Bacterial GWAS using machine learning

T.Tien MAI

Oslo Centre for Biostatistics and Epidemiology (OCBE),
Department of Biostatistics,

UiO : University of Oslo

October 28, 2019

Overview

- A research direction in the Jukka Corander's group at UiO.

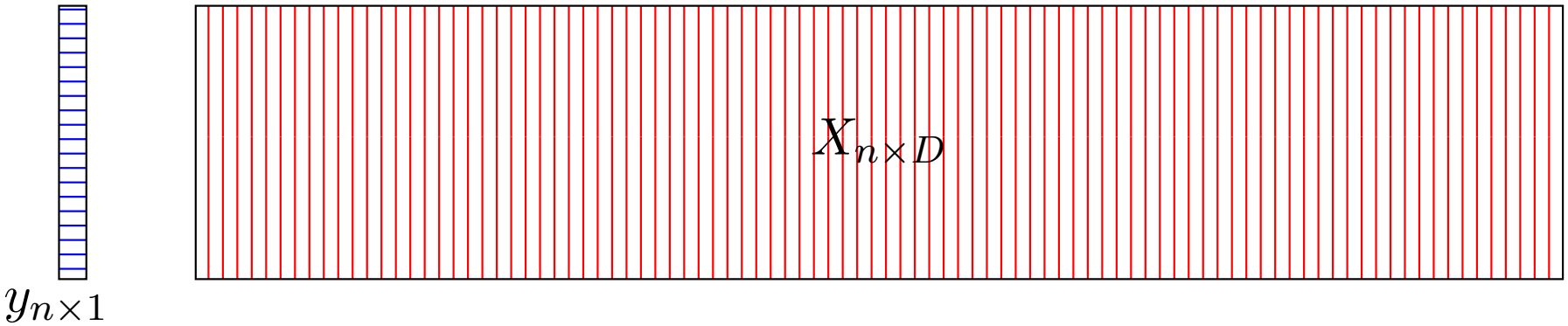
- My background is a PhD in stats: PAC-Bayesian analysis for low-rank matrices.

The GWAS problem

Given

- a phenotype (binary/cont.) $y_{n\times 1}$ response of n samples,
- a genetic data $X_{n\times D}$ (biomakers, e.g SNPs), with $n \ll D$.

Goal: detect which genetic variants $X_{.j}$ are importantly relevant to y.



The most popular approach is using marginal single test for each $X_{.j}$.

Challenging with bacterial data

- the design matrix is with linkage disequilibrium (LD): *highly correlated, cluster structures in* X.
- X is a binary matrix (single allele).

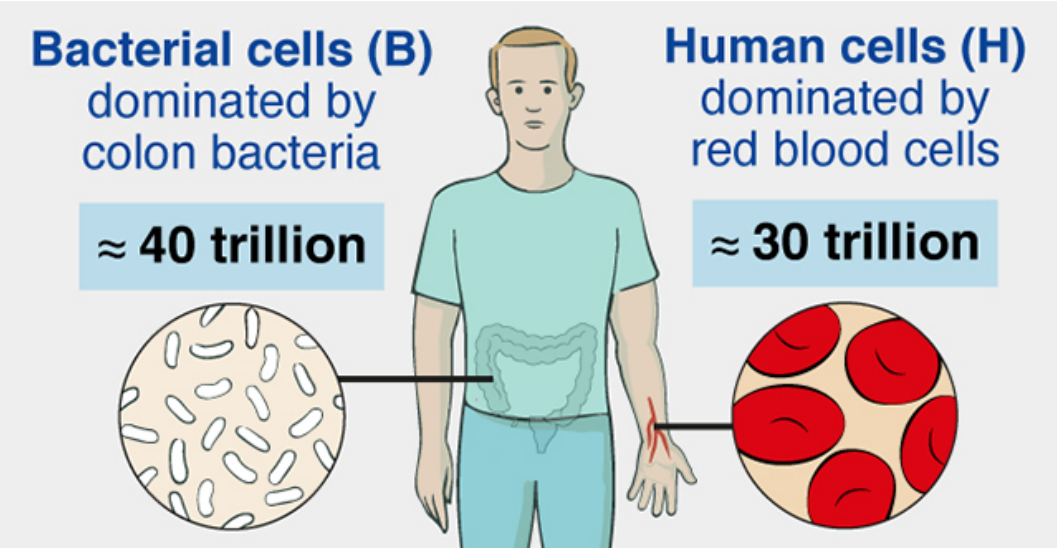


Figure: https://www.weizmann-usa.org/news-media/news-releases/germs-humans-and-numbers

Univariate approach for bacterial GWAS

Bacterial GWAS is done by testing

$$H_0 : \beta_j = 0$$

in the univariate marginal regression

$$y = f(\beta_0 + X_j\beta_j + \gamma C + \varepsilon_j), \quad j = 1, \ldots, D$$

where C is the "population structure correction".

Lees, J. A., et al. "Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes." *Nature communications 7 (2016): 12797.*

Multivariate approach using Elastic Net

Jointly selection approach does not need population correction and can improve the power when the sample size increase.

$$\min_{\beta} \left\{ -log.likelihood(y, X\beta) + \lambda \left[0.5(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \right\}$$

Elastic Net inherits both interesting features of ℓ_1 and ℓ_2 norm:

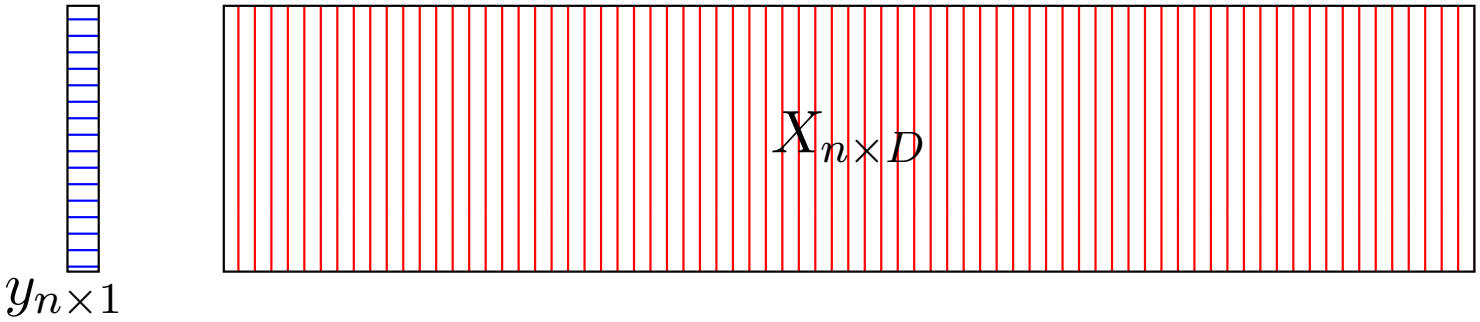
- ℓ_1 generates a sparse model ($\|\beta\|_0 := s \leq n$),
- ℓ_2 removes the limitation on the number of selected variables, encourages grouping effect (correlation) and stabilizes the ℓ_1 regularization path.

Problem: can not run if D is too large !!!

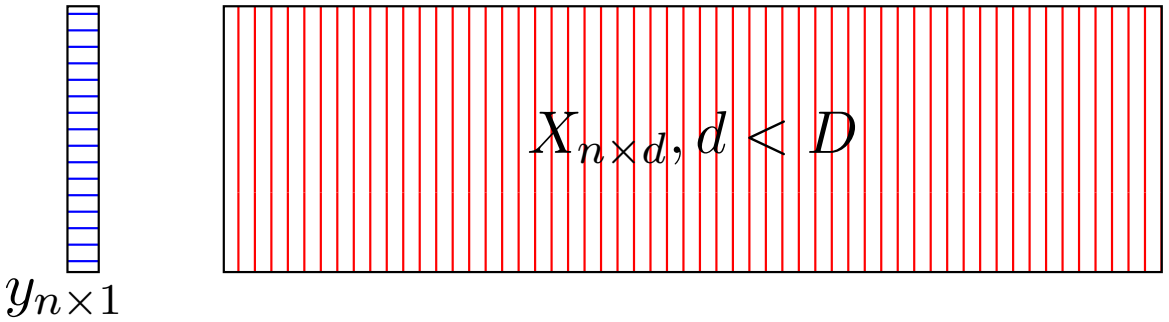
glmnet R package

Friedman, Hastie,& Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent". *Journal of statistical software, 33(1), 1.*

Screening to reduce irrelevant features



Remove all X_j whose the sample correlation with y are smaller than a threshold.



Fan & Lv (2008). "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849-911.*

Elastic Net with screening

Procedure. *Enet with pre-selection screening*

1. *Calculate the sample correlation between y and $X_{.j}$, as j varies across all predictors.*
2. *Retain the set B of predictors whose the sample correlation are bigger than the first quantile of all of the sample correlations.*
3. *Run the elastic net to select the relevant predictors from the set B.*

Lees, John A., et al. "pyseer: a comprehensive tool for microbial pangenome-wide association studies." *Bioinformatics 34.24 (2018): 4310-4312.*

Numerical results

Maela data: 3000 samples, 121014 SNPs (after cleaning), simulated phenotypes using GCTA.

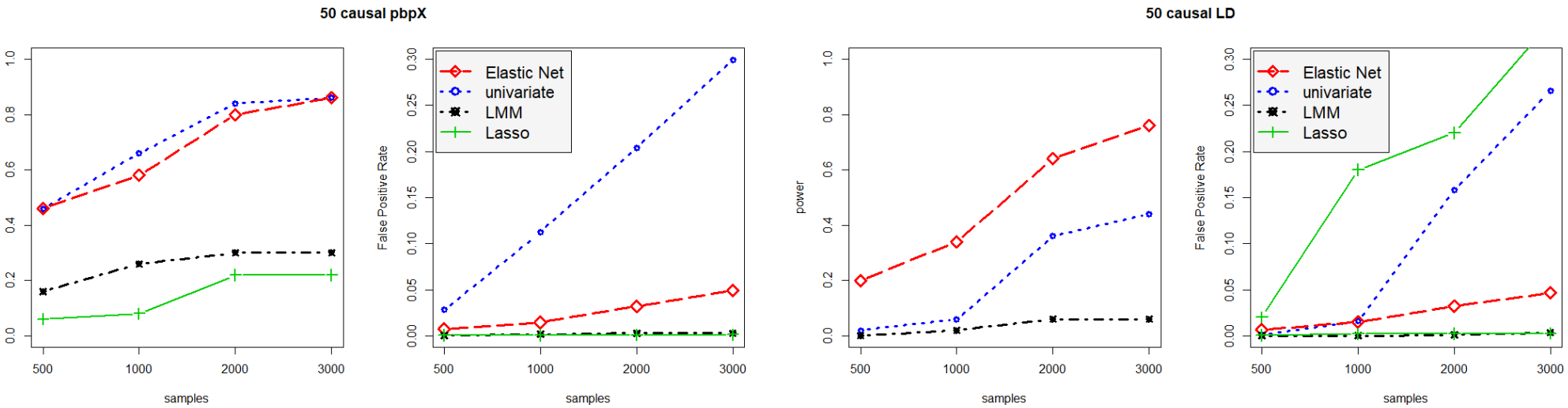


Figure: Higher in power (left) is better, lower in False Positive Rate (right) is better.

Heritability estimation

In linear model

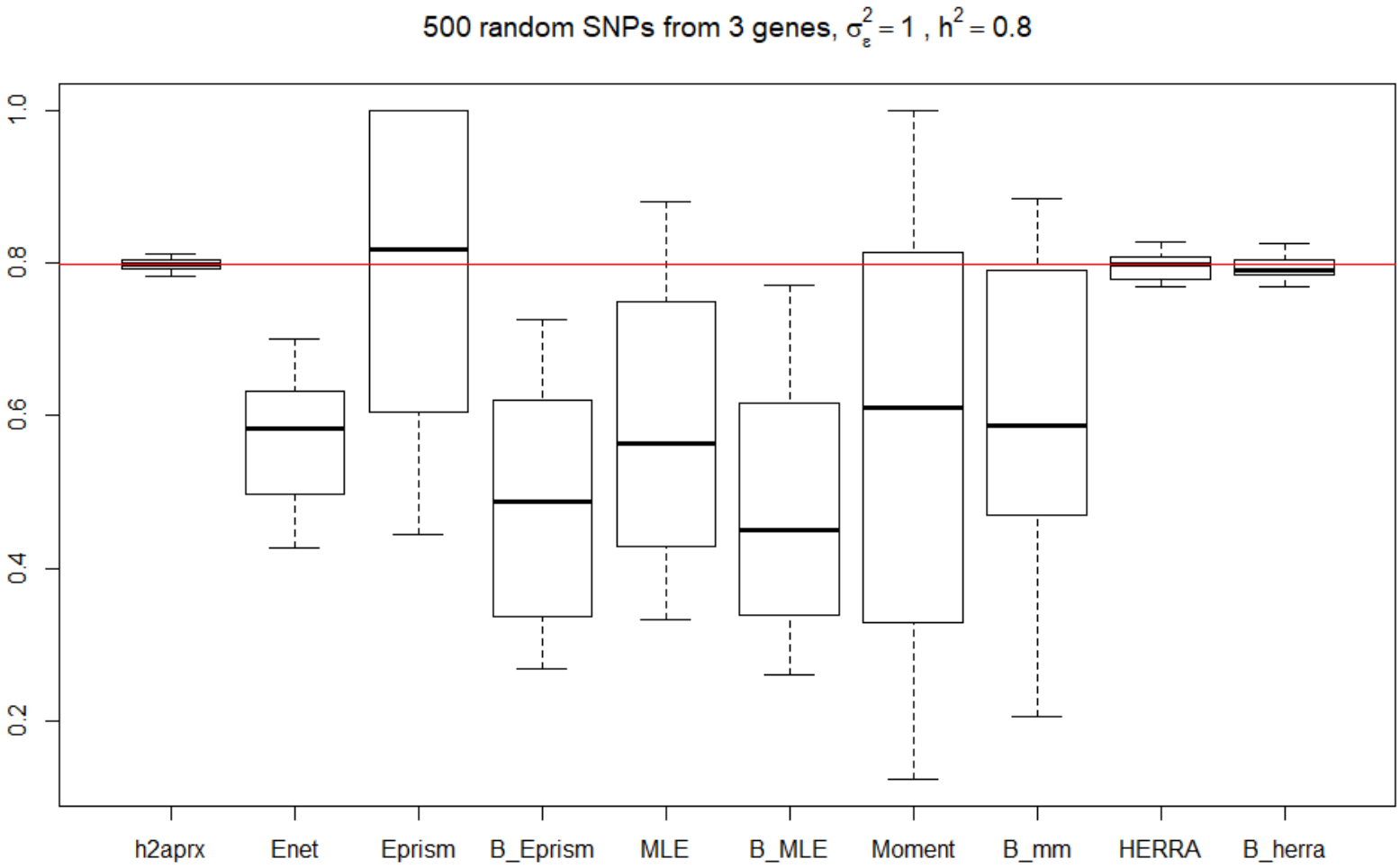
$$y_i = X_{i\times p}\beta + \varepsilon_i, i = 1, \ldots, n$$

where $X_{i\cdot} \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$ and are independent of $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

$$\mathrm{Var}(y_i) = \mathrm{Var}(X_{i\cdot}\beta) + \sigma_\varepsilon^2 = \beta^\top \Sigma \beta + \sigma_\varepsilon^2.$$

We are interested in estimating heritability for y defined as

$$\boxed{h^2 = \frac{\beta^\top \Sigma \beta}{\beta^\top \Sigma \beta + \sigma_\varepsilon^2}} = \frac{\beta^\top \Sigma \beta / \sigma_\varepsilon^2}{\beta^\top \Sigma \beta / \sigma_\varepsilon^2 + 1} = 1 - \frac{\sigma_\varepsilon^2}{\mathrm{Var}(y)}.$$



The Tien Mai and Jukka Corander (2019) "Boosting heritability: estimating the genetic component of phenotypic variation with multiple sample splitting." *arXiv 1910.11743*

Thank you!