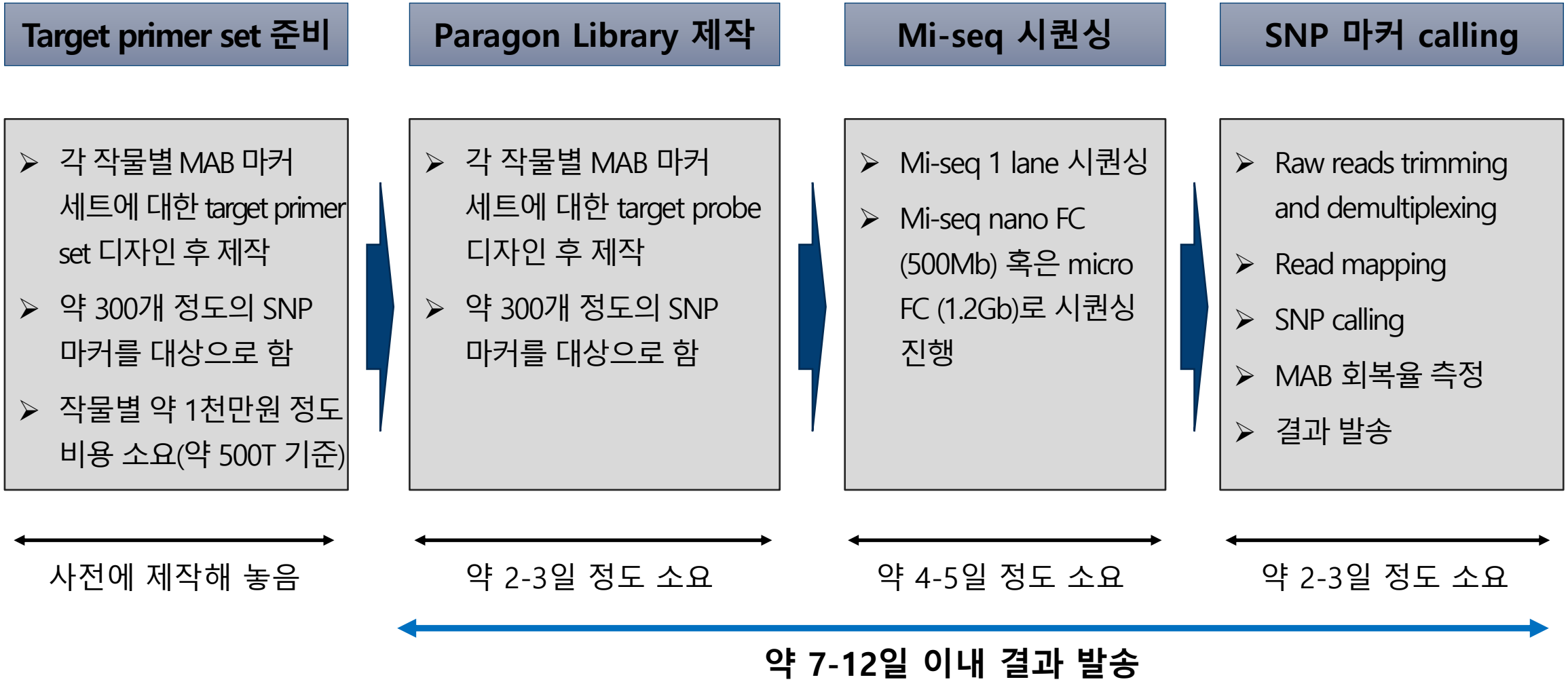
Targeted sequencing

PARAGON GENOMICS

➢ Targeted genotyping 라이브러리 제작 과정

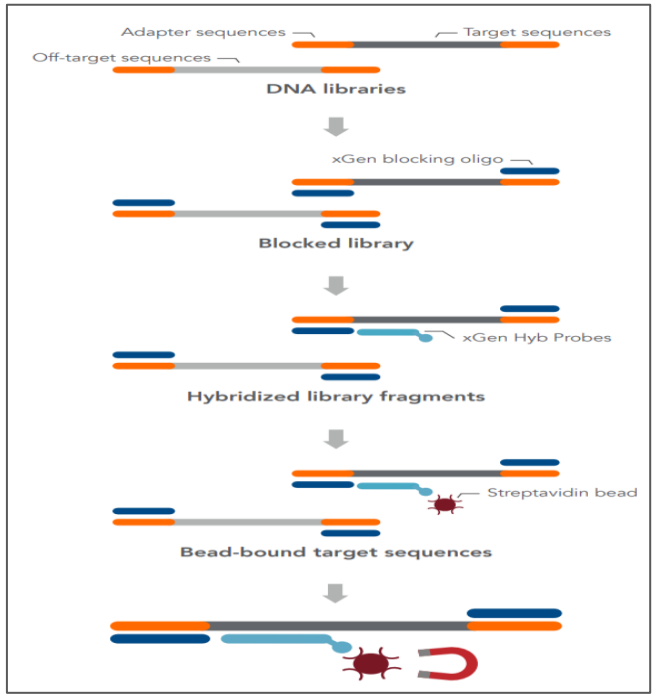


MAB genotyping workflow

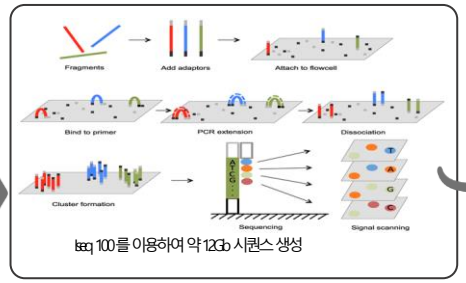


Targeted sequencing_Hybrid bait 방식

Hybrid bait 방식의 library 제작

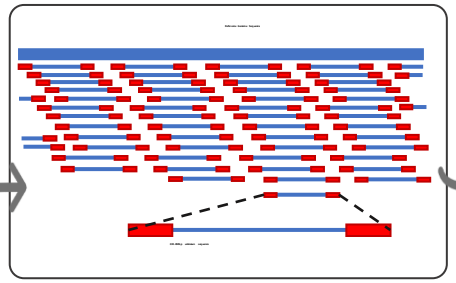


NGS sequencing

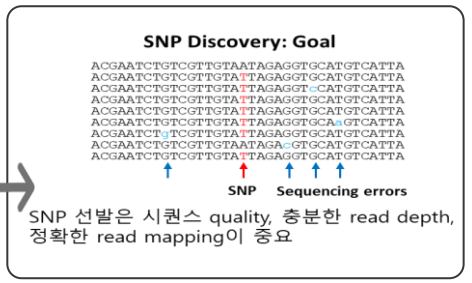


Illumina 100를 이용하여 약 12Gb 시퀀스 생성

Short read mapping



Extract consensus sequence



SNP 선발은 시퀀스 quality, 충분한 read depth, 정확한 read mapping이 중요

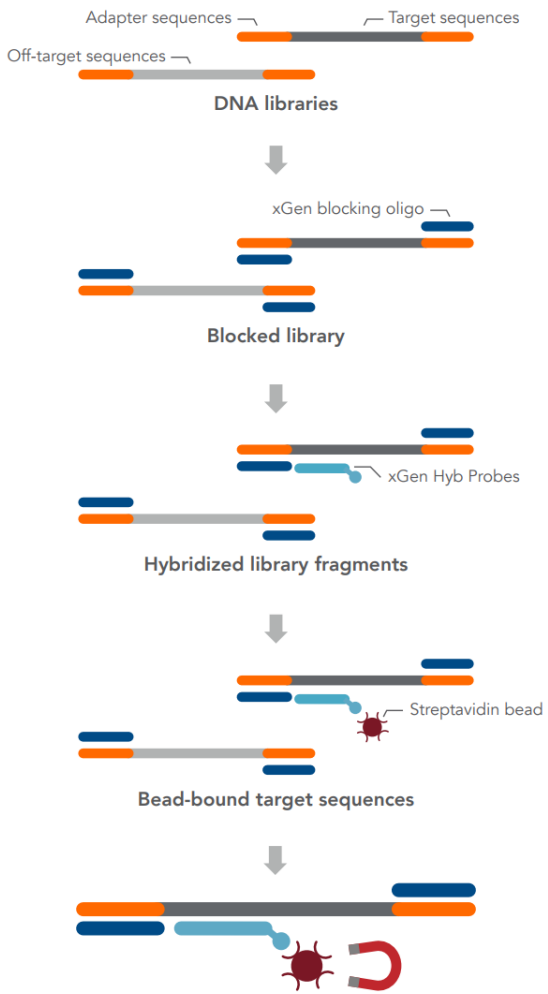
Applied analysis

- Hybrid bait 기반의 targeted 시퀀싱은 target 부위를 대상으로 oligo bait를 제작하여 hybrid 방식으로 게놈 내 특정 영역의 contig만을 시퀀싱 함.
- 집단 시료를 대상으로 타겟 유전자 부위의 분석에 강점을 가지고 있음.
- 주요 응용 분석 : Target gene sequencing, Multi-locus sequencing, Exome sequencing , Gene discovery, High throughput genotyping

- 두 방식 모두 96샘플/1세트로 분석을 수행하며, reference 게놈 서열내 타겟 영역의 염기서열 정보를 이용하여 사전에 **Amplicon primer 세트** 혹은 **Hybrid bait 세트**의 제작이 필요함.
- **Hybrid bait 방식** : 96샘플 기준으로 약 100Kb 내의 타겟 영역을 효과적으로 시퀀싱 및 분석하는데 유용함.

Hybrid bait 방식의 library 제작

Target capture workflow



Add blocking oligos and Cot-1 DNA

Hybridize targets to capture probes

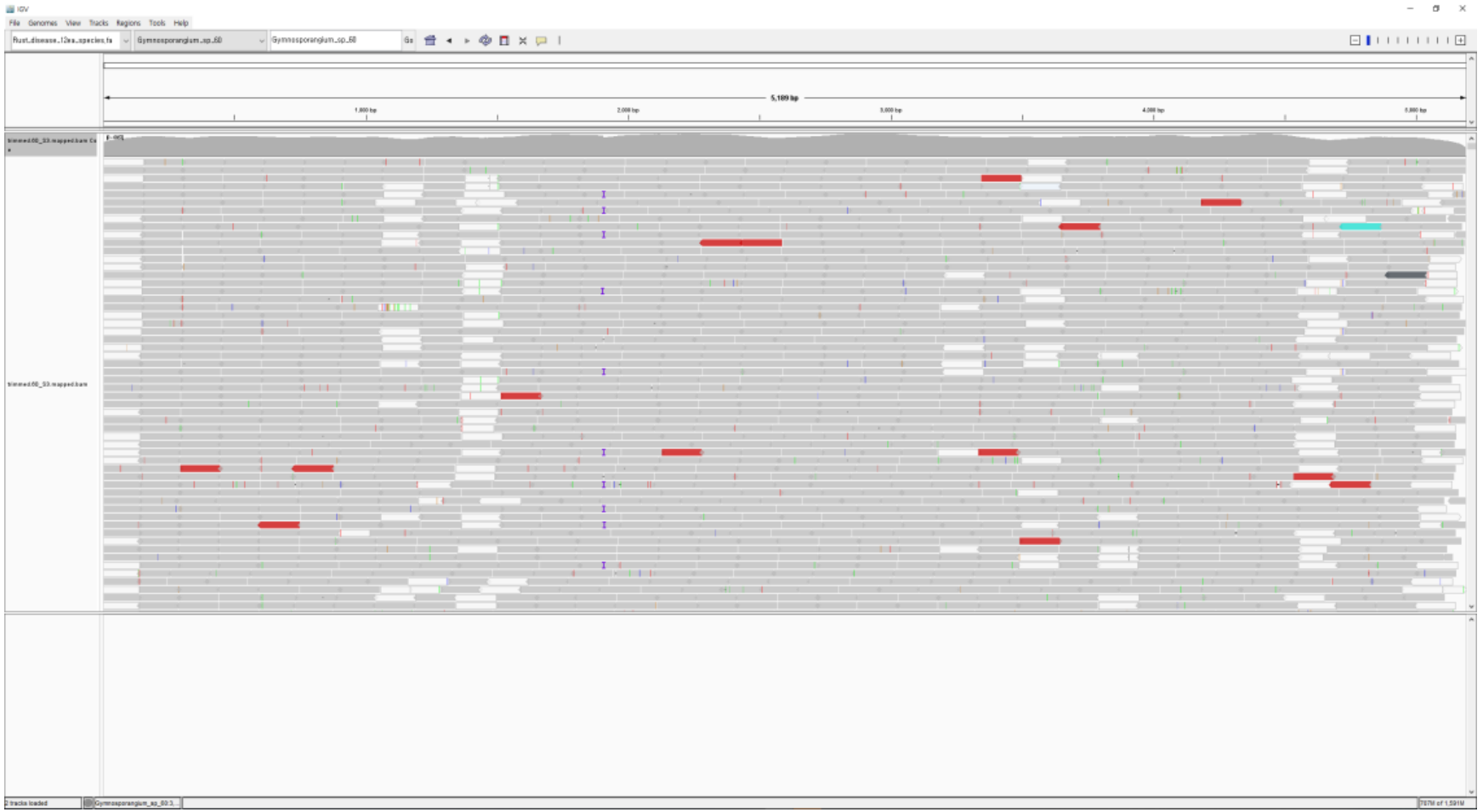
Incubate with magnetic streptavidin beads

Isolate targets with magnet

1	Combine DNA with blockers Dry down DNA ⛔ Perform hybridization incubation	Total time: 15 minutes Total time: Variable Total time: 4–16 hours
2	Prepare buffers	Total time: 15 minutes*
3	Wash streptavidin beads	Total time: 15 minutes*
4	Bead capture incubation	Total time: 45 minutes
5	Perform post-capture washes	Total time: 30 minutes
6	Perform post-capture PCR ⛔	Total time: 30 minutes
7	Post-capture PCR clean up ⛔	Total time: 30 minutes

* Perform during hybridization reaction

Target 부위의 read mapping 결과 (예시)



- Target 부위를 대상으로 hybrid bait를 제작하여 targeted sequencing 진행 후 read mapping 결과 위와 같이 target 부위에 특이적으로 read mapping이 된 것을 확인 할 수 있었음.

요약

- 차세대염기서열분석방법(NGS)의 개발을 통해 다양한 분자마커 선발 기술이 개발됨.
- 분자마커 선발 기술은 Whole genome sequencing(WGS) 보다는 Reduced representation sequencing(RRS) 기술을 접목하여 집단 시료를 대상으로 polymorphic SNP 마커를 탐색 및 스크리닝 하는 분석으로 진화하고 있음.
- 대표적인 RRS 분석은 GBS 분석과 MIG-seq 분석이 주로 활용되고 있으며, MIG-seq 분석의 경우 시료의 특성상 gDNA의 qualit가 낮고 양이 적을 경우 매우 유용한 분석 방식임.
- GBS 혹은 MIG-seq 분석을 통해 polymorphic SNP 마커 세트가 특정될 경우에는 targeted sequencing 기술을 통해 동일한 마커를 대상으로 SNP genotyping이 가능한 분석 기술도 최근 많이 활용되고 있음.
- 다양한 분자마커 선발 기술을 이용하여 선발된 polymorphic SNP 마커는 이후에 유연관계분석, 품종(원산지) 구분 마커 선발, GWAS 분석, QTL-mapping, Linkage map 제작 등 다양한 응용 분석에 활용 가능함.