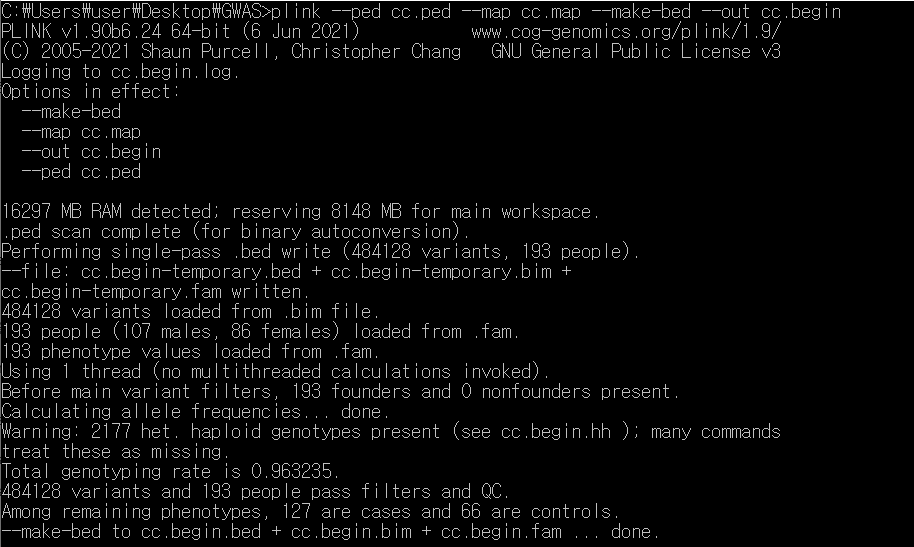
**사용데이터**

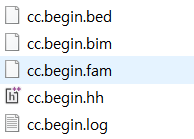
데이터는 <https://ibg.colorado.edu/cdrom2019/colodro_grasby/GWAS_QC_part1/>에서 얻었다.

데이터는 193명의 유전정보를 담고 있다.



이제 ped와 map 파일을 plink 형식의 파일로 변환해준다. 명령어는 plink --ped cc.ped --map cc.map --make-bed --out cc.begin로 한다.

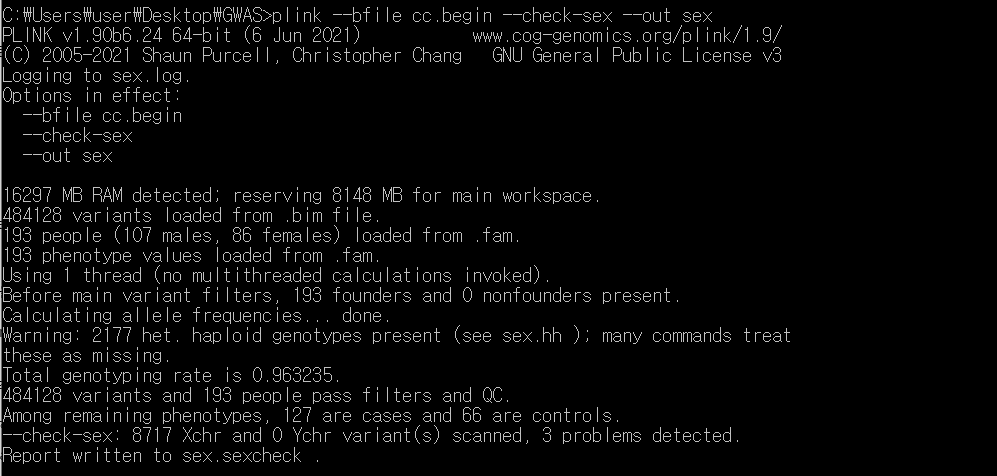




위와 같이 생성된 것을 확인할 수 있다. 이제 이 데이터를 이용해서 분석을 시작한다.

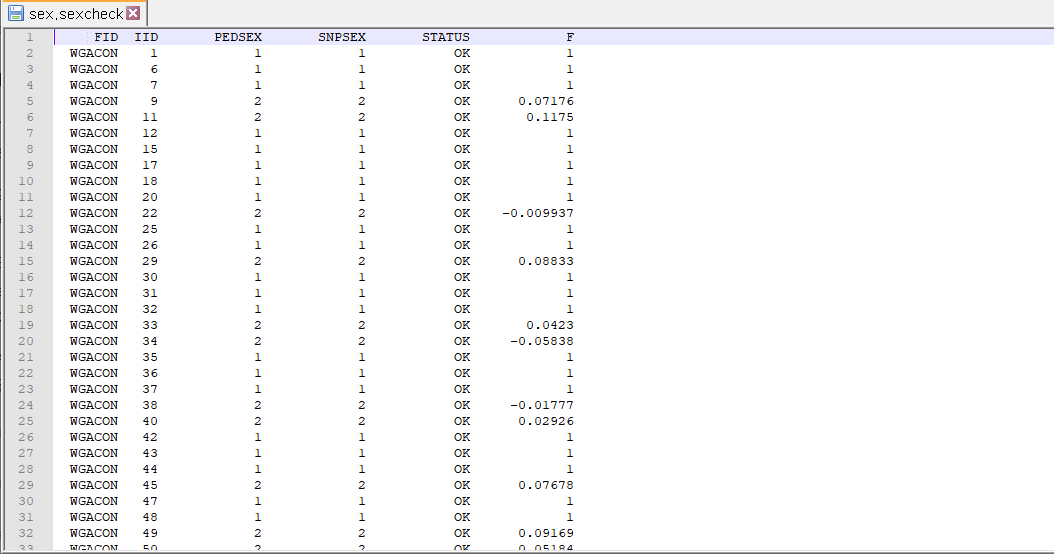
**Genotyping error 데이터 삭제**

plink --bfile cc.begin --check-sex --out sex 명령어를 통해서 성별 분석을 실시한다.

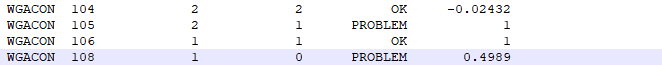


위에서 보이는 것처럼 193명 중 107명이 남자 86명이 여자인 것으로 나왔다.

생성된 파일을 확인해보면 다음과 같다.



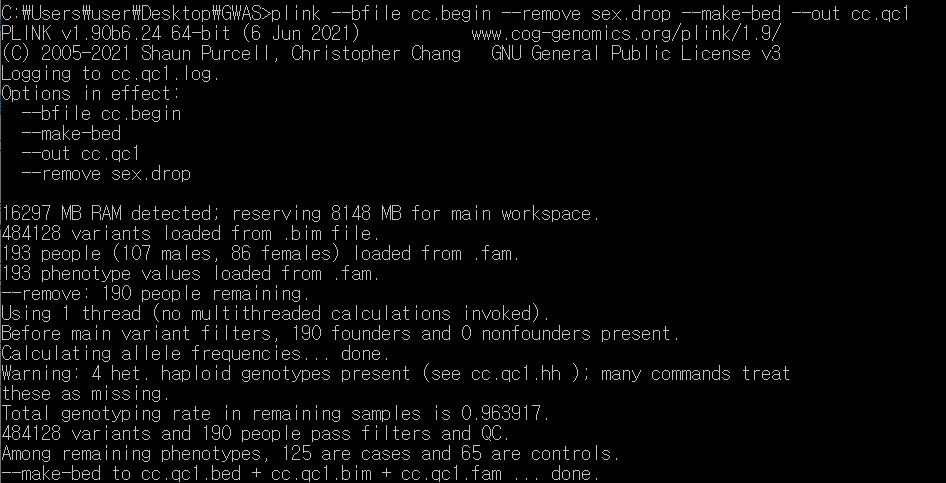
세명이 문제가 있다고 나왔는데 확인해보면 PED에서 나온 성별과 SNP에서 나온 성별이 다른 것으로 나왔다.

sex.sexcheck 파일에서 PROBLEM 상태인 3 데이터를 삭제한 데이터를 생성한다.

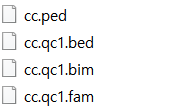
명령어는 다음과 같다.

plink --bfile cc.begin --remove sex.drop --make-bed --out cc.qc1



193명에서 3명이 줄어들어 190명의 데이터가 남은 것을 확인할 수 있다.

제거된 데이터가 저장되어 다음과 같이 생성된 것을 확인할 수 있다.



이제 나머지 문제가 pseudo-autosomal 영역 때문인지를 확인하기 위해서 split-x를 사용한다.

명령어는 다음과 같다.

plink --bfile cc.qc1 --split-x b37 no-fail --make-bed --out cc.qc2

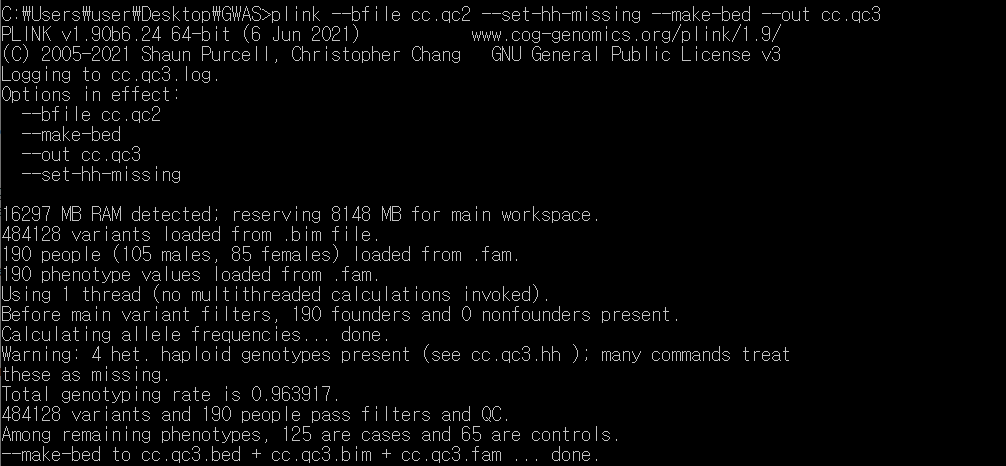


염색체를 확인하기 위해 다음과 같은 awk 코드를 입력한다.

awk “NR==FNR{a[$2]=$1; next} $3 in a{print $0,a[$3]}” cc.begin.bim cc.qc2.hh > check.hh

나머지 het haploid 문제를 누락으로 설정한다.

plink --bfile cc.qc2 --set-hh-missing --make-bed --out cc.qc3



이것들은 이제 분석에서 제외된다.

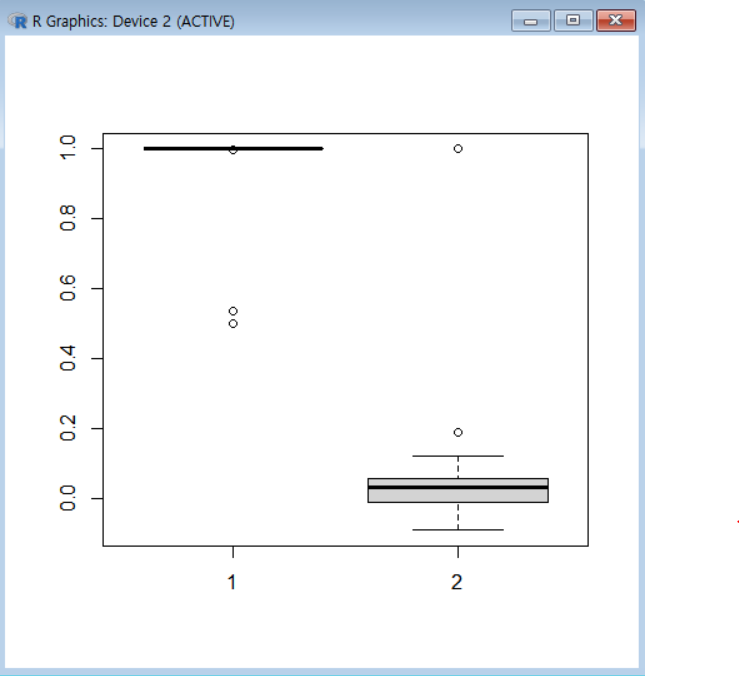
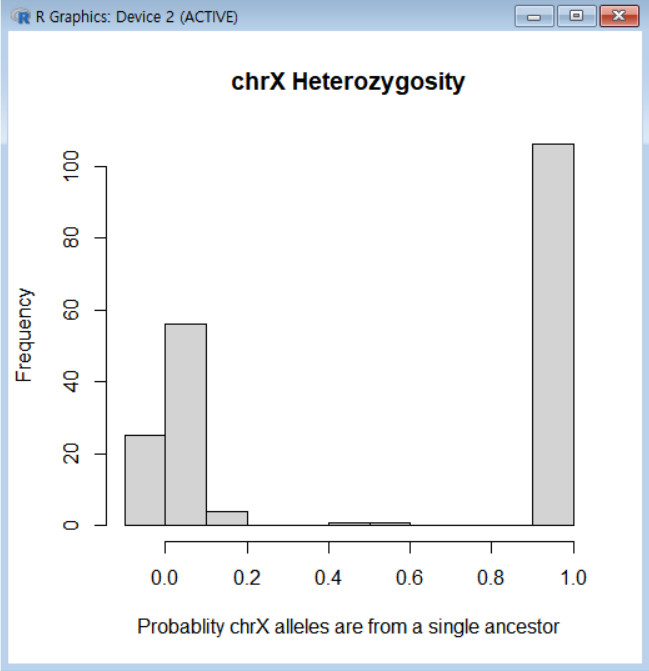
이제 plink --bfile cc.qc3 --het --out het 명령어를 통해 heterozygosity 정보를 얻는다.



이제 다음과 같은 명령어로 각 사람에 대한 heterozygosity의 비율로 파일을 만든다.

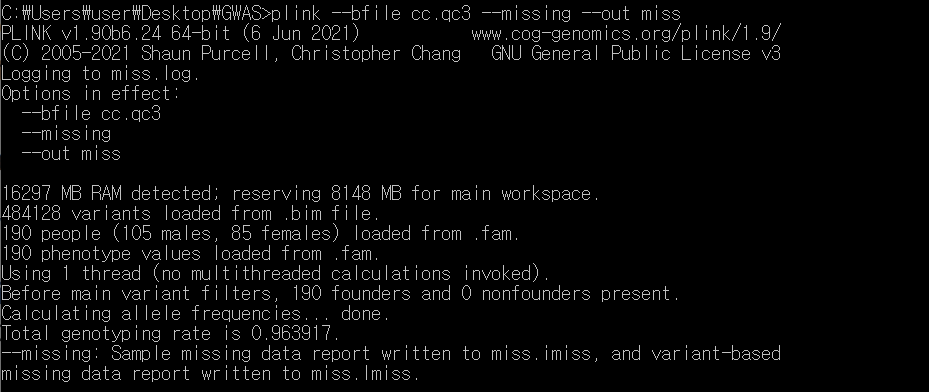
echo "FID IID obs\_HOM N\_SNPs prop\_HET" > het.txt

awk "NR>1{print $1,$2,$3,$5,($5-$3)/$5}" het.het >> het.txt



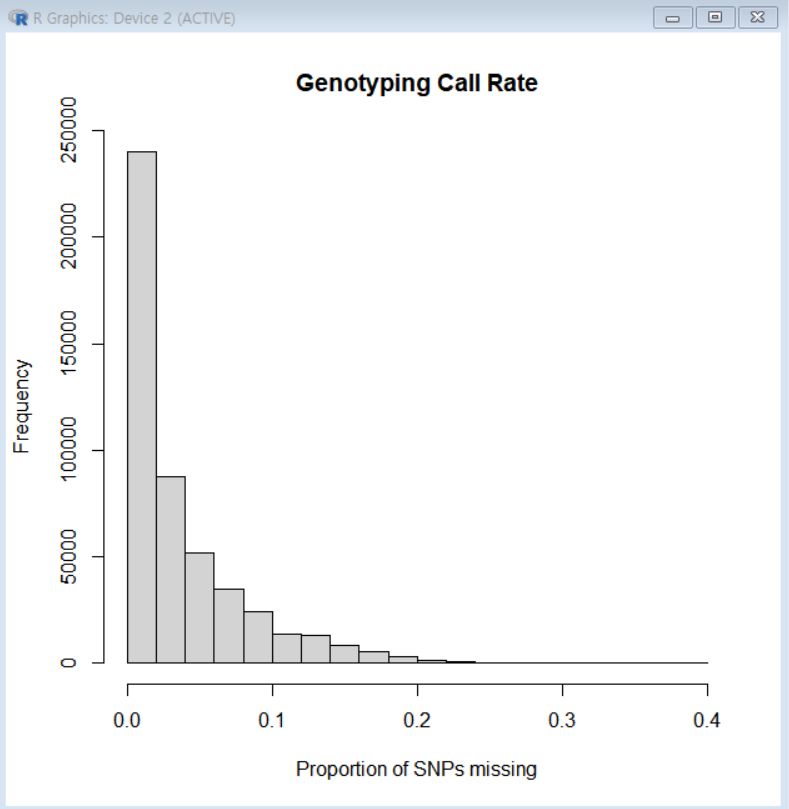
이제 개인 및 SNPs에 대한 누락에 대한 파일을 생성한다.

plink --bfile cc.qc3 --missing --out miss



이제 위에서 만든 heterozygosity의 비율과 누락에 대한 파일을 모두 포함하는 파일을 생성한다.

awk "NR==FNR{a[$1,$2]=$5;next}($1,$2) in a{print $1,$2,$6,a[$1,$2]}" het.txt miss.imiss > het.imiss.txt



**데이터 QC방법**

1) geno: SNP를 Filter하는 옵션으로 전체 샘플에서 각각의 SNP가 존재하는지 확인한다. genotype의 missing 비율에 따라 genotype marker를 제거하는 기능이다.

2) mind: 샘플을 Filter하는 옵션으로 각각의 샘플에 대해 전체 SNPs의 비율을 확인한다. 샘플에서 변이 marker의 Missing 비율이 많은 샘플을 제거해주는 기능이다.

3) maf: 각 SNP의 Frequency를 확인하는 옵션으로 minor allele frequency를 확인한다. Minor한 allele의 비율이 너무 작을 경우 phenotype에 영향을 줄 가능성이 적고 genotype error일 가능성이 있기 때문에 제거를 해주는 기능이다.

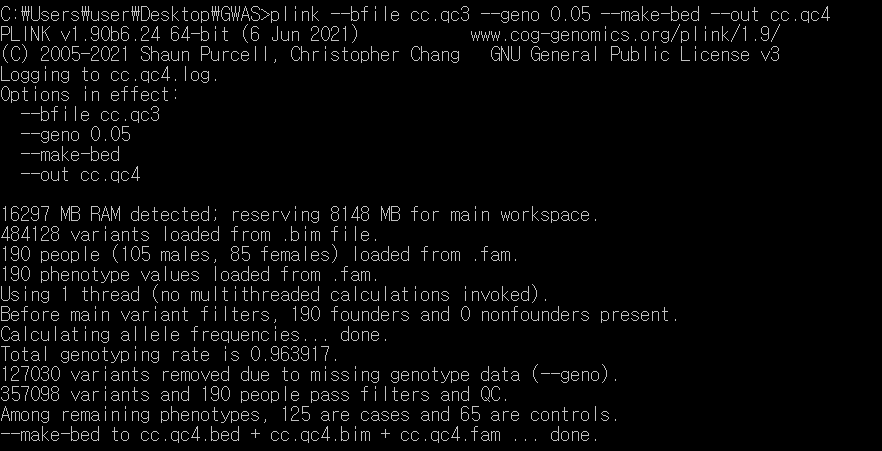
4) hwe: hardy 명령어를 통해 pvalue와 genotype의 개수를 확인이 가능한데 hwe의 p-value가 1로 갈수록 HWE 만족하고 p-value가 낮은 변이 마커일수록 genotype error라고 생각을 하는 것이다

**Low quality or quantity of DNA 제거**

(1) geno를 통해서 너무 많은 데이터가 누락된 데이터를 제거한다.

plink --bfile cc.qc3 --geno 0.05 --make-bed --out cc.qc4

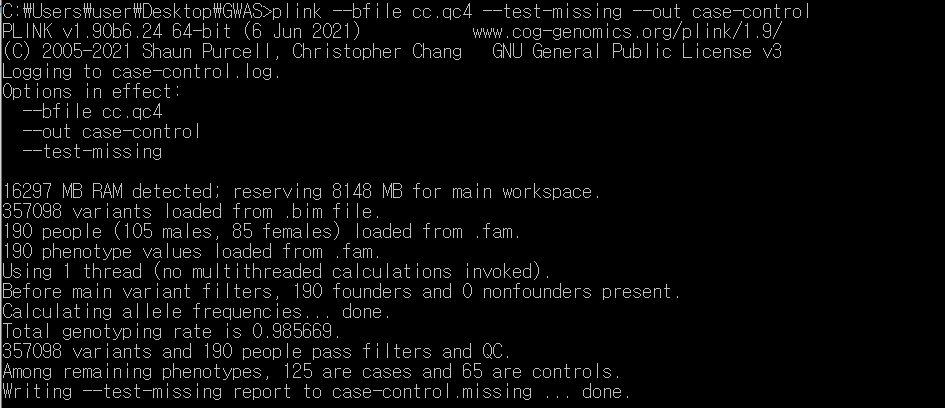
missing rate이 5% 이상이 되는 SNPs들을 제거해준다.



127030개의 SNPs들이 제거되고 357098개의 SNPs와 190명의 사람들의 정보가 남아 저장된다.

Case-Control 데이터를 사용해서 SNPs의 call rate가 실험군과 대조군 간에 유의하게 다르지 않은 것을 확인한다.

plink --bfile cc.qc4 --test-missing --out casecontrol



다음 명령어를 통해 p-value가 0.00001보다 작은 경우 제외하기 위한 파일을 생성해 저장한다.

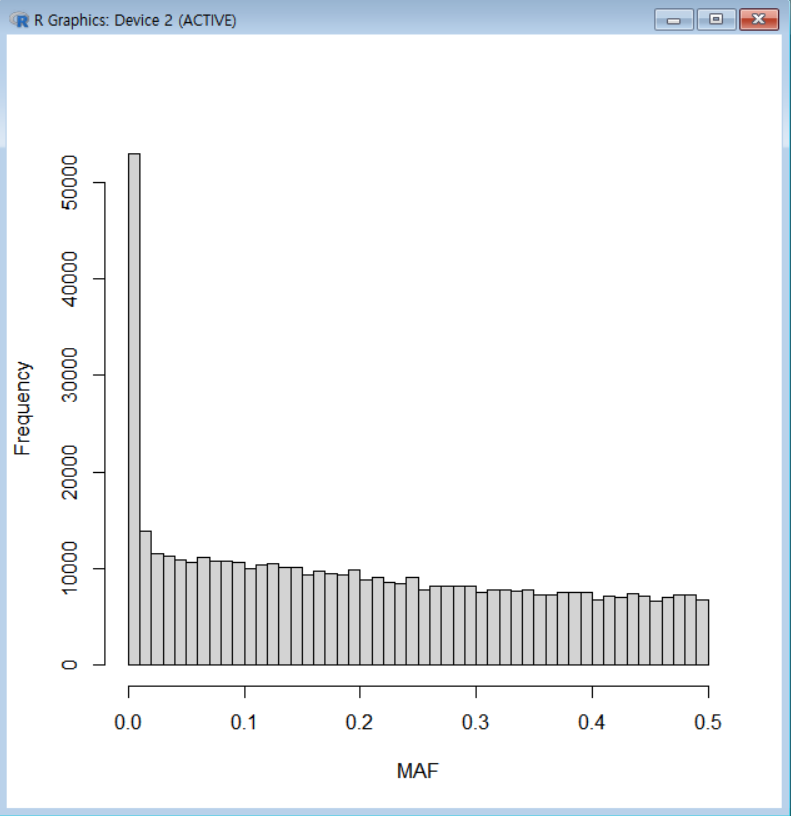
awk "$5<=0.00001" case-control.missing > case-control.missing.drop

(2) maf와 hwe를 통해서 불필요한 데이터를 제거한다.

MAF 정보를 얻는다.

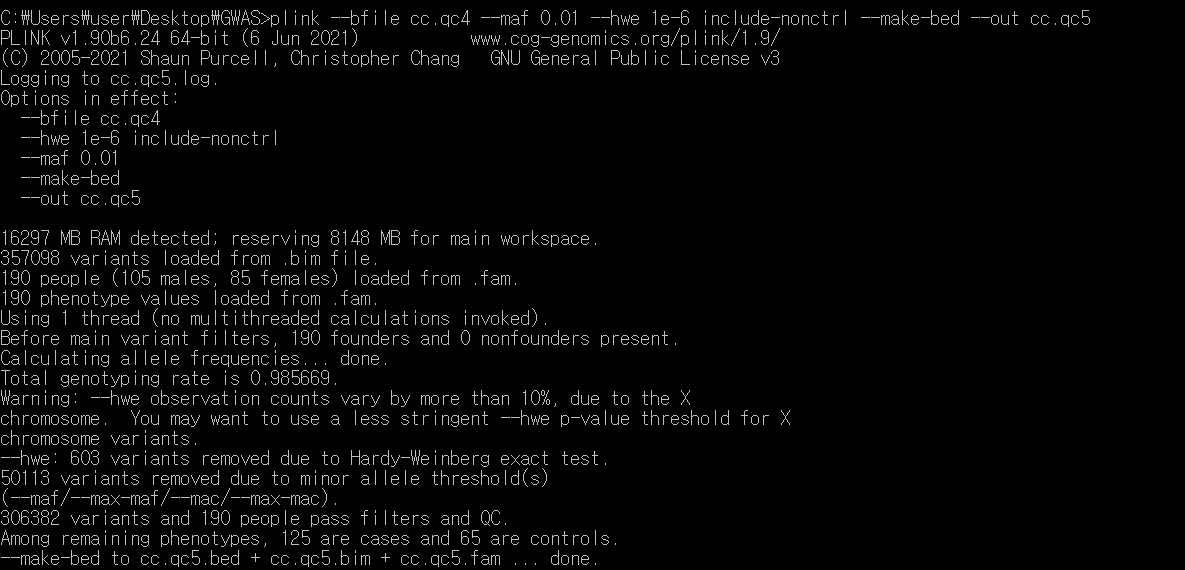
plink --bfile cc.qc3 --freq --out maf





minor allele frequency가 1% 이하로 나타나는 low frequency나 rare한 유전변이들을 제거하고 p-value가 0.000001 이하로 평형이 깨진 경우를 제거하기 위해 다음과 같은 명령어를 사용한다,

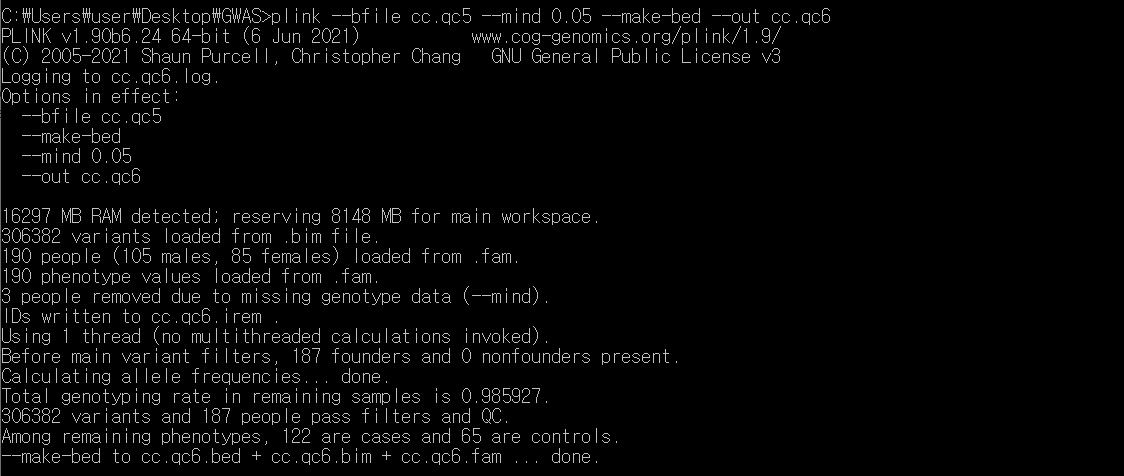
plink --bfile cc.qc4 --maf 0.01 --hwe 1e-6 include-nonctrl --make-bed --out cc.qc5



hwe를 통해 603개의 SNPs들이 제거되었다. 그리고 maf를 통해 50113개의 SNPs들이 제거되었다. 따라서 306382개의 SNPs과 190명의 사람이 남아 정보가 저장된다.

(3) mind를 통해 genotype의 missing rate이 5% 이상이 되는 데이터들을 제거한다.

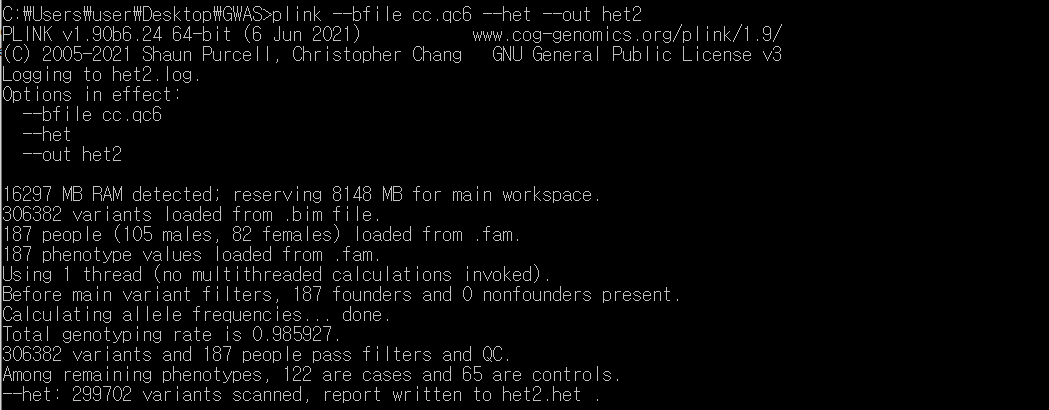
plink --bfile cc.qc5 --mind 0.05 --make-bed --out cc.qc6



mind를 통해 3명이 제거된 것을 확인할 수 있다. 따라서 306382개의 SNPs과 187명의 사람이 남아서 정보가 남게 된다.

이제 데이터를 정리했으므로 heterozygosity 점수를 구해본다.

plink --bfile cc.qc6 --het --out het2



echo "FID IID obs\_HOM N\_SNPs prop\_HET" > het2.txt

awk “NR>1{print $1,$2,$3,$5,($5-$3)/$5}” het2.het >> het2.txt

awk “NR>1{sum+=$5;sq+=$5^2}END{avg=sum/(NR-1);print avg-3\*(sqrt(sq/(NR-2)-2\*avg\*(sum/(NR-2))+(((NR-1)\*(avg^2))/(NR-2)))),avg+3\*(sqrt(sq/(NR-2)-2\*avg\*(sum/(NR-2))+(((NR-1)\*(avg^2))/(NR-2))))}” het2.txt

awk “$5<=0.29957 || $5>= 0.326526” het2.txt > het.drop

위와 같은 명령어를 통해서 het2.txt 파일을 생성하고 불필요한 데이터를 제거하기 위한 het.drop파일을 생성한다.

plink --bfile cc.qc6 --remove het.drop --make-bed --out cc.clean

명령어를 통해 het.drop에 있는 정보를 가진 데이터들을 삭제한다.

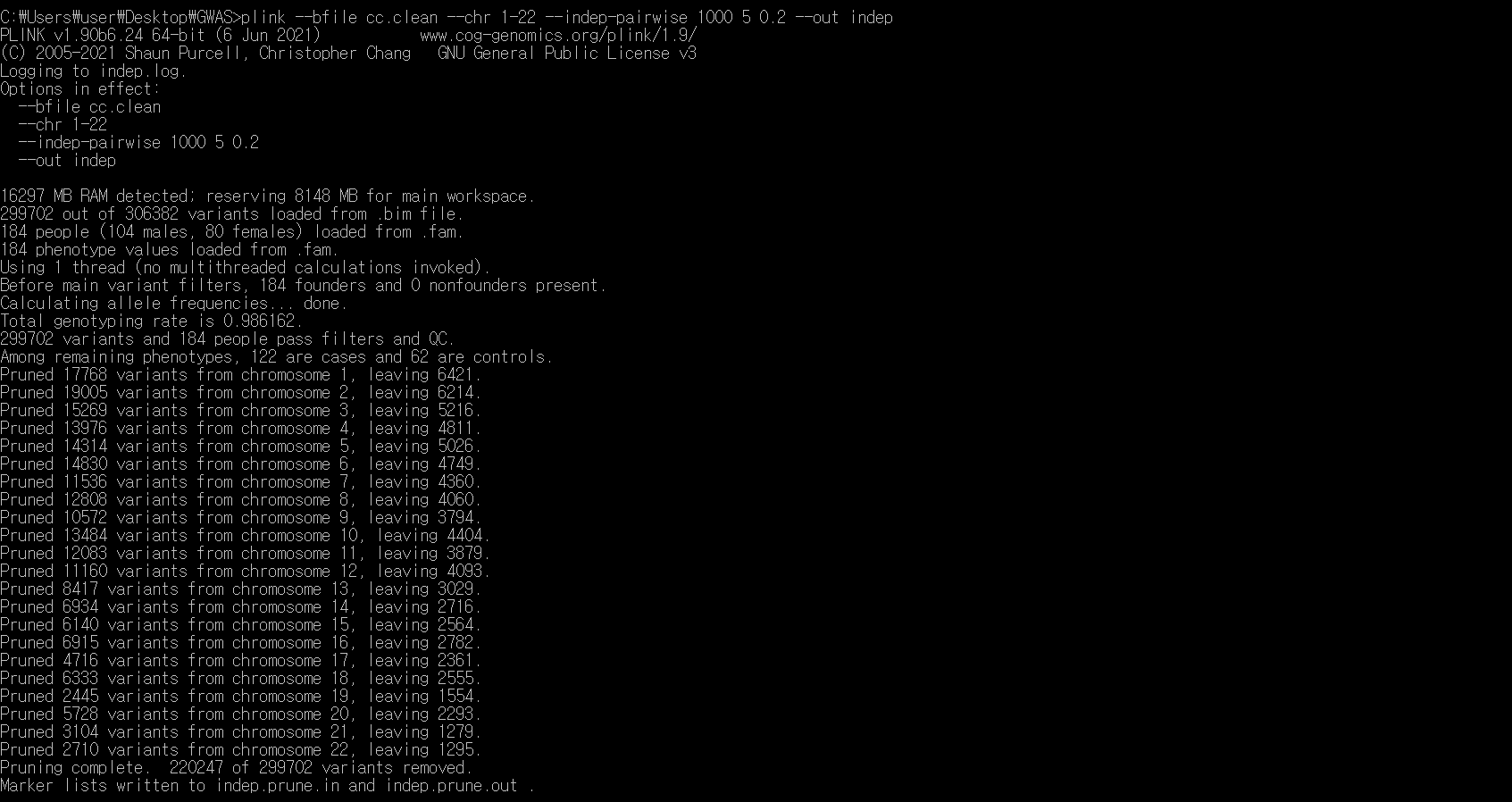


3명이 제거되어 306382개의 SNPs과 184명의 사람의 정보가 남게 된다.

**데이터 관계 알아보기**

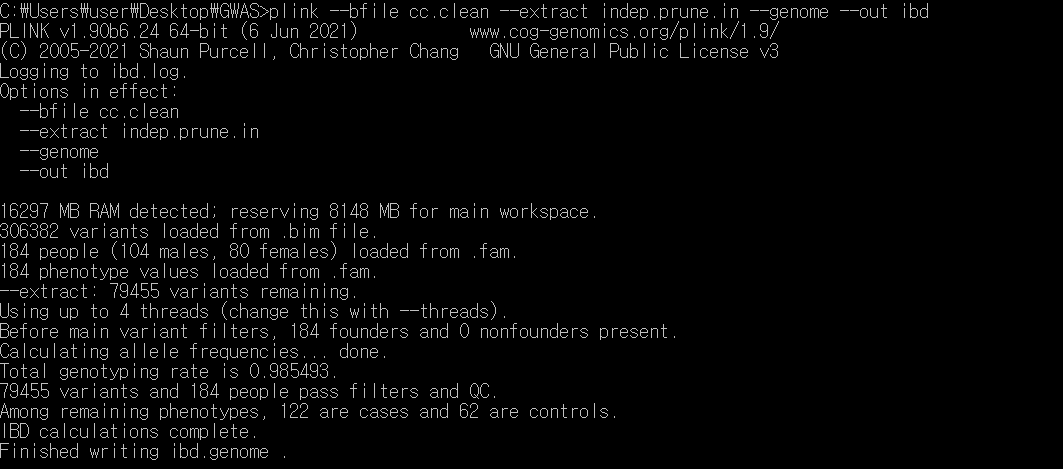
먼저 SNPs를 잘라내고 chr 1-22를 선택한다. 독립적으로 정리된 SNPs 파일이 생성된다.

plink --bfile cc.clean --chr 1-22 --indep-pairwise 1000 5 0.2 --out indep

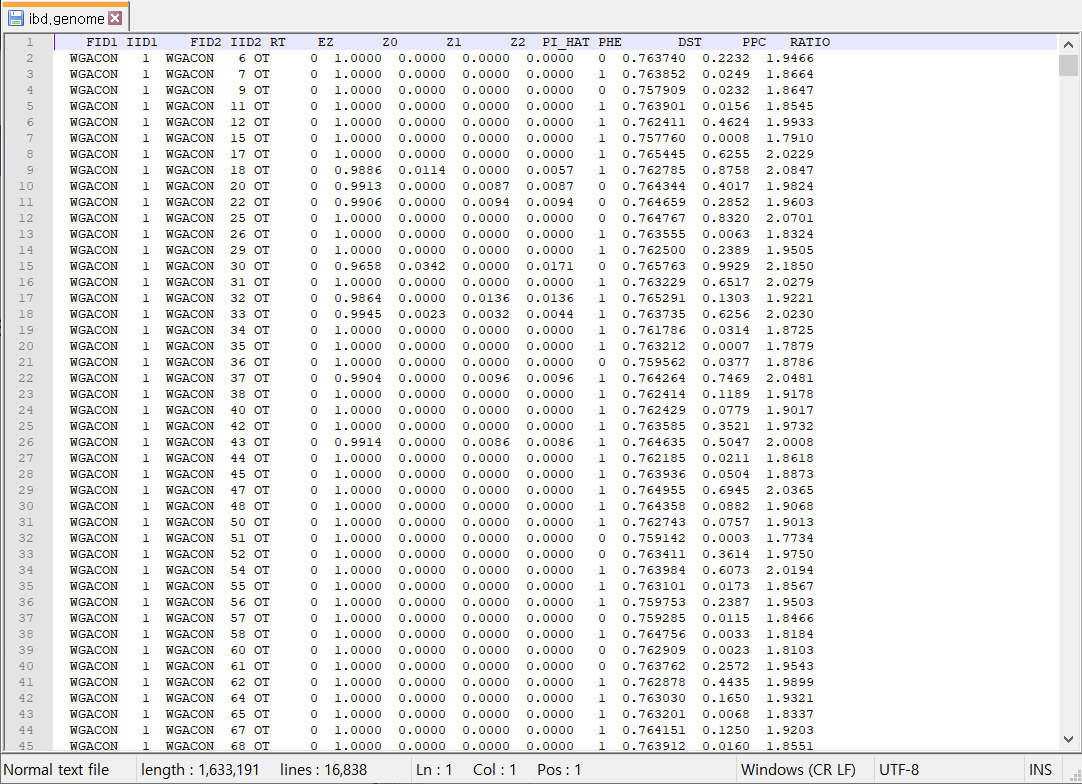


제거된 SNPs에서 IBD에 의한 식별성을 계산한다.

plink --bfile cc.clean --extract indep.prune.in --genome --out ibd



여기서 생성된 ibd.genome 파일을 확인한다.



<여기서 Z0 Z1 Z2의 의미는 다음과 같다.>

1, 0, 0 = 관계없음

0, 1, 0 = 부모-자식 관계

.25, .5, .25 = 형제 등

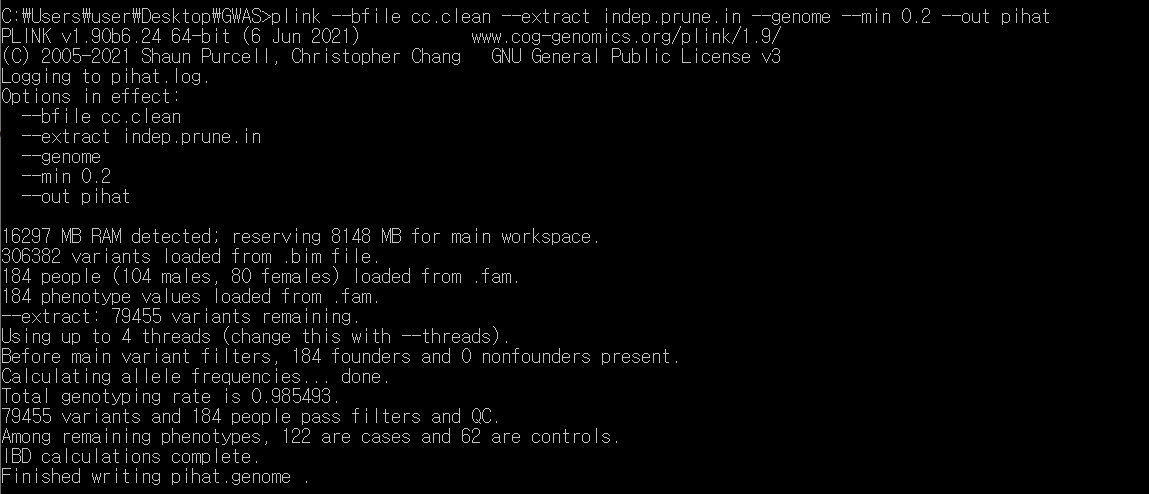
<PI\_HAT의 의미는 IBD 비율과 같고 이것은 P(IBD=2) + 0.5\*P(IBD=1)과 같다.>

0 = 관계없음

.5 = 부모-자식 관계 혹은 형제관계

.125 = 사촌관계

PI\_HAT이 0.2이상인 관계에 대한 개인들의 목록을 구한다.



**사용데이터**

<https://ibg.colorado.edu/cdrom2019/colodro_grasby/GWAS_QC_part2/GWAS_QC_part2_practical/>에서 데이터를 얻었다. 이 데이터는 358명의 관계없는 참가자들의 유전정보가 담겨있다.

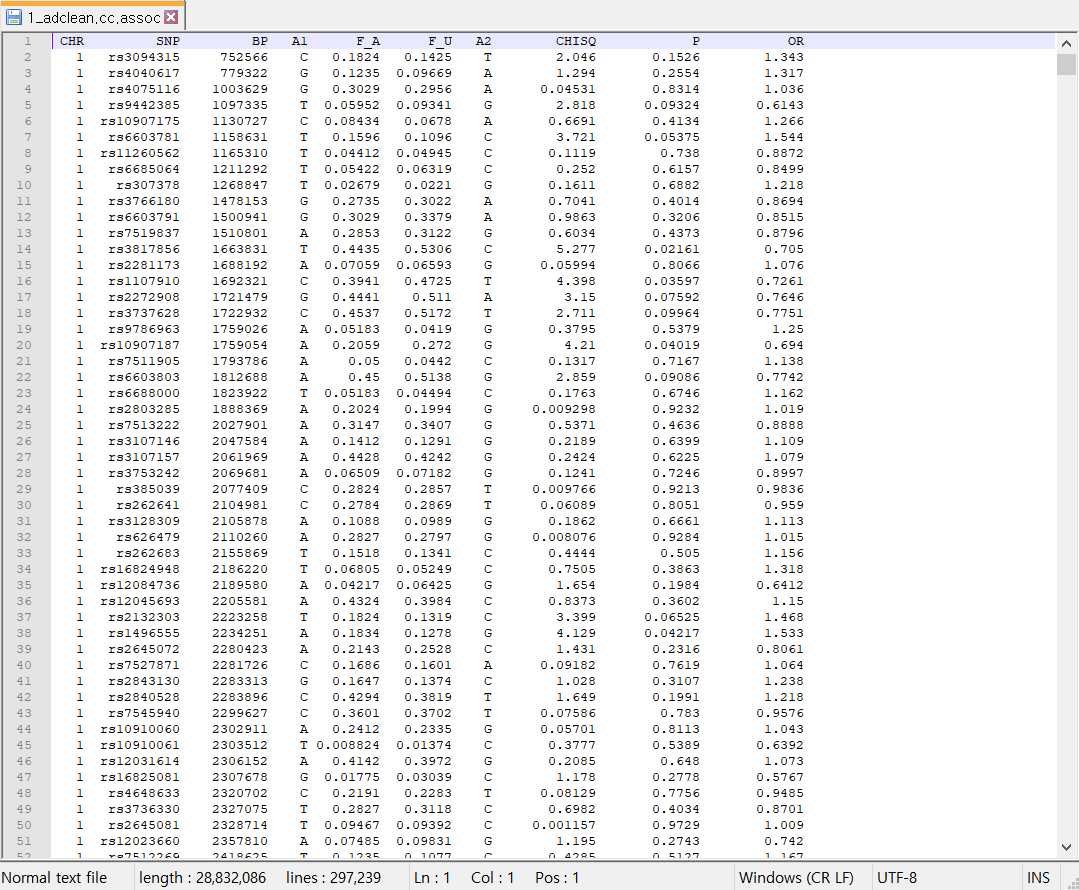
**연관성 테스트**

Case-Control 변수 AD에 대한 연관성 분석을 실시한다.

plink --bfile adclean.cc --assoc --out 1\_adclean.cc 명령어를 통해 실행한다.



170명은 case군, 182명은 control군이다.

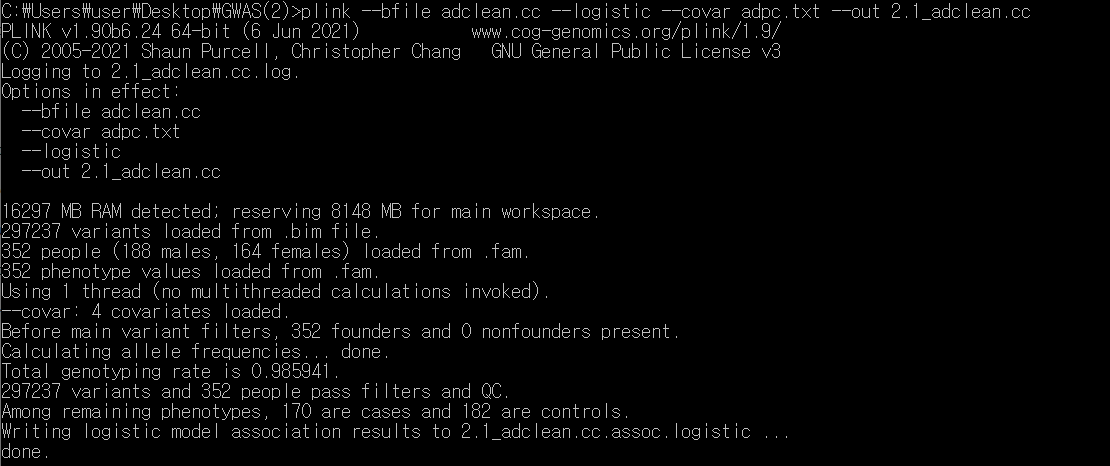


297237개의 SNPs과 352명의 유전정보에 대해서 연관성분석을 실시한 파일 1\_adclean.cc.assoc이 생성된다.

**Logistic regression 분석**

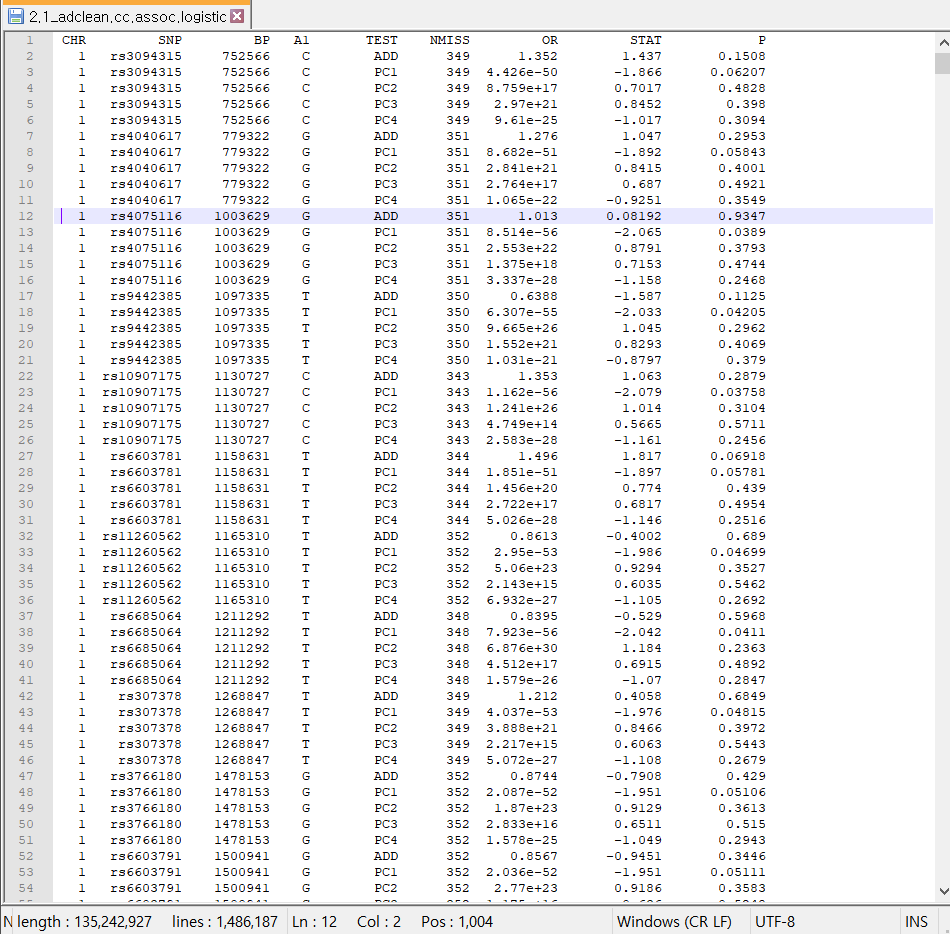
공변량 파일 adpc.txt파일을 사용해서 case-control 변수 AD에 대한 Logistic regression 분석을 실행한다.

plink --bfile adclean.cc --logistic --covar adpc.txt --out 2.1\_adclean.cc



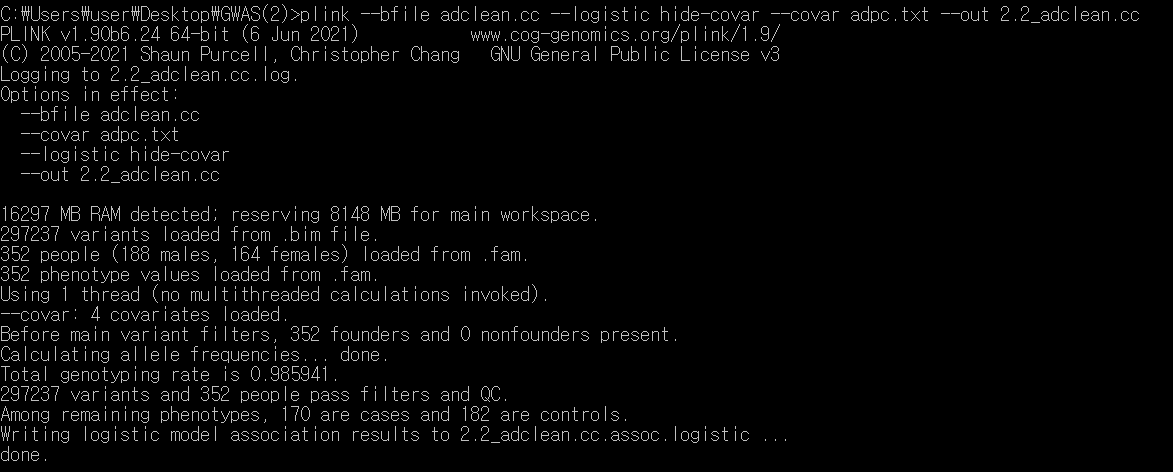
4개의 공변량이 load되고 Logistic regression 분석을 실시한 것을 확인할 수 있다.

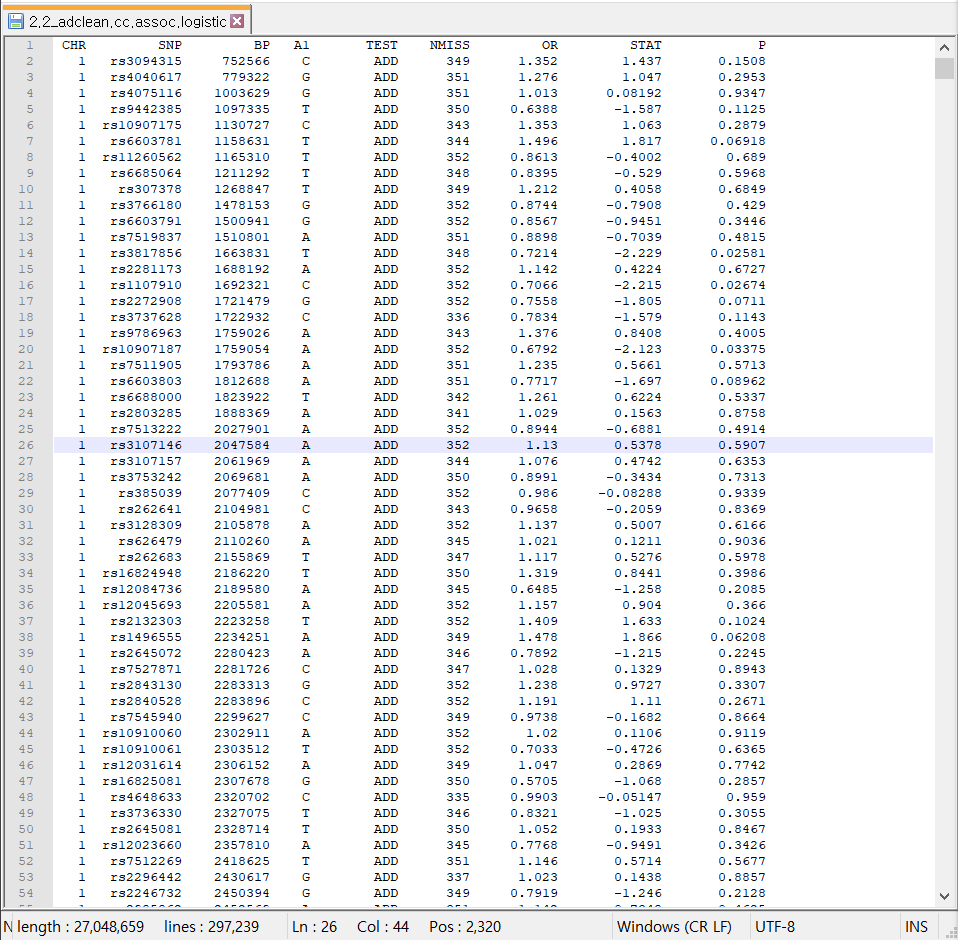
결과물은 다음과 같다.



공변량의 결과를 숨기고 공변량을 주성분으로 포함하는 case-control 변수 AD에 대한 Logistic regression 분석을 실행한다.

plink --bfile adclean.cc --logistic hide-covar --covar adpc.txt --out 2.2\_adclean.cc





QQ plot에 필요한 열 CHR, BP, P-Value를 가져온다. 헤더와 데이터 누락 없이 파일을 읽어온다.

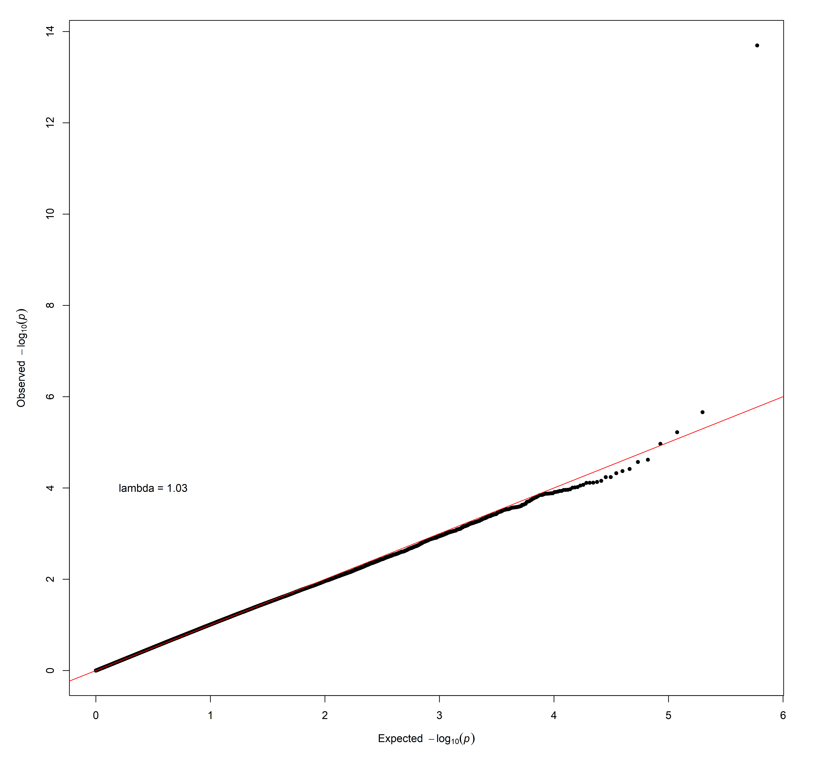
명령어: awk "{if (NR>1) print $1,$3,$9}" 2.2\_adclean.cc.assoc.logistic | grep -v NA > plot.adclean.cc.logistic.txt

LD링크에 19번 염색체 데이터를 업로드해서 LD 패턴을 탐색한다. CHR, BP, P-Value가 포함되어 있는 파일을 준비한다.

명령어: awk "{if (NR==1 || $1==19) print $1,$3,$2,$9}" 2.2\_adclean.cc.assoc.logistic | grep -v NA > ld19.adclean.cc.logistic.txt

가장 낮은 P-Value를 가진 SNPs의 번호를 찾는다.

명령어: sort -k4 -r ld19.adclean.cc.logistic.txt | head

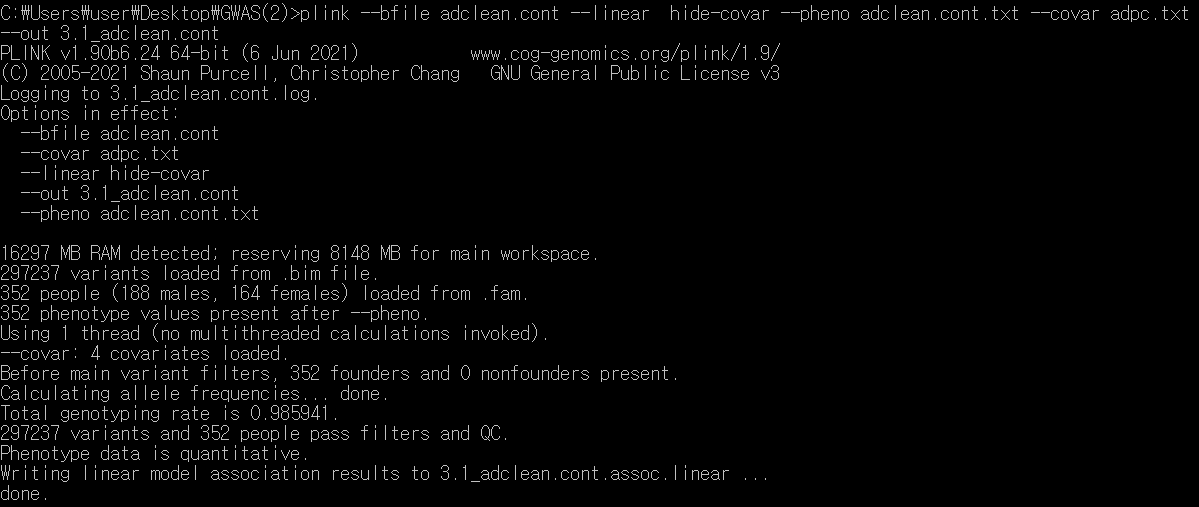


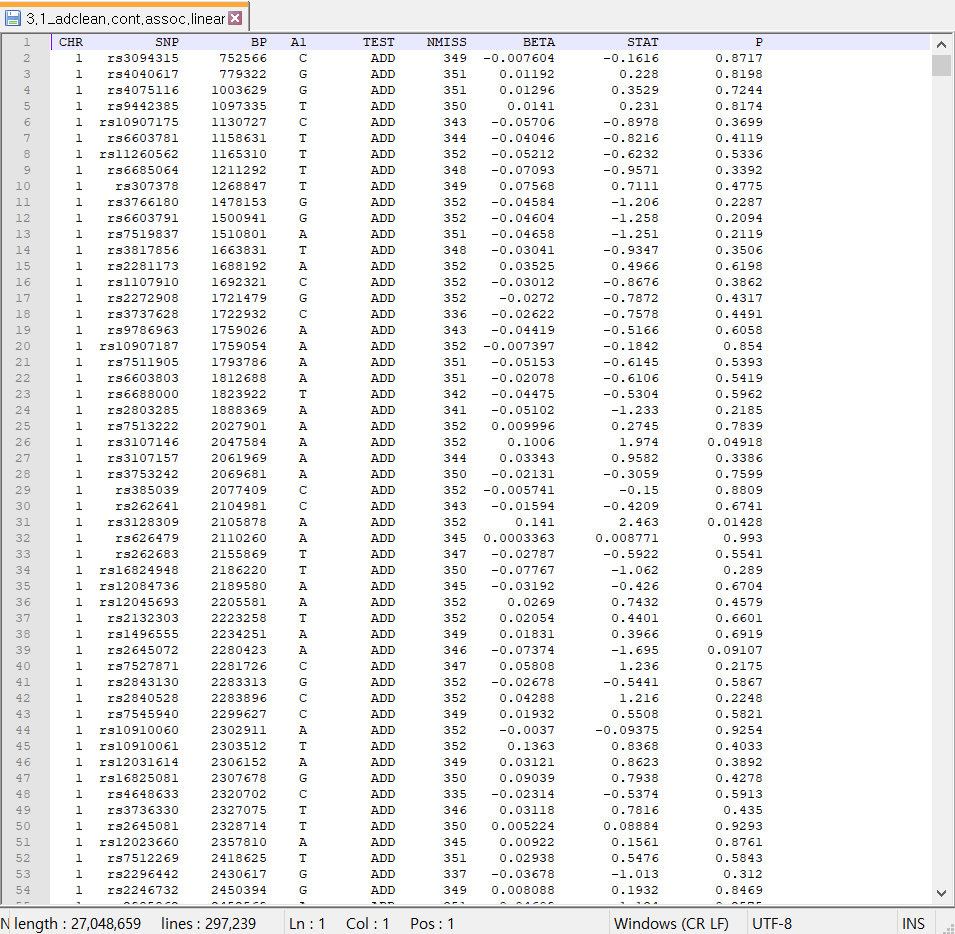
공변량 PC1-PC4를 이용해 Logistic regression 분석한 결과를 QQ plot으로 나타낸 그래프이다.

**Linear regression 분석**

공변량의 결과를 숨기고 pheno 옵션을 사용해서 주성분을 공변량으로 포함하는 연석형 특성에 대한 Linear regression 분석을 실행한다.

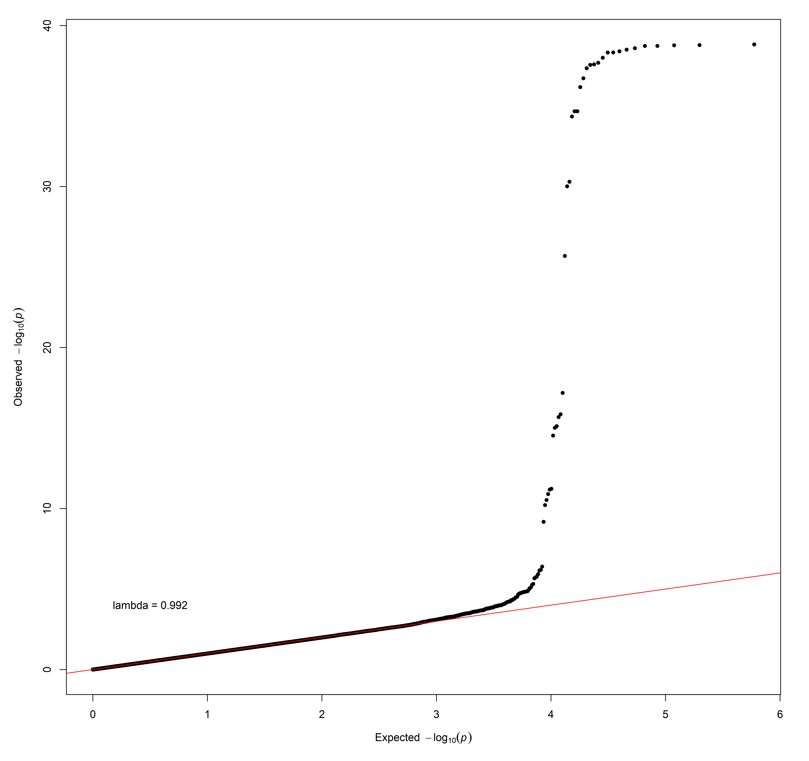
명령어: plink --bfile adclean.cont --linear hide-covar --pheno adclean.cont.txt --covar adpc.txt --out 3.1\_adclean.cont





QQ plot에 필요한 열 CHR, BP, P-Value를 가져온다. 헤더와 데이터 누락 없이 파일을 읽어온다.

명령어: awk "{if (NR>1) print $1,$3,$9}" 3.1\_adclean.cont.assoc.linear | grep -v NA > plot.adclean.cont.linear.txt



공변량 PC1-PC4를 이용해 Linear regression 분석한 결과를 QQ plot으로 나타낸 그래프이다.

LD링크에 20번 염색체 데이터를 업로드해서 LD 패턴을 탐색한다. CHR, BP, P-Value가 포함되어 있는 파일을 준비한다.

명령어: awk "{if (NR==1 || $1==20) print $1,$3,$2,$9}" 3.1\_adclean.cont.assoc.linear | grep -v NA > ld20.adclean.cont.linear.txt

가장 낮은 P-Value를 가진 SNPs의 번호를 찾는다.

명령어: sort -k4 -r ld20.adclean.cont.linear | head

공변량의 결과를 숨기고 --pheno 옵션을 사용하여 공변량으로 PC1만 포함하는 연속 특성에 대한 Linear regression 분석을 실행합니다.

명령어: plink --bfile adclean.cont --linear hide-covar --pheno adclean.cont.txt --covar adpc.txt --covar-name PC1 --out 3.2\_adclean.cont

