**개별연구 최종보고서**

**유전체 데이터 분석 및 결과 해석**

**담당교수: 주종화 교수님**

**학과: 컴퓨터공학과**

**학번: 2017112138**

**학년: 3학년**

**이름: 정여준**

사용데이터는 두 개를 구해서 진행하였다. 193명의 관계없는 참가자의 유전정보가 담긴 데이터와

358명의 관계없는 참가자들의 유전정보와 공변량 데이터가 담긴 데이터를 구하였다.

데이터의 출처는 순서대로 나열하면 다음과 같다.

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

<https://ibg.colorado.edu/cdrom2019/colodro_grasby/GWAS_QC_part2/GWAS_QC_part2_practical/>

일단 첫번째 데이터로는 성별분석을 실행하였다. 성별분석을 통해서 나온 데이터들 중 오류가 있는 데이터를 삭제하는 과정을 실행하였다. 그 이후에 문제가 pseudo-autosomal region 때문인지 확인하기 위해서 split-x를 사용해서 확인하는 과정을 진행했다. 문제가 있는 것들은 분석에서 제외시킨다. 이제 남은 데이터들의 heterozygosity 정보를 얻고 각 사람에 대한 heterozygosity의 비율로 파일을 만든다. 그리고 위에서 만든 비율과 누락에 대한 파일을 모두 포함하는 파일을 생성한다. 그 이후 데이터에서 질이 낮거나 양이 적은 데이터를 제거한다. geno를 통해 너무 많은 데이터가 누락된 데이터를 제거한다. 그 이후 maf와 hwe 명령어를 통해서 불필요한 데이터들을 제거한다. 이후 mind 명령어를 통해 genotype missing rate이 5% 이상이 되는 데이터들을 제거한다. 이렇게 데이터를 정리했으면 heterozygosity 점수를 구한다. 그리고 불필요한 데이터를 다시 제거해준다. 이후에는 데이터간 관계를 알아보는 과정을 진행했다. SNPs를 잘라내고 염색체번호를 선택한다. 독립적으로 정리된 SNPs 파일을 생성한다. 제거된 SNPs에서 IBD에 의한 식별성을 계산한다. 그렇게 나온 것을 확인하면 관계를 확인할 수 있다.

두번째 데이터로는 첫번째 연관성 테스트를 진행했다. Case-Control 변수 AD에 대한 연관성 분석을 실시해서 Case군 수와 Control군 수를 파악한다. 그 이후에는 Logistic regression 분석을 실행한다. 제공받은 공변량 파일을 사용해서 Case-Control 변수 AD에 대한 Logistic regression을 실행한다. 이후 공변량의 결과를 숨기고 공변량을 주성분으로 포함하는 Case-Control 변수 AD에 대한 Logistic regression 분석을 실행한다. 이후 QQ plot에 필요한 열 Chr, BP, P-Value를 가져오고 헤더와 데이터 누락 없이 파일을 읽어온다. 그리고 공변량 PC1-PC4를 이용해 Logistic regression 분석한 결과를 QQ plot으로 나타낸 그래프를 구한다. 다음으로 Linear regression 분석을 실시한다. 공변량의 결과를 숨기고 pheno 옵션을 사용해서 주성분을 공변량으로 포함하는 연속형 특성에 대한 Linear regression 분석을 실행한다. 그 이후 QQ plot에 필요한 열 Chr, BP, P-Value를 가져온다. 그 이후 헤더와 데이터 누락 없이 파일을 읽어온다. 그리고 공변량 PC1-PC4를 이용해 Linear regression 분석한 결과를 QQ plot으로 나타낸 그래프를 구한다.