Association Test

서울대학교 정보의학교실
최 선

연관분석(Association study)

- 표현형-유전자형 연관성에 대해 통계적 유의성을 판단하는 분석법.
- 유전체의 SNP좌위에 대해 둘 이상의 표현형을 갖는 집단(population)에 대하여 모든 개체로부터 얻은 유전자형(genotype) 정보를 이용해 집단의 표현형과 유전자형의 연관성을 검정. 크게 single SNP level, haplotype에 대한 검정이 가능
- SNP은 질병의 진단목적 이외에도 유전적 질환 추적에 이용될 수 있으며, SNP이 유전자 내 regulatory region에 존재하는 경우 유전자의 기능, 단백질의 생물학적 기능을 변화시켜 직접적으로 질병을 유발시킬 수 있음.

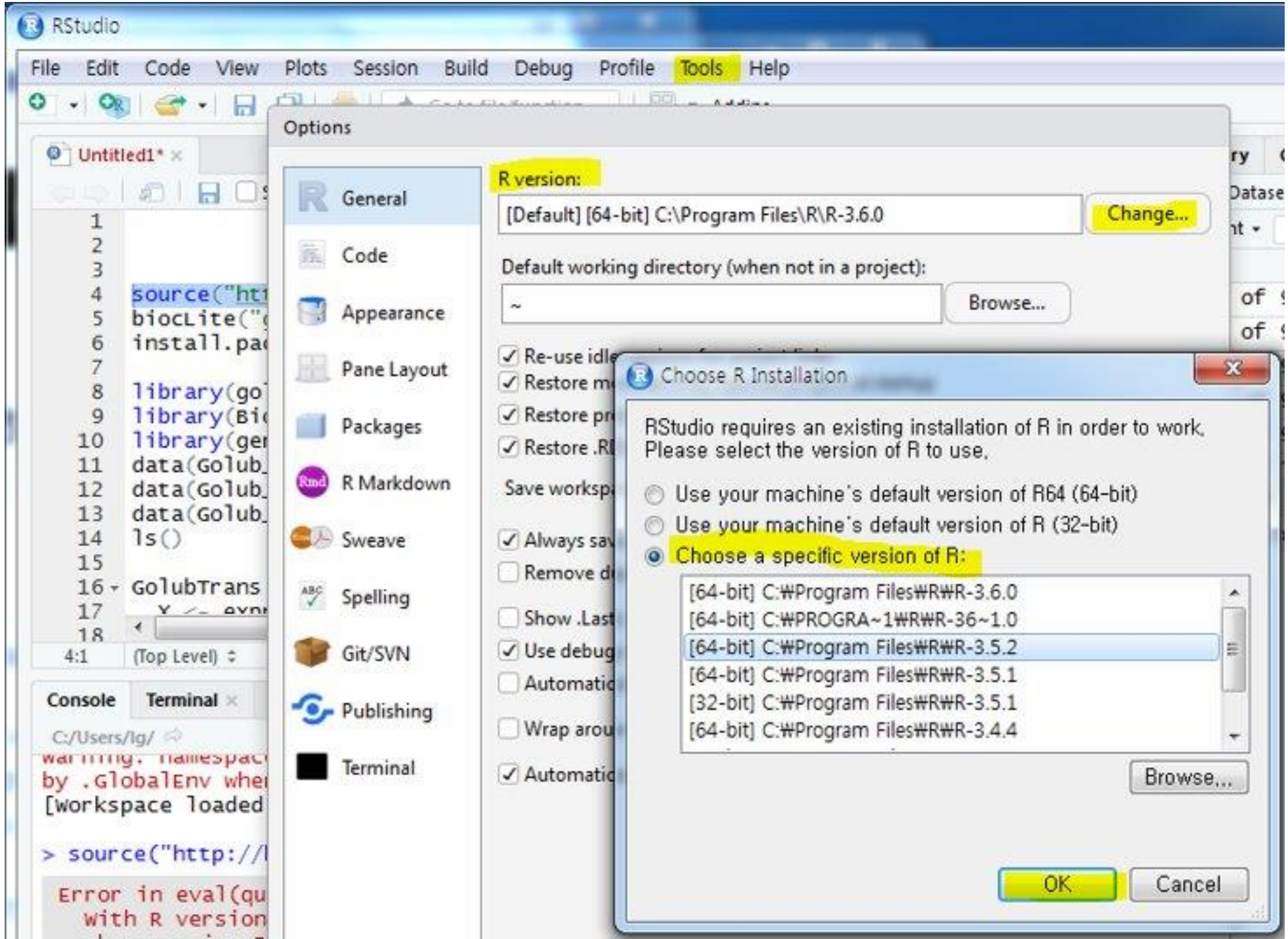
1. Statistical tests for association

    1) Fisher's exact test and Chi-squared test
    2) Odds Ratio (OR)
    3) Cochran-Armitage Trend Test (CATT)
    4) Regression methods
    5) Hardy-Weinberg Equilibrium (HWE) test
    6) MANHATTAN Plots

2. Rare variant association tests

    1) Burden tests
    2) SKAT test
    3) SKAT-O test

3. LD calculation based on 1000 Genomes data
4. HLA imputation


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1*
Options
General
Code
Appearance
Pane Layout
Packages
R Markdown
Sweave
Spelling
Git/SVN
Publishing
Terminal
R version:
[Default] [64-bit] C:\Program Files\R\R-3.6.0
Change...
Default working directory (when not in a project):
~
Browse...
Choose R Installation
RStudio requires an existing installation of R in order to work.
Please select the version of R to use.
Use your machine's default version of R64 (64-bit)
Use your machine's default version of R (32-bit)
Choose a specific version of R:
[64-bit] C:\Program Files\R\R-3.6.0
[64-bit] C:\PROGRA~1\R\R-36~1.0
[64-bit] C:\Program Files\R\R-3.5.2
[64-bit] C:\Program Files\R\R-3.5.1
[32-bit] C:\Program Files\R\R-3.5.1
[64-bit] C:\Program Files\R\R-3.4.4
Browse...
OK
Cancel
Console
Terminal
C:/Users/lg/
[Workspace loaded
> source("http://
Error in eval(qu
With R version

강의자료 다운로드

https://github.com/SEONCHOE/GDA

Virtualbox 이용 시 입력하기
source activate lecture2
R

패키지 설치

```
install.packages("SNPassoc")
install.packages("coin")
install.packages("qqman")
install.packages(“devtools")
install.packages("seqMeta")
library(SNPassoc)
library(coin)
library(qqman)
data(SNPs)
library(seqMeta)
data(seqMetaExample)
```

R STUDIO 단축키

- # 1. 코드실행
- # ctrl + enter
- # 2. 소스 저장
- # ctrl + s
- # 3. command 창 지우기
- # Ctrl+L
- # 4. 주석 처리/해제
- # 해당 라인에 커서를 두고 ctrl + shift + c
- # 5. 함수 또는 R 소스파일의 내용보기
- # 확인하려는 함수 또는 R 소스파일에 커서를 올리고 F2
- # 6. 실행중인 명령어 중지
- # ESC
- # 7. 이전 실행 명령어 창에 띄우기
- # UP / DOWN
- # 8. 이전 실행 명령어 확인
- # CTRL + UP

Single Nucleotide Polymorphisms (SNPs)

- 인간의 유전자는 99.9% 유사하고, 나머지의 0,1%가 특정의 질환에 관한 감수성이나 약제에 대한 부작용 등의 개체 차이에 관여하고 있다고 추정됨. 이 0.1%의 부분에 포함되는 유전자 다형의 대부분을 차지하는 것이 SNP.
- SNPs는 2 개의 대립유전자형 (bi-allele)이 서로 조합을 이루어 존재하는 유전변이형으로, 출현빈도가 높고 genome을 탐색하는데 필요한 신뢰성이 높은 유전자변형으로 알려져 있음.
- 290 base pair (bp)당 하나의 SNP가 존재하는 것으로 추정-> 인간 유전체 전체 염기서열 30억 개 에 존재하는 SNP는 약 1,000만 개가 있을 것으로 추정. (Kruglyak and Nickerson)
- “Polymorphism”: 인구의 1% 이상에서 발생하는 변이 (1% 이하: “mutation”)

Statistical tests for association

1) Fisher's Exact Test

- Fisher's exact test는 실험군과 대조군의 allele count를 비교하여 연관성을 검증.

2) Chi-square Test

- Fisher's exact test뿐만 아니라 Pearson's Chi-squared test도 많이 사용됨. Chi-squared test는 2 x 2 contingency table에서 기대빈도 값이 최소 5를 넘어야 함.

3) Cochran-Armitage Trend Test (CATT)

- 성공률이 증가하거나 혹은 감소하는 일정 방향이 있다는 대립가설에 대해서 모집단의 성공 확률이 같은지 다른지를 검정.

Statistical tests for association

4) Regression Methods

(1) Linear regression

$y = a_1x_1 + a_2x_2 + .. + b$ 꼴로 나타내는 선형방정식 구하기

coefficient: 기울기, intercept : y절편

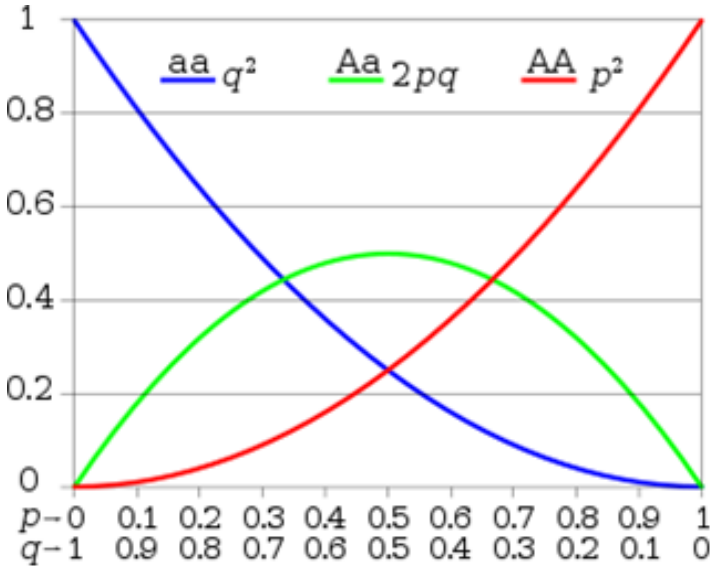
ex) x (snp10001의 변이) 1개 추가-> y(blood.pre) 값이 coefficient(0.10222)만큼 증가

(2) Logistic regression $\log_e(\frac{y}{1-y}) = a_1x_1 + a_2x_2 + .. + b$

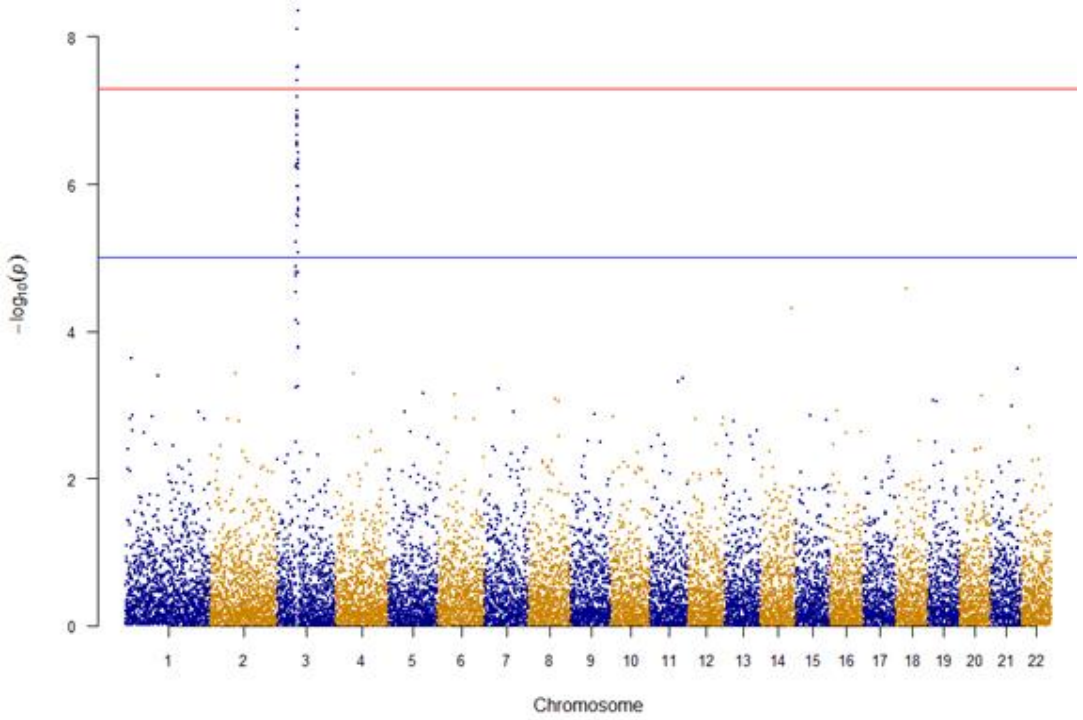
ex) x (snp10001의 변이) 1개 추가 -> case가 될 odds ratio가 coefficient(–0.1447)의 exponential 값(0.8653)

Statistical tests for association

5) Hardy-Weignberg Equilibrium (HWE)



6) MANHATTAN plot for GWAS result



Rare variant association test

- 지난 10년 간 복잡한 질병과 양적 특성의 유전적 구조를 해부하기 위해 GWAS (genome-wide association study)가 광범위하게 사용. 이러한 연구는 MAF (minor allele frequency)가 5% 이상인 일반적인 유전 변이 (Common variant or SNP)를 평가하기 위해서 앞서 살펴본 통계분석을 사용.
- GWAS는 주로 SNP에 초점을 맞추기 때문에, 희귀변이가 앞서 GWAS에서 설명하지 못했던 질병요인 및 특성을 설명해줄 것이라고 여겨지고 있으나 희귀변이에 대한 single-variant test는 샘플의 수를 확보하기 매우 어려워 통계적인 파워가 낮은 편.
- 따라서 희귀 변이는 주로 여러 변이를 유전자 혹은 특정 지역을 기반으로 통합하여 접근하는 방식이 주로 사용.

Burden testing

- A test collapses the variant data within a region by summing the minor allele counts for each marker in the region, and testing this against the phenotype. By contrast to Count Number of Variants (Per Gene), the counts are usually weighted by a function of each marker's minor allele frequency (MAF), so as to establish a contrast between rare and common variants.

$$Q_{\text{burden}} = \left(\sum_{j=1}^{m} w_j S_j \right)^2$$

https://doc.goldenhelix.com/SVS/tutorials/variant_analysis/skatAndSkato.html

Sequence Kernel Association Test (SKAT)

- SKAT is a SNP-set (e.g., a gene or a region) level test for association between a set of rare (or common) variants and dichotomous or quantitative phenotypes, SKAT aggregates individual score test statistics of SNPs in a SNP set and efficiently computes SNP-set level p-values. It collapses the variant data within a region by summing the squares of score statistics for testing individual markers. Weights based on each marker's MAF are usually used to establish a contrast between rare and common variants.

$$Q_{SKAT} = \sum_{j=1}^{m} w_j^2 S_j^2$$

Optimized SKAT (SKAT-O)

- a procedure which optimizes Generalized SKAT over a grid of N values of p between zero and 1, inclusive, in such a way as to count as only one test for multiple testing purposes instead of as N tests. (In Golden Helix SVS, seven grid points are used (N = 7), so we are talking about avoiding having to multiply the number of tests by 7 to get a proper multiple testing correction.)

$$Q_\rho = (1-\rho)Q_{\text{SKAT}} + \rho Q_{\text{burden}}, 0 \le \rho \le 1$$

PLINK를 이용한 연관분석

PLINK

- 표현형-유전자형 연관분석을 위해 일반적으로 사용되는 프로그램. Case/control 비교와 같은 일반적인 분석부터 셋 이상의 표현형 집단에 대해 정량, 정성데이터와 유전자형 간 연관성을 볼 수 있음. 병렬처리를 통한 빠른 속도를 장점

SNP2HLA

- SNP2HLA 알고리즘: HLA 유전자 사이에 있는 intergenic SNP 정보를 이용하여, reference data와 비교함으로써, HLA 유전형의 정보를 예측.

연관불평형(Linkage disequilibrium)

- 한 염색체에 위치하는 2개의 SNPs간 거리가 아주 멀다면, 감수분열 과정에서 교차(Crossover)가 빈번히 발생하는데 이때 재조합은 단일염기수준이 아니라 큰 블록의 단위로 일어나게 되며 이렇게 함께 유전되는 단위를 Haplotype
- 재조합의 일어남으로써 서로 다른 조합의 haplotype이 독립적으로 발생하게 되는데 이처럼 독립적으로 가능한 조합이 모두 발생하는 경우 두 SNPs는 linkage equilibrium(LE, 연관평형) 상태. 그러나 SNPs간의 거리가 매우 가깝다면 2개의 SNPs는 서로 연관되어 다음 세대에 같이 전달되게 되고 (Gabriel SB, 2002), 이때 haplotype의 조합이 모두 존재하지 않는다면 linkage disequilibrium(LD, 연관불평형) 상태.이런 이론을 바탕으로 LD는 한 집단에서 2개 이상의 Loci에 존재하는 유전자 간의 통계적 관련성을 나타내는 하나의 지표로 사용.
- 양적형질 유전자 발굴을 위한 LD mapping 연구 흐름: 유전체 상의 표지인자의 유전자형을 이용하여 LD block 구조를 분석. 각 LD block에 존재하는 haplotype의 규명. 대표적인 마커가 되는 htSNP (Haplotype-tagging SNP) 혹은 다른 tagging 표지인자를 찾아내어 통계모형으로 표현형과 유전자형간의 통계모형으로 연관분석을 수행하여 형질에 관련된 유의한 LD의 위치를 찾기

*.PED 파일

Column 1 = Family ID
Column 2 = Invididual ID
Column 3 = Parernal ID (0 인 경우는 missing)
Column 4 = Maternal ID (0 인 경우는 missing)
Column 5 = Sex
Column 6 = Phenotype (1,2, 또는 0 으로 구분. [1=unaffected, 2=affected, 0=missing])
Column 7+8 = SNP1의 genotype pair (0 인 경우는 missing)
Column 9+10 = SNP2의 genotype pair (0 인 경우는 missing)
....

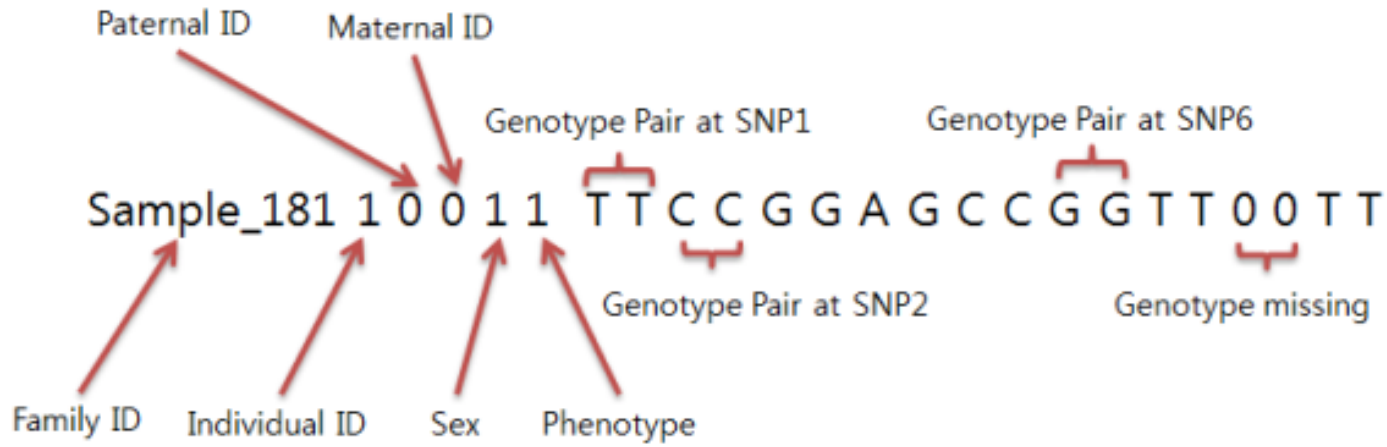
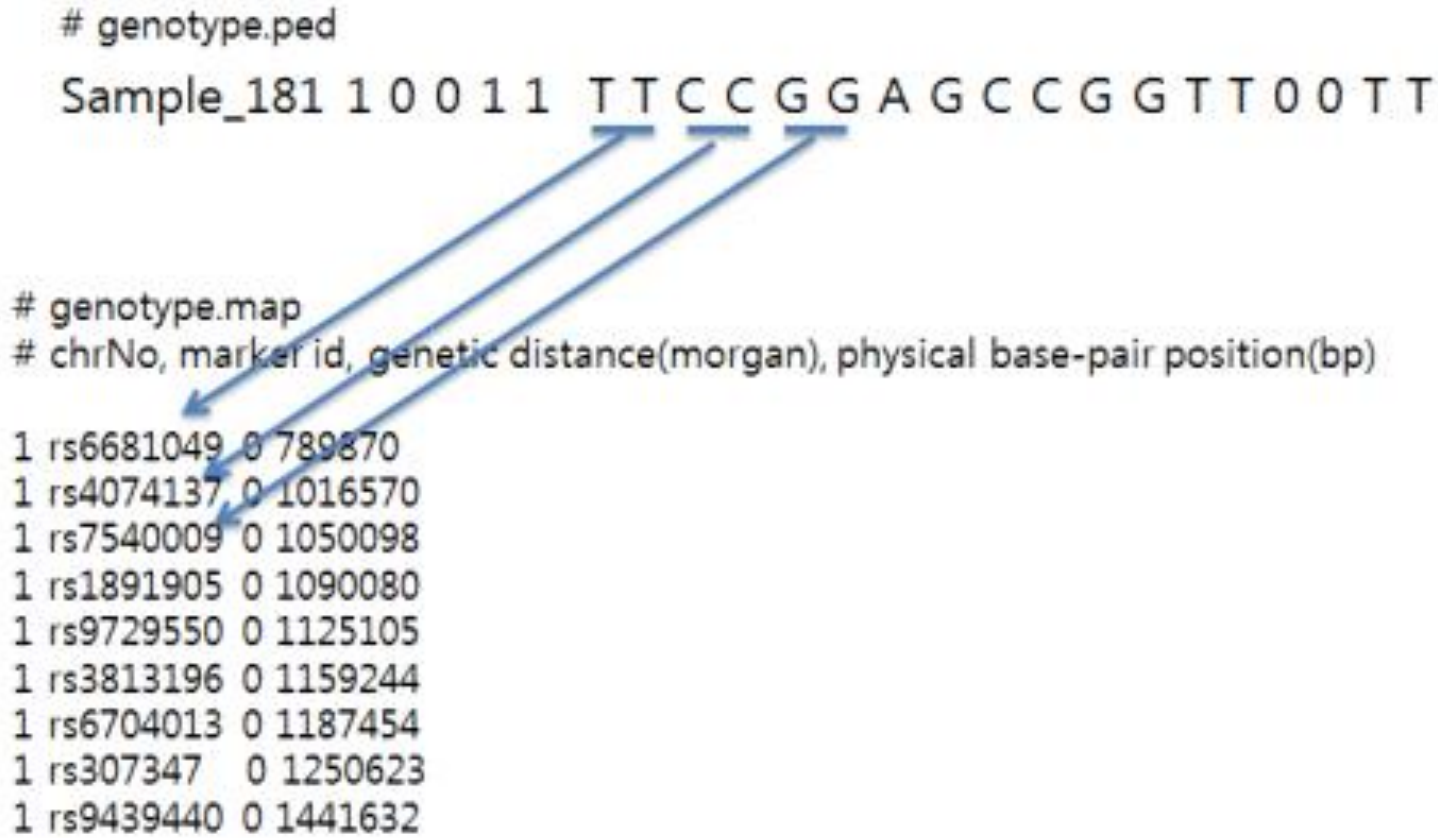


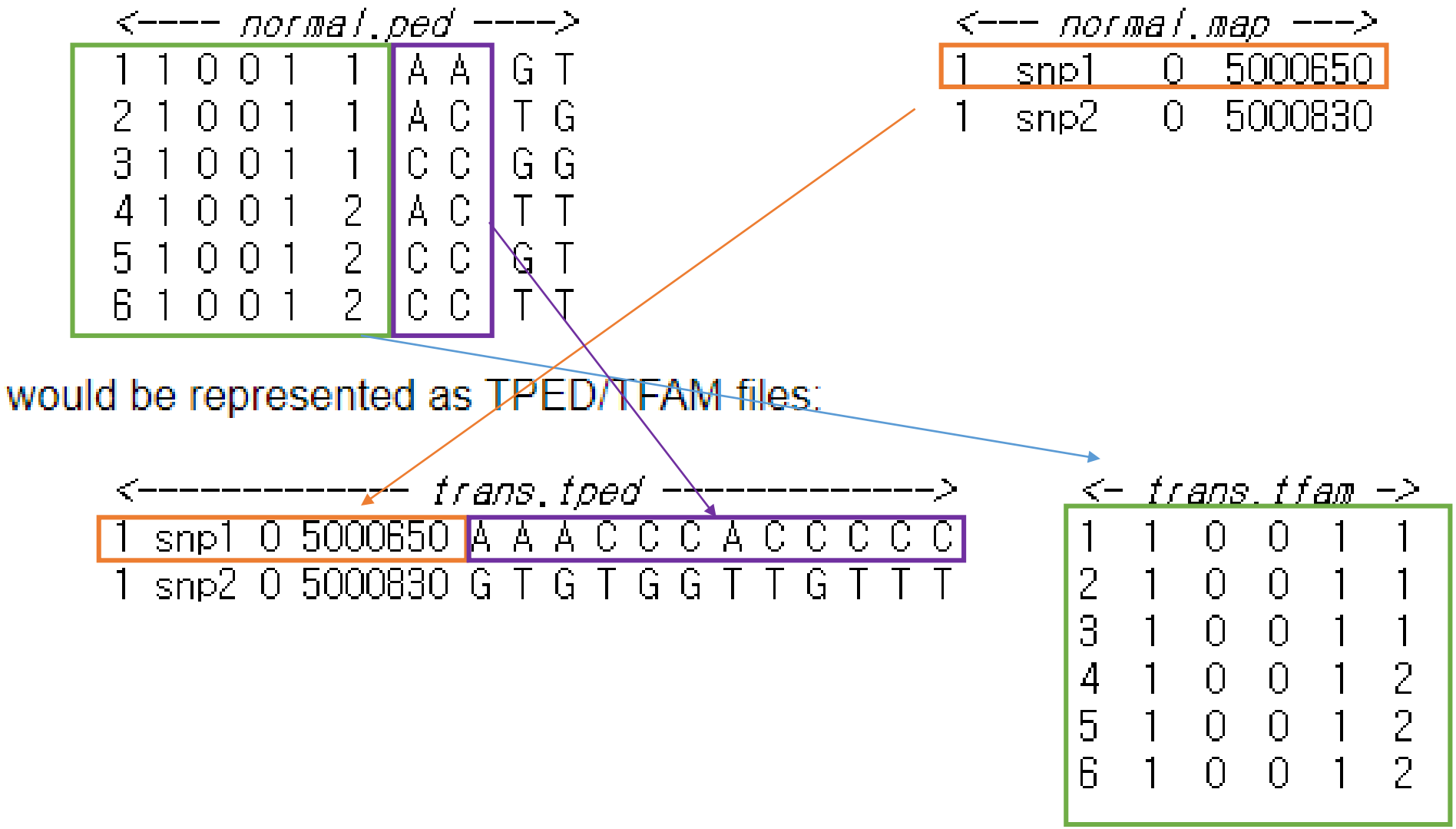
그림 PED 파일 형식

*.MAP

Column 1 = chromosome number
Column 2 = SNP ID
Column 3 = Genetic Distance (morgans)
Column 4 = physical base-pair position (bp)

따라서 2개의 파일은 아래와 같은 관계를 갖고 있다.





- TPED: one row -> SNP / TFAM: one row -> individual

*.BIM 파일

- Chromosome • Marker ID • Genetic distance • Physical positi on • Allele 1 • Allele 2

Example of a BIM file of the binary PLINK format:

• 21	rs115116470	26765	A	T
• X	rs3883674 0	32380	C	G
• X	rs122188820	48172	T	T
• 9	rs109040450	48426	A	T
• 9	rs107519310	49949	C	T
• 8	rs112521270	52087	A	C
• 10	rs127752030	52277	A	A
• 8	rs122556190	52481	G	T

*.FAM 파일

- Family ID • Sample ID • Paternal ID • Maternal ID • Sex (1=male; 2=female; other=unknown) • Affection (Phenotype) (0, -9 = missing, 1=unaffected, 2=affected)

```
EAS_JPT_NA18939_F EAS_JPT_NA18939_F 0 0 0 -9
EAS_JPT_NA18940_M EAS_JPT_NA18940_M 0 0 0 -9
EAS_JPT_NA18941_F EAS_JPT_NA18941_F 0 0 0 -9
```

Linkage disequilibrium 지수

(1) R- square : Linkage disequilibrium 지수

- Pairwise SNP 간에 계산된 correlation coefficient(=r)을 제곱한 값이 지정한 값보다 큰 SNP만을 tagging SNP로 간주
- 0에서 1 사이의 범위로 0일 때 perfect equilibrium, 1일 때 완전 LD 상태를 의미, 보통 1/3 이상의 값이면 강한 LD 상태.

(2) D’ : Linkage disequilibrium 지수

- D’ 값이 1이면 완전 LD 상태, D’ < 1이면 이전 세대 어디에선가 유전자 재조합이 일어났음을 의미, D’ 값이 0이면 연관 평형 상태.
- D’값이 낮을수록 두 site 간에 강한 유전자 재조합과 돌연변이가 빈번하게 나타났음을 의미.

R2 vs D’

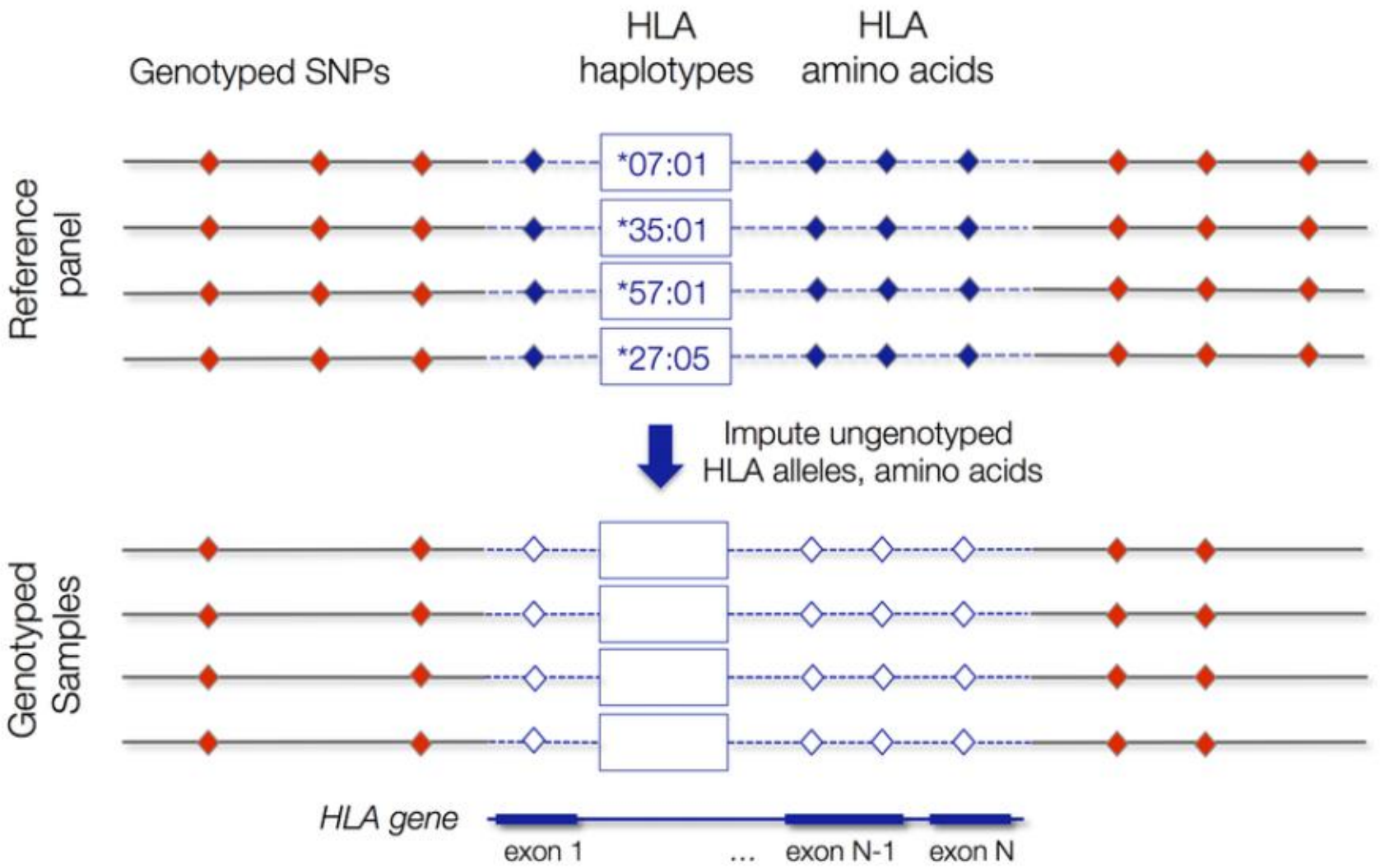
- R2 -> successfully identify LD for common variant
- But drop with low allele frequencies -> unlikely deal with rare alleles
- D’ -> theoretically can have better chance of identifying LD a mong rare variants (>R2)
- But may accumulate more false positives

plink --bfile JPT --chr 16 --from-bp 54000000 --to-bp 54050000 --ld-snp rs11646260 --r2 --ld-window-r2 0 --ld-window 999999 --ld-window-kb 99999 --noweb --out ./JPT

- --bfile : 이전 단계에서 생성한 binary file을 입력/ --chr: chromosome 위치
- --from-bp : 시작 base pair의 위치/ --to-bp : 끝 base pair의 위치
- --ld-snp : ld를 계산할 기준이 되는 snp를 입력
- --r2 : phased haplotypes을 이용하여 r2값을 계산
- --ld-window-r2 : 결과를 filtering 할 때 사용하는 r2값을 의미
- --ld-window : 결과를 filtering 할 때 사용하는 LD값을 계산하기 위해 사용된 SNP 들 사이의 최대 SNP 수
- --ld-window-kb : 결과를 filtering 할 때 사용하는 LD값을 계산하기 위해 사용된 SNP 사이의 최대 물리적 거리 (kb)

SNP2HLA: Imputation of Amino Acid Polymorphisms in Human Leukocyte Antigens

- SNP2HLA is a tool to impute amino acid polymorphisms and single nucl eotide polymorphisms in human luekocyte antigenes (HLA) within the major histocompatibility complex (MHC) region in chromosome 6.
- The unique feature of SNP2HLA is that it imputes not only the classical HLA alleles but also the amino acid sequences of those classical alleles, so that individual amino acid sites can be directly tested for association. This allows for facile amino-acid focused downstream analysis.
- SNP2HLA 알고리즘: HLA 유전자 사이에 있는 intergenic SNP 정보를 이용하여, reference data와 비교함으로써, HLA 유전형의 정보를 예측.
- 실습데이터: 1958년생 영국인 10명 샘플 (1958BC), reference: 1000 Genomes 의 CEU (HM_CEU_REF; Utah residents)



Overview of the SNP2HLA imputation procedure. The reference panel (top) contains SNPs in the MHC, classical HLA alleles at the class I and class II loci, and amino acid sequences corresponding to the 4-d igit HLA types at each locus. For a data set with genotyped SNPs across the MHC (bottom), we use the reference panel to impute classical alleles and their corresponding amino acid polymorphisms.

./ SNP2HLA.csh 1958BC HM_CEU_REF 1958BC_IMPUTED plink 2000 1000

-> CEU reference data 에 맞춰 10명 샘플 (1958BC) imputation

- 1958BC_IMPUTED.dosage: 모든 markers (HLA alleles, amino acids, SNPs)로부터 top statistical peak를 찾기 위한 imputation을 수행한 allele dosage data이다. Dosage 파일은 Beagle 프로그램의 포맷으로, 행은 markers, 열은 marker의 정보와 개인별 imputation 값 나타냄

- 1958BC_IMPUTED.bgl.phased: genotype의 imputation 수행한 파일. Beagle 포맷으로, 행은 markers 정보를 나타내며 2개의 열이 한 개인의 genotype 정보를 나타냄

- 1958BC_IMPUTED.bgl.gprobs: markers (HLA alleles, amino acids, SNPs) 에 대한 imputation posterior probabilities를 계산한 결과

- 1958BC_IMPUTED.bgl.r2: genotype을 이용하여 impute predicted r2 (correlation)를 계산한 결과

```
plink --noweb --dosage 1958BC_IMPUTED.dosage
noheader format=1 --fam 1958BC_IMPUTED.fam --
logistic --out 1958_IMPUTED
```

--noweb: 웹 접속을 차단하고 local 컴퓨터에서 실행.
--dosage OUTPUT.dosage: OUTPUT.dosage 파일을 생성.
noheader: header가 존재하지 않을 경우 사용하는 옵션
format=N: Dosage, two probabilities or three (N=1,2,3)
--fam OUTPUT.fam: OUTPUT.fam 파일을 생성.
--logistic: disease traits에 대해 logistic regression model 계산
--out OUTPUT.assoc: 결과 파일 OUTPUT.assoc 파일을 생산

.assoc.dosage 필드정보

```
gda@GDA:~/lecture02/snp2hla$ more 1958BC_IMPUTED.assoc.dosage
          SNP   A1   A2     FRQ    INFO      OR      SE       P
  rs13207673    A    G  0.5251  0.0132      NA      NA      NA
   rs9356991    T    G  0.7026  0.0216      NA      NA      NA
```

- SNP: SNP identifier,
- A1: Tested allele (minor allele by default),
- A2: Major allele,
- FRQ : Frequency of A1 from dosage data
- INFO : R-squared quality metric/ Information content
- OR: Odds ratio for association
- SE: Standard error of effect estimate
- P: P-value for association test