2023.03.31

GWAS
Quality Control

Seoungbeom Jin

GWAS Quality Control

Sample Quality Control

1) Remove mismatch gender information
2) Outlying missing genotype or Heterozygosity rates
3) Duplicated or Related individuals
4) Divergent ancestry

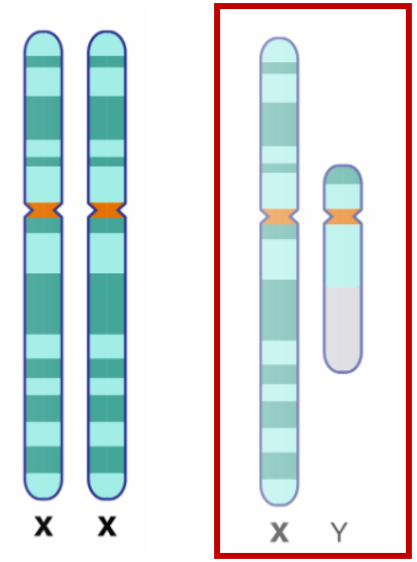
SNP Quality Control

1) High missing genotype
2) Low MAF
3) Significant deviation from HWE
4) Significantly different missing genotype rates between Cases and Controls

GWAS Quality Control

- 1. Sample Quality Control

1) Mismatch gender information



All the X chromosome SNPs is homozygous.

Male samples to have a homozygosity rate of 1.

(Inbreeding Coefficient(F) = 1)

But, homozygosity rate of male samples is not 1 in raw data set.

↓

Calculate the Inbreeding Coefficient(F)

↓

Compare gender of input data and gender of F values

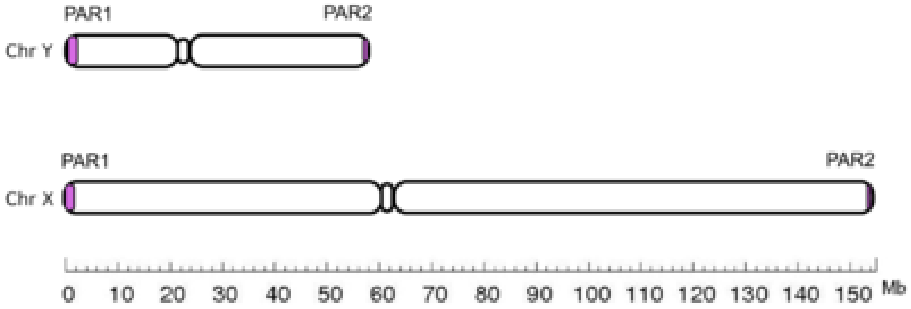
https://www.genome.gov/genetics-glossary/X-Chromosome(NCBI), 2023-02-16

GWAS Quality Control

- 1. Sample Quality Control

1) Mismatch gender information

$$F = 1 - \frac{Observed\ Homozygosity\ Rates}{Expected(Hardy\text{–}Weinberg\ Equibrium)\ Homozygosity\ Rates}$$



Pseudoautosomal region (PAR)

(유사상동염색체 영역)

X and Y chromosomes have regions similar to homologous chromosomes.

Therefore, the F value may not be 0 or 1.

F < 0.2 Female

F ≥ 0.8 Male

https://en.wikipedia.org/wiki/Pseudoautosomal_region, 2023-02-24

GWAS Quality Control

- 1. Sample Quality Control

2) Outlying missing genotype or Heterozygosity rates

Samples of low DNA quality or concentration → Below average call rates and genotype accuracy.

More than 3 – 7% missing genotypes are removed

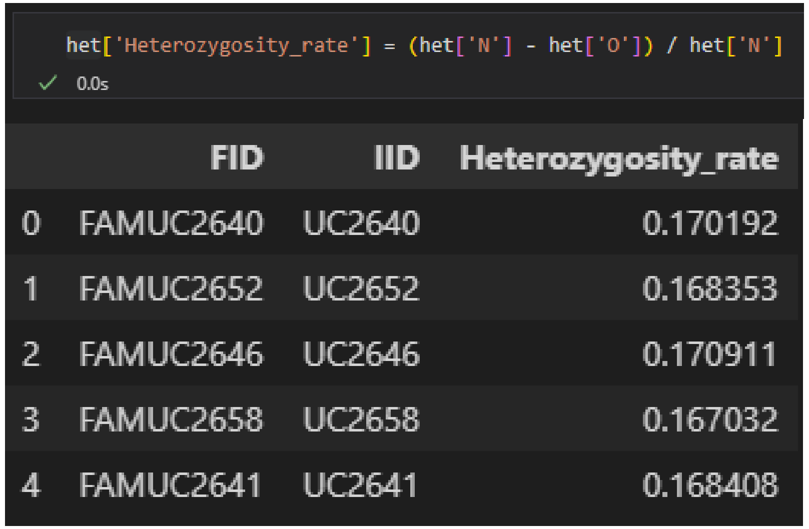
Excessive or reduced proportion of heterozygosity rates → Respectively DNA sample contamination or inbreeding.

$$\text{Heterozygosity rate} = \frac{\text{Non missing genotypes(N) - Observed homozygous genotypes(O)}}{\text{Non missing genotypes(N)}}$$

GWAS Quality Control

- 1. Sample Quality Control

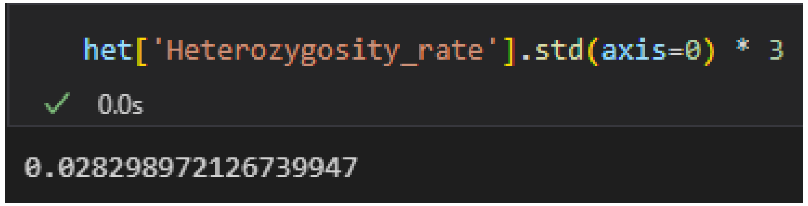
2) Outlying missing genotype or Heterozygosity rates



```
het['Heterozygosity_rate'] = (het['N'] - het['O']) / het['N']
✓ 0.0s
```

	FID	IID	Heterozygosity_rate
0	FAMUC2640	UC2640	0.170192
1	FAMUC2652	UC2652	0.168353
2	FAMUC2646	UC2646	0.170911
3	FAMUC2658	UC2658	0.167032
4	FAMUC2641	UC2641	0.168408

The each sample calculate Heterozygosity rates.

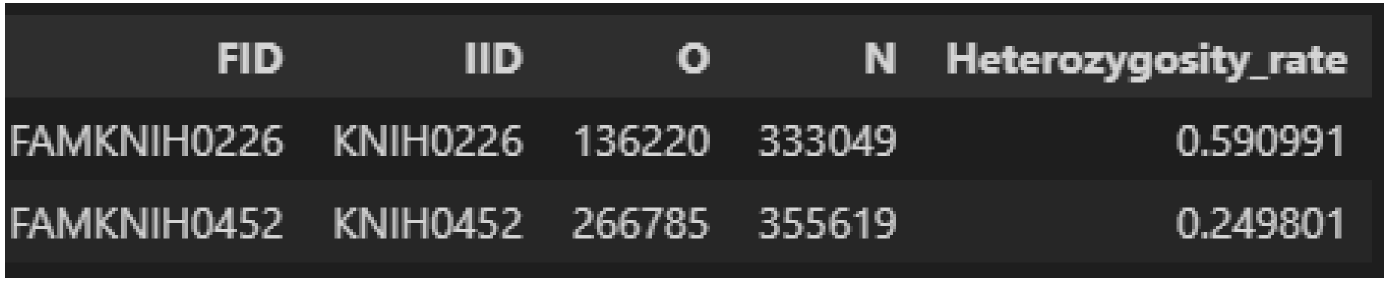


```
het['Heterozygosity_rate'].std(axis=0) * 3
✓ 0.0s

0.028298972126739947
```

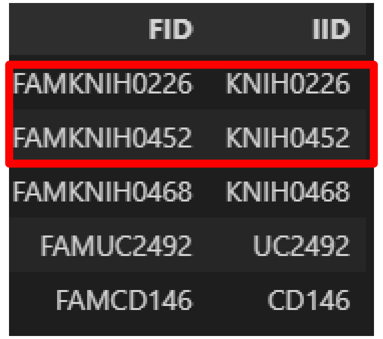
3 s.d.

Heterozygosity rates ±3 s.d. from the mean :
$0.142 < \text{Heterozygosity rate} < 0.198$



FID	IID	O	N	Heterozygosity_rate
FAMKNIH0226	KNIH0226	136220	333049	0.590991
FAMKNIH0452	KNIH0452	266785	355619	0.249801

Data not present in Heterozygosity rates ±3 s.d. from the mean = 2 Outliers Control Data



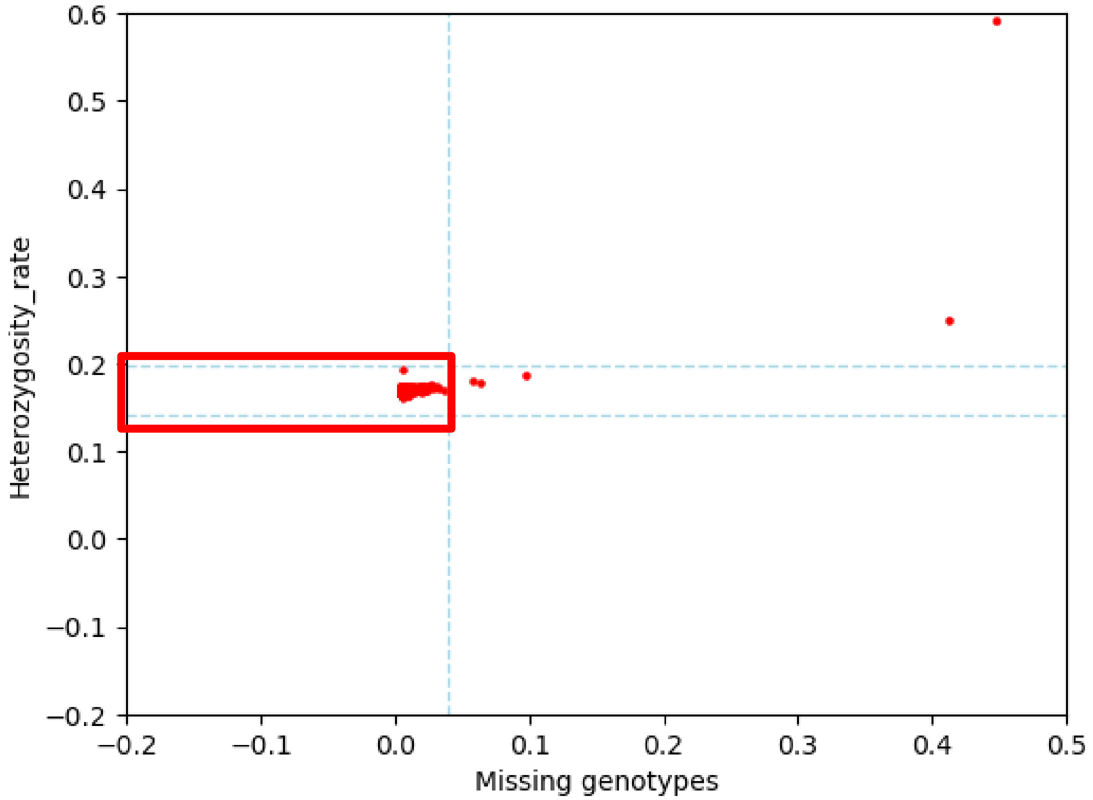
FID	IID
FAMKNIH0226	KNIH0226
FAMKNIH0452	KNIH0452
FAMKNIH0468	KNIH0468
FAMUC2492	UC2492
FAMCD146	CD146

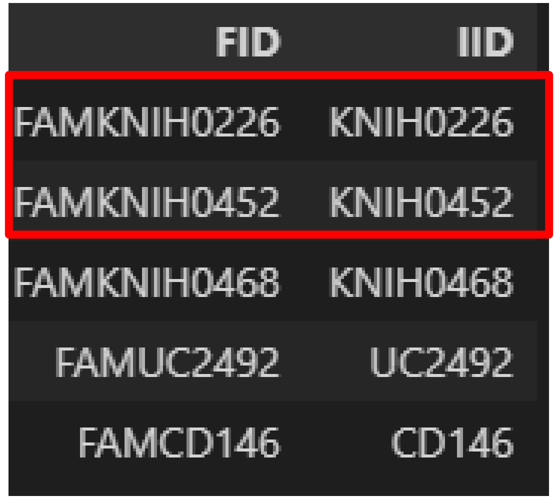
Outliers heterozygosity rates of 2 data is included in the missing genotype

GWAS Quality Control

- 1. Sample Quality Control

2) Outlying missing genotype or Heterozygosity rates





FID	IID
FAMKNIH0226	KNIH0226
FAMKNIH0452	KNIH0452
FAMKNIH0468	KNIH0468
FAMUC2492	UC2492
FAMCD146	CD146

Axis x : Missing genotype ≥ 3%

Axis y : Mean heterozygosity rates ±3 s.d. :
$0.142 <$ Heterozygosity rate < 0.198

GWAS Quality Control

- 1. Sample Quality Control

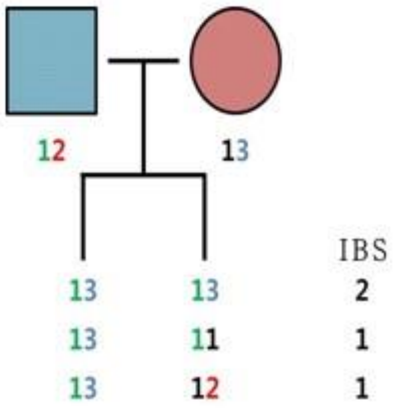
3) Duplicated or Related individuals

A basic feature of samples → All samples are unrelated

Duplicated or Related가 있으면 data에 bias가 생긴다.

- IBS(Identity By State) : 두 sample의 동일한 alleles의 frequency를 비교

- IBD(Identity By Descent) : 가계도에서 공통 조상에게 물려받은 alleles frequency를 비교



IBD(PI-HAT)
= 1(duplicate)
= 0.5(first-degree relatives)
= 0.25(for second-degree relatives)
= 0.125(for third-degree relatives)

Threashold : IBD((PI-HAT) > 0.2(0.185)

GWAS Quality Control

- 1. Sample Quality Control

3) Duplicated or Related individuals

- IBS 와 IBD가 다른 경우

모집단 내에서 비교 대상이
IBS는 높은데
IBD가 낮은 경우

→

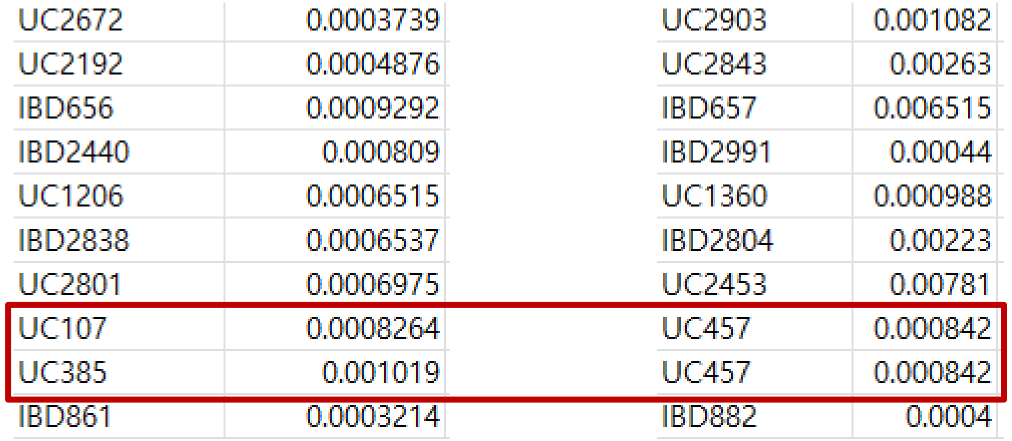
가계도가 다르지만 **우연의 일치**로
동일한 allele이 존재하는 경우

GWAS Quality Control

- 1. Sample Quality Control

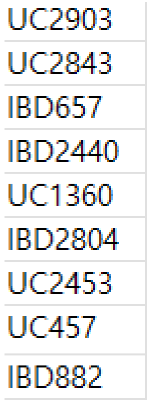
3) Duplicated or Related individuals

Removing high missing genotypes
from duplicate samples



UC2672	0.0003739	UC2903	0.001082
UC2192	0.0004876	UC2843	0.00263
IBD656	0.0009292	IBD657	0.006515
IBD2440	0.000809	IBD2991	0.00044
UC1206	0.0006515	UC1360	0.000988
IBD2838	0.0006537	IBD2804	0.00223
UC2801	0.0006975	UC2453	0.00781
UC107	0.0008264	UC457	0.000842
UC385	0.001019	UC457	0.000842
IBD861	0.0003214	IBD882	0.0004

Remove list



UC2903
UC2843
IBD657
IBD2440
UC1360
IBD2804
UC2453
UC457
IBD882

(UC107 and UC385)과 UC457은 genetic related가 있을 가능성이 있다. → 때문에 missing genotypes rate와는 상관 없이 UC457만 제거한다.

GWAS Quality Control

- 1. Sample Quality Control

4) Divergent ancestry

Population structure

- 한 데이터에서 집단의 구조가 생긴 것을 말합니다.
- Ancestry 마다 allele frequency가 달라서 disease와의 association이 disease risk의 효과로 나타나는게 아니라 인종차이로 나타날 수 있습니다.

Removal of Population structure.

↓

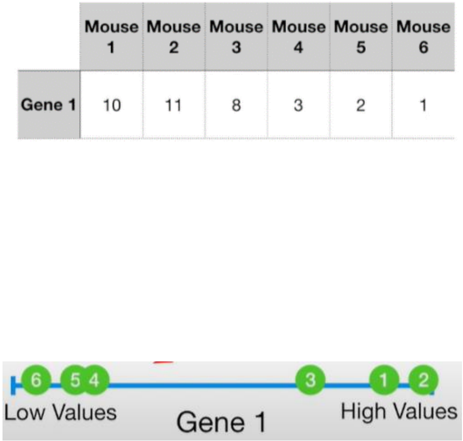
The most common method for identifying individuals with large-scale differences in ancestry **PCA(Principal component analysis).**

GWAS Quality Control

- 1. Sample Quality Control

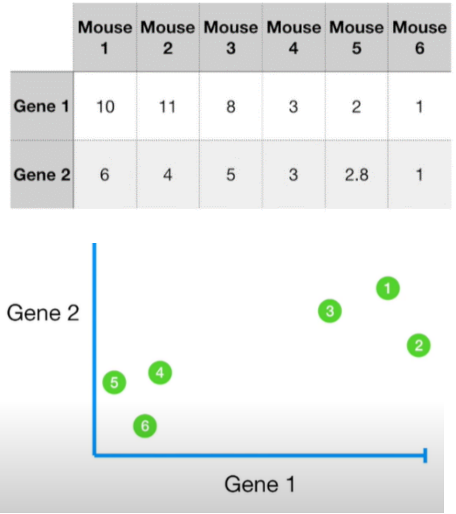
4) Divergent ancestry

PCA(Principal Component Analysis)



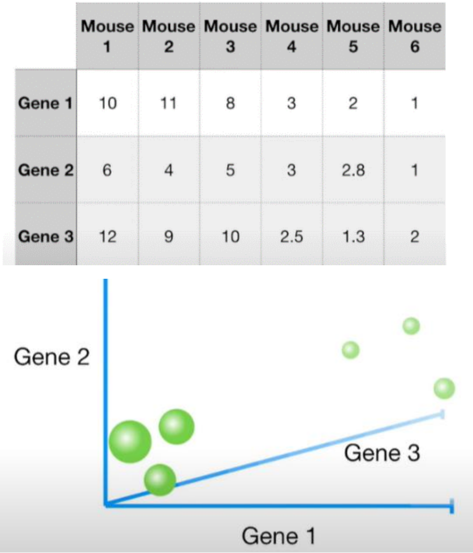
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

1 Dimension



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

2 Dimension



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

3 Dimension

PCA는 data의 분포를 가능한 유지하면서 data의 차원을 고차원에서 저차원으로 축소하여 sample들의 유사성을 확인하는 기법이다.

https://www.youtube.com/watch?v=FgakZw6K1QQ&list=PL1xAVbQe7FkzjepOObDkiwqvKjYDlX2OC&index=1&t=847s&ab_channel=StatQuestwithJoshStarmer, 2023-03-15

GWAS Quality Control

- 1. Sample Quality Control

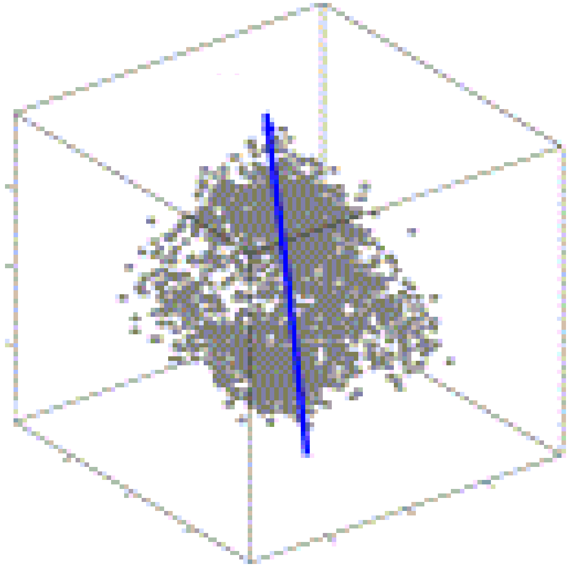
4) Divergent ancestry

PCA(Principal Component Analysis)

- Data의 분포를 가장 잘 설명할 수 있는 선을 찾음
- Data 사이에 line을 그렸을 때 data와 line 사이에 거리의 합이 최소인 line

↓

PC1



Daffertshofer Andreas et al, *ScienceDirect*, 2004

GWAS Quality Control

- 1. Sample Quality Control

4) Divergent ancestry

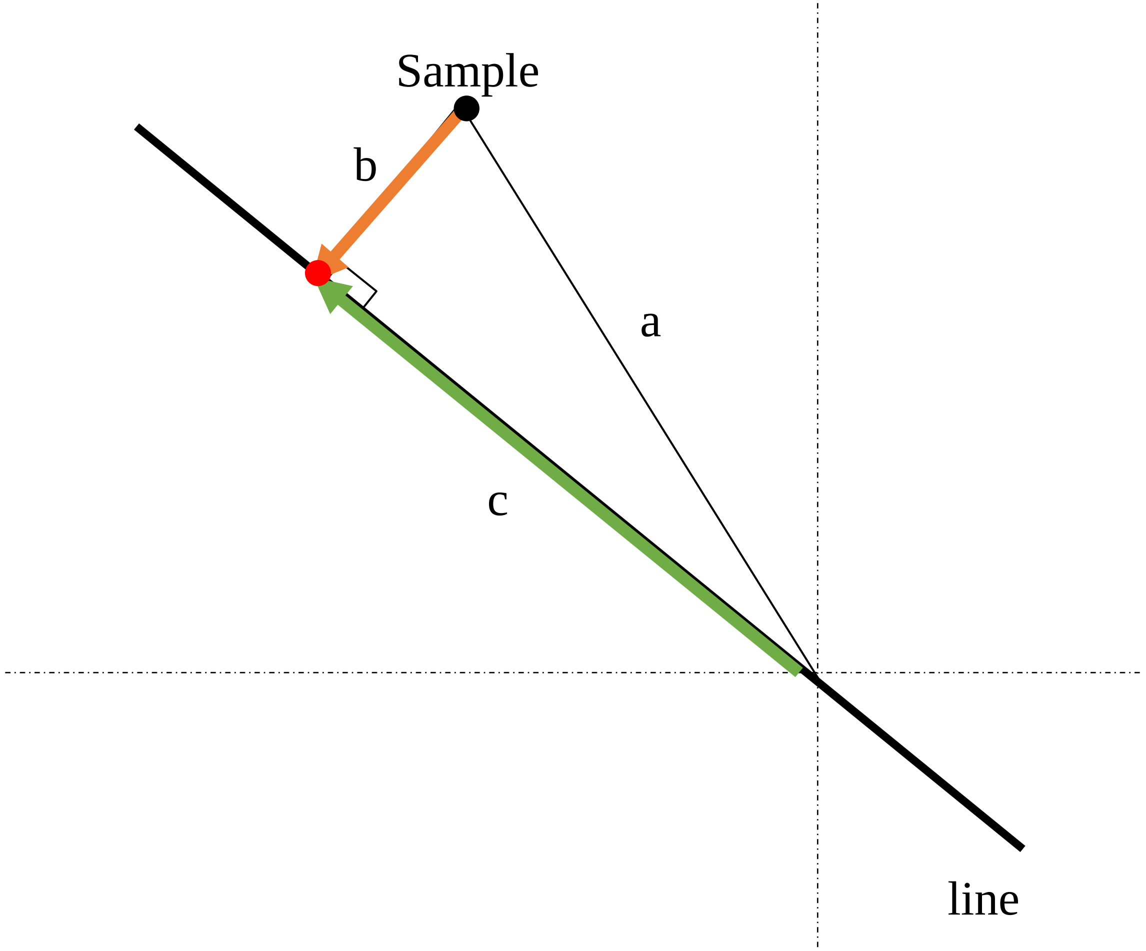
PCA(Principal Component Analysis)

- a : 0에서 sample까지의 거리
- b : Sample 에서 line 까지 거리 (손실된 분산)
- c : 0에서 sample을 line에 일직선으로 내린 점까지의 거리 (보존된 분산)

Datas의 b의 합을 minimum 또는 c의 합을 maximum으로 하는 line을 찾는 과정이다.

↓

line = PC1

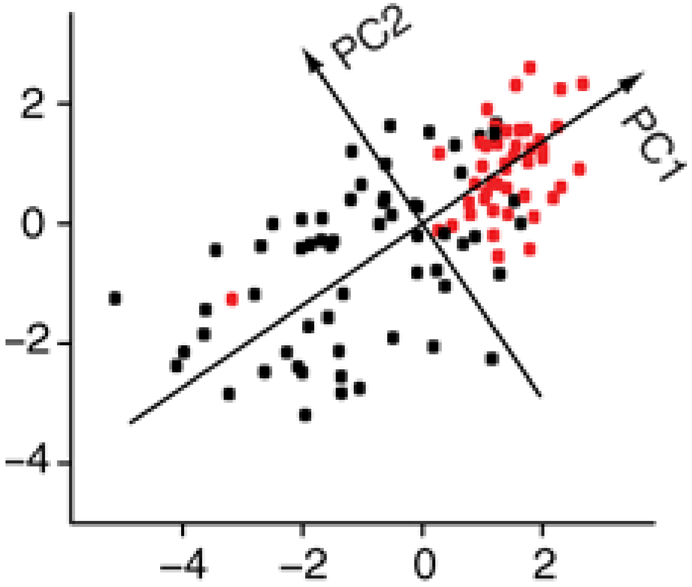


GWAS Quality Control

- 1. Sample Quality Control

4) Divergent ancestry

PCA(Principal Component Analysis)



PC1 다음으로 보존된 분산 값을 가지는 line 중 가장 data의

분포를 잘 표현한 line은 PC1을 직교하는 line

↓

PC2

결론적으로 PC들 중에서 PC1과 PC2가 데이터의 분포를 가장 잘 표현한 PC이다.

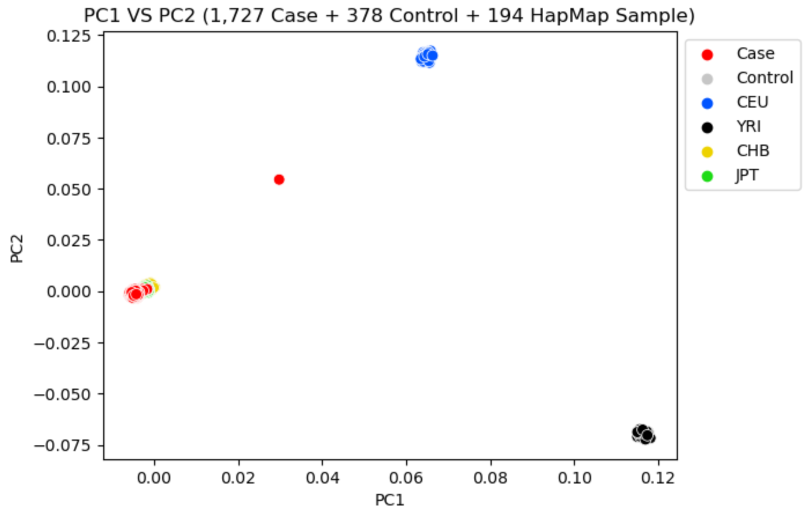
→

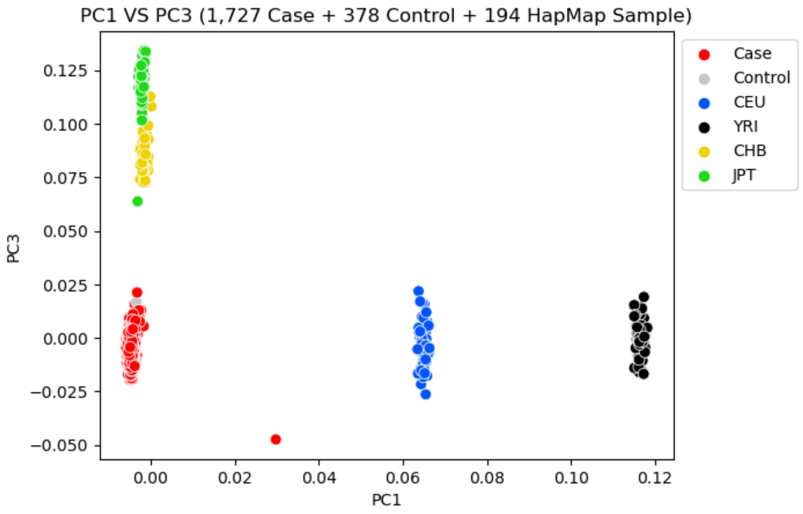
PC1과 PC2를 활용해 시각화 하여 Population structure을 확인하고 제거한다.

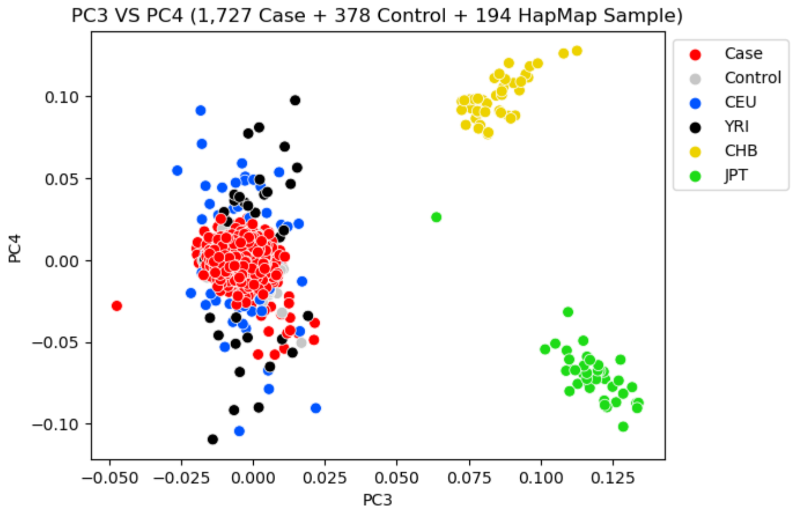
Markus Ringnér, *Nature Biotechnology*, 2008.

PCA(Principal Component Analysis) – Sample + HapMap Project sample

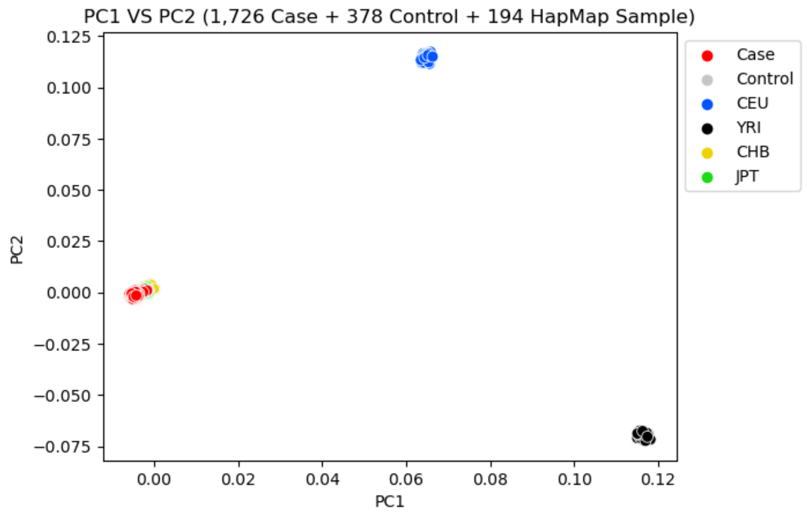
- 1,727 Case(1,001 UC / 726 CD) + 378 Control + 194 HapMap Sample

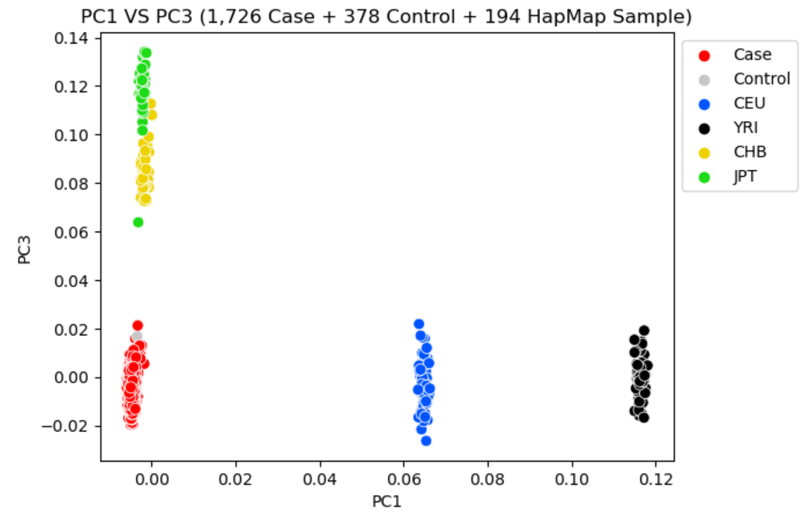


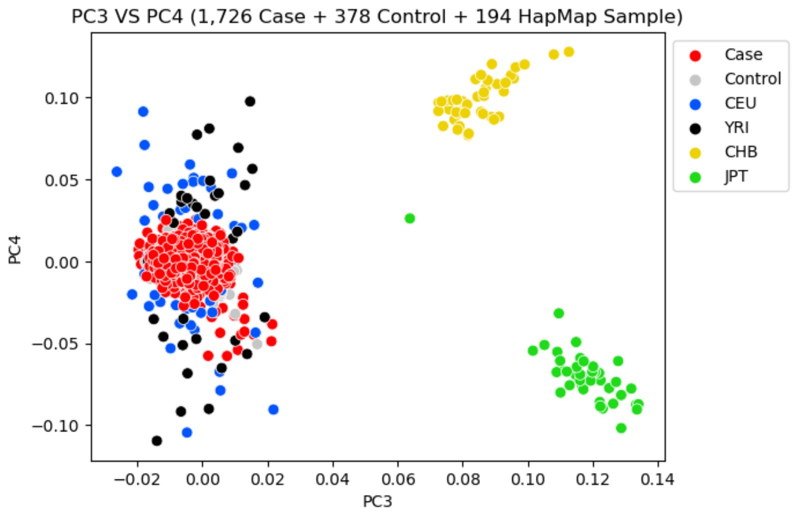




- 1,726 Case(1,001 UC / 725 CD) + 378 Control + 194 HapMap Sample

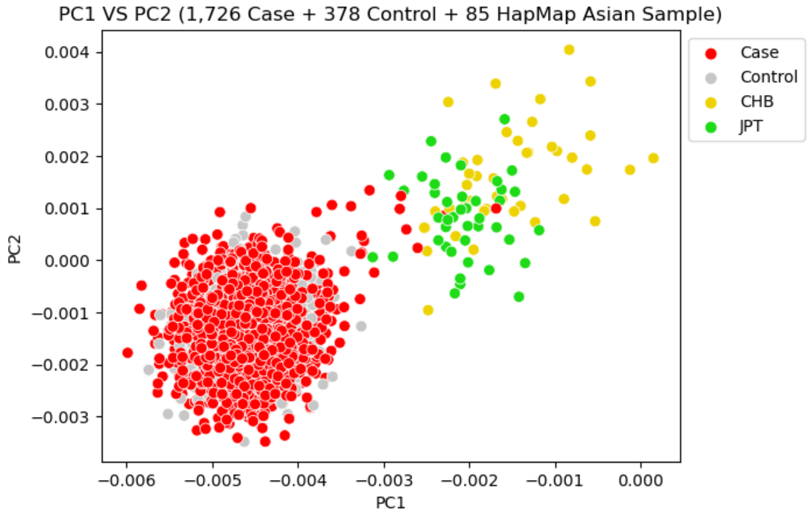




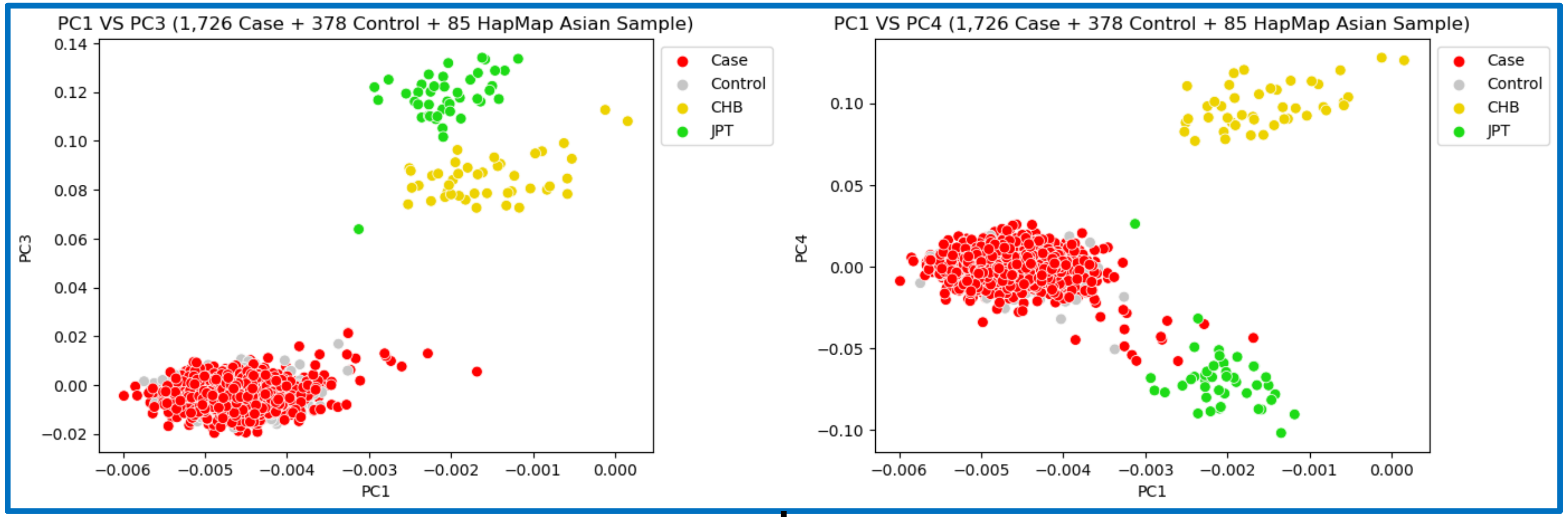


PCA(Principal Component Analysis) – Sample + HapMap Asian sample

- 1,726 Case(1,001 UC / 725 CD) + 378 Control + 85 HapMap Asian Sample

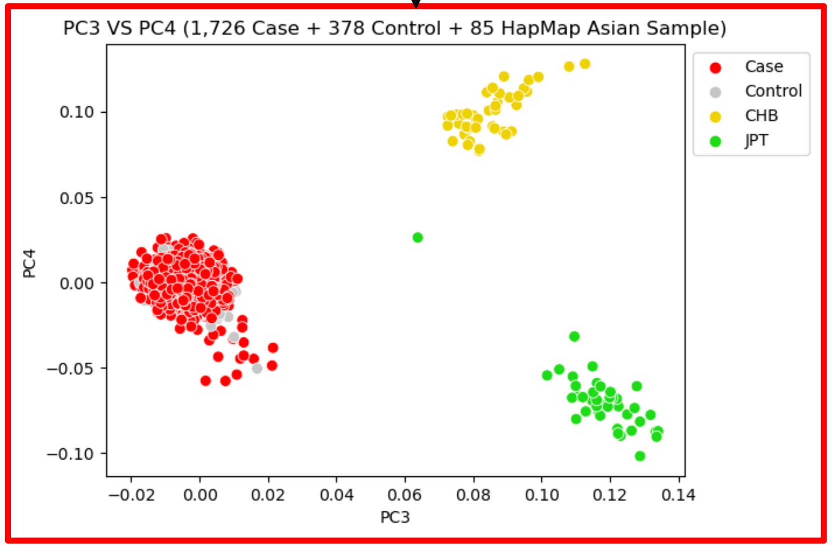


PC1 VS PC2 (1,726 Case + 378 Control + 85 HapMap Asian Sample)



PC1 VS PC3 (1,726 Case + 378 Control + 85 HapMap Asian Sample)

PC1 VS PC4 (1,726 Case + 378 Control + 85 HapMap Asian Sample)

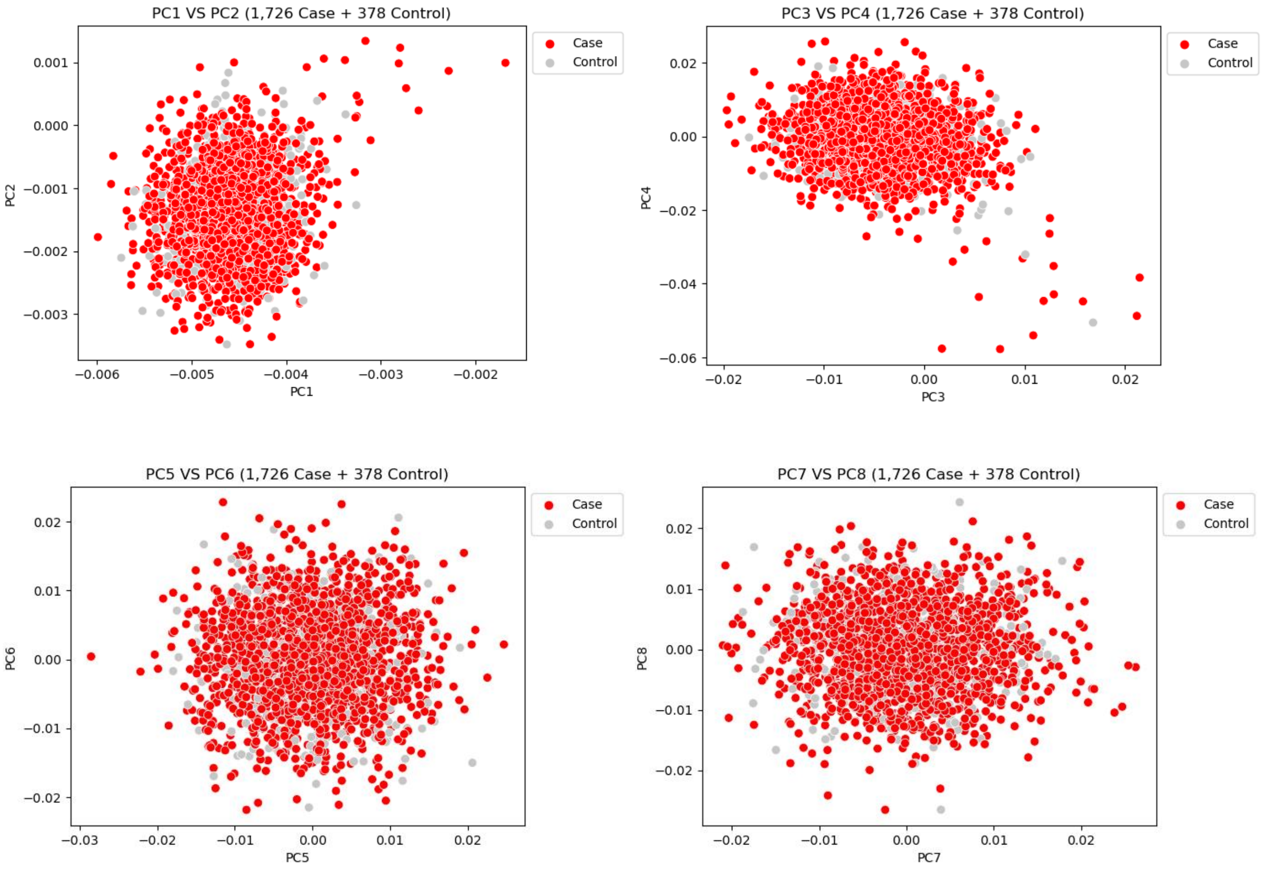


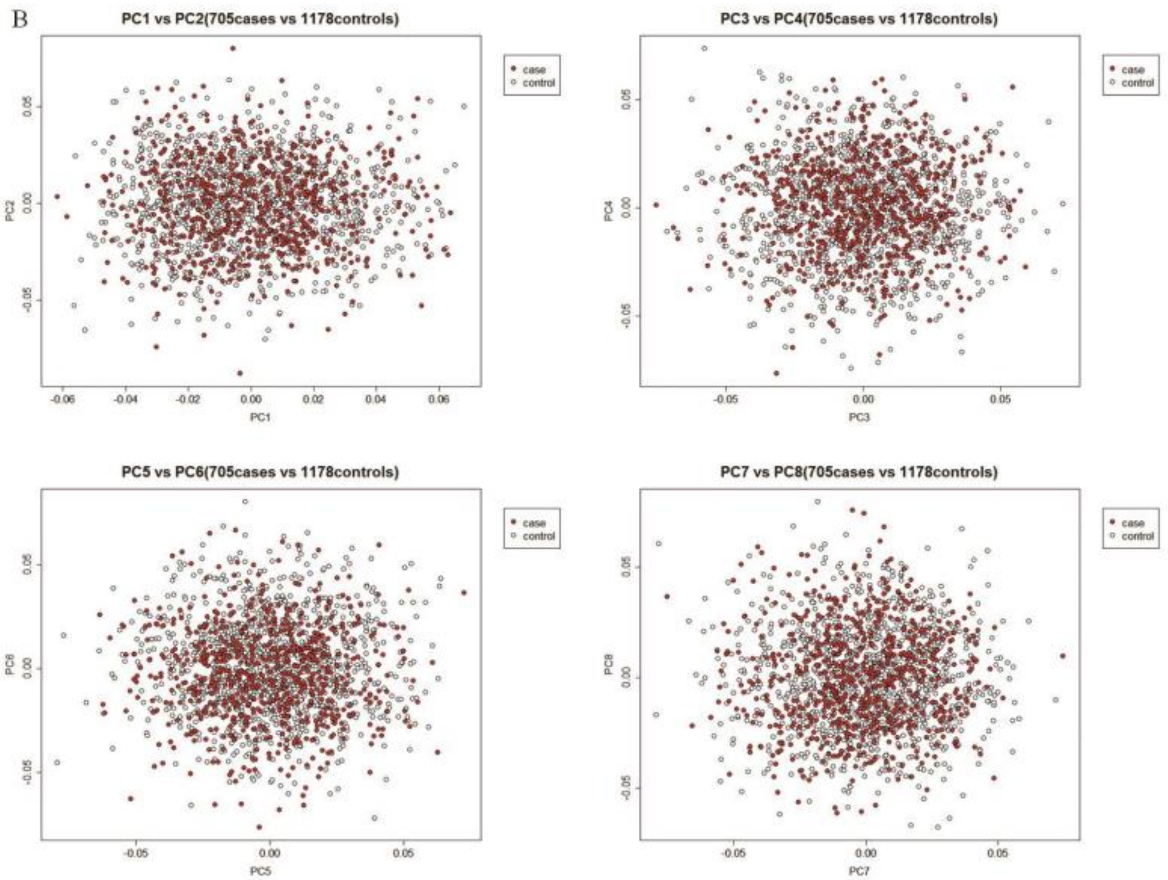
PC3 VS PC4 (1,726 Case + 378 Control + 85 HapMap Asian Sample)

PCA(Principal Component Analysis) – Sample + HapMap Asian sample

- 1,726 Case(1,001 UC / 725 CD) + 378 Control

PCA was used again to detect population stratification among the cases and controls.





PCA analysis suggested minimal genetic mismatch between the cases and controls.

Ye Byong-Duk, *Inflammatory Bowel Diseases*, 2016

GWAS Quality Control

- 2. SNP Quality Control

1) Low MAF(Minor Allele Frequency)

- MAF is actually the second most frequent allele.
- Low MAF는 GWAS 분석에서 noise를 일으킨다.
  - ➢ Case-control association tests에서 False positive association이 나타난다.
  - ➢ 너무 낮은 MAF는 association을 탐지하는 power을 감소시킨다.

❖ **일반적으로 1% 미만의 MAF를 제거한다.**

https://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs_attributes.html, 2023-03-30

GWAS Quality Control

- 2. SNP Quality Control

1.1) MAF 1% fix – Sample size

Allele frequency : $p + q = 1$

Genotype frequency : $p^2 + 2pq + q^2 = 1$

Allele frequency : 0.99 + 0.01 = 1

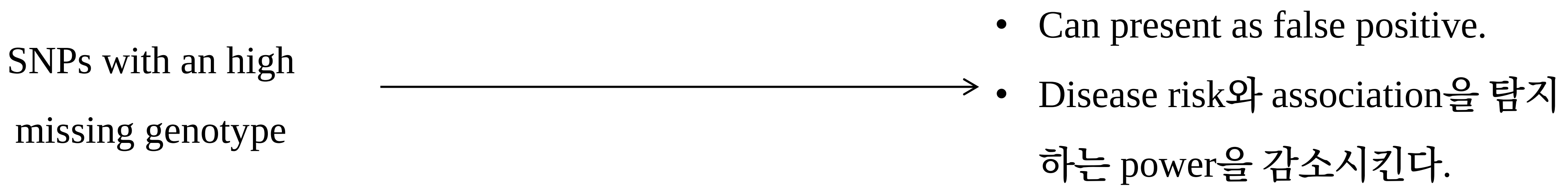
Genotype frequency : 0.9801 + 0.0198+ 0.0001 = 1

Sample	Heterozygous(pq) sample size	Homologous(q^2) sample size
100	1.98	0.01
1,000	19.8	0.1
10,000	198	1
100,000	1980	10
1,000,000	19800	100

GWAS Quality Control

- 2. SNP Quality Control

2) High missing genotype



SNPs with an high missing genotype

- Can present as false positive.
- Disease risk와 association을 탐지하는 power을 감소시킨다.

SNP call rate less than 95~99% are remove.

GWAS Quality Control

- 2. SNP Quality Control

A, a	Allele
p, q	Allele frequency
AA, Aa, aa	Genotype
p^2, pq, q^2	Genotype frequency

3) Significant deviation from HWE

Hardy-Weinberg Equilibrium : Conditions에 만족할 때 집단에서 시간이 지나 세대가 바뀌어도 allele frequency가 유지된다.

Conditions :

- In a large population
- Random mating
- Mutations
- Natural selection
- Migration

Allele frequency : $p + q = 1$

		Female	
		A(p)	a(q)
Male	A(p)	AA(p^2)	Aa(pq)
	a(q)	Aa(pq)	aa(q^2)

$$\begin{array}{ccc} AA & Aa & aa \\ \downarrow & \downarrow & \downarrow \\ p^2 & 2pq & q^2 \end{array}$$

$$p` = p^2 + ½(2pq) = p(p + q) = p$$

$$q` = ½(2pq) + q^2 = q(p + q) = q$$

- Control sample : HWE *P*-value < 0.00001 are removed.
- Case sample : Disease와 연관된 loci가 HWE 상태에서 벗어난 SNP을 제거하면 역효과가 날 수 있으므로 control sample 보다 더 엄격한 threshold를 적용하여 제거한다.

GWAS Quality Control

- 2. SNP Quality Control

4) Significantly different missing genotype rates between Cases and Controls

Present as false-positive associations.

↓

각 SNP 별로 missing genotype rate를 case와 control 샘플에서 각각 계산 후, significant한 차이를 보이는 SNP들을 제거한다.

GWAS Quality Control

- 3. Quality Control measures of Asian Screening Array data

		Samples		SNPs
		Cases (UC / CD)	Controls	
Initial counts		1,746(1,012 / 734)	384	659,184
Pre-QC:	Gender mis-matched samples	8 (5 / 3)	3	
Successfully genotyped		1,738 (1,007 / 731)	381	
SNPs exclusion criteria:	Non-autosomal SNPs			33,446
	In/Del SNPs			8,500
	SNP call rate < 98%			19,180
	MAF < 0.01			137,518
	HWE p < 1E-05 for controls, p < 5E-08 for cases			445
	Duplicated SNPs			2,716
Remaining SNPs				457,379
Samples exclusion criteria:	Sample call rate < 96%	2 (1 / 1)	3	
	PI-HAT > 0.2	9 (5 / 4)	0	
Remaining Samples		1,727(1,001 / 726)	378	
	Different missing genotype rates < 1E-05			84
	PCA	1 (0 / 1)	0	
Final QC data		1,726(1,001 / 725)	378	457,295

laboratory’s pipeline

		Samples		SNPs
		Cases (UC / CD)	Controls	
Initial counts		1,746 (1,012 / 734)	384	659,184
Samples exclusion criteria:	Gender mis-matched samples	8 (5 / 3)	3	
	Sample call rate < 96% Herozygosity rate 3 s.d.	2 (1 / 1)	3	
	PI-HAT > 0.2	9 (5 / 4)	0	
	PCA	1 (0 / 1)	0	
Remaining Samples		1,726 (1,001 / 725)	378	
SNPs exclusion criteria:	SNP call rate < 98%			19,389
	Different missing genotype rate s < 1E-05			128
	MAF < 0.01			162,698
	HWE p < 1E-05 for controls, p < 5E-08 for cases			723
Remaining Samples				476,246
Final QC data		1,726 (1,001 / 725)	378	476,246

Carl A Anderson’s pipeline