Multi-task group Lasso correcting for population stratification in Genome Wide Association Studies

Asma Nouira

Chloé-Agathe Azencott

Centre for Computational Biology (CBIO)

U900 Lab Meeting

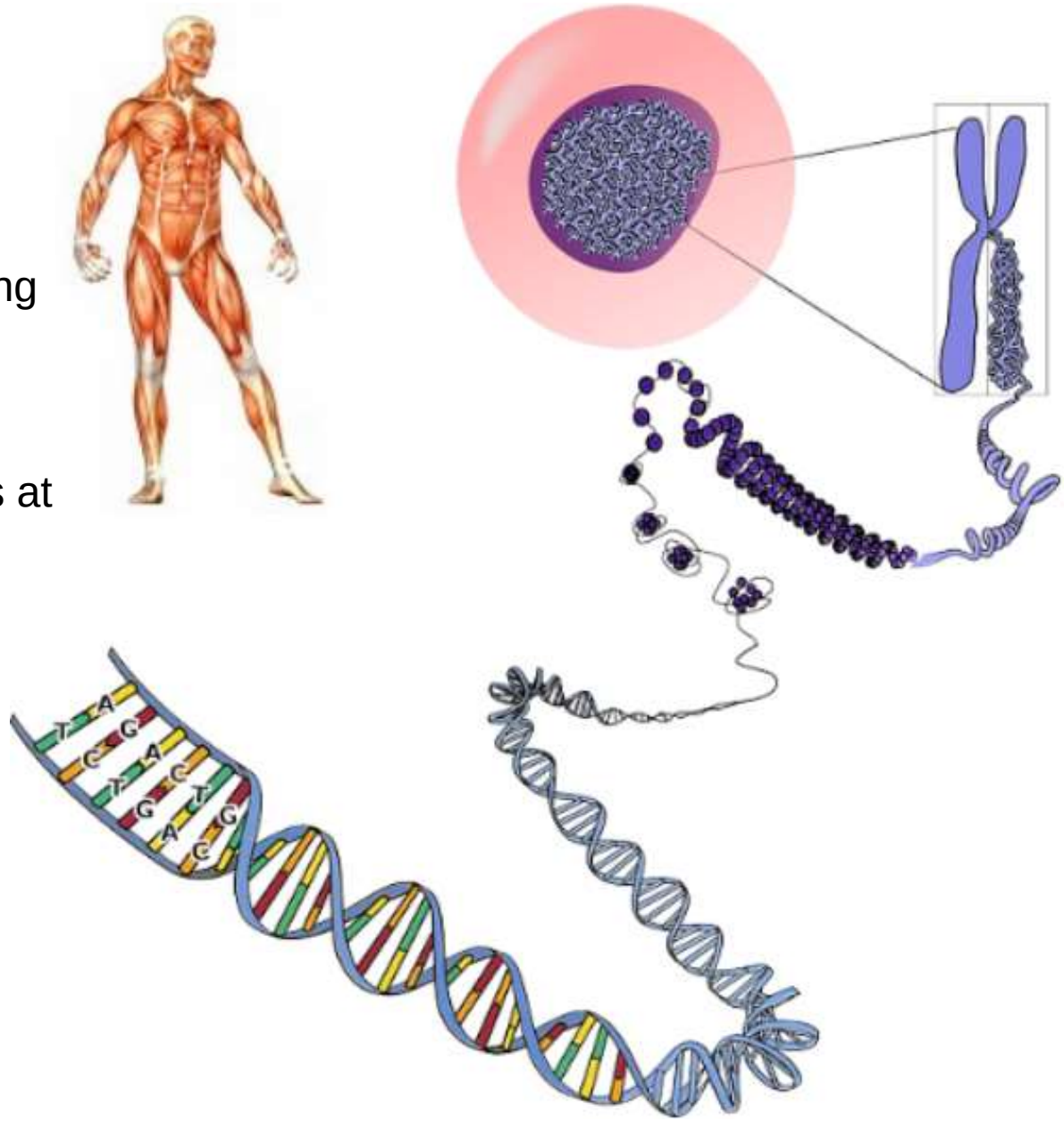
March 11, 2020

- The Human Genome Project (HGP) provides a good mapping to decode the whole genome at Single Nucleotide Variant (SNVs) level.

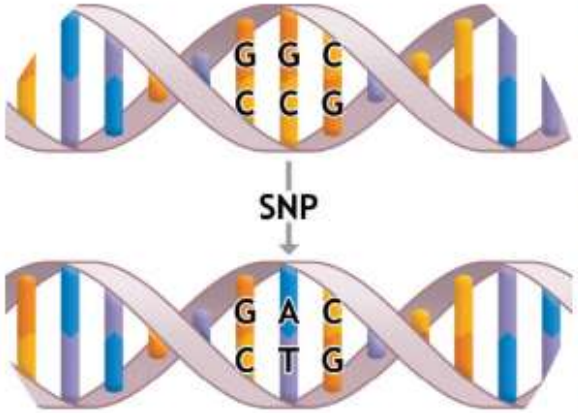
- Single Nucleotide Polymorphism (SNPs) are common SNVs at a frequency of 1%.

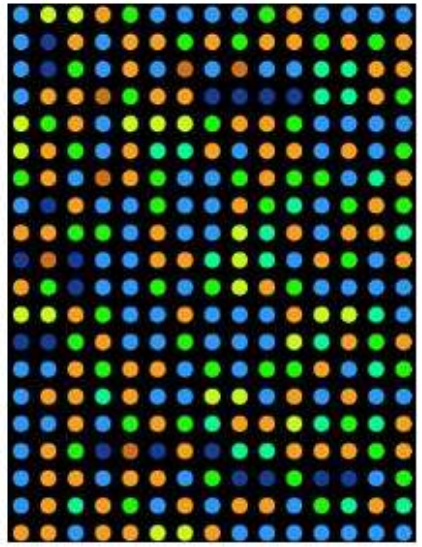
- 3 billions base pair divided in 24 chromosomes.
- 15 millions SNPs.

- Find association between genome and disease risk.



Biomarker discovery





- Microarray data: SNP arrays.
- Case-control studies.

➔ Find association between genotype and phenotype.

1 Breast cancer datasets

GWAS of CIDR Breast Cancer in the African Diaspora

- 3,827 samples x 2,379,855 SNPs
- 1,681 cases and 2,085 controls
- 330 African Barbadian ; 2,073 African American; 1,363 African Nigerian
- **Covariates:** Age group, height, weight, BMI, age of menarche, parity, age of first birth, menopause, age of menopause, alcohol, contraceptive, estrogen rate…

DRIVE Breast Cancer OncoArray Genotypes Distribution set

- 28,281 participants x 528,620 SNPs
- 13,846 cases and 14,435 controls
- Populations: USA – Uganda – Nigeria – Cameroon – Australia – Denmark
- **Covariates:** Age, estrogen rate, study, histological type…

2 Classic GWAS

➢ Preprocessing

Quality control

- MAF < 5%
- HWE-P-Value < 0.0001
- Remove samples with missing case/control criterion
- Sex check

Imputation

- Fill missing SNPs.
- Reference dataset: 1000 Genomes Project (GP) Phase 3
- Exclude SNPs with 10% rate of missing values.

Linkage Disequilibrium (LD) pruning

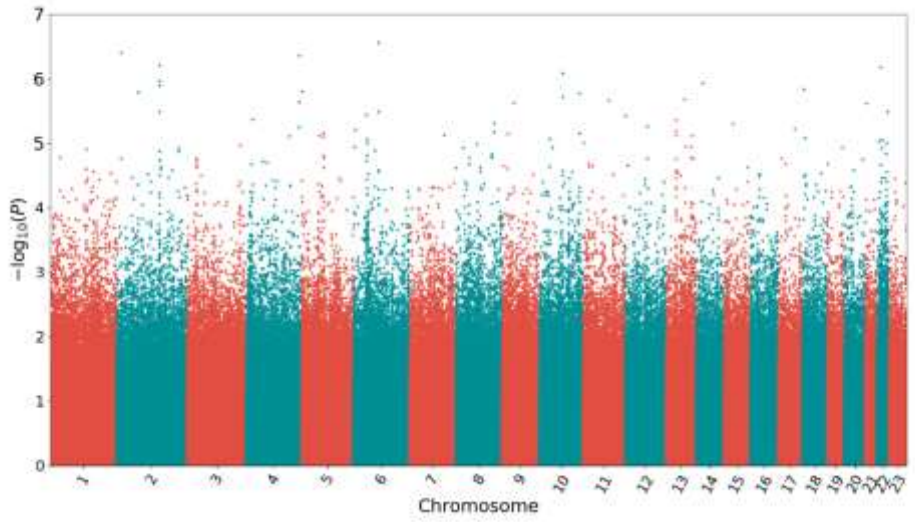
- Consider a window of 50 SNPs
- Calculate LD between each pair of SNPs in the window
- Remove one of a pair of SNPs if the LD is greater than 0.5
- Shift the window 5 SNPs forward

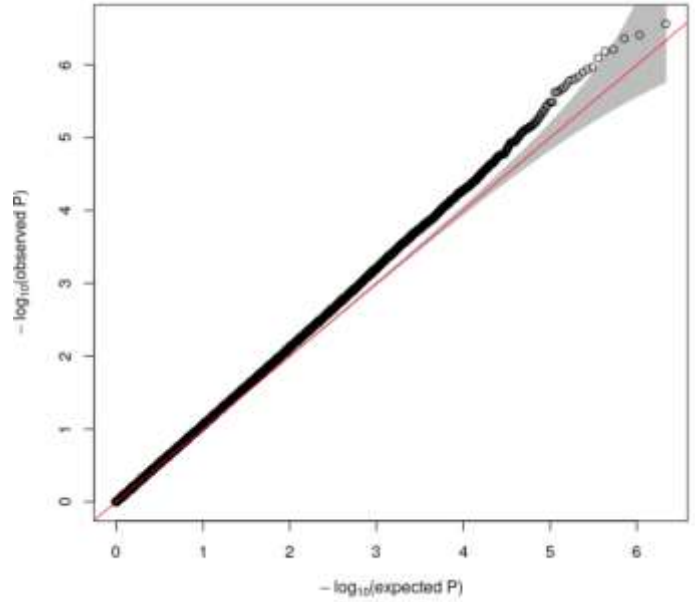
```
--indep-pairwise 50 5 0.5
```

[1] PLINK package: https://www.cog-genomics.org/plink2
[2] https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

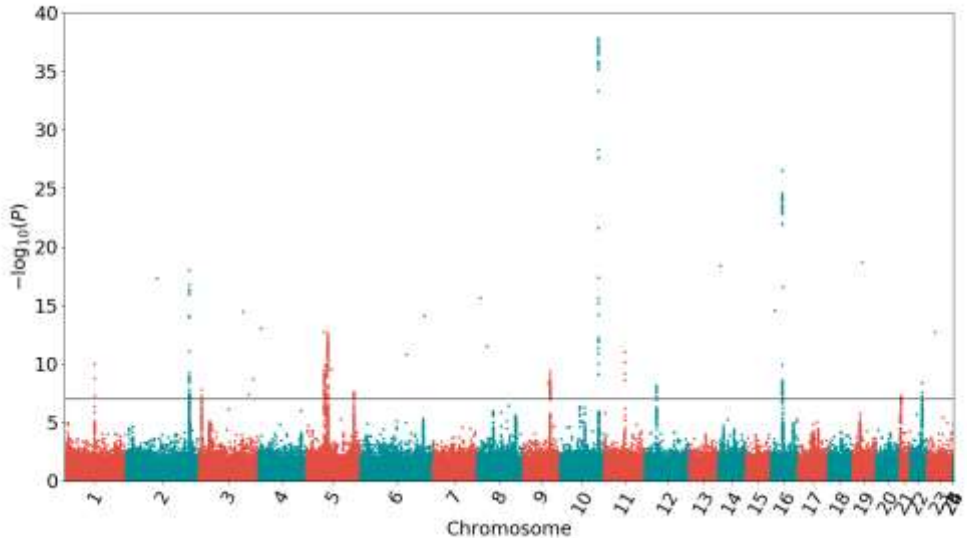
2 Classic GWAS

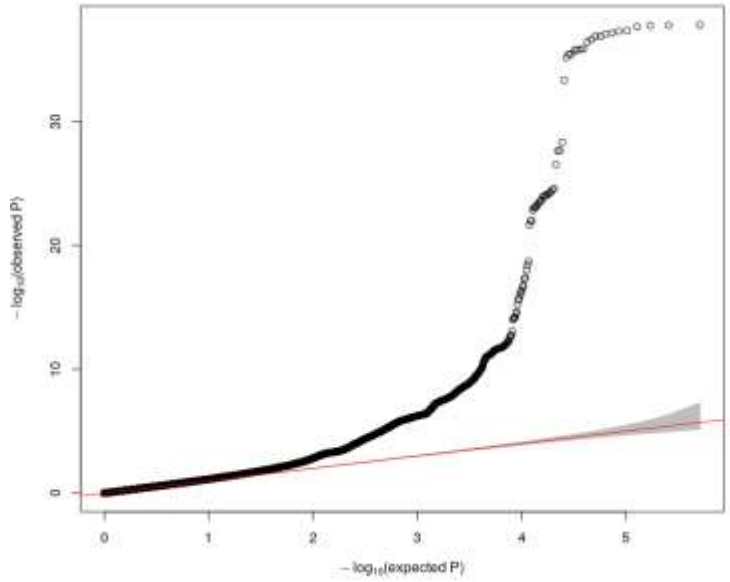
CIDR dataset





DRIVE-OncoArray dataset

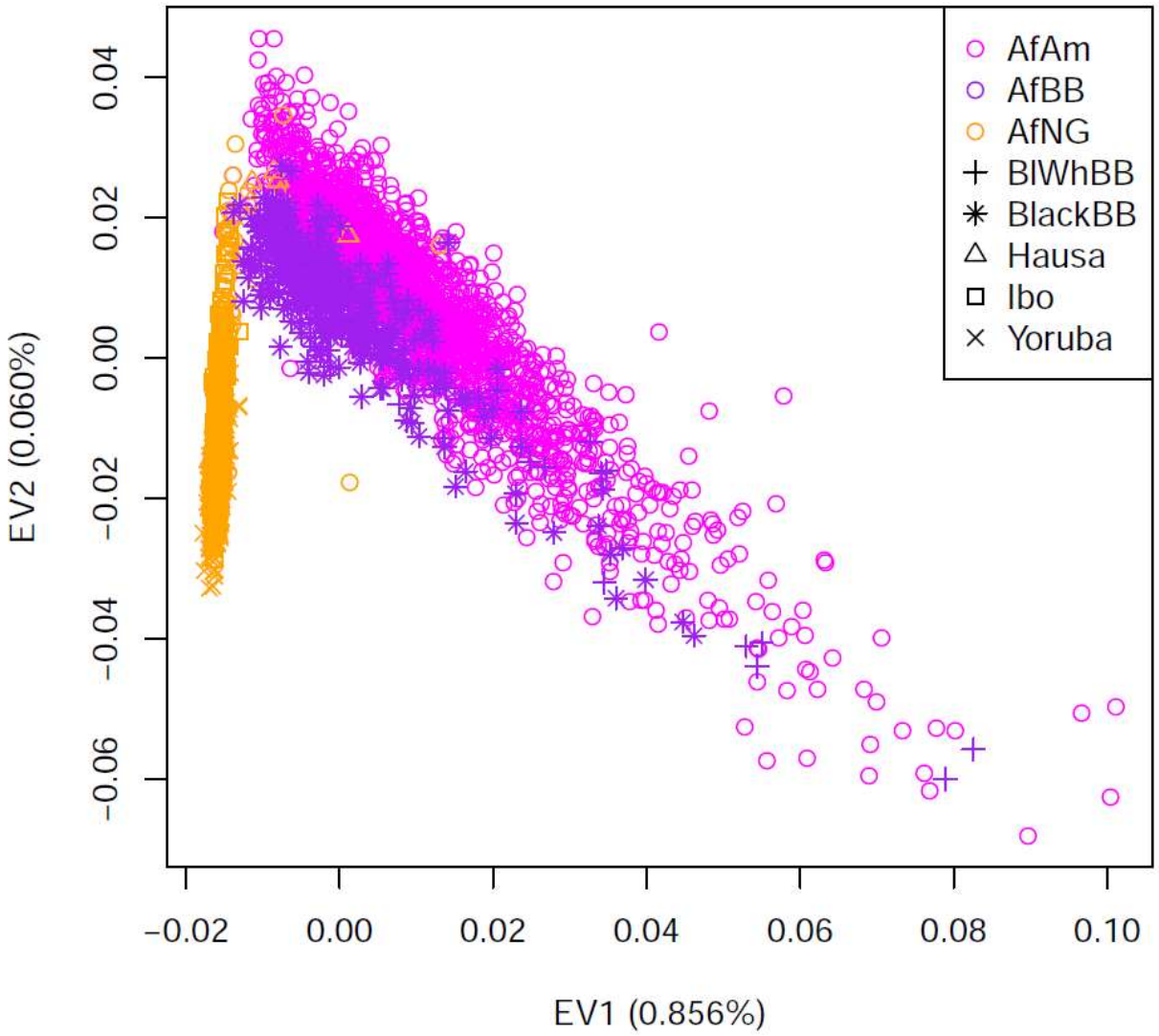


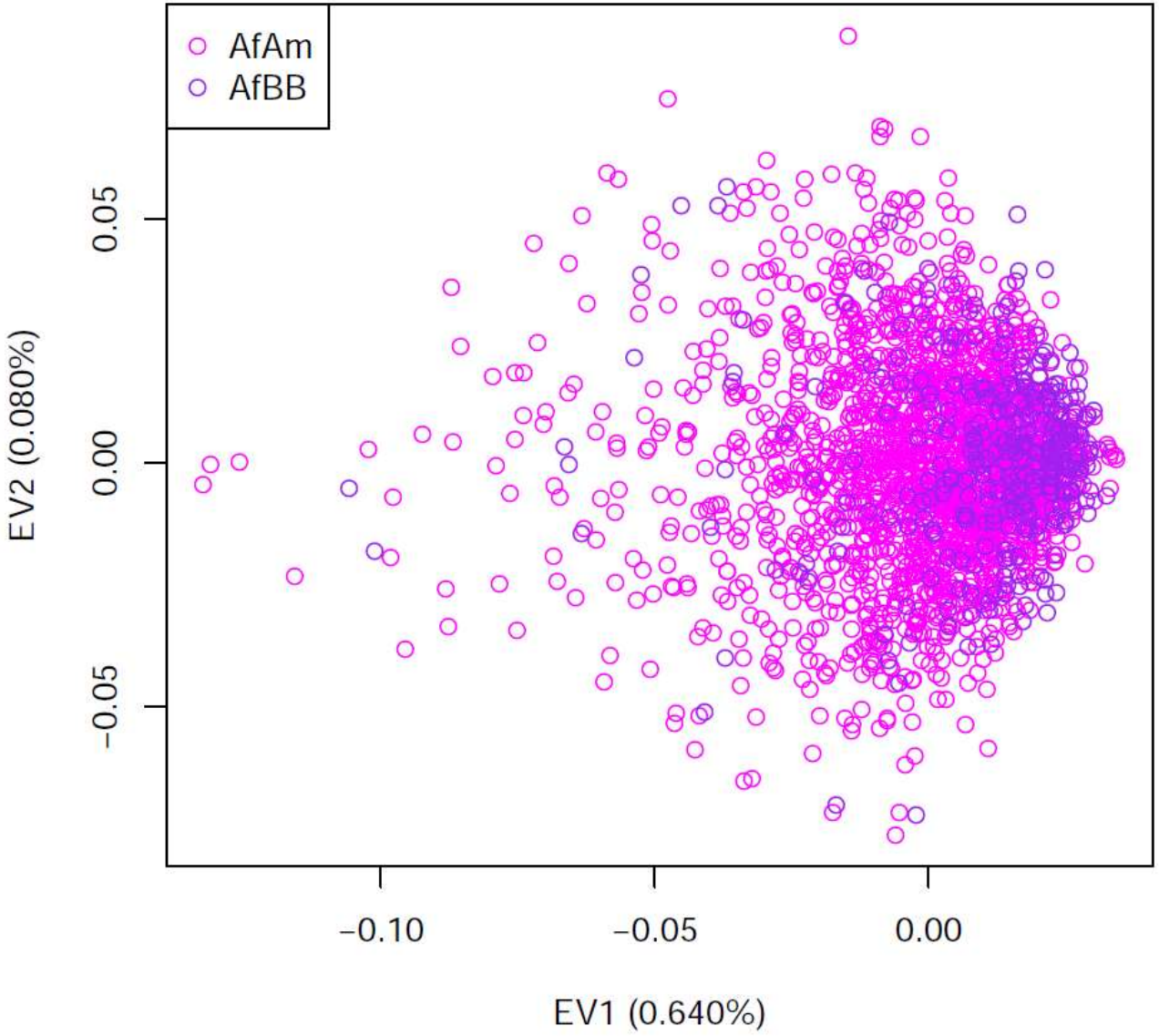


Population Stratification

3 Population structure

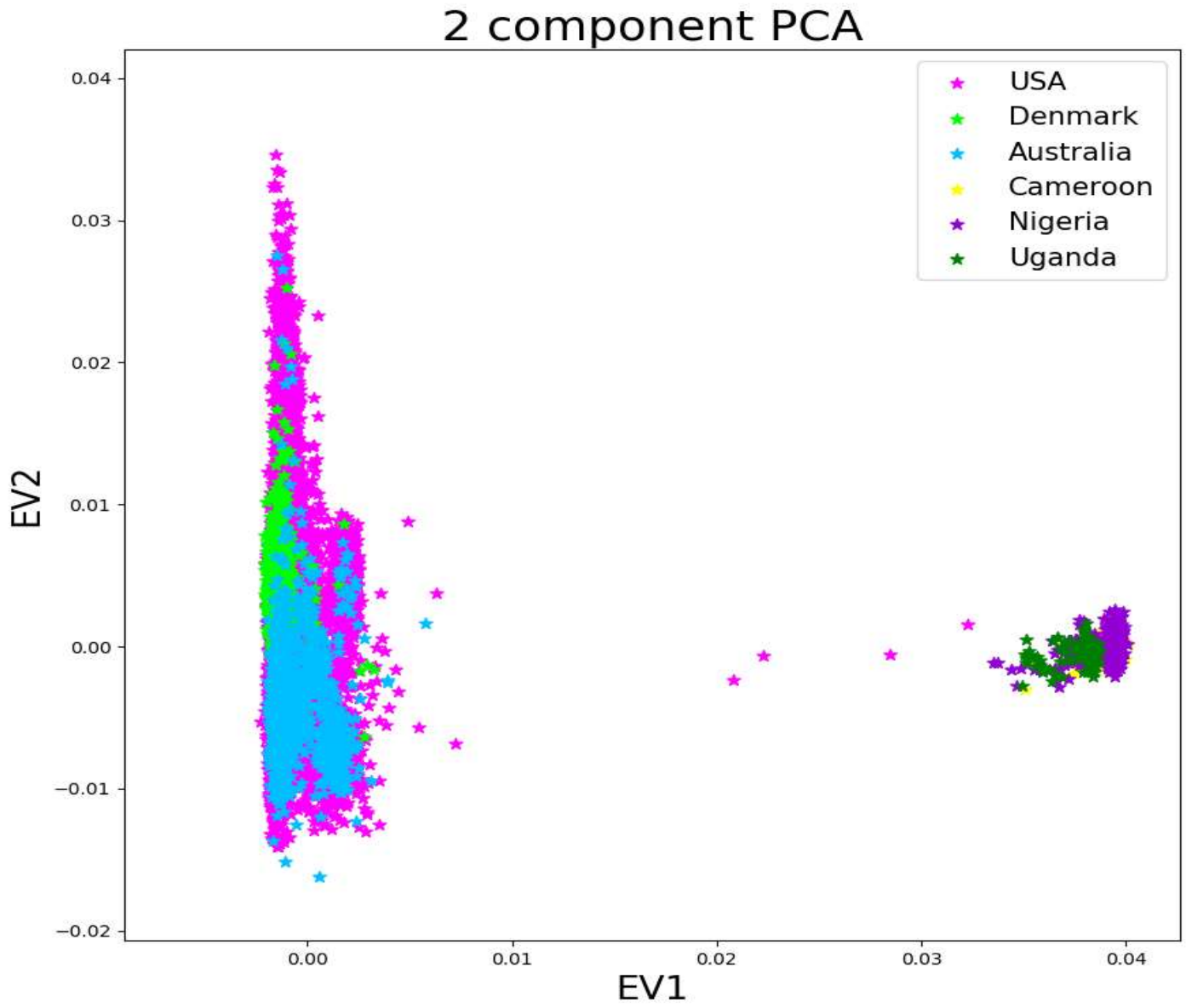
GWAS-CIDR dataset





3 Population structure

OncoArray DRIVE dataset



PCA-based methods

- **PCA-L**: Zeggini et al .(2008) ; Need et al. (2009)

$$\log\left(\frac{q}{1-q}\right) = \beta x + b_1\Phi_1 + b_2\Phi_2 + \cdots + b_d\Phi_d$$

- **EIGENSTRAT**: Price et al. (2006)

PCA-based methods

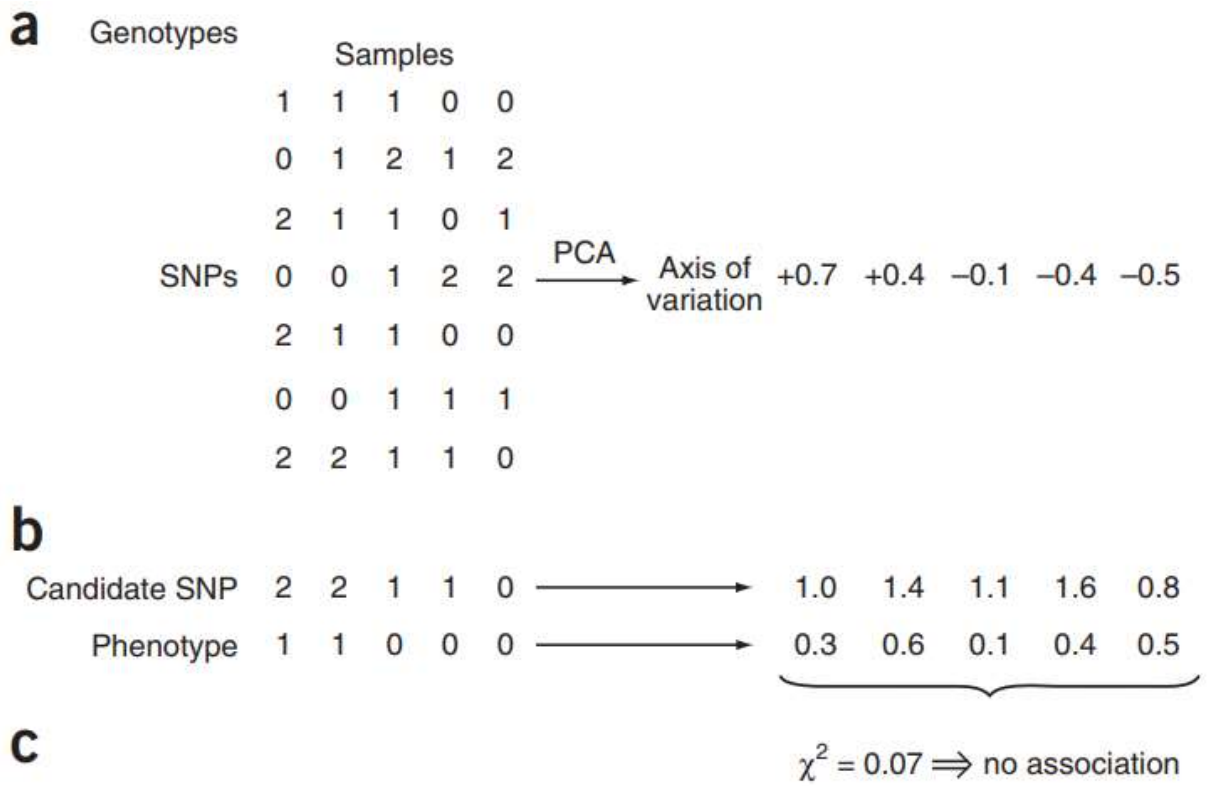
EIGENSTRAT:

- Detection of population structure via PCs.
- Adjustment of genotype and phenotype:

$$x_{ij}^{adj} = x_{ij} - \gamma_j a_i, \qquad \gamma_j = \frac{\sum_{i=1}^{n} a_i x_{ij}}{\sum_{i=1}^{n} a_i^2}$$

  ➔ Multivariate linear model.

- Verification of association.



Price, A. L., Patterson, N. J., Plenge, R. M., abd Nancy A Shadick, M. E. W., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*.

4 Simulated data via GWAsimulator

- Genotype data in case-control design.
- Population samples from genomic SNP chips.
- Specified multi locus disease model in specified regions.
- Similar LD patterns as the HapMap data and 1000 Genome Project.

HapMap3 data

POP	Num_samples
YRI	203
MKK	184
CEU	165
CHB	137
JPT	113
LWK	110
CHD	109
TSI	102
GIH	101
ASW	87
MXL	86
Consensus	**1397**

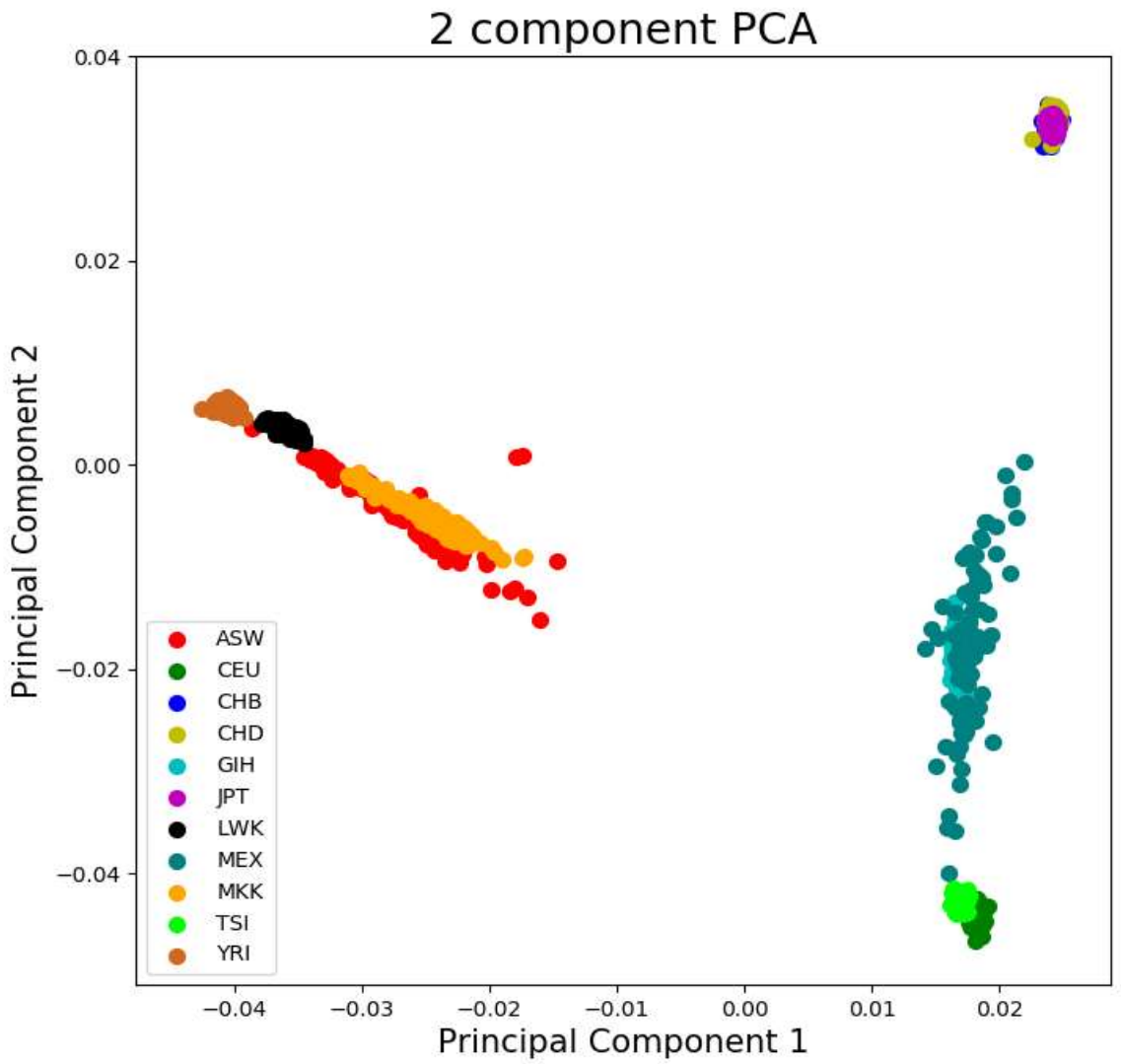
1000 Genome Project data

Population Code	Population Description	Super Population Code
CHB	Han Chinese in Beijing, China	EAS
JPT	Japanese in Tokyo, Japan	EAS
CHS	Southern Han Chinese	EAS
CDX	Chinese Dai in Xishuangbanna, China	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian Population in Spain	EUR
YRI	Yoruba in Ibadan, Nigeria	AFR
LWK	Luhya in Webuye, Kenya	AFR
GWD	Gambian in Western Divisions in the Gambia	AFR
MSL	Mende in Sierra Leone	AFR
ESN	Esan in Nigeria	AFR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Caribbeans in Barbados	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
PEL	Peruvians from Lima, Peru	AMR
GIH	Gujarati Indian from Houston, Texas	SAS
PJL	Punjabi from Lahore, Pakistan	SAS
BEB	Bengali from Bangladesh	SAS
STU	Sri Lankan Tamil from the UK	SAS
ITU	Indian Telugu from the UK	SAS

Li C, Li M, GWAsimulator: a rapid whole-genome simulation program.(2007)

4 Simulated data via GWAsimulator

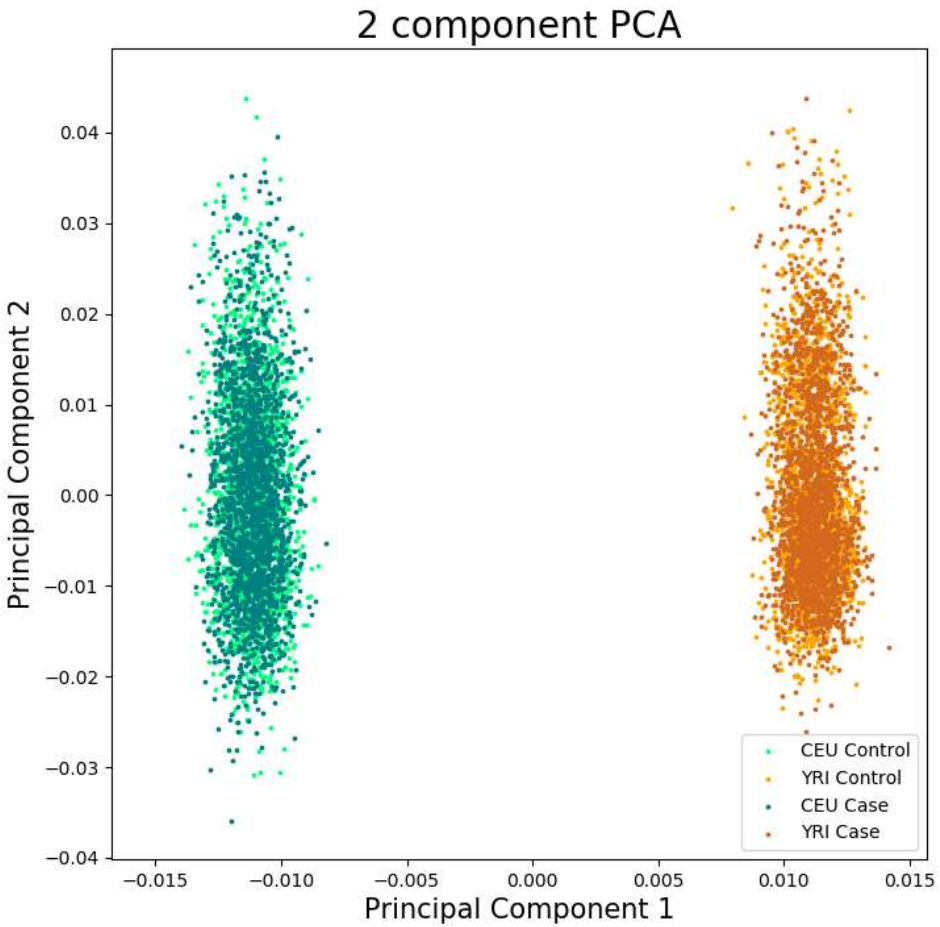
HapMap3 Data

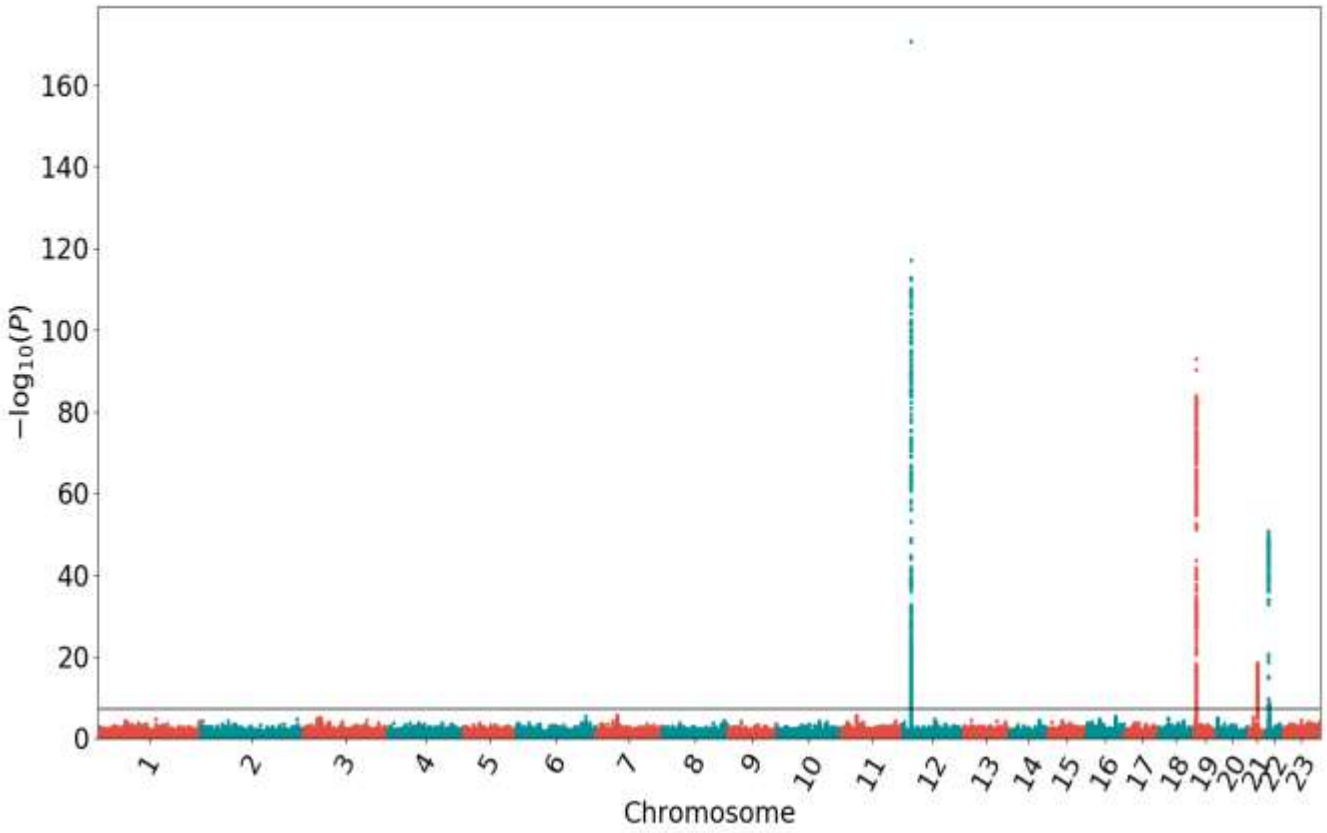


http://hapmap.ncbi.nlm.nih.gov/

4 Simulated data via GWAsimulator

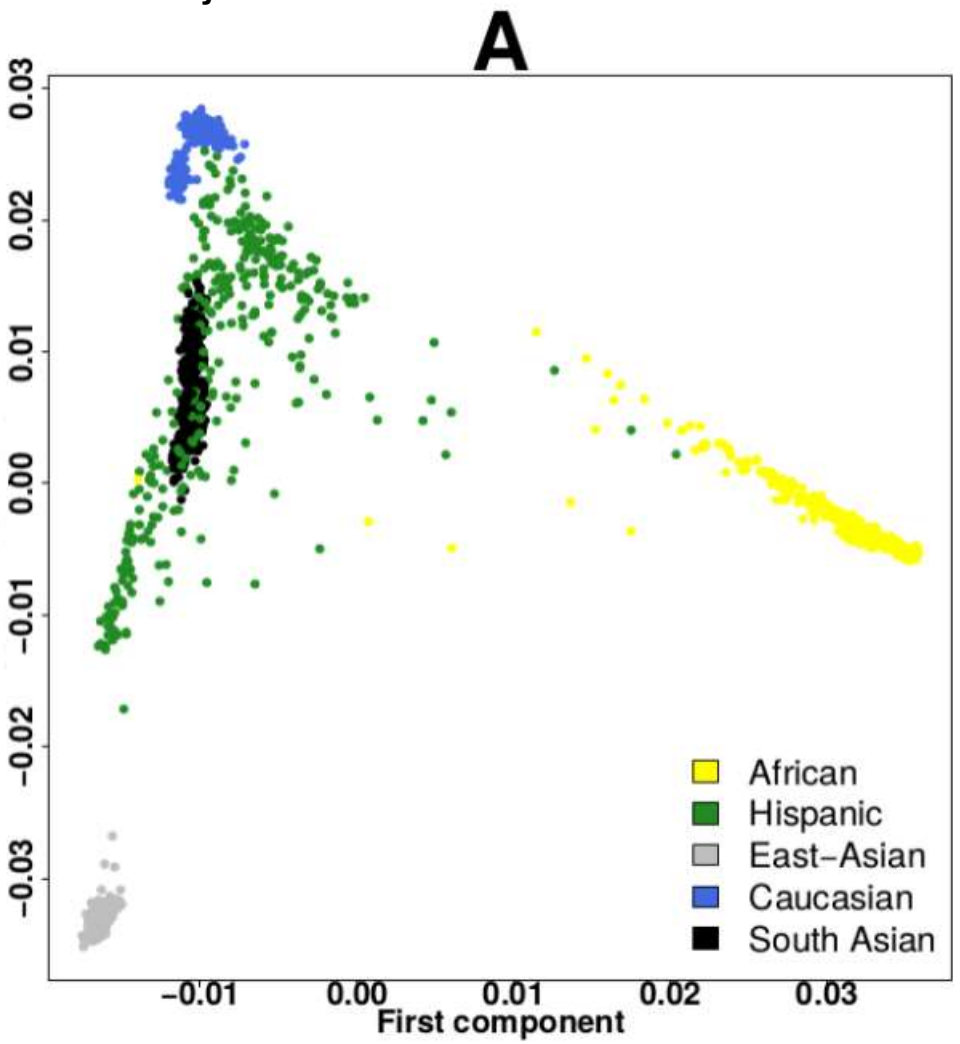
YRI & CEU Simulated data



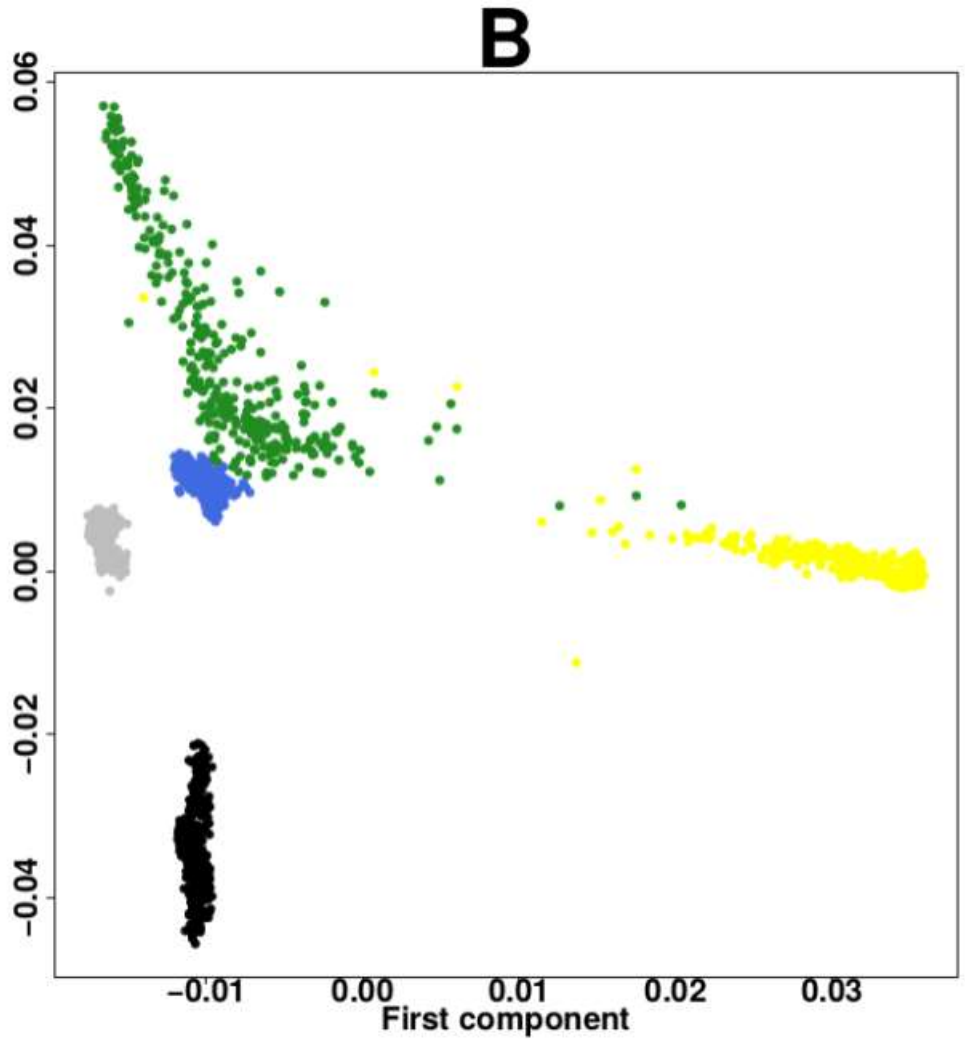


4 Linkage disequilibrium effect

1000 Genome Project data



Without LD pruning



After LD pruning

5 Linkage disequilibrium block partitioning

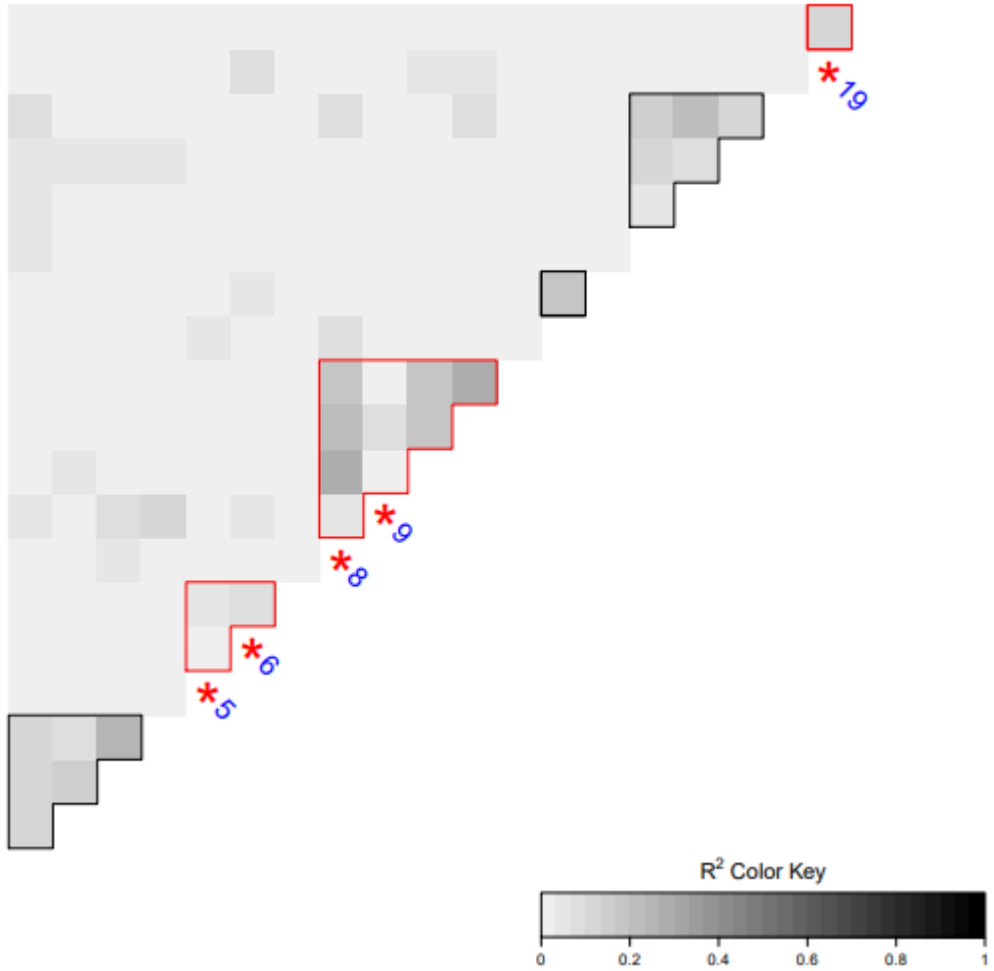
Spatial hierarchical clustering:

- Ward's Linkage criterion :

$$d_{wl}(A,B) = \frac{p_A \times p_B}{p_A + p_B} \|g_A - g_B\|_2^2$$

- Gap statistics to estimate the number of blocks

$$Gap(G) = \frac{1}{B}\sum_{b=1}^{B} \log\left(W_G^b\right) - \log(W_G)$$



➔ Feature selection at the block level in stead of single-SNP level.

A. Dehman, C. Ambroise & P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information, BMC Bioinformatics (2015).

Multi-task Regularized Regression

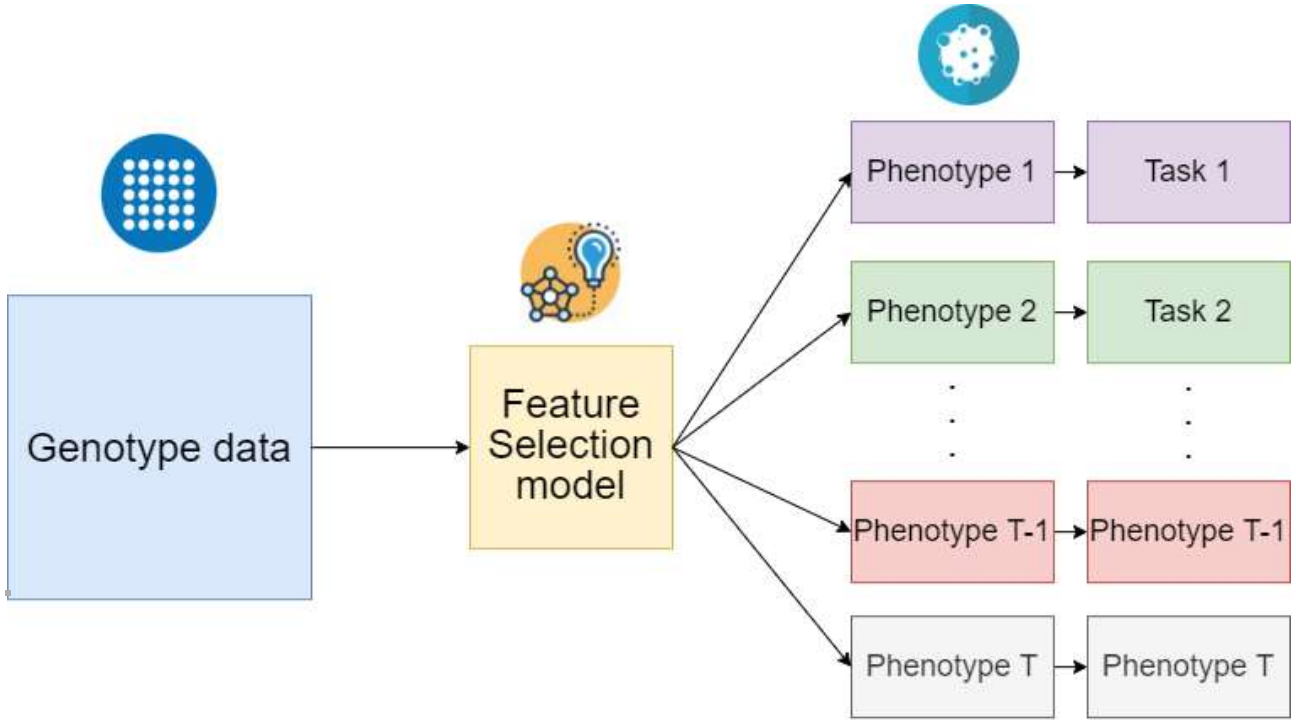
6 Multi-Task Learning formulation for GWAS data

- Distinct outcomes are used as tasks.

- Distinct samples populations are used as tasks.

6 Multi-Task Learning formulation for GWAS data

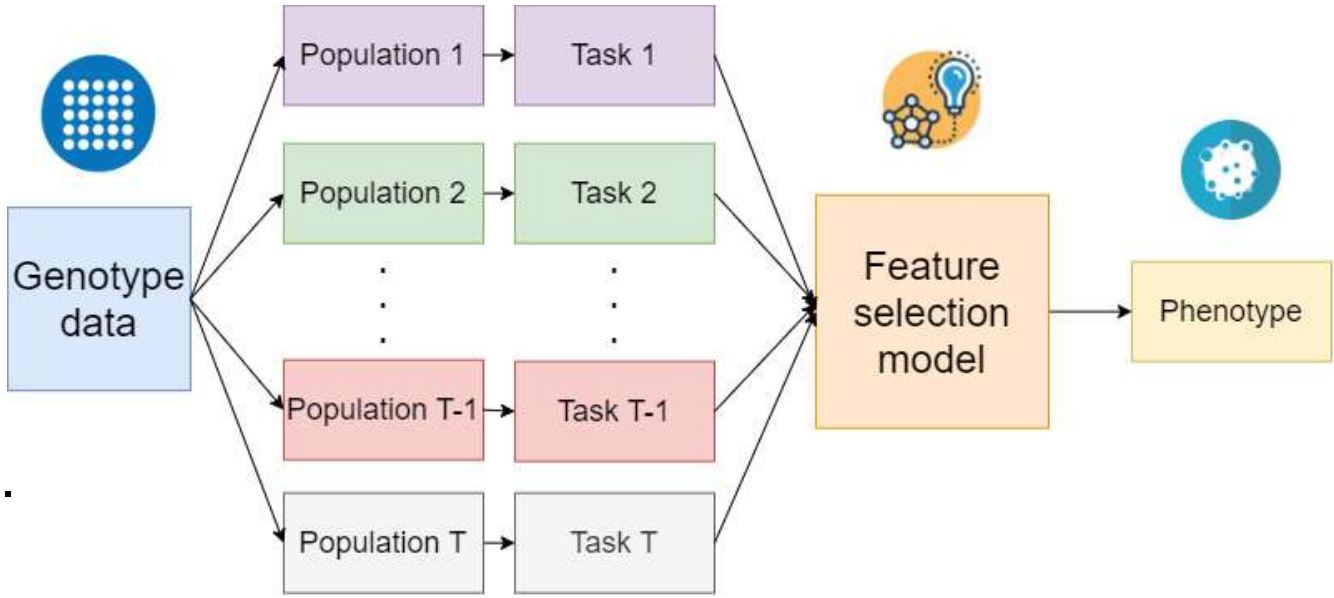
- Distinct outcomes are used as tasks.
- Distinct samples populations are used as tasks.



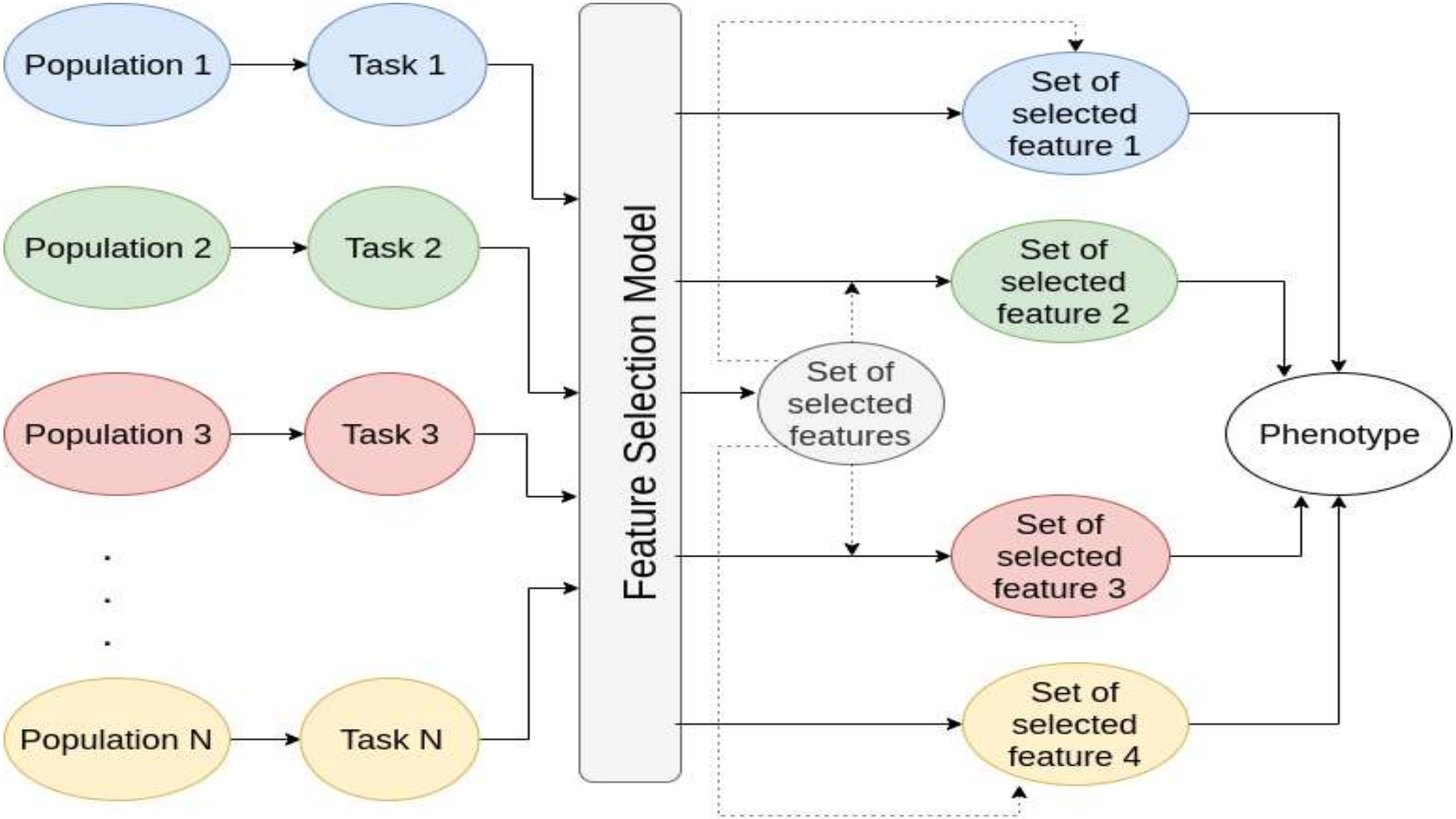
6 Multi-Task Learning formulation for GWAS data

- Distinct outcomes are used as tasks.

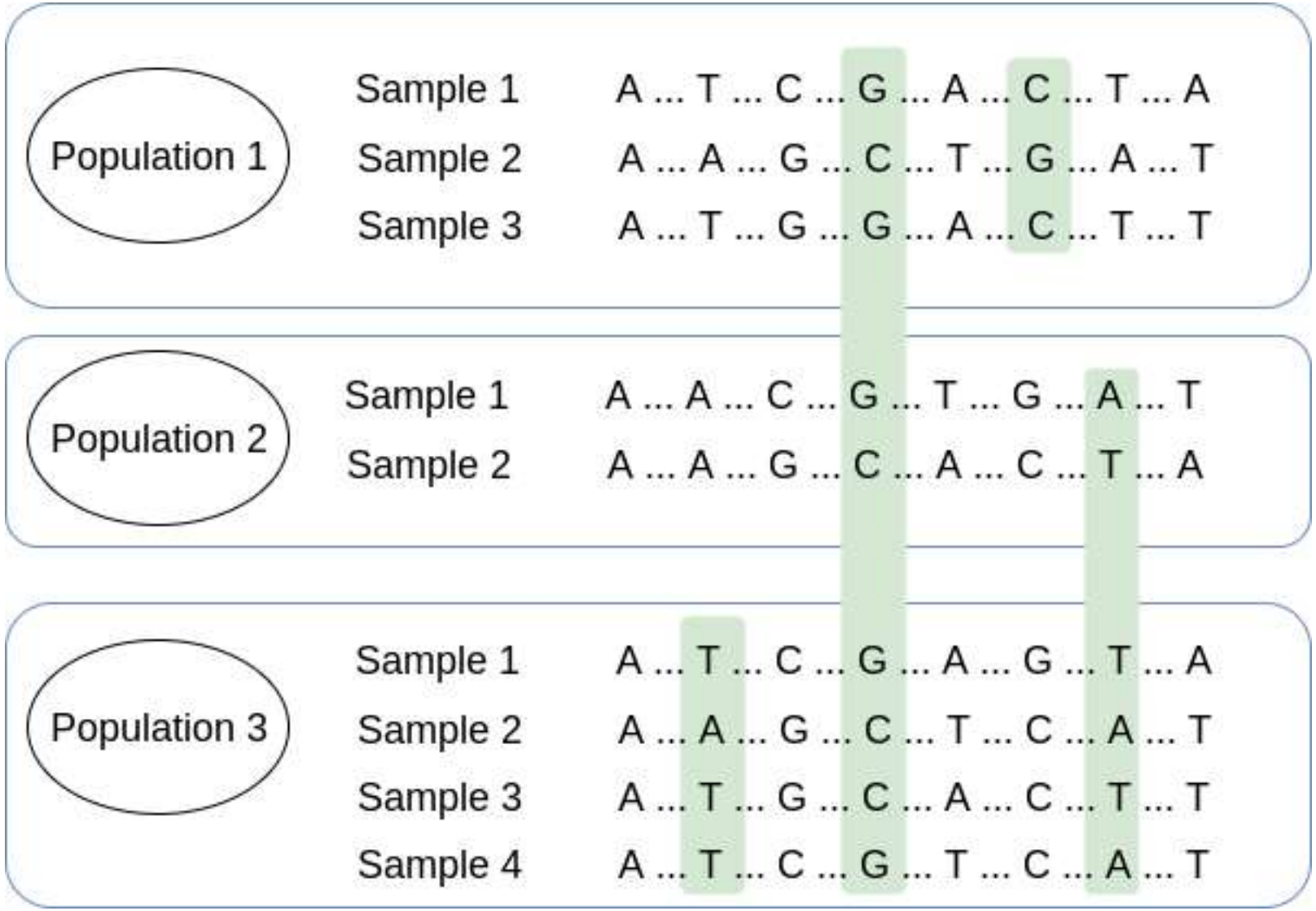
- Distinct samples populations are used as tasks.



6 Tasks assignment in Multi-task feature selection framework



6 Multi-task feature selection



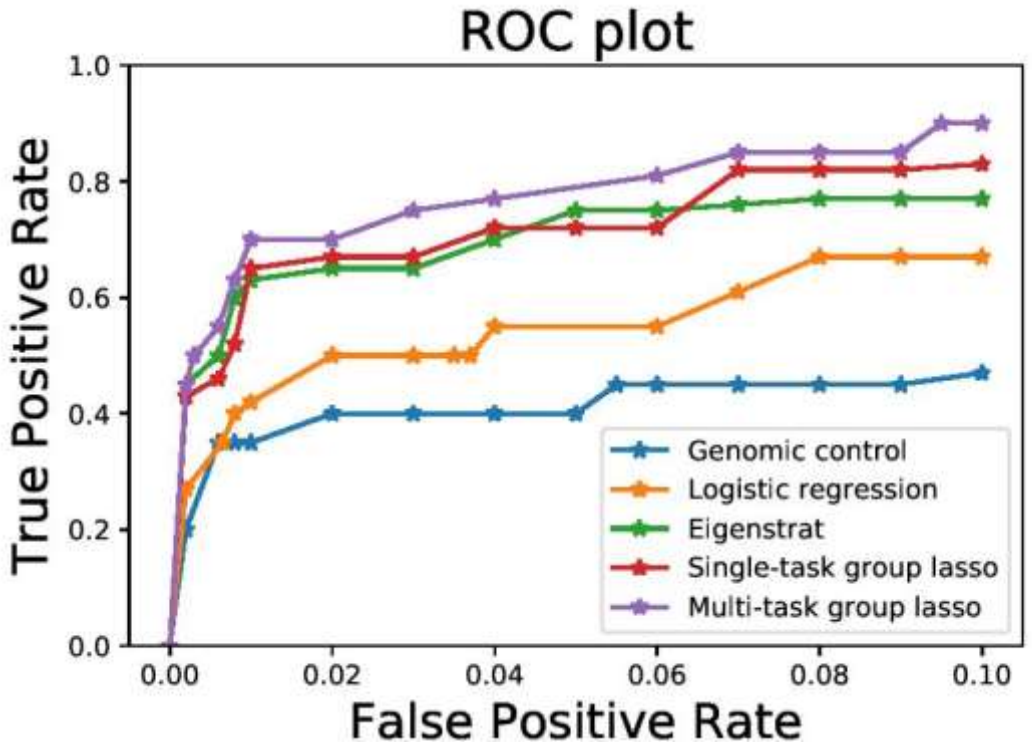
6 Multi-task group Lasso

- Clustering of SNPs into blocks following Linkage Disequilibrium (LD) patterns.
- Feature selection at the block level.
- Multi-task group Lasso where tasks are populations and groups are LD blocks.

$$\min_{\beta \epsilon \mathbb{R}^{T \times p}} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{m=1}^{n_t} \left\| Y^{(tm)} - \beta_{t0} + \sum_{j=1}^{p} \beta_j^{(t)} X_j^{(tm)}) \right\|_2^2 + \lambda \sum_{g=1}^{G} \sum_{t=1}^{T} \sqrt{p_g} \left\| \beta_g^{(t)} \right\|_2$$

6 Multi-task group Lasso

- Clustering of SNPs into blocks following Linkage Disequilibrium (LD) patterns.
- Feature selection at the block level.
- Multi-task group Lasso where tasks are populations and groups are LD blocks.



7 Conclusion

- Population heterogeneity can produce spurious associations.
- Population structure is influenced by Linkage Disequilibrium patterns.
- Markers can be partionated to blocks.
- Different populations may/may not share the same genetic patterns for the common phenotype.
- Applying Multi-task feature selection, using group Lasso, for joint association analysis of multiple populations at block level.

8 Future Work

- Apply the multi-task group Lasso to real data and other simulated data cases.
- Study the stability selection of the proposed method.
- Enforce the stability of the selection in the multi-task approach.

Thank you!