Pacific Symposium on Biocomputing (PSB) 2022

Human Intrigue: Meta Analysis Approaches for Big Questions with Big Data

Multitask group Lasso for Genome Wide Association Studies in diverse populations

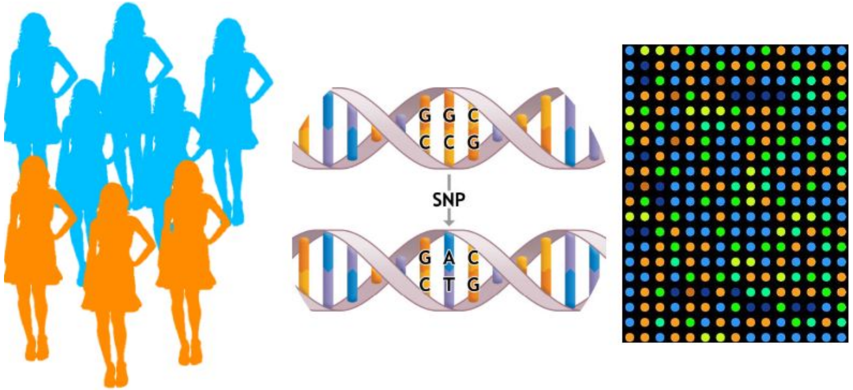
Asma Nouira

Chloé-Agathe Azencott

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

January 6th, 2022

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, height, etc.
  - Qualitative: Case-control study

1 From GWAS to Machine Learning

- **Single-marker analysis:**

Given a phenotype y, $\mathbf{X}$ is the genotype matrix:

For each feature $\mathbf{X}_j$, we fit a **single-predictor** equation $\mathbf{y}=\beta_0+\beta_j\mathbf{X}_j+\varepsilon \Rightarrow$ **p-value from a t-test** against an intercept-only model $H_0=\{\beta_j=0\}$.

- **Multi-variate approach:** Feature selection based on regularization

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

$$\underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\ \underbrace{\mathscr{L}\left(y,\beta X_j\right)}_{\text{loss}}+\lambda\underbrace{\sum_{j=1}^{p}\left|\beta_j\right|}_{\text{sparsity}}$$

where β is a $p\times 1$ vector corresponds to the SNP effects

Ω is the regularizer

λ is the penalization term

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

$$\underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\ \underbrace{\mathscr{L}\left(y,\beta X_j\right)}_{\text{loss}}+\lambda\underbrace{\sum_{g\in\mathscr{G}}\sqrt{p_g}\left\|\beta_g\right\|_2}_{\text{sparsity at the group level}}$$

where $\mathscr{G}$ is the set of groups

β_g is β restricted to the SNPs in g

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

- **Multi-task lasso:** allows fitting multiple regression problems jointly

$$\underset{\beta\in\mathbb{R}^{T\times p}}{\operatorname{argmin}}\ \underbrace{\sum_{t=1}^{T}\frac{1}{n_t}\sum_{m=1}^{n_t}\mathscr{L}\left(y^{(tm)},\left(\beta_0^{(t)}+\sum_{j=1}^{p}\beta_j^{(t)}X_j^{(tm)}\right)\right)}_{\text{loss}}+\lambda\underbrace{\sum_{j=0}^{p}\sum_{t=1}^{T}\left|\beta_j^{(t)}\right|}_{\text{task sharing}}$$

where T is the number of tasks to learn the training set

$\left\{\left(x_{tm},y_{tm}\right)\ \text{for}\ t=1..T\ \text{and}\ m=1..n_t\right\}$

2 Population stratification

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

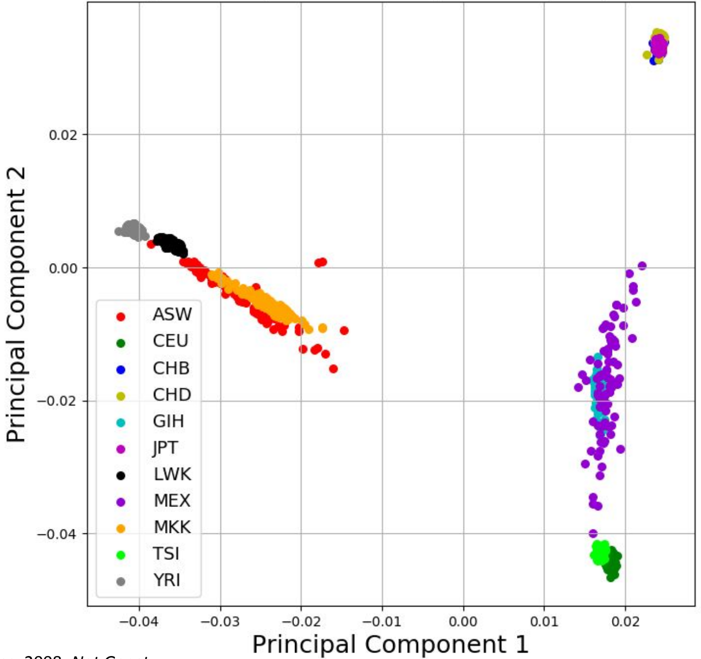
- **State-of-the art adjustment methods**
  - PCA-based methods

    Include Principal components (PCs) as covariates

    - Logistic Regression + Top PCs[1,2]
    - EIGENSTRAT[3]: multi-linear regression + 10 PCs
  - Linear mixed models

    Fast-LMM[4]

Principal Component Analysis - HapMap3 data



[1] Need et al.,A genome-wide investigation of snps and cnvs in schizophrenia. 2009, *PLoS Genet*.

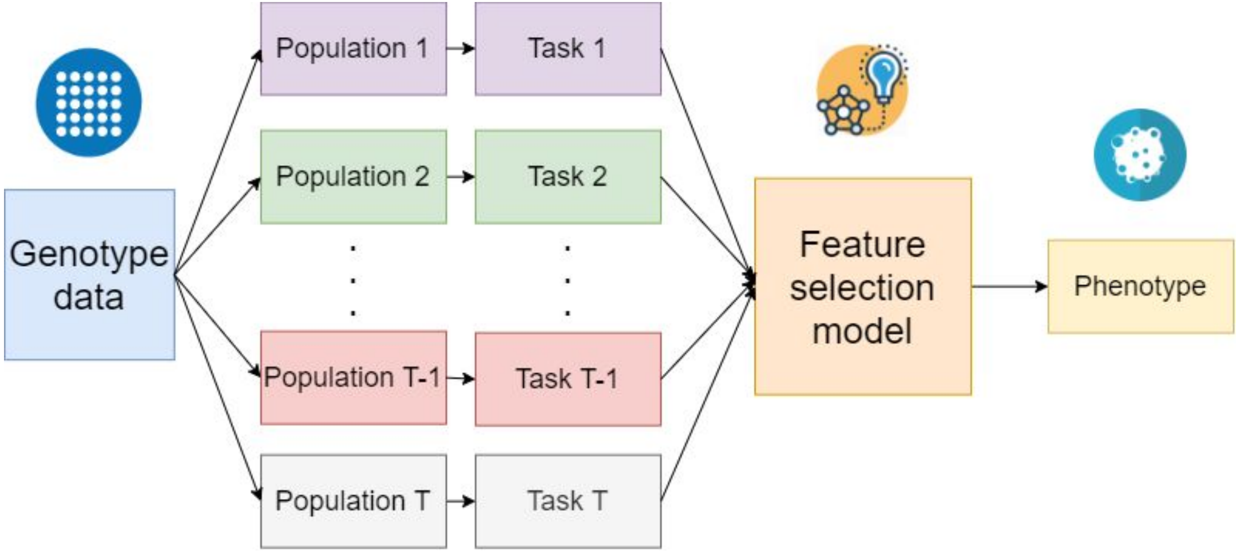
[2] Zeggini et al., Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. 2008, *Nat Genet.*

[3] Price et al.,Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.*

[4] Lippert et al., FaST linear mixed models for genome-wide association studies. 2011. *Nat Methods.*

2 Population stratification

- **Proposed adjustment method**: subpopulations assignment in multitask framework



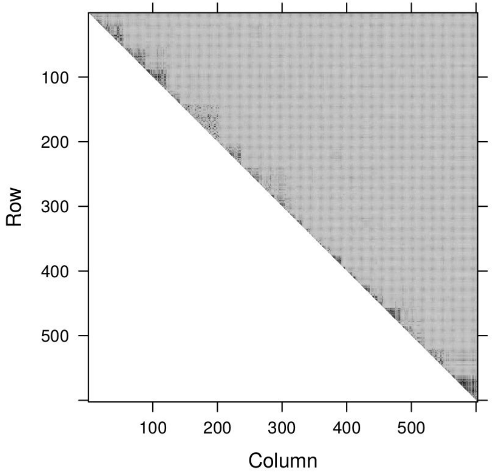
3 Linkage Disequilibrium groups clustering

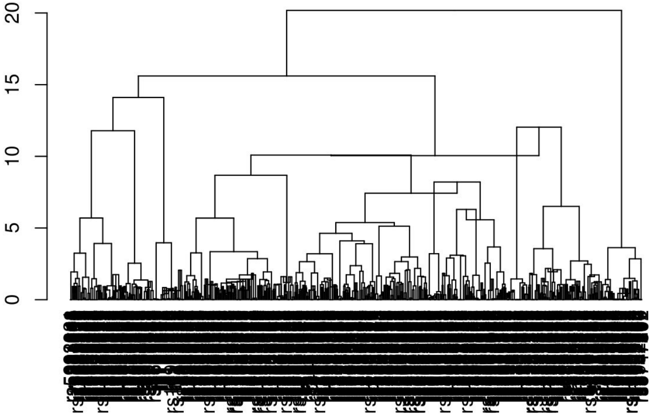
Linkage Disequilibrium (LD):

- Tendency of alleles to be transmitted together, more often than expected, by chance alone.
- Usually caused by nearby of genes in the same chromosome.

Hierarchical clustering approach[1]

Performing a **spatially-constrained hierarchical clustering**





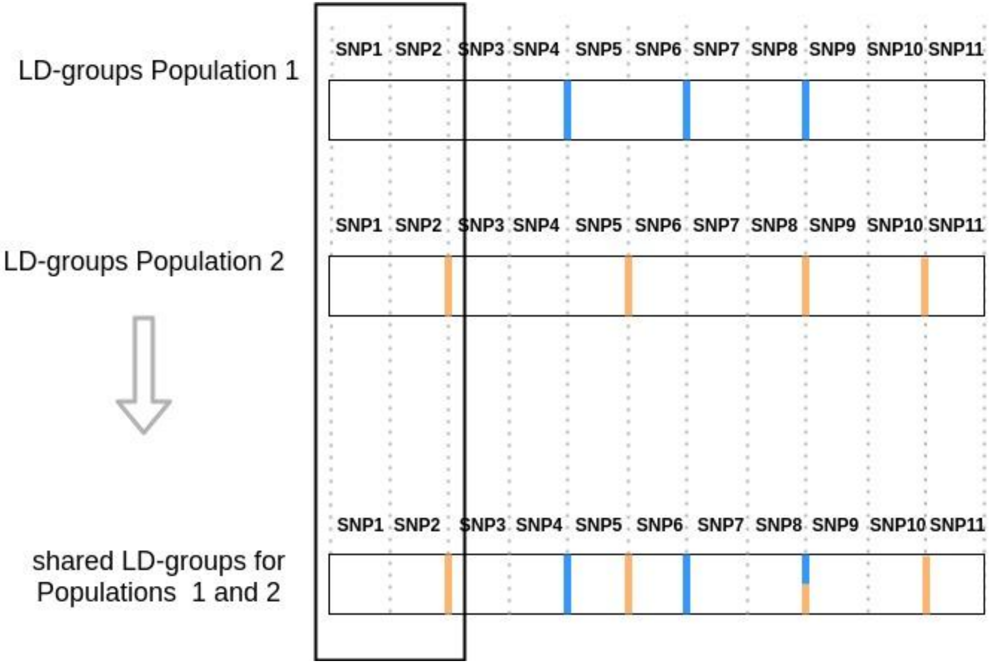
⇒ Selection at the **LD-group level** instead of single-SNP level.

[1] Ambroise et al.., Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. 2019. arXiv:1902.01596v1 [math.ST].

3 Linkage Disequilibrium groups clustering

- **Choice of LD-groups**

Linkage disequilibrium is different in different populations

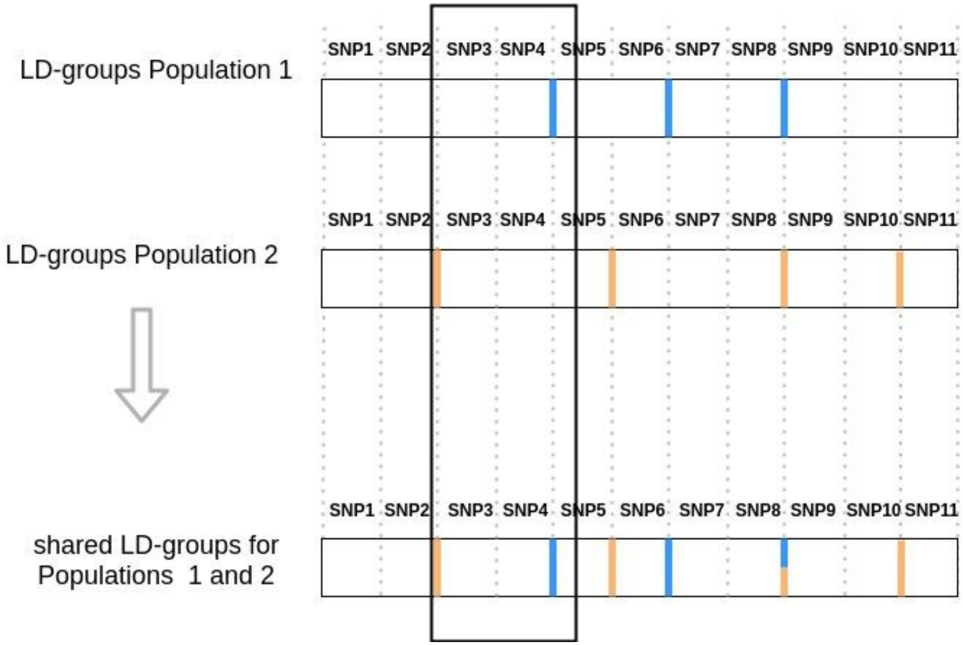


LD-groups Population 1
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11
LD-groups Population 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11
shared LD-groups for Populations 1 and 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11

3 Linkage Disequilibrium groups clustering

- **Choice of LD-groups**

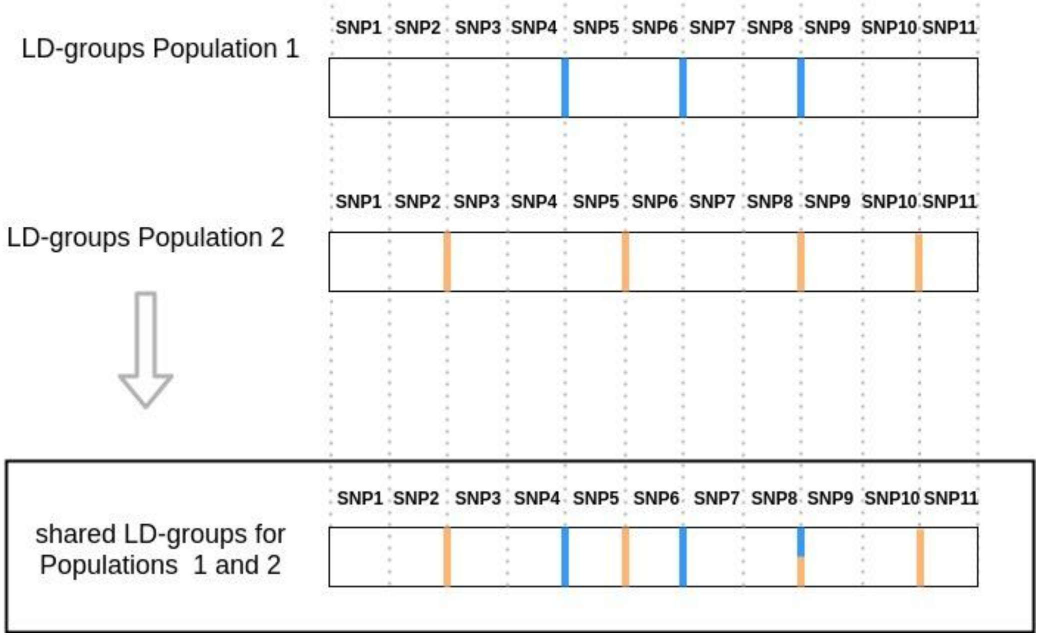
Linkage disequilibrium is different in different populations


LD-groups Population 1
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11
LD-groups Population 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11
shared LD-groups for Populations 1 and 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11

3 Linkage Disequilibrium groups clustering

- **Choice of LD-groups**

Linkage disequilibrium is different in different populations


LD-groups Population 1
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11
LD-groups Population 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11
shared LD-groups for Populations 1 and 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11

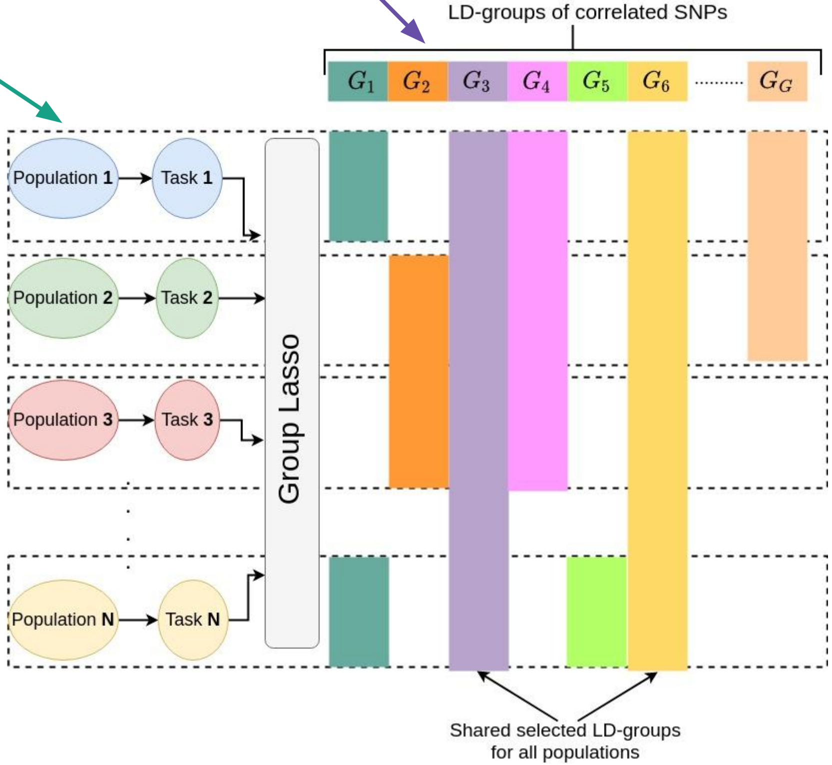
4 Multitask group Lasso for Genome Wide Association studies in diverse populations

Multitask group Lasso where **tasks** correspond to **subpopulations** and **groups** correspond to **LD-groups** of strongly correlated SNPs

$$\min_{B\in\mathbb{R}^{T\times(p+1)}} \boxed{\sum_{t=1}^{T}} \underbrace{\frac{1}{n_t}\sum_{m=1}^{n_t}\mathscr{L}\left(y^{(tm)},\left(\beta_0^{(t)}+\sum_{j=1}^{p}\beta_j^{(t)}x_j^{(tm)}\right)\right)}_{\text{loss for each task}} + \underbrace{\lambda\boxed{\sum_{g=1}^{G}}\sqrt{p_g}\boxed{\boxed{\|B_g\|_F}}}_{\text{sparsity at the LD-group level across tasks}}$$

where

- $\beta^{(t)}\in\mathbb{R}^{p+1}$ is a task-specific vector of regression coefficients
- $\mathscr{L}$ 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 LD-group g
- λ is the penalization parameter
- $\sqrt{p_g}$ scales the penalization factor according the group size



⇒ Selection of LD-groups associated with **the phenotype** across all tasks/populations, or specifically for some tasks/populations

5 Gap Safe screening rules

Gap Safe Screening rules[1]**:** eliminates features with associated coefficients are proved to be zero at the optimum in order to obtain **more speed up** and to **avoid memory errors**.

Ignoring some variables by exploiting geometric properties of the dual formulation of the following optimization problem:

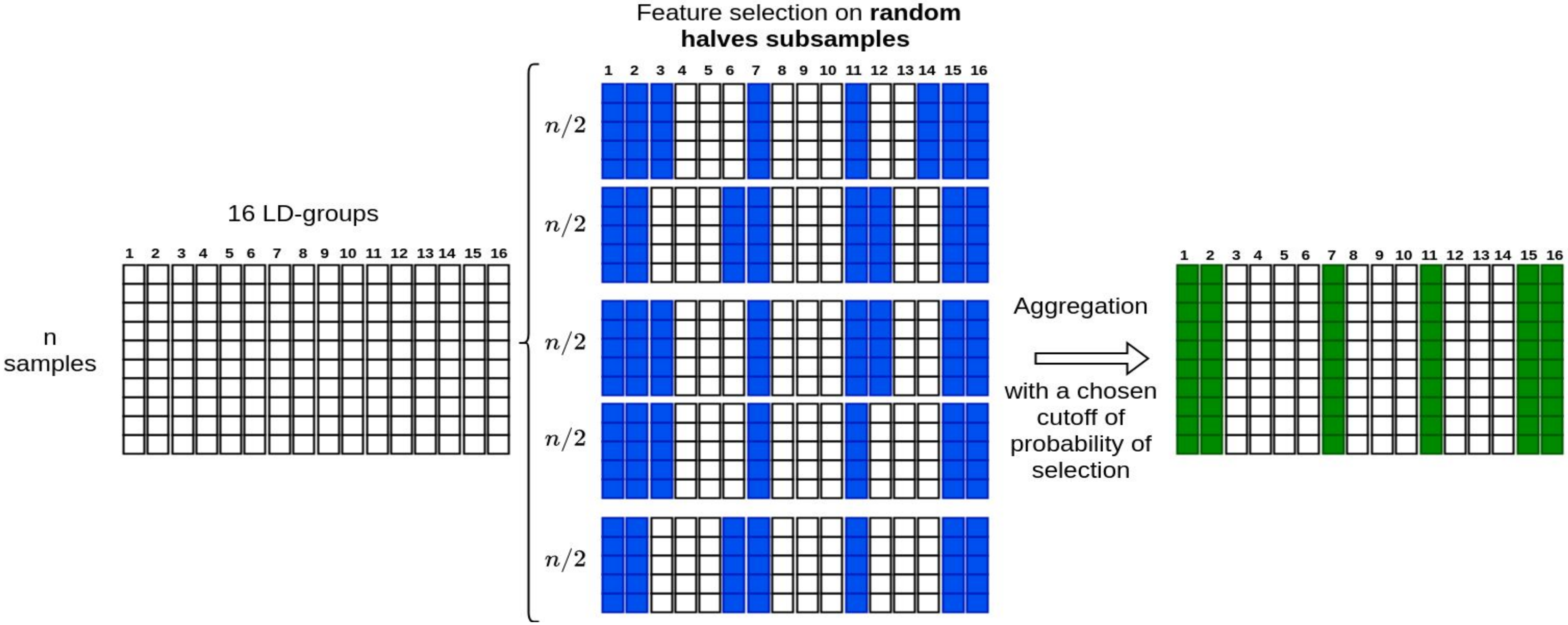
$$\widehat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} P_\lambda(\beta)\,, \text{ for } P_\lambda(\beta) := F(\beta) + \lambda\Omega(\beta) := \sum_{i=1}^{n} f_i\left(x_i^\top \beta\right) + \lambda\Omega(\beta)$$

where $f_i : \mathbb{R} \mapsto \mathbb{R}$ *are convex and differentiable functions and* $\Omega : \mathbb{R}^p \mapsto \mathbb{R}_+$ *is a group-decomposable norm:* $\Omega(\beta) = \sum_{g \in \mathcal{G}} \Omega_g(\beta_g)$
with Ω_g *a norm of* $\mathbb{R}^{n_g}$

[1] Ndiaye et al.,Gap Safe Screening Rules for Sparsity Enforcing Penalties. 2017, *Journal of Machine Learning Research 18*.

6 Stability Selection

Stability selection[1]**:** **bootstrap aggregation** procedure where feature selection is performed repeatedly on bootstrap subsamples, and the results of all repetitions are aggregated. It allows a **precise statement** of the significance of the selected features set and **reduce false positives**.


Feature selection on **random halves subsamples**
16 LD-groups
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
n samples
$n/2$
$n/2$
$n/2$
$n/2$
$n/2$
Aggregation
with a chosen cutoff of probability of selection

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

[1] Meinshausen et al,. Stability selection. 2010. *Journal of the Royal Statistical Society Series B-Statistical Methodology*.

Multitask group Lasso (MuGLasso) implementation

- Datasets

Realistic simulated data using GWAsimulator[1]

- **Dimension:** 4,000 samples x 1,400,000 SNPs
- **Populations:** 2000 European (CEU), 2000 African (YRI)
- **Phenotype:** 1100 CEU cases, 900 CEU controls, 900 YRI cases, 1100 controls.
- **Disease loci:** chromosomes: **2** (located on 1,000-50,000 SNPs), **12** (located on 10-40,000 SNPs), **19** (1000-50,000 SNPs), **21** (10-10,000 SNPs) and **22** (10-2000 SNPs)

Real data: DRIVE Breast Cancer OncoArray[2]

- **Dimension:** 28,281 samples x 528,620 SNPs
- **Phenotype:** 13,846 cases and 14,435 controls
- **Populations:** USA – Uganda – Nigeria – Cameroon – Australia – Denmark

[1] Li et al., GWAsimulator: a rapid whole-genome simulation program.2008.*Bioinformatics, Volume 24, Issue 1, 1 January 2008, Pages 140–142*.

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

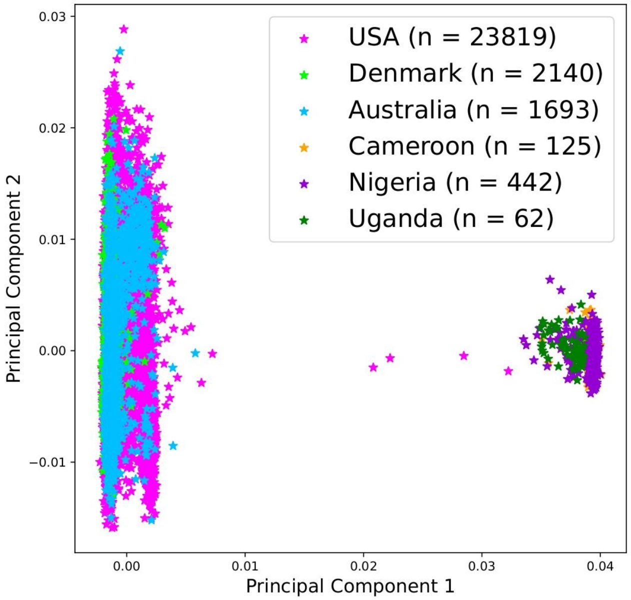
MuGLasso implementation

- Quality control and preprocessing
  - 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 of missing values: IMPUTE2

MuGLasso implementation

- Quality control and preprocessing
  - 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 of missing values: IMPUTE2
- Subpopulations definition

Assign subpopulations in multitask framework according to PCA patterns using **K-means clustering**.



POP1: USA, Denmark and Australia and **POP2:** Cameroon, Nigeria and Uganda

MuGLasso implementation

- Quality control and preprocessing

  - 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 of missing values: IMPUTE2

- Evaluation

  - **Validation using simulated data**

  Generate simulations with specified multi locus disease model in specified regions

  ⇒ Compute **false positives rate**

  - **Estimation of the stability of the selection** [1,2]

  $$Stability = \widehat{\Phi}(s_1, s_2, \dots s_M) = \frac{1}{M(M-1)} \sum_{i} \sum_{j \neq i} sim(s_i, s_j)$$

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

[1] Kuncheva et Al., A stability index for feature selection. 2008, *IASTED International Conference on Artificial Intelligence and Applications*.

[2] Nogueira et Al., On the Stability of Feature Selection Algorithms. 2018, *Journal of Machine Learning Reasearch 18.*

MuGLasso implementation

- Quality control and preprocessing
  - 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 of missing values: IMPUTE2
- Evaluation
  - **Validation using simulated data**

    Generate simulations with specified multi locus disease model in specified regions.

    ⇒ Compute **false positives rate**
  - **Estimation of the stability of the selection**[1,2]

$$Stability = \widehat{\Phi}(s_1, s_2, \dots s_M) = \frac{1}{M(M-1)} \sum_{i} \sum_{j \neq i} sim(s_i, s_j)$$

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

1. **Adjusted Lasso: after PCA adjustment** for population stratification at the **SNP level**
2. **Adjusted group Lasso: after PCA adjustment** for population stratification at **LD-groups level**
3. **Stratified group Lasso** for each subpopulation at **LD-groups level**
4. **Stratified Lasso** for each subpopulation at **the SNP level**
5. **Adjusted GWAS:** Classic GWAS after PCA adjustment

[1] Kuncheva et Al., A stability index for feature selection. 2008, *IASTED International Conference on Artificial Intelligence and Applications*.

[2] Nogueira et Al., On the Stability of Feature Selection Algorithms. 2018, *Journal of Machine Learning Reasearch 18.*

MuGLasso implementation

- Quality control and preprocessing
  - 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 of missing values: IMPUTE2

- Evaluation
  - **Validation using simulated data**

    Generate simulations with specified multi locus disease model in specified regions.

    ⇒ Compute **false positives rate**

  - **Estimation of the stability of the selection**[1,2]

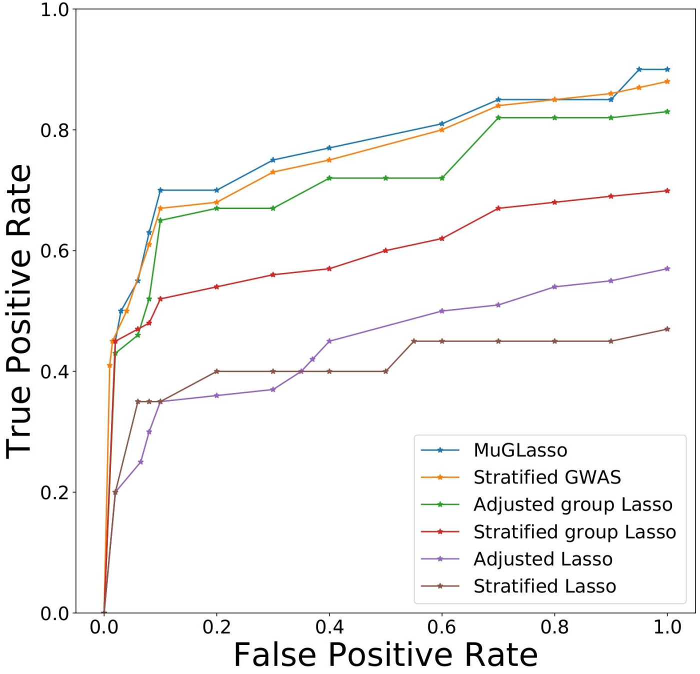
$$Stability = \widehat{\Phi}(s_1, s_2, \dots s_M) = \frac{1}{M(M-1)} \sum_{i} \sum_{j \neq i} sim(s_i, s_j)$$

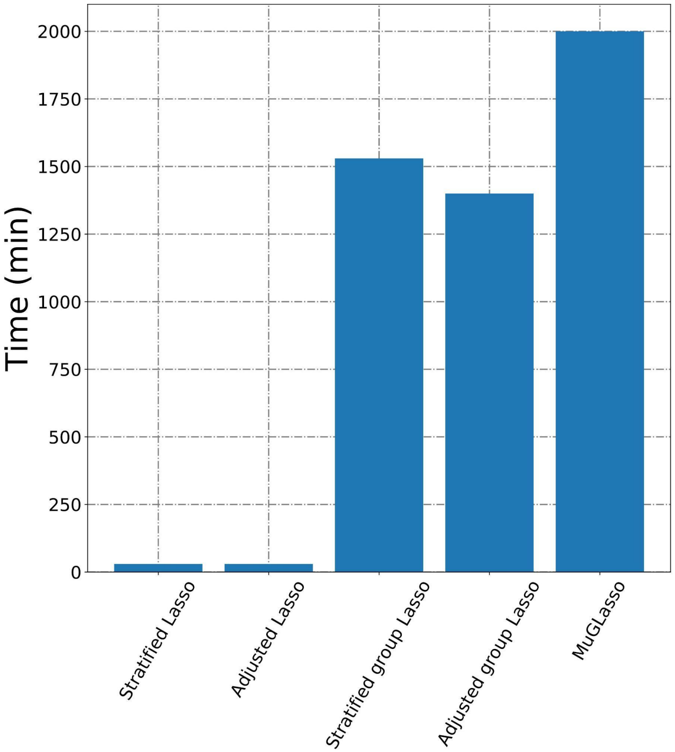
  - **Comparison with the state-of-the art methods**
  - **Computational time**

[1] Kuncheva et Al., A stability index for feature selection. 2008, *IASTED International Conference on Artificial Intelligence and Applications*.

[2] Nogueira et Al., On the Stability of Feature Selection Algorithms. 2018, *Journal of Machine Learning Reasearch 18.*

MuGLasso outperforms the state-of-the-art methods on simulated data

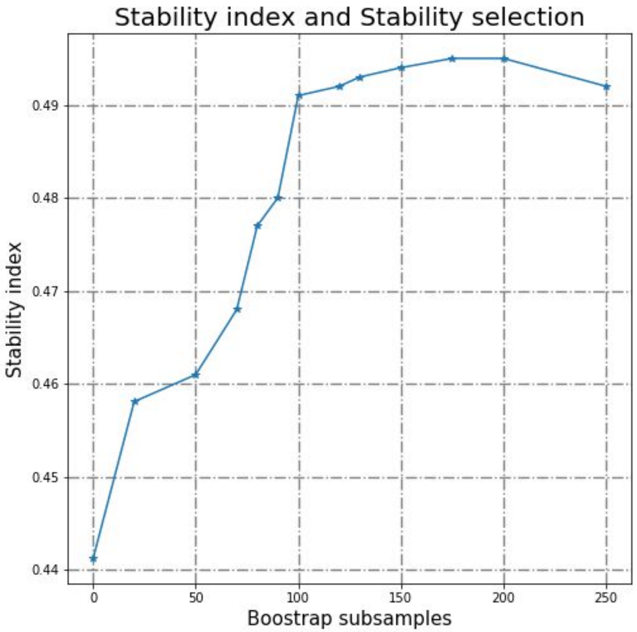




MuGLasso improve the stability of the selection on DRIVE data

Real data: DRIVE Breast Cancer OncoArray[1]: n=28,282 ; p=313,237 ; LD-groups = 17,782

Methods	Number of selected LD-groups	Stability index	Selection level
MuGLasso (100 boostraps)	62	0.4312	LD-groups
Adjusted group Lasso	59	0.3234	LD-groups
Stratified group Lasso	58	0.2498	LD-groups
Adjusted Lasso	41	0.2068	Single-SNP
Stratified Lasso	38	0.1581	Single-SNP
Adjusted GWAS	16	-	Single-SNP



⇒ The feature selection at the LD-groups level alleviate the curse of dimensionality and the lack of stability.

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

Breast cancer risk loci detected by MuGLasso on DRIVE

- All SNPs/genes found by adjusted GWAS were also selected by MuGLasso.
- **9 genes** were discovered by adjusted GWAS and 32 genes were discovered by MuGLasso.
- **17 of 32 genes** had been previously identified by a meta-GWAS containing the DRIVE data.
- **7 genes** were found in the literature prior evidence of relationship with breast cancer risk or tumor growth.

Genes found by adjusted GWAS	**ITPR1, MRPS30, MAP3K1, SETD9, MIER3, EBF1, FGFR2, TOX3, MKL1**
Genes found by MuGLasso	**ITPR1, MRPS30, MAP3K1, SETD9, MIER3, EBF1, FGFR2, TOX3, MKL1**, **ADSL**, **ASTN2**, C7orf73, **CACNA1I**, **CCDC170**, **CCDC91**, CCSER1, CD2AP, **CDYL2**, **DIRC3**, **ELL**, **ESR1**, **FTO**, **GRHL1**, HK1, HRSP12, **KCNU1**, LUC7L3, MED21, **NEK10**, **NUP205**, **PAX9**, **POP1**,**PPFIBP1**, **PTHLH**, REP15, **SGSM3**, **SSBP4**, **TGFBR2**, **TNRC6B**, **ZMIZ1**, **ZNF365**

Genes discovered for subpopulation POP1	**ESR1**, **SGSM3**, MED21, REP15
Genes discovered for subpopulation POP2	**DIRC3**, LUC7L3

POP1: USA, Denmark and Australia and **POP2:** Cameroon, Nigeria and Uganda

Conclusion and future work

01

Multi-variate approach

- Consider the effect of SNPs jointly

02

Multitask assignment

- Address population stratification by assigning an input task to each subpopulation

03

LD-groups clustering

- Address high correlation between SNPs
- Alleviate the curse of dimensionality

04

Safe screening rules

- Speed up the solvers and avoid memory errors in high scale

05

Stability selection

- Improve the stability of the feature selection using subsampling procedure

Paper: Multitask group Lasso for Genome Wide Association Studies in diverse populations, PSB 2022, https://www.biorxiv.org/content/10.1101/2021.08.02.454499
Code: https://github.com/asmanouira/MuGLasso_GWAS
Contact: asma.nouira@mines-paristech.fr

Conclusion and future work

01 Multi-variate approach

- Consider the effect of SNPs jointly

02 Multitask assignment

- Address population stratification by assigning an input task to each subpopulation

03 LD-groups clustering

- Address high correlation between SNPs
- Alleviate the curse of dimensionality

04 Safe screening rules

- Speed up the solvers and avoid memory errors in high scale

05 Stability selection

- Improve the stability of the feature selection using subsampling procedure

Future work

Sparse MuGLasso (SMuG Lasso)

- Add an L1-norm sparsity penalty to improve the LD-groups selection for specific-populations
- Extend MuGLasso to general applications

Paper: Multitask group Lasso for Genome Wide Association Studies in diverse populations, PSB 2022, https://www.biorxiv.org/content/10.1101/2021.08.02.454499
Code: https://github.com/asmanouira/MuGLasso_GWAS
Contact: asma.nouira@mines-paristech.fr

Acknowledgements

- CBIO (Mines ParisTech)

- GWAS team

- U900 (Institut Curie)

- This work was supported in part by Agence Nationale de la Recherche (ANR-18-CE45-0021-01 and ANR19-P3IA-0001)

THANK YOU!

Gap Safe screening rules

Gap Safe Screening rules[1]**:** eliminates features with associated coefficients are proved to be zero at the optimum in order to obtain **more speed up** and to **avoid memory errors**.

Ignoring some variables by exploiting geometric properties of the dual formulation of the following optimization problem:

$$\widehat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\, P_\lambda(\beta)\,,\ \text{for } P_\lambda(\beta) := F(\beta) + \lambda\Omega(\beta) := \sum_{i=1}^{n} f_i\left(x_i^\top \beta\right) + \lambda\Omega(\beta)$$

where $f_i : \mathbb{R} \mapsto \mathbb{R}$ *are convex and differentiable functions and* $\Omega : \mathbb{R}^p \mapsto \mathbb{R}_+$ *is a group-decomposable norm:* $\Omega(\beta) = \sum_{g \in \mathcal{G}} \Omega_g(\beta_g)$ *with* Ω_g *a norm of* $\mathbb{R}^{n_g}$

For group Lasso: the data fitting term is $F(\beta) = \frac{\mathcal{L}\left(y, \beta X_j\right)}{2}$,

The *L1/L2-norm* is defined by $\Omega(\beta) = \Omega_w(\beta)$:

$$\Omega_w(\beta) := \sum_{g \in \mathcal{G}} w_g \|\beta_g\|_2 \quad \text{and} \quad \Omega_w^D(\xi) := \max_{g \in \mathcal{G}} \frac{\|\xi_g\|_2}{w_g}$$

where $w = \left(w_g\right)_{g \in \mathcal{G}}$ *are weights satisfying* $w_g > 0$ *for all* $g \in \mathcal{G}$ *and* $\Omega_w^D(\xi)$ *is the dual norm along the regularization path.*

[1] Ndiaye et al.,Gap Safe Screening Rules for Sparsity Enforcing Penalties. 2017, *Journal of Machine Learning Research 18*.

Stability Selection

Stability selection[1]**: bootstrap aggregation** procedure where feature selection is performed repeatedly on bootstrap subsamples, and the results of all repetitions are aggregated. It allows a **precise statement** of the significance of the selected features set and **reduce false positives**.

Procedure:

- Identify $S=\left\{k:\beta_k \neq 0\right\}$ a set of non-zero inputs of a sparse parameter vector β of observed data (X, y)
- Feature selection is performed on randomly $\left|\mathrm{I}\right|=\frac{n}{2}$ of observations, where $\mathrm{I} \subset \{1, \ldots, n\}$
- Selection Path: Probability of the selection of a feature $k \in \{1, \ldots, p\}$

  $\pi_k^\lambda = Pr^*\left[k \in \widehat{S}^\lambda(I)\right]$, where $\widehat{S}^\lambda(\mathrm{I}) \subset \{1, \ldots, p\}$ denotes the selected features by a subsample I

  ⇒ Captures random selection within feature selection algorithms

- For a chosen cut-off $\frac{1}{2} \leq \pi_{thre} \leq 1$, the set of stable features is:

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

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

[1] Meinshausen et al,. Stability selection. 2010. *Journal of the Royal Statistical Society Series B-Statistical Methodology*.