Summary & More

: Whole genome sequencing (NGS)

Sung Wook Chi

Division of Life Sciences, Korea University

What we learned

1. Introduction (from Genomics to Functional Genomics)

2. Genome Projects
3. Human Genome Projects

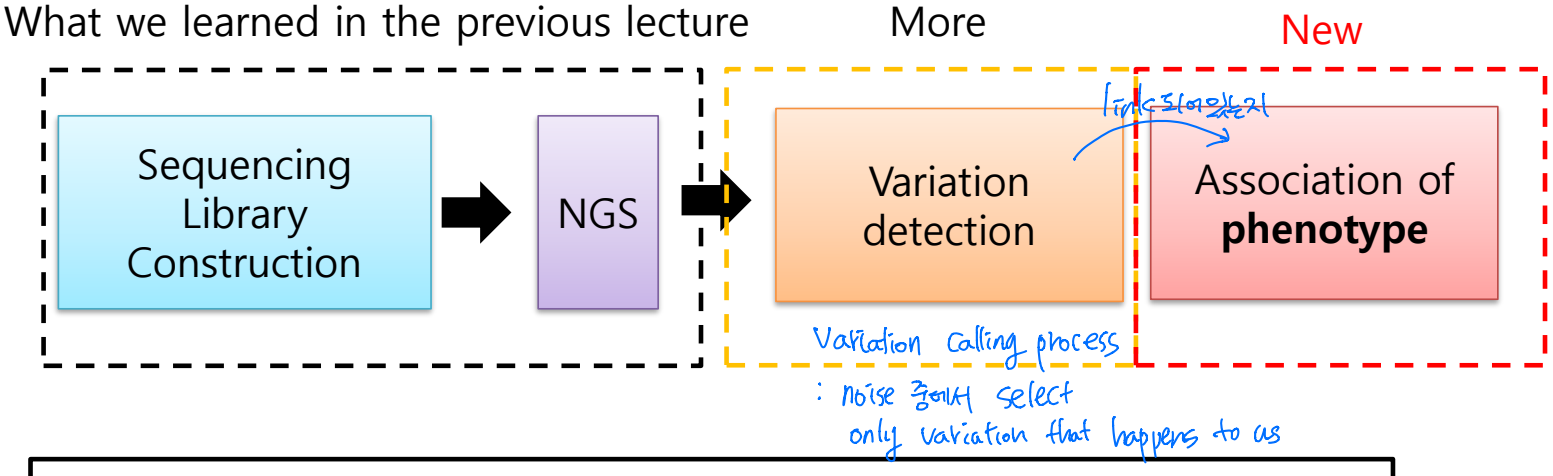
4. Next-generation Sequencing
5. NGS & Sequence Alignment

Variation of genome sequence

1. Haplotype & SNPs
2. Whole genome sequencing Variation 연구를 위한 NGS의 적용

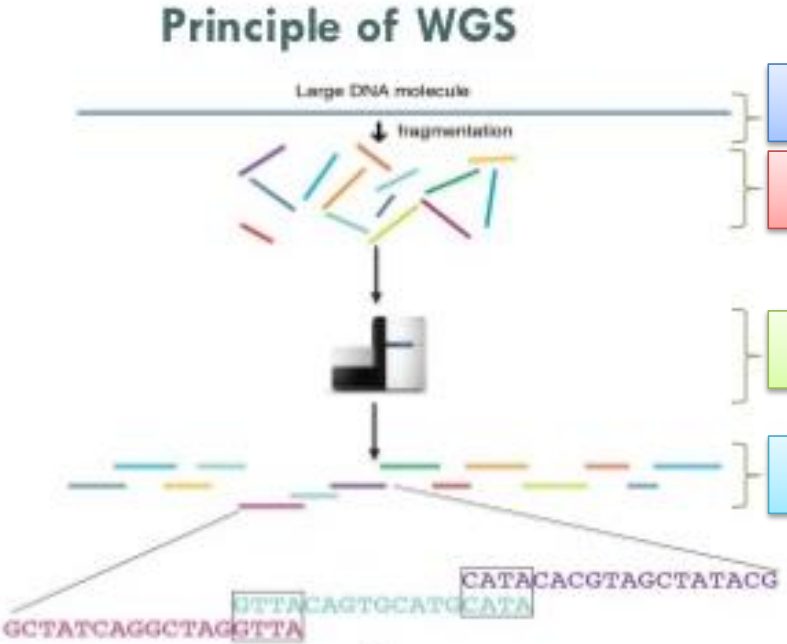
Analysis of genomic variation

Variation of genome sequence



1. Whole genome sequencing (WGS by NGS)
2. Variation Analysis for NGS data
3. Genome-wide association study (GWAS)

Whole Genome Sequencing (NGS) for variant identification



DNA fragmentation

Library preparation

NGS → read length is very short
ex) illumina: 200bp 정도

Sequenced reads

Require high coverage for variant identification

Mapping reads → to human genom data base

Downstream Analysis

Sequence analysis

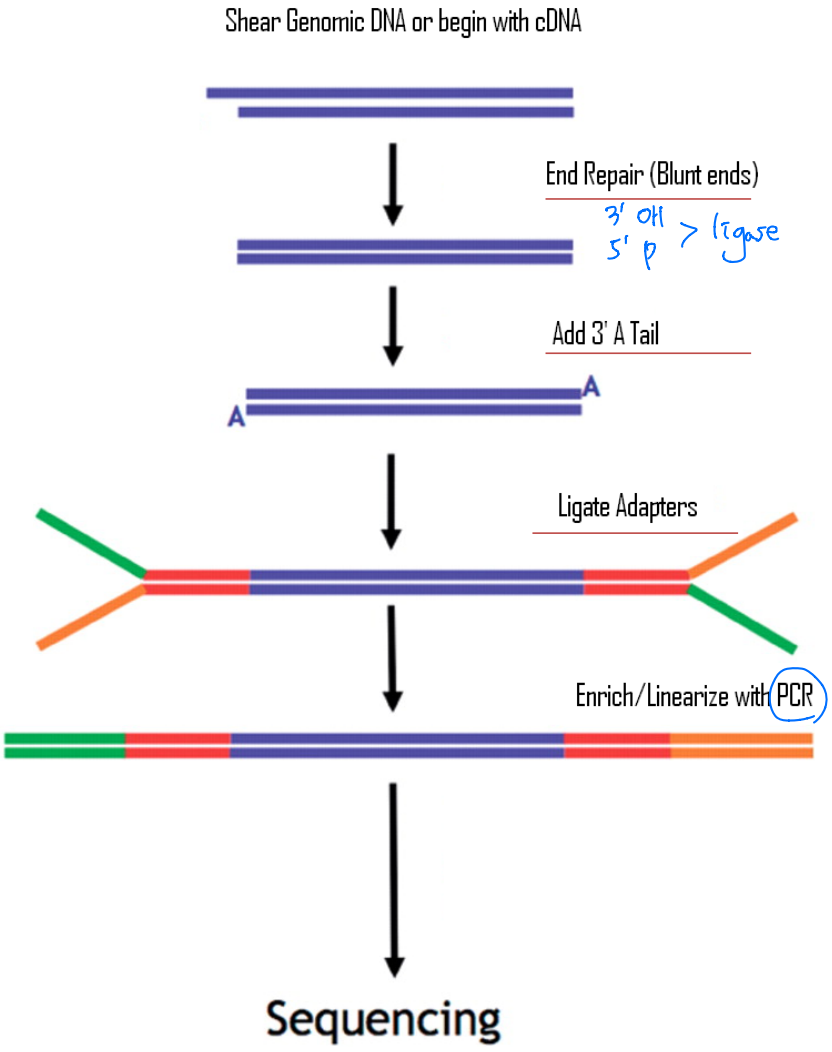
Sequencer - Illumina HiSeq 1500
Technique - Paired-End sequencing
Coverage - 100x
Read Length – 100-250 basepairs

Advantages	Disadvantages
-Point mutations -CNVs -Small indels -Rearrangements -Somatic mutations in non-coding regions (promoters, enhancers, and non-coding RNAs)	-Point mutations and indels: >30-fold haploid coverage variation의 sample에서 exact 한 값은 minimum depth -Rearrangements: >10-fold physical coverage accuracy ↓ → need high coverage

Detect every variation … unbiases

Error > high coverage → increase confidence

NGS Library preparation



1. DNA fragmentation / Target Selection
   PCR amplification for specific region of interest
2. Adapter ligation
3. Size selection
4. Library quantification (QC)

Sequencing data from NGS : FASTQ format & Q score

library
↓

NGS 기계

FASTQ

```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@:::??@@::FFAAAAACCAA:::BB@@?A?
```

Label — Sequence — Q scores (as ASCII chars)

base calling

Q score of each base calling

25 ↓ base = T

Q = : = 25

PHRED Q score :

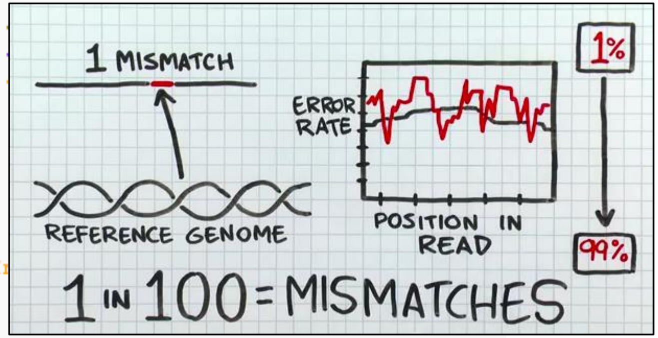
$Q = -10\log_{10}(P)$

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                             104                   126
0........................26...31.......40
                     -5....0........9.............................
                           0........9.............................
                             3.....9.............................
0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control I
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

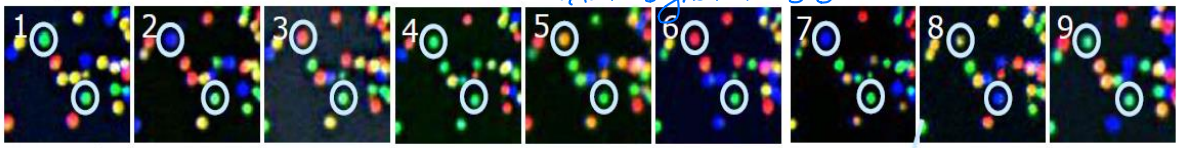
(25 circled at ":")

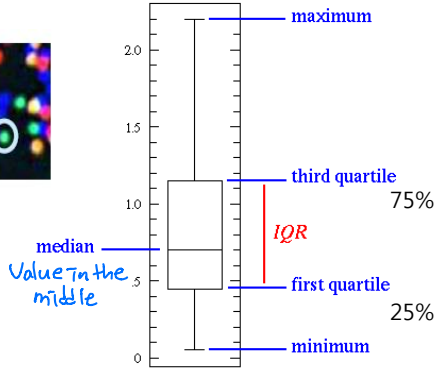
* SNP를 제외하고는 Human sequence가 같아야함 (99%는 constant) ⇒ Human genome에 mapping 했을 때 다르다면 sequence error 일 것이라고 생각 ⇒ low Qscore



Sequence Quality Control

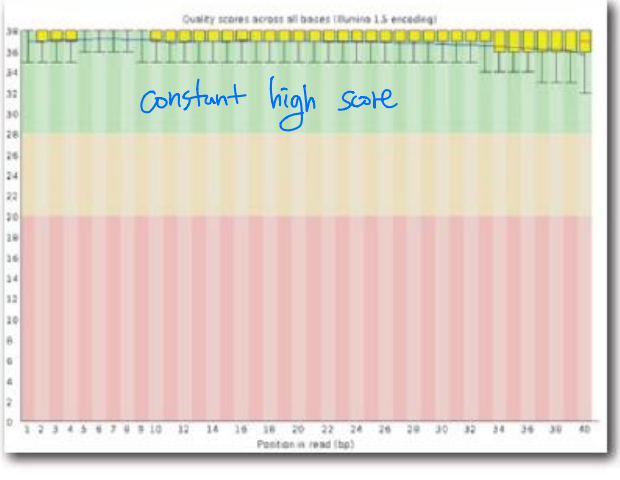
recording same position → quenching이 제대로 되지 않으면 남은 signal이 뒤의 signal에 영향



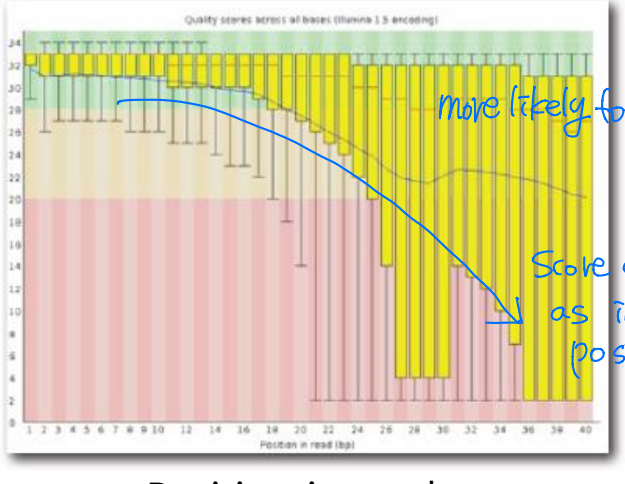


per base sequence quality

good

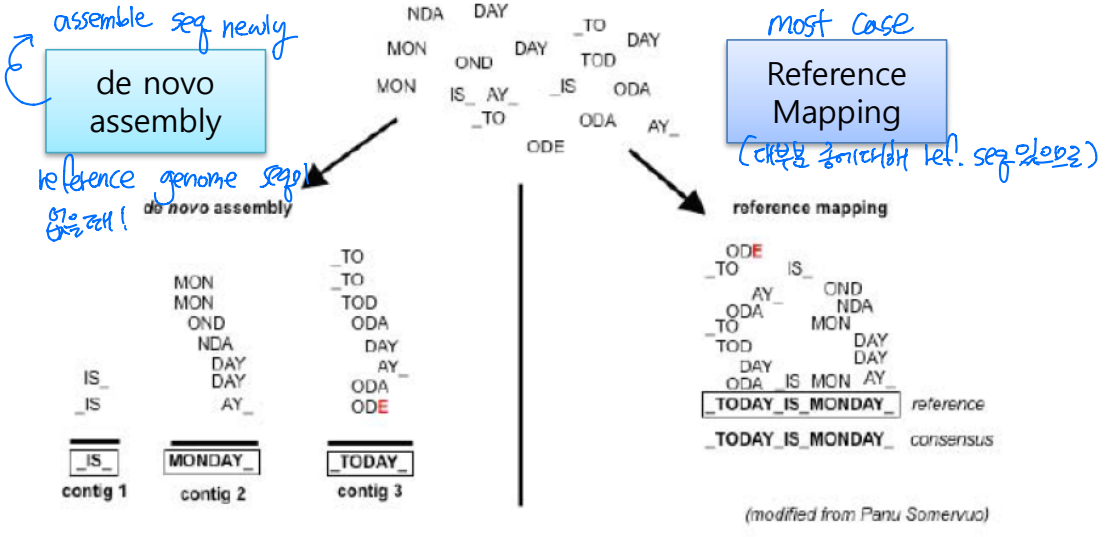


bad

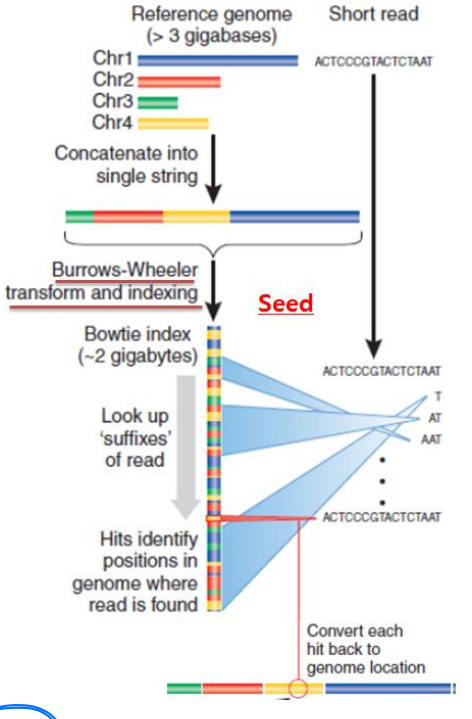


Q score tends to decrease depending on the increment of position

NGS Sequence mapping



Mapping billions of short reads on genome

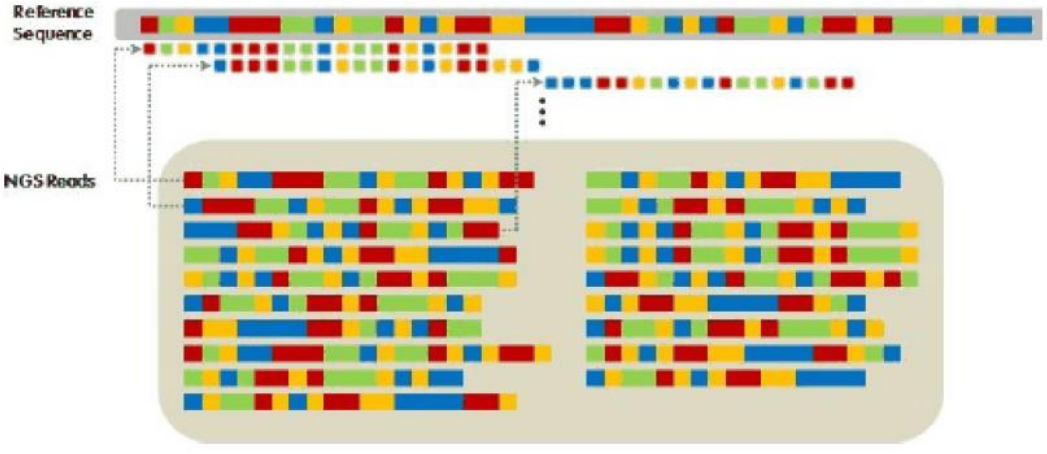


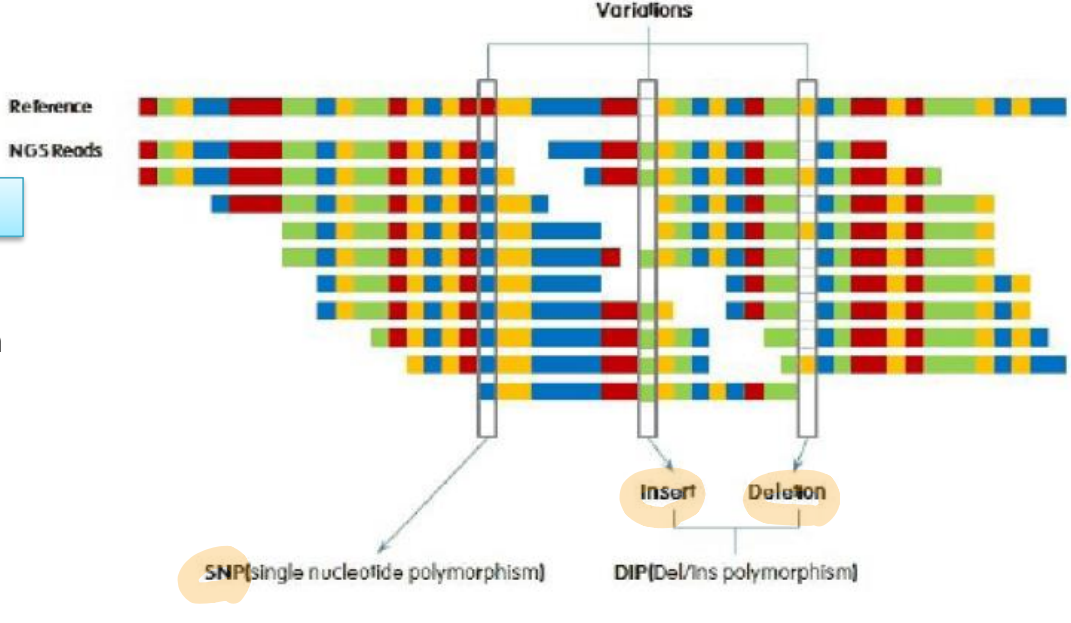
NGS WGS reads mapping (BWT based method)

- Transform genome sequences into data structure (condensed, indexed, fast search)
- Find billions of short reads sequence there (near perfect match)
- Location of matched reads (Genomic coordination, Alignment)

BWT (Burrow-Wheeler transform) based mapping algorithm

Sequence Alignment (Mapping)





Location

Genomic coordination

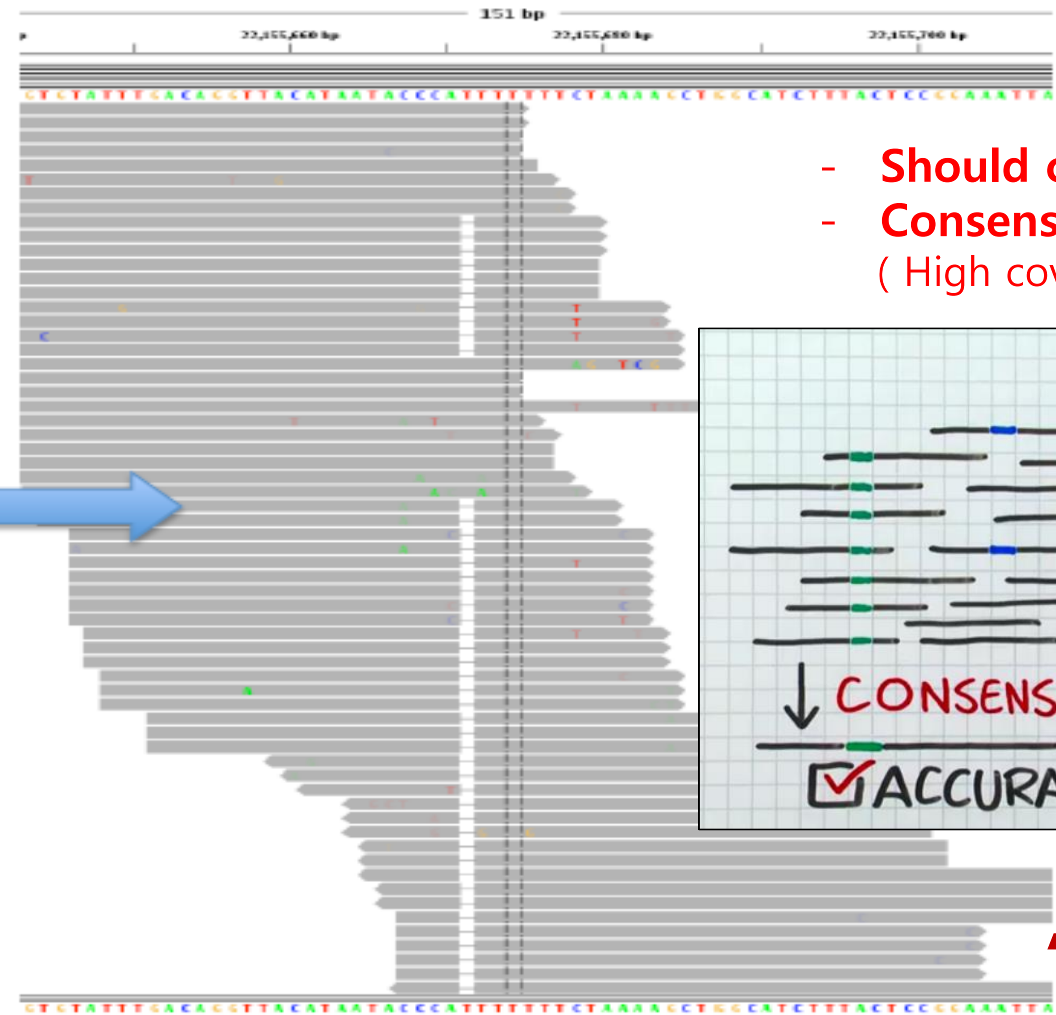
Alignment

Match
Mismatch
Deletion
Insertion

[BAM format]

Variation

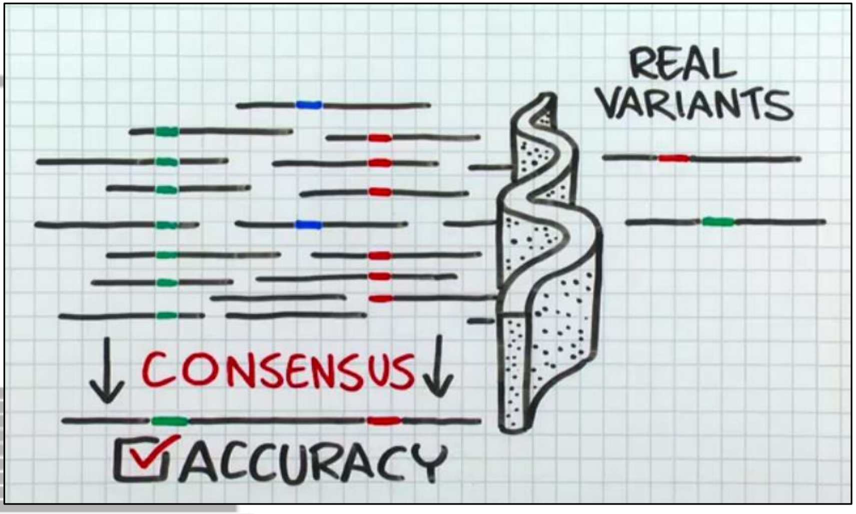
Variation in NGS reads : Real or not ?



Variation이 seq error인지 진짜인지

- **Should consider Q-score**
- **Consensus** ⇒ reproducably see this event

(High coverage > Statistics)



↝ statistically test

"Variant calling"

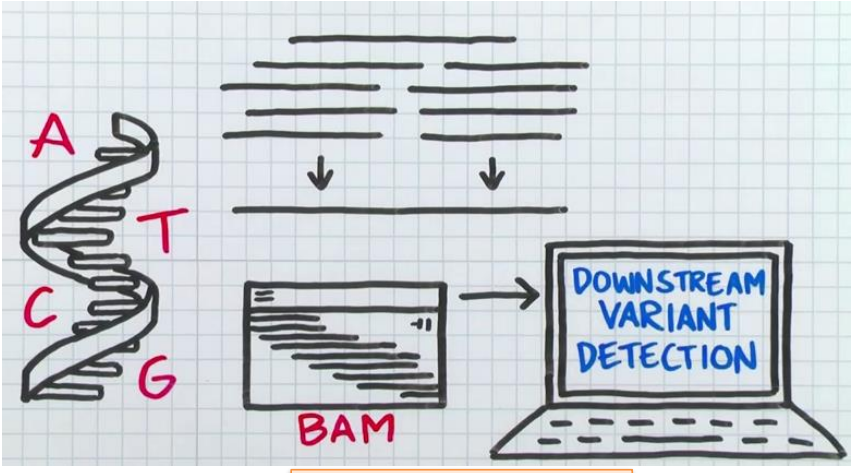
seq error가 아니라 sample에서 온거다

[VCF format]

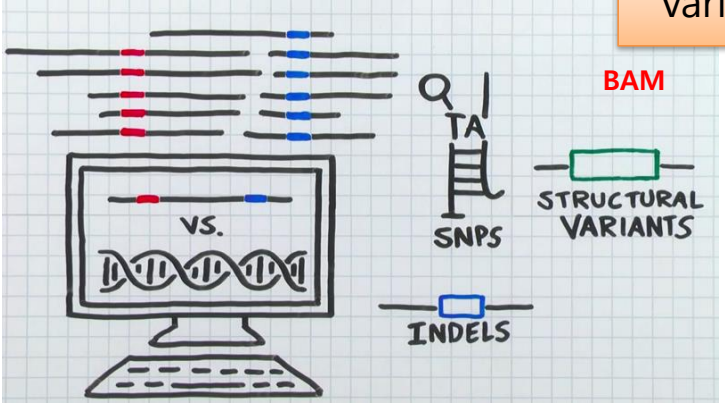
https://www.youtube.com/watch?v=0-5XtQOyVOA

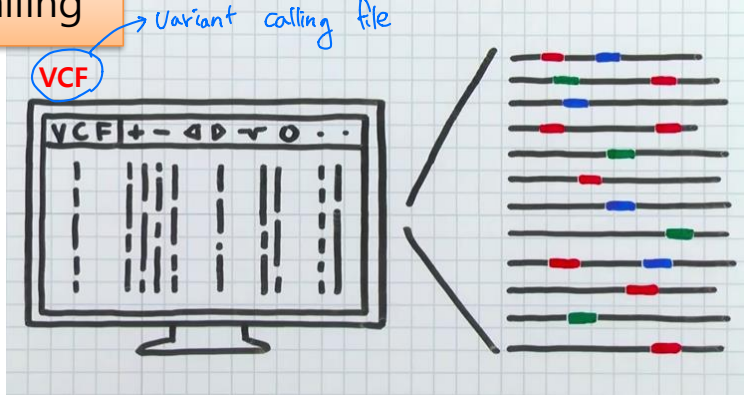
NGS WGS analysis for variation detection

Library preparation → NGS → FASTQ → Mapping → BAM



Variant calling





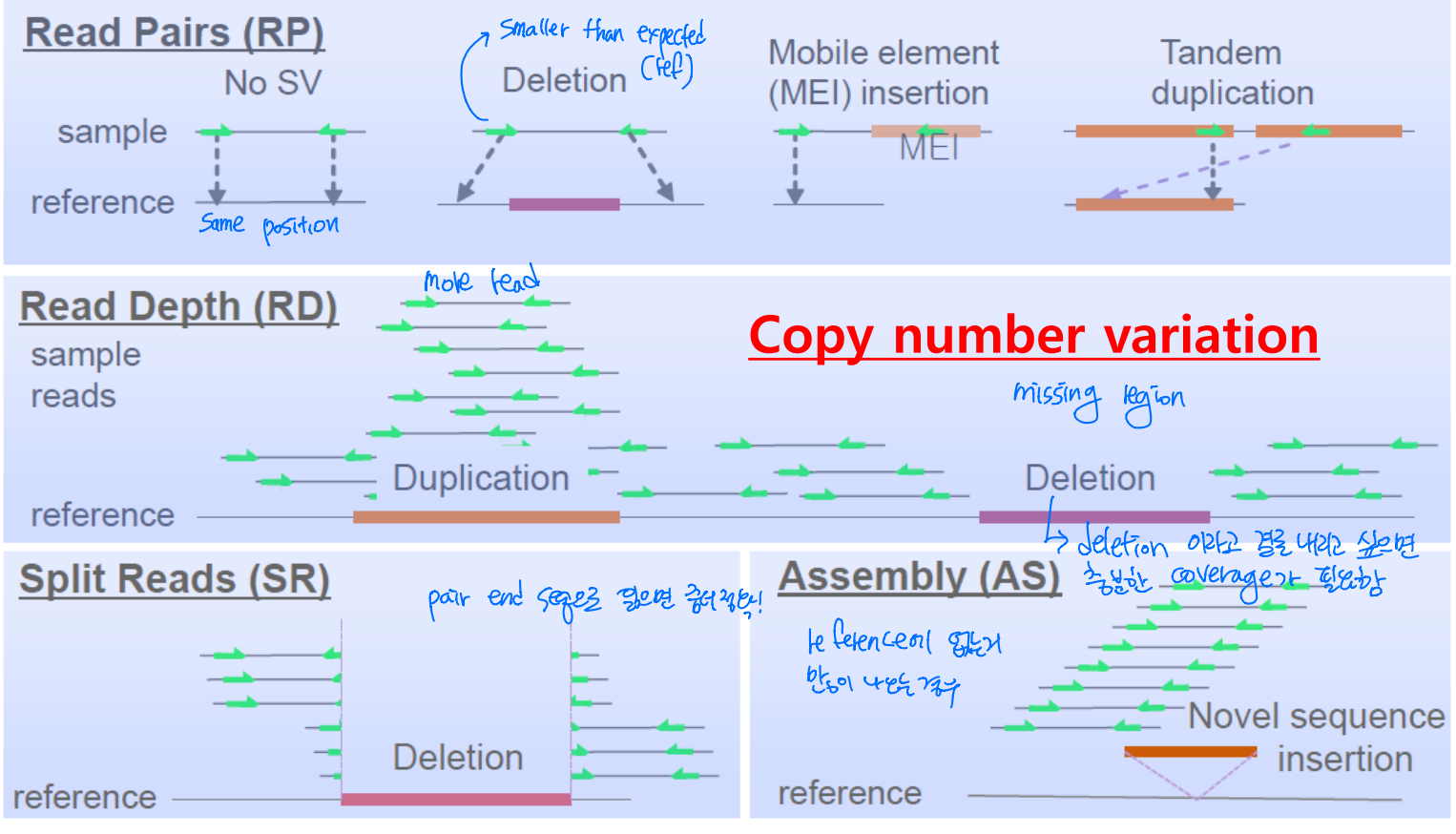
https://www.youtube.com/watch?v=6JfXU5iOAD0&list=PLGlvFEwL2wDFVLBA1egJFwHyetpVDPExK&index=11

SV Discovery with diverse resources

Structure variation

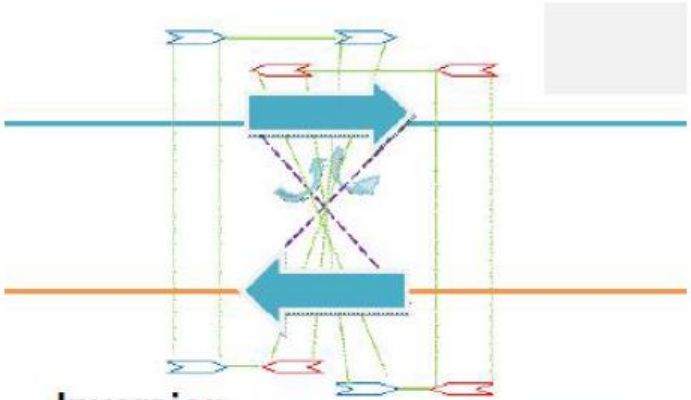
Discovery of Structural Variation

- Challenging in mapping
- Difficult for interpretation



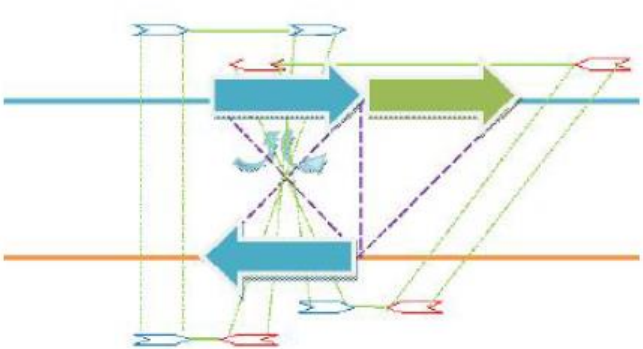
SV Discovery with diverse resources

Structure Variation 의 종류



Inversion

direction 의

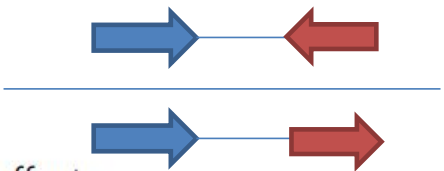


Inversion + Deletion

direction & size

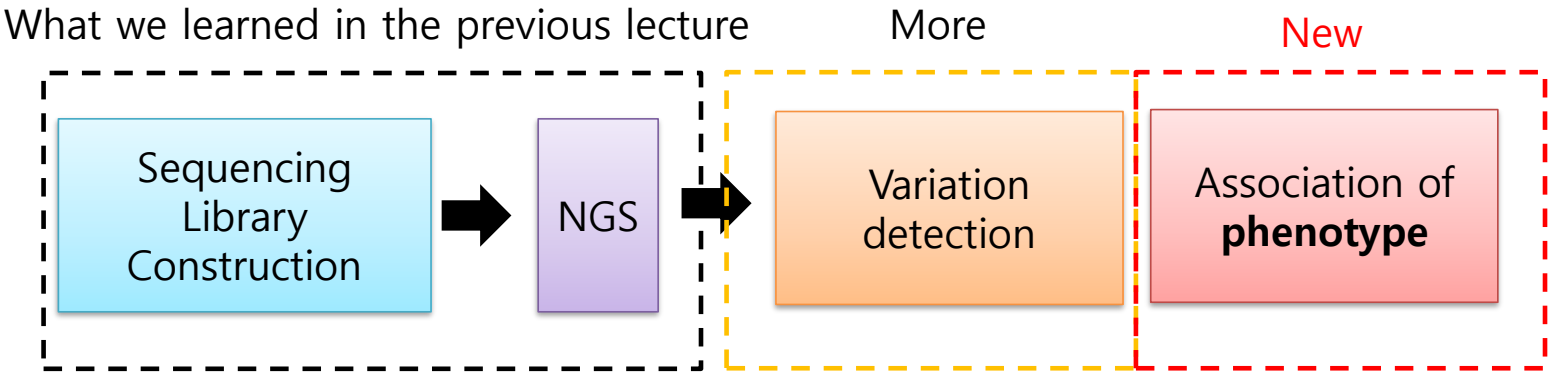
Faulty mapping of pair-end reads could be interpreted as "inversion" event !!

- **Inversion**
  - Does not involve a loss of genetic information
  - Simply rearranges the linear gene sequence
  - Generally considered to have no deleterious or harmful effects



Analysis of genomic variation

Variation of genome sequence



1. **Whole genome sequencing (WGS by NGS)**
2. **Variation Analysis for NGS data**
3. **Genome-wide association study (GWAS)**

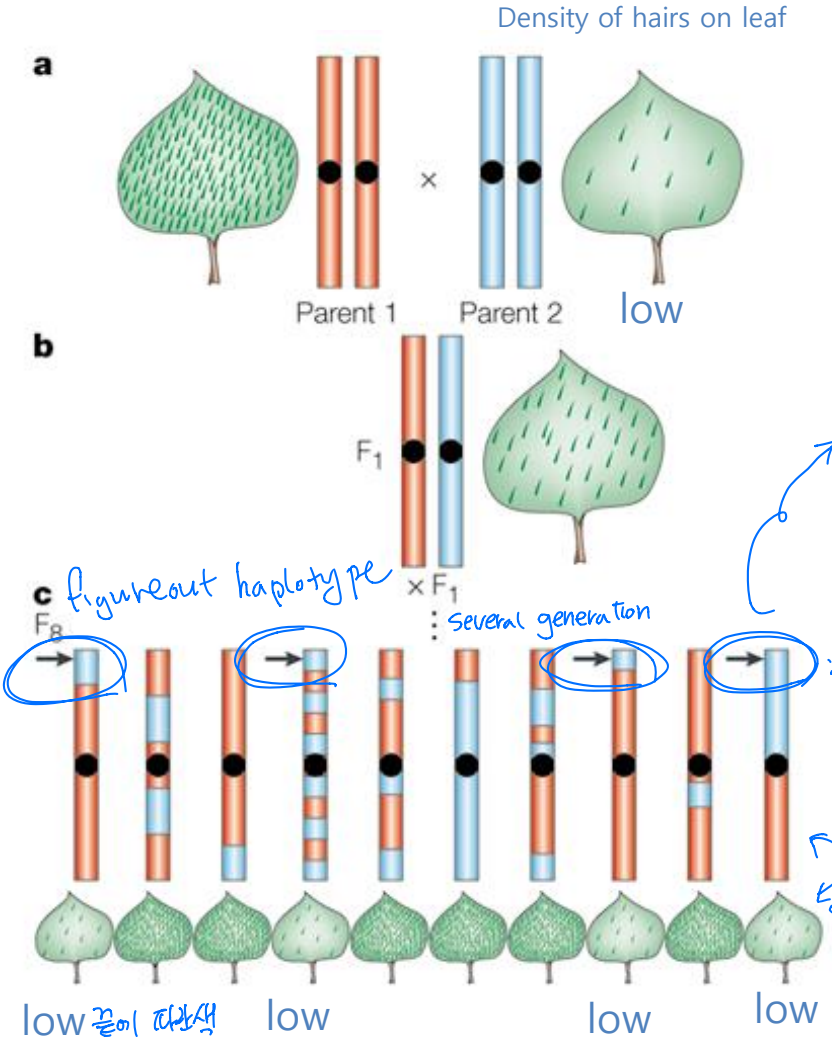
Genome-wide association study (GWAS)

Sung Wook Chi
Division of Life Sciences, Korea University

♂ phenotypeol 영향을 주는 loci
vary in degree

Variation and phenotype: Quantitative trait loci (QTL)

Variation (region) vs. Phenotype (trait)



Quantitative traits

: phenotypes (characteristics) that vary in degree and can be attributed to polygenic effects (i.e., product of two or more genes, and their environment)

Quantitative trait loci (QTLs)

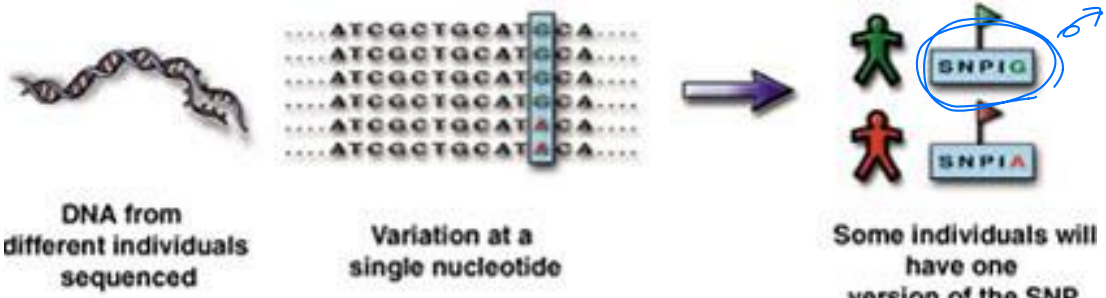
: stretches of DNA linked to, or containing, the genes or variations that underlie a quantitative trait.

How to identify QTLs ?

QTL linkage mapping

- SNPs , Haplotype
- association study with phenotypes

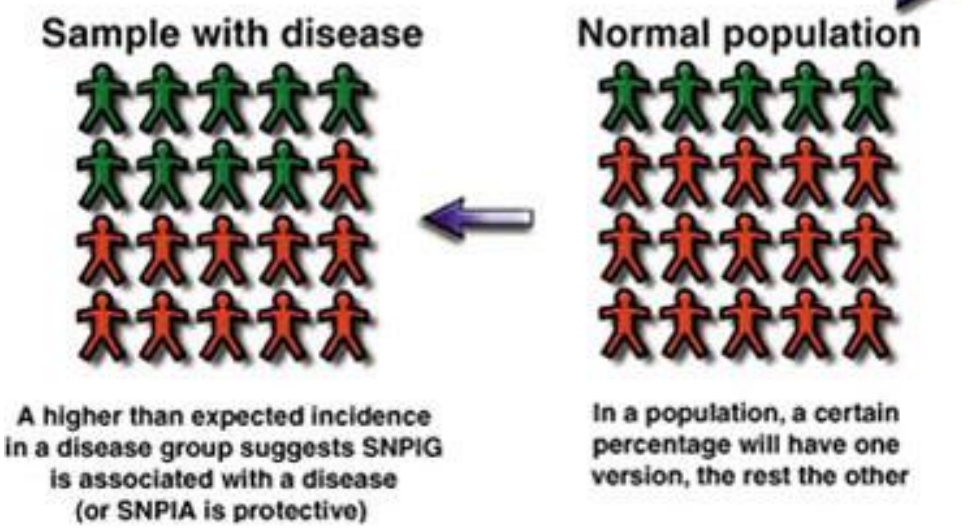
Using SNPs to Track Predisposition to Disease and other Genetic Traits



이 SNP가 disease와 관련있다면 patient > normal 일것

- Quantitative traits
- Diseases

Hypothesis-free approach



patient에서 더 많음
⇒ SNP가 disease phenotype과 관련있을것

polygenic event
→ nothing is tightly linked
⇒ 낮은 association이라도 중요!

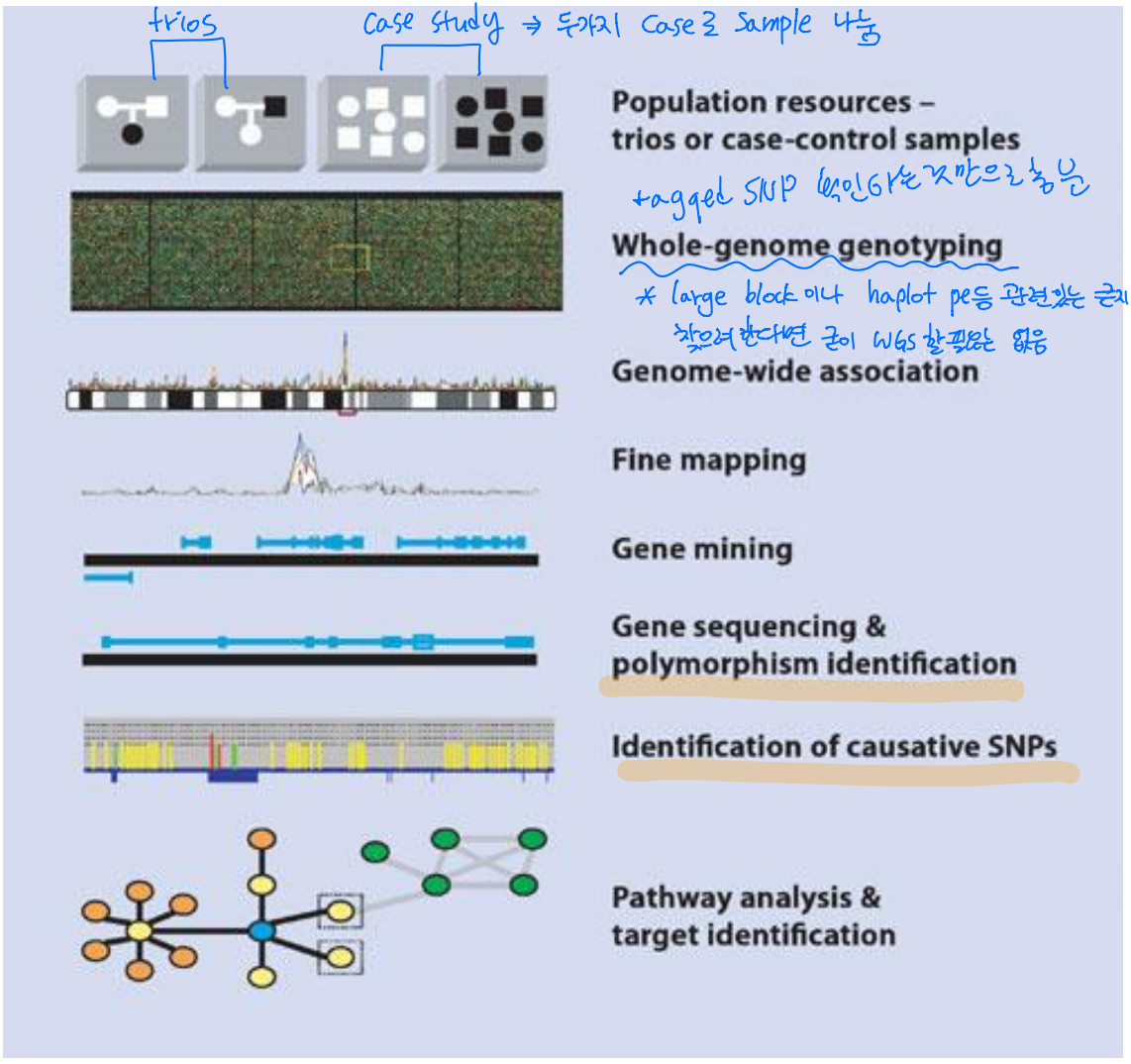
•Assumption

- **Bi-allelic SNPs**
- **Common ancestors**
- **Linkage disequilibrium and haplotypes**
- **Common disease-common variant**

:The hypothesis that genetic influences on susceptibility to common diseases are attributable to a limited number of variants present in more than 1% to 5% of the population.

이정도의 낮은 association 값
→ 이정도가 significant한지는 두고봐야

genome-wide association study (GWAS)

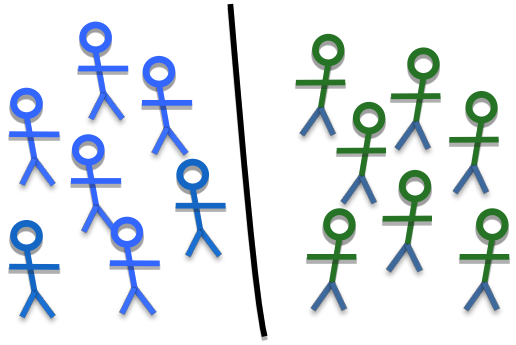


GWAS:
An examination of genetic variation across a given genome whether it associated with phenotypes (quantitative traits, diseases)

GWAS

Control Population

Disease Population



Haplotyping
(tagged SNPs)

SNP chip

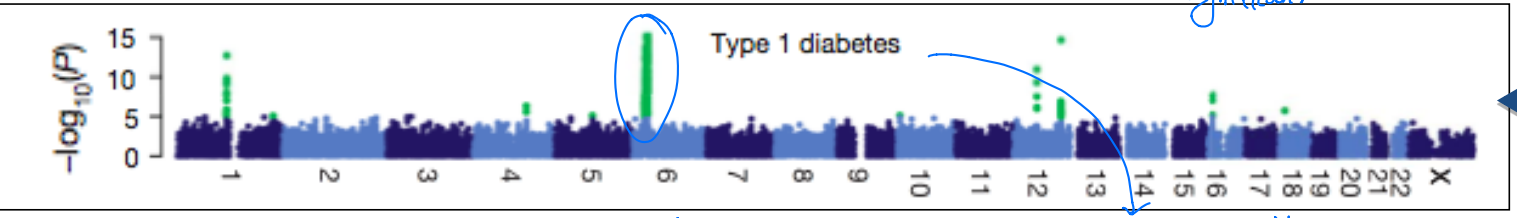
NGS sequencing

Whole genome sequencing (WGS)

D' 등이 매우 낮아 아무래도 어렵기 때문에 p-value를 사용

Linkage disequilibrium analysis (SNP & Phenotype) > Statistics (p-value) > low p-value > significant

= significant



→ statistical test를 통해 chr 6과 연관있다고 찾음

SNP variation & phenotypic change

- **Classification of SNPs**

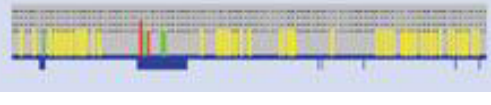
: transition, transversion, single-base indels

1. Noncoding SNPs, 5'UTR, 3'UTR, intron, intergenic regions;

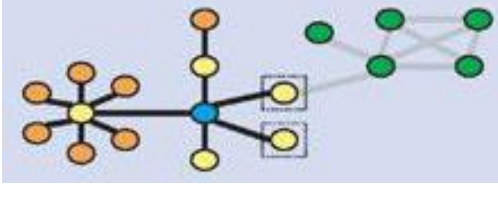
2. Coding SNPs,

- nonsynonymous or missense or replacement polymorphism
- Synonymous or sense polymorphism → a.a가 달라졌지만 same function
- Regulatory polymorphism : Synonymous and noncoding polymorphism

3. haplotype, a distinct combination of single nucleotide types on a single chromosome at a locus. used to narrow down region



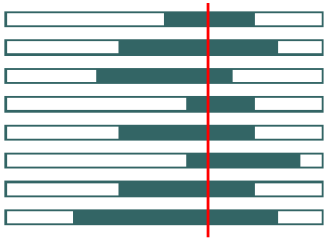
Identification of causative SNPs

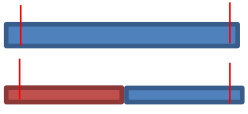


Pathway analysis & target identification

Need to explain the mechanism how SNP affects the phenotype !!
(gene expression, regulation)

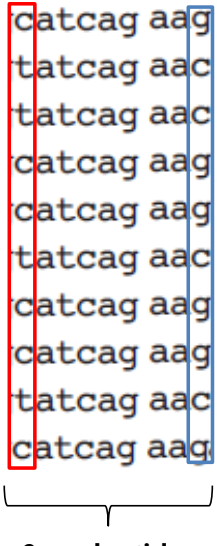
Haplotyping : linkage disequilibrium





6 x c 4 x t

n=10



9 nucleotides

1. Segregation sites ?

SNP1

SNP2

5 x g 5 x c

2. Minor allele frequencies?

SNP1 = 0.4

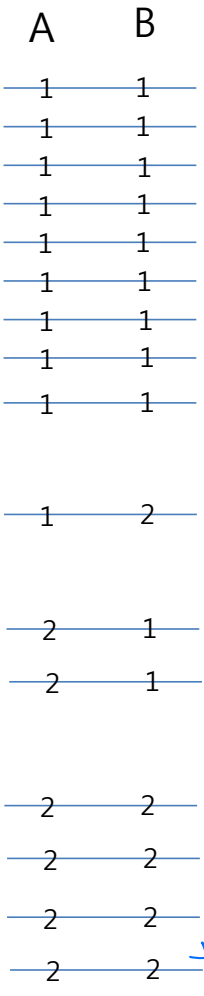
SNP2 = 0.5

3.Expected average heterozygosity (H)?

(0.4 x1 + 0.5 x1 + 0 x7) / 9 = 0.1

that your sequence results in this block can discriminate haplotypes from your ancestor

Haplotyping : linkage disequilibrium



	B_1	B_2	Total
A_1	$p_{11} = p_1q_1 + D$	$p_{12} = p_1q_2 - D$	p_1
A_2	$p_{21} = p_2q_1 - D$	$p_{22} = p_2q_2 + D$	p_2
Total	q_1	q_2	1

Expected

p1=10/16 = 0.625

q1=11/16 = 0.629

p1q1=0.39

observed

	B_1	B_2	Total
A_1	9/16 $= p_1q_1 + D$	1/16 $= p_1q_2 - D$	p_1
A_2	2/16 $= p_2q_1 - D$	4/16 $= p_2q_2 + D$	p_2
Total	q_1	q_2	1

D=0.133

Dmax
=min (0.625 x 0.371, 0.375 x 0.629)
= 0.23

D'=0.133 / 0.23 = 0.58

↳ linkage 되어 있는지 알아내려면 random 에서 나오는 수보다 커야 함

Whether it is significant ?

Statistical significance

Null hypothesis (H_0)

not associated

Independent 인가? (내가 세운 가정이 들어맞는 경우)

Random process
(Permutation)

If we know the null distribution, we just simply used it
(t-test, chi-square test...)

Alternative hypothesis (H_1)

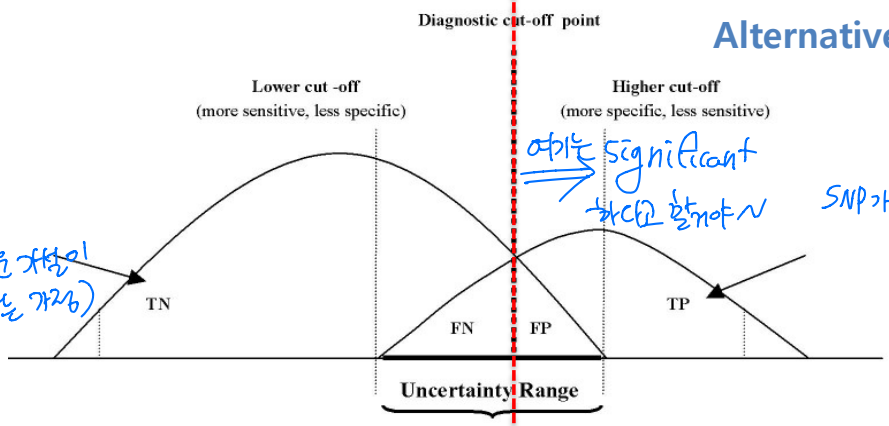
associated

SNP가 cosegregated 되는가?

관찰을 통해 얻은 값이 관찰값을 초월할 확률

actual no.

가정을 의미

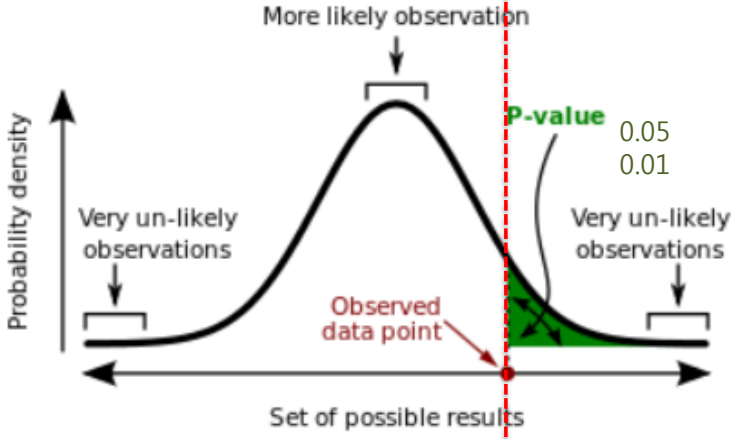


Null hypothesis (H_0) Alternative hypothesis (H_1)

위에서 얻은 D'가 상위 얻은두 있는 값인지 확인

D'가 낮다고 해도 random에서 안일어 난다면 Significant

⇒ disease랑 two SNP가 연관



Null hypothesis (H_0) **Observation**

Chi square test

카이 스퀘어

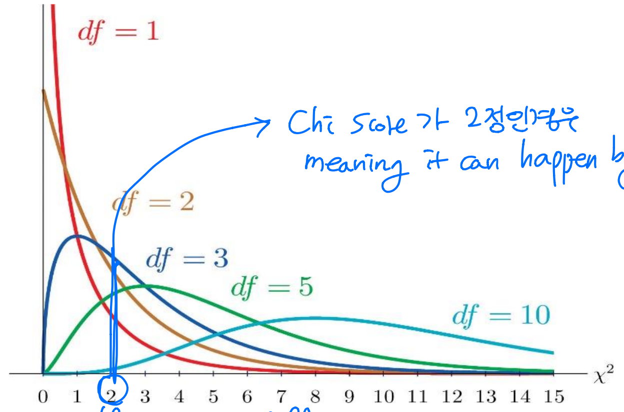
observed **Expected**

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

observed expected

χ^2 = the test statistic $\sum$ = the sum of

O = Observed frequencies E = Expected frequencies



Degrees of freedom (df) = n-1 = 4-1 = 3

catcag aag
tatcag aac
tatcag aac
catcag aag
catcag aag
tatcag aac
catcag aag
catcag aag
tatcag aac
catcag aag

observed

	c	t	total
g	Pcg	Ptg	Pg
c	Pcc	Ptc	Pc
	Pc	Pt	1

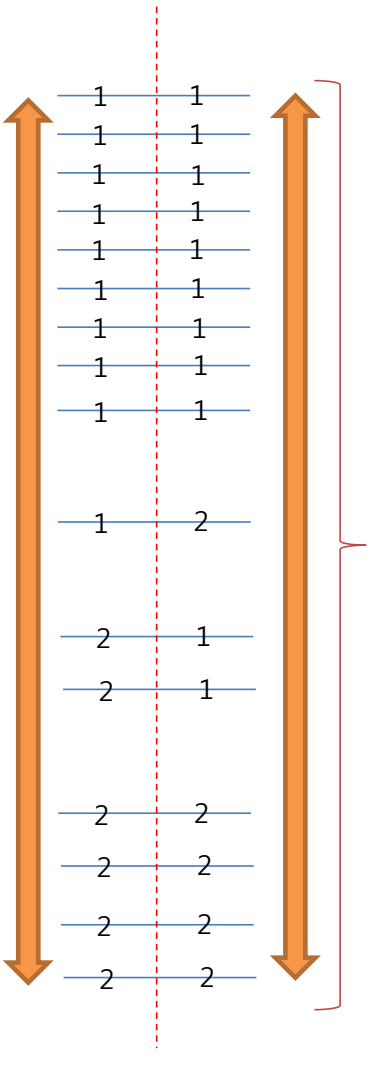
Expected

	c	t	total
g	Pc X Pg	Pt X Pg	Pg
c	Pc X Pc	Pt X Pc	Pc
	Pc	Pt	1

Chi square test (p=0.067889)

large D' value 가 나왔더라도 Chi square test가 significant 하지 않다면 random 하게 얻을수 있는결과임

Fisher's exact test



Shuffling
(Enumerate)

observed

	c	t	total
g	Pcg	Ptg	Pg
c	Pcc	Ptc	Pc
	Pc	Pt	1

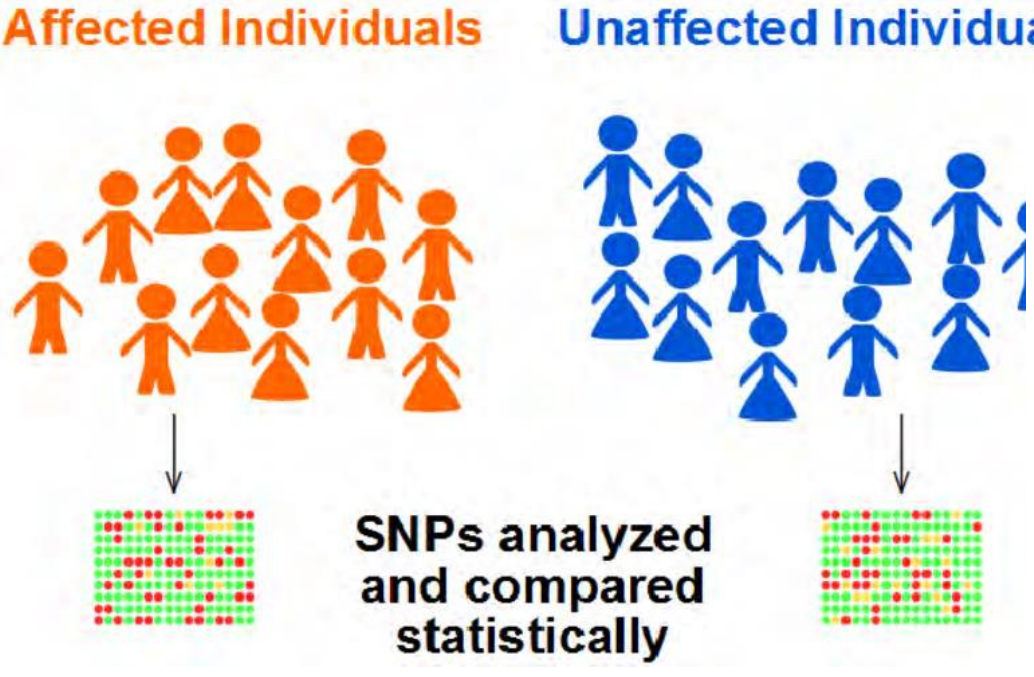
	1	2
1	9	1
2	2	4

Hypergeometric distribution

Table B

n_{11}	n_{12}	n_{21}	n_{22}	D	Probability	Cumulative probability
10	0	1	5	0.195	0.001	0.001
9	1	2	4	0.133	0.034	0.035
8	2	3	3	0.070	0.206	0.241
7	3	4	2	0.008	0.412	0.653
6	4	5	1	−0.055	0.288	0.941
5	5	6	0	−0.117	0.058	1.000

GWAS: Genome-wide association study



sample	seq	type	grade
1	actgtacttc	normal	0
2	actgtacttc	normal	0
3	actgtacttc	normal	0
4	actgtccttc	normal	0
5	actgtccttc	normal	0
6	actgtccttc	tumor	2
7	actgtccttc	tumor	3
8	actctactta	tumor	2
9	actgtccttc	tumor	3
10	actctcctta	tumor	4

Distribution of SNPs : Population genetics

- SNPs – phenotypic variation (Quantitative genetics)
- QTL (quantitative trait loci) = polygenic phenotypic variation

Linked?

SNPs ⟷ Phenotype

GWAS (practice)

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
B_2	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

observed frequency

	A	a	Total
Affected	9	1	10
Unaffected	2	4	6
Total	11	5	16

	A	a
Affected	e_{11}	e_{21}
Unaffected	e_{12}	e_{22}

expected frequency

e11= 16 x p1 x q1 = 6.875

p1=10/16 = 0.625
p2=0.374
q1=11/16=0.6875
q2=0.3125

D=9/16-0.625x0.6875= 0.1328

D_{max}=min(0.625x0.3125, 0.374x0.6875)
= 0.195313

	A	a	
Affected	6.875	3.125	10
Unaffected	4.125	1.875	6
	11	5	16

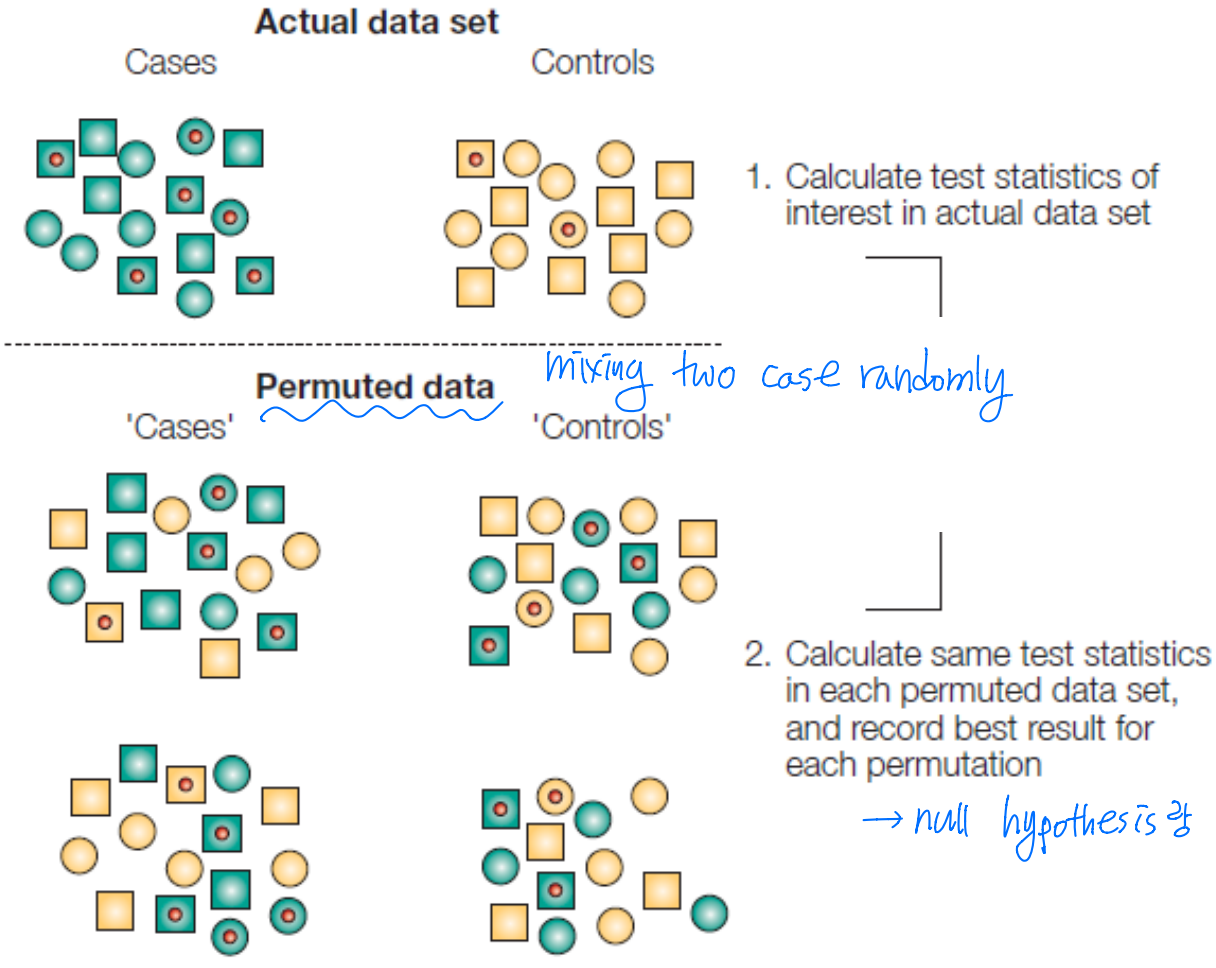
D' = 0.1328/0.195313 = 0.679934

Chi square test (p=0.017911)
Fisher's exact test

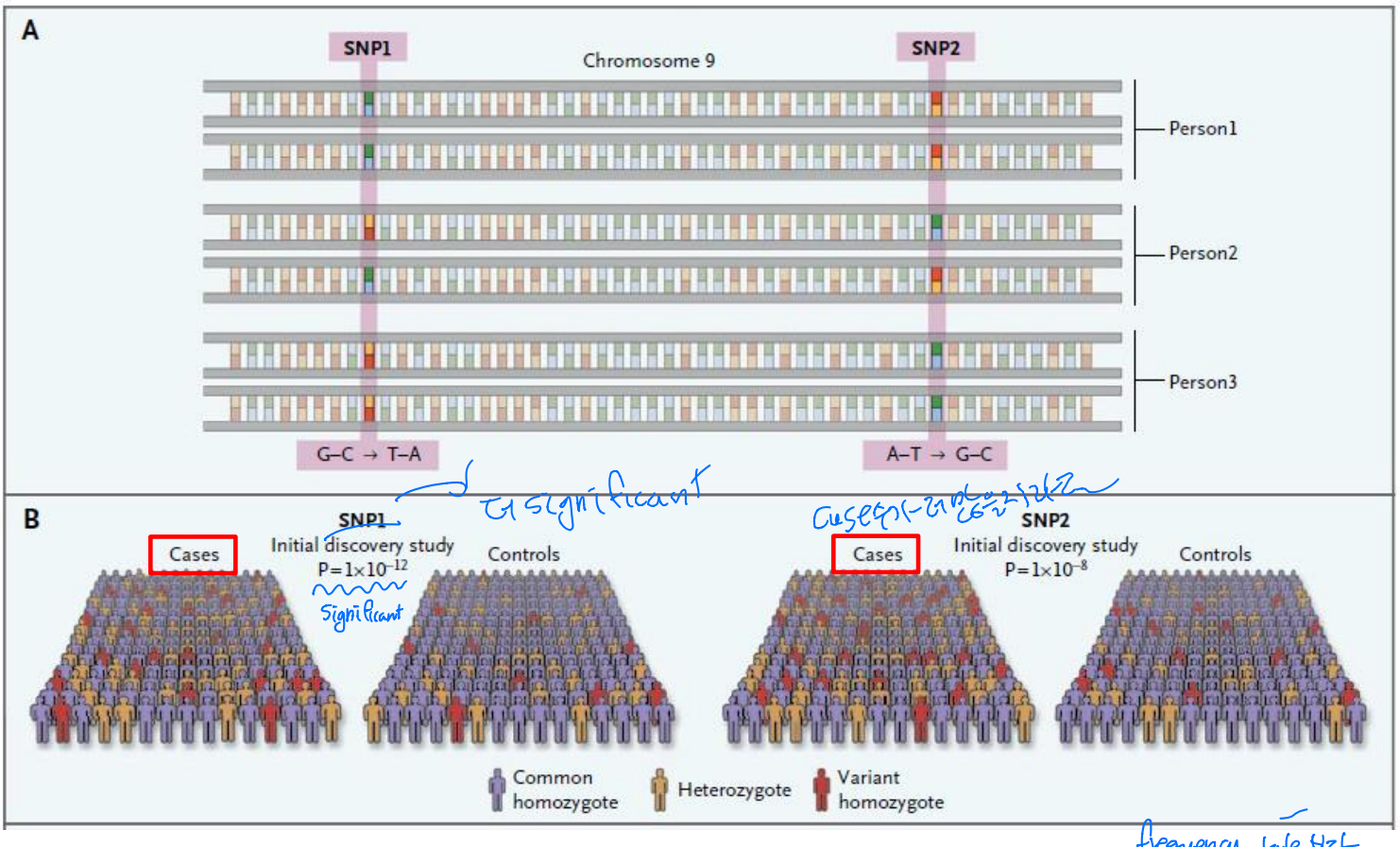
Permutation
(False-discovery rate)

Significant Linkage disequilibrium
(A is significantly associated with disease)

Permutation testing



The Genome-wide Association Study (GWAS)

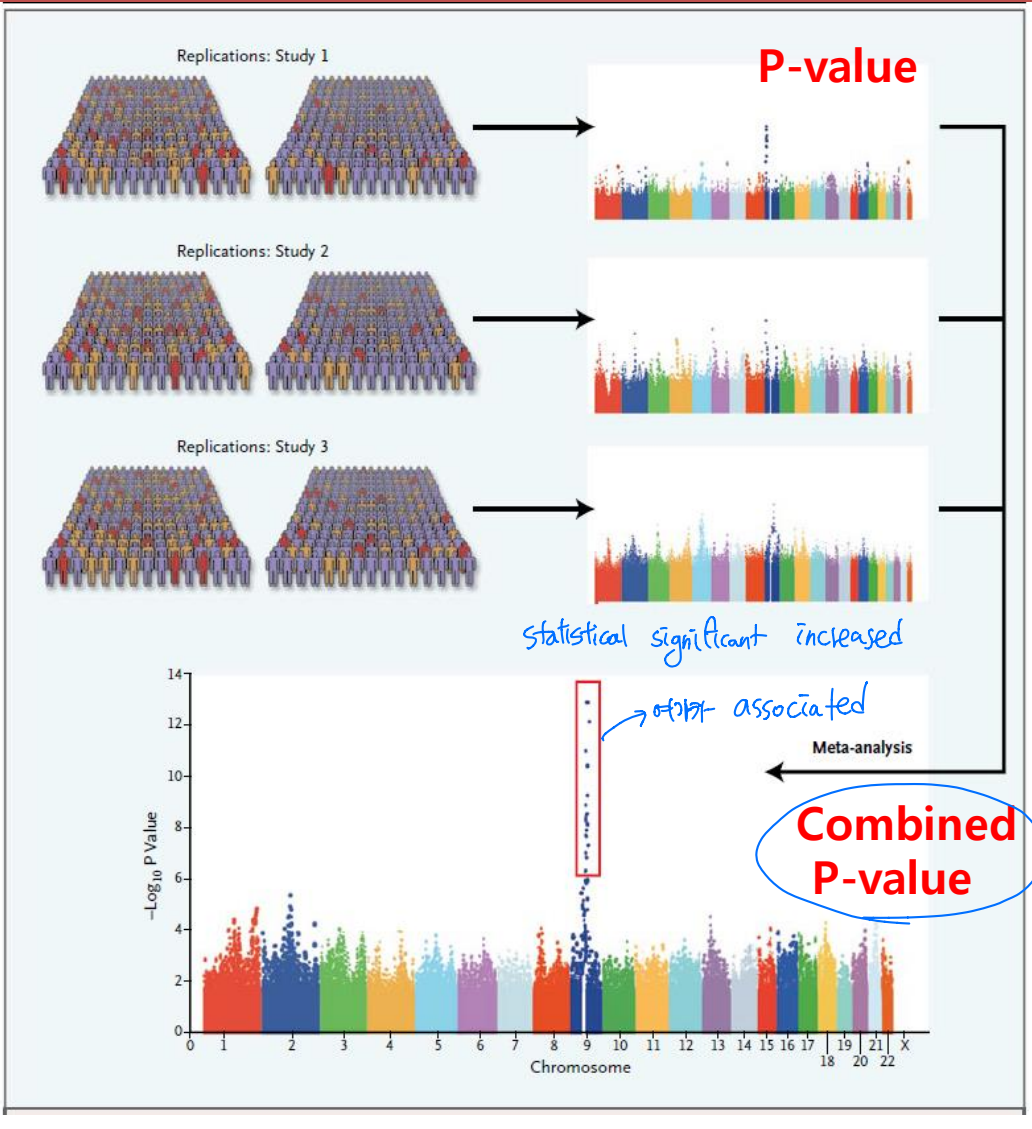


- An examination of *genetic variation* across a given genome
- Designed to identify genetic associations with observable traits –Such as blood pressure or weight, –or why some people get a disease or condition

The Genome-wide Association Study (GWAS)

Meta-analysis is required

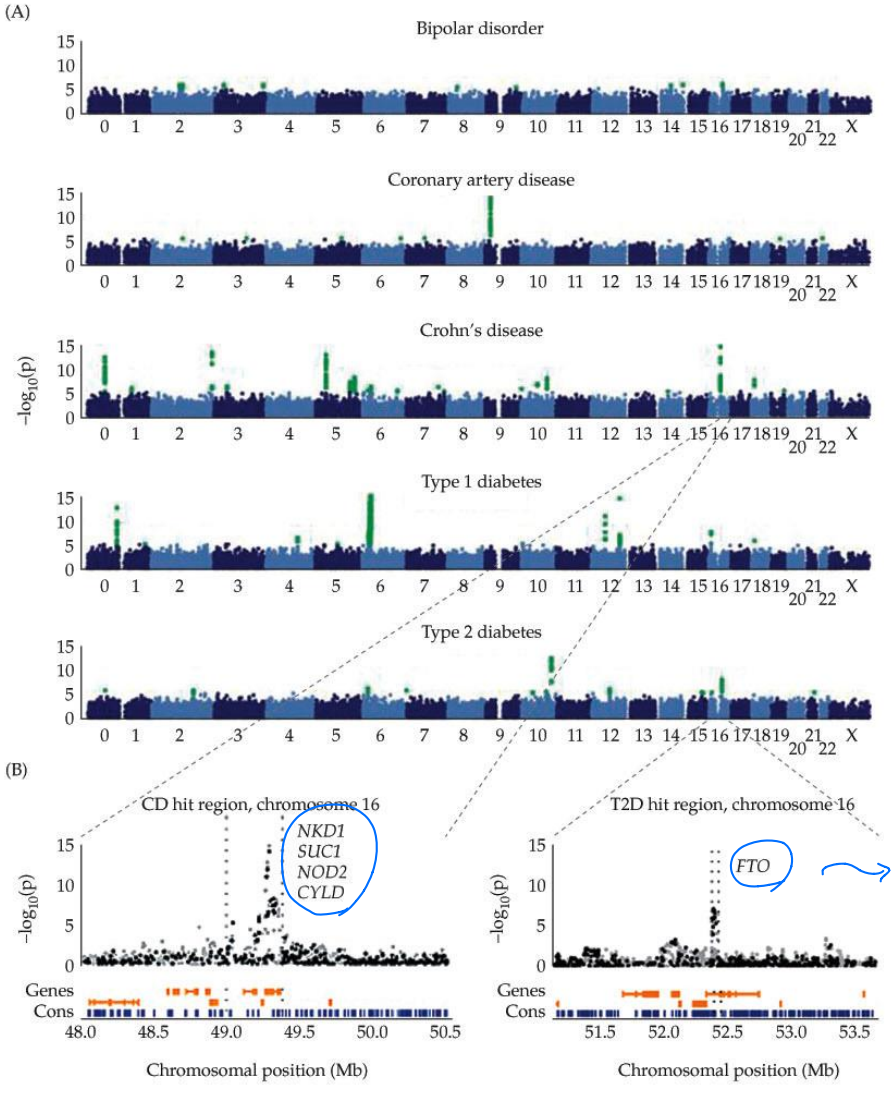
Statistical significance



Genome-wide Association Studies (example)

Y축 = $-(10\log_{10}(P))$

Statistical significance



→ p value가 높은 곳들 = associated

each dot : SNP & have own p-value

→ 각 병에 연관된 Variation