Stable Multi-task feature selection approach for Genome Wide Association Studies

Asma Nouira

Chloé-Agathe Azencott

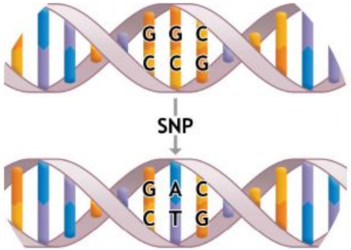
Centre for Computational Biology

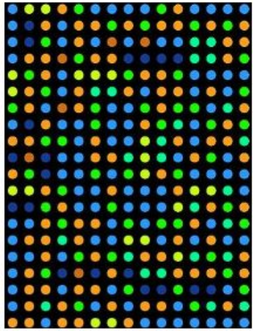
(CBIO)

U900 Lab Meeting

January 28, 2021

Genome Wide Association Studies





Goal: Find association between the genotype and the phenotype.

- The genotype: Single nucleotide polymorphism (SNP) arrays.
- The phenotype:
  - Quantitative: BMI, weight, age...
  - **Qualitative:** Case-control study

1 Breast Cancer datasets

CIDR Breast Cancer in the African Diaspora

Dimension: 3,827 samples x 2,379,855 SNPs

Phenotype: 1,681 cases and 2,085 controls

Populations: African Barbadian - African American - 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

Dimension: 28,281 samples x 528,620 SNPs

Phenotype: 13,846 cases and 14,435 controls

Populations: USA – Uganda – Nigeria – Cameroon – Australia – Denmark

Covariates: Age, estrogen rate, study, histological type…

Simulated data using GWAsimulator[1]

Dimension: 2,000 samples x 1,400,000 SNPs

Populations: 1000 European (CEU), 1000 African (YRI)

Phenotype: 500 CEU cases, 500 CEU controls, 500 YRI cases, 500 controls.

Disease loci: chromosomes 12, 19, 21 and 22.

[1] https://github.com/asmanouira/GWAS-admixed-population-simulator
http://biostat.mc.vanderbilt.edu/GWAsimulator

2 Classic GWAS analysis

- Preprocessing

Quality control

- MAF < 5%
- HWE-P-Value < 0.0001
- Remove samples with missing case/control criterion
- Sex check
- Remove samples and/or variants with high genotypic missing rate

Imputation

- Fill missing SNPs.
- Package: IMPUTE5[1]
- Reference dataset: 1000 Genomes Project (GP) Phase 3
- Exclude SNPs with 10% rate of missing values.

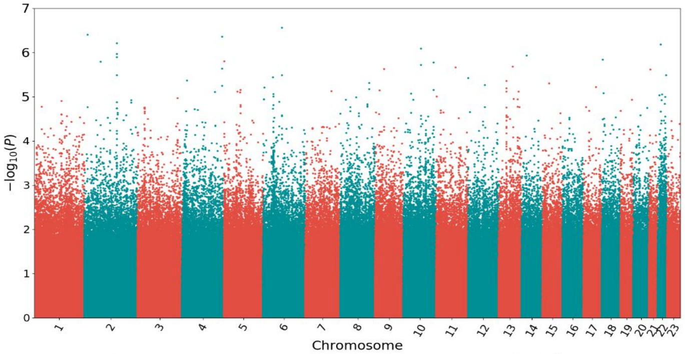
Linkage disequilibrium pruning

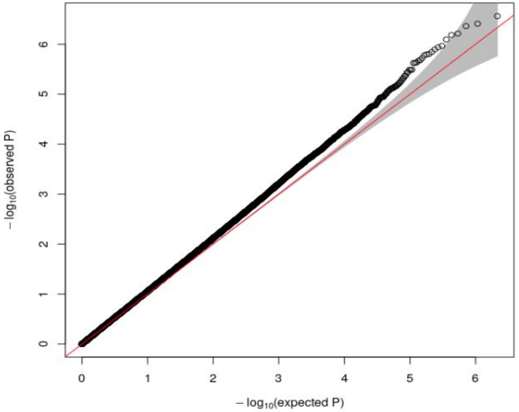
- Consider a window of 50 SNPs
- Calculate LD between each pair of SNPs in the window
- Remove one of a pair of SNPs if the LD is greater than 0.5
- Shift the window 5 SNPs forward

[1] https://jmarchini.org/impute5/

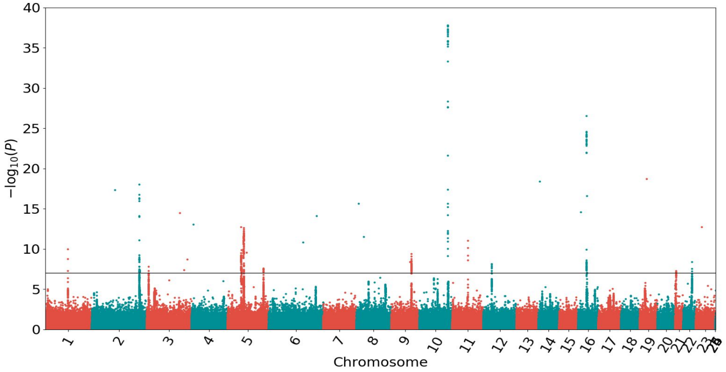
2 Classic GWAS analysis

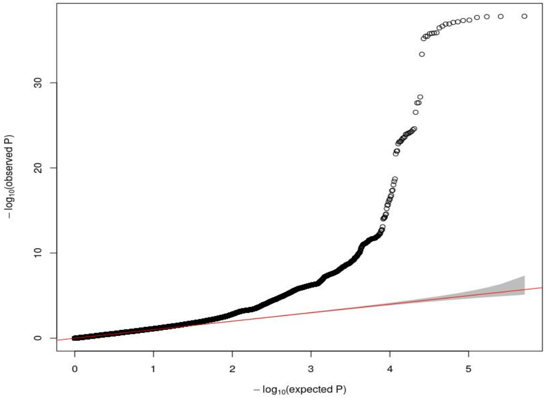
CIDR dataset





DRIVE-OncoArray dataset

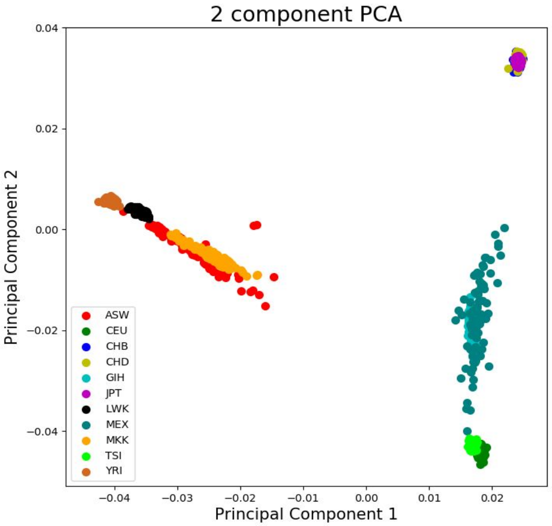


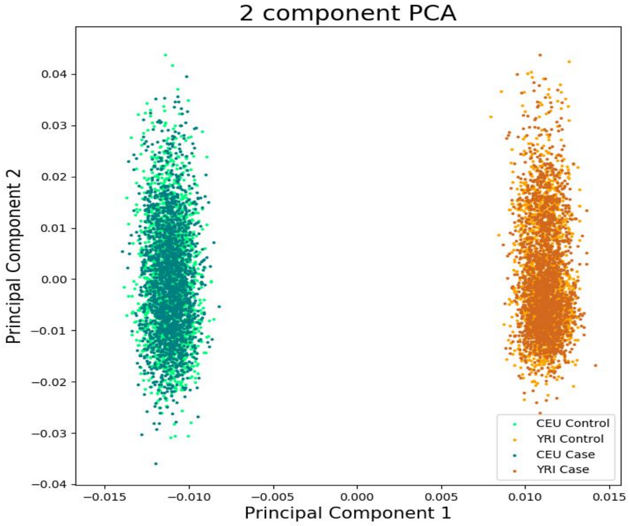


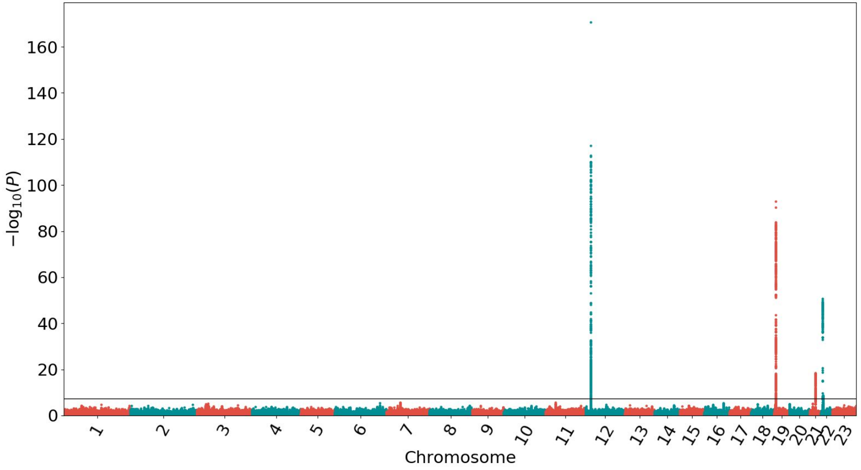
2 Classic GWAS analysis

Simulated case/control data using HapMap3 Data

- 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.



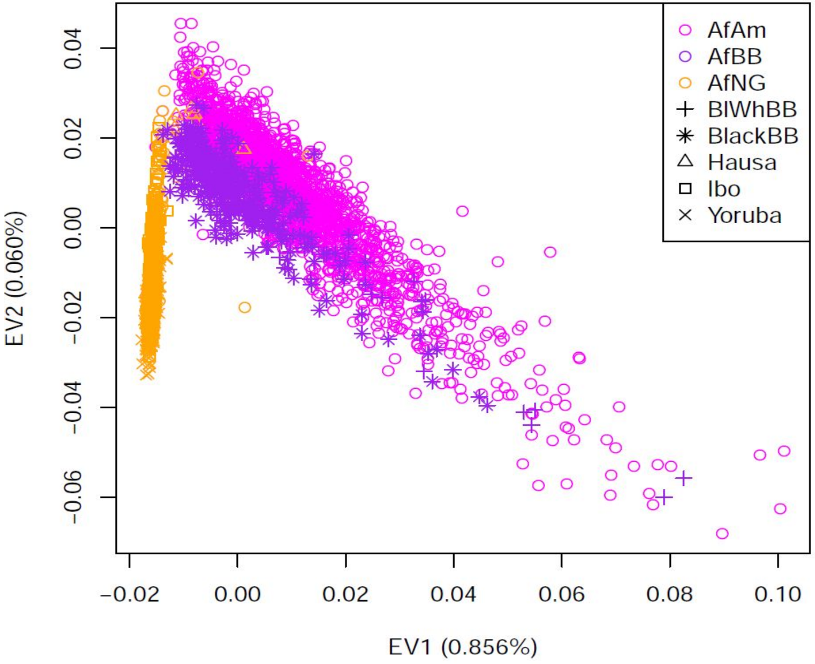


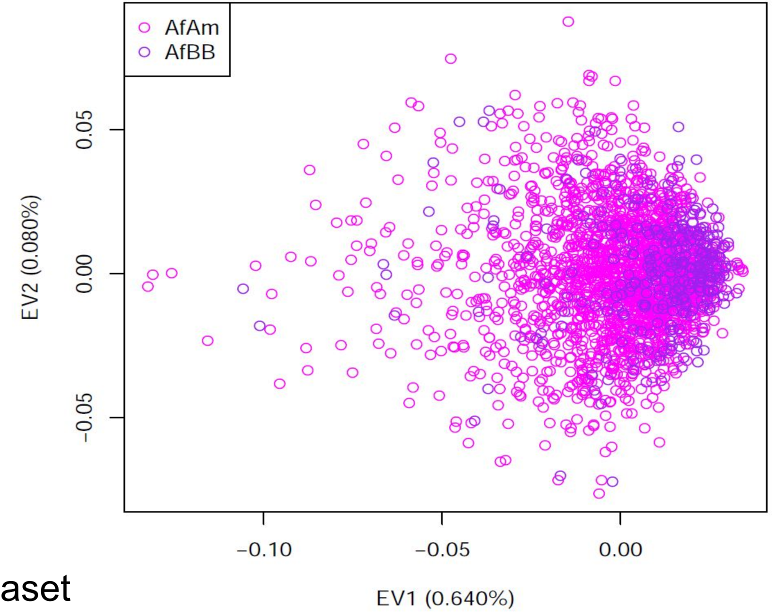


3 Population stratification

Population stratification refers to the presence of differences in allele frequencies between subpopulations within samples due to different ancestry.

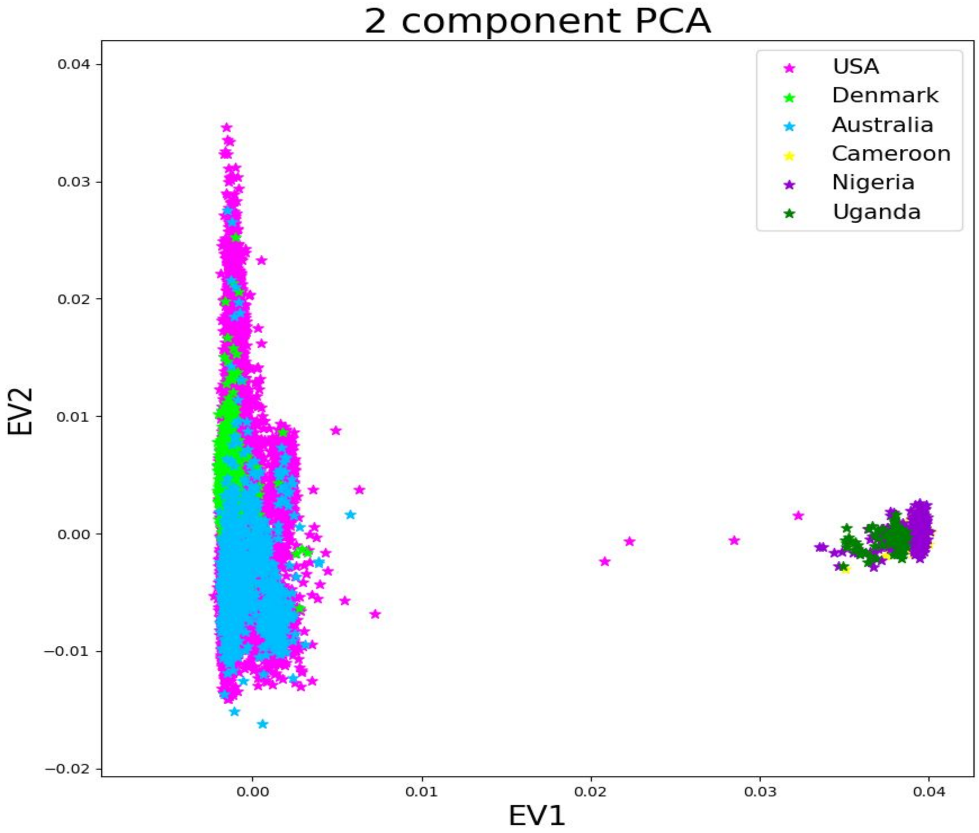
Principal Component Analysis (PCA)





CIDR dataset

3 Population stratification



DRIVE-OncoArray dataset

3 Population stratification

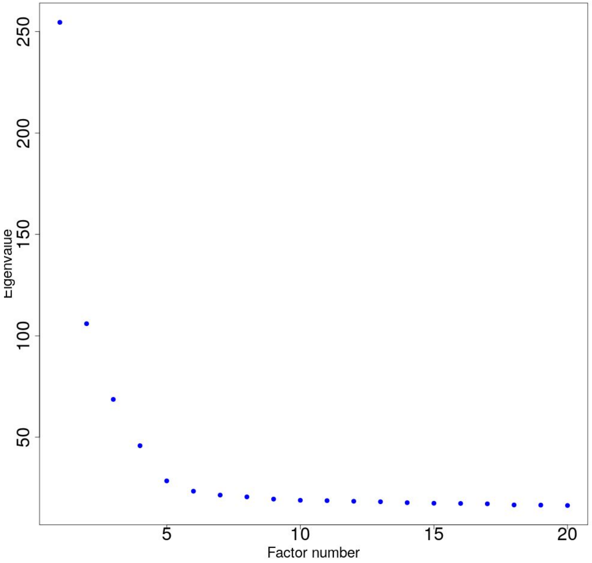
Stratification adjustment with PCA-based methods

- **PCA-L:**[1] Logistic regression with TOP PCs as covariates

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

- **EIGENSTRAT:**[2] Multivariate linear model

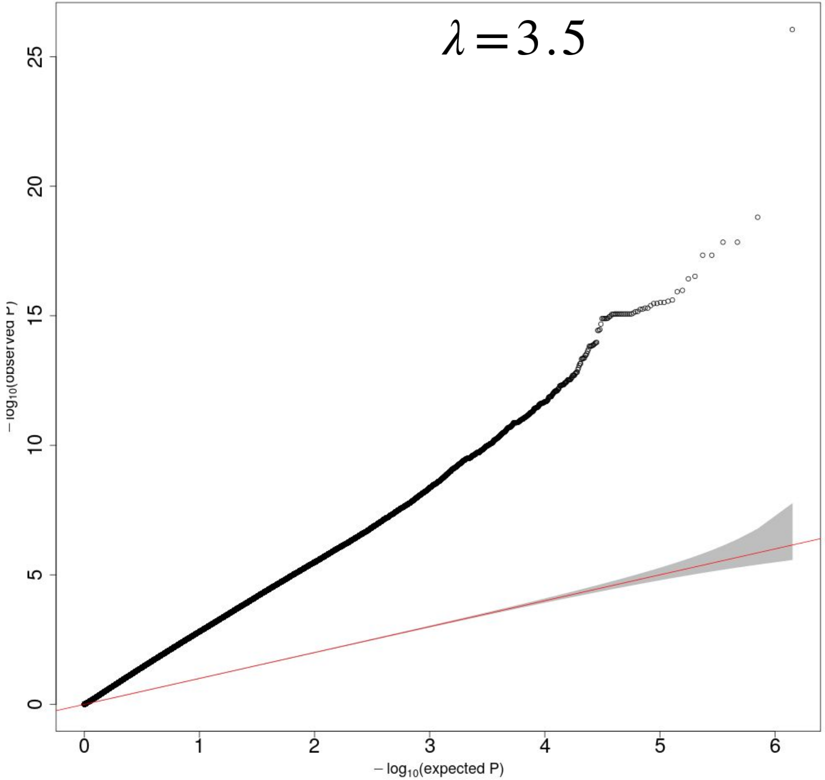
$$Y=\beta x+b_1\Phi_1+b_2\Phi_2+\ldots+b_d\Phi_d$$

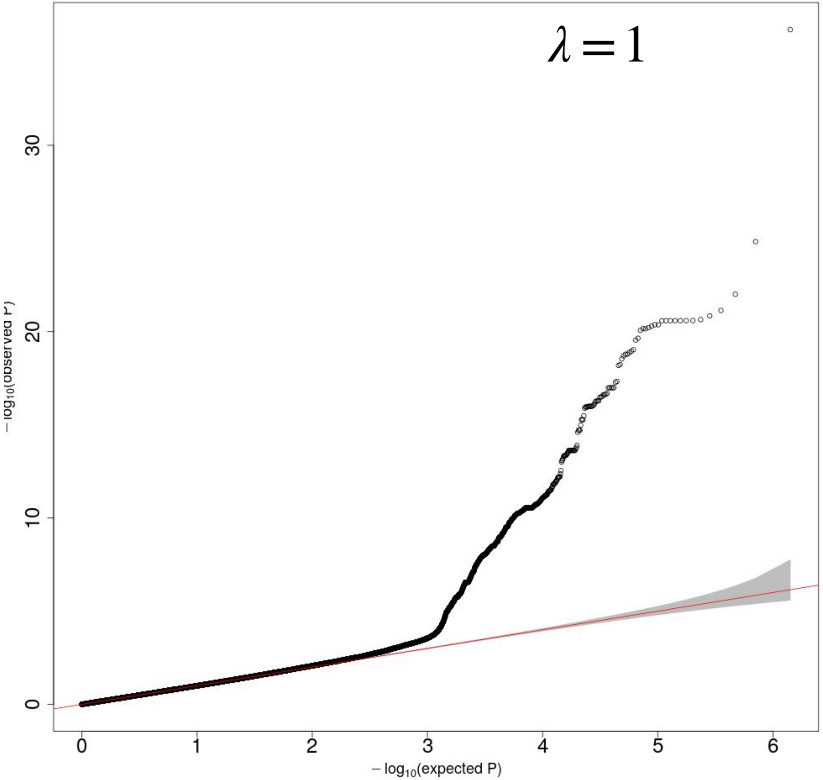


[1] Zeggini et al., 2008; Need et al., 2009
[2] https://github.com/DReichLab/EIG

3 Population stratification

Stratification adjustment with PCA-based methods





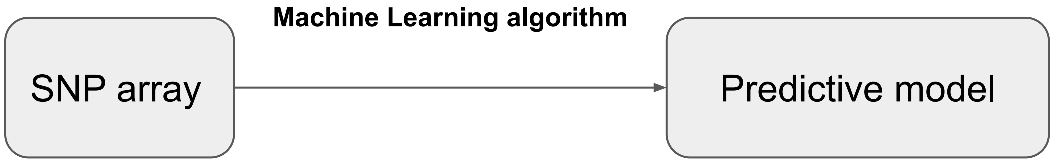
4 From GWAS to Machine learning

- **Microarray data: SNP arrays**

Curse of dimensionality (p>>N): $p \approx 10^5 - 10^7\ ,\ N \approx 10^2 - 10^4$

- **Notations** $y_i \in \{1, 2\}$

  $x_{i,j} \in \{0, 1, 2\}$

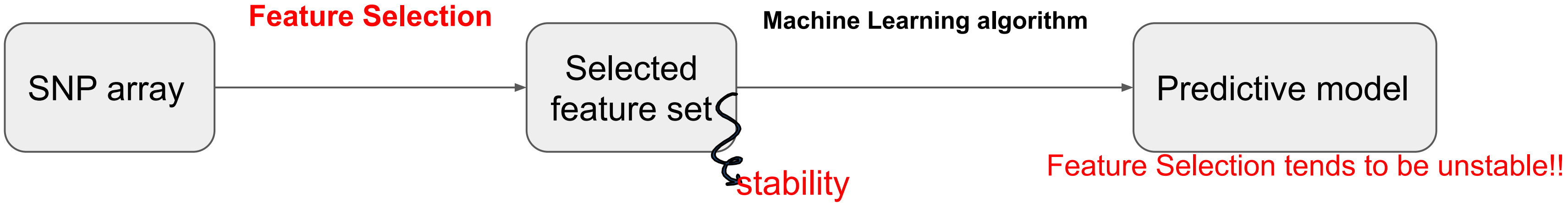


SNP array

Machine Learning algorithm

Predictive model

- **Biomarker identification : Explore feature selection models**



SNP array

Feature Selection

Selected feature set

Machine Learning algorithm

Predictive model

stability

Feature Selection tends to be unstable!!

4 From GWAS to Machine learning

Feature selection

- Regularization: adding an additional penalty term

$$\underset{\beta \in \mathbb{R}^p}{argmin} \underbrace{\|y - \beta X\|_2^2}_{\text{squared loss}} + \underbrace{\lambda \Omega(\beta_1, \beta_2, \ldots, \beta_j)}_{\text{regularization term}}$$

- Lasso: shrinkage and feature selection (L1-regularization)

$$\underset{\beta \in \mathbb{R}^p}{argmin} \|y - \beta X\|_2^2 + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{sparsity}}$$

- Group lasso: allow predefined groups of covariates to jointly be selected

$$\underset{\beta \in \mathbb{R}^p}{argmin} \|y - \beta X\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \underbrace{\eta \sum_{g \in \mathcal{G}} \|\beta_g\|}_{\text{sparsity on the group-level}}$$

-Multi-task lasso: allows to fit multiple regression problems jointly enforcing the selected features to be the same across task

$$\underset{\beta \in \mathbb{R}^{T \times p}}{argmin} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{m=1}^{n_t} \left\| Y^{(tm)} - \left(\beta_{t0} + \sum_{j=1}^{p} \beta_j^{(t)} X_j^{(tm)} \right) \right\|_2^2 + \underbrace{\lambda \sum_{j=0}^{p} \sum_{t=1}^{T} |\beta_j^{(t)}|}_{\text{sparsity for each task t}}$$

5 Linkage disequilibrium blocks clustering

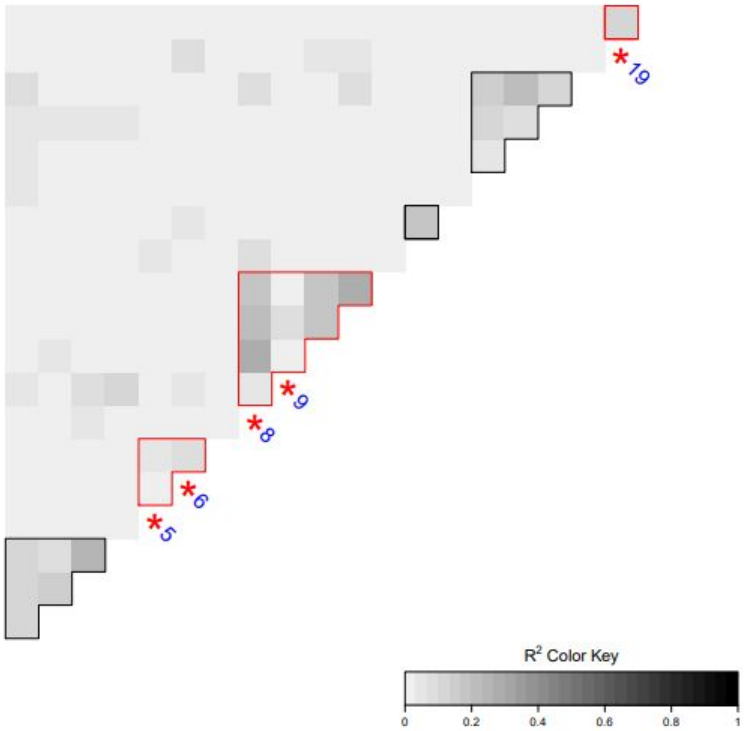
Spatial hierarchical clustering:

- 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

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

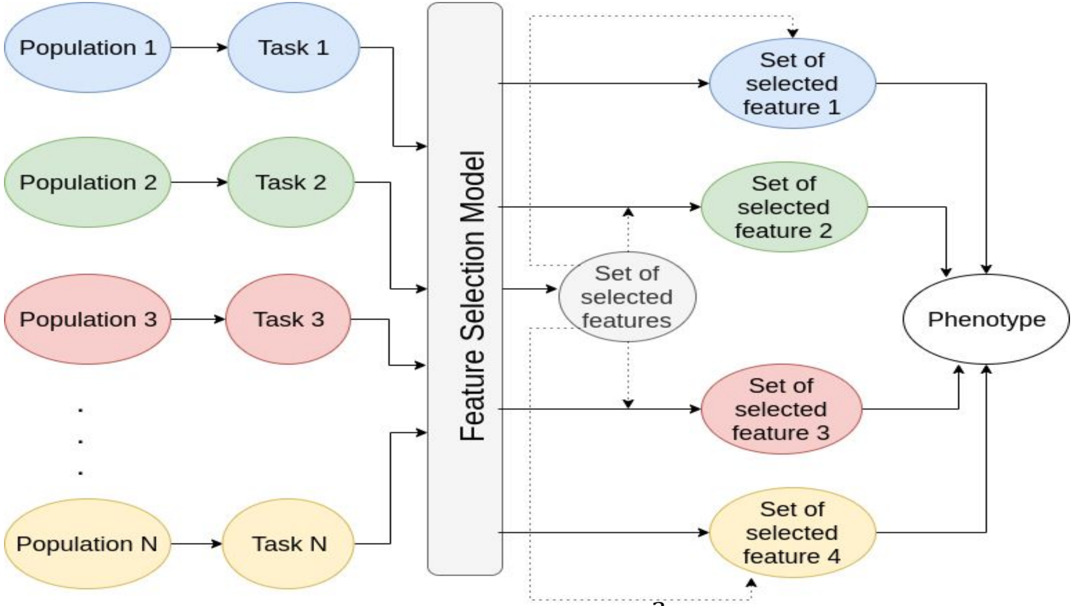


⇒ Feature selection on the block-level instead of single-SNP level.

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

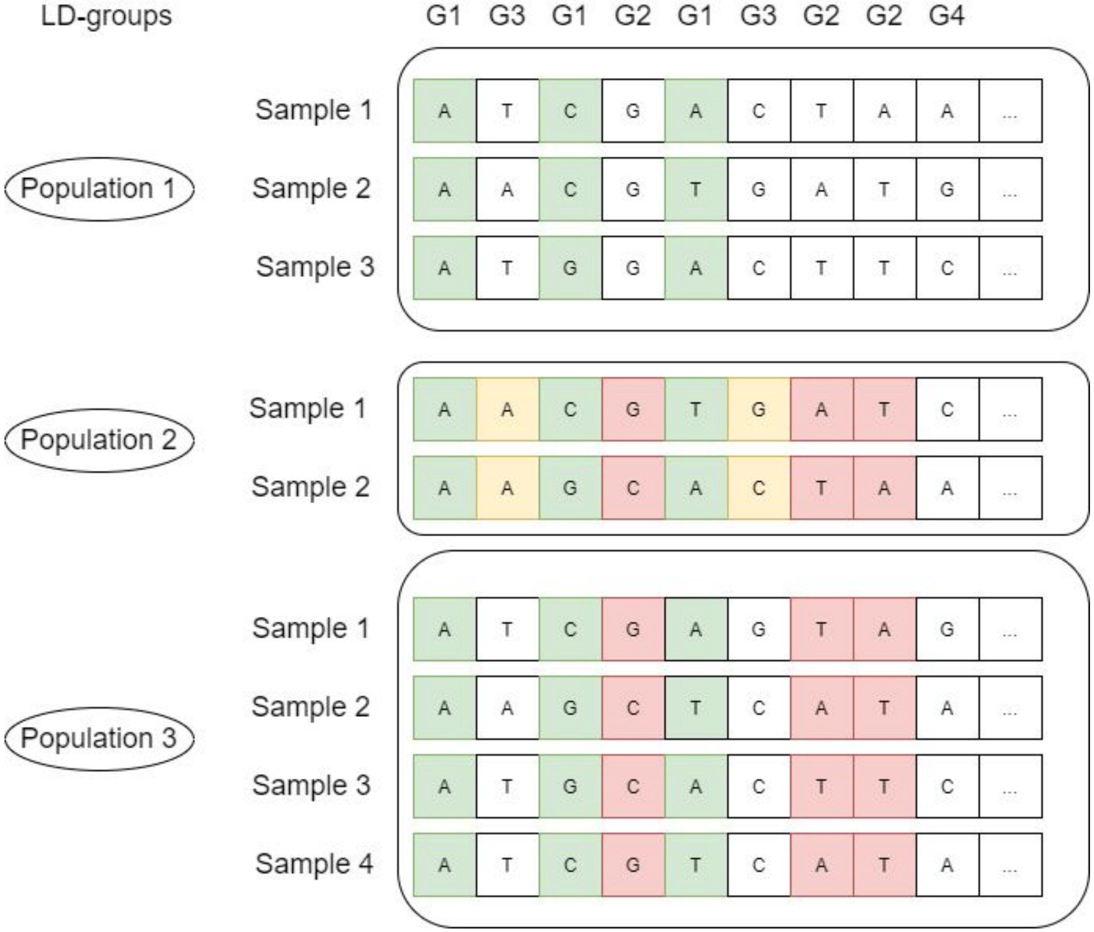
6 Multi-task group lasso

- 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.



$$\min_{\beta\epsilon\mathbb{R}^{T\times p}} \underbrace{\sum_{t=1}^{T}\frac{1}{n_t}\sum_{m=1}^{n_t}\left\|Y^{(tm)} - \beta_{t0} + \sum_{j=1}^{p}\beta_j^{(t)}X_j^{(tm)})\right\|_2^2}_{\text{loss for each task}} + \lambda \underbrace{\sum_{g=1}^{G}\sum_{t=1}^{T}\sqrt{p_g}\left\|\beta_g^{(t)}\right\|_2}_{\text{sparsity on the group-level for each task t}}$$

6 Multi-task group lasso



LD-groups		G1	G3	G1	G2	G1	G3	G2	G2	G4	
Population 1	Sample 1	A	T	C	G	A	C	T	A	A	...
	Sample 2	A	A	C	G	T	G	A	T	G	...
	Sample 3	A	T	G	G	A	C	T	T	C	...
Population 2	Sample 1	A	A	C	G	T	G	A	T	C	...
	Sample 2	A	A	G	C	A	C	T	A	A	...
Population 3	Sample 1	A	T	C	G	A	G	T	A	G	...
	Sample 2	A	A	G	C	T	C	A	T	A	...
	Sample 3	A	T	G	C	A	C	T	T	C	...
	Sample 4	A	T	C	G	T	C	A	T	A	...

Multi-task group lasso

Stability selection

[Meinshausen and Bühlmann, 2010]

Bootstrap aggregation procedure:

- Feature selection is performed on bootstrap subsamples

⇒ The results of the repetition are aggregated

- Very precise statement of the significance of the selected features set

⇒ Reduce the false positives selection

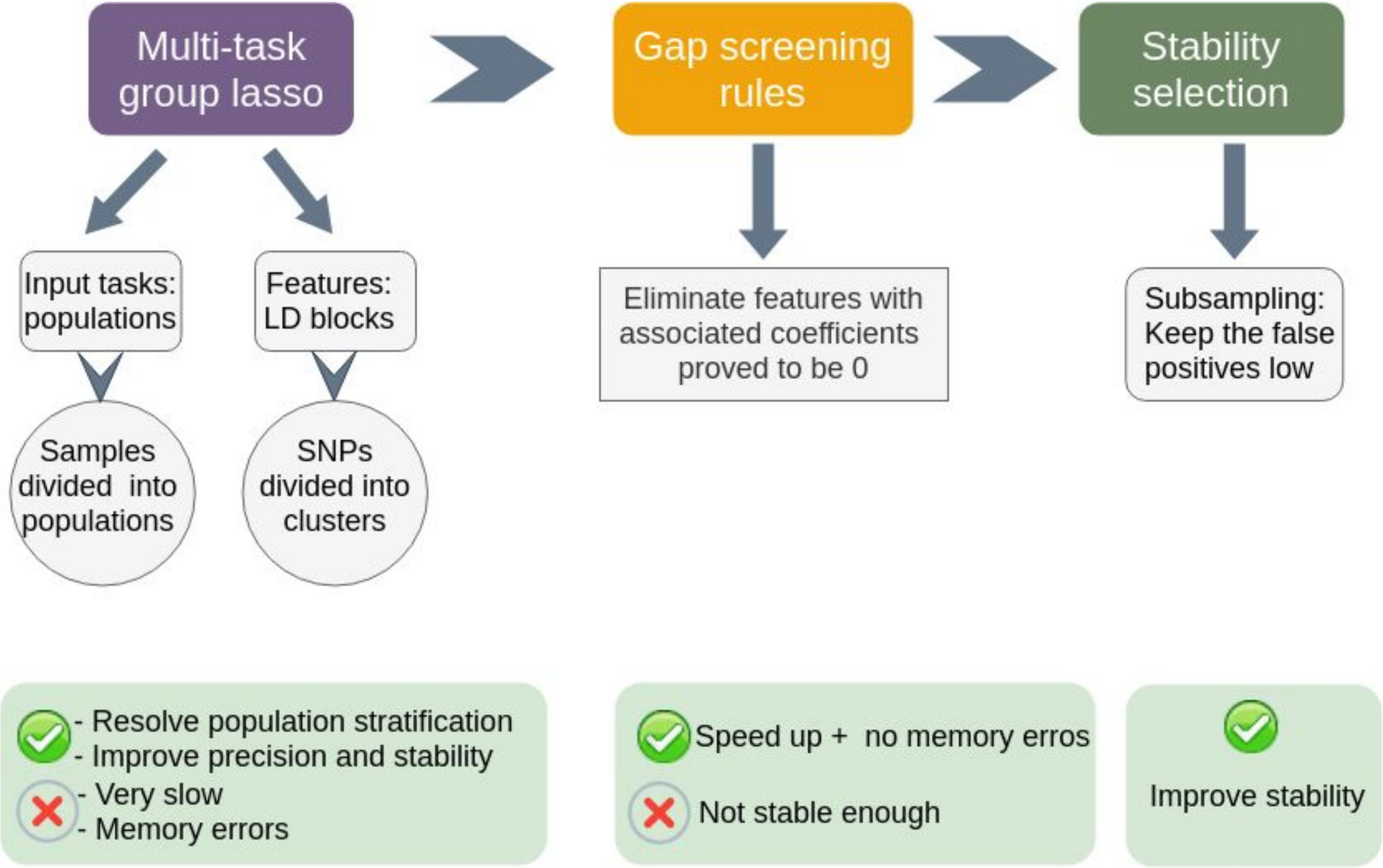
Procedure:

- We compute the probability of the selection of a variable $k \in \{1, \ldots, p\}$: $\pi_k^\lambda = Pr^* \left[k \in \widehat{S}^\lambda(I) \right]$
- For a chosen cut-off $\frac{1}{2} \leq \pi_{thre} \leq 1$:

$$\widehat{S}^{stable} = \left\{ k: \pi_k^\lambda \geq \pi_{thre} \right\}$$

⇒ Only variables that are selected consistently across all the random halves remain.

Limitations and conclusion



Multi-task group lasso

Input tasks: populations

Samples divided into populations

Features: LD blocks

SNPs divided into clusters

Gap screening rules

Eliminate features with associated coefficients proved to be 0

Stability selection

Subsampling: Keep the false positives low

✓ - Resolve population stratification
- Improve precision and stability

✗ - Very slow
- Memory errors

✓ Speed up + no memory erros

✗ Not stable enough

✓ Improve stability

Futur Work

- Implement stability selection for multi-task group lasso.
- Apply the multi-task group lasso for real data (breast cancer phenotype).
- More speed up.

Acknowledgement

- CBIO team
- GWAS team: Chloé, Héctor and Vivien.
- U900
- ANR

THANK YOU!