BICF Nanocourse

Population Genetics Workshop

1 Contents

- Quality control of germline mutations
- Detect population stratification
- Run association test

2 Tools

- Linux – we will use some basic commands to process our dataset.
- PLINK - Whole genome association analysis toolset
  Website: https://www.cog-genomics.org/plink2

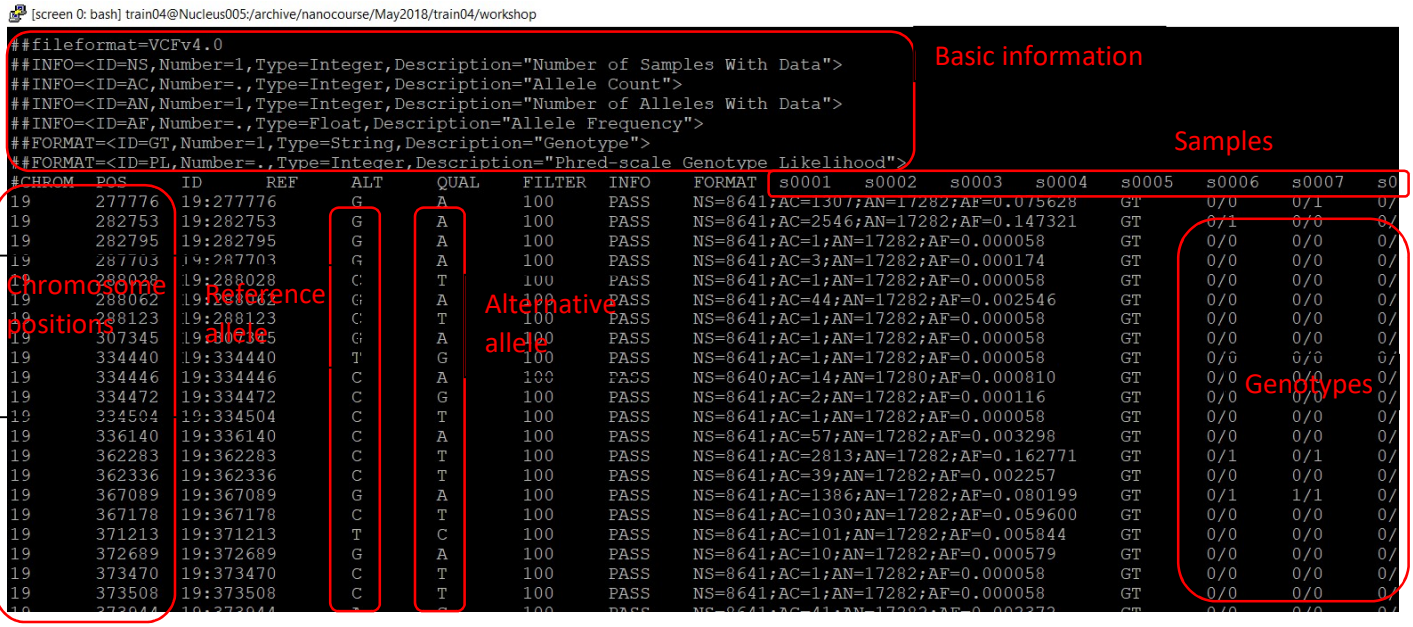
3 Prepare the environment and data

```
ln -s /archive/nanocourse/May2018/shared/train04/data/* .
export PATH=/archive/nanocourse/May2018/shared/train04/bin/:$PATH
module load R/3.4.1-gccmkl
alias ll='ls -lhtr'
```

4 Quality control

4.1 Check the vcf file

```
zcat gws.vcf.gz | less –S
```



4.2 Format the file to plink format

```
plink --vcf gws.vcf.gz --recode --out gws
less -S gws.ped
```

Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype, Genotypes ….

```
s0001 s0001 0 0 0 -9 G G A G G G G G C C G G C C G G T T C C C C C C C C T C
s0002 s0002 0 0 0 -9 A G G G G G G G C C G G C C G G T T C C C C C C C C T C
s0003 s0003 0 0 0 -9 G G G G G G G G C C G G C C G G T T C C C C C C C C T C
s0004 s0004 0 0 0 -9 G G G G G G G G C C G G C C G G T T C C C C C C C C T C
s0005 s0005 0 0 0 -9 G G A G G G G G C C G G C C G G T T C C C C C C C C C C
```

```
less -S gws.map
```

Chromosome, SNP ID, Genetic distance (morgans), Base-pair position (bp units)

```
19      19:277776       0       277776
19      19:282753       0       282753
19      19:282795       0       282795
19      19:287703       0       287703
19      19:288028       0       288028
19      19:288062       0       288062
```

4.3 Check minor allele frequency (MAF)

```
plink --file gws --freq --out gws

sort -k5,5g gws.frq | less
```

```
 CHR         SNP   A1   A2          MAF  NCHROBS
  19  19:1004741    0    G            0     3954
  19  19:1005173    0    G            0     3952
  19  19:1005337    0    G            0     3954
  19  19:1005388    0    C            0     3954
  19 19:10084884    0    G            0     3954
  19 19:10091365    0    C            0     3954
```

4.4 Keep only the SNPs with MAF > 0.01

```
plink --file gws --maf 0.01 --recode --out gws.maf_0.01

plink --file gws.maf_0.01 --freq --out gws.maf_0.01

sort -k5,5g gws.maf_0.01.frq | less
```

```
 CHR         SNP   A1   A2          MAF  NCHROBS
  19 19:37241167    T    C      0.01012     3954
  19 19:41884327    C    T      0.01012     3954
  19 19:44470147    A    G      0.01012     3954
  19 19:48807255    T    C      0.01012     3954
  19 19:49090527    A    G      0.01012     3954
  19 19:53269678    A    T      0.01012     3954
  19 19:53854713    A    G      0.01012     3954
  19 19:55365378    A    G      0.01012     3954
  19   19:757887    A    G      0.01012     3954
```

4.5 Calculate the p-value of hardy weinberg equilibrium

```
plink --file gws.maf_0.01 --hardy --out gws.maf_0.01

sort -k9,9g gws.maf_0.01.hwe | less
```

```
 CHR         SNP     TEST   A1   A2                 GENO   O(HET)   E(HET)            P
  19 19:45448465  ALL(NP)    T    G      335/1047/595   0.5296   0.4914     0.000595
  19 19:18569411  ALL(NP)    A    G        4/53/1920  0.02681  0.03038    0.0009592
  19 19:54414637  ALL(NP)    C    T       65/696/1216    0.352   0.3305     0.003427
  19 19:17025292  ALL(NP)    G    A       251/826/900   0.4178   0.4461     0.004773
  19 19:37383253  ALL(NP)    G    T        9/145/1823  0.07334  0.07905     0.005084
  19 19:33600764  ALL(NP)    T    C      372/1040/565    0.526   0.4952     0.006417
  19 19:48602948  ALL(NP)    T    C      378/1039/560   0.5255   0.4958     0.008491
  19 19:44118188  ALL(NP)    C    A      123/652/1202   0.3298   0.3511     0.008515
  19 19:44118353  ALL(NP)    A    G      123/652/1202   0.3298   0.3511     0.008515
```

A1A1 / A1A2 / A2A2

4.6 Keep only the SNPs with p-value(HWE) > 0.001

```
plink --file gws.maf_0.01 --hwe 0.001 --recode --out gws.maf_0.01.hwe_0.001
```

5 Population stratification

5.1 Combine GWS and 1000G

```
plink --file gws.maf_0.01.hwe_0.001 --merge 1000g --recode --out gws.maf_0.01.hwe_0.001.1000g
```

5.2 Calculate principal components of GWS and 1000 Genomes

```
plink --file gws.maf_0.01.hwe_0.001.1000g --pca --out gws.maf_0.01.hwe_0.001.1000g

less -S gws.maf_0.01.hwe_0.001.1000g.eigenvec
```

FID IID PC1 PC2 PC3 ….

```
HG00096 HG00096 0.024578 -0.0338613 0.002025 -0.00252931 0.0185213 0.0336488 0.006572
HG00097 HG00097 0.0233558 -0.0402901 0.00233885 -0.00937903 -0.00606553 -0.00801432 0
HG00099 HG00099 0.0241395 -0.0350457 0.00179585 -0.0103106 0.00191366 -7.95147e-05 0.
HG00100 HG00100 0.0246163 -0.0323874 -0.0245086 0.0128275 0.0111517 -0.0284391 -0.023
HG00101 HG00101 0.028928 -0.0277637 -0.0230846 -0.0118824 -0.00350972 -0.000614455 -0
HG00102 HG00102 0.0295604 -0.0281759 -0.0240386 0.0216821 0.0190777 0.0229553 0.01959
HG00103 HG00103 0.0167108 -0.0364489 0.0272802 0.04 -0.00457986 -0.0136702 0.0132983
```

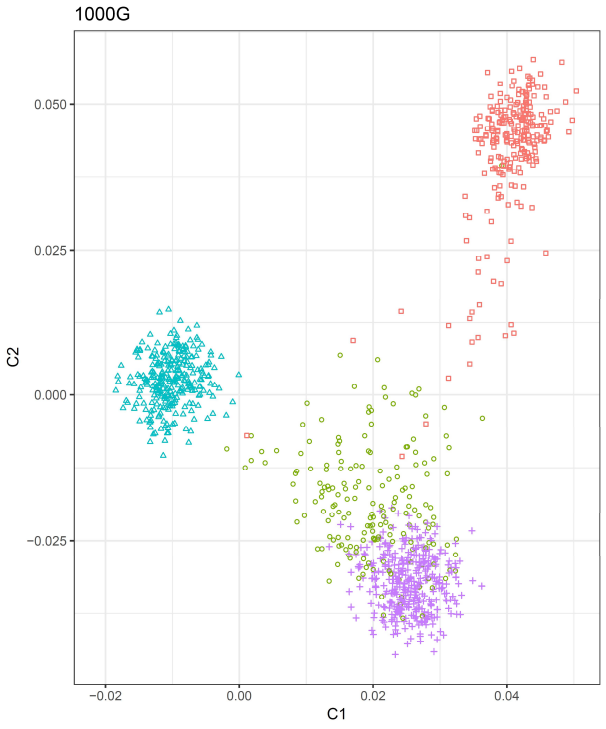
5.3 Combine the PCs and population groups

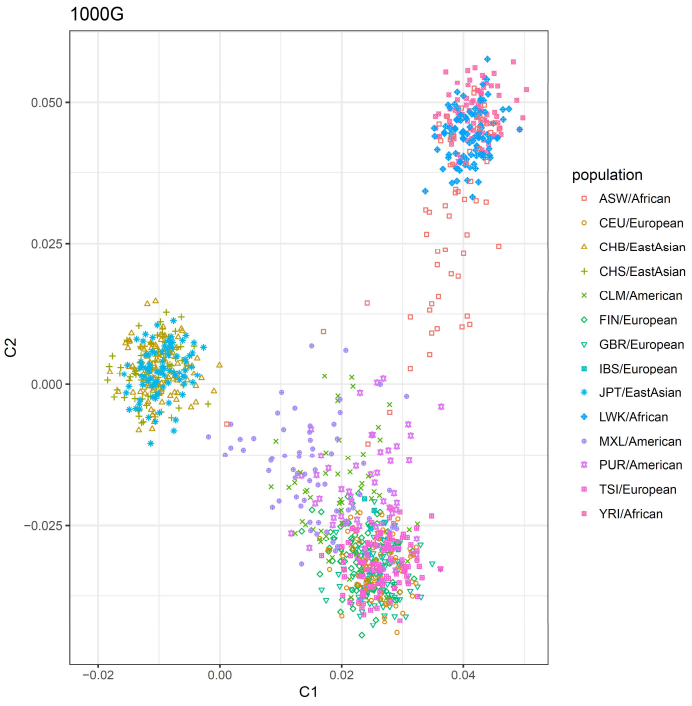
```
sort gws.maf_0.01.hwe_0.001.1000g.eigenvec > gws.maf_0.01.hwe_0.001.1000g.eigenvec.sort

paste -d' ' gws.1000g.pop gws.maf_0.01.hwe_0.001.1000g.eigenvec.sort | cut -f1,2,3,6- -d' ' >
gws.maf_0.01.hwe_0.001.1000g.eigenvec.sort.pop
```

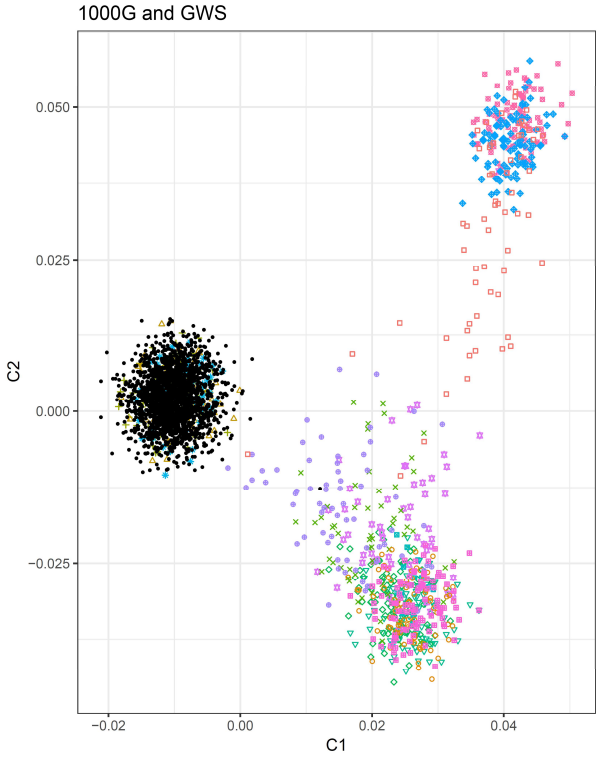
```
HG00096 GBR/European European 0.024578 -0.0338613 0.002025 -0.00252931 0.0185213 0.0336488 0.0
HG00097 GBR/European European 0.0233558 -0.0402901 0.00233885 -0.00937903 -0.00606553 -0.00801
HG00099 GBR/European European 0.0241395 -0.0350457 0.00179585 -0.0103106 0.00191366 -7.95147e-
HG00100 GBR/European European 0.0246163 -0.0323874 -0.0245086 0.0128275 0.0111517 -0.0284391 -
HG00101 GBR/European European 0.028928 -0.0277637 -0.0230846 -0.0118824 -0.00350972 -0.0006144
HG00102 GBR/European European 0.0295604 -0.0281759 -0.0240386 0.0216821 0.0190777 0.0229553 0.
HG00103 GBR/European European 0.0167108 -0.0364489 0.0272802 0.04 -0.00457986 -0.0136702 0.013
HG00104 GBR/European European 0.0245496 -0.0344032 0.00267052 0.0231587 0.00427451 0.010256 0.
```

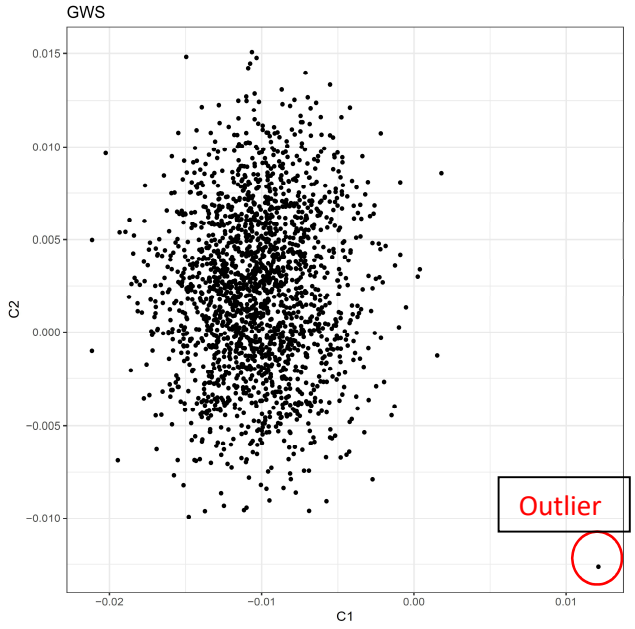
5.4 Plot the PCs for 1000G, GWS and 1000G+GWS

```
./plot_pc.sh gws.maf_0.01.hwe_0.001.1000g.eigenvec.sort.pop
gws.maf_0.01.hwe_0.001.1000g.eigenvec.sort.pop.pdf
```


1000G
0.050
0.025
0.000
−0.025
C2
−0.02
0.00
0.02
0.04
C1
super
African
American
EastAsian
European


1000G
0.050
0.025
0.000
−0.025
C2
−0.02
0.00
0.02
0.04
C1
population
ASW/African
CEU/European
CHB/EastAsian
CHS/EastAsian
CLM/American
FIN/European
GBR/European
IBS/European
JPT/EastAsian
LWK/African
MXL/American
PUR/American
TSI/European
YRI/African


1000G and GWS
0.050
0.025
0.000
−0.025
C2
−0.02
0.00
0.02
0.04
C1
population
ASW/African
CEU/European
CHB/EastAsian
CHS/EastAsian
CLM/American
FIN/European
GBR/European
IBS/European
JPT/EastAsian
LWK/African
MXL/American
PUR/American
TSI/European
YRI/African


GWS
0.015
0.010
0.005
0.000
−0.005
−0.010
C2
−0.02
−0.01
0.00
0.01
C1
Outlier

5.5 Exclude PC outliers

```
cat gws.maf_0.01.hwe_0.001.1000g.eigenvec.sort.pop | awk '{if($2=="GWS" && $3>0.01)print $1,$1}' > gws.maf_0.01.hwe_0.001.1000g.eigenvec.outlier

plink --file gws.maf_0.01.hwe_0.001 --remove gws.maf_0.01.hwe_0.001.1000g.eigenvec.outlier --recode --out gws.maf_0.01.hwe_0.001.no_pc_outlier
```

5.6 We finished QC

```
ln -s gws.maf_0.01.hwe_0.001.no_pc_outlier.ped gws.pass.ped

ln -s gws.maf_0.01.hwe_0.001.no_pc_outlier.map gws.pass.map
```

6 Association test

6.1 Check phenotype file

```
head pheno.pass.ldl
```

FID IID Low-density lipoprotein (LDL) cholesterol

```
FID IID ldl
s0001 s0001 2.9
s0002 s0002 2.2
s0003 s0003 2.5
s0004 s0004 1.9
s0005 s0005 2.08
```

```
head pheno.pass.age_sex
```

```
FID IID age sex
s0001 s0001 54 2
s0002 s0002 61 2
s0003 s0003 65 1
s0004 s0004 54 2
s0005 s0005 58 1
```

6.2 Re-generate PCs for QC passed data

```
plink --file gws.pass --pca --out gws.pass
```

6.3 Generate covariate file (age + sex + pc1 + pc2)

```
cat gws.pass.eigenvec | cut -f1-4 -d' ' > gws.pass.pc2

echo "FID IID pc1 pc2" > gws.pass.pc2.fmt

cat gws.pass.pc2 >> gws.pass.pc2.fmt

head gws.pass.pc2.fmt
```

```
FID IID pc1 pc2
s0001 s0001 -0.0296586 0.00356785
s0002 s0002 0.00165701 -0.014572
s0003 s0003 0.00212168 -0.0266324
s0004 s0004 -0.0283681 0.0097966
s0005 s0005 0.0355808 0.0139361
```

```
paste -d' ' pheno.pass.age_sex gws.pass.pc2.fmt | cut -f1-4,7,8 -d' ' > pheno.pass.age_sex_pc2

head pheno.pass.age_sex_pc2
```

```
FID IID age sex pc1 pc2
s0001 s0001 54 2 -0.0296586 0.00356785
s0002 s0002 61 2 0.00165701 -0.014572
s0003 s0003 65 1 0.00212168 -0.0266324
s0004 s0004 54 2 -0.0283681 0.0097966
s0005 s0005 58 1 0.0355808 0.0139361
```

6.4 LDL association test (linear regression)

Model: LDL ~ Genotype + Age + Sex + PC1 + PC2

```
plink --file gws.pass --allow-no-sex --pheno pheno.pass.ldl --covar pheno.pass.age_sex_pc2 --linear --out gws.pass.ldl
```

```
less gws.pass.ldl.assoc.linear
```

CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
19	19:277776	277776	A	ADD	1976	0.01533	0.2751	0.7832
19	19:277776	277776	A	age	1976	-0.002162	-1.042	0.2977
19	19:277776	277776	A	sex	1976	0.3331	8.27	2.429e-16
19	19:277776	277776	A	pc1	1976	0.2833	0.3189	0.7498
19	19:277776	277776	A	pc2	1976	1.747	1.966	0.04944
19	19:282753	282753	A	ADD	1976	-0.01881	-0.4711	0.6376
19	19:282753	282753	A	age	1976	-0.002162	-1.042	0.2976
19	19:282753	282753	A	sex	1976	0.3338	8.287	2.127e-16
19	19:282753	282753	A	pc1	1976	0.2926	0.3294	0.7419
19	19:282753	282753	A	pc2	1976	1.74	1.958	0.05041
19	19:362283	362283	T	ADD	1976	0.0004421	0.01144	0.9909
19	19:362283	362283	T	age	1976	-0.002173	-1.047	0.2952
19	19:362283	362283	T	sex	1976	0.3334	8.277	2.305e-16
19	19:362283	362283	T	pc1	1976	0.286	0.3216	0.7478
19	19:362283	362283	T	pc2	1976	1.743	1.961	0.04997
19	19:367089	367089	A	ADD	1976	-0.005858	-0.1068	0.915

```
cat gws.pass.ldl.assoc.linear | awk '{if(($5=="ADD" || $5=="TEST") && $9!="NA")print}' | sort -k9,9g > gws.pass.ldl.assoc.linear.res
```

```
head gws.pass.ldl.assoc.linear.res
```

CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
19	19:45415640	45415640	A	ADD	1976	-0.3125	-6.004	2.286e-09
19	19:52034506	52034506	T	ADD	1976	-0.2337	-3.145	0.001686
19	19:11257018	11257018	T	ADD	1976	-0.1009	-3.139	0.001718
19	19:39876735	39876735	A	ADD	1976	0.4303	3.072	0.002153
19	19:4236996	4236996	A	ADD	1976	0.08456	2.967	0.003042
19	19:15990431	15990431	T	ADD	1976	-0.09055	-2.878	0.004048
19	19:52000672	52000672	A	ADD	1976	0.07987	2.821	0.004837
19	19:14083761	14083761	C	ADD	1976	-0.07842	-2.763	0.00578
19	19:16060117	16060117	A	ADD	1974	-0.1213	-2.742	0.006159

6.5 QQ plot

```
./plot_qq.sh gws.pass.ldl.assoc.linear.res gws.pass.ldl.assoc.linear.res.qqplot.pdf
```