Quantitative genetics from genome assemblies to neural network aided omics-based prediction of complex traits

Jan Freudenthal

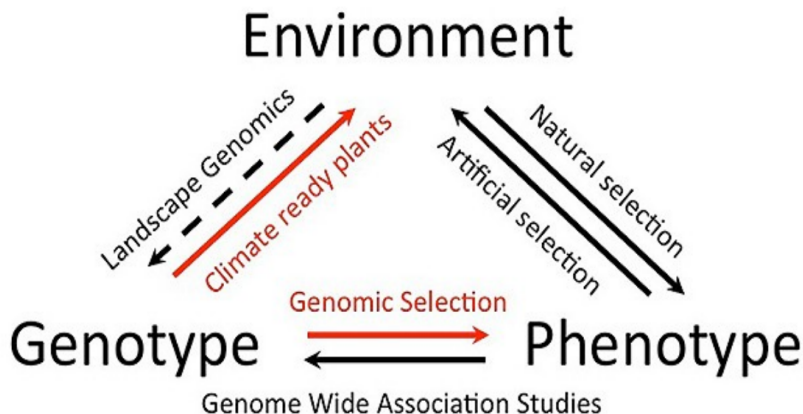CCTB
Evolutionary genomics
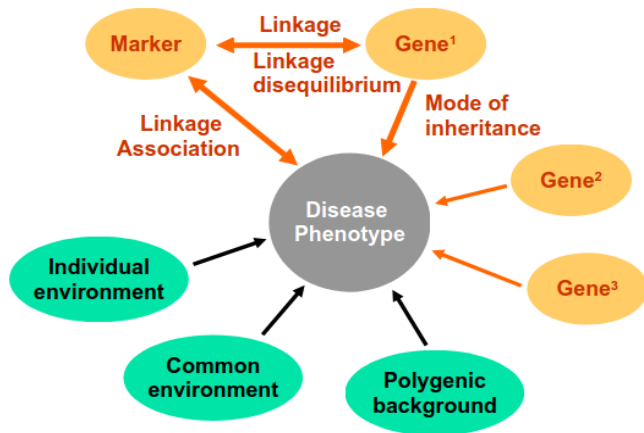Julius-Maximilians-Universität Würzburg

31. Jan 2020

Quantitative genetics

Quantitative genetics aims to explain the heritable parts of traits that follow certain statistical distributions.

Decomposition of phenotypic variance

$$\sigma_P = \sigma_G + \sigma_E + \sigma_{GxE}$$

$$\sigma_G = \sigma_A + \sigma_D + \sigma_I$$

$$\sigma_I = \sigma_{AA} + \sigma_{AD} + \sigma_{DD}$$

$$h^2 = \frac{\sigma_A}{\sigma_P}$$

Decomposition of phenotypic variance

$$\sigma_P = \sigma_G + \sigma_E + \sigma_{GxE}$$

$$\sigma_G = \sigma_A + \sigma_D + \sigma_I$$

$$\sigma_I = \sigma_{AA} + \sigma_{AD} + \sigma_{DD}$$

$$h^2 = \frac{\sigma_A}{\sigma_P}$$

Decomposition of phenotypic variance

$$\sigma_P = \sigma_G + \sigma_E + \sigma_{GxE}$$

$$\sigma_G = \sigma_A + \sigma_D + \sigma_I$$

$$\sigma_I = \sigma_{AA} + \sigma_{AD} + \sigma_{DD}$$

$$h^2 = \frac{\sigma_A}{\sigma_P}$$
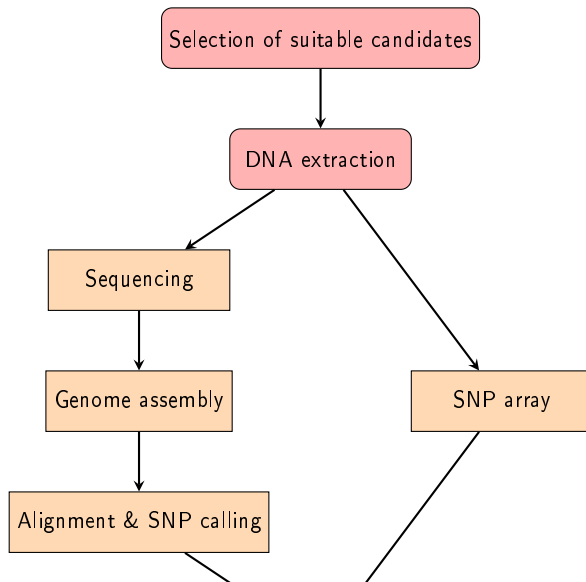
Decomposition of phenotypic variance

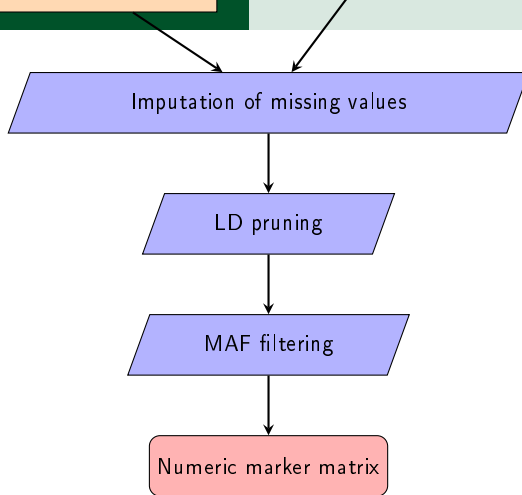$$\sigma_P = \sigma_G + \sigma_E + \sigma_{GxE}$$

$$\sigma_G = \sigma_A + \sigma_D + \sigma_I$$

$$\sigma_I = \sigma_{AA} + \sigma_{AD} + \sigma_{DD}$$

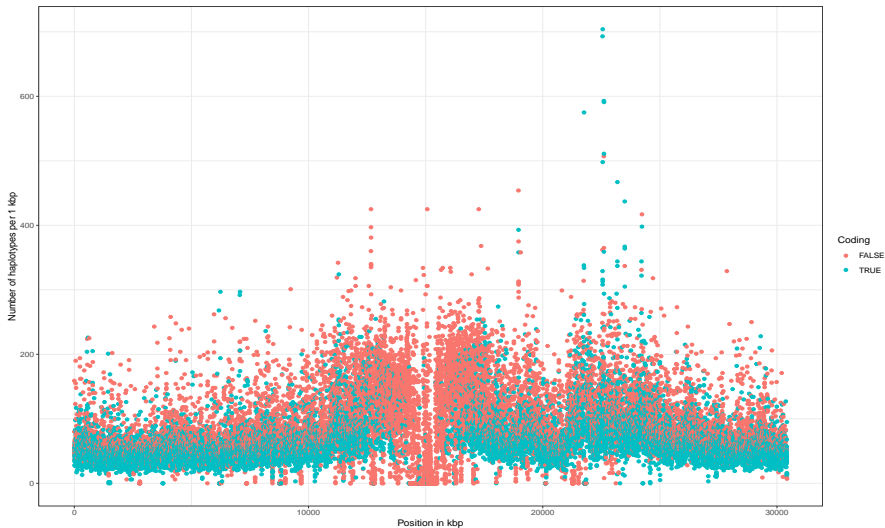$$h^2 = \frac{\sigma_A}{\sigma_P}$$

Workflow in quantitative Genetics

Schematic process of genotyping for quantitative genetics analyses with its crucial steps

Numeric marker matricies

Schematic representation of a genotype matrix as used in genetic analysis like GWAS and genomic prediction

	M-1	M-2	M-3	M-4
Acc1	0	1	1	0
Acc2	1	0	1	0
Acc3	0	1	0	1
Acc4	1	0	0	1

GWAS

Genomic Selection

Objectives

1. Improve GWAS methodology
2. Apply non-parametric statistical methods to genomic selection

Objectives

1. Improve GWAS methodology
2. Apply non-parametric statistical methods to genomic selection

GWAS

- GWAS is the main method used to link traits/phenotypes to genetic polymorphisms
- GWAS utilizes mixed-model linear equations to account for structured populations
- Significant testing is commonly done with bonferroni thresholds

GWAS

- GWAS is the main method used to link traits/phenotypes to genetic polymorphisms
- GWAS utilizes mixed-model linear equations to account for structured populations
- Significant testing is commonly done with bonferroni thresholds

GWAS

- GWAS is the main method used to link traits/phenotypes to genetic polymorphisms
- GWAS utilizes mixed-model linear equations to account for structured populations
- Significant testing is commonly done with bonferroni thresholds

Problems with GWAS and bonferroni correction

- With increasing phenotypes and markers the computational time increases exponentially
- Bonferroni assumes independent testing
- Due to LD markers are not independent from each other
- Permutation based thresholds are better suited to account for LD and structured population
- Permutations have to be repeated 100 times with shuffled phenotypes

Problems with GWAS and bonferroni correction

- With increasing phenotypes and markers the computational time increases exponentially
- Bonferroni assumes independent testing
- Due to LD markers are not independent from each other
- Permutation based thresholds are better suited to account for LD and structured population
- Permutations have to be repeated 100 times with shuffled phenotypes

Problems with GWAS and bonferroni correction

- With increasing phenotypes and markers the computational time increases exponentially
- Bonferroni assumes independent testing
- Due to LD markers are not independent from each other
- Permutation based thresholds are better suited to account for LD and structured population
- Permutations have to be repeated 100 times with shuffled phenotypes

Problems with GWAS and bonferroni correction

- With increasing phenotypes and markers the computational time increases exponentially
- Bonferroni assumes independent testing
- Due to LD markers are not independent from each other
- Permutation based thresholds are better suited to account for LD and structured population
  - Permutations have to be repeated 100 times with shuffled phenotypes

# Problems with GWAS and bonferroni correction

- With increasing phenotypes and markers the computational time increases exponentially
- Bonferroni assumes independent testing
- Due to LD markers are not independent from each other
- Permutation based thresholds are better suited to account for LD and structured population
- Permutations have to be repeated 100 times with shuffled phenotypes

GWAS Flow

- GWAS consists of a series of matrix operations that can be highly parallelized
- GWAS Flow uses the TensorFlow's Python API
- The calculations can be run on both GPU and CPU
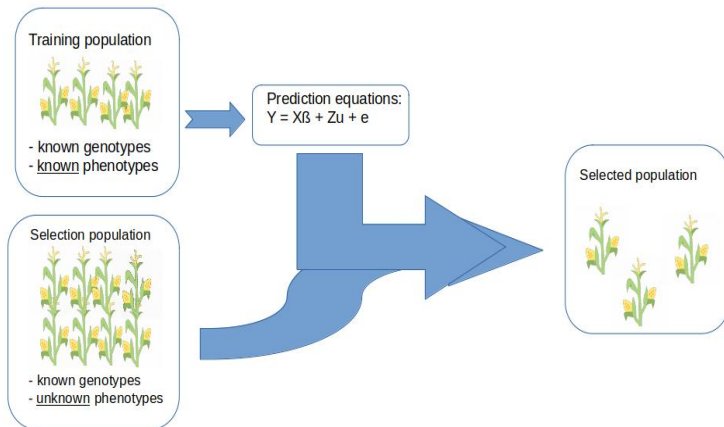
GWAS Flow

- GWAS consists of a series of matrix operations that can be highly parallelized
- GWAS Flow uses the TensorFlow's Python API
- The calculations can be run on both GPU and CPU

GWAS Flow

- GWAS consists of a series of matrix operations that can be highly parallelized
- GWAS Flow uses the TensorFlow's Python API
- The calculations can be run on both GPU and CPU

Performance of GWAS-Flow
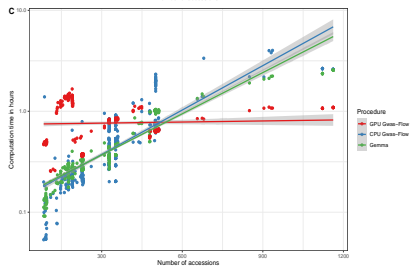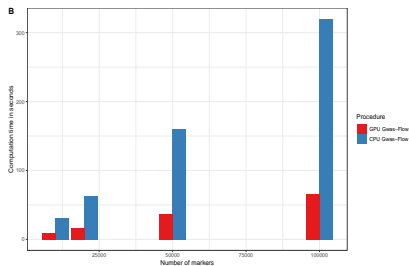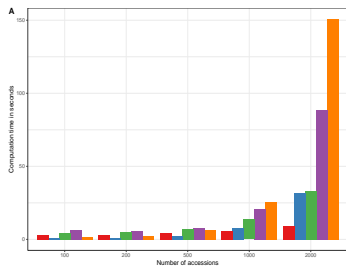
Genomic selection

Advantages of genomic selection

1. Selection from larger populations
2. Stricter selection intensity
3. Acceleration of the breeding cycle
4. Reduction in phenotyping costs

Advantages of genomic selection

1. Selection from larger populations
2. Stricter selection intensity
3. Acceleration of the breeding cycle
4. Reduction in phenotyping costs

Advantages of genomic selection

1. Selection from larger populations
2. Stricter selection intensity
3. Acceleration of the breeding cycle
4. Reduction in phenotyping costs

Advantages of genomic selection

1. Selection from larger populations
2. Stricter selection intensity
3. Acceleration of the breeding cycle
4. Reduction in phenotyping costs

Prediction methods in genomic selection

1. GBLUP - the gold-standard in genomic selection
2. Bayesian methods
3. Bayesian methods differ in the a prior assumptions on marker effect sizes
4. RHKS, random forest, support vector machines etc.

Prediction methods in genomic selection

1. GBLUP - the gold-standard in genomic selection
2. Bayesian methods
3. Bayesian methods differ in the a prior assumptions on marker effect sizes
4. RHKS, random forest, support vector machines etc.

Prediction methods in genomic selection

1. GBLUP - the gold-standard in genomic selection
2. Bayesian methods
3. Bayesian methods differ in the a prior assumptions on marker effect sizes
4. RHKS, random forest, support vector machines etc.

Prediction methods in genomic selection

1. GBLUP - the gold-standard in genomic selection
2. Bayesian methods
3. Bayesian methods differ in the a prior assumptions on marker effect sizes
4. RHKS, random forest, support vector machines etc.
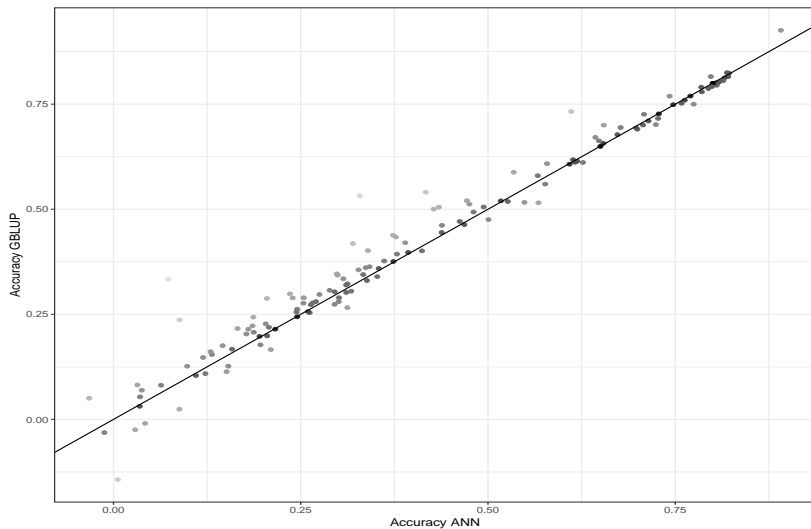
Neural networks in genomic prediction

- With the advent of high performing GPUs neural networks become popular in all branches
- In the scope of the present study the usability of neural networks for genomic selection was assessed
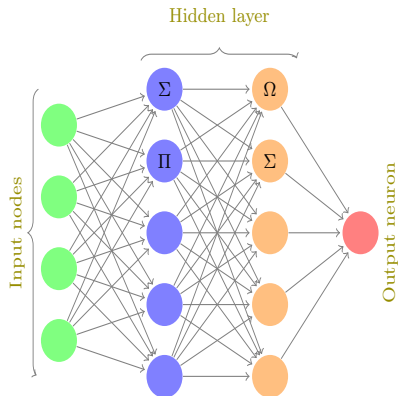
Datasets

- *A. thaliana* data from the 1001 genome project with 10 mio markers and 164 different phenotypes
- Doubled-haploid maize populations derived from two maize landraces

Datasets

- *A. thaliana* data from the 1001 genome project with 10 mio markers and 164 different phenotypes
- Doubled-haploid maize populations derived from two maize landraces
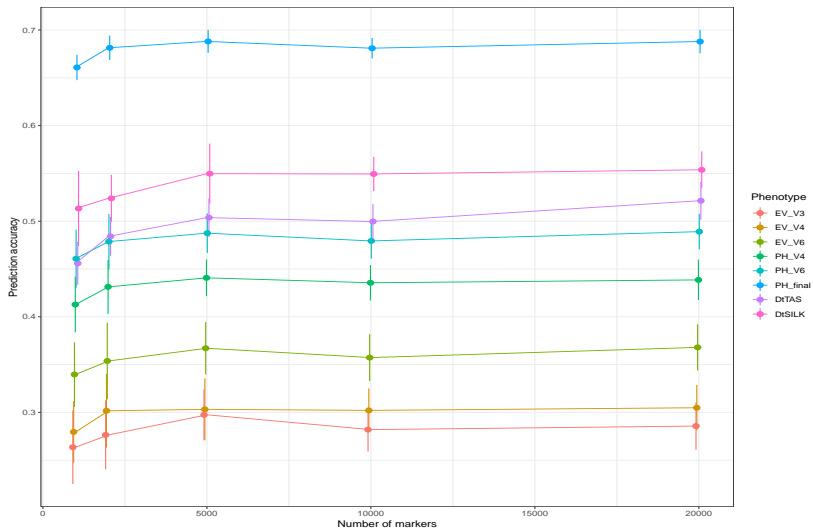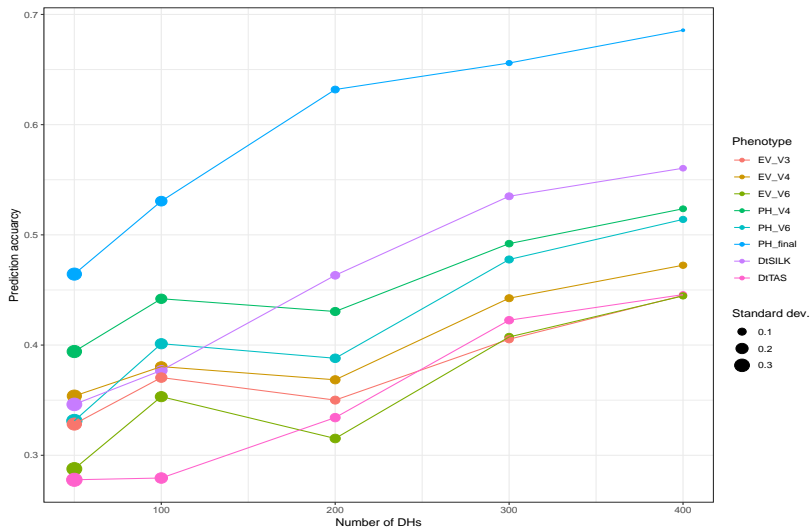
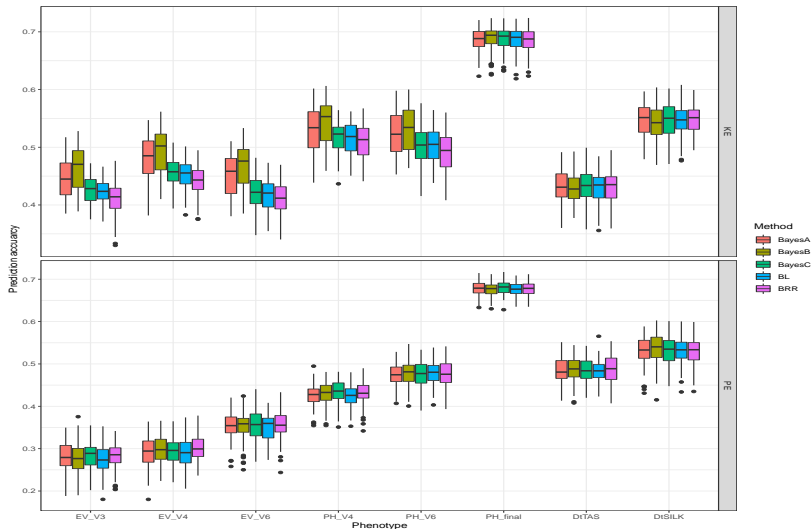# Genomic prediction *A. thaliana*
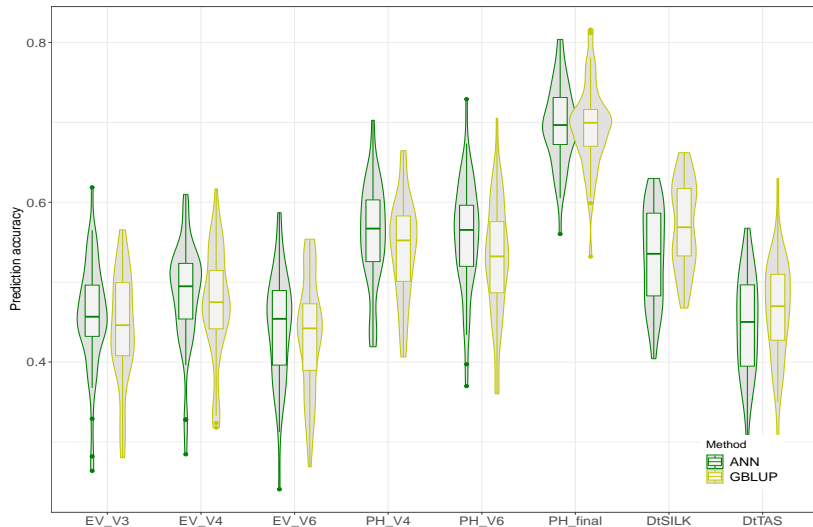
Number of markers versus prediction accuracy

Number of phenotypes versus prediction accuracy

Comparison of bayesian methods

Conclusions

- Neural networks are well-suited for genomic selection
- All methods fail to outperform all the other methods
- The limiting factors are the number of accessions and the effective population size
- Mostly additive effects are captured by the prediction equations

Conclusions

- Neural networks are well-suited for genomic selection
- All methods fail to outperform all the other methods
- The limiting factors are the number of accessions and the effective population size
- Mostly additive effects are captured by the prediction equations

Conclusions

- Neural networks are well-suited for genomic selection
- All methods fail to outperform all the other methods
- The limiting factors are the number of accessions and the effective population size
- Mostly additive effects are captured by the prediction equations

Conclusions

- Neural networks are well-suited for genomic selection
- All methods fail to outperform all the other methods
- The limiting factors are the number of accessions and the effective population size
- Mostly additive effects are captured by the prediction equations

Thanks for your attention!