Population Stratification Practical

Overview

- Use the SNP data to empirically investigate the ancestry of the samples
- Assign individuals to groups
- Capture Chinese/Japanese distinction?
- Generate a multi-dimensional scaling plot

1. Create Independent Set of SNPs

- clustering/MDS/PCA/etc... work best on sets of SNPs that are independent

- plink option
  - --indep-pairwise <*N*> <*M*> <*r2*>
    - N = windows size
    - M = window shift
    - r2 = threshold for removal

Filter SNPs based on LD

```
plink –-bfile wgas3
      --indep-pairwise 50 10 0.2
      --out prune1
```

Output: prune1.pruned.in

```
rs4040617
rs4075116
rs9442385
rs11260562
rs6685064
rs7519837
rs3855951
....
```

Output: run1.pruned.out

```
rs3094315
rs3766180
rs6603791
rs3737628
rs7511905
rs3107157
...
....
```

2. Measure IBS between Individual

- IBS = Identity-by-State
- Measure of total genotype similarity
- plink commands
  - --genome
  - --extract prune1.prune.in *or*

--exclude prune1.prune.out

Measure IBS between all samples

```
plink –-bfile wgas3
      --extract prune1.prune.in
      --genome
      --out ibs1
```

Output: ibs1.genome

FID1	IID1	FID2	IID2	RT	EZ	Z0	Z1	Z2	PI_HAT	PHE	DST	PPC	RATIO
CH18526	NA18526	CH18524	NA18524	UN	NA	0.9855	0.01205	0.00244	0.008465	-1	0.775486	0.2186	1.94
CH18526	NA18526	CH18529	NA18529	UN	NA	0.9836	0	0.01641	0.01641	-1	0.776529	0.5053	2.001
CH18526	NA18526	CH18558	NA18558	UN	NA	0.9874	0.01255	0	0.006277	-1	0.774003	0.3914	1.978
CH18526	NA18526	CH18532	NA18532	UN	NA	0.9631	0.03686	0	0.01843	-1	0.773503	0.432	1.987
CH18526	NA18526	CH18561	NA18561	UN	NA	0.9721	0.02785	0	0.01393	-1	0.774229	0.9224	2.117
CH18526	NA18526	CH18562	NA18562	UN	NA	0.9927	0.007251	0	0.003626	-1	0.773934	0.4424	1.989
CH18526	NA18526	CH18537	NA18537	UN	NA	0.9912	0.008208	0.0006042	0.004708	0	0.774734	0.6533	2.031
CH18526	NA18526	CH18603	NA18603	UN	NA	0.9949	0.001806	0.003256	0.004159	0	0.774771	0.4686	1.994

...

ibs1.genome columns

- RT = relationship from pedigree
- EZ = expected IBD sharing
- PI_HAT = observed IBD sharing

  = P(IBD=2)+0.5*P(IBD=1)
- PHE = pairwise phenotypic code

(1,0,-1 = AA, AU and UU pairs)

- RATIO = HETHET : IBS 0 SNPs (expected 2)

(HETHET = Number of IBS2 het/het SNP pairs)

- PPC = p-value of deviation of RATIO from 2

Clustering

- Use the IBS values to perform clustering

- plink commands
  - --read-genome <*genome file*>
  - --cluster
  - --mds-plot <*N*>

- Add restrictions
  - PPC test (--ppc 1e-3)
  - Clusters contain 1 case and 1 control (--cc)

Clustering

```
plink –-bfile wgas3
      --read-genome ibs1.genome
      --cluster
      --ppc 1e-3
      --cc
      --mds-plot 2
      --out strat1
```

Output: strat1.genome

```
CH18526 NA18526    0
CH18524 NA18524    0
CH18529 NA18529    0
CH18558 NA18558    1
CH18532 NA18532    1
CH18561 NA18561    0
....
JA18987 NA18987    3
JA18990 NA18990    2
JA18991 NA18991    2
JA18994 NA18994    2
JA18992 NA18992    4
JA18997 NA18997    2
...
```

Clustering

```
plink –-bfile wgas3
      --read-genome ibs1.genome
      --cluster
      --ppc 1e-3
      --cc
      --mds-plot 2
      --out strat1
```

Output: strat1.mds

FID	IID	SOL	C1	C2
CH18526	NA18526	0	-0.0245884	0.00917367
CH18524	NA18524	0	-0.0242271	-0.0278564
CH18529	NA18529	0	-0.0182157	0.00810387
CH18558	NA18558	1	-0.0233174	-0.0318428
CH18532	NA18532	1	-0.023291	-0.000359224
CH18561	NA18561	0	-0.0219817	-0.00855256
CH18562	NA18562	1	-0.0203434	-0.0379835
CH18537	NA18537	0	-0.0202644	-0.0320186
CH18603	NA18603	1	-0.0247671	0.0113186
CH18540	NA18540	1	-0.0236313	0.00147393
CH18605	NA18605	0	-0.0223527	-0.00800508

...

Make MDS Plot

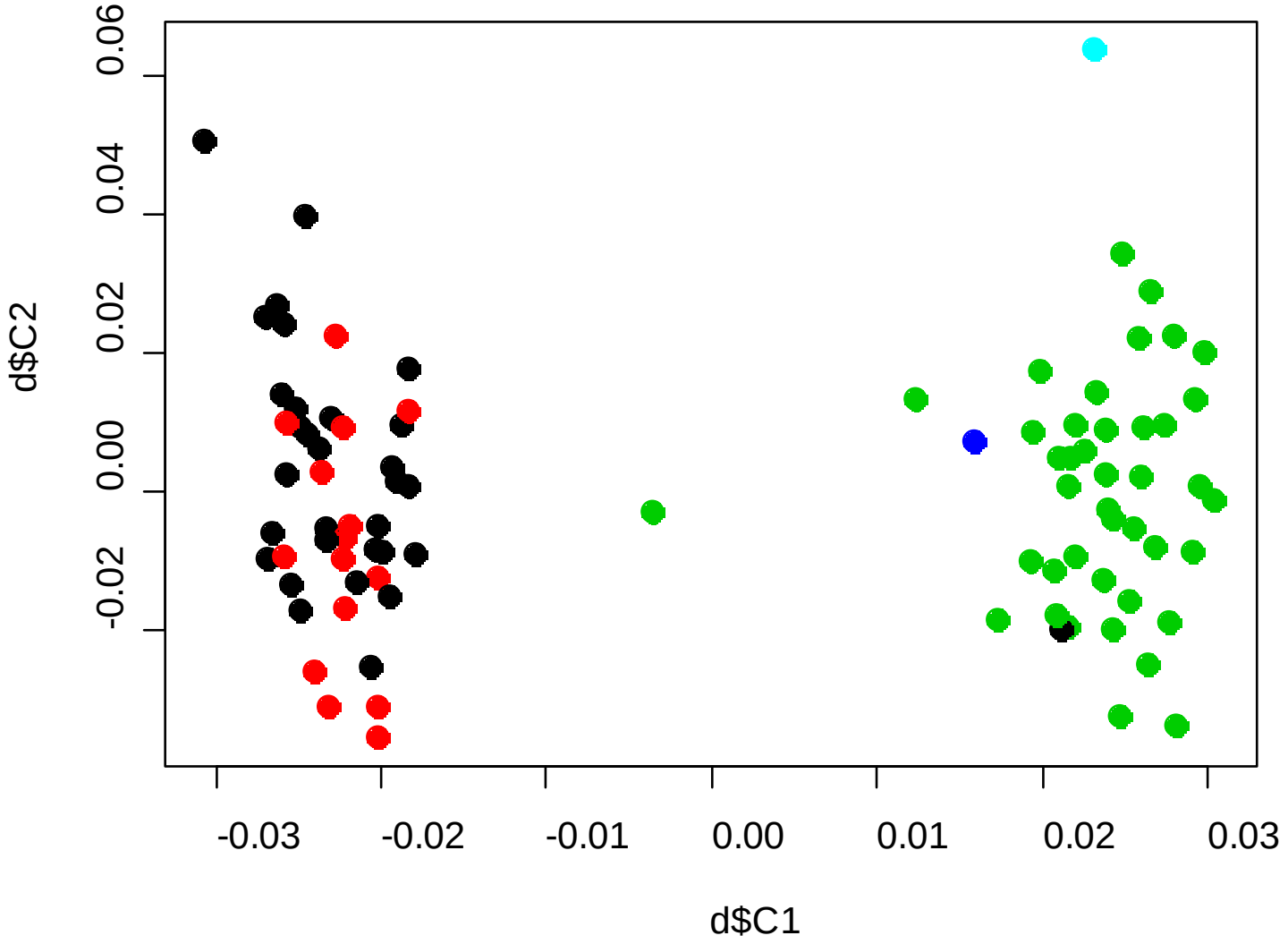
- Open R
- Navigate to your working directory

```
> d <- read.table("strat1.mds", header=TRUE)
> plot( d$C1, d$C2, pch=20, cex=2, col = d$SOL+1)



> pop = numeric(nrow(d))
> pop[which(substr(d[,1],1,1)==”C”)] = 1
> pop[which(substr(d[,1],1,1)==”J”)] = 2
> plot( d$C1, d$C2, pch=20, cex=2, col = pop)
```

MDS Plot



Association using MDS

```
plink –-bfile wgas3
      --mh
      --within strat1.cluster2
      --out mdsaccoc1
```

```
plink –-bfile wgas3
      --logistic
      --covar strat1.mds
      --covar-number 1,2
      --out mdsaccoc2
```