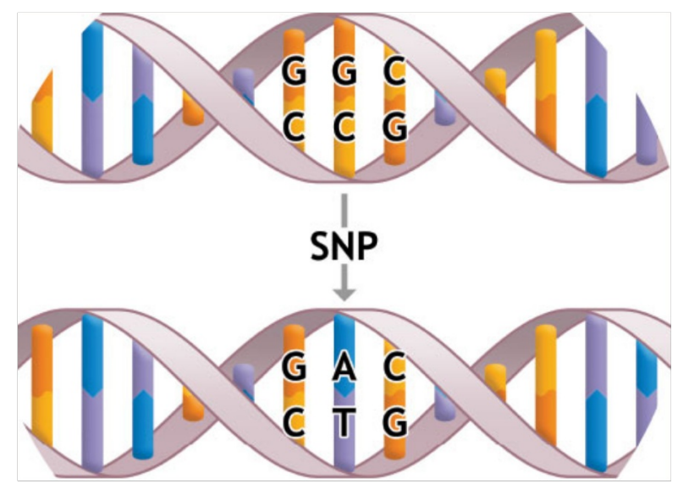
Multitask group Lasso for Genome Wide Association Studies in admixed populations

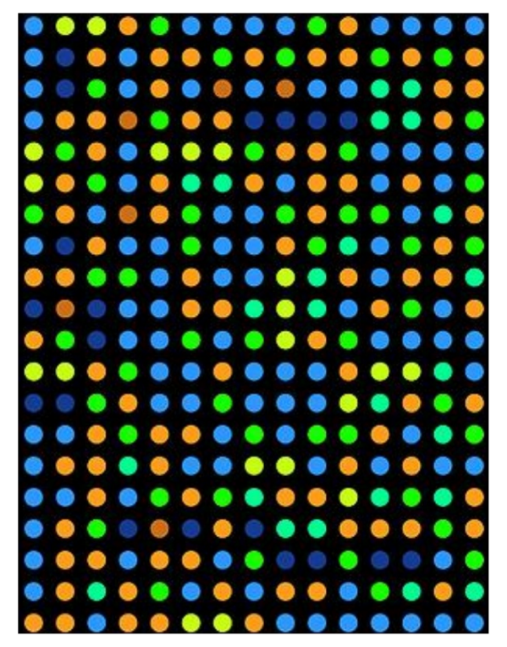
Asma Nouira, Chloé-Agathe Azencott

MINES ParisTech, CBIO-Centre for Computational Biology, Institut Curie, INSERM, U900, PSL Research University

SCAPHE ANR-18-CE45-0021-01

Genome Wide Association Studies (GWAS)



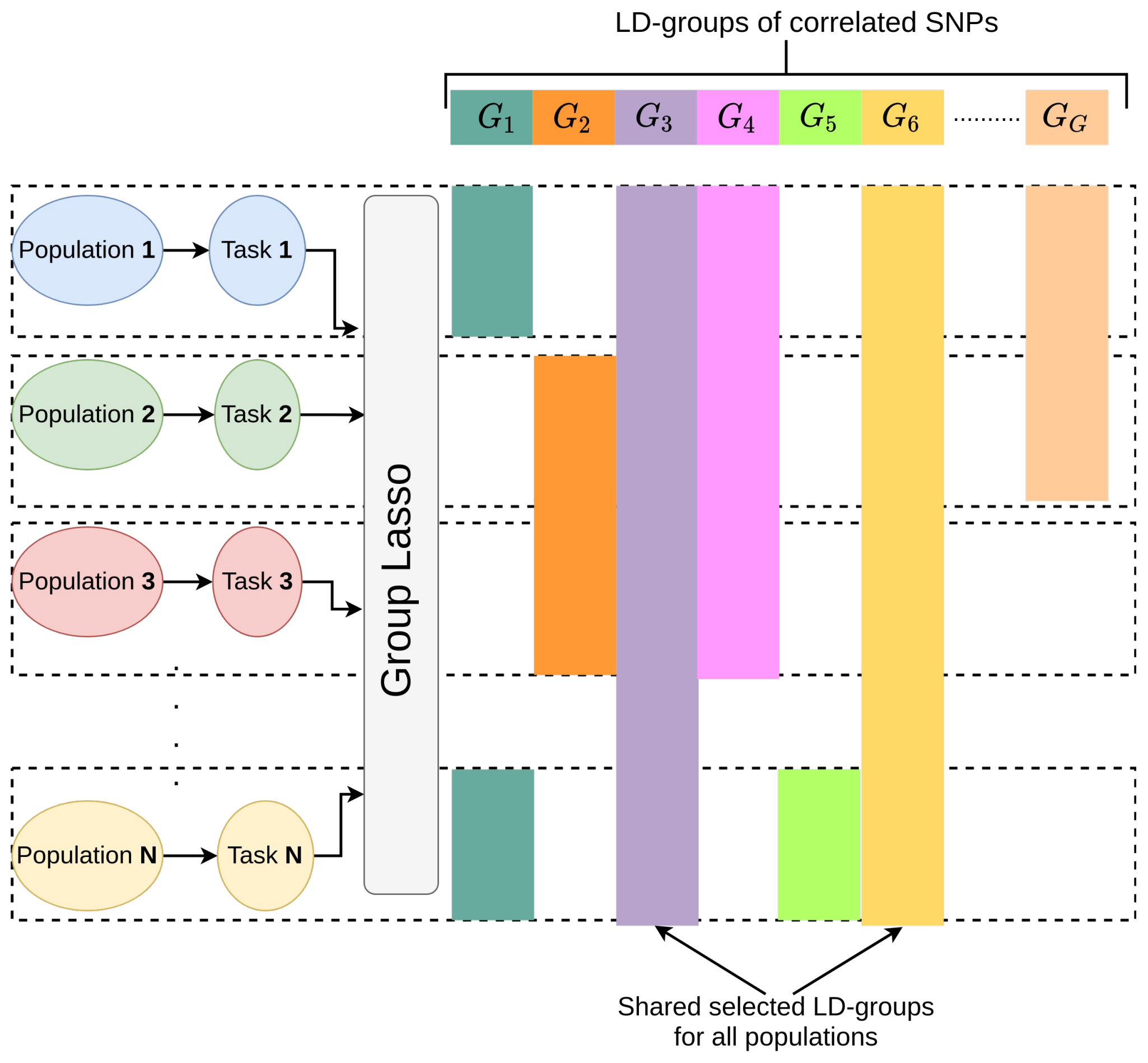


Find associations between the genotype represented by single-nucleotide polymorphisms (SNPs) and the phenotype (e.g. the disease)

The model: Multi-task group Lasso

Tasks correspond to subpopulations and groups correspond to LD-groups

$$\min_{B\in\mathbb{R}^{T\times(p+1)}} \boxed{\sum_{t=1}^{T}} \frac{1}{n_t}\sum_{m=1}^{n_t}\mathcal{L}\left(y^{(tm)},\left(\beta_0^{(t)}+\sum_{j=1}^{p}\beta_j^{(t)}x_j^{(tm)}\right)\right)+\lambda\boxed{\sum_{g=1}^{G}}\sqrt{p_g}\boxed{\boxed{\|B_g\|_F}}$$



where

$\beta^{(t)} \in \mathbb{R}^{p+1}$ is a task-specific vector of regression coefficients,

$\mathcal{L}$ is the loss function (quadratic or logistic regression),

B_g is a $T \times p_g$ matrix of the regression coefficients, across all tasks T for the SNPs of an LD-group g,

λ is the penalization parameter,

$\sqrt{p_g}$ scales the penalization factor according the group size.

Our goal is to **select LD-groups** associated with the phenotype across all tasks/populations, or specifically for some tasks/populations

Evaluation

- Validation using simulated data with predifined disease loci

Ability to detect false positives

- Comparison with the state-of-the-art methods:

1- Lasso after PCA adjustment* for population stratification

2- Group Lasso after PCA adjustment* for population stratification

3- Separate Lasso for each subpopulation

4- Separate group Lasso for each subpopulation

- Estimation of the stability of the selection: Pearson index[4]

- Computational time

*Include Top Principles Componenents (PCs) as covariates in regression models

GWAS challenges

Single marker analysis

Testing each SNP individually

Population stratification

Difference in allele frequecies between subpopulations

Linkage disequilibirum

Dependence relationship between two alleles at two different loci

Computational limitations

For complex methods:

- Memory errors
- Very slow

Lack of stability

Susceptibility to small perturbations in the data set

Methods

Multi-variate approach

Feature selection based on regularization

Multitask assignment

Assign a task for each subpopulation in a multitask framework

Hierarchical clustering[1]

Clustering of strongly correlated SNPs in LD-groups

Gap safe[2] screening rules

Eliminates useless coefficients: avoid memory errors and get more speed up

Stability selection[3]

Subsampling procedure to improve the stability

Data and results

Case-Control simulated data using GWAsimulator[5]

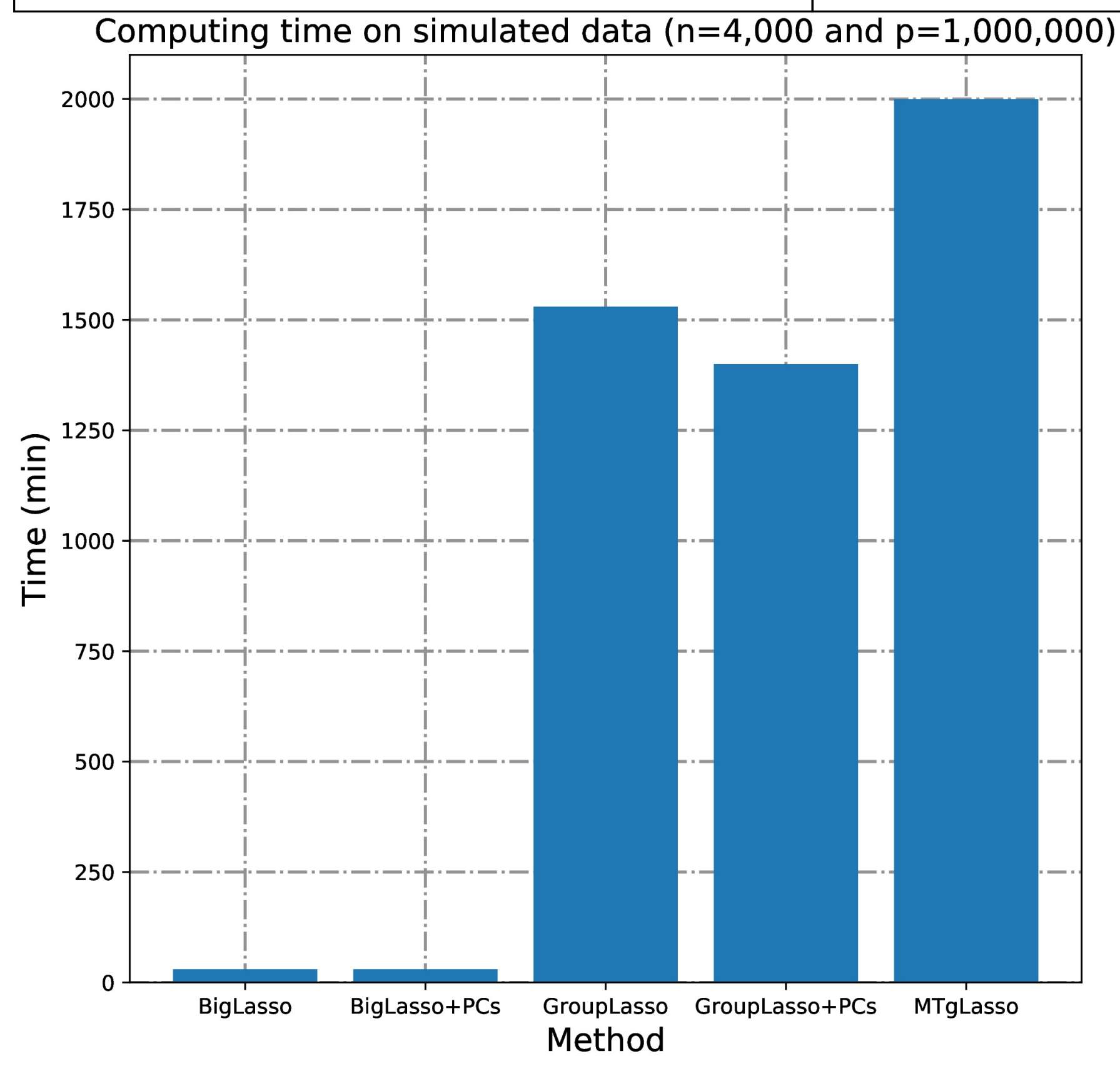
4,000 samples (European CEU and African YRI) x 1,000,000 SNPs

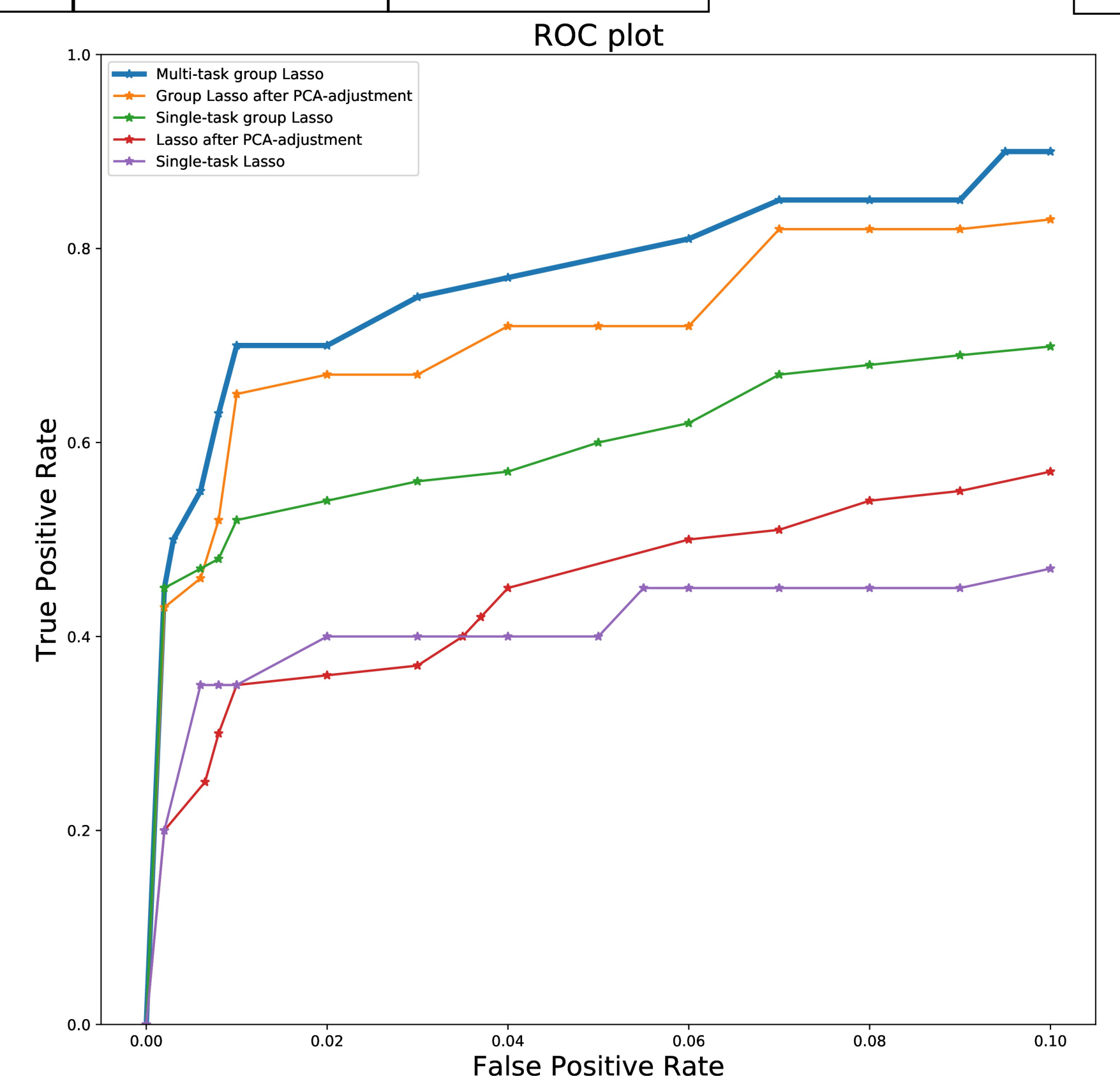
- Disease loci: chromosomes: 2, 12, 19, 21 and 22

2 (1,000-50,000 SNPs), 12 (10-40,000 SNPs), 19 (1000-50,000 SNPs), 21 (10-10,000 SNPs) and 22 (10-2000 SNPs)

- LD-groups: **35,792 groups**

	Number of selected features/groups	Stability index	Selection level
Multitask group Lasso (100 boostraps)	5,623	0.4912	LD-groups
Group Lasso after PCs adjustment	6,054	0.4134	LD-groups
Single task group Lasso	4,836	0.3398	LD-groups
Lasso after PCs adjustment	158,856	0.2368	Single-SNP
Single task Lasso	168,158	0.1742	Single-SNP





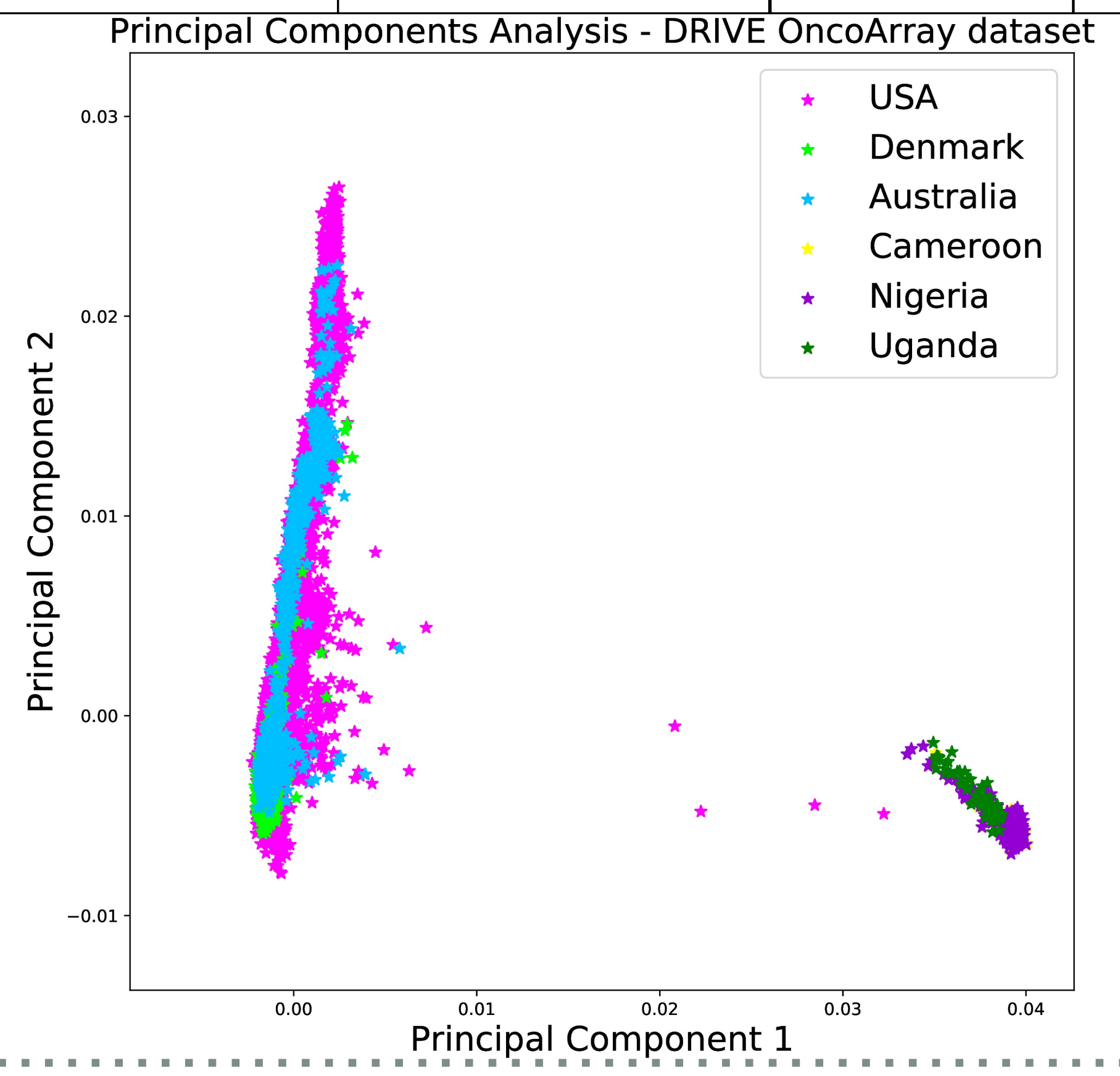
Real data: DRIVE Breast Cancer OncoArray[6]

28,282 samples x 313,237 SNPs

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

- LD-groups: **17,782 groups**

	Number of selected features/groups	Stability index	Selection level
Multitask group Lasso (100 boostraps)	62	0.4312	LD-groups
Group Lasso after PCs adjustment	59	0.3234	LD-groups
Single task group Lasso	58	0.2498	LD-groups
Lasso after PCs adjustment	874	0.2068	Single-SNP
Single task Lasso	789	0.1581	Single-SNP



References

[1] C. Ambroise et al., Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics, Algorithms Mol Biol (2019).

[2] E. Ndiaye et al., Gap safe screening rules for sparsity enforcing penalties, JMLR 18 (2017).

[3] N. Meinshausen and P. Bühlmann, Stability selection, J. R. Statist. Soc. B (2009).

[4] Nogueira et Al., On the Stability of Feature Selection Algorithms, JMLR 18 (2018).

[5] C. Li and M. Li, GWAsimulator: a rapid whole-genome simulation program, Bioinformatics (2008).

[6] DRIVE: "General Research Use" dataset in DRIVE Breast Cancer OncoArray Genotypes, available from dbGaP (study accession: phs001265/GRU), accessed under project #17707.

correspondence to:
asma.nouira@mines-paristech.fr