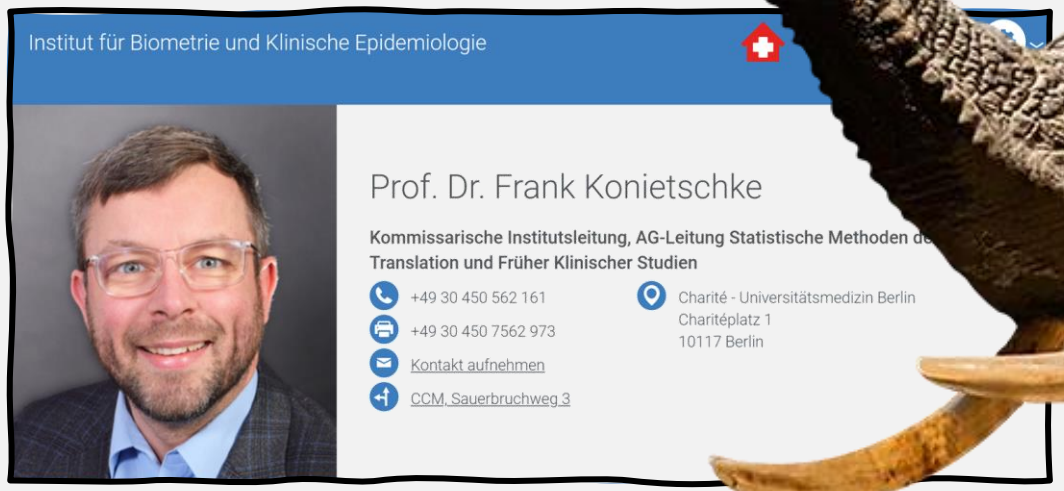
(Artificial) intelligent and non-intelligent methods to explore multidimensional genomic data

Joanna Szyda & The THETA Biostatistic Group

Different worlds



Small datasets

Huge datasets

Different worlds

- Assymptotic test properties (type I, type II errors)
- Missing data patterns
- Do not remove observations

- Redundant, correlated information
- Fitting patterns to „noise”
- Input, interim, final data storage problem
- Statistical modeling problem
- Computationally intensive
- Extensive data preprocessing

Outline

multidimensional

genomic

AI

data

Outline

multidimensional

genomic

AI

data

1. cattle
2. humans

Outline

multidimensional

genomic

1. SNPs from WGS

AI

data

1. cattle
2. humans

Outline

multidimensional

genomic

1. SNPs from WGS

AI

1. DL based classification

data

1. cattle
2. humans

Outline

Dimensionality reduction by feature selection

1. basic
2. tagSNP
3. 1D-SRA
4. MD-SRA

genomic

1. SNPs from WGS

AI

1. DL based classification

data

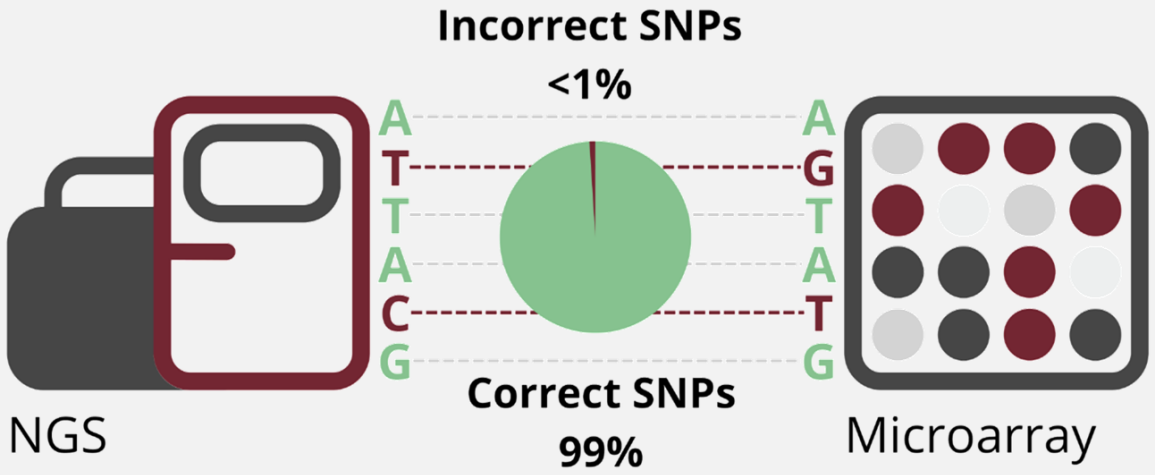
1. cattle
2. humans

basic approach to feature selection

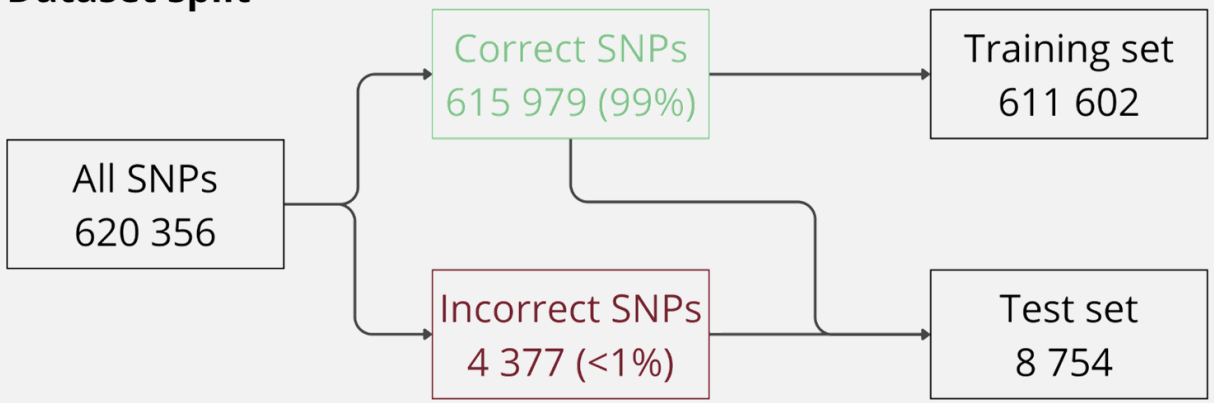
PCA

PCA

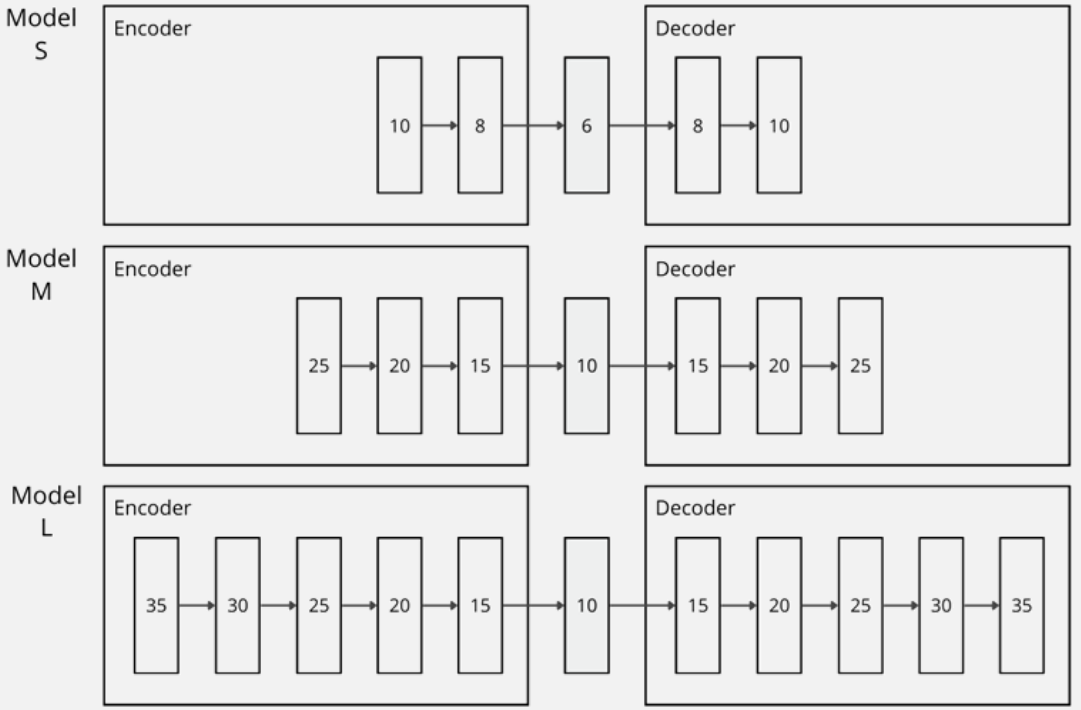
- SNP genotypes of 20 HF cows



Dataset split



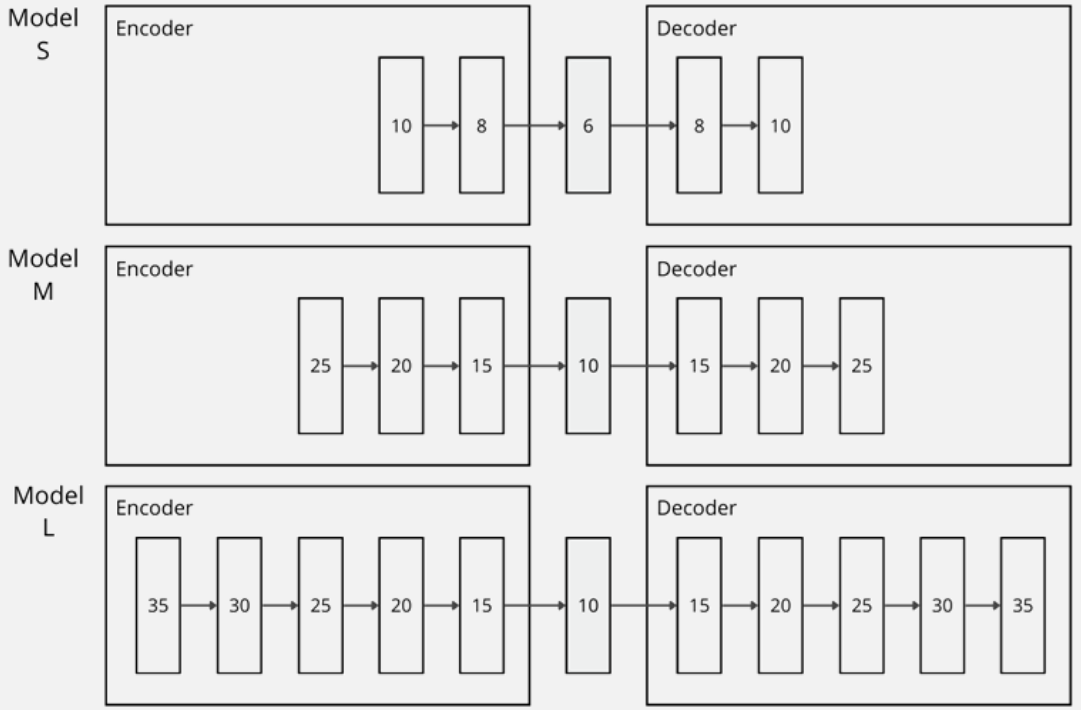
- 3 AutoEncoder architectures
- Classification of SNPs
  - correct
  - incorrect



PCA

- **Considered features**
  - Reference allele
  - Alternative allele
  - Sequencing depth
  - SNP genotype quality
  - Sequence context
    - 4 nucleotides downstream
    - 4 nucleotides upstream

- 3 AutoEncoder architectures
- Classification of SNPs
  - correct
  - incorrect



PCA

- **Considered features**
  - Reference allele
  - Alternative allele
  - Sequencing depth
  - SNP genotype quality
  - Sequence context
    - 4 nucleotides downstream
    - 4 nucleotides upstream

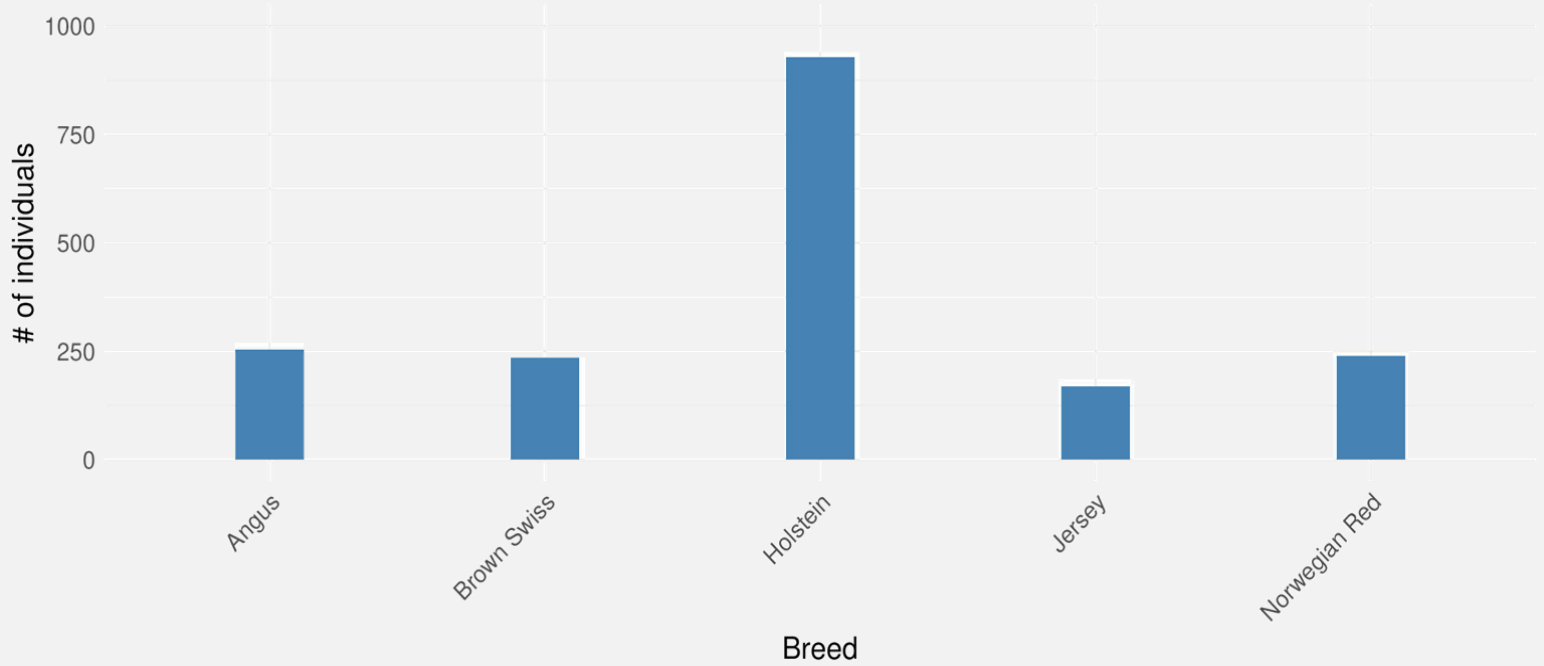
- Features measured on different scales

1. Original encoding → poor classification
2. **PCA**

feature selection based on SNP tagging, 1D-SRA, MD-SRA

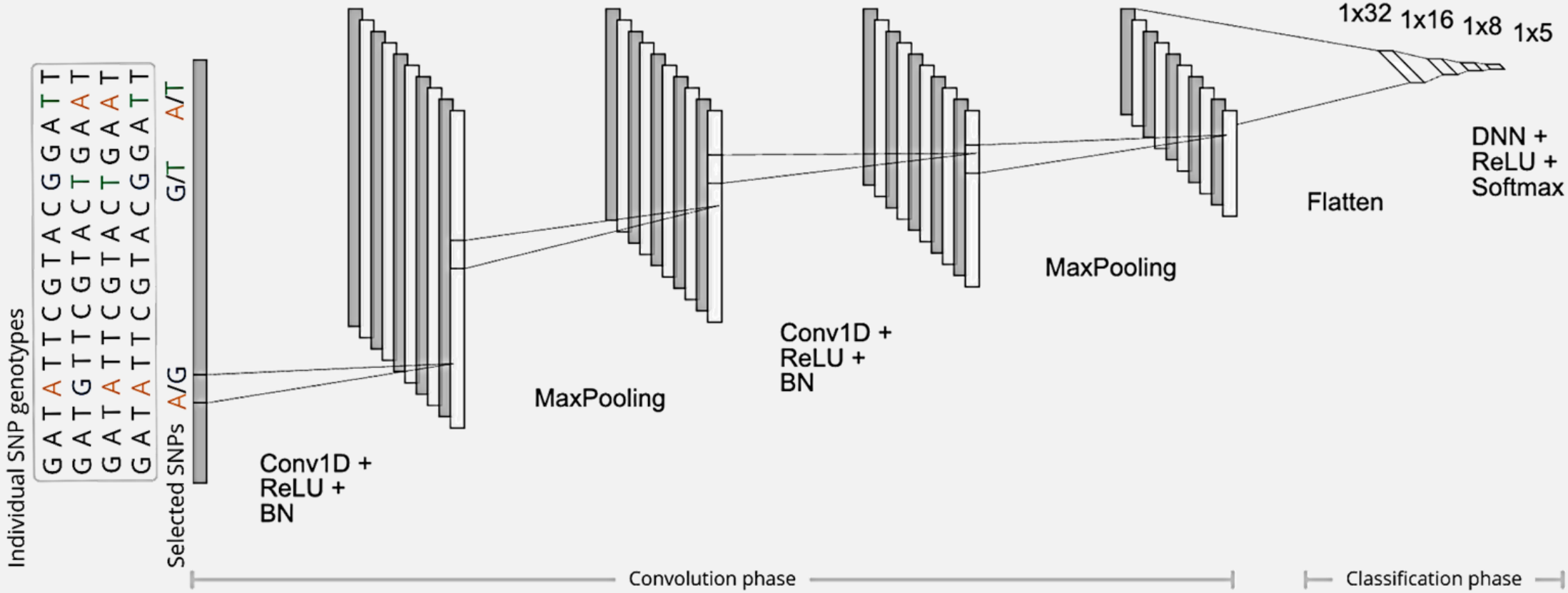
material

- 1000 Bull Genomes Project, Run 9
- Full data → 5,063 bulls with WGS → 33,595,340 SNPs
- Edited data → 1,825 bulls, 5 breeds → 11,915,233 SNPs = features



goal

- Classification of bulls to breeds → multi-class
- DL based on CNNs and DNNs



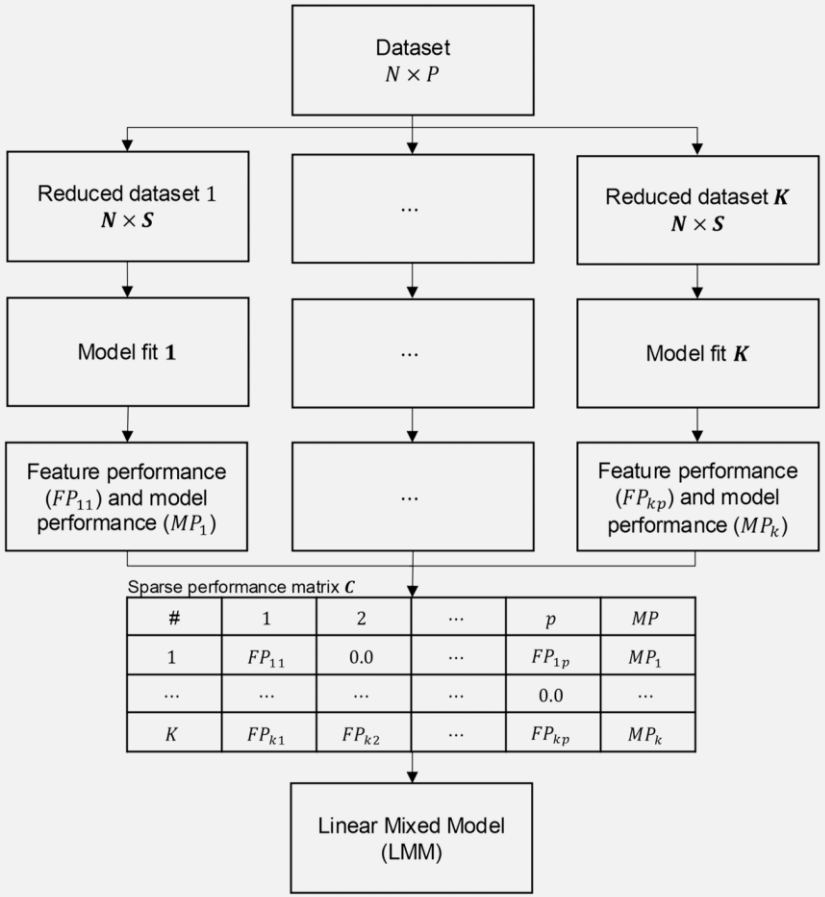
SNP tagging

- Selection of representative SNPs based on pairwise LD
- PLINK 1.9
  - LD (R^2) threshold → 0.5
  - windowsize → 100 000 bp
  - stepsize → 1 000 bp

1D - SRA

Supervised Rank Aggregation

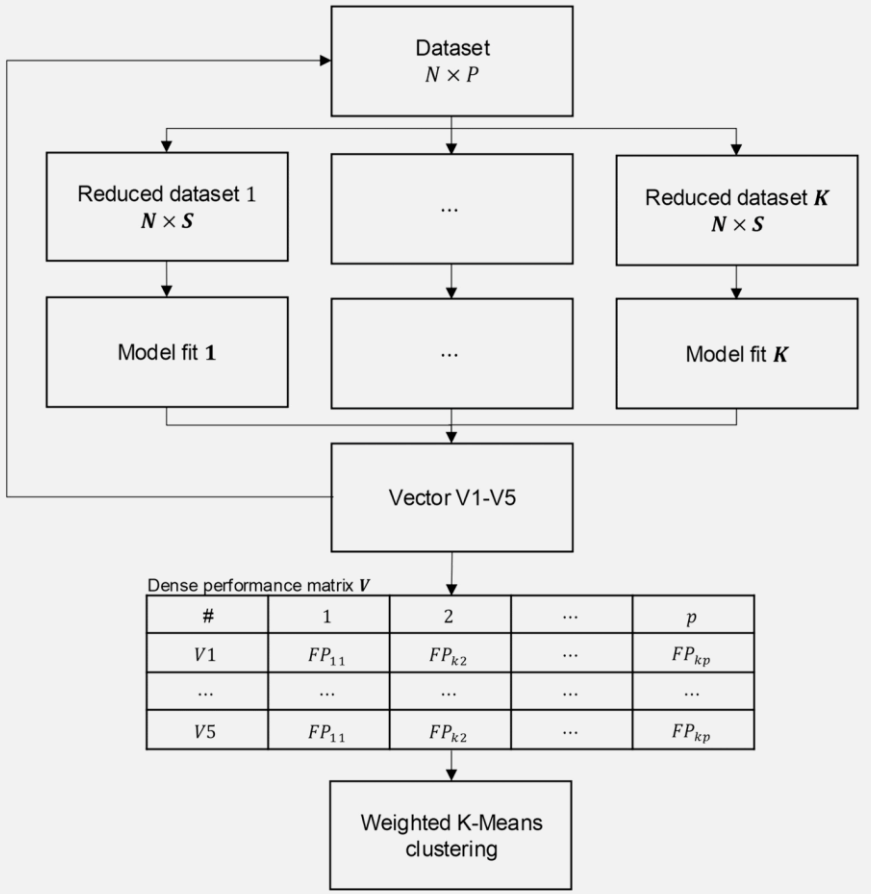
1. Fitting multiple **sub**models to **sub**sets of features → multinomial logistic regression
2. Last model → aggregation of features' importance → linear mixed model
   - estimates $\boldsymbol{y = 1 + Zu + e}$
   - submodel fit
3. Selecting important features
   - **1D-K-means clustering**
4. DL based classification (CNN + DNN)



MD - SRA

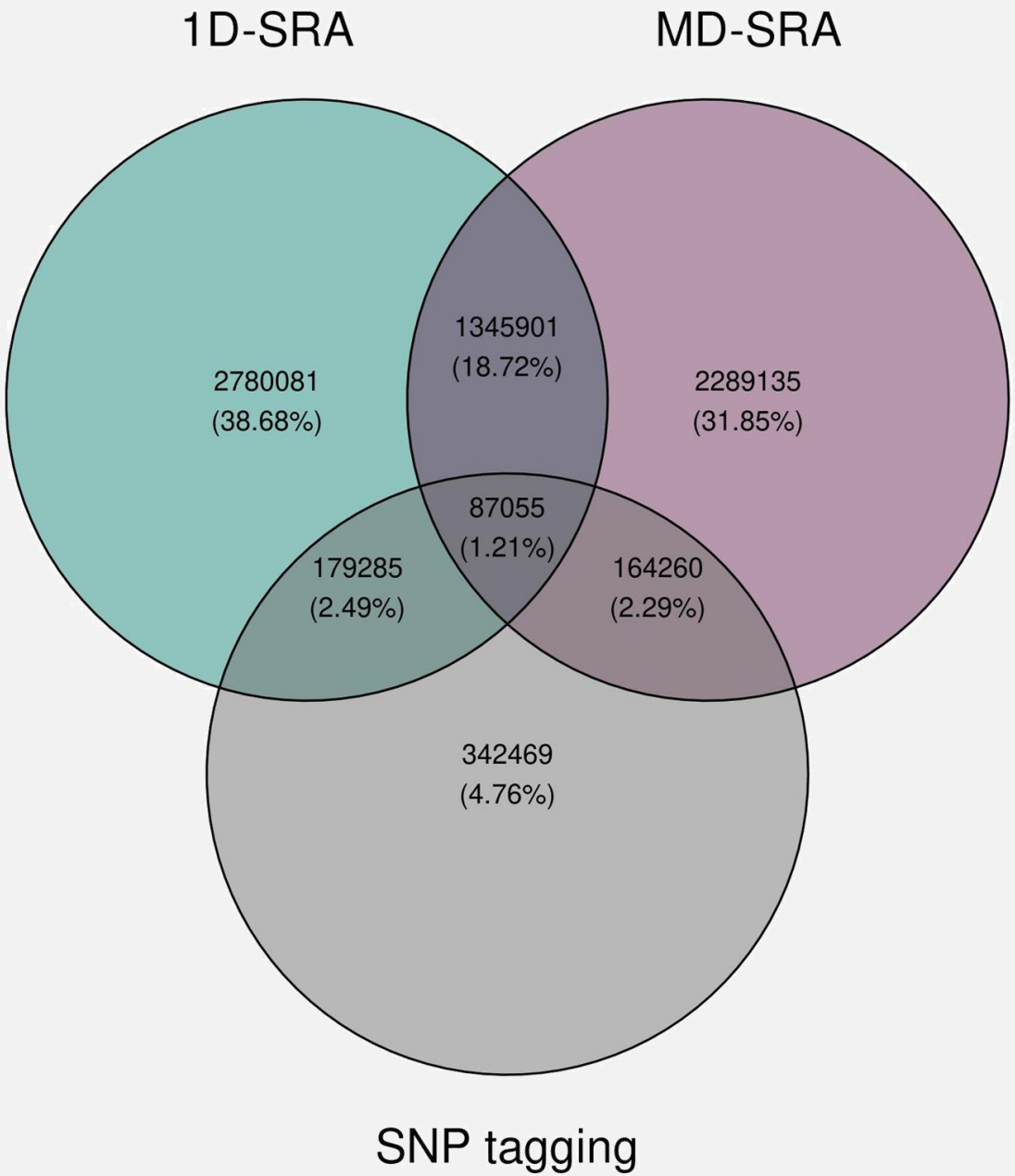
Supervised Rank Aggregation

1. Fitting multiple **sub**models to **sub**sets of features → multinomial logistic regression
2. ~~Last model → aggregation of features' importance → linear mixed model~~
   - ~~estimates~~
   - ~~submodel fit~~
3. Selecting important features
   - **MD-K-means clustering weighted by submodel's fit**
4. DL based classification (CNN + DNN)



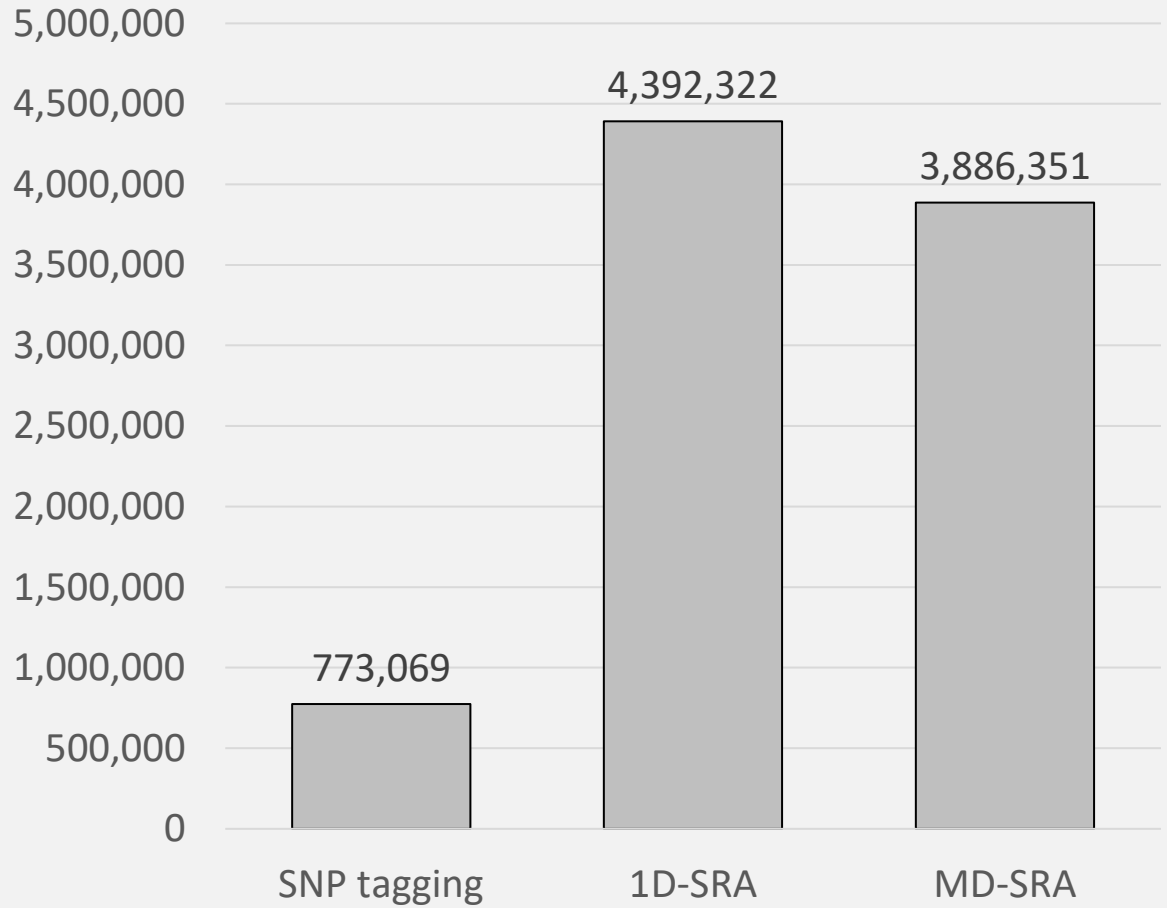
Comparison

Low overalp

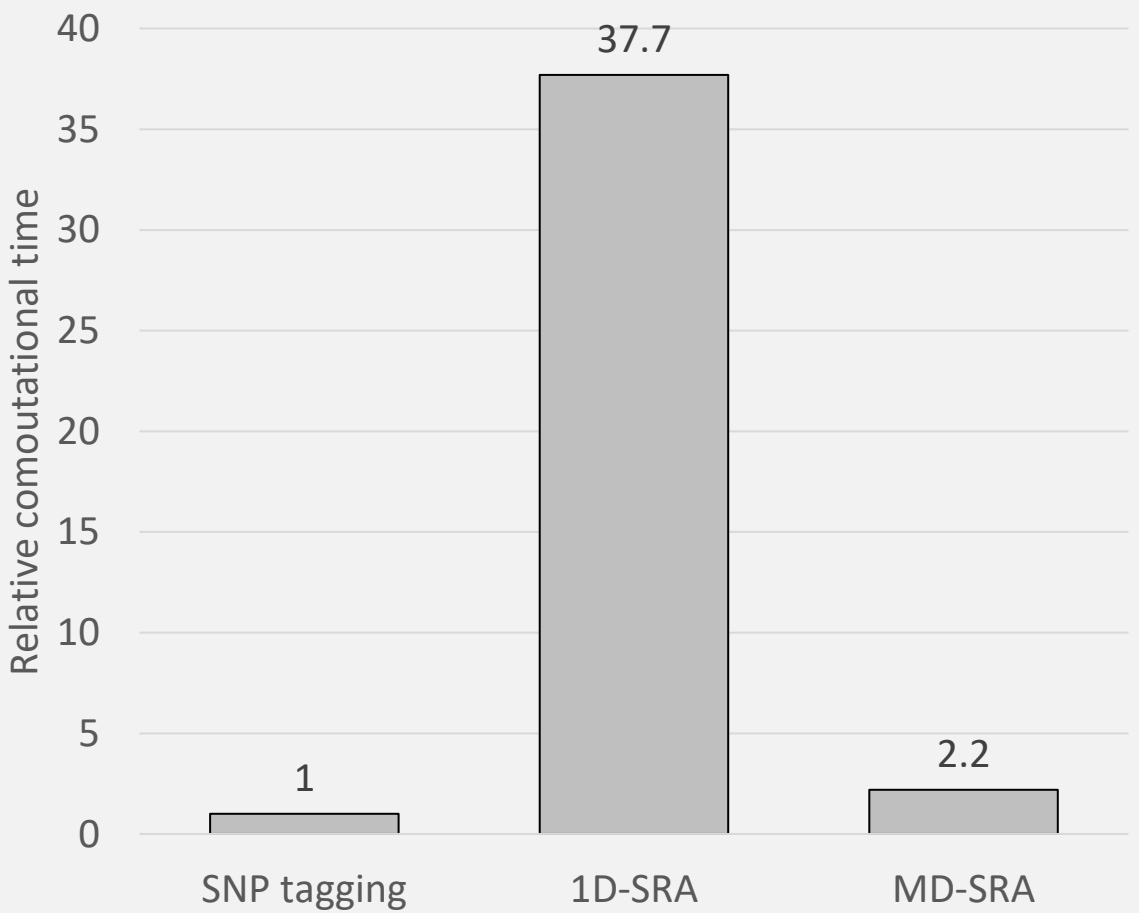


Comparison

Very different numbers of SNPs selected

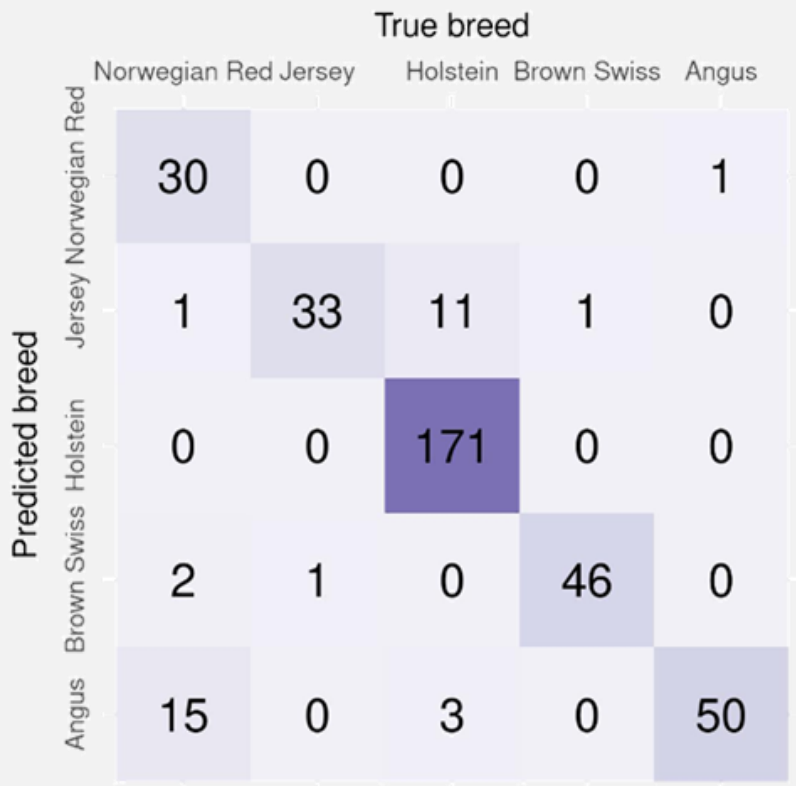


Very different computational times

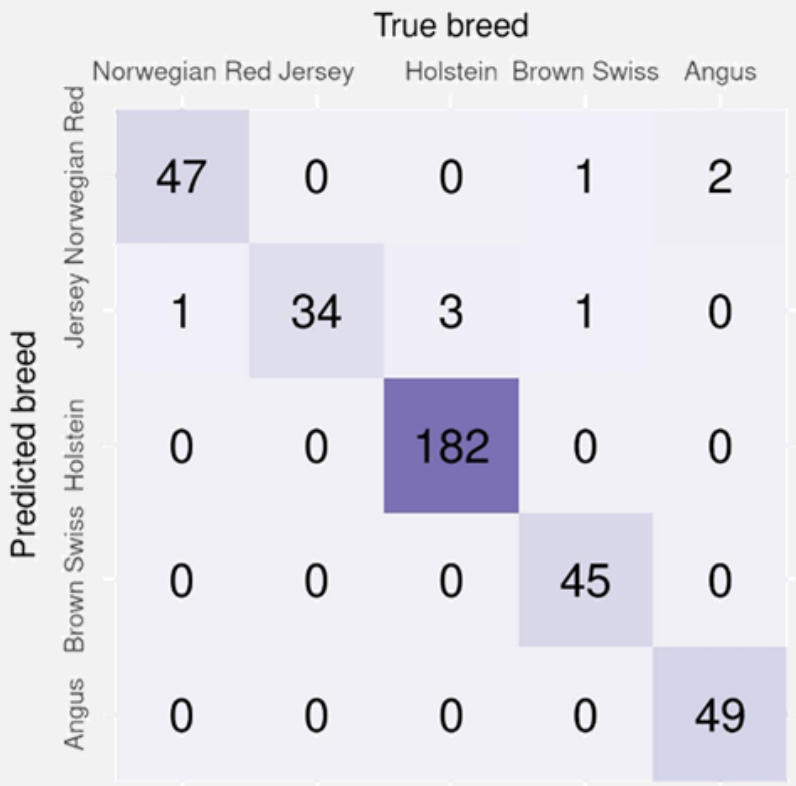


Comparison

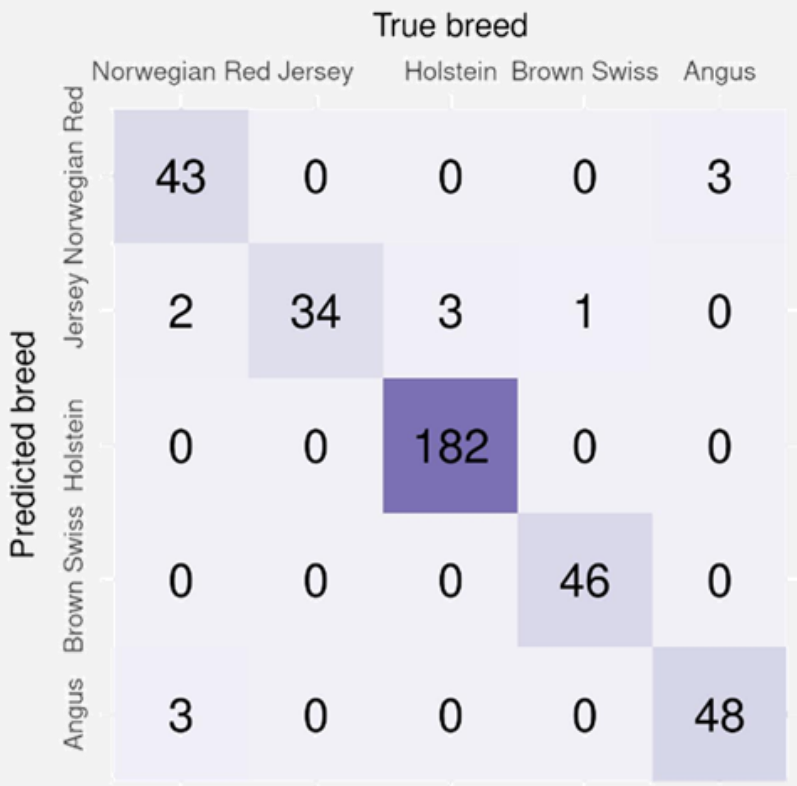
SNP tagging



1D-SRA



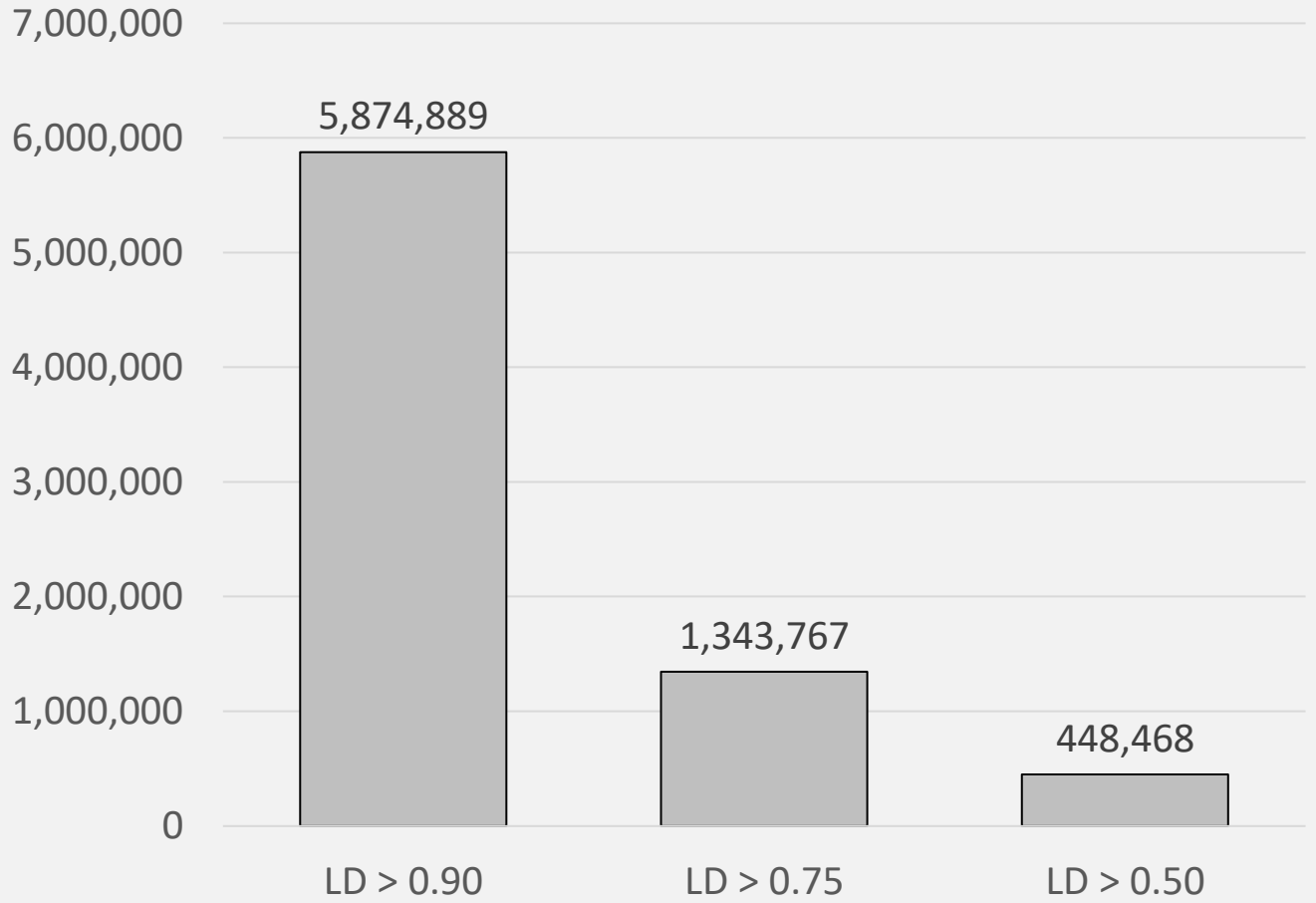
MD-SRA



Other applications

material

- 1000 Bull Genomes Project, Run 9
- 528 HF bulls
  - EBVs for stature
  - 10,771,020 SNPs
  - SNP subsets selected by LD



GWAS

$$y = 1 + Zu + e$$

y EBVs

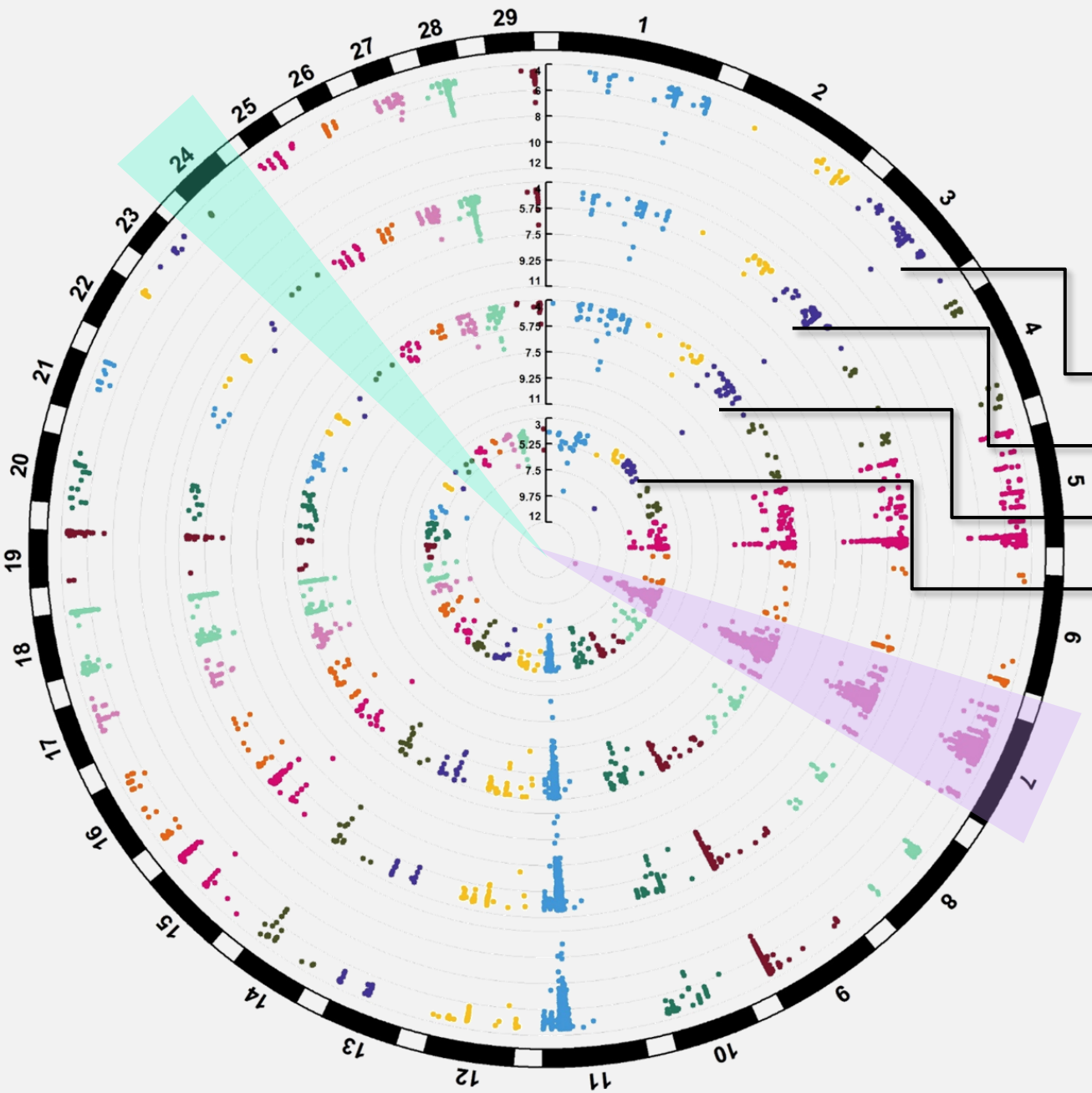
Z SNP genotypes

u SNP effect

e residual

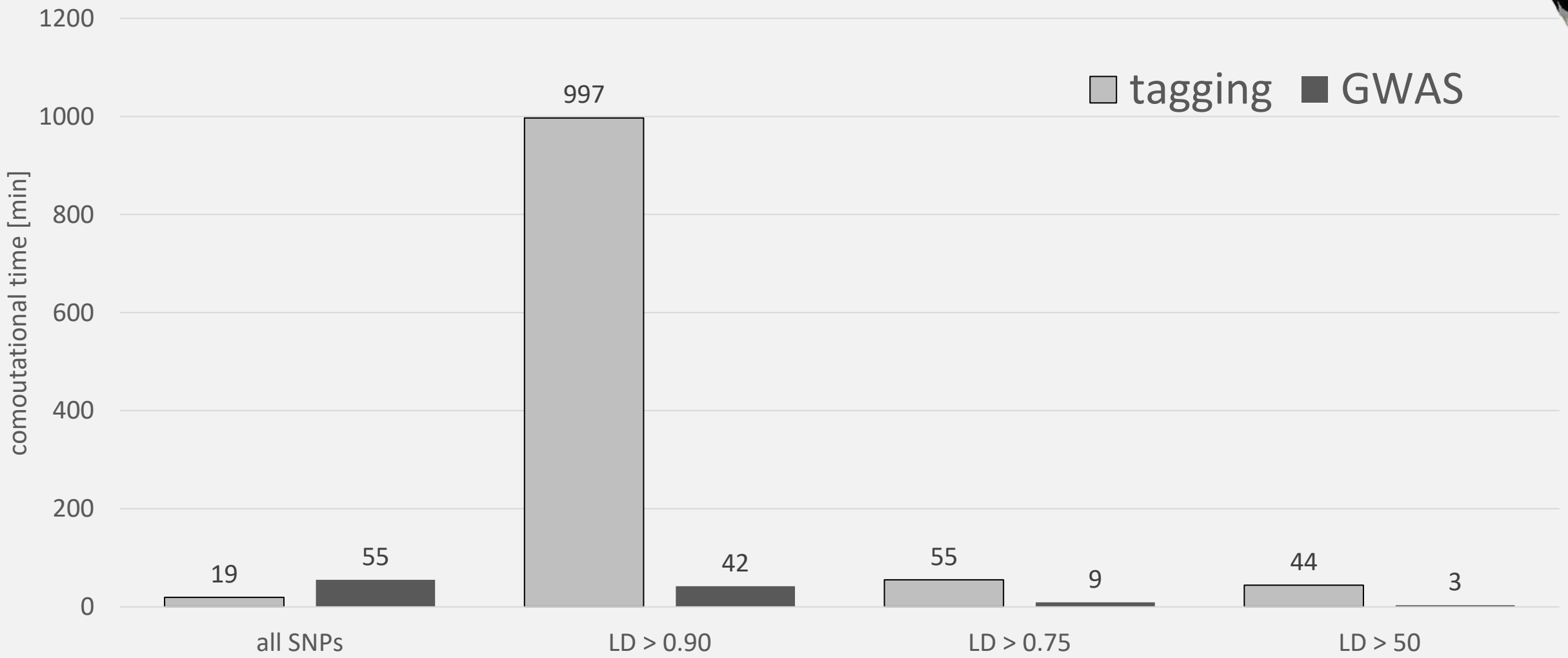
$$u \sim N(0, I\sigma_u^2)$$

$$e \sim N(0, I\sigma_e^2)$$



GWAS

Differences in computational times



1D-SRA humans

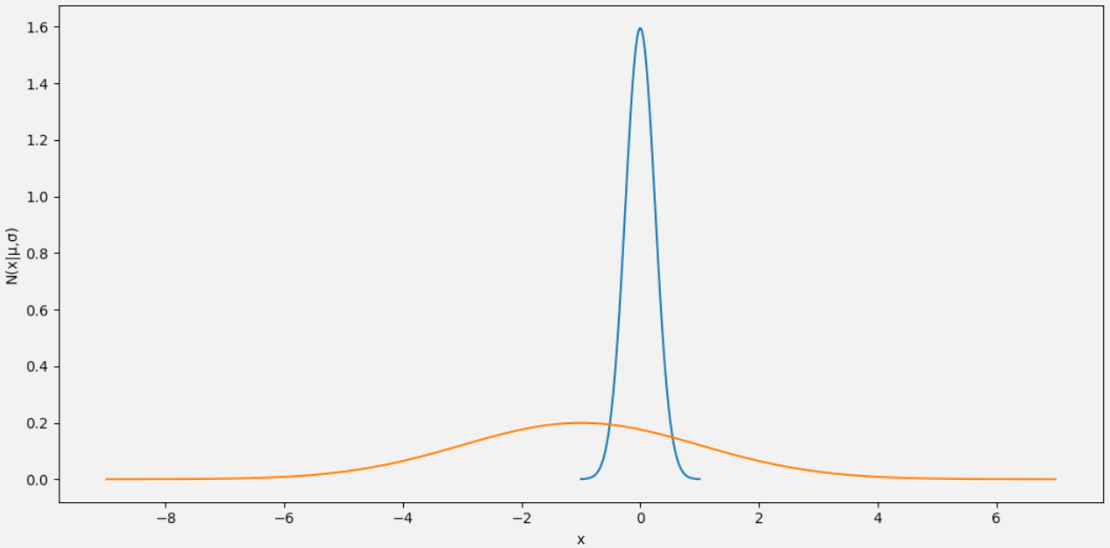
- 1000 Polish Genomes
- 1,222 persons with WGS → 41,836,187 SNPs → ~7,000,000 SNPs
- Severe / non severe COVID-19 infection → binary classification

1. Submodels → logistic regression
2. Important features → 1D-K-means / ~~Gaussian mixture~~
3. Last model → linear mixed model $\boldsymbol{y = 1 + Zu + e}$
4. DL based classification (DNN)

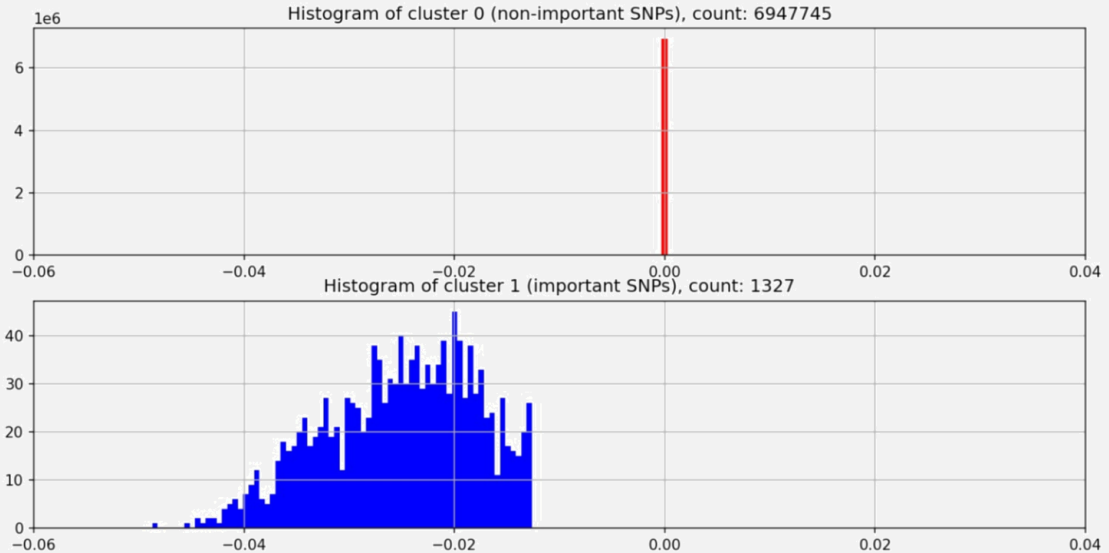
GWAS

Approaches to 1-D clustering

Gaussian mixture



K-means



Conclusions

1. Feature selection necessary for ultra-high-dimensional data
    - Fitting pattern to noise → redundant information
    - Fitting models statistically impossible p>>n
    - Fitting models computationally impossible → numerical or time constraints
2. Feature selection requires design
    - Selection process computationally intensive as well
    - Retain informative features
3. MD-SRA → promising algorithm

People

Wrocław University of Environmental and Life Sciences

Respect the data …

Krzysztof Kotlarz

Magda Mielczarek

Dawid Słomian

Jakub Liu