

JULIUS-MAXIMILIANS UNIVERSITÄT  
WÜRZBURG

DOCTORAL THESIS

---

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

---

*Author:*

Jan Alexander  
FREUDENTHAL

*Supervisor:*

Prof. Arthur KORTE

*A thesis submitted in fulfillment of the requirements  
for the degree of Ph.D.*

*in the*

Research group for evolutionary genomics  
GSLs

November 3, 2019

## **Declaration of Authorship**



*“Wit beyond measure is man’s greatest treasure”*

Rowena Rawenclaw



JULIUS-MAXIMILIANS UNIVERSITÄT WÜRZBURG

# *Abstract*

Faculty Name

GSLs

Ph.D.

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

by Jan Alexander FREUDENTHAL

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...





# *Acknowledgements*

The acknowledgments and the people to thank go here, don't forget to include your project advisor...



# Contents

<b>Declaration of Authorship</b>	<b>iii</b>
<b>Abstract</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>1 Benchmarking of Chloroplast Genome Assembly tools</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.1.1 Motivation . . . . .	1
1.1.2 Extraction of chloroplast reads from whole genome data and general assembly workflow . . . . .	4
Purpose and scope of benchmarking the landscape of chloro- plast assembly tools . . . . .	5
1.2 Material and Methods . . . . .	6
1.2.1 Methods . . . . .	6
Data and code availability . . . . .	6
Tools . . . . .	6
Standardization and reproducibility . . . . .	7
1.2.2 Data . . . . .	8
Simulated . . . . .	8
Real data set . . . . .	8
Novel data sets . . . . .	9
1.2.3 Evaluation . . . . .	9
Quantitative . . . . .	9

	Consistency . . . . .	10
1.3	Results . . . . .	11
1.3.1	Quantitative . . . . .	11
	Simulated data . . . . .	11
	Real data sets . . . . .	13
	Consistency . . . . .	15
	Novel . . . . .	16
1.4	Discussion . . . . .	17
1.5	Conclusion & outlook . . . . .	22
<b>2</b>	<b>Understanding the haplotype structure of <i>Arabidopsis thaliana</i></b>	<b>23</b>
2.1	Introduction . . . . .	23
2.2	Haplotyping of <i>A. thaliana</i> . . . . .	23
2.3	Results . . . . .	23
2.4	Disucssion . . . . .	23
<b>3</b>	<b>GWAS-Flow a gpu-accelerated software for large-scale genome-wide as-</b>	
	<b>sociation studies</b>	<b>31</b>
3.1	Introduction . . . . .	31
3.2	Methods . . . . .	33
	GWAS Model . . . . .	33
	The GWAS-Flow Software . . . . .	33
	Calculation of permutation-based thresholds for GWAS . . . . .	34
	Benchmarking . . . . .	35
3.3	Results . . . . .	36
3.4	Disucssion . . . . .	37
<b>4</b>	<b>Genomic prediction of phenotypic values of quantitative traits using</b>	
	<b>artificial neural networks</b>	<b>41</b>
4.1	Introduction . . . . .	41

4.1.1	A brief history of machine learning . . . . .	41
	Basic perceptron model . . . . .	41
	Activation functions . . . . .	44
	Gradient descent algorithm . . . . .	46
	Optimizers . . . . .	49
	Backpropagation . . . . .	50
	Regularization parameters . . . . .	50
4.1.2	On the nature of quantitative traits . . . . .	52
4.1.3	Artificial selection in plant and animal breeding in the ge- nomics era . . . . .	56
	Introduction to genomic selection . . . . .	56
	Genomic selection in recurrent selection and the breeders equation . . . . .	58
	Genomic BLUP and Bayesian methods . . . . .	61
	cross validation . . . . .	68
4.1.4	Genomic selection using artificial neural networks . . . . .	68
4.2	Proof of concept for ANN-based genomic selection . . . . .	71
4.3	Material . . . . .	73
4.3.1	DH populations derived from maize landraces . . . . .	73
	Genomic maize data . . . . .	74
	Phenotypic maize data . . . . .	74
4.3.2	A. thaliana . . . . .	75
	Genomic data . . . . .	75
	Phenotypic data . . . . .	75
4.4	Methods . . . . .	75
	CV . . . . .	76
4.4.1	ANN . . . . .	76
4.4.2	GBLUP . . . . .	77
4.5	Results . . . . .	77

4.6	Discussion	80
<b>5</b>	<b>GWAS</b>	<b>81</b>
5.1	Reevaluation of 463 phenotypes from the AraPheno database	81
5.1.1	Introduction	81
5.1.2	Material and Methods	81
5.1.3	Results	81
5.1.4	Results	81
5.1.5	Disucssion	81
5.2	GWAS in DH landrace populatios of maze across and within envi- ronments	81
5.2.1	Introduction	81
5.2.2	Material and Methods	81
5.2.3	Results	81
5.2.4	Results	81
5.2.5	Disucssion	81
<b>A</b>	<b>Source code</b>	<b>83</b>
A.1	GBLUP example	83
A.2	gwas.py	83
A.3	main.py	87
A.4	herit.py	91
<b>B</b>	<b><i>A. thaliana</i> phenotypic data</b>	<b>93</b>
<b>C</b>	<b>Genomic prediction</b>	<b>101</b>
C.1	GP Scripts	101
C.2	Results of GP	108
	<b>Bibliography</b>	<b>117</b>

# List of Figures

1.1	Structure of a chloroplast genome . . . . .	3
1.2	Chloroplast genome assembly workflow . . . . .	5
1.3	Score of assemblies of simulated data sets . . . . .	12
1.4	Scores of assemblies from real data sets . . . . .	15
1.5	Comparison between two runs with the same assembler for consistency testing . . . . .	16
1.6	Upset plot comparing the success rates for novel data sets . . . . .	17
1.7	Upset plot comparing the success rates of of all assemblers . . . . .	19
1.8	AliTV plot of alignments of assemblies of <i>Oryza brachyantha</i> of all assemblers . . . . .	21
2.1	Haplotype strutcture of chromosome 1 of <i>A. thaliana</i> . . . . .	24
2.2	Haplotype strutcture of chromosome 2 of <i>A. thaliana</i> . . . . .	25
2.3	Haplotype strutcture of chromosome 3 of <i>A. thaliana</i> . . . . .	26
2.4	Haplotype strutcture of chromosome 4 of <i>A. thaliana</i> . . . . .	27
2.5	Haplotype strutcture of chromosome 5 of <i>A. thaliana</i> . . . . .	28
2.6	blabl . . . . .	29
3.1	Computation time vs number of markers . . . . .	37
3.2	Computations time vs accessions . . . . .	38
3.3	Computational time of GWA Analyses on real <i>A. thaliana</i> data sets . . . . .	40
4.1	Basic perceptron model . . . . .	41
4.2	Schematic layout of a simple multi-layer perceptron . . . . .	43

4.3	Popular activation functions for neural networks . . . . .	44
4.4	Training vs. validation loss over time . . . . .	51
4.5	Truncation selection of a normal distributed phenotype . . . . .	59



# List of Tables

1.1	Data selection criteria for real data sets from SRA . . . . .	9
1.2	Scores of assemblies of simulated data . . . . .	13
1.3	Mean scores of chloroplast genome assemblers . . . . .	14
4.1	Overview of properties of a variety of commonly applied Bayesian methods for genomic prediction. Table altered after <i>Kärkkäinen and Sillanpää, 2012</i> . . . . .	68
4.2	Simple simulated phenotypes and genotypes for genomic prediction with genotypes $G_1 \dots G_4$ , $M_1$ and $M_2$ and phenotypes based on additive effects or <i>and</i> , <i>or</i> , <i>xor</i> logic gates. . . . .	71
4.3	Results of genomic prediction from phenotypes and genotypes in table 4.2 . . . . .	72



# List of Abbreviations

<b>Adadelata</b>	<b>Adaptive delta</b>
<b>Adagrad</b>	<b>Adaptive Gradient Algorithm</b>
<b>Adam</b>	<b>Adaptive Moment estimation</b>
<b>ANN</b>	<b>Artificial Neural Network</b>
<b>API</b>	<b>Application Programming Interface</b>
<b>AUC</b>	<b>Area Under the Curve</b>
<b>BLUE</b>	<b>Best Linear Unbiased Estimator</b>
<b>BLUP</b>	<b>Best Linear Unbiased Predictor</b>
<b>BP</b>	<b>Base Pair</b>
<b>CPU</b>	<b>Core Processing Unit</b>
<b>DH</b>	<b>Doubled Haploid</b>
<b>DNA</b>	<b>DeoxyriboNucleic Acid</b>
<b>RNA</b>	<b>RiboNucleic Acid</b>
<b>EMMA</b>	<b>Efficient Mixed Model Associations</b>
<b>FCL</b>	<b>Fully Connected Layer</b>
<b>GBLUP</b>	<b>Genomic Best Linear Unbiased Predictor</b>
<b>GD</b>	<b>Gradient Descent</b>
<b>GEBC</b>	<b>Genomic Estimated Breeding Values</b>
<b>GPL</b>	<b>General Public License</b>
<b>GP</b>	<b>Genomic Prediction</b>
<b>GPU</b>	<b>Graphical Processing Unit</b>
<b>GRM</b>	<b>Genomic Relationship Matrix</b>
<b>GS</b>	<b>Genomic Selection</b>

<b>GUI</b>	<b>G</b> raphical <b>U</b> ser <b>I</b> nterface
<b>GWAIS</b>	<b>G</b> enome <b>W</b> ide <b>I</b> nteraction <b>A</b> ssociation <b>S</b> tudies
<b>GWAS</b>	<b>G</b> enome <b>W</b> ide <b>A</b> ssociation <b>S</b> tudies
<b>HDF</b>	<b>H</b> ierarchical <b>D</b> ata <b>F</b> ormat
<b>HPC</b>	<b>H</b> igh <b>P</b> erformance <b>C</b> omputing
<b>IR</b>	<b>I</b> nverted <b>R</b> epet
<b>LCL</b>	<b>L</b> ocally <b>C</b> onnected <b>L</b> ayer
<b>LD</b>	<b>L</b> inkage <b>D</b> isequilibrium
<b>LMM</b>	<b>L</b> inear <b>M</b> ixed <b>M</b> odel
<b>LSC</b>	<b>L</b> arge <b>S</b> ingle <b>C</b> opy
<b>MAF</b>	<b>M</b> inor <b>A</b> llele <b>F</b> requncy
<b>MBP</b>	<b>M</b> ega <b>B</b> ase <b>P</b> air
<b>MCMC</b>	<b>M</b> arkov <b>C</b> hain <b>M</b> onte <b>C</b> arlo
<b>MLP</b>	<b>M</b> ulti <b>L</b> ayer <b>P</b> erceptron
<b>ML</b>	<b>M</b> achine <b>L</b> earning
<b>MSE</b>	<b>M</b> ean <b>S</b> quare <b>E</b> rror
<b>Nadam</b>	<b>N</b> esterov- <b>a</b> ccelerated <b>A</b> daptive <b>M</b> oment <b>E</b> stimation
<b>NAG</b>	<b>N</b> esterov <b>A</b> ccelerated <b>M</b> omentum
<b>NCBI</b>	<b>N</b> ational <b>C</b> enter for <b>B</b> iotechnological <b>I</b> nformation
<b>QTL</b>	<b>Q</b> uantitative <b>T</b> rait <b>L</b> ocus
<b>ReLU</b>	<b>R</b> ectified <b>L</b> inear <b>U</b> nits
<b>RKHS</b>	<b>R</b> eproducing <b>K</b> ernel <b>H</b> ilbert <b>S</b> paces
<b>RMSE</b>	<b>R</b> oot <b>M</b> ean <b>S</b> quare <b>E</b> rror
<b>RMSPProp</b>	<b>R</b> oot <b>M</b> ean <b>S</b> quare <b>P</b> ropagation
<b>ROC</b>	<b>R</b> eciever <b>O</b> perating <b>C</b> haracteristics
<b>RSS</b>	<b>R</b> esidual <b>S</b> um of <b>S</b> quares
<b>SGD</b>	<b>S</b> tochastic <b>G</b> radient <b>D</b> escent
<b>SNP</b>	<b>S</b> ingle <b>N</b> ucleotide <b>P</b> olymorphism

<b>SRA</b>	<b>Sequence Read Archive</b>
<b>SSC</b>	<b>Small Single Copy</b>
<b>TRN</b>	<b>TRaiNing subset</b>
<b>TST</b>	<b>TeSTing subset</b>
<b>WGS</b>	<b>Whole Genome Sequencing</b>
<b>XOR</b>	<b>eXclusive OR</b>



*For/Dedicated to/To my...*





# 1 Benchmarking of Chloroplast Genome Assembly tools

This chapter orientates on *Freudenthal et al., 2019b* only the chapters from the publication which the author majorly contributed are included. If not cited otherwise the plots even though they were published along the aforementioned paper, were designed and generated by the author.

## 1.1 Introduction

### 1.1.1 Motivation

Certain organelles like mitochondria and chloroplasts contain their own genetic information from which they are able to synthesize certain proteins independent of the core genome. Evolutionary this developed during endosymbiosis. Its underlying theory seeks to explain how eukaryotic cells formed from prokaryotes *Mereschkowsky, 1905; Kutschera and Niklas, 2005*. The widely acknowledged theory explains how in the early evolution of eukaryotes, other cells were incorporated in those cells from which today's known organelles descent. In the case of a chloroplast this was most likely a photosynthetic bacteria or similar organism *Archibald, 2015*. Until today the structure of chloroplast genomes resembles more that of a prokaryotic genome than that of its eukaryotic host cells. A typical chloroplast genome consists of circular DNA with a size between 120 kBP to 160 kBP *Palmer, 1985*. The first chloroplast have been sequenced as early as 1986 and were isolated from *Marchantia polymorpha* and *Nico Ohyama et al., 1986; Shinozaki*

et al., 1986. Complete reviews of the genome structure of chloroplast genomes were written by Green, 2011; Wicke et al., 2011. Chloroplast genomics is widely applied in evolutionary studies Martin et al., 2002; Xiao-Ming et al., 2017. The genome has been reduced in size over the course of evolution through endosymbiotic gene transfer Martin et al., 2002; Deiner et al., 2017. This is still an ongoing process, which can be observed *in vitro* Bock, 2017; Fuentes et al., 2014; Stegemann and Bock, 2009. Up to 14 % of the the core genome of *Arabidopsis thaliana* is made up of genes previously located on the chloroplast (fancy citation), while 100-120 genes remain on the chloroplast Wicke et al., 2011 itself, which by far would not be enough to function independently. Chloroplast genomes are much smaller than core genomes e.g *A. thaliana* has a core genome size of ca. 125 MBP and are highly conserved with a large gene content, therefore polymorphisms are more likely to cause functional changes in the metabolism. Single chloroplast contain up to hundreds of copies of its genome Kumar, Oldenburg, and Bendich, 2014; Bendich, 1987. Photosynthetic activate cells contain multiple chloroplasts, therefore the copy number of the chloroplast genomes is much higher than the number of core genomes per cell which is most cases is 1.

Structurally chloroplast genomes are made up of 4 distinct regions: two inverted repeats (IR) -  $IR_A$  and  $IR_B$  - ranging from 10 kbp to 76 kbp. They divide the genome into two areas: the large single copy (LSC) and the small single copy (SSC) as shown in 1.1 Palmer, 1985. Taking into account that the majority of assembly tools has been designed to assemble linear core genomes, the structure of chloroplast genomes is a major obstacle when wanting to assemble those genomes with modern short read technologies, especially solving and aligning the IR Wang et al., 2018.

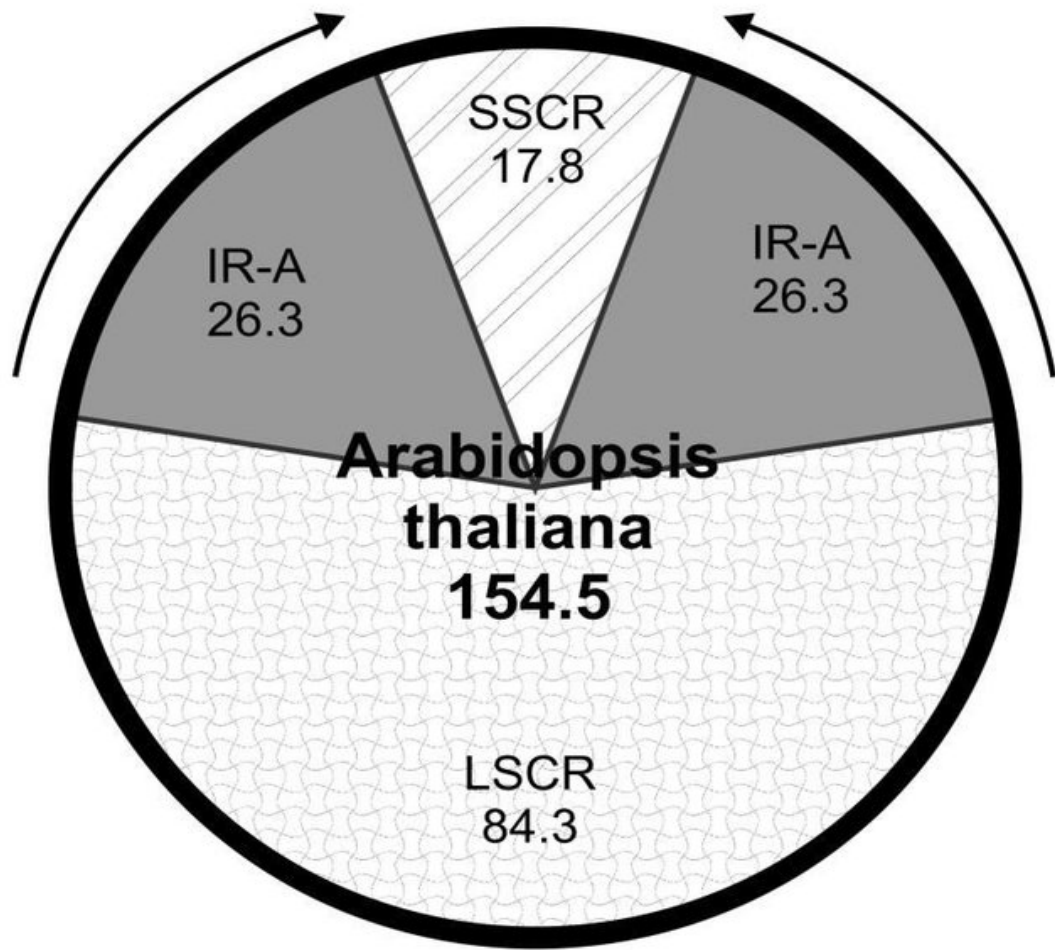


FIGURE 1.1: Structure of the chloroplast genome of *A. thaliana* with SSC-Region and LSC-Region. Length of the genome and its parts in kbp. Graphic from Olejniczak et al., 2016

Chloroplast genomes present biologist with many obstacles of which heteroplasmy stands out and immensely complicated genome assemblies and downstream analyses Corriveau and Coleman, 1988; Chat et al., 2002. Heteroplasmy describes the phenomenon of co-existence of multiple versions of the chloroplast genomes in a single organism and single cells of that respective organism. The underlying evolutionary mechanisms behind heteroplasmy are not fully elucidated and eventually existing fitness advantages fueling heteroplasmy cannot be explained satisfactorily by evolutionary methods Scarcelli et al., 2016. There is a variety of databases containing short read data for species without an chloroplast

genome available. For example NCBI's the short read archive (SRA) *Leinonen et al., 2010*. Because most plant DNA extraction protocols use green leaf tissue as basis, they usually contain a large amount of plastid DNA. Having larger number of assembled and annotated chloroplast genomes publicly available would be beneficial for evolutionary studies and could be a useful addition to bar-coding and super-barcoding *Coissac et al., 2016* and other biotechnological applications *Daniell et al., 2016*. To achieve this there is a variety of tools available. The study presented in this chapter, assesses the availability, usability and overall performance of 7 of those tools and ultimately makes use of the newly gained insights to attempt *de novo* assemble more than 100 chloroplasts.

### **1.1.2 Extraction of chloroplast reads from whole genome data and general assembly workflow**

There is a variety of strategies to assemble chloroplast genomes from raw sequencing data *Twyford and Ness, 2017*. In general the process involves three steps: (i) extraction of plastid reads from the WGS data, (ii) assembly of the plastid genome, (iii) solving the circular structure of the chromosome with the IRs. There are two distinct ways to tackle step (i): The first one is to map all the reads to a reference chloroplast *Vinga et al., 2012*, which works reasonably well if there is one available for the same, or at least a closely related species. The second is to make use of the much larger coverage of chloroplast DNA compared to core DNA, with a k-mer analysis *Chan and Ragan, 2013*, this is for example done by *chloroExtractor* *Ankenbrand et al., 2018*. The third way to accomplish this is to combine as done by *NOVOPlasty* *Dierckxsens, Mardulyn, and Smits, 2017*. Figure 1.2 shows the general workflow of chloroplast assembly too.

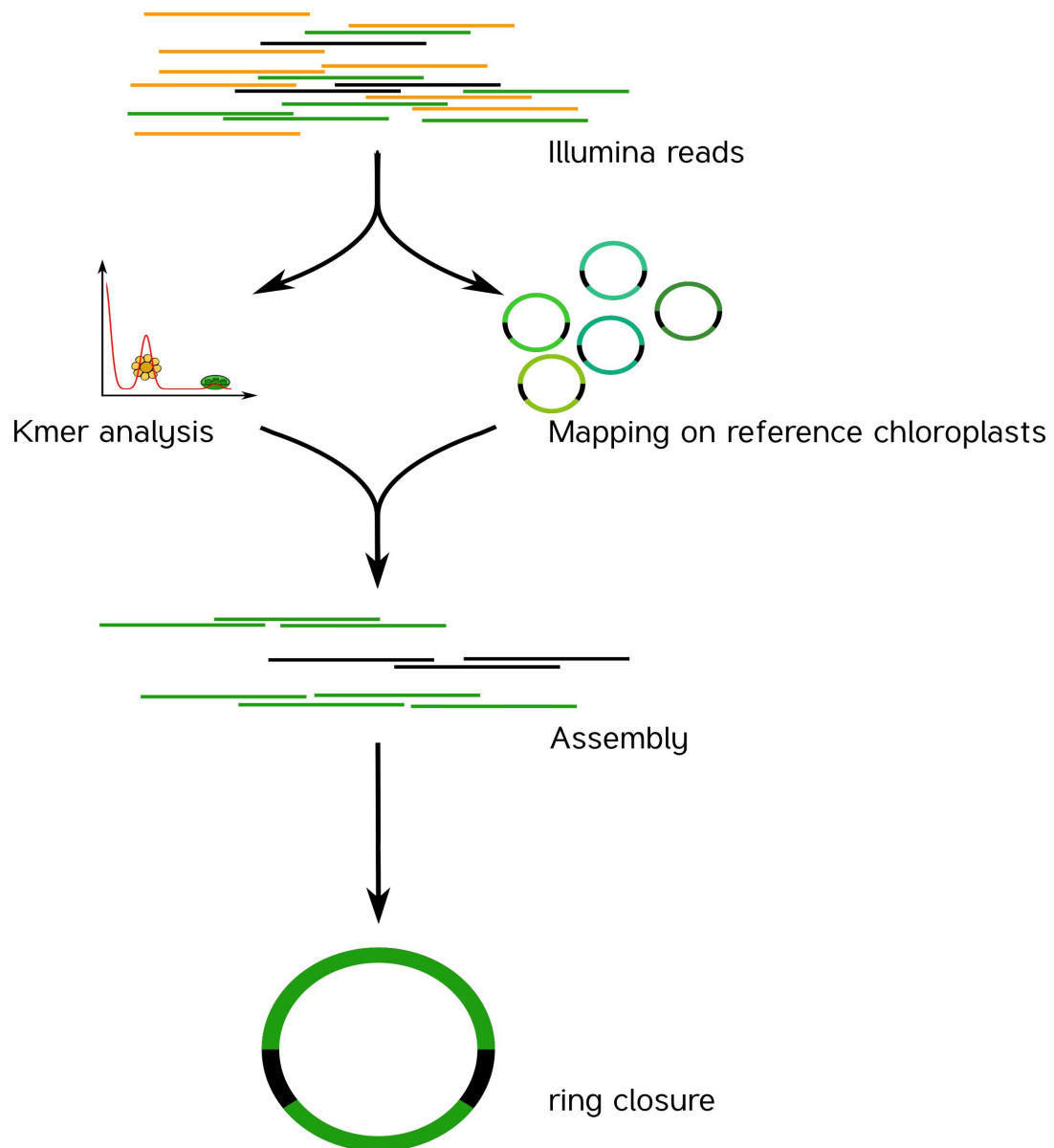


FIGURE 1.2: Standard workflow of chloroplast genome assembly.  
Graphic from Ankenbrand et al., 2018

### Purpose and scope of benchmarking the landscape of chloroplast assembly tools

The purpose of this study is to provide insights into the landscape of chloroplast assembly tools, to recommend best practices for organelle genome assemblies

and to provide *de novo* assemblies for many species and family's without a reference chloroplast available so far. To be included into this study the software including the source code must be publicly available and published under the terms of a liberal free open source software e.g. GPL or MIT-license. The study was also limited to paired-end Illumina data sets as their sole input source, because they are abundantly available for this benchmark. The seven tools that met the requirements and were therefore included in this study were: chloroExtractor, ORG.Asm, Fast-Plast, IOGA, NOVOPlasty, GetOrganelle and Chloroplast assembly protocol. As later thoroughly described in section 1.3, there are huge differences between those tools and some will not be recommended to be used in scientific applications, while others stand out for most but not all cases.

## 1.2 Material and Methods

### 1.2.1 Methods

#### Data and code availability

All the source code and data used in the scope of this study is publicly available under the terms of the MIT-License. The source code has been published on github [GitHub Repository for Benchmark Project](#) and archived on zenodo [Förster and Ankenbrand, 2019](#). The docker images are available on dockerhub [Docker Hub Group for Benchmark Project](#)

#### Tools

There were certain requirements to be met to be included into the study as aforementioned. The only technical requirement was, being able to assemble chloroplast genomes from paired-end Illumina reads. The other requirements were due to reproducibility. (i) The software must be open-source and available under

the terms of a liberal software license. In the authors opinion obscuring reproducibility under paywalls cannot be considered good scientific practice. (ii) It must be a command line tool, GUI-only tools are not suited for highly repetitive automated analyses. In total there were 7 tools that met those requirements: (i) chloroExtractor *Ankenbrand et al., 2018*; (ii) Chloroplast assembly protocol *Sancho et al., 2018*; (iii) GetOrganelle *Jin et al., 2018*; (iv) ORG.Asm *Coissac et al., 2016*; (v) IOGA *Bakker et al., 2016*; (vi) Fast-Plast *McKain and Afinit, 2017*; (vii) NOVOPlasty *Dierckxsens, Mardulyn, and Smits, 2017*. There are other tools available, that did not meet the requirements.

### Standardization and reproducibility

The main goal of the study was to provided deep insights into the landscape of chloroplast assembly tools. To accomplish that we tried to use the highest standards in bioinformatics when it comes to standardization and reproducibility. Along the study we also wanted to publish easy and ready-to-use versions of all the involved programs, working with standardized input. To accomplish that we decided to make us of containerization in form of docker containers *Merkel, 2014* to work with the containers in a closed HPC environments we decided a related software - singularity *Kurtzer, Sochat, and Bauer, 2017*. Therefore users do not have anything else to do than provide two files one for the forward reads (`forward.fq`) and one for the reverse reads (`reverse.fq`). Both files are required to be in FASTQ format. If docker or singularity are installed no further setup is requirement, because all dependencies and packages are installed and run from the containers on any given environment. Besides the individual output files, recording the process of the respective program, all programs write the assembly products in to files called `output.fa` in FASTA format. For the quantitative and consistency measurements the singularity containers were run on the Julia HPC-Cluster of the University of Würzburg using the SLURM workload manager *Jette, Yoo, and Grondona, 2002*. All runs for all assemblies were set with a time limit of

48h. This was necessary because some assemblers e.g. IOGA, if not finishing after at least 12 hours appear to continue in some kind of loop and never finish.

### 1.2.2 Data

Three different data sets were used for this study. (i) Simulated data from *A. thaliana* chloroplasts. (ii) real data with known reference chloroplast to rate the success of the assemblies. (iii) Novel data sets from NCBI's SRA without a known reference chloroplast to apply the gained knowledge for the *de novo* assembly of more than 100 chloroplasts.

#### Simulated

For first steps in any benchmarking process it is always useful to start with simulated data, where the investigators have full control over all the parameters involved. In the case of the present study the data's input parameters thought to be influential on the outcome where: The read length, the ratio between chloroplast and core genome reads. The simulations were based on data from the TAIR10 genome *Berardini et al., 2015* and performed using seqkit *Shen et al., 2016*. Ratios simulated were: 0:1, 1:10, 1:1000 and 1:1000, with read length of 150 and 250 bp. The simulated data consisted either of 2 million read pairs or the full data available. The simulation process is documented and the code and the data is available on github and zenodo *Ankenbrand and Förster, 2019*

#### Real data set

Real data was selected from the SRA database table 1.1 lists the search terms that had to be met according to the aforementioned search terms.



TABLE 1.1: Data selection criteria for real data sets from SRA

choice	option	explanation
green plants	Organism	include only photosynthetic plants e.g. no algae
wgs	Strategy	only data from wgs projects included
Illumina	Platform	include only paired-en Illumina reads
biomol DNA	Properties	include only biomol. DNA samples (e.g. no RNA)
paired	Layout	exclude single-end reads
random	Selection	
public	Access	Only publicly available data included

In total this resulted in 369 data sets representing a broad variety of the plant kingdom with many different families and genera included.

### Novel data sets

To assess the performance on assemblies without a published chloroplast on CpBase *CpBase* 105 data sets were selected from SRA, it was emphasized that such data sets were selected, with the next relatives with a reference chloroplast as distantly related as possible in taxonomic terms according to NCBI *NCBI Taxonomy*

## 1.2.3 Evaluation

### Quantitative

Each assembly from each assembler was compared to their respective reference genome by alignment using minimap2 *Li, 2018* and based on those alignments scores were calculated following equation 1.1 from 0 to 100 with 100 being a perfect score. Four different metrics contributed equally to the final score. (i) The coverage of the assembled genome compared to the reference genome  $cov_{ref}$  as an estimate for the completeness. (ii) The vice versa case  $cov_{qry}$  as a measure for the correctness of the assembly. (iii) The success of resolving the IR correct,

estimated from the size difference from the reference and the newly assembled genome is:  $\min \left\{ \frac{cov_{qry}}{cov_{ref}}, \frac{cov_{ref}}{cov_{qry}} \right\}$ . (iv) The number of total contigs were weighted as  $\frac{1}{n_{contigs}}$  giving a chloroplast with 1 contig the optimal score.

$$score = \frac{1}{4} \cdot \left( cov_{ref} + cov_{qry} + \min \left\{ \frac{cov_{qry}}{cov_{ref}}, \frac{cov_{ref}}{cov_{qry}} \right\} + \frac{1}{n_{contigs}} \right) \cdot 100 \quad (1.1)$$

While it is difficult to assess the success or failure of assemblies on a continuous scale, equation 1.1 allows for objective and unbiased measurements. SNPs or other small variants do not influence the outcome of the score, because it is much more likely that there are due to in-species variation of the plastid's genome and not caused by the assembly, and even if the latter is true it would be difficult to evaluate that.

## Consistency

In any given bioinformatical application consistency is a desired trait. All software (exceptions excluded) should given repeatedly the same input provide the same output. The same obviously or even more so holds true for assembly tools. To evaluate the reproducibility of the 7 tools. All the 369 real data sets described in section 1.2.2 has been used twice for assembly with each assembler and scored twice with equation 1.1. The correlation between the first and the second scores and runs was used the measure for the robustness of a program.

## 1.3 Results

### 1.3.1 Quantitative

#### Simulated data

The simulated data was assembled and scored with all the tools as described above. Figure 1.3 shows as a tile plot the results for all data sets and assemblers. While at first sight there is no clear correlation with the input data sets. It is obvious that there are grave differences between the assemblers. Two programs namely Chloroplast assembly protocol and IOGA failed to correctly assemble a single chloroplast genome. IOGA even fails to provide output at all for the majority of the data sets. While those two stand out as a negative example Fast-Plast and GetOrganelle stand out as a positive example perfectly or nearly perfectly assembling all the data sets, with GetOrganelle surpassing the performance of Fast-Plast. In the middle of the field are chloroExtractor, ORG.As and NOVOPlasty performing reasonably well, but sometimes lacking to solve the IRs and the circular structure.

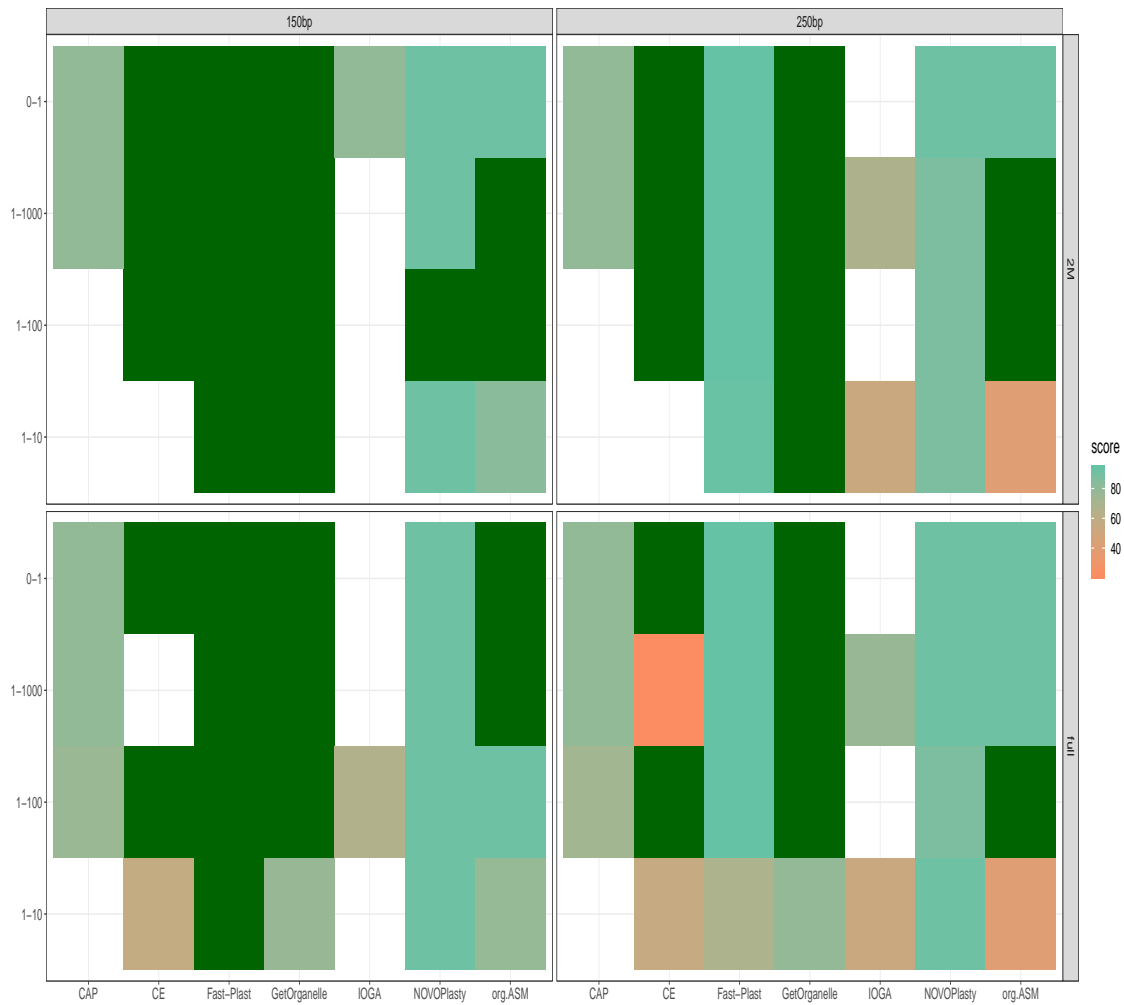


FIGURE 1.3: Results of assemblies executed with simulated data sets.

While there is a significant difference between the assemblers the same is not necessarily true in general for the other varying parameters. While Fast-Plast deals with the shorter reads of 150 bp much better than with the longer reads of 250 bp. The outcome of the other assemblies does not seem to be influenced by that. There is no difference between the full and the subsampled data sets. And while all assemblers appear to be more challenge by low chloroplast to core genome ratios of 1:10 larger ratios do not affect the quality of the result as expected. Table 1.2 shows all the individual results for all data sets and assemblers. For the fields with no entry the respective assembler failed to provide an output.

TABLE 1.2: Scores of assemblies of simulated data

	data set	CAP	CE	Fast-Plast	GetOrganelle	IOGA	NOVOPlasty	org.ASM
1	sim_150bp.0-1	79.10	100.00	99.48	100.00		91.52	100.00
2	sim_150bp.0-1.2M	79.10	100.00	99.72	100.00	79.10	91.52	91.50
3	sim_150bp.1-10		56.44	100.00	76.98		91.52	78.00
4	sim_150bp.1-10.2M			99.97	100.00		91.52	82.72
5	sim_150bp.1-100	75.72	100.00	99.48	100.00	66.09	91.52	91.50
6	sim_150bp.1-100.2M		100.00	99.47	100.00		100.00	100.00
7	sim_150bp.1-1000	79.10		99.72	100.00		91.52	100.00
8	sim_150bp.1-1000.2M	79.10	100.00	99.72	100.00		91.52	100.00
9	sim_250bp.0-1	79.10	100.00	93.82	100.00		91.52	91.50
10	sim_250bp.0-1.2M	79.10	100.00	93.83	100.00		91.52	91.50
11	sim_250bp.1-10		54.98	68.45	78.89	52.71	91.52	40.20
12	sim_250bp.1-10.2M			93.00	100.00	52.67	87.40	40.20
13	sim_250bp.1-100	72.81	100.00	93.82	100.00		87.40	100.00
14	sim_250bp.1-100.2M		100.00	93.83	100.00		87.40	100.00
15	sim_250bp.1-1000	79.10	21.30	93.83	100.00	76.96	91.52	91.50
16	sim_250bp.1-1000.2M	79.10	100.00	93.83	100.00	67.55	87.40	100.00

### Real data sets

Table 1.3 gives the results from the assemblies of 369 data sets with 7 assemblers. Similar to the results of the previous section there is a significant difference between the tools. Likewise GetOrganelle is the most successful assembler by a large margin with 210 of 369 perfectly assembled chloroplast genomes, failing to provide output for only 9 data sets, resulting in a media score > 99. Contrary to GetOrganelle, Chloroplast assembly protocol and IOGA fail to completely assemble a single genome. The performance of Fast-Plast could be considered reasonably well, completing approximately half as many genomes as GetOrganelle and being the only other tool whose average score is larger than 90. Similar to the trials with the simulated data in chapter 1.3.1 in the middle of the field are chloroExtractor, NOVOPlasty and ORG.Asm.

TABLE 1.3: Mean scores of chloroplast genome assemblers

	assembler	Median	IQR	N_perfect	N_tot
1	CAP	45.25	50.19	0	369
2	CE	56.55	71.50	14	369
3	Fast-Plast	92.80	23.59	113	369
4	GetOrganelle	99.83	20.94	210	360
5	IOGA	71.10	11.21	0	338
6	NOVOPlasty	75.95	48.69	58	369
7	org.ASM	67.35	91.69	46	348

Figure 1.4 emphasizes the large differences between the assemblers shown in table 1.3. The swarm plots show some distinct bands for some assemblers e.g. NOVOPlasty and ORG.Asm, suggesting that multiple assemblies fail to be solved into a single contig genome at a certain point. As thoroughly discussed in section 1.4 just from this swarm plot it is debatable if all the tools could be recommended for the purpose they were designed for.

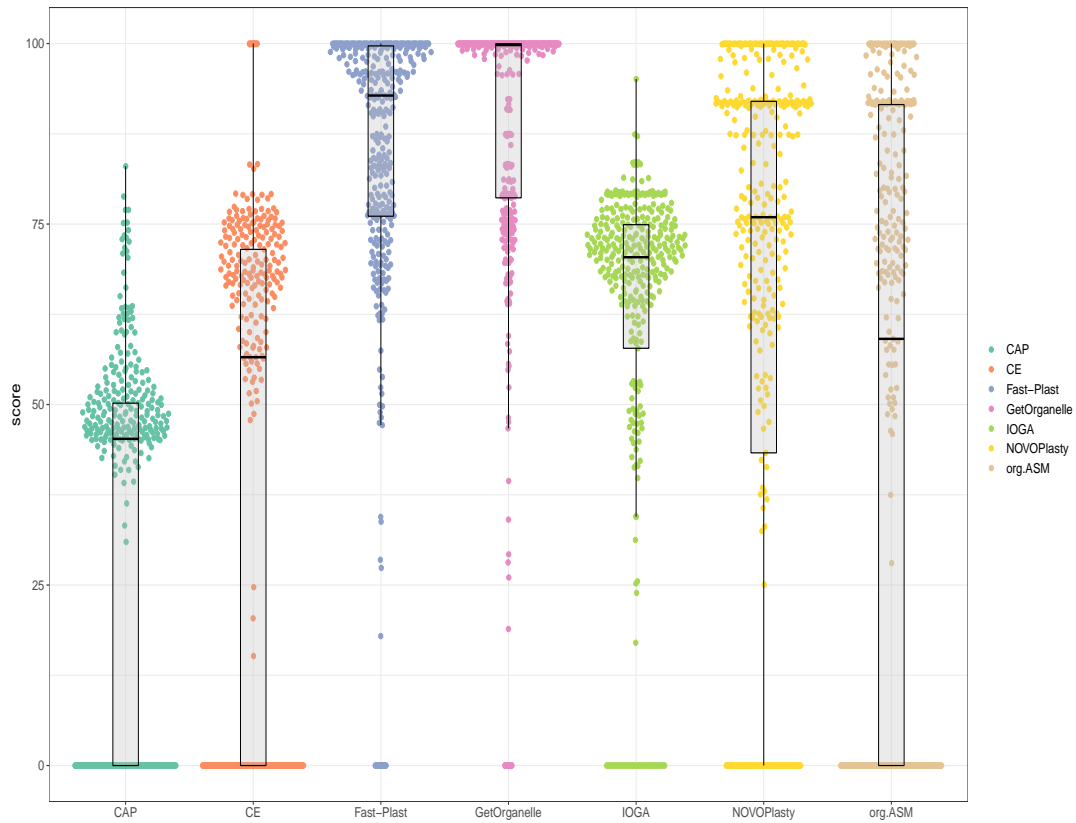


FIGURE 1.4: Box and swarm plots depict the results from the scoring shown in 1.1

## Consistency

Consistency testing was done by re-running every assembly for the real data sets twice and comparing the scores. `chloroExtractor` was the only tool that was 100 % consistent over 2 runs. Followed by `Fast-Plast` and `NOVOPlasty`. The consistency plot figure 1.5 for both of them results in an arrowhead shaped plot. With the main differences between the first and second run in the best scores. All other assemblers appear to produce the same output in the two runs except if they fail to complete the assembly at all. This is less pronounced for `Chloroplast` assembly protocol and `GetOrganelle` and is a grave issue for `ORG.ASM` and `IOGA`.

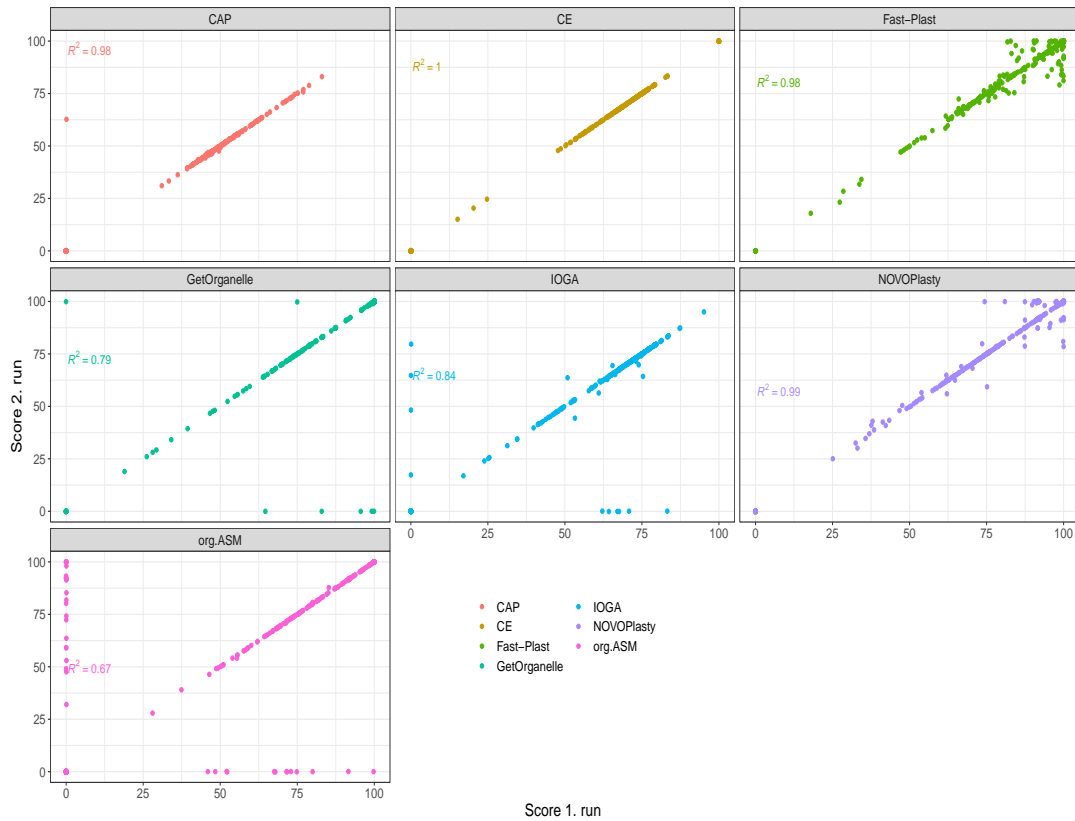


FIGURE 1.5: Swarm plots depict the results from the scoring shown in 1.1 for two independent runs for each assembler on each of the data sets

## Novel

The final assessment in the scope of the study was to test the assemblers on novel data sets without a published chloroplast. This step is important for two reasons. (i) There could be the possibility that certain tools perform well on known chloroplasts because they somehow have knowledge over those and simply apply it during those assemblies and therefore the previous results will not be able to generalize on unknown data. (ii) To apply and test the gained insights in the hope of providing the scientific community with a larger variety of published chloroplast genomes. Again the most successful assembler was GetOrganelle, with 49 out of 105 data sets completely assembled. Lacking a reference genome the success had to be defined differently and equation 1.1 was not suitable to evaluate the novel



assemblies. Metrics influencing the score of the novel assemblies were the number of contigs, solving the IRs and the size of the SSC and LSC. This known to the authors might be very biased and not true for all chloroplast and assumes that all chloroplast genomes follow the general structure for their genomes described in chapter 1.1. Figure 1.6 compares the results for the assemblies with at least one successful assembly.

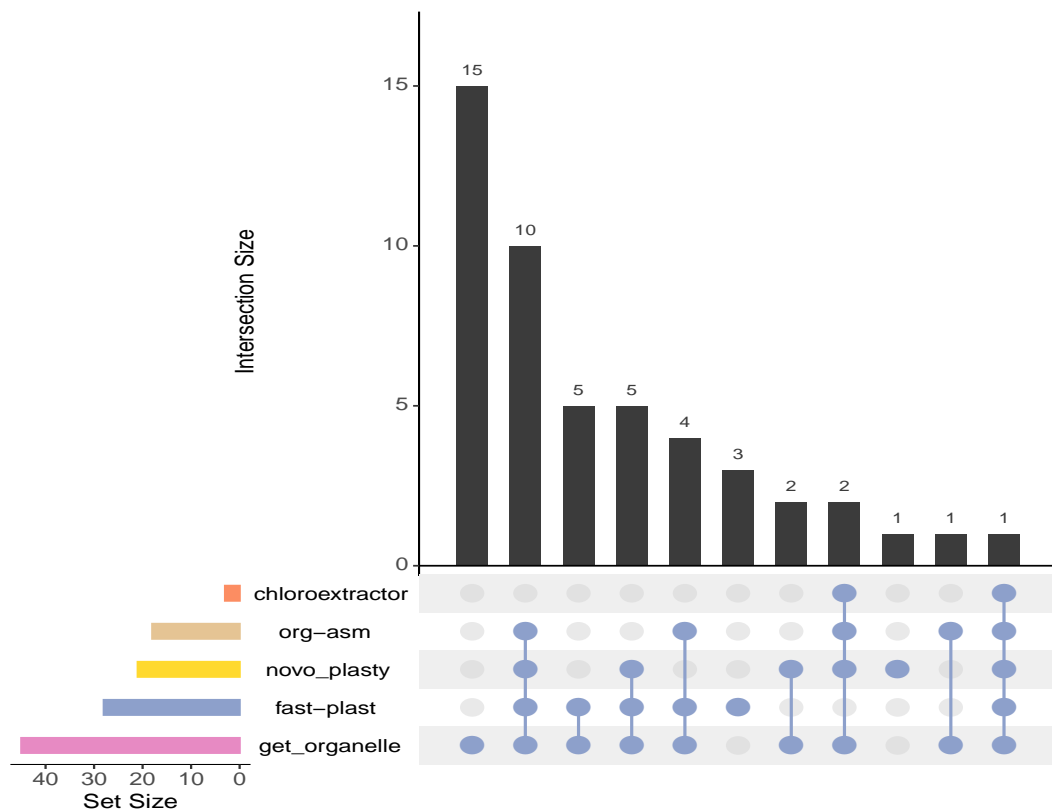


FIGURE 1.6: Upset plot comparing the success rates of the different assemblers

## 1.4 Discussion

The study compromised of two goals. (i) to assess the overall performance of a variety of tools designed specifically for the assembly of circular chloroplast genomes from paired-end Illumina data and (ii) to *de novo* assemble a variety of yet unpublished chloroplast genomes from existing data. To accomplish the

first goal 16 simulated and 369 simulated data sets were used adding up to a total of 5166 assemblies for the real data sets and 112 for the simulated data, along 735 assemblies for the novel data sets, thus underlying the statistical powers of this benchmarking study. The most successful tools were GetOrganelle and Fast-Plast which are recommended to be used complementary, because as shown in figure 1.7 they succeed for the most data sets compared to other assemblers and accomplish to satisfactory assemble chloroplast genomes where the other fails. If both of them fail it might be worthwhile to repeat the runs because other results could be expected as shown in the scatter plots of figure 1.5, especially Fast-Plast might be able to improve the previously achieved score. Only if both of them fail it might, even though improbable, that NOVOPlast might lead to a successful assembly. The other assemblers should be used with caution. While chloroExtractormight be good for a quick overview due to its relatively low demand in computational time *Freudenthal et al., 2019b*; Chloroplast assembly protocol, ORG.Asmand IOGA are not recommended to be used as the primary tools in organelle genome assembly projects, as used in this study.

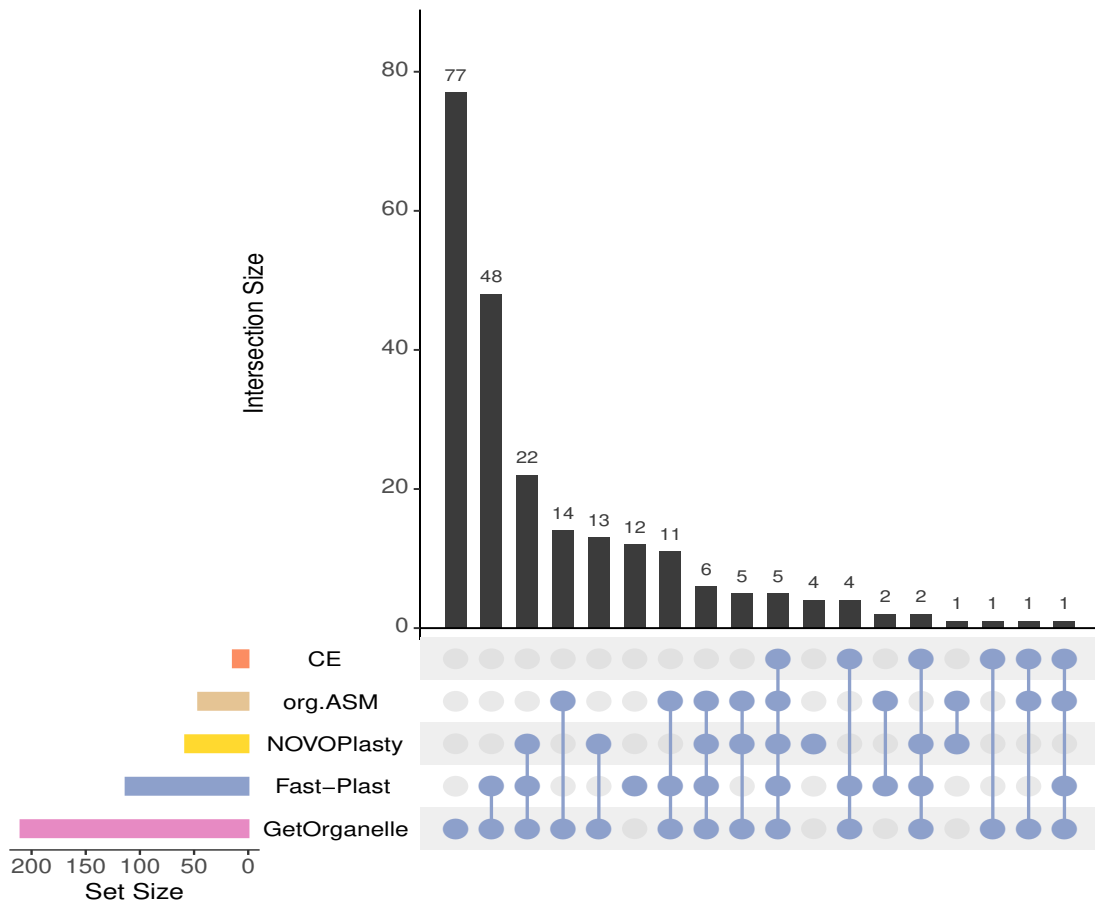


FIGURE 1.7: Upset plot showing the intersections of success rates between assemblers. A successful assembly was defined with a score  $> 99$  according to equation 1.1. The colored, horizontal barplot indicates the total number of successful assemblies for an individual tool. The black, vertical barplot gives the magnitude of the intersection between the assemblers indicated with the dot. Therefore there first bar is to be interpreted as follows: 77 data sets were only successfully assembled by chloroExtractor, likewise 48 genomes were assembled completely by GetOrganelle and Fast-Plast and so on.

It might be possible that overall performance of a specify tool might change significantly by fine tuning the input parameters of the tool, which was purposely not done in the scope of the present study, because this study was designed to

mimic the behavior have end-users and not developers of such tools and the assumption is proposed that end-users with little experience in bioinformatics would be inclined to use the basic configurations of such a tool.

While there is a huge differences for all assemblers they are presented with the same challenges and the bottlenecks are similar for each of them, the success rate of passing those differs however. Figure 1.8 shows the alignment of the 7 chloroplast genomes of *Oryza brachyantha* a grass distantly related to cultivated rice *Oryza sativa* and the respective reference genome. For the need of a linear representation of the circular genome the convention is to present chloroplast genomes in the order LSC - IRa - SSC -IRb. *O. brachyantha* was chosen because multiple tools successfully or at least almost assembled the full genome. Only Chloroplast assembly protocolis singled out, which only managed to assemble a few fragments on the SCC and the IRs on many contigs. A common mistake is to return 3 contigs as IOGA did. They represent the LSC the SSC and one IR but failed to resolve those regions into a one circular contig. GetOrganelle and Fast-Plast were able to reproduce the structure of the reference, while chloroExtractor flipped the LSC and NOVOPlastyand ORG.Asm were not able to construct the single contig into the conventional structure. All of this are common mistakes appearing more or less rare in all the assemblers. This could be a good starting point for the developers to further improve their tools. In this example all but Chloroplast assembly protocol were able to construct all the parts of the chloroplast's genome, and the main mistake was to resolve the structure of the genome into a circular, one contig version.

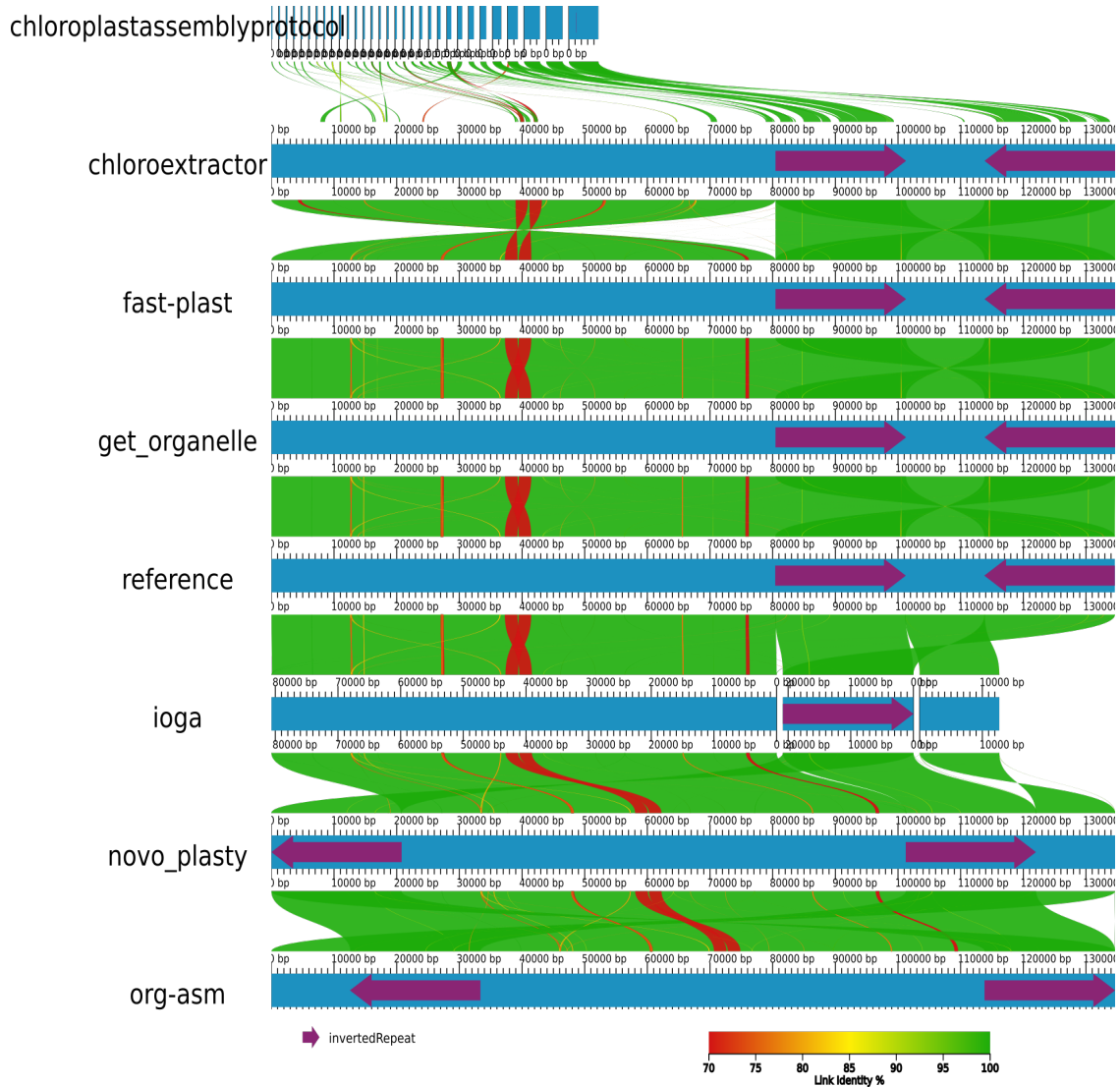


FIGURE 1.8: AliTV plot Ankenbrand et al., 2017 from Freudenthal et al., 2019b showing the alignments of *Oryza brachyantha* chloroplast genomes from all 7 assemblers. Regions in adjacent assemblies are connected with colored ribbons for similar regions in alignment with the identity coded as described in the legend. The purple arrows indicate the IR regions

## 1.5 Conclusion & outlook

Organelle genomics is promising field in plant genetics. As described in section 1.1 chloroplast genomes are well-suited for applications in evolutionary sciences, taxonomy and barcoding applications. Alike its mother branch genomics

for comparative chloroplast genomics its just as crucial to obtain high quality genomes. And the quality is mainly influenced by two major factors: the quality of the genome sequencing protocol and the quality of the assembly. As shown the latter varies massively between tools and not all tools are recommend equally from the conclusions drawn for the above experiments. All tools have room for improvement, this is meant to criticize the respectable work of the developers, but to encourage them to further develop tools and publish them under terms of liberal software licenses for the greater benefit of the entire scientific community.

## 2 Understanding the haplotype structure of *Arabidopsis thaliana*

### 2.1 Introduction

Recombination and LD in *A. thaliana* Kim et al., 2007 LD in *A. thaliana* Nordborg et al., 2002 Evolution of selfing Tang et al., 2007 Evolution and genetic differentiation among relatives of *Arabidopsis thaliana* Koch and Matschinger, 2007 FLC haplotypes Li et al., 2014

### 2.2 Haplotyping of *A. thaliana*

### 2.3 Results

### 2.4 Discussion

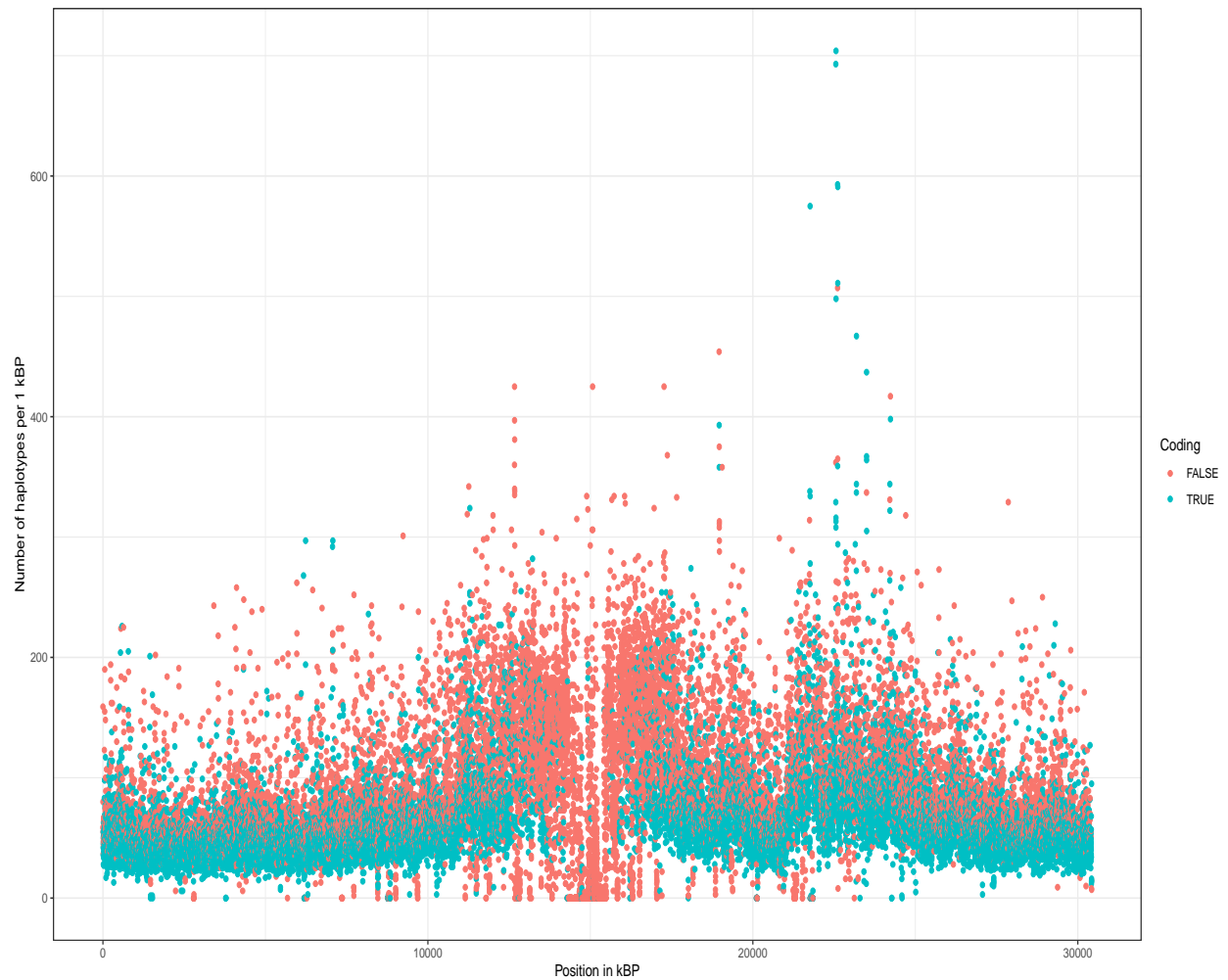


FIGURE 2.1: The number of segregating haplotypes with a polymorphism in at least one position over a stretch of 1 kBP.



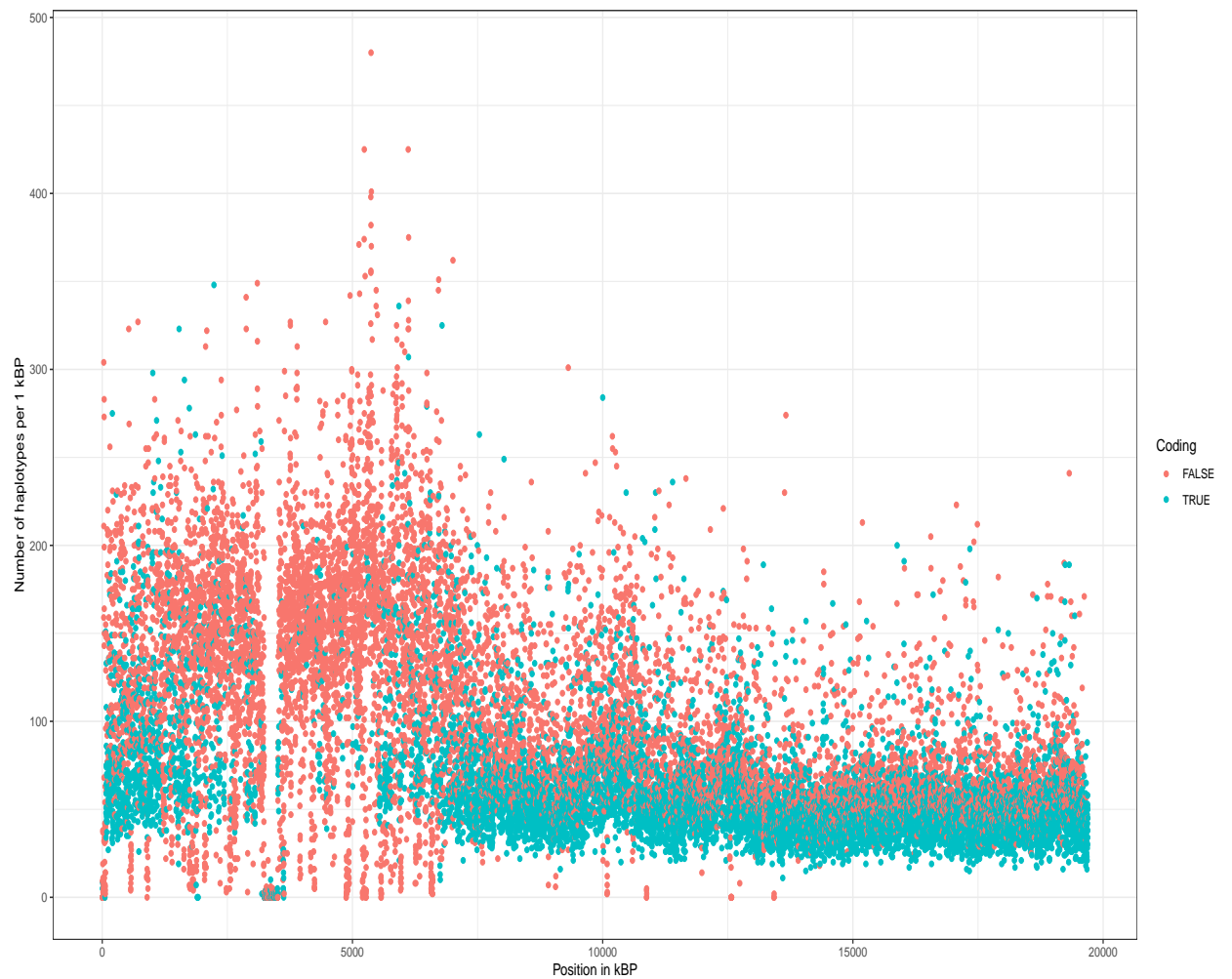


FIGURE 2.2: Number of segregating haplotypes with a polymorphism in at least one position over a stretch of 1 kBP.

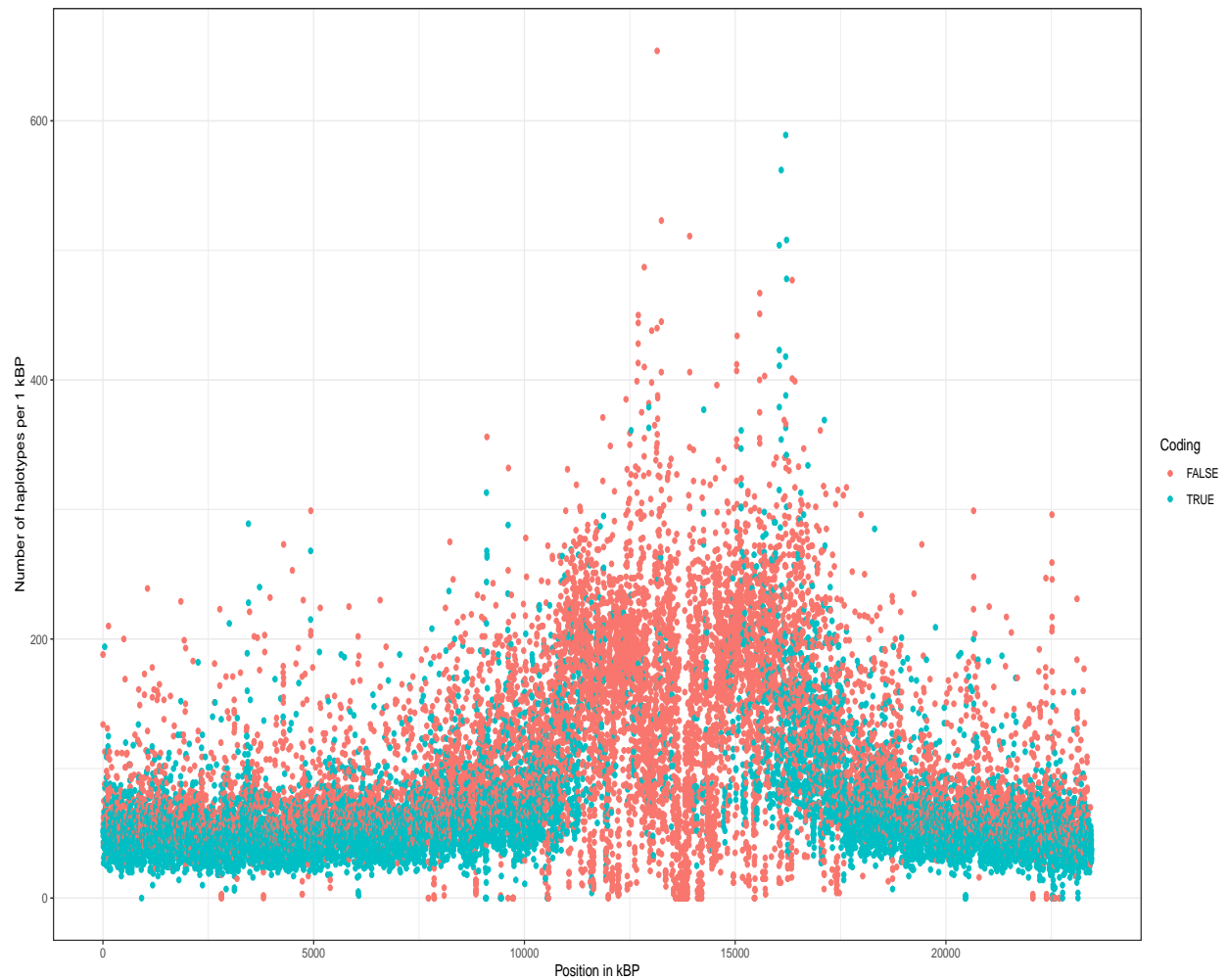


FIGURE 2.3: Number of segregating haplotypes with a polymorphism in at least one position over a stretch of 1 kBP.

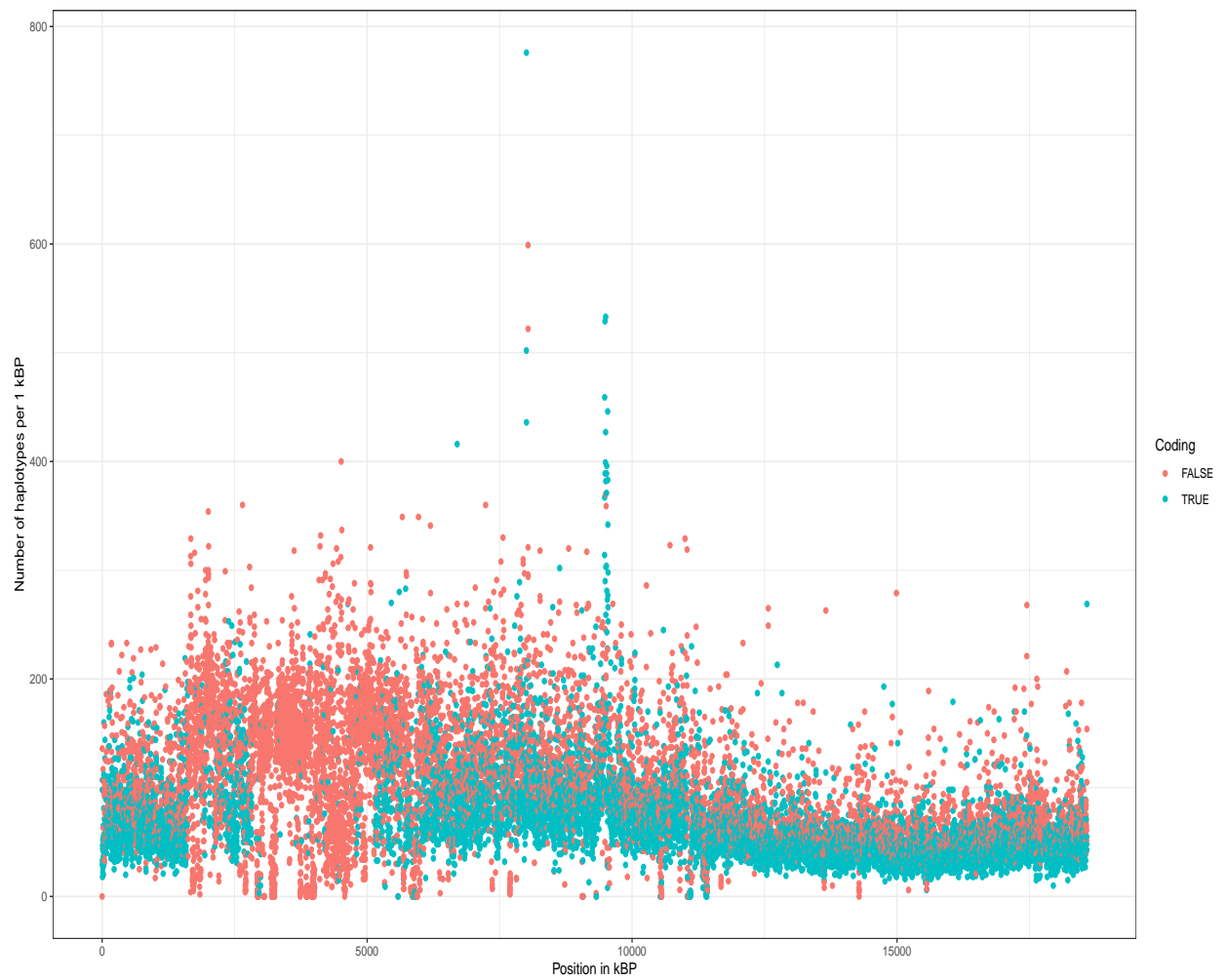


FIGURE 2.4: Number of segregating haplotypes with a polymorphism in at least one position over a stretch of 1 kBP.

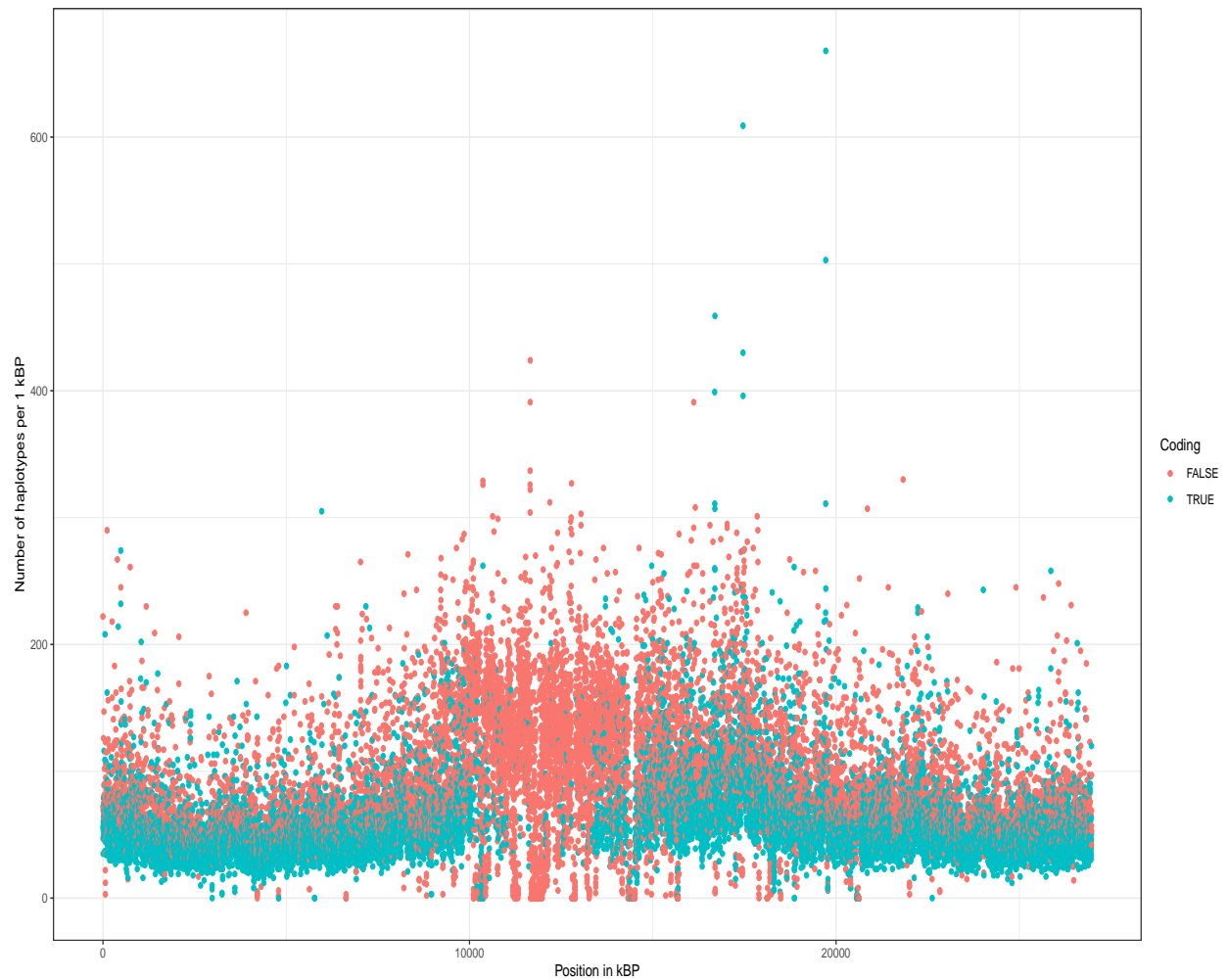


FIGURE 2.5: Number of segregating haplotypes with a polymorphism in at least one position over a stretch of 1 kBP.

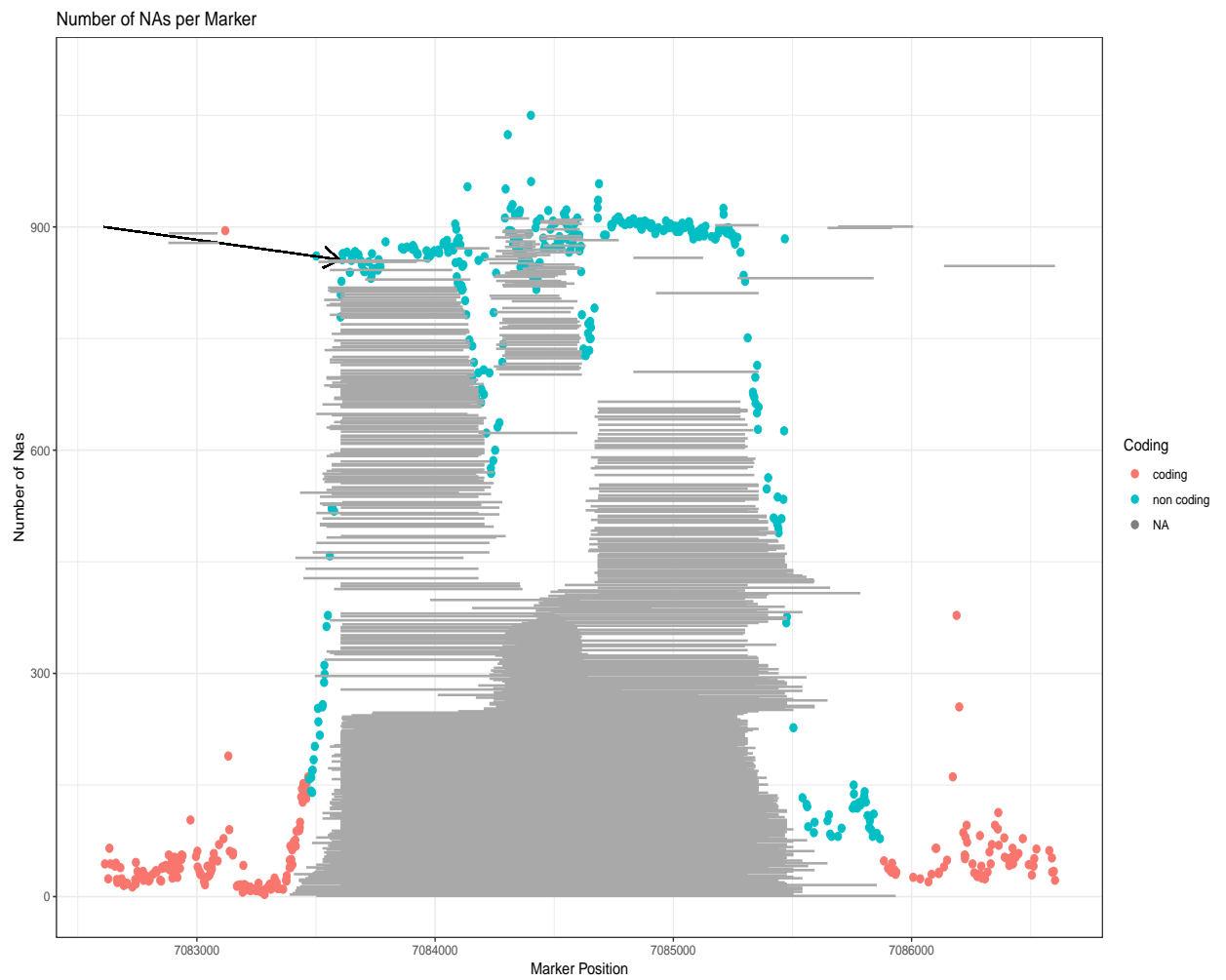


FIGURE 2.6: Number of segregating haplotypes with a polymorphism in at least one position over a stretch of 1 kBP.



## 3 GWAS-Flow a gpu-accelerated software for large-scale genome-wide association studies

The following chapter has been published in a similar version on the bioRxiv preprint server *Freudenthal et al., 2019a* and has been submitted for publication to Oxford Bioinformatics. The experiments and the software have been designed and conducted author. The manuscript has been prepared by the author, with minor corrections from Prof. Arthur Korte & Dominik Grimm. All authors approved of the final manuscript.

### 3.1 Introduction

Genome-wide association studies, pioneered in human genetics *Hirschhorn and Daly, 2005* in the last decade, have become the predominant method to detect associations between phenotypes and the genetic variations present in a population. Understanding the genetic architecture of traits and mapping the underlying genomic polymorphisms is of paramount importance for successful breeding

both in plants and animals, as well as for studying the genetic risk factors of diseases. Over the last decades, the cost for genotyping have been reduced dramatically. Early GWAS consisted of a few hundred individuals which have been phenotyped and genotyped on a couple of hundreds to thousands of genomic markers. Nowadays, marker density for many species easily exceed millions of genomic polymorphisms. Albeit commonly SNPs are used for association studies, standard GWAS models are flexible to handle different genomic features as input. The *Arabidopsis* 1001 genomes project features for example 1135 sequenced *Arabidopsis thaliana* accessions with over 10 million genomic markers that segregate in the population *Alonso-Blanco et al., 2016*. Other genome projects also yielded large amounts of genomic data for a substantial amount of individuals, as exemplified in the 1000 genomes project for humans *Siva, 2008*, the 2000 yeast genomes project or the 3000 rice genomes project *Li, Wang, and Zeigler, 2014*. Thus, there is an increasing demand for GWAS models that can analyze these data in a reasonable time frame. One critical step of GWAS is to determine the threshold at which an association is termed significant. Classically the conservative Bonferroni threshold is used, which accounts for the number of statistical tests that are performed, while many recent studies try to significance thresholds that are based on the false-discovery rate (FDR) *Storey and Tibshirani, 2003*. An alternative approach are permutation-based thresholds *Che et al., 2014*. Permutation-based thresholds estimate the significance by shuffling phenotypes and genotypes before each GWAS run, thus any signal left in the data should not have a genetic cause, but might represent model mis-specifications or uneven phenotypic distributions. Typically this process is repeated hundreds to thousands of times and will lead to a distinct threshold for each phenotype analyzed *Togninalli et al., 2017*. The computational demand of permutation-based thresholds is immense, as per analysis not one, but at least hundreds of GWAS need to be performed. Here the main limitation is the pure computational demand. Thus, faster GWAS models could easily make the estimation of permutation-based thresholds the default



choice.

## 3.2 Methods

### GWAS Model

The GWAS model used for GWAS-Flow is based on a fast approximation of the linear-mixed-model described in *Kang et al., 2010; Zhang et al., 2010*, which estimates the variance components  $\sigma_g$  and  $\sigma_e$  only once in a null model that includes the genetic relationship matrix, but no distinct genetic markers. These components are thereafter used for the tests of each specific marker. Here, the underlying assumption is, that the ratio of these components stays constant, even if distinct genetic markers are included into the GWAS model. This holds true for nearly all markers and only markers which possess a big effect will alter this ratio slightly, where now  $\sigma_g$  would become smaller compared to the null model. Thus, the p-values calculated by the approximation might be a little higher (less significant) for strongly associated markers.

### The GWAS-Flow Software

The GWAS-Flow software was designed to provide a fast and robust GWAS implementation that can easily handle large data and allows to perform permutations in a reasonable time frame. Traditional GWAS implementations that are implemented using Python *Van Rossum and Drake Jr, 1995* or R *R Core Team, 2019* cannot always meet these demands. We tried to overcome those limitations by using TensorFlow *Abadi et al., 2015*, a multi-language machine learning framework published and developed by Google. GWAS calculations are composed of a series of matrix computations that can be highly parallelized, and easily integrated into the architecture provided by TensorFlow. Our implementation allows both, the classical parallelization of code on multiple processors (CPUs) and the

use of graphical processing units (GPUs). GWAS-Flow is written using the Python TensorFlow API. Data import is done with *pandas* McKinney, 2010 and/or *HDF5* for Python Collette, 2013. Preprocessing of the data (e.g filtering by minor Allele count (MAC)) is performed with *numpy* Oliphant, 2006. Variance components for residual and genomic effects are estimated with a slightly altered function based on the Python package *limix* Lippert et al., 2014. The GWAS model is based on the following linear mixed model that takes into account the effect of every marker with respect to the kinship:

$$Y = \beta_0 + X_i\beta_i + u + \epsilon, u \sim N(0, \sigma_g K), \epsilon \sim N(0, \sigma_e I) \quad (3.1)$$

From this LMM the residual sum of squares for marker  $i$  are calculated as described in 3.2

$$RSS_i = \sum Y - (X_i\beta_0 + I_i\beta_1) \quad (3.2)$$

The residuals are used to calculate a p-value for each marker according to an overall F-test that compares the model including a distinct genetic effect to a model without this genetic effect:

$$F = \frac{RSS_{env} - R1_{full}}{\frac{R1_{full}}{n-3}} \quad (3.3)$$

Apart from the p-values that derive from the F-distribution, GWAS-Flow also report summary statistics, such as the estimated effect size ( $\beta_i$ ) and its standard error for each marker.

### Calculation of permutation-based thresholds for GWAS

To calculate a permutation-based threshold, we essentially perform  $n$  repetitions ( $n > 100$ ) of the GWAS on the same data with the sole difference that before each GWAS we randomize the phenotypic values. Thus any correlation between the

phenotype and the genotype will be broken and indeed for over 90% of these analyses the estimated pseudo-heritability is close to zero. On the other hand, the phenotypic distribution will stay unaltered by this randomization. Hence, any remaining signal in the GWAS has to be of a non-genetic origin and could be caused by e.g. model mis-specifications. Now we take the lowest p-value (after filtering for the desired minor allele count) for each permutation and take the 5% lowest value as the permutation-based threshold for the GWAS.

### Benchmarking

For benchmarking of GWAS-Flow we used data from the *Arabidopsis* 1001 Genomes Project *Alonso-Blanco et al., 2016*. The genomic data we used were subsets between 10,000 and 100,000 markers. We chose not to include subsets that exceed 100,000 markers, because there is a linear relationship between the number of markers and the computational time demanded, as all markers are tested independently. We used phenotypic data for flowering time at ten degrees (FT10) for *A. thaliana*, published and downloaded from the AraPheno database *Seren et al., 2016*. We down- and up-sampled sets to generate phenotypes for sets between 100 and 5000 accessions. For each set of phenotypes and markers we ran 10 permutations to assess the computational time needed. All analyses have been performed with a custom R script that has been used previously *Togninalli et al., 2017*, GWAS-Flow using either a CPU or a GPU architecture and GEMMA *Zhou and Stephens, 2012*. GEMMA is a fast and efficient implementation of the mixed model that is broadly used to perform GWAS. All calculations were run on the same machine using 16 i9 virtual CPUs. The GPU version ran on an NVIDIA Tesla P100 graphic card. Additionally to the analyses of the simulated data, we compared the times required by GEMMA and both GWAS-Flow implementations for > 200 different real datasets from *A. thaliana* that have been downloaded from the AraPheno *Seren et al., 2016* database and have been analyzed with the available fully imputed genomic dataset of ca. 10 million markers, filtered for a minor

allele count greater five.

### 3.3 Results

The two main factors influencing the computational time for GWAS are the number of markers incorporated in such an analysis and the number of different accessions, while the latter has an approximate quadratic effect in classical GWAS implementations *Zhou and Stephens, 2012*. Figure 1A shows the time demand as a function of the number of accessions used in the analysis with 10,000 markers. The quadratic increase in time demand is clearly visible for the custom R implementation, as well as for the CPU-based GWAS-Flow implementation and *GEMMA*. The GWAS-Flow implementation and *GEMMA* clearly outperforms the R implementation in general, while for a small number of accessions GWAS-Flow is slightly faster than *GEMMA*. For the GPU-based implementation the increase in run-time with larger sample sizes is much less pronounced. While for small ( $< 1,000$  individuals) data, there is no benefit compared to running GWAS-Flow on CPUs or running *GEMMA*, the GPU-version clearly outperforms the other implementations if the number of accessions increases. Figure 1B shows the computational time in relation to the number of markers and a fixed amount of 2000 accessions for the two different GWAS-Flow implementations. Here, a linear relationship is visible in both cases. To show the performance of GWAS-Flow not only for simulated data, we also run both implementations on more than 200 different real datasets downloaded from the AraPheno database. Figure 1C shows the computational time demands for all analyses comparing both GWAS-Flow implementation to *GEMMA*. Here, the CPU-based GWAS-Flow performs comparable to *GEMMA*, while the GPU-based implementation outperforms both, if the number of accessions is above 500. Importantly all obtained GWAS results (p-values, beta estimates and standard errors of the beta estimates) are nearly (apart from

some mathematical inaccuracies) identical between the three different implementations.

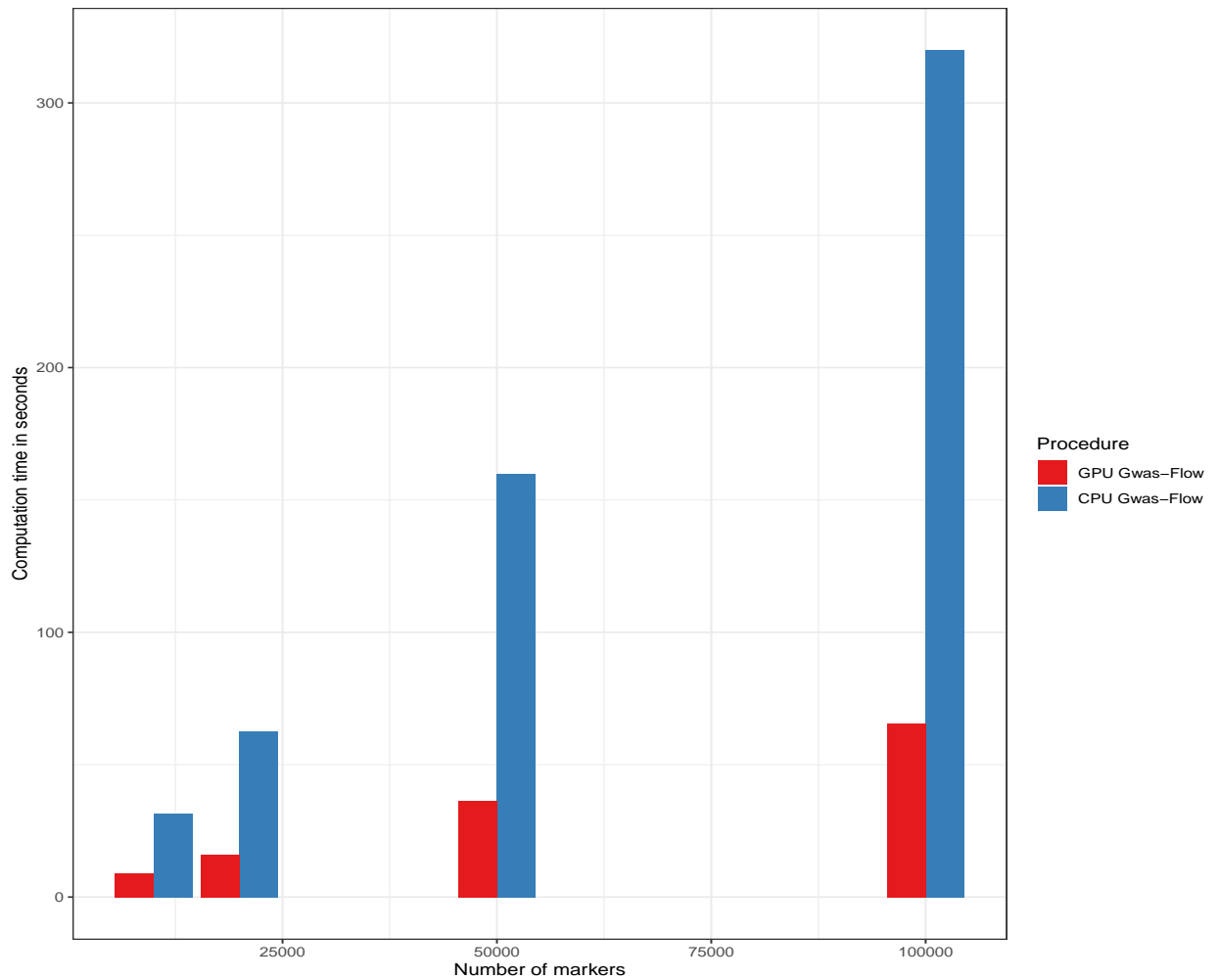


FIGURE 3.1: Computational time as a function of the number of genetic markers with constantly 2000 accessions for both GWAS-Flow versions

## 3.4 Disucssion

We made use of recent developments of computational architecture and software to cope with the increasing computational demand in analyzing large GWAS datasets. With GWAS-Flow we implemented both, a CPU- and a GPU-based version of the classical linear mixed model commonly used for GWAS. Both implementations outperform custom R scripts on simulated and real data. While the

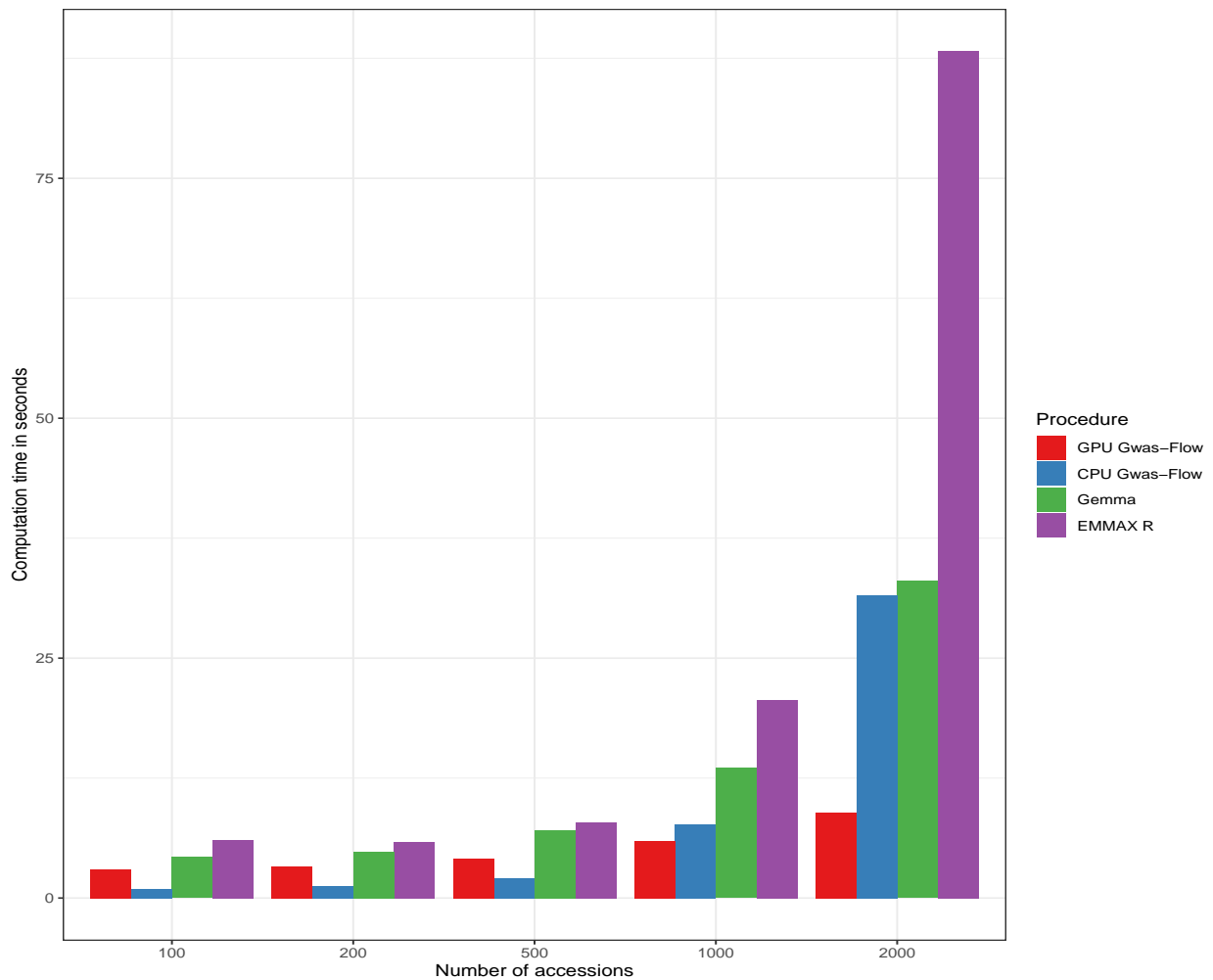


FIGURE 3.2: Computational time as a function of the number of accessions with 10000 markers each.

CPU-based version performs nearly identical compared to *GEMMA*, a commonly used GWAS implementation, the GPU-based implementation outperforms both, if the number of individuals, which have been phenotyped, increases. For analyzing big data, here the main limitation would be the RAM of the GPU, but as the individual test for each marker are independent, this can be easily overcome programmatically. The presented GWAS-Flow implementations are markedly faster compared to custom GWAS scripts and even outperform efficient fast implementations like *GEMMA* in terms of speed. This readily enables the use of permutation-based thresholds, as with GWAS-Flow hundred permutations can be performed in

a reasonable time even for big data. Thus, it is possible for each analyzed phenotype to create a specific, permutation-based threshold that might present a more realistic scenario. Importantly the permutation-based threshold can be easily adjusted to different minor allele counts, generating different significance thresholds depending on the allele count. This could help to distinguish false and true associations even for rare alleles. GWAS-Flow is a versatile and fast software package. Currently GWAS-Flow is and will remain under active development to make the software more versatile. This will e.g. include the compatibility with TensorFlow v2.0.0 and enable data input formats, such as PLINK *Purcell et al., 2007*. The whole framework is flexible, so it is easy to include predefined co-factors e.g. to enable multi-locus models *Segura et al., 2012* or account for multi-variate models like the multi-trait mixed model *Korte et al., 2012*. Standard GWAS are good in detecting additive effects with comparably large effect sizes, but lack the ability to detect epistatic interactions and their influence on complex traits *McKinney and Pajewski, 2012; Korte and Farlow, 2013*. To catch the effects of these gene-by-gene or SNP-by-SNP interactions, a variety of genome-wide association interaction studies (GWAIS) have been developed, thoroughly reviewed in *Ritchie and Van Steen, 2018*. Here, GWAS-Flow might provide a tool that enables to test the full pairwise interaction matrix of all SNPs. Although this might be a statistic nightmare, it now would be computationally feasible.

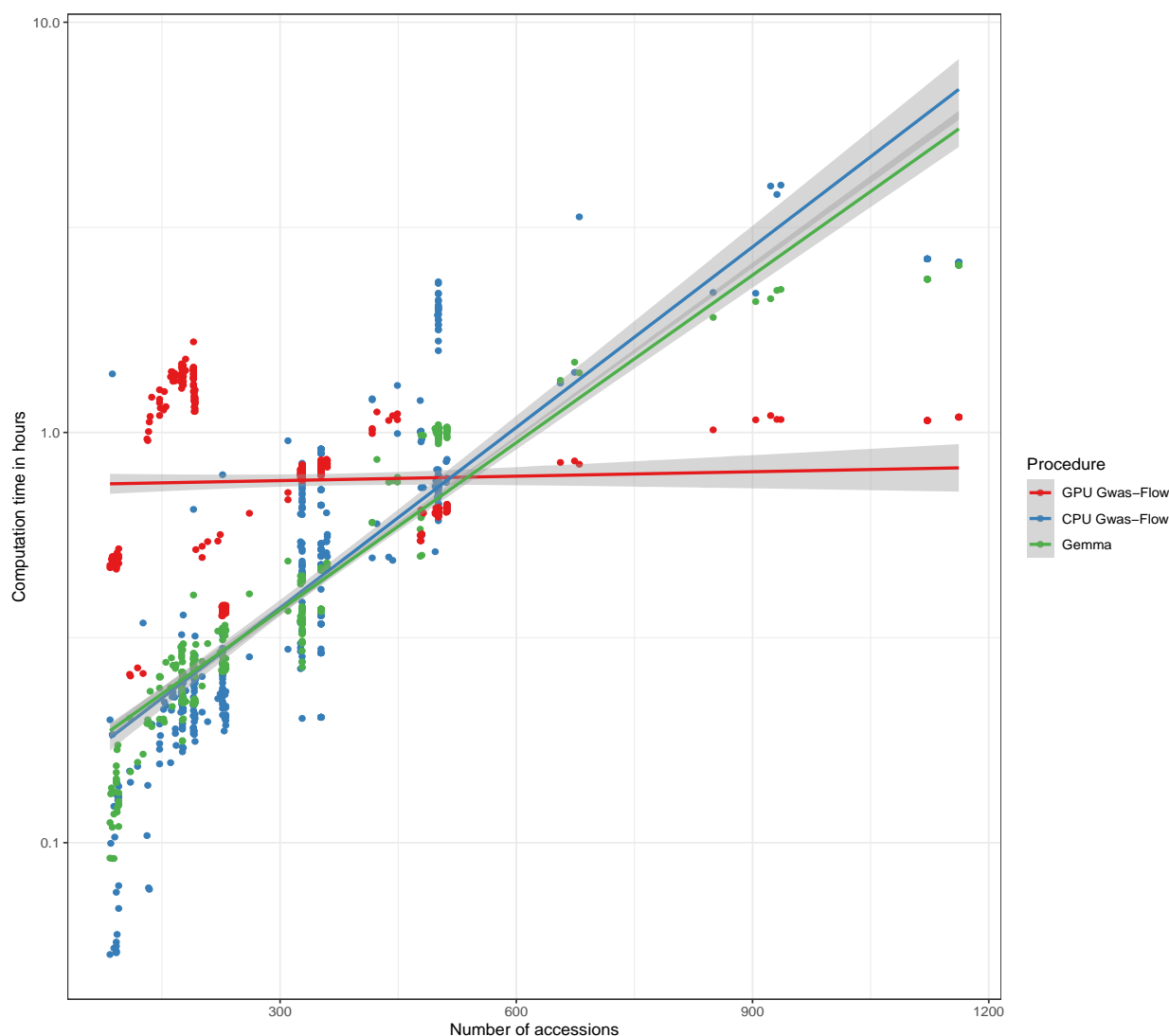


FIGURE 3.3: Comparison of the computational time for the analyses of  $> 200$  phenotypes from *Arabidopsis thaliana* as a function of the number of accessions for GEMMA and the CPU- and GPU-based version of GWAS-Flow. GWAS was performed with a fully imputed genotype matrix containing 10.7 M markers and a minor allele filter of  $MAC > 5$



## 4 Genomic prediction of phenotypic values of quantitative traits using artificial neural networks

### 4.1 Introduction

#### 4.1.1 A brief history of machine learning

##### Basic perceptron model

While machine learning, neural networks, deep learning became essential tools for many applications in more recent years, their mathematical principals date back to the early 1950s and 1960s. Figure 4.1 schematically show the basic perceptron model as proposed by Rosenblatt, which was designed to mimic the information flow in biological nervous systems *Rosenblatt, 1961*.

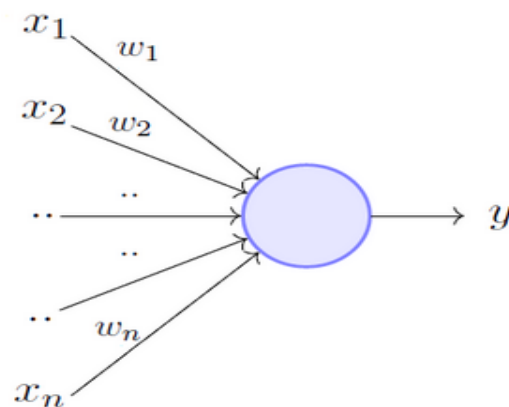


FIGURE 4.1: Basic perceptron model as proposed by Rosenblatt

This basic perceptron, which contrary to perceptrons used nowadays does not have an embedded activation function, takes  $n$  binary inputs  $x_1, x_2, \dots, x_n$  and produces a single, likewise binary, output  $y$  after being processed by the perceptron or neuron. To achieve this Rosenblatt introduced the concept of weights which indicated a certain relative importance to the outcome of the output  $w_1, w_2, \dots, w_n$ . The output  $y$  is determined by the weighted sum of the weights and biases  $\sum_i w_i x_i$ . If a certain threshold value is met the neuron is either activated and outputs 1 or not activated resulting in an output of 0. This is algebraically represented in equation 4.1:

$$0 = \text{if } \sum_i^n w_i x_i - \theta \leq 0 \quad (4.1a)$$

$$1 = \text{if } \sum_i^n w_i x_i - \theta > 0 \quad (4.1b)$$

Next to the weights  $w_n$  and the inputs  $x_n$  a third term  $\theta$  is introduced in equation 4.1 which represents the activation threshold and per definition is a negative value. A single perceptron is a linear classifier and can only be trained on linearly separable functions and can be used as shown by Rosenblatt, 1961 to solve simple logical operations as AND, OR and not. The simple perceptron fails, due to non-linearity, to perform XOR operations as shown by Marvin and Seymour, 1969. This discovery led to a near stillstand in the research of artificial neural networks in the 1970s. This time period is now often referred to as the first AI-winter. Another reason that massively hindered the applications and research of machine learning during that time, was the compared to modern times the incredibly small amount of computational power available Nguyen and Widrow, 1990.

More complex decision making, like solving XOR problems, requires more complex structures than a single perceptron. Continuing the trend of mimicking human neural networks, multiple artificial neurons are stacked into layers and these

layers, are connected to each other allowing communication between the many perceptrons in a such generated network. Figure 4.2 shows schematically the basic structure of such a network, now harboring three types of layers. (i) the input layer, (ii) one or more hidden layers and (iii) one output layer, which in this case only consists of one neuron.

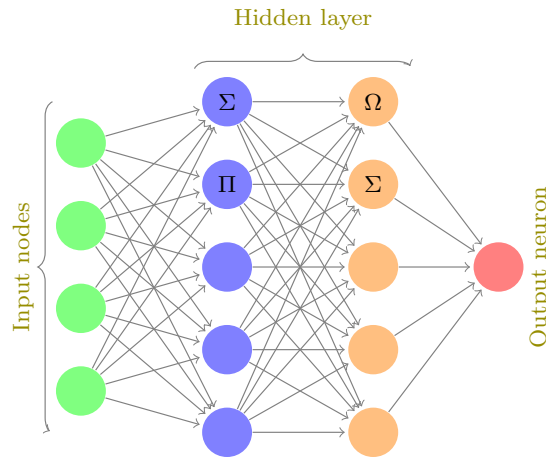


FIGURE 4.2: Schematic layout of a simple multi-layer perceptron

In the sample layout of figure 4.2 the neurons in the first column weigh the inputs and pass those the the neurons in the second layer. In this case all neurons on the first layer or connected to all neurons on the second layer, such layers are referred to fully-connected layers (FLC) , and their resulting networks are often called multi-layer perceptrons (MLP). This architecture enables the network to perform more complex calculations and result in more abstract decisions than single neurons or single layer architectures. There are also layers were neurons in the previous layer are only connected with neighboring neurons in the succeeding layers. Those are known as locally-connected layers (LCL) or related to them are convolutional layers which shared weights between selected neurons, building convolutional neural networks (CNN) *LeCun et al., 1999* .

## Activation functions

The neurons discussed so far are only capable of outputting binary results. Either 0 or 1, depending on whether threshold values are being reached or not. For more complex estimations it is desirable that small changes in the input also result in small changes of the output. This requirement can not be met with binary outputs. Activation functions for a given node provide more sophisticated rules for the output in accordance to their inputs *Žilinskas, 2006*.

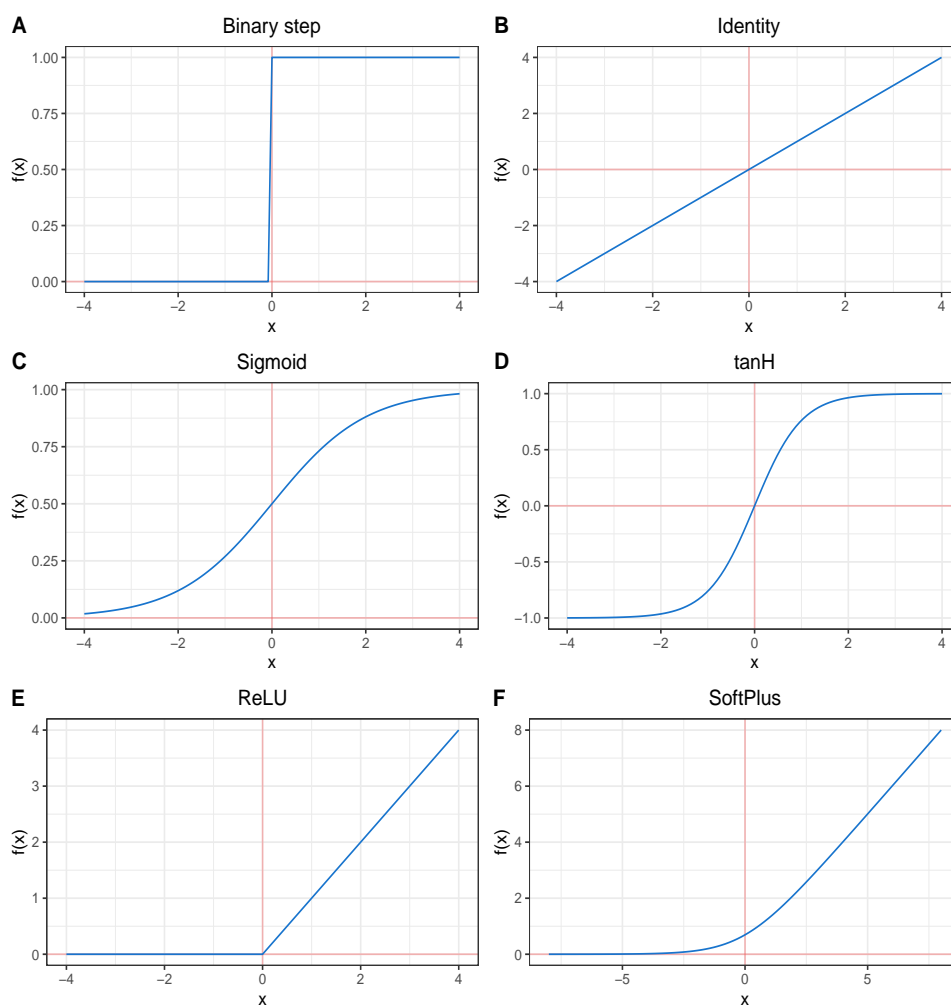


FIGURE 4.3: Popular activation functions used in neural networks. **A** Binary step activation function. **B** Identity activation function. **C** Sigmoid or logistic activation function. **D** tangens hyperbolicus activation function. **E** rectified linear units activation function . **F** SoftPlus activation function.

Figure 4.3 shows six of the most commonly used activation functions Warner and Misra, 1996. The simplest one was introduced, is the binary step activation function A in equation 4.2, which properties have been discussed along the perceptron model. All other activation produce continuous outputs from any given input. Any mathematical function can serve as an activation function in neural nets, starting with a simple identity function 4.3, 4.3 B. The sigmoid functions 4.3 C, equation 4.4 and tanh figure 4.3 D, equation 4.5, when  $x \rightarrow \infty$  or  $x \rightarrow -\infty$  they have similar properties to the binary function, but produce continuous output around 0.

$$f(x) = \sigma(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases} \quad (4.2)$$

$$f(x) = \sigma(x) = x \quad (4.3)$$

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \quad (4.4)$$

$$f(x) = \sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (4.5)$$

$$f(x) = \sigma(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (4.6)$$

$$f(x) = \ln(1 + e^x) \quad (4.7)$$

ReLU (equation 4.6) and the softplus (equation 4.7) share similar properties as well, the latter one being a smoothed version of ReLU. Rectifiers as activation functions have been introduced in 2000s Hahnloser et al., 2000 and have since then overtaken all others as the most popular activations functions in neural networks and deep learning today LeCun, Bengio, and Hinton, 2015 and they have proven

to be superior in many deep-learning applications over sigmoid or logistic functions. One of the advantages leading to the superiority of ReLUs is that with randomly initialized weights only half of the ReLU neurons are activated at start, compared to tanh and sigmoid activation *Glorot, Bordes, and Bengio, 2011*. All activation functions shown in figure 4.3, but the binary step function, share one common property: a small change of the input weight will result in small changes in the output, while a small change of the input for the binary step function leads to either no or a complete change of the output. This property is, as described below, is an important prerequisite for networks being able to learn.

### Gradient descent algorithm

Let the network shown in 4.2 be for the classification of a arbitrary phenotype like blue petals with  $x_1 \dots x_4$  on the input layers being genetic markers as features. And the output layer displaying a value from 0 to 1, meaning yes: blue petals from 0 - 0.5 and no blue petals from 0.5 to 1. To quantify how well the network performs on achieving that goal a loss function is applied *Schmidhuber, 2015*. There is a large variety of different loss functions available for neural networks like mean squared error (MSE), root mean squared error (RMSE), cross-entropy and many others. In general MSE and RMSE are commonly used for regression problems, with the latter being less popular and cross-entropy also called log-loss is used for binary or multi-class classification problems *Janocha and Czarnecki, 2017*. Since all problems presented in due course are regression problems, that use MSE as their loss function, this will be the only loss function used.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y})^2 \quad (4.8)$$

Equation 4.8 shows the MSE function which is the sum of the squares of the differences of all the predicted and the real values. The same function can be

rewritten with the previously used terminology of weights and biases in equation 4.8.

$$L(w, b) = \frac{1}{2n} \sum_x \|y(x) - \tilde{y}\|^2. \quad (4.9)$$

With  $w$  and  $b$  as the collection of all the weights and the biases in the network used to optimize the function  $y(x)$ . Giving the quadratic nature of the function the  $L(w, b)$  will always be positive. And if  $L(w, b) \rightarrow 0$  the loss is minimal, meaning that the real and predicted values are close together and the network found weights and biases that explain the output well.

A widely used function to find the optimum for such a loss function is gradient descent. Its objective is to find the minimum for the loss function *Bottou, 1991*. The behind gradient descent or other optimizing algorithms is start with randomly initialized weights and biases and repeatedly move them in direction  $\Delta w$  and  $\Delta b$ . This results in a change of the loss function as shown in equation 4.10, making use of partial derivatives.

$$\Delta L = \frac{\partial L}{\partial w} \Delta w + \frac{\partial L}{\partial b} \Delta b \quad (4.10)$$

Ideally  $\Delta L$  is negative and the optimization algorithm found  $\Delta w$  and  $\Delta b$  that lead to a reduction of the loss. To simplify this problem let  $\Delta d$  be the vector of changes:  $\Delta d = (\Delta w, \Delta b)^T$  and  $\nabla L$  the vector of the partial derivatives: equation 4.11

$$\nabla L = \left( \frac{\partial L}{\partial w}, \frac{\partial L}{\partial b} \right)^T \quad (4.11)$$

Having defined  $\nabla L$  and  $\Delta d$  the term 4.10 can be simplified as equation 4.12

$$\Delta C = \nabla L * \Delta d \quad (4.12)$$

Now the task of gradient descent or any other optimizer is to find  $\Delta d$  that results in  $\Delta C$  being negative as shown in equation 4.13

$$\Delta d = -\eta \nabla L \quad (4.13)$$

In this case  $\eta$  is a small positive decimal number, commonly referred to as the learning rate, which usually, but not exclusively ranges from 0.1 to 0.001. However it can be larger or much smaller in some cases. Having found a way to ensure that  $\Delta L$  always decreases according to equation 4.13 it is utilized to repeatedly update the gradient  $\nabla L$ . To make the gradient descent algorithm efficient the learning rate  $\eta$  must be chosen correctly. If  $\eta$  is too large, the gradient  $\Delta L$  might end up being larger than zero, leading to an increase of the loss, and if the step size is too small convergence will either take too long or not take place at all Bergstra et al., 2011. In practical machine learning approaches different learning rates are tested. There are also algorithmic approaches. While equation 4.10 only accounts for two inputs features, it can be generalized to compute  $n$  inputs shown in equation 4.14.

$$\nabla L = \left( \frac{\partial L}{\partial w_1}, \dots, \frac{\partial L}{\partial w_n} \right)^T \quad (4.14)$$

Equation 4.15 shows the gradient descent how it is used to repetitively update the weights and biases to optimize the loss function  $L(c, w)$  with  $w$  and  $b$  as the weight and bias matrices and the learning rate  $\eta$ . In machine learning each



iterational update of the network is often called epoch or training epoch.

$$w = w_i - \eta \frac{\partial}{\partial w} L(w) \quad (4.15a)$$

$$b = b_i - \eta \frac{\partial}{\partial b} L(b) \quad (4.15b)$$

$$(4.15c)$$

Substituting the partial differentials with  $\nabla L$  equation 4.15 a simplifies to:

$$w = w_i - \eta \nabla L \quad (4.16)$$

## Optimizers

The previous section introduced the concept of gradient descent, an algorithm to minimize the loss function of the weights and biases of a neural network. All other optimizers introduced here, are either variations or extensions of the basic gradient descent algorithm (GD) shown in 4.15. One disadvantage of gradient descent is that if the data sets grow larger, the demand in memory for computation increases exponentially. Taking into consideration machine learning is a popular method in big data applications this is a serious drawback. Methods to overcome that are stochastic gradient descent and mini-batch gradient descent. The idea behind the latter is to randomly divide the entity of the training data in sub-samples called mini-batches *Bottou and Bousquet, 2008*. The network is then trained iteratively over the mini batches. The batch size influences the accuracy and the training speed and is another hyperparameter which has to be tuned. If the batch size is 1 mini batch GD is also referred to as stochastic gradient descent (SGD). During the optimization process optimizers can find local minima in the cost function without being able to overcome them to find the desired global minimum. An algorithm extending GD to accelerate the search of

the global minimum is momentum. Which allows the GD to speed up when the loss is decreasing and to slow down when going in the wrong direction - increasing the loss function  $L(w, b)$ . This is achieved by accounting for the gradient of the previous step in the calculation of the current step. This concept was introduced by *Polyak, 1964* and re-popularized alongside backpropagation learning by *Rumelhart, Hinton, Williams, et al., 1988*.

$$w = w_i - \eta \nabla L + \alpha \Delta w \quad (4.17)$$

Equation 4.17 shows how the momentum is mathematically represented in GD to update the weights  $w$  or likewise the biases the delta of the weights multiplied by the coefficient  $\alpha$  - the momentum, which usually ranges from 0.1 to 0.9 and is another parameter to tuned for successful training. If the momentum is too small the GD will not be able to overcome local minima and if  $\alpha$  is too large the loss functions tends to oscillate without finding an optimum *LeCun, Bengio, and Hinton, 2015*. For both of the momentum and the learning rate it is impractical to remain on the same level during all training epochs. Because after each epoch the loss function is either closer or further away from its global minima and depending on the distance to that minimum it is desirable to have larger or smaller learning rates and momenta. This can be achieved with naive approaches for example using a step function to gradually decrease those values after each iteration, or to utilize algorithmic approaches *Michie, Spiegelhalter, Taylor, et al., 1994*. There is a large variety of optimizers trying to find optimal values for  $\alpha$  and  $\eta$  and till today this field is under active research *Goodfellow, Bengio, and Courville, 2016*. Popular among those are: RMSprop *Hinton, Srivastava, and Swersky, 2012*; Nesterov momentum *Dozat, 2016*; Adadelata *Zeiler, 2012*; Adagrad *Ruder, 2016* and Adam *Kingma and Ba, 2014*. With Adam being the most widely used optimizer today. Nesterov momentum is slight change to the normal momentum capable

of having huge impacts in practical applications, because it helps avoiding oscillations around the minimum by using intermediate information to adapt the momentum.

RMSProp - root mean square propagation - is a method aiming to adapt the learning rate algorithmically, by choosing  $\eta$  for each parameter. And lastly the widespread Adam optimizer combines both of the features of momentum and RMSProp and adapts the learning rate as well as the momentum iteratively *Kingma and Ba, 2014*.

### Backpropagation

maybe i will leave out backpropagation ☺

Backpropagation *Rumelhart, Hinton, Williams, et al., 1988*

### Regularization parameters

When applying the combined aforementioned algorithms and optimizers to find global minima of a loss function of a neural network the problem of overfitting arises, because optimizers like Adam work “too” well. This issue is due to the fact that neural networks have 100s of thousand of free parameters to be trained, deep neural networks have billions and trillions of parameters. If training of the neural net continues for enough epochs eventually the loss function will approach a minimum and as  $L(w, b) \rightarrow 0$  the initially drawn conclusion could mislead to assuming that training was quite successful, but when trying to apply the network trained on the training data set (TRN) to a testing data set (TST) the loss and accuracy of the prediction of TST and TST are very large or accordingly small. This phenomenon is known as overfitting and a lot of fine tuning of hyperparameters is devoted to minimizing this effect *Tetko, Livingstone, and Luik, 1995*. Figure 4.4 visualizes the effects of overfitting during training *Goodfellow, Bengio, and Courville, 2016*.

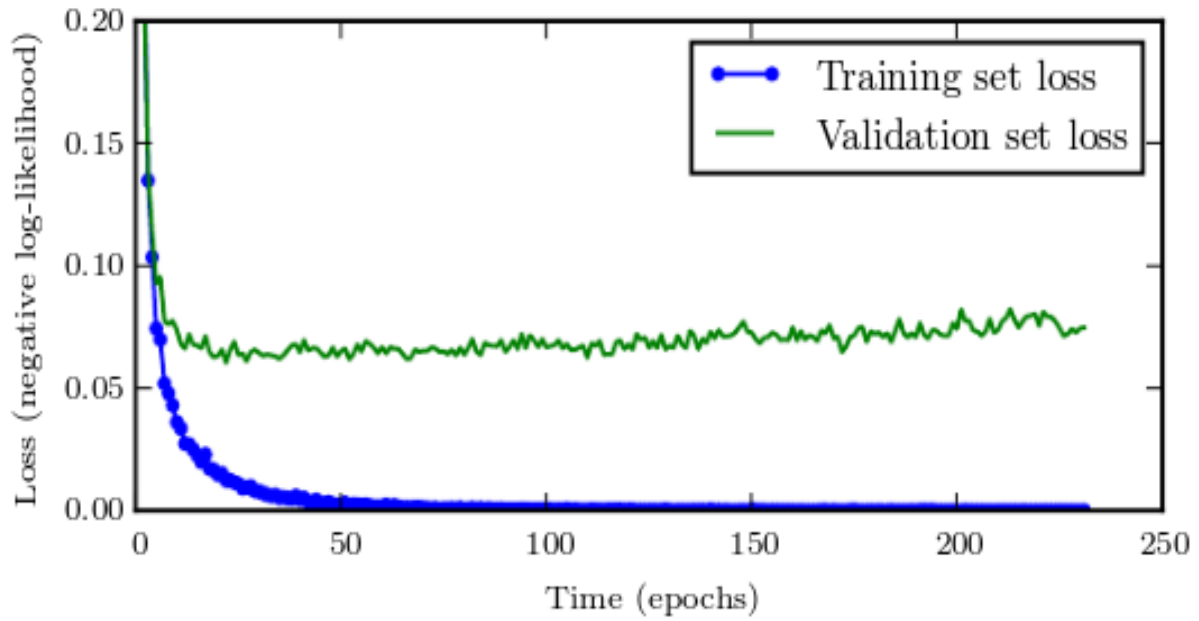


FIGURE 4.4: Learning curves showing how a loss function changes during training in the loss and validation data set. While the training loss approaches 0 the validation loss starts increasing after hitting a minimum. This effect is due to overfitting on the training data set. Figure from Goodfellow, Bengio, and Courville, 2016.

### Cross-validation

A method that is used in basically every training of neural network is splitting up the data sets in multiple sub-sets. More specifically a training set (TRN) and a testing set (TST). The training set is used to minimize the loss functions and its success is evaluated on the TRN set, by comparing the predicted values  $\hat{y}$  with the real values in TST  $y$ . For all neural nets in this study person's correlation coefficient was chosen as performance metric, as in equation 4.18 Soper et al., 1917.

$$\rho(y, \hat{y}) = \frac{\text{cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} \quad (4.18)$$

There are other popular performance metrics, especially for classification problems, like AUC (area under the curve) and ROC (receiver operating characteristics), which basically evaluate by weighing sensitivity and specificity. In cross-validation compared to single validation the initial data set is split into TRN and TST multiple times e.g. if the ratio is 80:20 5 times, and each TRN-TST pair is evaluated individually. Sometimes it becomes necessary to use a third subset - the validation data. Because hyperparameter tuning is performed with the TRN and TST sets, a third portion of the data needs to be assessed to check whether the neural network is able to generalize on global data.

### **L1 and L2 loss**

L1 and L2

**Dropout**

## **4.1.2 On the nature of quantitative traits**

According to the omnigenic model which is an extension of the polygenic model proposed by *Boyle, Li, and Pritchard, 2017* and thoroughly reviewed in *Timpson et al., 2018* all traits or phenotypic values are influenced by a great number or all genes in the genome. Therefore resulting in traits following certain gradual statistical distributions instead of being binned in classes or even binary. Intuitively this might be contradicting with the foundation of modern Genetics - Mendel's three laws. That were derived from observations with where mainly influenced by one locus. But staying with one of Mendel's examples the round or wrinkled surfaces of peas *Pisum sativum*, an assessment of a couple of thousands peas, would most likely inevitably lead to the conclusion that from the "roundest" to the "wrinkliest" pea any gradual step between those is possible and observable. Mendel's third law of independent segregation also only holds true under certain assumptions. The most simplest one being that the traits under investigation

have to be located on different linkage groups. Otherwise for the 7 traits used in Mendel's initial studies would not have segregated independently. The odds of 7 randomly selected traits being on 7 different linkage groups are rather small, especially taking into account, that the genome of the *P. sativum* consists of only 7 chromosomes itself Kalo et al., 2004. Mendel probably new about traits not following its own laws, as well as being aware of the quantitative nature of traits such as the constitution of surfaces of peas or the color of petals. But being the pioneer of a then rather unexplored field of science, some of which big questions we fail to satisfactory answer today, he did not have the resources or the knowledge to explain behavior's not "mendeling", that were only able to be deciphered in later decades and centuries based on his ground-breaking work.

Initially thought to be contradicting to Mendel's ideas Darwin proposed the concept's of evolution due to natural selection which also introduce the idea of traits following a gradual distribution Darwin, 1859. This contrast led to a long lasting debate in the scientific community in the early 1900s, between the Mendelians and the biometricians who believed in the quantitative nature of continuous traits. This conflict has eventually been solved by Fisher's fundamental work published in 1918 Fisher, 1919. His theories combined the then in all fields of science popular research of distributions with genomics. He he mathematically proved that traits influenced by many genes, with randomly-sampled alleles follow a continuous normal distribution in a population. While this combined the ideas of Mendel and the biometricians it opened an other long debated question of effect size and the overall architecture of complex traits. While in the theory of monogenic traits the effect size of the single gene on the trait is 1 or 100 % with an increasing number of genes influencing a complex traits the *per se* contribution of single gene has to decrease with an increasing number of loci determining the value a given trait. In the 1990s it has been thought, that complex traits are predominantly controlled from few genes with a large to medium effect size, while others had a minimal influence Zhang et al., 2018.

With the upcoming popularity of GWAS as the favored method to decipher genetic architectures of traits, or having pioneered in human genetics it became clear that the majority of the effect sizes are tiny  $< 1\%$  while there are very few loci which have a moderate effect on the phenotypic variance of a population with around  $10\%$  or less *Korte and Farlow, 2013, Stringer et al., 2011*. This nature of quantitative traits present great challenges to animal *Goddard and Hayes, 2009* and plant breeding *Würschum, 2012*, in further improving crop or livestock performances, as well complicating the decomposition of genomic causes for diseases like schizophrenia or autism in human medicine *De Rubeis et al., 2014, Purcell et al., 2014*.

While the complex nature of the architecture of quantitative traits provide enough challenges as is, all traits will also be influenced by the environment from which an individual originates. Therefore the distribution of trait values in a given population can be expressed as the addition of the variances of its genetic and the environmental effects 4.19.

$$\sigma_P = \sigma_G + \sigma_E \quad (4.19)$$

The genomic and the environmental effects not only influence the phenotypic variance directly, but the environment also has an influence on gene expression methylation of DNA bases etc. and therefore the equation 4.19 needs to be extended by the variance of the gene-environment interactions  $\sigma_{G \times E}$  4.20, *Lynch, Walsh, et al., 1998, Walsh and Lynch, 2018b*.

$$\sigma_P = \sigma_G + \sigma_E + \sigma_{G \times E} \quad (4.20)$$

Equation 4.20 shows the decomposition of the phenotypic variance, to thoroughly understand complex genetic architectures of traits the genetic variance needs to be decomposed further in its additive, dominance and epistatic components 4.21

$$\sigma_G = \sigma_A + \sigma_D + \sigma_I \quad (4.21)$$

The additive effects are caused by single, for this model mostly homozygous, loci while the variance caused by dominance effects, is caused by heterozygous loci and their resulting interactions being full-, over-, co- or underdominant. And lastly the interaction effects that are a result of two or more genes only having an impact if the involved genes co-occur in a certain state. The resulting variance is commonly known as gene-gene interactions and/or epistasis *Falconer and Mackay, 1996*.

Since possible interactions in a genome can happen between additive or dominant or a combination of those loci. The variance due to interaction effects  $\sigma_I$  can be further dissembled in the variance resulting from additive-additive  $\sigma_{AA}$  dominant-dominant  $\sigma_{DD}$  and additive-dominant  $\sigma_{AD}$  terms as represented in equation 4.22.

$$\sigma_I = \sigma_{AxA} + \sigma_{DxD} + \sigma_{AxD} \quad (4.22)$$

Knowledge of the variance components involved in the expression of a trait in population, lead up to the estimation of the total influence of all genetic variances and the environmental variance on the phenotypic distribution. This concept is called heritability. The heritability of a trait  $H^2$  accounts for the proportion of the phenotypic variance controlled by the total genetic variance as shown in equation 4.23. This is also referred to as broad sense heritability, because all genetic effects including additive, dominance and epistatic effects are included *Brooker, 1999*.

$$H^2 = \frac{\sigma_A + \sigma_D + \sigma_I}{\sigma_P} \quad (4.23)$$

The concept of narrow-sense heritability 4.24 is similar to the broad-sense heritability, but only the additive genetic effects are included in the genetic part of



the equation. This differentiation is important for natural and artificial selection and thus is commonly used in evolutionary genomics and breeding. Because in diploid species each parent only passes down one allele of a given locus, dominance effects or interaction effects are not commonly inherited from one parent. Therefore it is mainly the additive genetic effects of a parent that influence its offspring. While the dominance and epistatic variances are controlled by the combination of the parents *Falconer and Mackay, 1996, Walsh and Lynch, 2018b*.

$$h^2 = \frac{\sigma_A}{\sigma_P} \quad (4.24)$$

### 4.1.3 Artificial selection in plant and animal breeding in the genomics era

#### Introduction to genomic selection

Genomic prediction has been applied to almost all relevant crop and model species. Including: *A.thaliana* Hu et al., 2015; Shen et al., 2013. Alfalfa (*Medicago sativa*) Li and Brummer, 2012; Annicchiarico et al., 2015; Li et al., 2015; Biazzi et al., 2017; Hawkins and Yu, 2018. Barley Neyhart, Lorenz, and Smith, 2019; Oakey et al., 2016; Zhong et al., 2009. Cassava (*Manihot esculenta*) Elias et al., 2018a; Elias et al., 2018b. Cauliflower (*Brassica olearacea* spp) Thorwarth, Yousef, and Schmid, 2018. Cotton (*Gossypium* spp.) Gapare et al., 2018. Maize (*Zea mays*) Moeiniazade et al., 2019; Allier et al., 2019; Brauner et al., 2018; Schrag et al., 2018; Schopp et al., 2017b; Sousa et al., 2017; Schopp et al., 2017a; Kadam et al., 2016; Bustos-Korts et al., 2016a; Montesinos-López et al., 2015; Owens et al., 2014; Lehermeier et al., 2014; Technow et al., 2014; Peiffer et al., 2014; Riedelsheimer et al., 2013; Guo et al., 2013; Technow, Bürger, and Melchinger, 2013; Windhausen et al., 2012; Rincent et al., 2012. Potato (*Solanum tuberosum*); Enciso-Rodriguez et al., 2018; Endelman et al., 2018. Rapeseed (*Brassica napus*) Würschum, Abel, and Zhao, 2014; Jan et al., 2016; Luo et al., 2017; Werner et al.,

2018; Snowdon and Iniguez Luy, 2012; Qian, Qian, and Snowdon, 2014. Rice (*Oryza sativa*) Momen et al., 2019; Hassen et al., 2018; Xu, 2013; Grenier et al., 2015. Rye (*Secale cereale*) Auinger et al., 2016; Bernal-Vasquez et al., 2014; Wang et al., 2014; Bernal-Vasquez et al., 2017; Marulanda et al., 2016. Sugar beet (*Beta vulgaris*), Würschum et al., 2013; Biscarini et al., 2014. Sugar cane (*Saccharum officinarum*) Gouy et al., 2013. Soybean (*Glycine max*) Stewart-Brown et al., 2019; Jarquin, Specht, and Lorenz, 2016; Xavier, Muir, and Rainey, 2016. Switchgrass (*Panicum virgatum*) Poudel et al., 2019; Ramstein and Casler, 2019; Ramstein et al., 2016. Wheat (*Triticum aestivum*) Cuevas et al., 2019a; Howard et al., 2019; Krause et al., 2019; Rincen et al., 2018; Norman et al., 2018; Belamkar et al., 2018; Ovenden et al., 2018; Sukumaran et al., 2016; Bustos-Korts et al., 2016b; Gianola et al., 2016; Crossa et al., 2016; Thavamanikumar, Dolferus, and Thumma, 2015; Lopez-Cruz et al., 2015. As well as various tree species Almeida Filho et al., 2019; Rincen et al., 2018; Kainer et al., 2018; Ratcliffe et al., 2017; El-Dien et al., 2016; Kumar et al., 2015; Jaramillo-Correa et al., 2014; Zapata-Valenzuela et al., 2013; Holliday, Wang, and Aitken, 2012; Resende et al., 2012.

Even though GS finds broad application in plant breeding it has been originally developed for the use in animal breeding Hayes and Goddard, 2010; Goddard, Hayes, and Meuwissen, 2011. The gold standard is a method known as genomic BLUP VanRaden, 2008 which utilizes a relationship matrix based on the co-occurrence of genetic markers. This method is derived from the pre-genomic era in animal breeding where the relationship matrix was constructed after pedigrees according to the best linear unbiased predictors based linear mixed models developed by Henderson, 1975. GBLUP accounts only for additive-genetic effects VanRaden, 2008. There are other methods that are able to account for more complex genomic effects that are non-additive. Popular among those are reproducing Kernel Hilbert Spaces (RKHS) Gianola and Kaam, 2008. Alternatively to linear mixed models a variety of different Bayesian methods became popular, basically differing in the degree of shrinkage of the assumed marker distribution Hayes, Goddard, et al., 2001; Gianola et al., 2009; Habier et al., 2011; Gianola, 2013; Crossa et al., 2017;

Azodi et al., 2019.

### Genomic selection in recurrent selection and the breeders equation

While the quantitative genetic methods breeders utilize are complex their goals can be defined in one sentence: To genetically improve plant germplasms for agriculture. The breeding process started at the same time as farming was first practiced 10.000 BC in the region between the Euphrat and Tigris rivers known as the fertile crescent. This changed the phenotypic appearance of the early crops dramatically to the point that they share little external traits with their wild ancestors. Those changes have also been deeply carved into the genomes, that underwent serious alterations, including hybridization, duplications etc. Leading to most crop plants not having any wild ancestors with whom they could naturally mate. For example wheat (*Triticum aestivum*), one of the three most important sources of food on a global scale underwent multiple hybridization steps Ozkan, Levy, and Feldman, 2001. Wheat is a hybrid from either the diploid emmer (*T. diccoides*) or durum wheat (*T. durum*) and *Aegilops tauschii*, while emmer and durum are hybrids derived from wild emmer which is a hybrid of wild grass of the genus of *Aegilops* and *T. urata* Friebe et al., 2000; Feldman and Levy, 2012. While being ignorant of modern genetics early “plant breeders” must have had an intuitive, yet naive, understanding of the general concept of heritability in a way that they must have established that offsprings share properties of their parent generation, which lead to regrowing individuals with desired traits generations after generation. This lead to many changes including that artificial selected plants are commonly largely inbred. This process could be considered an early form of recurrent truncation selection. Truncation selection on a normal distributed phenotype is shown in figure 4.5.

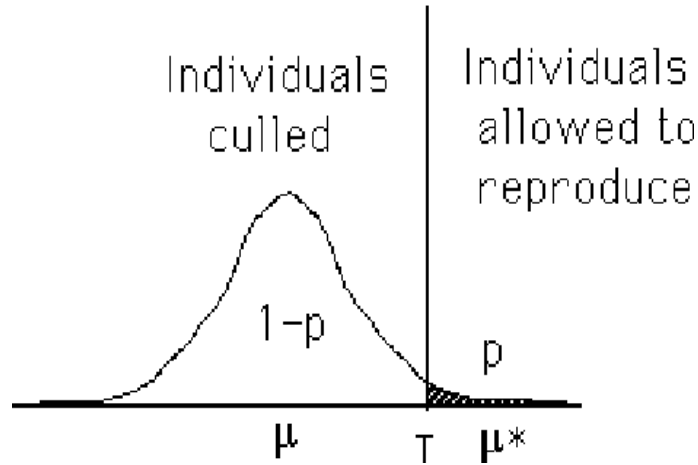


FIGURE 4.5: Truncation selection from a normal distributed phenotype with a threshold value of  $T$ ,  $\mu$  as the mean of the total population and  $\mu^*$  as the mean of the selected phenotypes. Graphic from Walsh and Lynch, 2018a

Like the early breeders modern breeders of to determine the selection threshold  $T$  to divide the total population with the mean  $\mu$  into two groups: the individuals culled and the ones allowed to reproduce with the mean  $\mu^*$ . The difference between those two is the selection differential  $S$ :

$$S = \mu^* - \mu \quad (4.25)$$

which in the case of normal distributed data as depicted in figure ?? can be expressed as:

$$S = \varphi\left(\frac{T - \mu}{\sigma}\right) \frac{\sigma}{p} \quad (4.26)$$

From which we can obtain the selection intensity  $i$ , which makes  $i$  solely a function of  $p$

$$i = \frac{S}{\sigma} = \frac{\varphi(z_{|1-p|})}{p} \quad (4.27)$$

With recurrent truncation selection over many generations the population

mean of the trait  $\mu$  will change (hopefully in the desired direction), if the heritability (in this case the narrow sense heritability)  $h^2 > 0$ . It is impossible to breed for traits that do not contain any genetic components in its architecture *Walsh and Lynch, 2018b*. Next to  $i$  the intensity and  $h^2$  the additive portion of the heritability the accuracy of the selection  $r_{uA}$  is important for the success of a breeding program. Those three terms can be applied to compute the gain of selection  $R$  over one generation (equation 4.28). Due to its importance in the evaluation of breeding schemes it is known as the breeder's equation *Mousseau and Roff, 1987; Falconer and Mackay, 1996; Kingsolver et al., 2001*.

$$R = ir_{uA}\sigma_A \quad (4.28)$$

The accuracy  $r_{uA}$  of equation 4.28 in cases only phenotypic selection is conducted the heritability and in cases where the selection process is aided by genomic prediction it is the prediction accuracy. According to breeder's equation there are three parameters which can be influenced through genomic prediction. (i) The prediction accuracy, which is usually smaller than the heritability, varies for different prediction equations and an increase in the accuracy will lead to an proportional increase in  $R$  the gain per generation cycle. For this reasons since 2001 in quantitative genetics one very active field of research was and still is to find new algorithms that are superior to others as presented in the next chapter (4.1.3. As later evaluated on more than 150 phenotypes in chapter ??  $h^2$  is always larger than  $r_{uA}$ . Which if it was the only variable factor in equation 4.28 make genomic selection inferior to phenotypic selection. Which from a certain point of view it is. Phenotypic trials are better approximation for phenotypic appearance as GEBVs. However as the cost of genotyping have decreased dramatically in the last 20 years. Phenotyping in the field remains tedious, laborious and mostly

vastly expensive. Taking into account that field trials have to be repeated in several years and locations to produce reliable accounts it becomes clear that genotyping 10th of thousands of accessions is much cheaper than conducting field trials with 1000 of them.

(ii) The selection intensity can much stricter if the total population that is selected from is larger and in genomic prediction settings they are, because it allows selection from two pools. The first the pool of plants with known phenotyped and known genotype and from those where just genomic data is available. When selecting from a pool of 1000 with  $p = 0.05$  with the goal to keep 50 plants in the next breeding cycle, the same goal can be reached when genomically selecting from a pool of 10000 with an intensity of  $p = 0.05$ .

(iii) The decreased time per generation is probably the largest advantage of genomic selection when applied to breeding. While in field trials it is only possible to have one generation per year. Genomic selection does not require the plants to be grown in the field. For selection it is only necessary to grow enough so that DNA can be extracted from the tissue and evaluated. After selection only the ones above the threshold are grown until they bear seeds (or any other reproductive organ) and be used for the next selection cycle, allowing up to ten generations per year. This development has led to the rise to a new branch of breeding: speed breeding Ghosh et al., 2018; Watson et al., 2018. In practical, company-level genomic prediction has largely contributed to an increase by a factor of 2 of the gain in selection in recent years (personal communication with breeding company employees).

The last term in equation 4.28 the additive genetic variance  $\sigma_A$  is not directly yet heavily influenced by the described breeding scheme. Artificial selection has similar effects on the genetic variance as bottlenecks in natural selection have: it decreases, thus making it harder to increase  $R$  in later selection cycles Walsh and Lynch, 2018b.

### Genomic BLUP and Bayesian methods

All methods share a common statistical obstacle which is commonly referred to as the  $n \gg p$  problematic, which arises because the number  $n$  marker is usually significantly larger than the number of observations  $p$ . In practical applications it is not uncommon to have more than 100k markers while the number of phenotypes is no larger than 100. This does not allow to obtain genomic estimated breeding values (GEBV) by single marker regression as done by GWAS, which have highly inflated over all SNP-effects *Korte and Farlow, 2013*. One possible is to include effect sizes as random effects and make prior assumptions about their distribution. The difference in prior distribution is the main distinction between the methods of the Bayesian alphabet *Gianola, 2013*.

#### Genomic BLUP

In the early years of research on genomic prediction algorithms they were not solely benchmarked against each other, but had to compete with the previously popular pedigree methods. Quickly in the course of the first decade of this millennium the superiority of the genomic method became elucidated in livestock and plant breeding *Habier, Fernando, and Dekkers, 2007; VanRaden, 2008; VanRaden et al., 2008; Harris, Johnson, Spelman, et al., 2009*. While the genomic methods are superior to non-genomic methods, there is no clear evidence that either of the genomic methods are superior to each other and there is lack of empirical evidence that the Bayesian methods generally outperform GBLUP *Moser et al., 2009; Bernardo, 2010; Azodi et al., 2019*. Alike in the pedigree BLUP in genomic BLUP the co-variance between related individuals is used for the prediction. In the later case it is constructed from marker information. In the GWAS terminology the relationship matrix is referred to as  $K$  for kinship in GWAS literature, while in GS circumstances it is also called GRM (genomic relationship matrix) or

simply abbreviated as  $G$ . This study will remain consistent with the circumstantial literature and therefore purposely inconsistent within itself. In the chapter addressing GWAS it will be called  $K$  for kinship matrix and in the following chapter elucidating GBLUP it will be referred to as  $G$ . The general genomic prediction model 4.29 is derived from the mixed model *Henderson, 1975; VanRaden, 2008* and implemented as:

$$Y = X\beta + Zu + \varepsilon \quad (4.29)$$

where  $Y$  is an  $n \times 1$  vector of phenotypic observations,  $X$  the matrix of the fixed effects and  $\beta$  the vector of the fixed effects.  $Z$  the incidence matrix for the combined marker effects and  $u$  a  $(n \times 1)$  vector of the additive genetic effect the vector of the residuals  $\varepsilon$ . To construct the model let's assume a matrix of size  $(n \times m)$  with  $n$  individuals and  $m$  loci  $M$  containing marker information for 3 individuals on 4 loci, thus being of size  $3 \times 4$ . The 4 markers of matrix 4.30 can take values of  $-1, 0$  and  $1$  translating into minor allele, heterozygous locus and major allele. The following example calculation has been adapted from *Isik, 2013*.

$$M = \begin{pmatrix} -1 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 \end{pmatrix} \quad (4.30)$$

The  $M$  matrix contains all the information that is necessary for the computation of the  $K$  matrix and other viable genetic parameters. The  $MM'$  matrix of size  $n \times n$  (??) bears additional parameters.

$$MM' = \begin{pmatrix} 3 & 1 & 2 \\ -1 & 1 & 0 \\ 2 & 0 & 3 \end{pmatrix} \quad (4.31)$$

The diagonal shows the number of homozygous loci per individual, while the



other elements of the matrix indicate the number of markers shared by related individuals and is an indicator for the distance of the relationship as defined by identity-by-descent *VanRaden, 2008; Misztal et al., 2013*. While matrix 4.31 calculates the metrics per individual the  $M'M$  matrix (4.32) counts those metrics per marker. Likewise the diagonal contains the number of homozygous individuals per marker.

$$M'M = \begin{pmatrix} 3 & -1 & 0 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix} \quad (4.32)$$

The next step is to obtain a matrix of the allele frequencies at each locus of the size  $(n \times m)$  as matrix  $M$ . For the design of matrix  $P$  (4.33) let the minor allele frequencies  $p_1 \dots p_4$  be  $\{0.3, 0.2, 0.1, 0.15\}$ . The allele frequency of the  $i^{th}$  column of  $P$  is expressed, according the  $i^{th}$  marker of matrix  $M$  as  $P_i = 2(p_i - 0.5)$  resulting in:

$$P = \begin{pmatrix} -0.4 & -0.6 & -0.8 & -0.7 \\ -0.4 & -0.6 & -0.8 & -0.7 \\ -0.4 & -0.6 & -0.8 & -0.7 \end{pmatrix} \quad (4.33)$$

The allele frequencies as in this simulated example should be taken from the entire population and not the subsample used in this calculation *VanRaden, 2008*. The final step to obtain the  $Z$  matrix of equation 4.29 is to subtract the  $P$  matrix from the  $M$  matrix  $Z = M - P$  resulting in:

$$Z = \begin{pmatrix} 1.4 & 0.6 & 1.8 & -0.3 \\ -0.6 & 0.6 & 0.8 & 0.7 \\ 0.4 & 1.6 & 1.8 & -0.3 \end{pmatrix} \quad (4.34)$$

In  $Z$  the mean values of the allele effects are set to 0 and the subtraction of

$P$  emphasizes the effect of rare variants *VanRaden, 2008*. There is a large variety of methods to generate the genomic relationship matrices and here lies the major difference between different genomic BLUP methods. But they all have in common that  $K$  is always of size  $n \times n$ .

(i) The naive is approach to iterate over each individual and count the common markers with every other individual. This approach is not uncommon and suited for inbred or doubled-haploid populations, less so for outcrossed populations with high degrees of heterozygosity because in the sample implementation it does only account for homozygous loci. This method becomes computationally intense when the data sets grow larger as common today (personal observation).

(ii) Probably the most popular method in GS is to obtain  $K$  as proposed by *VanRaden, 2008* designed after Wright's *Wright, 1922* equations for the covariance in structured populations, as described by equation 4.35 with  $Z$  as in 4.34.

$$G = \frac{ZZ'}{2\sum p_i(1 - p_i)} \quad (4.35)$$

(iii) The unified additive relationship  $G_{UAR}$  according to *Yang et al., 2010* and equation 4.36

$$G_{UAR} = A_{jk} = \frac{1}{N} \sum_i A_{ijk} = \begin{cases} \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}, j \neq k \\ 1 + \frac{1}{N} \sum_i \frac{x_{ij}^2(1 + 2p_i)x_{ij} + 2p_i^2}{2p_i(1 - p_i)}, j = k \end{cases} \quad (4.36)$$

where  $p_i$  is the allele frequency at locus  $i$  and  $x_{ij}$  the genotype for the  $j^{th}$  individual at the  $i^{th}$  locus. Another method also proposed by *Yang et al., 2010* is to adjust  $G_{UAR}$  with  $\beta$  as in equation ??

$$G_{UARadj} = \begin{cases} \beta A_{jk}, j \neq k \\ 1 + \beta(A_{jk} - 1), j = k \end{cases} \quad (4.37)$$

(iv) Another approach is to weigh marker by the reciprocals of their expected variance according to the model 4.38 originally designed to investigate population structures in human data *Leutenegger et al., 2003; Amin, Van Duijn, and Aulchenko, 2007.*

$$G = ZDZ', \text{ with} \quad (4.38)$$

$$D_{ii} = \frac{1}{m|2p_i(1-p_i)|}$$

(v) Other methods like the gaussian kernel compute kinship between individuals by the euclidean distance between the respective genotypes *Morota and Gianola, 2014.*

$$K(x_i, x_j) = \exp(-\theta d_{ij}^2) \quad (4.39)$$

$$= \prod_{k=1}^m \exp(-\theta(x_{ik} - x_{jk})^2)$$

with  $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{ik} - x_{jk})^2 + \dots + (x_{im} - x_{ja})^2}$  and  $x_{ik}(i, j = 1, \dots, n, k = 1, \dots, m)$  and  $x_{ik}$  as the  $i^{th}$  individual at SNP  $k$ .

The linear model in equation 4.29  $Y = X\beta + Zu + \varepsilon$  with  $\beta$  as the vector fixed effects and  $u$  as the vector of additive genetic effects. The mixed model can be solved to obtain genomic estimated breeding values as:

$$\begin{pmatrix} X'X & X'Z & 0 \\ Z'X & Z'Z + G^{11} & G^{12} \\ 0 & G^{21} & G^{22} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \\ 0 \end{pmatrix} \quad (4.40)$$

with  $G^{12}$  as the part of  $G^{-1}$  containing individuals with phenotypic data. with  $G^{22}$  as the part of  $G^{-1}$  containing individuals without phenotypic data and just SNP information.

The GEBV of the unknown phenotypes  $\hat{y}_2$  can thus be predicted as:

$$\hat{y}_2 = - \left( G^{22} \right)^{-1} G^{21} \hat{y}_1 \quad (4.41)$$

GBLUP is fairly easy compared to more complex Bayesian methods and be quickly implemented in any programming language capable of solving liner equations. Computational as the the number of phenotypes in the study increases in numbers the timed demand grows exponentially, because the kinship matrix quadruples in size and it becomes more complicated to compute the inverse of  $G$  (personal observations).

### Bayesian methods

Next to the universal GBLUP a set of related algorithms became popular for solving the mixed models involved in genomic selection, known as the Bayesian alphabet *Gianola et al., 2009; Gianola, 2013*. They are all based on Bayes' fundamental theorem in equation 4.42

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)} \quad (4.42)$$

with  $P(\theta)$  as the prior distribution,  $P(y|\theta)$  as the likelihood and  $P(y)$  as the marginal density of  $y$ . The prior distribution in GS assume that  $y$  was drawn from a certain distribution. Infinitesimal models assume that the genetic effects follow a normal distribution *Legarra, Lourenco, and Vitezica, 2018*. The Bayesian frameworks however will assume non-normal distributed marker effects. This can be explained by a two-step hierarchical distribution. Stage one assumes that every marker has *a priori* a different variance *Legarra, Lourenco, and Vitezica, 2018*.

$$p(a_i|\sigma_{ai}^2) = N(0, \sigma_{ai}^1) \quad (4.43)$$

The second stage assumes prior distributions for the variances.

$$p(a_i|variable) = P(\dots) \quad (4.44)$$

with *variable* standing for the large variety of prior distribution. In total there are more than >20 different Bayesian models known to the authors with unknown number of different methods proposed. Their main difference “simply” lies in the a priori assumptions. This change can make some methods mathematically much more advanced than others. And as shown in later chapters none of the methods is completely superior over others in terms of prediction accuracy. Approximation to the solution of the linear equations is usually performed by Gibb’s sampling using Markov Chain Monte Carlo (MCMC) simulations *De Los Campos et al., 2009; Campos and Rodriguez, 2016*. Table 4.1 summarizes commonly applied Bayesian methods for genomic prediction indicating the key differences between them.

TABLE 4.1: Overview of properties of a variety of commonly applied Bayesian methods for genomic prediction. Table altered after *Kärkkäinen and Sillanpää, 2012*

Name	Reference	Prior	Indicator	Hierarchy	Hyperprior	Estimation
BayesA	<i>Hayes, Goddard, et al., 2001</i>	Student	No	Yes	No	MCMC
BayesB	<i>Hayes, Goddard, et al., 2001</i>	Student	Yes	Yes	No	MCMC
BayesC	<i>Verbyla et al., 2009</i>	Student	Yes	Yes	No	MCMC
BL	<i>Xu, 2010</i>	Laplace	No	Yes	No	EM
BayesD $\pi$	<i>Habier et al., 2011</i>	Student	Yes	Yes	Yes	MCMC

The name is given by the author. The prior column tells which shrinkage prior is used.

## cross validation

### 4.1.4 Genomic selection using artificial neural networks

Genomic selection (GS) has been successfully applied in animal *Gianola and Rosa, 2015; Hayes and Goddard, 2010* and plant breeding *Crossa et al., 2010; Desta and*

Ortiz, 2014; Heffner et al., 2010; Crossa et al., 2017 as well as in medical applications, since it was first reported Hayes, Goddard, et al., 2001. Since then the repertoire of methods for predicting phenotypic values has increased rapidly e.g. De Los Campos et al., 2009; Habier et al., 2011; Gianola, 2013; Crossa et al., 2017. The most commonly applied methods include GULP and a set of related algorithms known as the bayesian alphabet Gianola et al., 2009. Genomic prediction in general has repeatedly been shown to outperform pedigree-based methods Crossa et al., 2010; Albrecht et al., 2011 and is nowadays used in many plant and animal breeding schemes. It has also been shown that using whole-genome information is superior to using only feature-selected markers with known QTLs for a given trait Bernardo and Yu, 2007; Heffner, Jannink, and Sorrells, 2011 in some cases. A more recent study Azodi et al., 2019 compared 11 different genomic prediction algorithms with a variety of data sets and found contradicting results, indicating that feature selection can be useful in some cases when the whole genome regression is performed by neural nets. While every new method is a valuable addition to the tool-kits for genomic selection, some fundamental problems remain unsolved, of which the  $n \gg p$  problematic stands out. Usually in genomic selection settings the size of the training population (TRN) with  $n$  phenotypes is substantially smaller than the number of markers ( $p$ ) Fan, Han, and Liu, 2014. Making the number of features immensely large, even when SNP-SNP interactions are not considered. Furthermore each marker is treated as an independent observation neglecting collinearity and linkage disequilibrium (LD). Further difficulties arise through non-additive, epistatic and dominance marker effects. The main problem with epistasis in quantitative genetics is the almost infinite amount of different marker combinations, that cannot be represented within the size of TRN in the thousands, the same problem arises for example in GWA studies Korte and Farlow, 2013. With already large  $p$  the number of possible additive SNP-SNP interactions potentiates to  $p^{(p-1)}$ . Methods that attempt to overcome those issues are EG-BLUP, using an enhanced epistatic kinship matrix and reproducing kernel

Hilbert space regression (RKHS) *Jiang and Reif, 2015; Martini et al., 2017.*

In the past 10 years, due to increasing availability of high performance computational hardware with decreasing costs and parallel development of free easy-to-use software, most prominent being googles library TensorFlow *Abadi et al., 2016* and Keras *Chollet et al., 2015*, machine learning (ML) has experienced a renaissance. ML is a set of methods and algorithms used widely for regression and classification problems. popular among those are e.g. support vector machines, multi-layer perceptrons (MLP) and convolutional neural networks. The machine learning mimics the architecture of neural networks and are therefore commonly referred to as artificial neural networks (ANN). Those algorithms have widely been applied in many biological fields *Min, Lee, and Yoon, 2017; Lan et al., 2018; Mamoshina et al., 2016; Angermueller et al., 2016; Webb, 2018; Rampasek and Goldenberg, 2016.*

A variety of studies assessed the usability of ML in genomic prediction *González-Camacho et al., 2018; González-Camacho et al., 2016; Ogutu, Piepho, and Schulz-Streeck, 2011; Montesinos-López et al., 2019a; Grinberg, Orhobor, and King, 2018 ; Cuevas et al., 2019b; Montesinos-López et al., 2019b; Ma et al., 2017; Qiu et al., 2016; González-Camacho et al., 2012 Li et al., 2018.* Through all those studies the common denominator is that there is no such thing as a gold standard for genomic prediction. No single algorithm was able to outperform all the others tested in a single of those studies, let alone in all. While the generally aptitude of ML for genomic selection has been repeatedly shown, how no evidence exists that neural networks can outperform or in many cases perform on that same level as mixed-model approaches as GBLUP *Hayes, Goddard, et al., 2001.* While in other fields like image classification neural networks have up to 100s of hidden layers *He et al., 2016* the commonly used fully-connected networks in genomic prediction of 1 - 3 hidden layers. With 1 layer networks often being the most successful among those. Contradicting to the idea behind machine learning in genomic selection 1 hidden layer networks will be inapt to capture interactions between loci and thus

only account for additive effects. As shown in Azodi et al., 2019 convolutional networks perform worse than fully-connected networks in genomic selection, which again is contradicting to other fields where convolutional layers are applied successfully, e.g natural language processing Dos Santos and Gatti, 2014 or medical image analysis Litjens et al., 2017. Instead of using convolutional layers and fully-connected layers only, as show in Pook et al 2019, we also propose to use locally-connected layer in combination with fully-connected layers. While CL and LCL are closely related they have a significant difference. While in CL weights are shared between neurons in LCLs each neuron as its own weight. This leads to a reduced number of parameters to be trained in the following FCLs, and should therefore theoretically lead to a decrease in overfitting a common problem in machine learning. To evaluate the results of Pook et al. 2019 accomplished with simulated data we used the data sets generated in the scope of the 1001 genome project of *Arabidopsis thaliana* Alonso-Blanco et al., 2016

## 4.2 Proof of concept for ANN-based genomic selection

Having established the quantitative architecture of traits in section 4.1.2 an the basics of machine learning and neural nets in section 4.1.1, that knowledge can be used to provide a proof of concept that neural networks are a candidate for GP. Table 4.2 provides also the possible genotypes that can be derived by two bi-allelic markers  $G_1 \dots G_4$  on a fictional haploid organism. In this simulation the effect sizes for each marker  $\beta_1$  and  $\beta_2$  are constant with a value of 1.



TABLE 4.2: Simple simulated phenotypes and genotypes for genomic prediction with genotypes  $G_1 \dots G_4$ ,  $M_1$  and  $M_2$  and phenotypes based on additive effects or *and*, *or*, *xor* logic gates.

	$M_1$	$M_2$	$Y_{ADD}$	$Y_{AND}$	$Y_{OR}$	$Y_{XOR}$
$G_1$	0	0	0	0	0	0
$G_2$	0	1	1	0	1	1
$G_3$	1	0	1	0	1	1
$G_4$	1	1	2	1	1	0

The four phenotypes  $Y_{ADD}$ ,  $Y_{AND}$ ,  $Y_{OR}$  and  $Y_{XOR}$ , which were derived from their respective marker effects.  $Y_{ADD}$  is a phenotype with purely additive effects. So in the nomenclature introduced in chapter 4.1.2  $\sigma_A = \sigma_G$  and  $\sigma_I = 0$ . Since the hypothetical organism is haploid there are dominance effects to be accounted for  $\sigma_D = 0$ . Since all the genetic effects are caused by additive effects and there are now environmental effects  $\sigma_E$ , the narrow sense heritability  $h^2$  - equation 4.24 - and the broad sense heritability  $H^2$  - equation 4.23 - are equally 1. The other three phenotypes are base on epistatic effects  $\sigma_I$  generated by passing the markers  $M_1$  and  $M_2$  through their respective logic gates. This theoretically in results in  $h^2 = 0$  and  $H^2 = 1$ , because there are no additive effects. For  $y_{AND}$  however  $h \approx 0.5$ , because there is a correlation between  $Y_{ADD}$  and  $Y_{AND}$ . In practical applications this allows methods like GBLUP, designed to account for additive genetic effects to capture some of the epistatic effects of  $\sigma_I$  *Vieira et al., 2017*.

According to chapter 4.1.1 a single perceptron would fail to solve *xor* gates. While a network with multiple nodes and layers should be able to overcome that deficit. A relatively simple neural network with two fully-connected hidden layers with 10 and 5 nodes, was trained for the prediction of each phenotypes. To keep the simulation as possible, no regularization parameters, dropout etc. was included. The activation function was ReLU (4.6) with an Adam optimizer. The results of the prediction are shown in table 4.3.

TABLE 4.3: Results of genomic prediction from phenotypes and genotypes in table 4.2

	$M_1$	$M_2$	$\hat{Y}_{ADD}$	$\hat{Y}_{AND}$	$\hat{Y}_{OR}$	$\hat{Y}_{XOR}$
$G_1$	0	0	0.01	0.00	0.00	0.01
$G_2$	0	1	0.99	0.01	0.99	0.98
$G_3$	1	0	0.99	0.00	0.99	1.01
$G_4$	1	1	1.99	0.98	1.01	0.02

Not surprisingly, the simple network is able to solve all four phenotypes and predicting the phenotypes accurately. The task was rather easy because the training data set and the testing data set were the same, but it served the purpose of showing that neural networks are generally apt to solve different marker interactions. *In natura* those interactions and the overall genetic architecture is much more complex. Effect sizes are not constant and epistasis may be caused by interactions by more than just two markers, and with an increasing number of markers  $n$  the number of possible two-way interactions increases even more so to  $2^{n-1}$ . Smaller interaction effects could be obscured under larger additive effects, gene-environment might have a significant influence leading to a model that does not converge.

### 4.3 Material

Two different data sets were used for the genomic prediction trials. A set of maize doubled-haploid (DH) populations, derived from MAZE landraces. And *A. thaliana* data sets, with genomic data procured along the 1001 genomic project Alonso-Blanco et al., 2016 and various phenotypic trials.

### 4.3.1 DH populations derived from maize landraces

The DH populations were produced, propagated and phenotyped in the scope of the MAZE project phase I, funded by the Federal Ministry of Education and Research (BMBF) (Funding ID: 031B0195, project “MAZE”) as well as the KWS SAAT SE, by various project partners at the Technical University of Munich, University of Hohenheim and the KWS. A thorough description of the germplasm selection and phenotyping was recently published by *Hölker et al.*, 2019.

Modern maize cultivars are almost exclusively high-performing hybrids from two inbred lines from different heterotic pools. Commonly hybrids are derived from the cross European Flint x American Dent *Santos Dias et al.*, 2004; *Brauner et al.*, 2019. Before hybrid breeding became the predominant method in maize breeding in the 1960s landraces were grown by farmers. Landraces are dynamic, open-pollinated, locally highly-adapted populations. That did not derive from modern breeding but from locally confined selection and adaption by farmers to often very specific needs *Arteaga et al.*, 2016. The hybrids grown today are derived from just a few landraces as founder lines, while the majority of landraces has been nearly forgotten. This and high intensity selection over many generations has led to a loss of genetic diversity  $\sigma_G$  in modern maize cultivars. Therefore the landrace germplasm presents an important, essential stock of genetic variability for continuous success in maize breeding. The utilization of those germplasms would be impossible without the invaluable work of institutions as the IPK Gatersleben whose goal as genebanks is to maintain and store genetic material for long time periods. Of those landraces three, representing large phenotypic heterogeneity were chosen: (i) Kemater Landmais Gelb (KE, Austria), Petkuser Ferdinand Rot (PE, Germany) and Lalin (LL, Spain). They represent 95 % of the molecular variance if a set of 35 landraces analyzed in a preceding project by *Mayer et al.*, 2017.

In total 1015 DH lines (516 KE, 432 PE, 67 LL) were produced with *in vivo* haploid

induction with an inducer line as described in *Roeber, Gordillo, and Geiger, 2005*.

### **Genomic maize data**

(The genomic data was provided by the TUM as described by *Hölker et al., 2019*!) Genotyping was performed with the 600k Affymetrix<sup>®</sup> Axiom<sup>®</sup> Maize array *Unterseer et al., 2014*. The markers were quality filters and missing values were imputed individually for each landrace population using Beagle 5.0 *Browning and Browning, 2007; Browning, Zhou, and Browning, 2018*. After LD pruning and further quality control 29833 markers remained and 471 Kemater DHs and 403 PE DHs. LL was excluded from further analyses due to insufficient amounts of genotypes.

### **Phenotypic maize data**

(The phenotype data was provided by the TUM as described by *Hölker et al., 2019*!)

The traits were evaluated with lattice design in 7 different locations across Europe. Those traits were: Early Vigor (EV) at three different stages (V3, V4, V6); Plant height (PH) at two stages (V4,V6) and the final plant height (PH\_final), as well as male and female flowering time (Days till Tasseling (DtTAS) and Days till Silking (DtSILK)) and root lodging (RL). To account for GxE best linear unbiased estimators were calculated according to Henderson's model *Henderson, 1975* and use for further prediction.

## **4.3.2 A. thaliana**

### **Genomic data**

The genomic data was generated during the course of the 1001 genome project of *A. thaliana* *Alonso-Blanco et al., 2016* producing completed sequenced and assembled genomes from 1035 genomes, along 600k marker data for 1307 accessions

with a overlap between those groups resulting in a total of 2029 genotyped accessions. With more than 10 mio. SNPs and Indels on the 5 chromosomes of *A. thaliana*. Imputation of missing data and upsampling of the 600k subsets was performed with Beagle3 *Browning and Browning, 2007*. For every one of the 164 phenotypes used subsets were sampled and LD pruned and MAF filtered. LD pruning was executed with the R-package SNPRelate *Zheng, 2013* with a relatively strict LD threshold of 0.65 and  $MAF > 10$ . This resulted in data sets with approximately 150.000 markers for each phenotype.

### Phenotypic data

A complete list of the phenotypes used can be found in Appendix B with the according study references. The phenotypic trials ranged from 100 to more than 1000 accessions per data set *Atwell et al., 2010; Li et al., 2010; Strauch et al., 2015; Meijó et al., 2014*.

## 4.4 Methods

The theoretical backgrounds of the methods used for genomic prediction were described in section 4.1.1 for the ANNs and section 4.1.3 for the Bayesian methods and GBLUP. The next sections are devoted to explaining who those methods were adapted and implemented for the prediction of maize and *Arabidopsis* traits.

### CV

#### 4.4.1 ANN

The scripts for ANN based GS were written in python using the lower level API TensorFlow *Abadi et al., 2016* and the higher level API Keras *Chollet et al., 2015*. Both are very versatile, well-documented and are capable of performing a large variety of machine learning applications. For those reasons they are the most

used ML libraries. Another advantage is that they work well on GPUs, which allows ML algorithms to run a reasonable amount of time compared to CPU-based calculations. Prior to training the data was split into TRN and TST. The markers of TRN served as the input layer for the network while the phenotypes were trained upon in the output layer, which in the present cases consisted of only one node, because GS in the cases applied here is a regression problem. Preliminary trials showed Adam is the superior optimizer for GS and hence was the only one further used. Likewise relu was the activation of choice being superior to sigmoid or other non rectifiers. All the weights and the biases of the kernel were initialized with truncated normal distributed values. The loss function used was always MSE.

Having a few hyperparameters fixed, the other ones were optimized via grid search. For each TRN multiple networks were trained to fine tune the input parameters. Those were the number of layers, the nodes per layer, the magnitude of the dropout, the type of dropout used, whether the first layer was locally-connected for fully-connected and the duration of training via the training epochs. This amount to a total of a little shy of 260000 trained networks for the 146 *A. thaliana* data sets alone.

After another set preliminary runs LCL as the first layer seemed to result in higher accuracies then FLC as the first layer and where henceforth exclusively used and applied with a stride length of 7. The stride length determines how many node of the input layer, in this case markers, where combined in the first hidden layer. The type of drop out used (alpha dropout, Gaussian noise or normal dropout) did not show an effect therefore the normal dropout function was used. The network training was iterated over the different number epochs, architecture, drop out values the the cross validation cycles, thus explaining the tremendous amount of total networks trained. Epochs from 5 to 60 in steps 5 and several 1, 2 or 3 Layer architectures following the locally-connected layer.

### 4.4.2 GBLUP

*Campos and Rodriguez, 2016*

## 4.5 Results

### C.2

Phenotype	GBLUP	ANN	Architecture	Epochs
YEL	0.9259	0.891	20, 10	20
FT16	0.8237	0.8215	100	10
2W	0.8156	0.8205	50, 30	35
FT10	0.8249	0.8191	48	50
LD	0.8128	0.8159	150	30
DTF sweden 2009 (1st experiment)	0.8063	0.8141	48	30
DTF sweden 2009 (2nd experiment)	0.8035	0.8091	50, 30	20
DTF sweden 2008 (2nd experiment)	0.7986	0.8057	150	25
4W	0.795	0.8052	50, 35, 15	30
FT22	0.8009	0.8043	150	15
DTF spain 2008 (2nd experiment)	0.7975	0.8032	150	40
LN16	0.7996	0.7999	50, 30	20
DTF spain 2009 (2nd experiment)	0.7917	0.7988	150	55
LDV	0.8158	0.7975	150	15
0W GH FT	0.7873	0.7942	50, 30	15
DTFmainEffect2009	0.7794	0.7855	50, 35, 15	35
SD	0.7905	0.7848	48	30
DTFplantingSummer2008	0.75	0.7746	50, 30	20
FT GH	0.7693	0.7702	50, 30	15
DTFlocSweden2009	0.7595	0.7626	50, 30	60
DTFplantingSummer2009	0.7521	0.7584	50, 30	50
0W	0.7488	0.7473	48	40
DTF spain 2009 (1st experiment)	0.7691	0.7425	48	40
DTF sweden 2008 (1st experiment)	0.727	0.728	50, 30	20
DTFlocSweden2008	0.7161	0.7271	50, 30	55
Seed Dormancy	0.7014	0.7241	50, 30	35
DTFmainEffect2008	0.7102	0.7142	50, 30	20
8W	0.7259	0.7083	150	50
LN22	0.7004	0.7069	50, 30	20
Size sweden 2009 (1st experiment)	0.6905	0.6994	48	50
LN10	0.6934	0.698	50, 30	20
DTF spain 2008 (1st experiment)	0.6944	0.677	150	25
SDV	0.6775	0.6728	150	15
8W GH FT	0.7001	0.6546	48	40
0W GH LN	0.6568	0.654	50, 30	20
Storage 7 days	0.6496	0.65	50, 30	25
Storage 28 days	0.6627	0.6483	50, 30	55
8W GH LN	0.671	0.6434	48	70

Size sweden 2009 (2nd experiment)	0.6114	0.6268	48	50
SizeLocSweden2009	0.6144	0.619	150	35
FLC	0.6118	0.6161	50, 30	30
LFS GH	0.6178	0.6136	150	35
FT Field	0.7324	0.6112	150	60
LY	0.6072	0.6088	150	60
Storage 56 days	0.6085	0.5788	150	15
LES	0.56	0.5764	150	50
M216T665	0.5155	0.5674	50, 30	50
LC Duration GH	0.5799	0.5664	150	55
M172T666	0.5165	0.5487	150	60
Trichome avg JA	0.588	0.5343	150	55
Secondary Dormancy	0.5184	0.5264	150	30
SizeMainEffect2009	0.52	0.5171	48	50
DSDS50	0.4754	0.5006	50, 30	60
avrPphB	0.5054	0.4942	150	60
Hypocotyl length	0.4934	0.4807	150	50
Size spain 2009 (1st experiment)	0.5121	0.4751	150	50
Yield spain 2009 (1st experiment)	0.5205	0.4719	50, 30	50
Leaf serr 10	0.4636	0.4683	150	55
Size spain 2009 (2nd experiment)	0.471	0.4623	48	50
Trichome avg C	0.4617	0.4385	48	40
Germ in dark	0.4447	0.4382	150	15
YieldMainEffect2009	0.505	0.4345	150	30
FT Diameter Field	0.5004	0.4274	150	15
Bacterial titer	0.5406	0.417	150	55
FRI	0.4011	0.4119	48	30
Rosette Erect 22	0.3973	0.3934	48	30
Area sweden 2009 (1st experiment)	0.4203	0.3895	50, 35, 15	30
Width 10	0.3932	0.3784	50, 30	60
Silique 22	0.4339	0.377	50, 30	50
avrRpt2	0.3757	0.3737	50, 30	30
M130T666	0.4381	0.3733	150	60
SizePlantingSummer2009	0.3769	0.3615	150	5
Area sweden 2009 (2nd experiment)	0.359	0.3542	48	45
FW	0.3397	0.3522	50, 30	25
P31	0.3632	0.3419	50, 30	45
MT GH	0.4016	0.3397	150	50
avrB	0.3304	0.3384	50, 30	30
avrRpm1	0.361	0.3368	50, 30	20
Seed bank 133-91	0.3446	0.3334	150	5
Mg25	0.5321	0.3288	50, 30	60
Leaf roll 10	0.3558	0.3272	48	40
Yield spain 2009 (2nd experiment)	0.4184	0.3197	20, 10	40
Noco2	0.3051	0.3174	48	30
Emwa1	0.3226	0.3124	50, 30	30
FT Duration GH	0.2659	0.3123	48	5



Leaf serr 22	0.3021	0.3108	150	60
Anthocyanin 10	0.3198	0.3107	50, 35, 15	60
Cd114	0.3345	0.3069	50, 30	50
Leaf serr 16	0.2895	0.3011	48	40
Fe56	0.2802	0.3006	150	35
YieldLocSweden2009	0.3431	0.2993	150	60
Width 16	0.3463	0.2983	150	50
Co59	0.2738	0.2953	50, 35, 15	25
K39	0.3036	0.2952	50, 30	60
Leaf roll 16	0.3072	0.2886	150	15
DTFplantingLoc2008	0.2971	0.275	50, 30	5
SizePlantingSummerLocSweden2009	0.2803	0.2704	50, 30	60
Mn55	0.2775	0.2662	50, 30	20
Anthocyanin 22	0.2731	0.2635	150	15
As75	0.254	0.2619	50, 30	35
Na23	0.2564	0.2598	50, 30	15
Ni60	0.2894	0.2539	150	25
Mo98	0.2765	0.2537	50, 30	35
Chlorosis 22	0.2622	0.2453	50, 35, 15	10
Hiks1	0.2441	0.2452	20, 10	20
Zn66	0.2553	0.2444	150	35
B11	0.2891	0.2392	48	40
Germ 16	0.2987	0.2356	50, 30	41
At2	0.2147	0.216	150	15
Emco5	0.166	0.2101	150, 30	20
Se82	0.2192	0.2075	150	25
Mature cell length	0.1987	0.2052	150	45
DW	0.2878	0.2048	50, 30	60
Yield sweden 2009 (1st experiment)	0.2274	0.2033	150	55
As2	0.1774	0.1962	150	15
Meristem zone length	0.1976	0.195	150	50
Germ 10	0.2073	0.1873	20, 10	40
Anthocyanin 16	0.2433	0.1867	20, 10	10
Width 22	0.2224	0.1856	50, 30	50
YieldPlantingSummerLocSweden2009	0.2146	0.18	150	55
DTFplantingSummerLocSweden2009	0.2032	0.1775	150	55
Bs	0.2161	0.1656	50, 30	60
Bs CFU2	0.1672	0.1584	50, 35, 15	15
Germ 22	0.1267	0.1533	50, 30	35
Leaf roll 22	0.1135	0.1511	48	45
RP GH	0.1755	0.1458	150	15
Cu65	0.1543	0.1315	150	5
Li7	0.1611	0.1297	150	60
As	0.1089	0.1227	100	20
At1	0.1473	0.1197	48	40
S34	0.1045	0.11	50, 30	60
YieldPlantingSummer2009	0.1265	0.0984	150	50

---

Silique 16	0.2366	0.0884	50, 30	60
Chlorosis 10	0.0243	0.088	50, 35, 15	55
Ca43	0.3333	0.0732	50, 35, 15	55
Seedling Growth	0.0813	0.0636	48	30
Vern Growth	-0.0096	0.0422	150	15
At2 CFU2	0.0694	0.0378	150	25
Yield sweden 2009 (2nd experiment)	0.0536	0.0355	150	25
As CFU2	0.0312	0.035	150	5
At1 CFU2	0.0818	0.0319	50, 30	50
Aphid number	-0.0246	0.029	50, 35, 15	10
After Vern Growth	-0.1433	0.0057	50, 35, 15	5
Chlorosis 16	-0.0313	-0.0121	150	5
As2 CFU2	0.0504	-0.0325	50, 30	60

## 4.6 Discussion

## **5 GWAS**

### **5.1 Reevaluation of 463 phenotypes from the Ara-Pheno database**

#### **5.1.1 Introduction**

#### **5.1.2 Material and Methods**

#### **5.1.3 Results**

#### **5.1.4 Results**

#### **5.1.5 Discussion**

### **5.2 GWAS in DH landrace populations of maize across and within environments**

#### **5.2.1 Introduction**

#### **5.2.2 Material and Methods**

#### **5.2.3 Results**

#### **5.2.4 Results**

#### **5.2.5 Discussion**



# A Source code

## A.1 GBLUP example

```

1 M <- matrix(c(1,-1,0,0,0,1,1,0,1,-1,0,-1),nrow=3)
2 MtM <- M %*% t(M)
3 MtM
4      [,1] [,2] [,3]
5 [1,]    3  -1    2
6 [2,]   -1    1    0
7 [3,]    2    0    3
8 tMM <- t(M) %*% t(M)
9 tMM
10     [,1] [,2] [,3] [,4]
11 [1,]    2    0    1  -1
12 [2,]    0    1    1  -1
13 [3,]    1    1    2  -2
14 [4,]   -1   -1   -2    2
15
16 P = matrix(rep(2*(mafs - 0.5),3),nrow=3,byrow=T)
17 P
18     [,1] [,2] [,3] [,4]
19 [1,] -0.4 -0.6 -0.8 -0.7
20 [2,] -0.4 -0.6 -0.8 -0.7
21 [3,] -0.4 -0.6 -0.8 -0.7

```

## A.2 gwas.py

```

1 import os
2 import sys
3 import time
4 import numpy as np

```

```

5 import pandas as pd
6 import main
7 import h5py
8
9 # set defaults
10 mac_min = 1
11 batch_size = 500000
12 out_file = "results.csv"
13 m = 'phenotype_value'
14 perm = 1
15 mac_min= 6
16
17 X_file = 'gwas_sample_data/AT_geno.hdf5'
18 Y_file = 'gwas_sample_data/phenotype.csv'
19 K_file = 'gwas_sample_data/kinship_ibs_binary_mac5.h5py'
20
21
22
23 for i in range (1,len(sys.argv),2):
24     if sys.argv[i] == "-x" or sys.argv[i] == "--genotype":
25         X_file = sys.argv[i+1]
26     elif sys.argv[i] == "-y" or sys.argv[i] == "--phenotype":
27         Y_file = sys.argv[i+1]
28     elif sys.argv[i] == "-k" or sys.argv[i] == "--kinship":
29         K_file = sys.argv[i+1]
30     elif sys.argv[i] == "-m":
31         m = sys.argv[i+1]
32     elif sys.argv[i] == "-a" or sys.argv[i] == "--mac_min":
33         mac_min = int(sys.argv[i+1])
34     elif sys.argv[i] == "-bs" or sys.argv[i] == "--batch-size":
35         batch_size = int(sys.argv[i+1])
36     elif sys.argv[i] == "-p" or sys.argv[i] == "--perm":
37         perm = int(sys.argv[i+1])
38     elif sys.argv[i] == "-o" or sys.argv[i] == "--out":
39         out_file = sys.argv[i+1]
40     elif sys.argv[i] == "-h" or sys.argv[i] == "--help":
41         print("-x , --genotype :file containing marker information
in csv or hdf5 format of size")

```

```

42     print("-y , --phenotype: file container phenotype
information in csv format" )
43     print("-k , --kinship : file containing kinship matrix of
size k X k in csv or hdf5 format")
44     print("-m : name of columnn containing the phenotype :
default m = phenotype_value")
45     print("-a , --mac_min : integer specifying the minimum
minor allele count necessary for a marker to be included.
Default a = 1" )
46     print("-bs, --batch-size : integer specifying the number of
markers processed at once. Default -bs 500000" )
47     print("-p , --perm : single integer specifying the number
of permutations. Default 1 == no perm ")
48     print("-o , --out : name of output file. Default -o results
.csv ")
49     print("-h , --help : prints help and command line options")
50     quit()
51     else:
52         print('unknown option ' + str(sys.argv[i]))
53         quit()
54
55
56
57 print("parsed commandline args")
58
59 start = time.time()
60
61 X,K,Y_,markers = main.load_and_prepare_data(X_file,Y_file,K_file,m)
62
63
64 ## MAF filterin
65 markers_used , X , macs = main.mac_filter(mac_min,X,markers)
66
67 ## prepare
68 print("Begin performing GWAS on ", Y_file)
69
70 if perm == 1:
71     output = main.gwas(X,K,Y_,batch_size)

```

```

72     if( X_file.split(".")[ -1] == 'csv'):
73         chr_pos = np.array(list(map(lambda x : x.split("- "),
markers_used)))
74     else:
75         chr_reg = h5py.File(X_file, 'r')['positions'].attrs['
chr_regions']
76         mk_index= np.array(range(len(markers)), dtype=int)[macs >=
mac_min]
77         chr_pos = np.array([list(map(lambda x: sum(x > chr_reg
[: ,1]) + 1, mk_index)), markers_used]).T
78         my_time = np.repeat((time.time()-start), len(chr_pos))
79     pd.DataFrame({
80         'chr' : chr_pos[:,0] ,
81         'pos' : chr_pos[:,1] ,
82         'pval': output[:,0] ,
83         'mac' : np.array(macs[macs >= mac_min], dtype=np.int) ,
84         'eff_size': output[:,1] ,
85         'SE' : output[:,2]}).to_csv(out_file, index=False)
86 elif perm > 1:
87     min_pval = []
88     perm_seeds = []
89     my_time = []
90     for i in range(perm):
91         start_perm = time.time()
92         print("Running permutation ", i+1, " of ", perm)
93         my_seed = np.asscalar(np.random.randint(9999, size=1))
94         perm_seeds.append(my_seed)
95         np.random.seed(my_seed)
96         Y_perm = np.random.permutation(Y_)
97         output = main.gwas(X, K, Y_perm, batch_size)
98         min_pval.append(np.min(output[:,0]))
99         print("Elapsed time for permutatation", i+1, " with p_min",
min_pval[i], " is", ":", round(time.time() - start_perm, 2))
100         my_time.append(time.time()-start_perm)
101     pd.DataFrame({
102         'time': my_time ,
103         'seed': perm_seeds ,
104         'min_p': min_pval }).to_csv(out_file, index=False)

```



```

105
106 print("done")
107
108 end = time.time()
109 eltime = np.round(end -start,2)
110
111 if eltime <= 59:
112     print("Total time elapsed", eltime, "seconds")
113 elif eltime > 59 and eltime <= 3600:
114     print("Total time elapsed", np.round(eltime / 60,2) , "minutes
115 ")
116 elif eltime > 3600 :
117     print("Total time elapsed", np.round(eltime / 60 / 60,2), "
118 hours")

```

## A.3 main.py

```

1     import pandas as pd
2     import numpy as np
3     from scipy.stats import f
4     import tensorflow as tf
5     import limix
6     import herit
7     import h5py
8     import limix
9     import multiprocessing as mlt
10
11     def load_and_prepare_data(X_file,Y_file,K_file,m):
12         type_K = K_file.split(".")[ -1]
13         type_X = X_file.split(".")[ -1]
14
15         ## load and preprocess genotype matrix
16         Y = pd.read_csv(Y_file,engine='python').sort_values(['
17         accession_id']).groupby('accession_id').mean()
18         Y = pd.DataFrame({'accession_id' : Y.index, 'phenotype_value'
19 : Y[m]})

```

```

18     if type_X == 'hdf5' or type_X == 'h5py' :
19         SNP = h5py.File(X_file, 'r')
20         markers = np.asarray(SNP['positions'])
21         acc_X = np.asarray(SNP['accessions'][:], dtype=np.int)
22     elif type_X == 'csv' :
23         X = pd.read_csv(X_file, index_col=0)
24         markers = X.columns.values
25         acc_X = X.index
26         X = np.asarray(X, dtype=np.float32)/2
27     else :
28         sys.exit("Only hdf5, h5py and csv files are supported")
29
30     if type_K == 'hdf5' or type_K == 'h5py':
31         k = h5py.File(K_file, 'r')
32         acc_K = np.asarray(k['accessions'][:], dtype=np.int)
33     elif type_K == 'csv':
34         k = pd.read_csv(K_file, index_col=0)
35         acc_K = k.index
36         k = np.array(k, dtype=np.float32)
37
38     acc_Y = np.asarray(Y[['accession_id']]).flatten()
39     acc_isec = [isec for isec in acc_X if isec in acc_Y]
40
41     idx_acc = list(map(lambda x: x in acc_isec, acc_X))
42     idy_acc = list(map(lambda x: x in acc_isec, acc_Y))
43     idk_acc = list(map(lambda x: x in acc_isec, acc_K))
44
45     Y_ = np.asarray(Y.drop('accession_id', 1), dtype=np.float32)[
46         idy_acc, :]
47
48     if type_X == 'hdf5' or type_X == 'h5py' :
49         X = np.asarray(SNP['snps'][0:(len(SNP['snps'])+1),:], dtype=
50             np.float32)[: , idx_acc].T
51         X = X[np.argsort(acc_X[idx_acc]), :]
52         k1 = np.asarray(k['kinship'][:]) [idk_acc, :]
53         K = k1[:, idk_acc]
54         K = K[np.argsort(acc_X[idx_acc]), :]
55         K = K[:, np.argsort(acc_X[idx_acc])]

```

```

54     else:
55         X = X[idx_acc,:]
56         k1 = k[idk_acc,:]
57         K = k1[:,idk_acc]
58
59
60     print("data has been imported")
61     return X,K,Y_,markers
62
63
64 def mac_filter(mac_min, X, markers):
65     ac1 = np.sum(X,axis=0)
66     ac0 = X.shape[0] - ac1
67     macs = np.minimum(ac1,ac0)
68     markers_used = markers[macs >= mac_min]
69     X = X[:,macs >= mac_min]
70     return markers_used, X, macs
71
72 def gwas(X,K,Y,batch_size):
73     n_marker = X.shape[1]
74     n = len(Y)
75     ## REML
76     K_stand = (n-1)/np.sum((np.identity(n) - np.ones((n,n))/n) * K)
77     * K
78     vg, delta, ve = herit.estimate(Y,"normal",K_stand,verbose =
False)
79     print(" Pseudo-heritability is " , vg / (ve + vg + delta))
80     print(" Performing GWAS on ", n , " phenotypes and ", n_marker
,"markers")
81     ## Transform kinship-matrix, phenotypes and estimate intercpt
82     Xo = np.ones(K.shape[0]).flatten()
83     M = np.transpose(np.linalg.inv(np.linalg.cholesky(vg * K_stand
+ ve * np.identity(n))))).astype(np.float32)
84     Y_t = np.sum(np.multiply(np.transpose(M),Y),axis=1).astype(np.
float32)
85     int_t = np.sum(np.multiply(np.transpose(M),np.ones(n)),axis=1).
astype(np.float32)
86     ## EMMAX Scan

```

```

86     RSS_env = (np.linalg.lstsq(np.reshape(int_t,(n,-1)) , np.
reshape(Y_t,(n,-1)))[1]).astype(np.float32)
87     ## calculate betas and se of betas
88     def stderr(a,M,Y_t2d,int_t):
89         x = tf.stack((int_t,tf.squeeze(tf.matmul(M.T,tf.reshape(a
,(n,-1))))),axis=1)
90         coeff = tf.matmul(tf.matmul(tf.linalg.inv(tf.matmul(tf.
transpose(x),x)),tf.transpose(x)),Y_t2d)
91         SSE = tf.reduce_sum(tf.math.square(tf.math.subtract(Y_t,tf
.math.add(tf.math.multiply(x[:,1],coeff[0,0]),tf.math.multiply(x
[:,1],coeff[1,0])))))
92         SE = tf.math.sqrt(SSE/(471-(1+2)))
93         StdERR = tf.sqrt(tf.linalg.diag_part(tf.math.multiply(SE ,
tf.linalg.inv(tf.matmul(tf.transpose(x),x)))[1]
94         return tf.stack((coeff[1,0],StdERR))
95     ## calculate residual sum squares
96     def rss(a,M,y,int_t):
97         x_t = tf.reduce_sum(tf.math.multiply(M.T,a),axis=1)
98         lm_res = tf.linalg.lstsq(tf.transpose(tf.stack((int_t,x_t)
,axis=0)),Y_t2d)
99         lm_x = tf.concat((tf.squeeze(lm_res),x_t),axis=0)
100        return tf.reduce_sum(tf.math.square(tf.math.subtract(tf.
squeeze(Y_t2d),tf.math.add(tf.math.multiply(lm_x[1],lm_x[2:]),
tf.multiply(lm_x[0],int_t)))))
101    ## loop over the batches
102    for i in range(int(np.ceil(n_marker/batch_size))):
103        tf.reset_default_graph()
104        if n_marker < batch_size:
105            X_sub = X
106        else:
107            lower_limit = batch_size * i
108            upper_limit = batch_size * i + batch_size
109            if upper_limit <= n_marker :
110                X_sub = X[:,lower_limit:upper_limit]
111                print("Working on markers ", lower_limit , " to ",
upper_limit, " of ", n_marker )
112            else:
113                X_sub = X[:,lower_limit:]

```

```

114         print("Working on markers ", lower_limit , " to ",
n_marker, " of ", n_marker )
115         config = tf.ConfigProto()
116         n_cores = mlt.cpu_count()
117         config.intra_op_parallelism_threads = n_cores
118         config.inter_op_parallelism_threads = n_cores
119         sess = tf.Session(config=config)
120         Y_t2d = tf.cast(tf.reshape(Y_t,(n,-1)),dtype=tf.float32)
121         y_tensor = tf.convert_to_tensor(Y_t,dtype = tf.float32)
122         StdERR = tf.map_fn(lambda a : stderr(a,M,Y_t2d,int_t),
X_sub.T)
123         R1_full = tf.map_fn(lambda a: rss(a,M,Y_t2d,int_t), X_sub.T
)
124         F_1 = tf.divide(tf.subtract(RSS_env, R1_full),tf.divide(
R1_full,(n-3)))
125         if i == 0 :
126             output = sess.run(tf.concat([tf.reshape(F_1,(X_sub.
shape[1],-1)),StdERR],axis=1))
127         else :
128             tmp = sess.run(tf.concat([tf.reshape(F_1,(X_sub.shape
[1],-1)),StdERR],axis=1))
129             output = np.append(output,tmp,axis=0)
130             sess.close()
131             F_dist = output[:,0]
132             pval = 1 - f.cdf(F_dist,1,n-3)
133             output[:,0] = pval
134         return output
135
136
137

```

## A.4 herit.py

```

1
2 def estimate(y, lik, K, M=None, verbose=True):
3     from numpy_sugar.linalg import economic_qs
4     from numpy import pi, var, diag
5     from glimix_core.glmm import GLMMExpFam

```

```

6     from glimix_core.lmm import LMM
7     from limix._data._assert import assert_likelihoood
8     from limix._data import normalize_likelihoood, conform_dataset
9     from limix.qtl._assert import assert_finite
10    from limix._display import session_block, session_line
11    lik = normalize_likelihoood(lik)
12    lik_name = lik[0]
13    with session_block("Heritability analysis", disable=not verbose
14    ):
15        with session_line("Normalising input...", disable=not
16        verbose):
17            data = conform_dataset(y, M=M, K=K)
18            y = data["y"]
19            M = data["M"]
20            K = data["K"]
21            assert_finite(y, M, K)
22            if K is not None:
23                # K = K / diag(K).mean()
24                QS = economic_qs(K)
25            else:
26                QS = None
27            if lik_name == "normal":
28                method = LMM(y.values, M.values, QS, restricted=True)
29                method.fit(verbose=verbose)
30            else:
31                method = GLMMExpFam(y, lik, M.values, QS, n_int=500)
32                method.fit(verbose=verbose, factr=1e6, pgtol=1e-3)
33            g = method.scale * (1 - method.delta)
34            e = method.scale * method.delta
35            if lik_name == "bernoulli":
36                e += pi * pi / 3
37            v = var(method.mean())
38            return g, v, e

```

## B *A. thaliana* phenotypic data

ID	Phenotype name	doi	Reference
1	FT Diameter Field	10.21958/phenotype:1	<i>Atwell</i> et al., 2010
2	At2 CFU2	10.21958/phenotype:2	<i>Atwell</i> et al., 2010
3	Leaf serr 16	10.21958/phenotype:3	<i>Atwell</i> et al., 2010
4	Seed bank 133-91	10.21958/phenotype:4	<i>Atwell</i> et al., 2010
5	Na23	10.21958/phenotype:5	<i>Atwell</i> et al., 2010
6	Leaf serr 10	10.21958/phenotype:6	<i>Atwell</i> et al., 2010
7	Emco5	10.21958/phenotype:7	<i>Atwell</i> et al., 2010
8	Leaf roll 16	10.21958/phenotype:8	<i>Atwell</i> et al., 2010
9	Leaf roll 10	10.21958/phenotype:9	<i>Atwell</i> et al., 2010
10	Bs	10.21958/phenotype:10	<i>Atwell</i> et al., 2010
11	2W	10.21958/phenotype:11	<i>Atwell</i> et al., 2010
12	Rosette Erect 22	10.21958/phenotype:12	<i>Atwell</i> et al., 2010
13	Cd114	10.21958/phenotype:13	<i>Atwell</i> et al., 2010
14	Width 16	10.21958/phenotype:14	<i>Atwell</i> et al., 2010
15	Storage 28 days	10.21958/phenotype:15	<i>Atwell</i> et al., 2010
16	LY	10.21958/phenotype:16	<i>Atwell</i> et al., 2010
17	avrRpm1	10.21958/phenotype:17	<i>Atwell</i> et al., 2010
18	Width 10	10.21958/phenotype:18	<i>Atwell</i> et al., 2010
19	Chlorosis 22	10.21958/phenotype:19	<i>Atwell</i> et al., 2010
20	Storage 7 days	10.21958/phenotype:20	<i>Atwell</i> et al., 2010
21	As2 CFU2	10.21958/phenotype:21	<i>Atwell</i> et al., 2010
22	Co59	10.21958/phenotype:22	<i>Atwell</i> et al., 2010

---

23	FW	10.21958/phenotype:23	<i>Atwell et al., 2010</i>
24	Cu65	10.21958/phenotype:24	<i>Atwell et al., 2010</i>
25	Bacterial titer	10.21958/phenotype:25	<i>Atwell et al., 2010</i>
26	Width 22	10.21958/phenotype:26	<i>Atwell et al., 2010</i>
27	Storage 56 days	10.21958/phenotype:27	<i>Atwell et al., 2010</i>
28	YEL	10.21958/phenotype:28	<i>Atwell et al., 2010</i>
29	FLC	10.21958/phenotype:29	<i>Atwell et al., 2010</i>
30	FT16	10.21958/phenotype:30	<i>Atwell et al., 2010</i>
31	FT10	10.21958/phenotype:31	<i>Atwell et al., 2010</i>
32	FT Duration GH	10.21958/phenotype:32	<i>Atwell et al., 2010</i>
33	Se82	10.21958/phenotype:33	<i>Atwell et al., 2010</i>
34	LDV	10.21958/phenotype:34	<i>Atwell et al., 2010</i>
35	Noco2	10.21958/phenotype:35	<i>Atwell et al., 2010</i>
36	8W GH LN	10.21958/phenotype:36	<i>Atwell et al., 2010</i>
37	0W	10.21958/phenotype:37	<i>Atwell et al., 2010</i>
38	MT GH	10.21958/phenotype:38	<i>Atwell et al., 2010</i>
39	After Vern Growth	10.21958/phenotype:39	<i>Atwell et al., 2010</i>
40	Aphid number	10.21958/phenotype:40	<i>Atwell et al., 2010</i>
41	LN22	10.21958/phenotype:41	<i>Atwell et al., 2010</i>
42	Bs CFU2	10.21958/phenotype:42	<i>Atwell et al., 2010</i>
43	avrRpt2	10.21958/phenotype:43	<i>Atwell et al., 2010</i>
44	Hypocotyl length	10.21958/phenotype:44	<i>Atwell et al., 2010</i>
45	Germ 22	10.21958/phenotype:45	<i>Atwell et al., 2010</i>
46	Leaf roll 22	10.21958/phenotype:46	<i>Atwell et al., 2010</i>
47	SD	10.21958/phenotype:47	<i>Atwell et al., 2010</i>
48	8W	10.21958/phenotype:48	<i>Atwell et al., 2010</i>
49	FT GH	10.21958/phenotype:49	<i>Atwell et al., 2010</i>
50	DSDS50	10.21958/phenotype:50	<i>Atwell et al., 2010</i>



51	Ca43	10.21958/phenotype:51	<i>Atwell et al., 2010</i>
52	LC Duration GH	10.21958/phenotype:52	<i>Atwell et al., 2010</i>
53	0W GH FT	10.21958/phenotype:53	<i>Atwell et al., 2010</i>
54	B11	10.21958/phenotype:54	<i>Atwell et al., 2010</i>
55	Chlorosis 10	10.21958/phenotype:55	<i>Atwell et al., 2010</i>
56	RP GH	10.21958/phenotype:56	<i>Atwell et al., 2010</i>
57	Chlorosis 16	10.21958/phenotype:57	<i>Atwell et al., 2010</i>
58	LFS GH	10.21958/phenotype:58	<i>Atwell et al., 2010</i>
59	Germ 10	10.21958/phenotype:59	<i>Atwell et al., 2010</i>
60	Germ 16	10.21958/phenotype:60	<i>Atwell et al., 2010</i>
61	Anthocyanin 16	10.21958/phenotype:61	<i>Atwell et al., 2010</i>
62	Anthocyanin 10	10.21958/phenotype:62	<i>Atwell et al., 2010</i>
63	At1 CFU2	10.21958/phenotype:63	<i>Atwell et al., 2010</i>
64	Ni60	10.21958/phenotype:64	<i>Atwell et al., 2010</i>
65	P31	10.21958/phenotype:65	<i>Atwell et al., 2010</i>
66	Emwa1	10.21958/phenotype:66	<i>Atwell et al., 2010</i>
67	As75	10.21958/phenotype:67	<i>Atwell et al., 2010</i>
68	Germ in dark	10.21958/phenotype:68	<i>Atwell et al., 2010</i>
69	FRI	10.21958/phenotype:69	<i>Atwell et al., 2010</i>
70	As CFU2	10.21958/phenotype:70	<i>Atwell et al., 2010</i>
71	Trichome avg C	10.21958/phenotype:71	<i>Atwell et al., 2010</i>
72	Vern Growth	10.21958/phenotype:72	<i>Atwell et al., 2010</i>
73	Mo98	10.21958/phenotype:73	<i>Atwell et al., 2010</i>
74	Hiks1	10.21958/phenotype:74	<i>Atwell et al., 2010</i>
75	Anthocyanin 22	10.21958/phenotype:75	<i>Atwell et al., 2010</i>
76	Zn66	10.21958/phenotype:76	<i>Atwell et al., 2010</i>
77	Trichome avg JA	10.21958/phenotype:77	<i>Atwell et al., 2010</i>
78	LES	10.21958/phenotype:78	<i>Atwell et al., 2010</i>

---

79	Silique 16	10.21958/phenotype:79	<i>Atwell et al., 2010</i>
80	Emoy*	10.21958/phenotype:80	<i>Atwell et al., 2010</i>
81	K39	10.21958/phenotype:81	<i>Atwell et al., 2010</i>
82	0W GH LN	10.21958/phenotype:82	<i>Atwell et al., 2010</i>
83	At2	10.21958/phenotype:83	<i>Atwell et al., 2010</i>
84	At1	10.21958/phenotype:84	<i>Atwell et al., 2010</i>
85	LN10	10.21958/phenotype:85	<i>Atwell et al., 2010</i>
86	FT Field	10.21958/phenotype:86	<i>Atwell et al., 2010</i>
87	LN16	10.21958/phenotype:87	<i>Atwell et al., 2010</i>
88	avrB	10.21958/phenotype:88	<i>Atwell et al., 2010</i>
89	LD	10.21958/phenotype:89	<i>Atwell et al., 2010</i>
90	Seedling Growth	10.21958/phenotype:90	<i>Atwell et al., 2010</i>
91	S34	10.21958/phenotype:91	<i>Atwell et al., 2010</i>
92	Leaf serr 22	10.21958/phenotype:92	<i>Atwell et al., 2010</i>
93	DW	10.21958/phenotype:93	<i>Atwell et al., 2010</i>
94	Seed Dormancy	10.21958/phenotype:94	<i>Atwell et al., 2010</i>
95	Mn55	10.21958/phenotype:95	<i>Atwell et al., 2010</i>
96	Silique 22	10.21958/phenotype:96	<i>Atwell et al., 2010</i>
97	avrPphB	10.21958/phenotype:97	<i>Atwell et al., 2010</i>
98	Fe56	10.21958/phenotype:98	<i>Atwell et al., 2010</i>
99	8W GH FT	10.21958/phenotype:99	<i>Atwell et al., 2010</i>
100	4W	10.21958/phenotype:100	<i>Atwell et al., 2010</i>
101	Li7	10.21958/phenotype:101	<i>Atwell et al., 2010</i>
102	FT22	10.21958/phenotype:102	<i>Atwell et al., 2010</i>
103	As2	10.21958/phenotype:103	<i>Atwell et al., 2010</i>
104	SDV	10.21958/phenotype:104	<i>Atwell et al., 2010</i>
105	Mg25	10.21958/phenotype:105	<i>Atwell et al., 2010</i>
106	Secondary Dormancy	10.21958/phenotype:106	<i>Atwell et al., 2010</i>

107	As	10.21958/phenotype:107	<i>Atwell et al., 2010</i>
108	Area Sweden 2009 (1st experiment)	10.21958/phenotype:108	<i>Li et al., 2010</i>
109	Size Planting Summer 2009	10.21958/phenotype:109	<i>Li et al., 2010</i>
110	Size Sweden 2009 (2nd experiment)	10.21958/phenotype:110	<i>Li et al., 2010</i>
111	Size Planting Summer Loc Sweden 2009	10.21958/phenotype:111	<i>Li et al., 2010</i>
112	Area Sweden 2009 (2nd experiment)	10.21958/phenotype:112	<i>Li et al., 2010</i>
113	DTF Sweden 2008 (1st experiment)	10.21958/phenotype:113	<i>Li et al., 2010</i>
114	Yield Sweden 2009 (2nd experiment)	10.21958/phenotype:114	<i>Li et al., 2010</i>
115	Size Loc Sweden 2009	10.21958/phenotype:115	<i>Li et al., 2010</i>
116	DTF planting Summer Loc Sweden 2009	10.21958/phenotype:116	<i>Li et al., 2010</i>
117	DTF loc Sweden 2008	10.21958/phenotype:117	<i>Li et al., 2010</i>
118	DTF loc Sweden 2009	10.21958/phenotype:118	<i>Li et al., 2010</i>
119	DTF Spain 2009 (1st experiment)	10.21958/phenotype:119	<i>Li et al., 2010</i>
120	DTF planting Loc 2008	10.21958/phenotype:120	<i>Li et al., 2010</i>
121	DTF Spain 2009 (2nd experiment)	10.21958/phenotype:121	<i>Li et al., 2010</i>
122	Yield Spain 2009 (2nd experiment)	10.21958/phenotype:122	<i>Li et al., 2010</i>
123	Size Sweden 2009 (1st experiment)	10.21958/phenotype:123	<i>Li et al., 2010</i>

---

124	Yield Spain 2009 (1st experiment)	10.21958/phenotype:124	<i>Li et al., 2010</i>
125	DTF main Effect 2009	10.21958/phenotype:125	<i>Li et al., 2010</i>
126	DTF main Effect 2008	10.21958/phenotype:126	<i>Li et al., 2010</i>
127	Size Spain 2009 (2nd experiment)	10.21958/phenotype:127	<i>Li et al., 2010</i>
128	Size Spain 2009 (1st experiment)	10.21958/phenotype:128	<i>Li et al., 2010</i>
129	DTF planting Summer 2009	10.21958/phenotype:129	<i>Li et al., 2010</i>
130	DTF planting Summer 2008	10.21958/phenotype:130	<i>Li et al., 2010</i>
131	Size Main Effect 2009	10.21958/phenotype:131	<i>Li et al., 2010</i>
132	DTF Spain 2008 (1st experiment)	10.21958/phenotype:132	<i>Li et al., 2010</i>
133	Yield Planting Summer 2009	10.21958/phenotype:133	<i>Li et al., 2010</i>
134	DTF Sweden 2009 (1st experiment)	10.21958/phenotype:134	<i>Li et al., 2010</i>
135	Yield Loc Sweden 2009	10.21958/phenotype:135	<i>Li et al., 2010</i>
136	DTF Spain 2008 (2nd experiment)	10.21958/phenotype:136	<i>Li et al., 2010</i>
137	Yield Main Effect 2009	10.21958/phenotype:137	<i>Li et al., 2010</i>
138	Yield Planting Summer Loc Sweden 009	10.21958/phenotype:138	<i>Li et al., 2010</i>
139	Yield Sweden 2009 (1st experiment)	10.21958/phenotype:139	<i>Li et al., 2010</i>
140	DTF Sweden 2009 (2nd experiment)	10.21958/phenotype:140	<i>Li et al., 2010</i>

---

141	DTF Sweden 2008 (2nd experiment)	10.21958/phenotype:141	Li et al., 2010
142	Mature cell length	10.21958/phenotype:142	Meijó et al., 2014
143	Meristem zone length	10.21958/phenotype:143	Meijó et al., 2014
144	M216T665	10.21958/phenotype:144	Strauch et al., 2015
145	M130T666	10.21958/phenotype:145	Strauch et al., 2015
146	M172T666	10.21958/phenotype:146	Strauch et al., 2015
261	FT10	10.21958/phenotype:261	Alonso-Blanco et al., 2016
262	FT16	10.21958/phenotype:262	Alonso-Blanco et al., 2016
269	Li7	10.21958/phenotype:269	Forsberg et al., 2015
270	B11	10.21958/phenotype:270	Forsberg et al., 2015
271	Na23	10.21958/phenotype:271	Forsberg et al., 2015
272	Mg25	10.21958/phenotype:272	Forsberg et al., 2015
273	P31	10.21958/phenotype:273	Forsberg et al., 2015
274	S34	10.21958/phenotype:274	Forsberg et al., 2015
275	K39	10.21958/phenotype:275	Forsberg et al., 2015
276	Ca43	10.21958/phenotype:276	Forsberg et al., 2015
277	Mn55	10.21958/phenotype:277	Forsberg et al., 2015
279	Co59	10.21958/phenotype:279	Forsberg et al., 2015
280	Ni60	10.21958/phenotype:280	Forsberg et al., 2015
281	Cu65	10.21958/phenotype:281	Forsberg et al., 2015
282	Zn66	10.21958/phenotype:282	Forsberg et al., 2015
283	As75	10.21958/phenotype:283	Forsberg et al., 2015
284	Se82	10.21958/phenotype:284	Forsberg et al., 2015

---



# C Genomic prediction

## C.1 GP Scripts

```

1 import os,sys,gc
2 import pandas as pd
3 import numpy as np
4 import timeit
5 from datetime import datetime
6 import keras
7 import tensorflow as tf
8 from keras import backend as K
9 from keras import layers
10 from keras.models import Sequential
11 from keras.layers import Dense, Dropout, GaussianNoise,
    AlphaDropout, Reshape
12 from keras.layers import Flatten, LocallyConnected1D,
    LocallyConnected2D
13 from keras.optimizers import Adam, Adagrad, Adadelta
14 from keras.backend.tensorflow_backend import set_session
15
16 ##set default values
17
18 learning_rate = 0.01
19 JobID = 1
20 ps = 25
21 optim = "adam"
22 X_file = "KE.geno.csv"
23 Y_file = "KE.pheno.csv"
24 CV_file = "KE_cv_pw.csv"
25 label = "DtSILK"
26 start_time = timeit.default_timer()
27 act="relu"

```

```

28 drop_rate = str('0.5,0.5,0.5')
29 arc = str('63,63')
30 DG = 'D,D,D,D,D,G'
31 LC = True
32 training_epochs = 25
33 hyp = False
34
35 ### parse command line arguments
36
37 for i in range (1,len(sys.argv),2):
38     if sys.argv[i] == "-x":
39         X_file = sys.argv[i+1]
40     elif sys.argv[i] == "-y":
41         Y_file = sys.argv[i+1]
42     elif sys.argv[i] == "-cv":
43         CV_file = sys.argv[i+1]
44     elif sys.argv[i] == "-JobID":
45         JobID = int(sys.argv[i+1])
46     elif sys.argv[i] == "-label":
47         label = sys.argv[i+1]
48     elif sys.argv[i] == "-act":
49         act = str(sys.argv[i+1])
50     elif sys.argv[i] == "-epochs":
51         training_epochs = int(sys.argv[i+1])
52     elif sys.argv[i] == "-lr":
53         learning_rate = float(sys.argv[i+1])
54     elif sys.argv[i] == "-arc":
55         arc = sys.argv[i+1]
56     elif sys.argv[i] == "-ps":
57         ps = int(sys.argv[i+1])
58     elif sys.argv[i] == "-dr":
59         drop_rate=str(sys.argv[i+1])
60     elif sys.argv[i] == "-LC":
61         LC = bool(sys.argv[i+1])
62     elif sys.argv[i] == "hyp":
63         hyp = bool(sys.argv[i+1])
64     else:
65         print('unknown option ' + str(sys.argv[i]))

```



```

66         quit()
67
68
69
70 ## change dir to data location
71
72
73 #os.chdir('/home/jaf81qa/jan_storage/tens')
74 x = pd.read_csv(X_file, index_col = 0)
75 #os.chdir("/storage/full-share/genoPred/maze")
76 y = pd.read_csv(Y_file, index_col = 0)
77 cv_folds = pd.read_csv(CV_file, index_col=0)
78
79 ## select column of phenotype file via columnname
80
81 y = y[[label]]
82 ## activity_regularizer=regularizers.l1(0.01)))
83
84 def build_network(arc, drop_rate, LC, DG):
85     def add_drops(model, drop_out, k):
86         if DG[k].upper() == 'D':
87             model.add(Dropout(drop_out[0]))
88         elif DG[k].upper() == 'G':
89             model.add(GaussianNoise(drop_out[k]))
90         elif DG[k].upper() == "A":
91             model.add(AlphaDropout(drop_out[k]))
92         else:
93             pass
94     return model
95
96 DG = DG.strip().split(",")
97 arc = arc.strip().split(",")
98 archit = []
99 for layer in arc:
100     archit.append(int(layer))
101 layer_number = len(archit)
102 drop_rate = drop_rate.strip().split(",")
103 drop_out = []
104 for drops in drop_rate:

```

```

104     drop_out.append(float(drops))
105     model = Sequential()
106     if LC == True:
107         model.add(Reshape(input_shape=(x_train.shape[1],),
108 target_shape=(x_train.shape[1],1)))
109         model.add(LocallyConnected1D(1,10, strides=7, input_shape=(
110 x_train.shape[1],1)))
111         model.add(Flatten())
112         start = 0
113         model = add_drops(model, drop_out, start)
114     elif LC == False:
115         model.add(Dense(archit[0], kernel_initializer='
116 truncated_normal', activation=act, input_shape=(x_train.shape
117 [1],)))
118         model = add_drops(model, drop_out, start)
119         start = 1
120         for k in range(start, len(archit)):
121             model.add(Dense(archit[k], kernel_initializer='
122 truncated_normal', activation=act))
123             model = add_drops(model, drop_out, k)
124         model.add(Dense(1, kernel_initializer='truncated_normal'))
125         return(model)
126
127
128
129 config = tf.ConfigProto()
130 #config.gpu_options.per_process_gpu_memory_fraction = 0.1
131 config.gpu_options.allow_growth = True
132 set_session(tf.Session(config=config))
133
134 if not os.path.isfile("RESULTScv50.txt"):
135     out2 = open("RESULTScv50.txt", 'w')
136     out2.write('DateTime\tCompTime\tDF\tGenos\tPhenos\tCV_fold\
137 \tArchit\tConv\tActFun\tEpochs\tdrop_rate\tAccuracy\n' )
138
139 for k in range(1,51):
140     print("Training on cv fold "+ str(k))

```

```

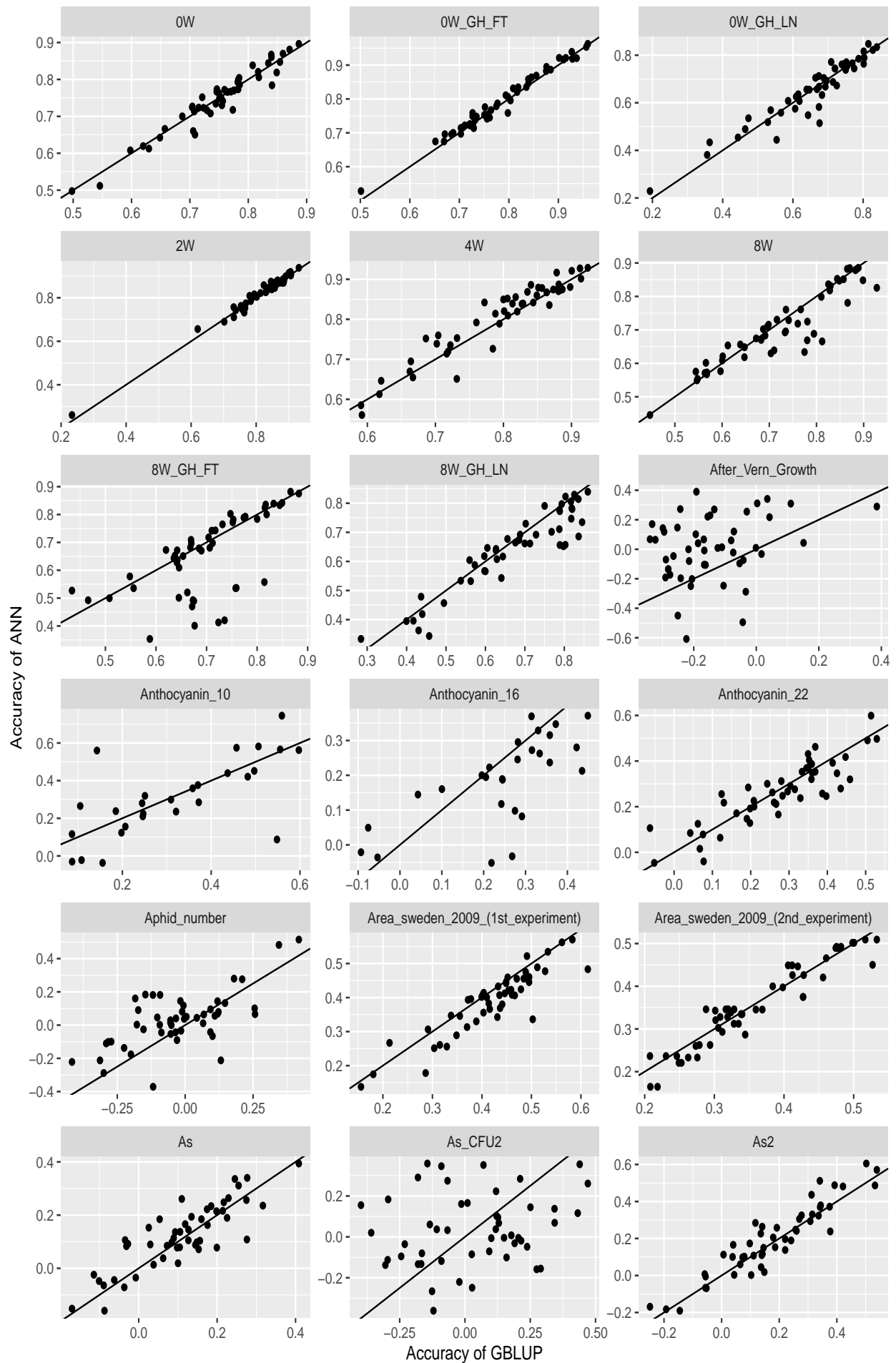
136     cv = cv_folds['cv_' + str(k)]
137     num_cvs = np.ptp(cv) + 1
138
139     i = 1
140     x_train = x[cv != i]
141     x_test = x[cv == i]
142     y_train = y[cv != i]
143     y_test = y[cv == i]
144
145     yhat = np.zeros(shape = y_test.shape)
146
147     model = build_network(arc,drop_rate,LC,DG)
148     model.compile(loss='mse', optimizer=Adam(lr=0.01,decay = 0.001)
149 ,metrics=['accuracy'])
150 #     score = model.evaluate(x_test, y_test, verbose=0)
151     bla = model.predict(x_test)
152     y_sub= y[np.asarray(cv == i)]
153
154     print(model.summary())
155     print('\n')
156     print(label)
157
158     comp_time = int(round(timeit.default_timer() - start_time,0))
159
160     DateTime = datetime.now().strftime('%Y-%m-%d %H:%M:%S')
161     acc = np.corrcoef(bla[:,0],np.asarray(y_sub)[:,0])[0,1]
162
163     out2 = open("RESULTScv50.txt", 'a')
164     out2.write('%s\t%i\t%s\t%s\t%s\t%i\t%s\t%s\t%s\t%i\t%s\t%0.5f\n'
165 , % (
166         DateTime, comp_time, label, X_file, Y_file, int(k), arc, LC
167 ,   act,int(training_epochs), drop_rate, round(acc,4)))
168
169     del model,bla, x_train, x_test, y_train, y_test
170     K.clear_session()
171     gc.collect()

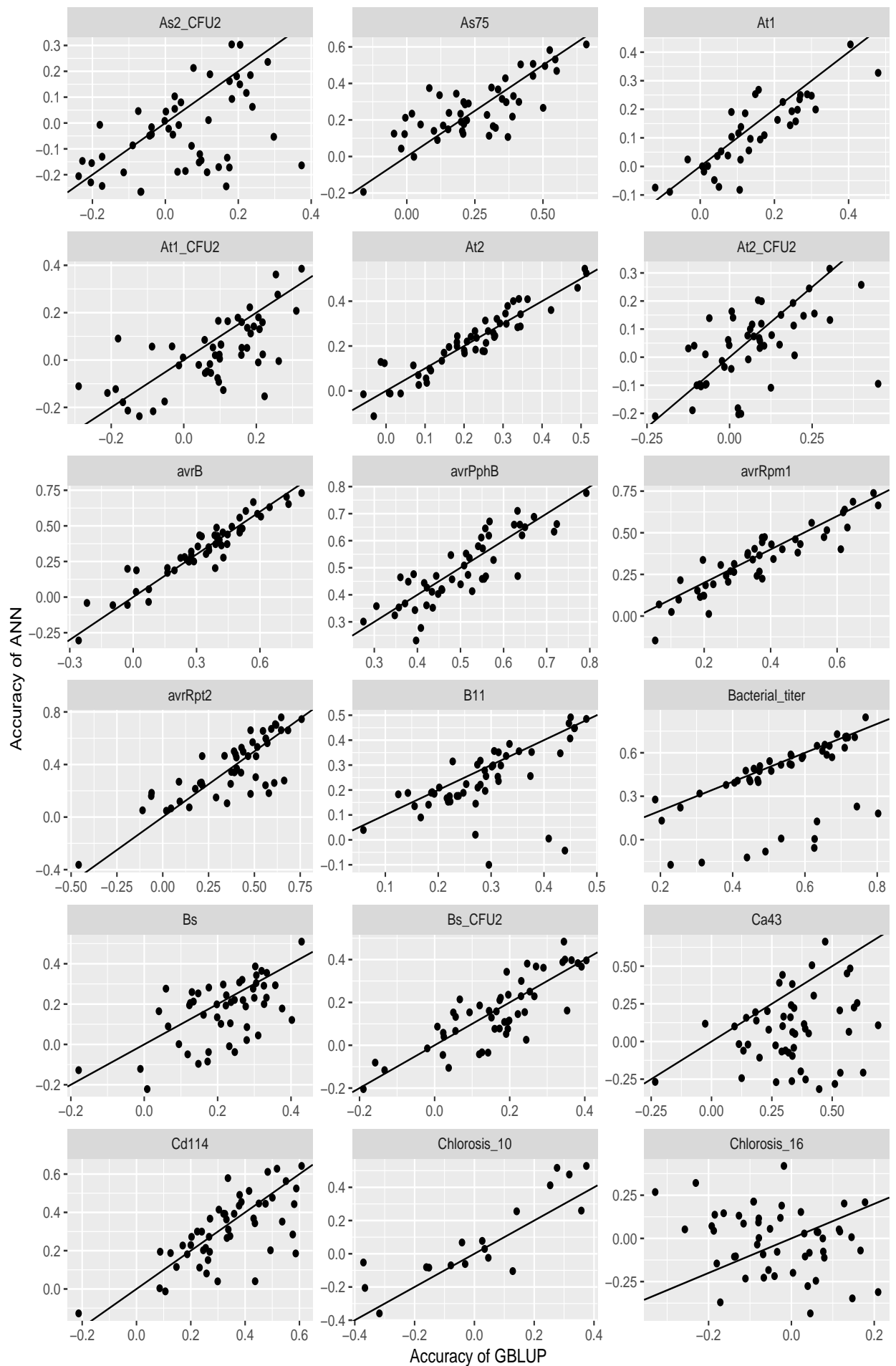
```

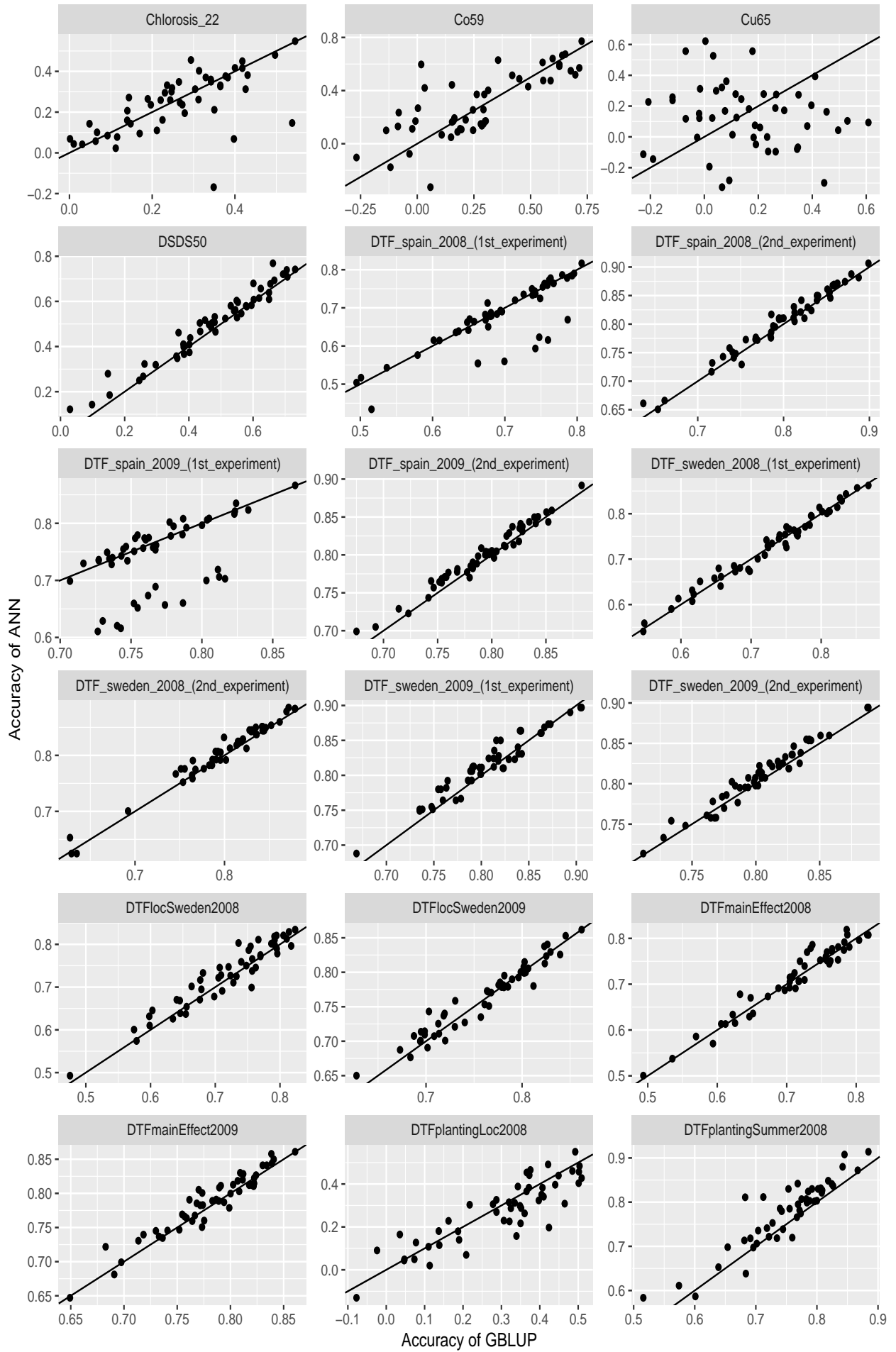
```
171 config = tf.ConfigProto()
172 #config.gpu_options.per_process_gpu_memory_fraction = 0.1
173 config.gpu_options.allow_growth = True
174 set_session(tf.Session(config=config))
175
```



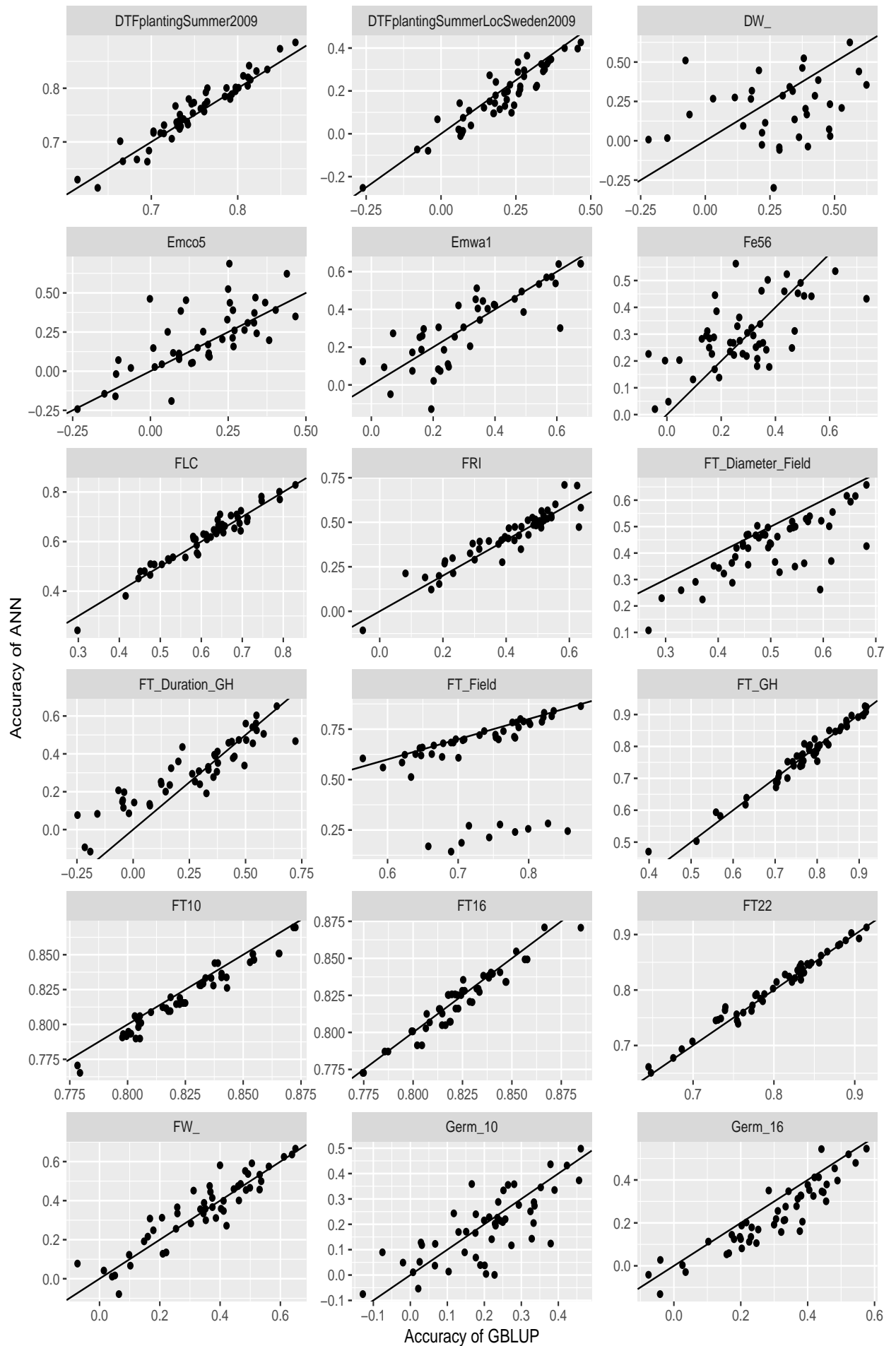
## C.2 Results of GP

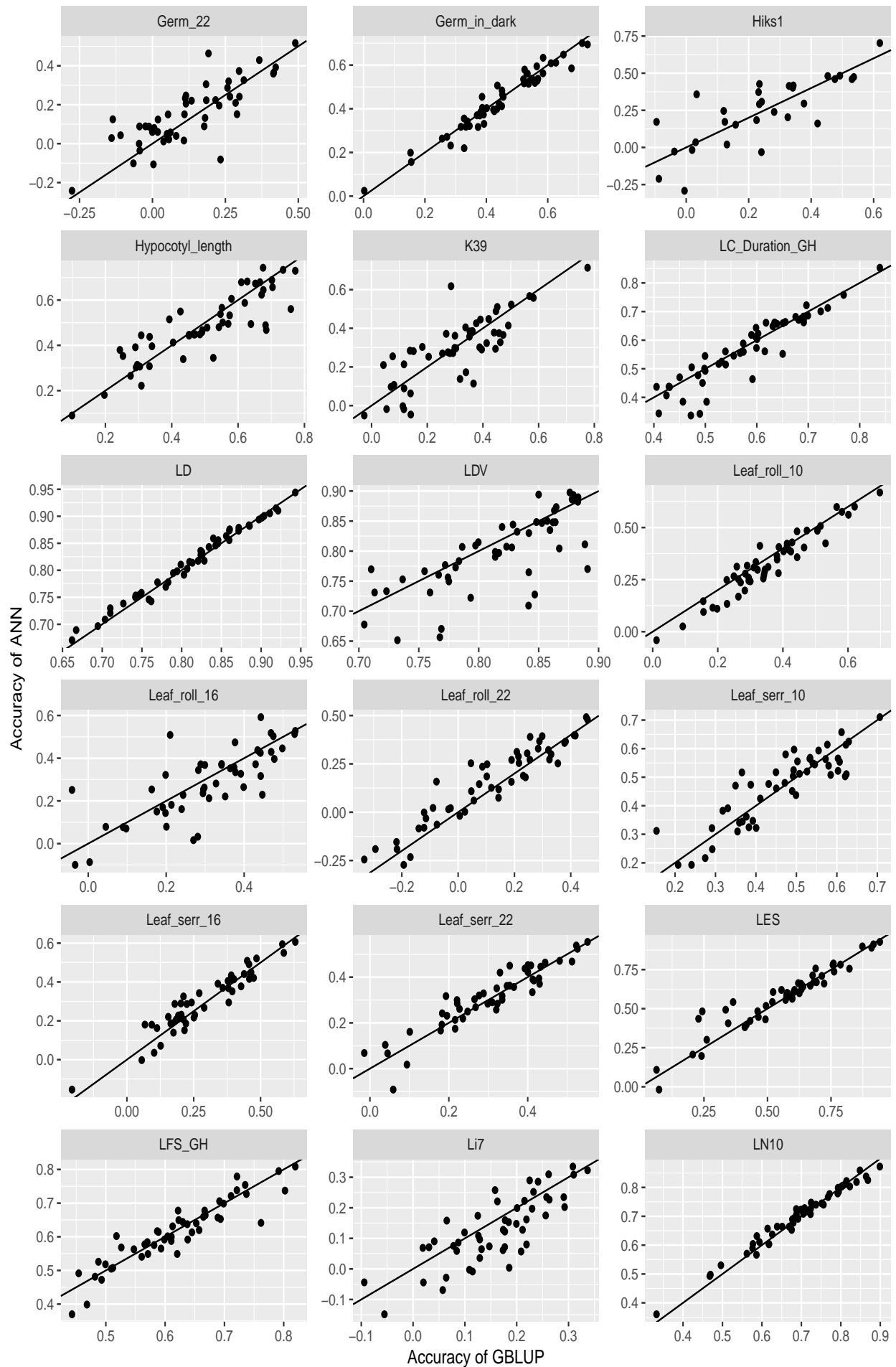


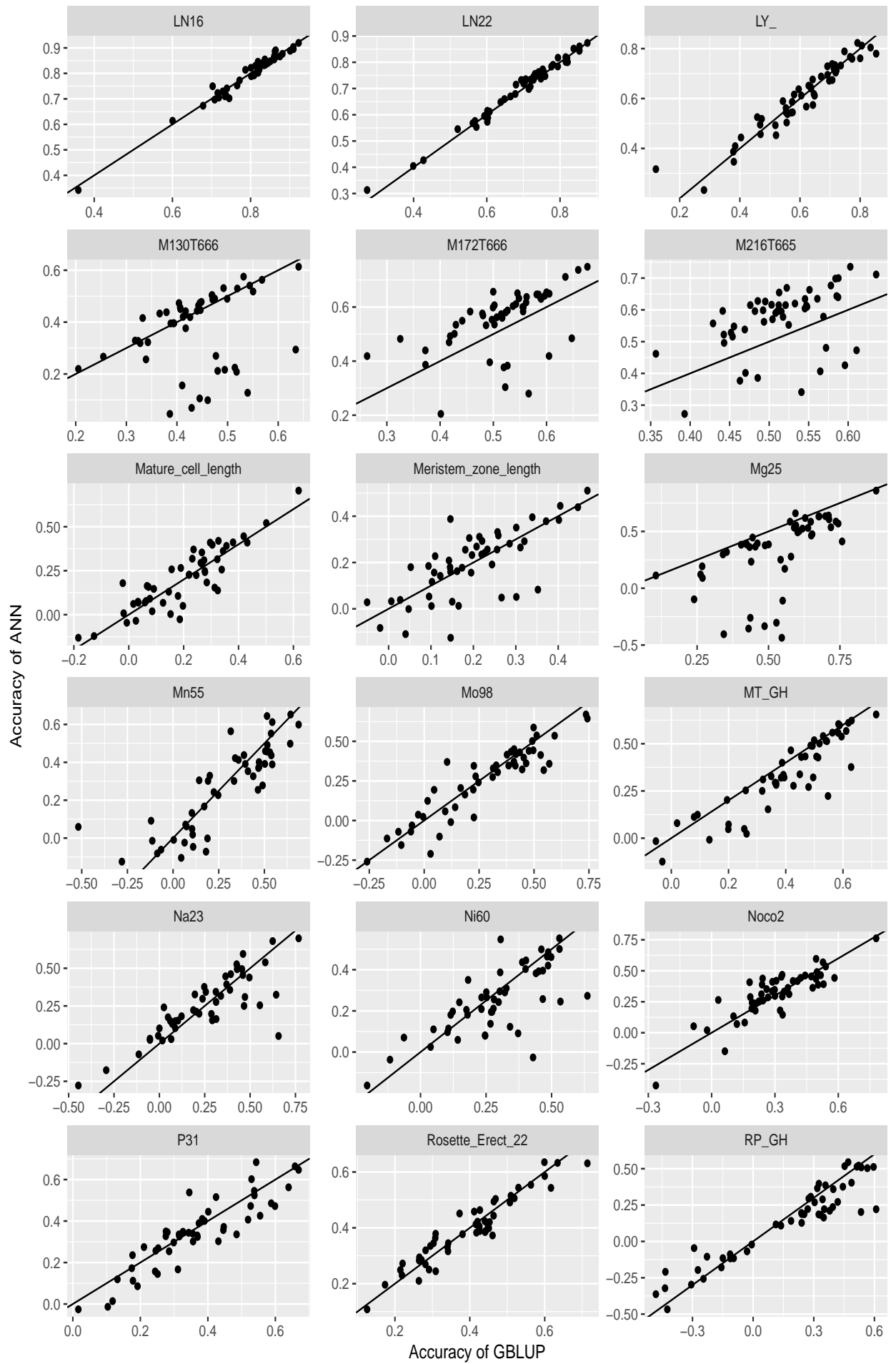


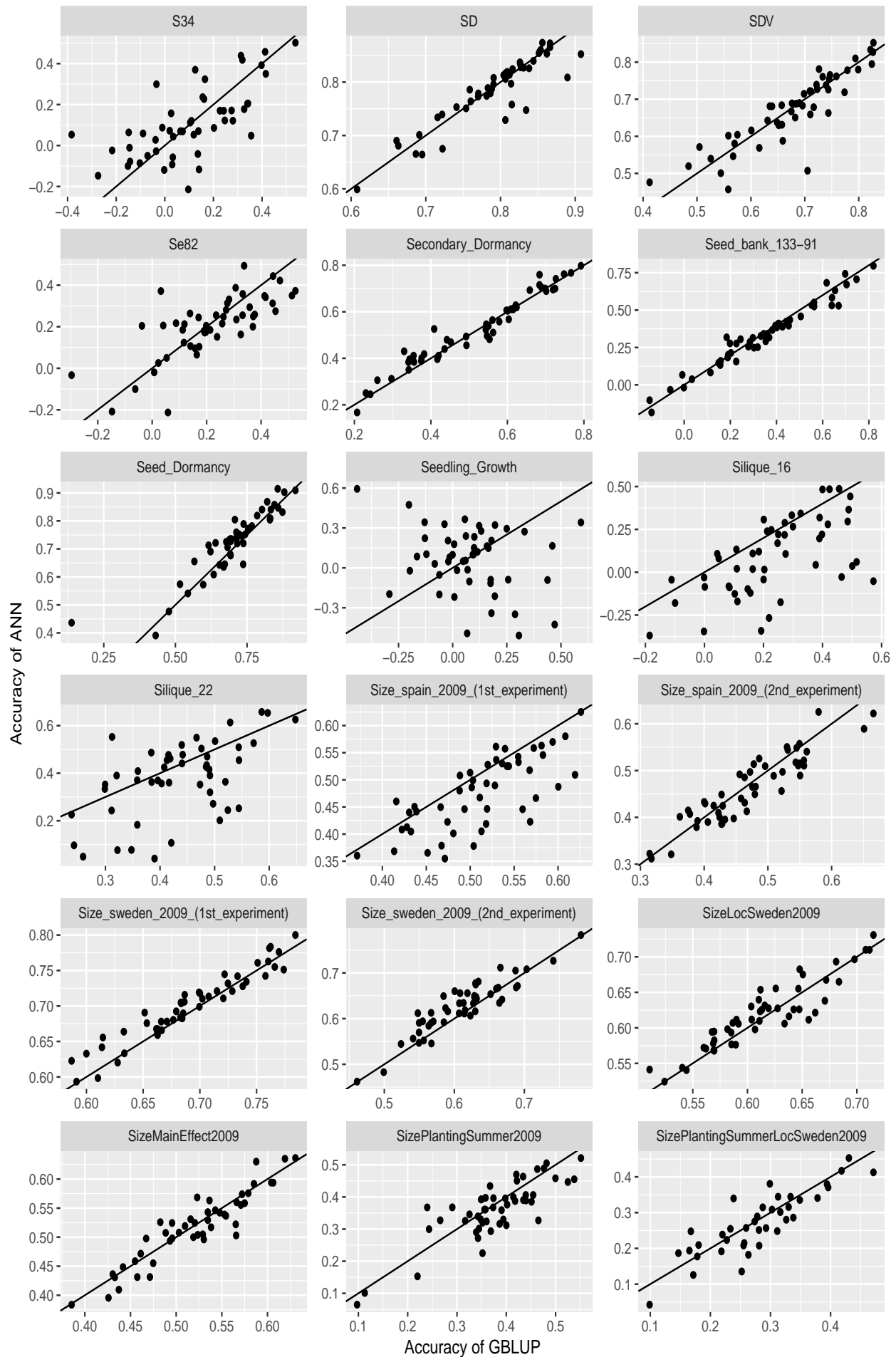


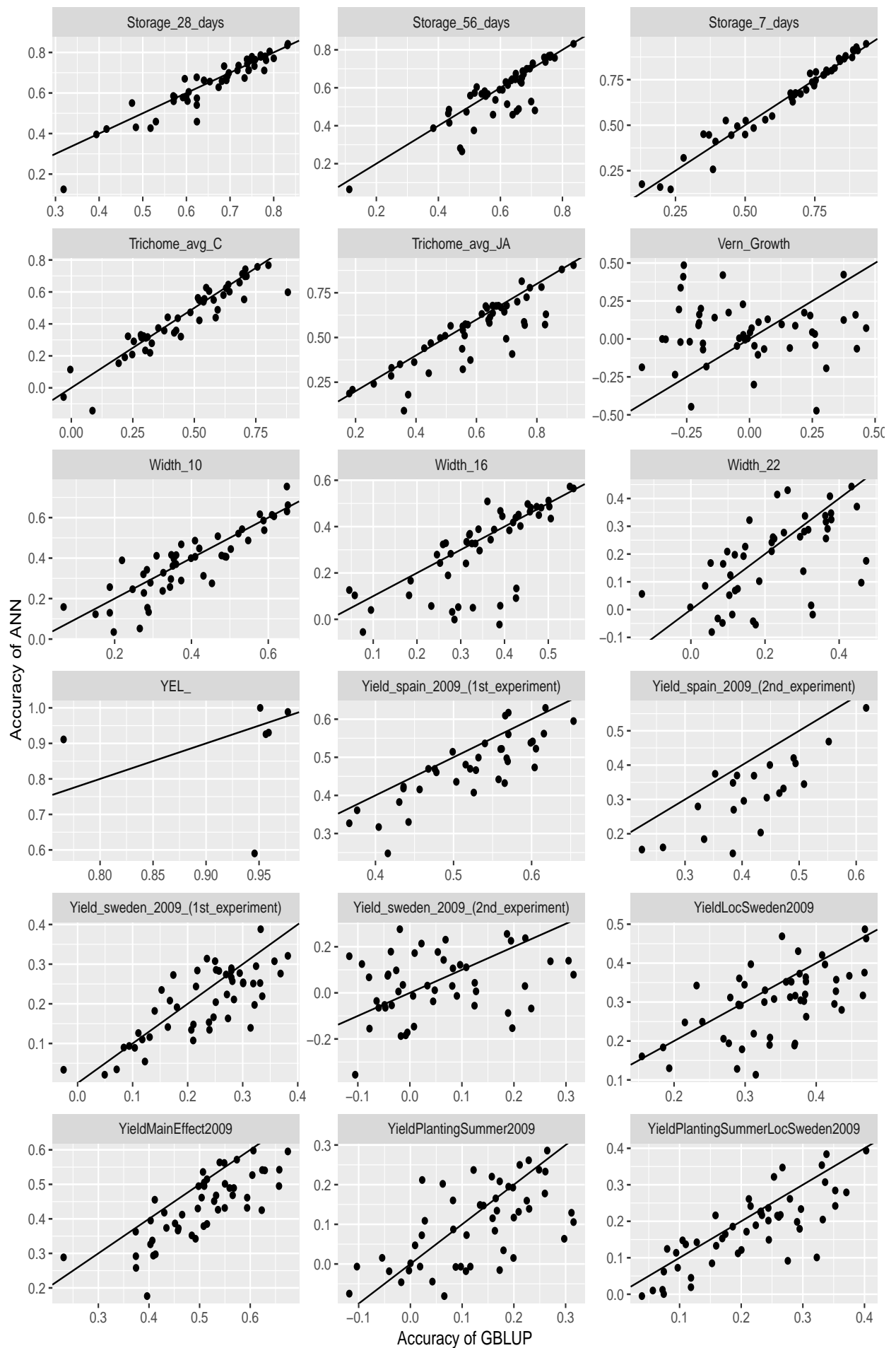


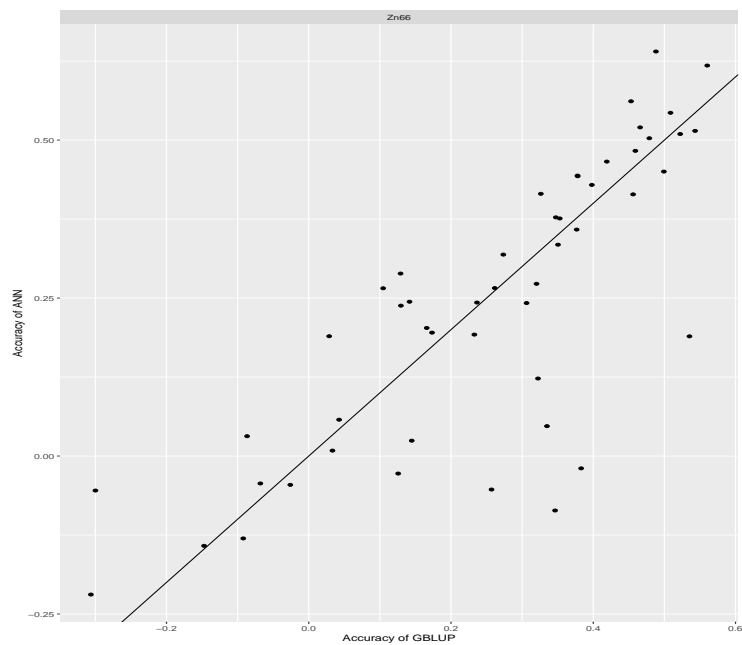












# Bibliography

- Abadi, Martín et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: <https://www.tensorflow.org/>.
- Abadi, Martín et al. (2016). “TensorFlow: A System for Large-scale Machine Learning”. In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. OSDI’16. Savannah, GA, USA: USENIX Association, pp. 265–283. ISBN: 978-1-931971-33-1. URL: <http://dl.acm.org/citation.cfm?id=3026877.3026899>.
- Albrecht, Theresa et al. (2011). “Genome-based prediction of testcross values in maize”. In: *Theoretical and Applied Genetics* 123.2, p. 339.
- Allier, Antoine et al. (2019). “Usefulness Criterion and post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression”. In: *G3: Genes, Genomes, Genetics* 9.5, pp. 1469–1479.
- Almeida Filho, Janeo Eustáquio de et al. (2019). “Genomic Prediction of Additive and Non-additive Effects Using Genetic Markers and Pedigrees”. In: *G3: Genes, Genomes, Genetics* 9.8, pp. 2739–2748. DOI: [10.1534/g3.119.201004](https://doi.org/10.1534/g3.119.201004). URL: <https://doi.org/10.1534>.
- Alonso-Blanco, Carlos et al. (2016). “1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*”. In: *Cell* 166.2, pp. 481–491.
- Amin, Najaf, Cornelia M Van Duijn, and Yurii S Aulchenko (2007). “A genomic background based method for association analysis in related individuals”. In: *PloS one* 2.12, e1274.
- Angermueller, Christof et al. (2016). “Deep learning for computational biology”. In: *Molecular systems biology* 12.7, p. 878.
- Ankenbrand, Markus et al. (Jan. 2018). “chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data”. In: *The Journal of*

- Open Source Software* 3.21, p. 464. ISSN: 2475-9066. DOI: [10.21105/joss.00464](https://doi.org/10.21105/joss.00464). URL: <http://joss.theoj.org/papers/10.21105/joss.00464>.
- Ankenbrand, Markus J. and Frank Förster (Apr. 2019). *Simulated Arabidopsis thaliana sequencing datasets for chloroplast assembler benchmarking*. DOI: [10.5281/zenodo.2622875](https://doi.org/10.5281/zenodo.2622875). URL: <https://doi.org/10.5281/zenodo.2622875>.
- Ankenbrand, Markus J. et al. (June 2017). "AliTV—interactive visualization of whole genome comparisons". In: *PeerJ Computer Science* 3, e116. ISSN: 2376-5992. DOI: [10.7717/peerj-cs.116](https://doi.org/10.7717/peerj-cs.116). URL: <https://doi.org/10.7717/peerj-cs.116>.
- Annicchiarico, Paolo et al. (2015). "Accuracy of genomic selection for alfalfa biomass yield in different reference populations". In: *BMC genomics* 16.1, p. 1020.
- Archibald, John M (2015). "Endosymbiosis and eukaryotic cell evolution". In: *Current Biology* 25.19, R911–R921.
- Arteaga, María Clara et al. (2016). "Genomic variation in recently collected maize landraces from Mexico". In: *Genomics Data* 7, pp. 38–45.
- Atwell, S et al. (2010). "Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines". In: *Nature* 465(7298). DOI: [10.1038/nature08800](https://doi.org/10.1038/nature08800).
- Auinger, Hans-Jürgen et al. (2016). "Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.)" In: *Theoretical and Applied Genetics* 129.11, pp. 2043–2053.
- Azodi, Christina B et al. (2019). "Benchmarking algorithms for genomic prediction of complex traits". In: *bioRxiv*, p. 614479.
- Bakker, Freek T. et al. (Jan. 2016). "Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline". en. In: *Biological Journal of the Linnean Society* 117.1, pp. 33–43. ISSN: 00244066. DOI: [10.1111/bij.12642](https://doi.org/10.1111/bij.12642). URL: <https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/bij.12642>.
- Belamkar, Vikas et al. (2018). "Genomic Selection in Preliminary Yield Trials in a Winter Wheat Breeding Program". In: *G3: Genes, Genomes, Genetics* 8.8, pp. 2735–2747. DOI: [10.1534/g3.118.200415](https://doi.org/10.1534/g3.118.200415). URL: <https://doi.org/10.1534>.
- Bendich, Arnold J. (1987). "Why do chloroplasts and mitochondria contain so many copies of their genome?" In: *BioEssays* 6.6, pp. 279–282. DOI: [10.1002/bies.950060608](https://doi.org/10.1002/bies.950060608). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.950060608>.



- 1002/bies.950060608. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.950060608>.
- Berardini, Tanya Z. et al. (2015). "The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome". In: *genesis* 53.8, pp. 474–485. DOI: 10.1002/dvg.22877. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/dvg.22877>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/dvg.22877>.
- Bergstra, James S et al. (2011). "Algorithms for hyper-parameter optimization". In: *Advances in neural information processing systems*, pp. 2546–2554.
- Bernal-Vasquez, Angela-Maria et al. (2014). "The importance of phenotypic data analysis for genomic prediction-a case study comparing different spatial models in rye". In: *BMC genomics* 15.1, p. 646.
- Bernal-Vasquez, Angela-Maria et al. (2017). "Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program". In: *BMC genetics* 18.1, p. 51.
- Bernardo, R (2010). *Breeding for quantitative traits in plants*. Tech. rep. Stemma Press.
- Bernardo, Rex and Jianming Yu (2007). "Prospects for genomewide selection for quantitative traits in maize". In: *Crop Science* 47.3, pp. 1082–1090.
- Biazzi, Elisa et al. (2017). "Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits". In: *PLoS One* 12.1, e0169234.
- Biscarini, Filippo et al. (2014). "Genome-enabled predictions for binomial traits in sugar beet populations". In: *BMC genetics* 15.1, p. 87.
- Bock, Ralph (2017). "Witnessing genome evolution: experimental reconstruction of endosymbiotic and horizontal gene transfer". In: *Annual review of genetics* 51, pp. 1–22.
- Bottou, Léon (1991). "Stochastic gradient learning in neural networks". In: *Proceedings of Neuro-Nimes* 91.8, p. 12.
- Bottou, Léon and Olivier Bousquet (2008). "The Tradeoffs of Large Scale Learning". In: *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. Ed. by J.C. Platt et al. NIPS Foundation (<http://books.nips.cc>), pp. 161–168. URL: <http://leon.bottou.org/papers/bottou-bousquet-2008>.

- Boyle, Evan A, Yang I Li, and Jonathan K Pritchard (2017). "An expanded view of complex traits: from polygenic to omnigenic". In: *Cell* 169.7, pp. 1177–1186.
- Brauner, Pedro C et al. (2018). "Genomic prediction within and among doubled-haploid libraries from maize landraces". In: *Genetics* 210.4, pp. 1185–1196.
- Brauner, Pedro C et al. (2019). "Testcross performance of doubled haploid lines from European flint maize landraces is promising for broadening the genetic base of elite germplasm". In: *Theoretical and Applied Genetics* 132.6, pp. 1897–1908.
- Brooker, Robert J (1999). *Genetics: analysis & principles*. Addison-Wesley Reading, MA.
- Browning, Brian L, Ying Zhou, and Sharon R Browning (2018). "A one-penny imputed genome from next-generation reference panels". In: *The American Journal of Human Genetics* 103.3, pp. 338–348.
- Browning, Sharon R and Brian L Browning (2007). "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering". In: *The American Journal of Human Genetics* 81.5, pp. 1084–1097.
- Bustos-Korts, Daniela et al. (2016a). "Improvement of predictive ability by uniform coverage of the target genetic space". In: *G3: Genes, Genomes, Genetics* 6.11, pp. 3733–3747.
- (2016b). "Improvement of Predictive Ability by Uniform Coverage of the Target Genetic Space". In: *G3: Genes, Genomes, Genetics* 6.11, pp. 3733–3747. DOI: [10.1534/g3.116.035410](https://doi.org/10.1534/g3.116.035410). URL: <https://doi.org/10.1534>.
- Campos, Gustavo De los and Paulino Perez Rodriguez (2016). *BGLR: Bayesian Generalized Linear Regression*. R package version 1.0.5. URL: <https://CRAN.R-project.org/package=BGLR>.
- Chan, Cheong Xin and Mark A. Ragan (2013). "Next-generation phylogenomics". In: *Biology direct* 8, pp. 3–3. ISSN: 1745-6150. DOI: [10.1186/1745-6150-8-3](https://doi.org/10.1186/1745-6150-8-3). URL: <https://www.ncbi.nlm.nih.gov/pubmed/23339707>.
- Chat, J. et al. (July 2002). "A Case of Chloroplast Heteroplasmy in Kiwifruit (*Actinidia deliciosa*) That Is Not Transmitted During Sexual Reproduction". In: *Journal of Heredity* 93.4, pp. 293–300. ISSN: 0022-1503. DOI: [10.1093/jhered/](https://doi.org/10.1093/jhered/)

- 93.4.293. eprint: <http://oup.prod.sis.lan/jhered/article-pdf/93/4/293/6454216/293.pdf>. URL: <https://doi.org/10.1093/jhered/93.4.293>.
- Che, Ronglin et al. (2014). "An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use". In: *BioData mining* 7.1, p. 9.
- Chollet, François et al. (2015). Keras. <https://keras.io>.
- Coissac, Eric et al. (2016). "From barcodes to genomes: extending the concept of DNA barcoding". In: *Molecular Ecology* 25.7, pp. 1423–1428. ISSN: 1365-294X. DOI: [10.1111/mec.13549](https://doi.org/10.1111/mec.13549). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13549> (visited on 05/16/2019).
- Collette, Andrew (2013). *Python and HDF5*. O'Reilly.
- Corriveau, Joseph L. and Annette W. Coleman (1988). "Rapid Screening Method to Detect Potential Biparental Inheritance of Plastid DNA and Results for Over 200 Angiosperm Species". In: *American Journal of Botany* 75.10, pp. 1443–1458. ISSN: 00029122, 15372197. URL: <http://www.jstor.org/stable/2444695>.
- Crossa, José et al. (2010). "Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers". In: *Genetics*.
- Crossa, José et al. (2016). "Genomic Prediction of Gene Bank Wheat Landraces". In: *G3: Genes, Genomes, Genetics* 6.7, pp. 1819–1834. DOI: [10.1534/g3.116.029637](https://doi.org/10.1534/g3.116.029637). URL: <https://doi.org/10.1534>.
- Crossa, José et al. (2017). "Genomic selection in plant breeding: Methods, models, and perspectives". In: *Trends in plant science*.
- Cuevas, Jaime et al. (2019a). "Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials". In: *G3: Genes, Genomes, Genetics* 9.9, pp. 2913–2924. DOI: [10.1534/g3.119.400493](https://doi.org/10.1534/g3.119.400493). URL: <https://doi.org/10.1534>.
- Cuevas, Jaime et al. (2019b). "Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials". In: *G3: Genes, Genomes, Genetics* 9.9, pp. 2913–2924.

- Daniell, Henry et al. (June 23, 2016). "Chloroplast genomes: diversity, evolution, and applications in genetic engineering". In: *Genome Biology* 17. ISSN: 1474-7596. DOI: [10.1186/s13059-016-1004-2](https://doi.org/10.1186/s13059-016-1004-2). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4918201/> (visited on 05/20/2019).
- Darwin, Charles (1859). *On the Origin of Species by Means of Natural Selection. or the Preservation of Favored Races in the Struggle for Life*. London: Murray.
- De Los Campos, Gustavo et al. (2009). "Predicting quantitative traits with regression models for dense molecular markers and pedigree". In: *Genetics* 182.1, pp. 375–385.
- De Rubeis, Silvia et al. (2014). "Synaptic, transcriptional and chromatin genes disrupted in autism". In: *Nature* 515.7526, p. 209.
- Deiner, Kristy et al. (2017). "Environmental DNA metabarcoding: Transforming how we survey animal and plant communities". In: *Molecular Ecology* 26.21, pp. 5872–5895. ISSN: 1365-294X. DOI: [10.1111/mec.14350](https://doi.org/10.1111/mec.14350). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14350> (visited on 05/21/2019).
- Desta, Zeratsion Abera and Rodomiro Ortiz (2014). "Genomic selection: genome-wide prediction in plant improvement". In: *Trends in plant science* 19.9, pp. 592–601.
- Dierckxsens, Nicolas, Patrick Mardulyn, and Guillaume Smits (Feb. 28, 2017). "NOVO-Plasty: de novo assembly of organelle genomes from whole genome data". In: *Nucleic Acids Research* 45.4, e18–e18. ISSN: 0305-1048. DOI: [10.1093/nar/gkw955](https://doi.org/10.1093/nar/gkw955). URL: <https://academic.oup.com/nar/article/45/4/e18/2290925> (visited on 01/26/2018).
- Docker Hub Group for Benchmark Project. URL: <https://cloud.docker.com/u/chloroextractorteam/>.
- Dos Santos, Cicero and Maira Gatti (2014). "Deep convolutional neural networks for sentiment analysis of short texts". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78.
- Dozat, Timothy (2016). "Incorporating nesterov momentum into adam". In:

- El-Dien, Omnia Gamal et al. (2016). "Implementation of the Realized Genomic Relationship Matrix to Open-Pollinated White Spruce Family Testing for Disentangling Additive from Nonadditive Genetic Effects". In: *G3: Genes, Genomes, Genetics* 6.3, pp. 743–753. DOI: [10.1534/g3.115.025957](https://doi.org/10.1534/g3.115.025957). URL: <https://doi.org/10.1534>.
- Elias, Ani A et al. (2018a). "Improving genomic prediction in cassava field experiments by accounting for interplot competition". In: *G3: Genes, Genomes, Genetics* 8.3, pp. 933–944.
- (2018b). "Improving genomic prediction in cassava field experiments using spatial analysis". In: *G3: Genes, Genomes, Genetics* 8.1, pp. 53–62.
- Enciso-Rodriguez, Felix et al. (2018). "Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*)". In: *G3: Genes, Genomes, Genetics* 8.7, pp. 2471–2481.
- Endelman, Jeffrey B. et al. (Mar. 2018). "Genetic Variance Partitioning and Genome-Wide Prediction with Allele Dosage Information in Autotetraploid Potato". In: *Genetics* 209.1, pp. 77–87. DOI: [10.1534/genetics.118.300685](https://doi.org/10.1534/genetics.118.300685). URL: <https://doi.org/10.1534/genetics.118.300685>.
- Falconer, DS and TFC Mackay (1996). "Introduction to quantitative genetics. 1996". In: Harlow, Essex, UK: Longmans Green 3.
- Fan, Jianqing, Fang Han, and Han Liu (2014). "Challenges of big data analysis". In: *National science review* 1.2, pp. 293–314.
- Feldman, Moshe and Avraham A Levy (2012). "Genome evolution due to allopolyploidization in wheat". In: *Genetics* 192.3, pp. 763–774.
- Fisher, Ronald A (1919). "XV.—The correlation between relatives on the supposition of Mendelian inheritance." In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52.2, pp. 399–433.
- Forsberg, Simon K. G. et al. (2015). "The Multi-allelic Genetic Architecture of a Variance-Heterogeneity Locus for Molybdenum Concentration in Leaves Acts as a Source of Unexplained Additive Genetic Variance". In: *PLOS Genetics* None. DOI: [10.1371/journal.pgen.1005648](https://doi.org/10.1371/journal.pgen.1005648).

- Förster, Frank and Markus J. Ankenbrand (May 2019). *chloroExtractorTeam/benchmark: Benchmark container setup v2.0.1*. DOI: [10.5281/zenodo.2628061](https://doi.org/10.5281/zenodo.2628061). URL: <https://doi.org/10.5281/zenodo.2628061>.
- Freudenthal, Jan A. et al. (2019a). "GWAS-Flow: A GPU accelerated framework for efficient permutation based genome-wide association studies". In: DOI: [10.1101/783100](https://doi.org/10.1101/783100). URL: <https://doi.org/10.1101>.
- Freudenthal, Jan A et al. (2019b). "The landscape of chloroplast genome assembly tools". In: *bioRxiv*, p. 665869.
- Friebe, B et al. (2000). "Development of a complete set of *Triticum aestivum*-*Aegilops speltoides* chromosome addition lines". In: *Theoretical and Applied Genetics* 101.1-2, pp. 51–58.
- Fuentes, Ignacia et al. (2014). "Horizontal genome transfer as an asexual path to the formation of new species". In: *Nature* 511.7508, p. 232.
- Gapare, Washington et al. (2018). "Historical Datasets Support Genomic Selection Models for the Prediction of Cotton Fiber Quality Phenotypes Across Multiple Environments". In: *G3: Genes, Genomes, Genetics* 8.5, pp. 1721–1732.
- Ghosh, Sreya et al. (2018). "Speed breeding in growth chambers and glasshouses for crop breeding and model plant research". In: *Nature protocols* 13.12, p. 2944.
- Gianola, Daniel (2013). "Priors in whole-genome regression: the Bayesian alphabet returns". In: *Genetics* 194.3, pp. 573–596.
- Gianola, Daniel and Johannes BCHM van Kaam (2008). "Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits". In: *Genetics* 178.4, pp. 2289–2303.
- Gianola, Daniel and Guilherme JM Rosa (2015). "One hundred years of statistical developments in animal breeding". In: *Annu. Rev. Anim. Biosci.* 3.1, pp. 19–56.
- Gianola, Daniel et al. (2009). "Additive genetic variability and the Bayesian alphabet". In: *Genetics* 183.1, pp. 347–363.
- Gianola, Daniel et al. (2016). "Genome-Wide Association Studies with a Genomic Relationship Matrix: A Case Study with Wheat and Arabidopsis". In: *G3: Genes, Genomes, Genetics* 6.10, pp. 3241–3256. DOI: [10.1534/g3.116.034256](https://doi.org/10.1534/g3.116.034256). URL: <https://doi.org/10.1534>.

- GitHub Repository for Benchmark Project. URL: <https://github.com/chloroExtractorTeam/benchmark>.
- Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323.
- Goddard, Michael E and Ben J Hayes (2009). "Mapping genes for complex traits in domestic animals and their use in breeding programmes". In: *Nature Reviews Genetics* 10.6, p. 381.
- Goddard, Michael E, Ben J Hayes, and Theo HE Meuwissen (2011). "Using the genomic relationship matrix to predict the accuracy of genomic selection". In: *Journal of animal breeding and genetics* 128.6, pp. 409–421.
- González-Camacho, JM et al. (2012). "Genome-enabled prediction of genetic values using radial basis function neural networks". In: *Theoretical and Applied Genetics* 125.4, pp. 759–771.
- González-Camacho, Juan Manuel et al. (2016). "Genome-enabled prediction using probabilistic neural network classifiers". In: *BMC genomics* 17.1, p. 208.
- González-Camacho, Juan Manuel et al. (2018). "Applications of machine learning methods to genomic selection in breeding wheat for rust resistance". In: *The plant genome* 11.2.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.
- Gouy, Matthieu et al. (2013). "Experimental assessment of the accuracy of genomic selection in sugarcane". In: *Theoretical and applied genetics* 126.10, pp. 2575–2586.
- Green, Beverley R. (2011). "Chloroplast genomes of photosynthetic eukaryotes". In: *The Plant Journal* 66.1, pp. 34–44. ISSN: 1365-313X. DOI: [10.1111/j.1365-313X.2011.04541.x](https://doi.org/10.1111/j.1365-313X.2011.04541.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2011.04541.x> (visited on 05/16/2019).
- Grenier, Cécile et al. (2015). "Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding". In: *PloS one* 10.8, e0136594.



- Grinberg, Nastasiya F, Oghenejokpeme I Orhobor, and Ross D King (2018). "An Evaluation of Machine-learning for Predicting Phenotype: Studies in Yeast, Rice and Wheat". In: *BioRxiv*, p. 105528.
- Guo, Zhigang et al. (2013). "Accuracy of across-environment genome-wide prediction in maize nested association mapping populations". In: *G3: Genes, Genomes, Genetics* 3.2, pp. 263–272.
- Habier, David, Rohan L Fernando, and Jack CM Dekkers (2007). "The impact of genetic relationship information on genome-assisted breeding values". In: *Genetics* 177.4, pp. 2389–2397.
- Habier, David et al. (2011). "Extension of the Bayesian alphabet for genomic selection". In: *BMC bioinformatics* 12.1, p. 186.
- Hahnloser, Richard HR et al. (2000). "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit". In: *Nature* 405.6789, p. 947.
- Harris, BL, DL Johnson, RJ Spelman, et al. (2009). "Genomic selection in New Zealand and the implications for national genetic evaluation." In: *ICAR Technical Series* 13, pp. 325–330.
- Hassen, Manel Ben et al. (May 2018). "Genomic Prediction Accounting for Genotype by Environment Interaction Offers an Effective Framework for Breeding Simultaneously for Adaptation to an Abiotic Stress and Performance Under Normal Cropping Conditions in Rice". In: *G3: Genes, Genomes, Genetics* 8.7, pp. 2319–2332. DOI: [10.1534/g3.118.200098](https://doi.org/10.1534/g3.118.200098). URL: <https://doi.org/10.1534/g3.118.200098>.
- Hawkins, Charles and Long-Xi Yu (2018). "Recent progress in alfalfa (*Medicago sativa* L.) genomics and genomic selection". In: *The Crop Journal* 6.6, pp. 565–575.
- Hayes, Ben and Mike Goddard (2010). "Genome-wide association and genomic selection in animal breeding". In: *Genome* 53.11, pp. 876–883.
- Hayes, BJ, ME Goddard, et al. (2001). "Prediction of total genetic value using genome-wide dense marker maps". In: *Genetics* 157.4, pp. 1819–1829.
- He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.



- Heffner, Elliot L, Jean-Luc Jannink, and Mark E Sorrells (2011). "Genomic selection accuracy using multifamily prediction models in a wheat breeding program". In: *The Plant Genome* 4.1, pp. 65–75.
- Heffner, Elliot L et al. (2010). "Plant breeding with genomic selection: gain per unit time and cost". In: *Crop science* 50.5, pp. 1681–1690.
- Henderson, Charles R (1975). "Best linear unbiased estimation and prediction under a selection model". In: *Biometrics*, pp. 423–447.
- Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky (2012). "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent". In: *Cited on* 14, p. 8.
- Hirschhorn, Joel N. and Mark J. Daly (2005). "Genome-wide association studies for common diseases and complex traits". In: *Nature Reviews Genetics* 6.2, pp. 95–108. ISSN: 1471-0064. DOI: [10.1038/nrg1521](https://doi.org/10.1038/nrg1521). URL: <https://doi.org/10.1038/nrg1521>.
- Hölker, Armin C et al. (2019). "European maize landraces made accessible for plant breeding and genome-based studies". In: *Theoretical and Applied Genetics*, pp. 1–13.
- Holliday, Jason A., Tongli Wang, and Sally Aitken (2012). "Predicting Adaptive Phenotypes From Multilocus Genotypes in Sitka Spruce (*Picea sitchensis*) Using Random Forest". In: *G3: Genes, Genomes, Genetics* 2.9, pp. 1085–1093. DOI: [10.1534/g3.112.002733](https://doi.org/10.1534/g3.112.002733). URL: <https://doi.org/10.1534>.
- Howard, Réka et al. (2019). "Joint Use of Genome, Pedigree, and Their Interaction with Environment for Predicting the Performance of Wheat Lines in New Environments". In: *G3: Genes, Genomes, Genetics* 9.9, pp. 2925–2934. DOI: [10.1534/g3.119.400508](https://doi.org/10.1534/g3.119.400508). URL: <https://doi.org/10.1534>.
- Hu, Yaodong et al. (2015). "Prediction of plant height in *Arabidopsis thaliana* using DNA methylation data". In: *Genetics* 201.2, pp. 779–793.
- Isik, F (2013). *Genomic Relationships and GBLUP*.
- Jan, Habib U et al. (2016). "Genomic prediction of testcross performance in canola (*Brassica napus*)". In: *PLoS One* 11.1, e0147769.
- Janocha, Katarzyna and Wojciech Marian Czarnecki (2017). "On loss functions for deep neural networks in classification". In: *arXiv preprint arXiv:1702.05659*.

- Jaramillo-Correa, Juan-Pablo et al. (2014). "Molecular Proxies for Climate Maladaptation in a Long-Lived Tree (*Pinus pinaster* Aiton, Pinaceae)". In: *Genetics* 199.3, pp. 793–807. DOI: [10.1534/genetics.114.173252](https://doi.org/10.1534/genetics.114.173252). URL: <https://doi.org/10.1534>.
- Jarquín, Diego, James Specht, and Aaron Lorenz (2016). "Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions". In: *G3: Genes, Genomes, Genetics* 6.8, pp. 2329–2341. DOI: [10.1534/g3.116.031443](https://doi.org/10.1534/g3.116.031443). URL: <https://doi.org/10.1534>.
- Jette, Morris A., Andy B. Yoo, and Mark Grondona (2002). "SLURM: Simple Linux Utility for Resource Management". In: *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003*. Springer-Verlag, pp. 44–60.
- Jiang, Yong and Jochen C Reif (2015). "Modeling epistasis in genomic selection". In: *Genetics* 201.2, pp. 759–768.
- Jin, Jian-Jun et al. (Mar. 2018). "GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data". In: *bioRxiv*. DOI: [10.1101/256479](https://doi.org/10.1101/256479). URL: <http://biorxiv.org/lookup/doi/10.1101/256479>.
- Kadam, Dnyaneshwar C et al. (2016). "Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline". In: *G3: Genes, Genomes, Genetics* 6.11, pp. 3443–3453.
- Kainer, David et al. (2018). "Accuracy of Genomic Prediction for Foliar Terpene Traits in *Eucalyptus polybractea*". In: *G3: Genes, Genomes, Genetics* 8.8, pp. 2573–2583. DOI: [10.1534/g3.118.200443](https://doi.org/10.1534/g3.118.200443). URL: <https://doi.org/10.1534>.
- Kalo, P et al. (2004). "Comparative mapping between *Medicago sativa* and *Pisum sativum*". In: *Molecular Genetics and Genomics* 272.3, pp. 235–246.
- Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". In: *Nature genetics* 42.4, p. 348.
- Kärkkäinen, Hanni P and Mikko J Sillanpää (2012). "Back to basics for Bayesian model building in genomic selection". In: *Genetics* 191.3, pp. 969–987.

- Kim, Sung et al. (2007). "Recombination and linkage disequilibrium in *Arabidopsis thaliana*". In: *Nature genetics* 39.9, p. 1151.
- Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.
- Kingsolver, Joel G et al. (2001). "The strength of phenotypic selection in natural populations". In: *The American Naturalist* 157.3, pp. 245–261.
- Koch, Marcus A and Michaela Matschinger (2007). "Evolution and genetic differentiation among relatives of *Arabidopsis thaliana*". In: *Proceedings of the National Academy of Sciences* 104.15, pp. 6272–6277.
- Korte, Arthur and Ashley Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1, p. 29.
- Korte, Arthur et al. (2012). "A mixed-model approach for genome-wide association studies of correlated traits in structured populations". In: *Nature genetics* 44.9, p. 1066.
- Krause, Margaret R. et al. (2019). "Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat". In: *G3: Genes, Genomes, Genetics*, g3.200856.2018. DOI: [10.1534/g3.118.200856](https://doi.org/10.1534/g3.118.200856). URL: <https://doi.org/10.1534/g3.118.200856>.
- Kumar, Rachana A., Delene J. Oldenburg, and Arnold J. Bendich (Sept. 2014). "Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development". In: *Journal of Experimental Botany* 65.22, pp. 6425–6439. ISSN: 0022-0957. DOI: [10.1093/jxb/eru359](https://doi.org/10.1093/jxb/eru359). eprint: <http://oup.prod.sis.lan/jxb/article-pdf/65/22/6425/16935653/eru359.pdf>. URL: <https://doi.org/10.1093/jxb/eru359>.
- Kumar, Satish et al. (2015). "Genome-Enabled Estimates of Additive and Nonadditive Genetic Variances and Prediction of Apple Phenotypes Across Environments". In: *G3: Genes, Genomes, Genetics* 5.12, pp. 2711–2718. DOI: [10.1534/g3.115.021105](https://doi.org/10.1534/g3.115.021105). URL: <https://doi.org/10.1534/g3.115.021105>.
- Kurtzer, Gregory M, Vanessa Sochat, and Michael W Bauer (2017). "Singularity: Scientific containers for mobility of compute". In: *PloS one* 12.5, e0177459.
- Kutschera, Ulrich and Karl J Niklas (2005). "Endosymbiosis, cell evolution, and speciation". In: *Theory in Biosciences* 124.1, pp. 1–24.

- Lan, Kun et al. (2018). "A survey of data mining and deep learning in bioinformatics". In: *Journal of medical systems* 42.8, p. 139.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, p. 436.
- LeCun, Yann et al. (1999). "Object recognition with gradient-based learning". In: *Shape, contour and grouping in computer vision*. Springer, pp. 319–345.
- Legarra, Andres, Danila A.L Lourenco, and Zulma G. Vitezica (2018). "Bases for Genomic Prediction". In:
- Lehermeier, Christina et al. (2014). "Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction". In: *Genetics* 198.1, pp. 3–16.
- Leinonen, Rasko et al. (Nov. 2010). "The Sequence Read Archive". In: *Nucleic Acids Research* 39.suppl\_1, pp. D19–D21. ISSN: 0305-1048. DOI: [10.1093/nar/gkq1019](https://doi.org/10.1093/nar/gkq1019). eprint: [http://oup.prod.sis.lan/nar/article-pdf/39/suppl\\_1/D19/7624335/gkq1019.pdf](http://oup.prod.sis.lan/nar/article-pdf/39/suppl_1/D19/7624335/gkq1019.pdf). URL: <https://doi.org/10.1093/nar/gkq1019>.
- Leutenegger, Anne-Louise et al. (2003). "Estimation of the inbreeding coefficient through use of genomic data". In: *The American Journal of Human Genetics* 73.3, pp. 516–523.
- Li, Bo et al. (2018). "Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods". In: *Frontiers in genetics* 9, p. 237.
- Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18, pp. 3094–3100.
- Li, Jia-Yang, Jun Wang, and Robert S Zeigler (2014). "The 3,000 rice genomes project: new opportunities and challenges for future rice research". In: *GigaScience* 3.1, p. 8.
- Li, Peijin et al. (2014). "Multiple FLC haplotypes defined by independent cis-regulatory variation underpin life history diversity in *Arabidopsis thaliana*". In: *Genes & Development* 28.15, pp. 1635–1640.
- Li, Xuehui and E Charles Brummer (2012). "Applied genetics and genomics in alfalfa breeding". In: *Agronomy* 2.1, pp. 40–61.

- Li, Xuehui et al. (2015). "Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population". In: *The Plant Genome* 8.2.
- Li, Yan et al. (2010). "Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*". In: *PNAS* 107. DOI: [10.1073/pnas.1007431107](https://doi.org/10.1073/pnas.1007431107).
- Lippert, Christoph et al. (2014). "LIMIX: genetic analysis of multiple traits". In: *bioRxiv*. DOI: [10.1101/003905](https://doi.org/10.1101/003905). eprint: <https://www.biorxiv.org/content/early/2014/05/22/003905.full.pdf>. URL: <https://www.biorxiv.org/content/early/2014/05/22/003905>.
- Litjens, Geert et al. (2017). "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42, pp. 60–88.
- Lopez-Cruz, Marco et al. (2015). "Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model". In: *G3: Genes, Genomes, Genetics* 5.4, pp. 569–582. DOI: [10.1534/g3.114.016097](https://doi.org/10.1534/g3.114.016097). URL: <https://doi.org/10.1534>.
- Luo, Xiang et al. (2017). "Genomic prediction of genotypic effects with epistasis and environment interactions for yield-related traits of rapeseed (*Brassica napus* L.)" In: *Frontiers in genetics* 8, p. 15.
- Lynch, Michael, Bruce Walsh, et al. (1998). *Genetics and analysis of quantitative traits*. Vol. 1. Sinauer Sunderland, MA.
- Ma, Wenlong et al. (2017). "DeepGS: Predicting phenotypes from genotypes using Deep Learning". In: *bioRxiv*, p. 241414.
- Mamoshina, Polina et al. (2016). "Applications of deep learning in biomedicine". In: *Molecular pharmaceuticals* 13.5, pp. 1445–1454.
- Martin, William et al. (Sept. 17, 2002). "Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.19, pp. 12246–12251. ISSN: 0027-8424. DOI: [10.1073/pnas.182432999](https://doi.org/10.1073/pnas.182432999). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC129430/> (visited on 05/20/2019).
- Martini, Johannes WR et al. (2017). "Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and

- properties of the categorical epistasis model (CE)". In: *BMC bioinformatics* 18.1, p. 3.
- Marulanda, Jose J et al. (2016). "Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale". In: *Theoretical and applied genetics* 129.10, pp. 1901–1913.
- Marvin, Minsky and Papert Seymour (1969). *Perceptrons*.
- Mayer, Manfred et al. (2017). "Is there an optimum level of diversity in utilization of genetic resources?" In: *Theoretical and applied genetics* 130.11, pp. 2283–2295.
- McKain, Michael and Afnit (Sept. 2017). *Mrmckain/Fast-Plast: Fast-Plast V.1.2.6*. DOI: [10.5281/zenodo.973887](https://doi.org/10.5281/zenodo.973887). URL: <https://zenodo.org/record/973887>.
- Mckinney, Brett and Nicholas Pajewski (2012). "Six degrees of epistasis: statistical network models for GWAS". In: *Frontiers in genetics* 2, p. 109.
- McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56.
- Meijó, Mónica et al. (2014). "Genome-wide association study using cellular traits identifies a new regulator of root development in Arabidopsis". In: *Nature Genetics* 46. DOI: [10.1038/ng.2824](https://doi.org/10.1038/ng.2824).
- Mereschkowsky, Constantin (1905). "Über natur und ursprung der chromatophoren im pflanzenreiche". In: *Biologisches Centralblatt* 25, pp. 293–604.
- Merkel, Dirk (2014). "Docker: lightweight linux containers for consistent development and deployment". In: *Linux Journal* 2014.239, p. 2.
- Michie, Donald, David J Spiegelhalter, CC Taylor, et al. (1994). "Machine learning". In: *Neural and Statistical Classification* 13.
- Min, Seonwoo, Byunghan Lee, and Sungroh Yoon (2017). "Deep learning in bioinformatics". In: *Briefings in bioinformatics* 18.5, pp. 851–869.
- Misztal, I et al. (2013). "Methods to approximate reliabilities in single-step genomic evaluation". In: *Journal of Dairy Science* 96.1, pp. 647–654.
- Moeinizada, Saba et al. (2019). "Optimizing Selection and Mating in Genomic Selection with a Look-Ahead Approach: An Operations Research Framework". In: *G3: Genes, Genomes, Genetics*, g3–200842.



- Momen, Mehdi et al. (Aug. 2019). "Predicting Longitudinal Traits Derived from High-Throughput Phenomics in Contrasting Environments Using Genomic Legendre Polynomials and B-Splines". In: *G3: Genes, Genomes, Genetics* 9.10, pp. 3369–3380. DOI: [10.1534/g3.119.400346](https://doi.org/10.1534/g3.119.400346). URL: <https://doi.org/10.1534/g3.119.400346>.
- Montesinos-López, Osval A et al. (2015). "Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding". In: *G3: Genes, Genomes, Genetics* 5.2, pp. 291–300.
- Montesinos-López, Osval A et al. (2019a). "A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding". In: *G3: Genes, Genomes, Genetics* 9.2, pp. 601–618.
- (2019b). "New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes". In: *G3: Genes, Genomes, Genetics* 9.5, pp. 1545–1556.
- Morota, Gota and Daniel Gianola (2014). "Kernel-based whole-genome prediction of complex traits: a review". In: *Frontiers in genetics* 5, p. 363.
- Moser, Gerhard et al. (2009). "A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers". In: *Genetics Selection Evolution* 41.1, p. 56.
- Mousseau, Timothy A and Derek A Roff (1987). "Natural selection and the heritability of fitness components". In: *Heredity* 59.2, p. 181.
- National Center for Biotechnology Information. NCBI Taxonomy. Accessed: 2019-10-01. URL: <https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>.
- Neyhart, Jeffrey, Aaron J Lorenz, and Kevin P Smith (2019). "Multi-Trait Improvement by Predicting Genetic Correlations in Breeding Crosses". In: *bioRxiv*, p. 593210.
- Nguyen, Derrick and Bernard Widrow (1990). "The truck backer-upper: An example of self-learning in neural networks". In: *Advanced neural computers*. Elsevier, pp. 11–19.
- Nordborg, Magnus et al. (2002). "The extent of linkage disequilibrium in *Arabidopsis thaliana*". In: *Nature genetics* 30.2, p. 190.

- Norman, Adam et al. (2018). "Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy". In: *G3: Genes, Genomes, Genetics* 8.9, pp. 2889–2899. DOI: [10.1534/g3.118.200311](https://doi.org/10.1534/g3.118.200311). URL: <https://doi.org/10.1534>.
- Oakey, Helena et al. (2016). "Genomic selection in multi-environment crop trials". In: *G3: Genes, Genomes, Genetics* 6.5, pp. 1313–1326.
- Ogutu, Joseph O, Hans-Peter Piepho, and Torben Schulz-Streeck (2011). "A comparison of random forests, boosting and support vector machines for genomic selection". In: *BMC proceedings*. Vol. 5. 3. BioMed Central, S11.
- Ohyama, Kanji et al. (Aug. 1986). "Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA". In: *Nature* 322.6079, p. 572. ISSN: 1476-4687. DOI: [10.1038/322572a0](https://doi.org/10.1038/322572a0). URL: <https://www.nature.com/articles/322572a0> (visited on 05/20/2019).
- Olejniczak, Szymon Adam et al. (2016). "Chloroplasts: state of research and practical applications of plastome sequencing". In: *Planta* 244.3, pp. 517–527.
- Oliphant, Travis E (2006). *A guide to NumPy*. Vol. 1. Trelgol Publishing USA.
- Ovenden, Ben et al. (2018). "Accounting for Genotype-by-Environment Interactions and Residual Genetic Variation in Genomic Selection for Water-Soluble Carbohydrate Concentration in Wheat". In: *G3: Genes, Genomes, Genetics* 8.6, pp. 1909–1919. DOI: [10.1534/g3.118.200038](https://doi.org/10.1534/g3.118.200038). URL: <https://doi.org/10.1534>.
- Owens, Brenda F et al. (2014). "A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels". In: *Genetics* 198.4, pp. 1699–1716.
- Ozkan, Hakan, Avraham A Levy, and Moshe Feldman (2001). "Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group". In: *The Plant Cell* 13.8, pp. 1735–1747.
- Palmer, Jeffrey D. (1985). "COMPARATIVE ORGANIZATION OF CHLOROPLAST GENOMES". In: *Annual Review of Genetics* 19.1. PMID: 3936406, pp. 325–354. DOI: [10.1146/annurev.ge.19.120185.001545](https://doi.org/10.1146/annurev.ge.19.120185.001545). eprint: <https://doi.org/10.1146/annurev.ge.19.120185.001545>. URL: <https://doi.org/10.1146/annurev.ge.19.120185.001545>.



- Peiffer, Jason A et al. (2014). "The genetic architecture of maize height". In: *Genetics* 196.4, pp. 1337–1356.
- Polyak, Boris T (1964). "Some methods of speeding up the convergence of iteration methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5, pp. 1–17.
- Poudel, Hari P. et al. (2019). "Genomic Prediction for Winter Survival of Lowland Switchgrass in the Northern USA". In: *G3: Genes, Genomes, Genetics*, g3.400094.2019. DOI: [10.1534/g3.119.400094](https://doi.org/10.1534/g3.119.400094). URL: <https://doi.org/10.1534>.
- Purcell, Shaun et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3, pp. 559–575.
- Purcell, Shaun M et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". In: *Nature* 506.7487, p. 185.
- Qian, Lunwen, Wei Qian, and Rod J Snowdon (2014). "Sub-genomic selection patterns as a signature of breeding in the allopolyploid *Brassica napus* genome". In: *BMC genomics* 15.1, p. 1170.
- Qiu, Zhixu et al. (2016). "Application of machine learning-based classification to genomic selection and performance improvement". In: *International Conference on Intelligent Computing*. Springer, pp. 412–421.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rampasek, Ladislav and Anna Goldenberg (2016). "Tensorflow: Biology's gateway to deep learning?" In: *Cell systems* 2.1, pp. 12–14.
- Ramstein, Guillaume P. and Michael D. Casler (2019). "Extensions of BLUP Models for Genomic Prediction in Heterogeneous Populations: Application in a Diverse Switchgrass Sample". In: *G3: Genes, Genomes, Genetics*, g3.200969.2018. DOI: [10.1534/g3.118.200969](https://doi.org/10.1534/g3.118.200969). URL: <https://doi.org/10.1534>.
- Ramstein, Guillaume P. et al. (2016). "Accuracy of Genomic Prediction in Switchgrass (*Panicum virgatum*L.) Improved by Accounting for Linkage Disequilibrium". In: *G3: Genes, Genomes, Genetics* 6.4, pp. 1049–1062. DOI: [10.1534/g3.115.024950](https://doi.org/10.1534/g3.115.024950). URL: <https://doi.org/10.1534>.

- Ratcliffe, Blaise et al. (2017). "Single-Step BLUP with Varying Genotyping Effort in Open-Pollinated *Picea glauca*". In: *G3: Genes, Genomes, Genetics* 7.3, pp. 935–942. DOI: [10.1534/g3.116.037895](https://doi.org/10.1534/g3.116.037895). URL: <https://doi.org/10.1534>.
- Resende, M. F. R. et al. (2012). "Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda* L.)" In: *Genetics* 190.4, pp. 1503–1510. DOI: [10.1534/genetics.111.137026](https://doi.org/10.1534/genetics.111.137026). URL: <https://doi.org/10.1534>.
- Riedelsheimer, Christian et al. (2013). "Genomic predictability of interconnected biparental maize populations". In: *Genetics* 194.2, pp. 493–503.
- Rincent, Renaud et al. (2012). "Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.)" In: *Genetics* 192.2, pp. 715–728.
- Rincent, Renaud et al. (2018). "Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar". In: *G3: Genes, Genomes, Genetics*, g3.200760.2018. DOI: [10.1534/g3.118.200760](https://doi.org/10.1534/g3.118.200760). URL: <https://doi.org/10.1534>.
- Ritchie, Marylyn D and Kristel Van Steen (2018). "The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation". In: *Annals of translational medicine* 6.8.
- Rocaps Lab. CpBase. Accessed: 2019-04-01, Version: 8/20/2017. URL: [http://rocaplab.ocean.washington.edu/old\\_website/tools/cpbase](http://rocaplab.ocean.washington.edu/old_website/tools/cpbase).
- Roeber, FK, GA Gordillo, and HH Geiger (2005). "In vivo haploid induction in maize. Performance of new inducers and significance of doubled haploid lines in hybrid breeding [*Zea mays* L.]" In: *Maydica (Italy)*.
- Rosenblatt, Frank (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Tech. rep. Cornell Aeronautical Lab Inc Buffalo NY.
- Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747*.
- Rumelhart, David E, Geoffrey E Hinton, Ronald J Williams, et al. (1988). "Learning representations by back-propagating errors". In: *Cognitive modeling* 5.3, p. 1.

- Sancho, Rubén et al. (June 2018). "Comparative plastome genomics and phylogenomics of *Brachypodium* : flowering time signatures, introgression and recombination in recently diverged ecotypes". en. In: *New Phytologist* 218.4, pp. 1631–1644. ISSN: 0028646X. DOI: [10.1111/nph.14926](https://doi.org/10.1111/nph.14926). URL: <http://doi.wiley.com/10.1111/nph.14926>.
- Santos Dias, Luiz Antônio dos et al. (2004). "A priori choice of hybrid parents in plants". In: *Genet. Mol. Res* 3.3, pp. 356–368.
- Scarcelli, N. et al. (2016). "Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it?" In: *Molecular Ecology Resources* 16.2, pp. 434–445. DOI: [10.1111/1755-0998.12462](https://doi.org/10.1111/1755-0998.12462). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12462>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12462>.
- Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural networks* 61, pp. 85–117.
- Schopp, Pascal et al. (2017a). "Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium". In: *Genetics* 205.1, pp. 441–454.
- Schopp, Pascal et al. (2017b). "Genomic prediction within and across biparental families: means and variances of prediction accuracy and usefulness of deterministic equations". In: *G3: Genes, Genomes, Genetics* 7.11, pp. 3571–3586.
- Schrag, Tobias A et al. (2018). "Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize". In: *Genetics* 208.4, pp. 1373–1385.
- Segura, Vincent et al. (2012). "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations". In: *Nature genetics* 44.7, p. 825.
- Seren, Ümit et al. (2016). "AraPheno: a public database for *Arabidopsis thaliana* phenotypes". In: *Nucleic acids research*, gkw986.
- Shen, Wei et al. (Oct. 2016). "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation". In: *PLOS ONE* 11.10, pp. 1–10. DOI: [10.1371/](https://doi.org/10.1371/journal.pone.0162850)

- journal.pone.0163962. URL: <https://doi.org/10.1371/journal.pone.0163962>.
- Shen, Xia et al. (2013). "A novel generalized ridge regression method for quantitative genetics". In: *Genetics* 193.4, pp. 1255–1268.
- Shinozaki, K. et al. (1986). "The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression". In: *The EMBO Journal* 5.9, pp. 2043–2049. ISSN: 1460-2075. DOI: [10.1002/j.1460-2075.1986.tb04464.x](https://doi.org/10.1002/j.1460-2075.1986.tb04464.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1460-2075.1986.tb04464.x> (visited on 05/20/2019).
- Siva, Nayanah (2008). *1000 Genomes project*.
- Snowdon, Rod J and Federico L Iniguez Luy (2012). "Potential to improve oilseed rape and canola breeding in the genomics era". In: *Plant breeding* 131.3, pp. 351–360.
- Soper, HE et al. (1917). "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and RA Fisher". In: *Biometrika* 11.4, pp. 328–413.
- Sousa, Massaine Bandeira e et al. (2017). "Genomic-enabled prediction in maize using kernel models with genotype  $\times$  environment interaction". In: *G3: Genes, Genomes, Genetics* 7.6, pp. 1995–2014.
- Stegemann, Sandra and Ralph Bock (2009). "Exchange of genetic material between cells in plant tissue grafts". In: *science* 324.5927, pp. 649–651.
- Stewart-Brown, Benjamin B. et al. (2019). "Genomic Selection for Yield and Seed Composition Traits Within an Applied Soybean Breeding Program". In: *G3: Genes, Genomes, Genetics* 9.7, pp. 2253–2265. DOI: [10.1534/g3.118.200917](https://doi.org/10.1534/g3.118.200917). URL: <https://doi.org/10.1534>.
- Storey, John D. and Robert Tibshirani (2003). "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16, pp. 9440–9445. ISSN: 0027-8424. DOI: [10.1073/pnas.1530509100](https://doi.org/10.1073/pnas.1530509100). eprint: <https://www.pnas.org/content/100/16/9440.full.pdf>. URL: <https://www.pnas.org/content/100/16/9440>.

- Strauch, Rene et al.* (2015). "Discovery of a novel amino acid racemase through exploration of natural variation in *Arabidopsis thaliana*". In: *PNAS* 112. DOI: [10.1073/pnas.1503272112](https://doi.org/10.1073/pnas.1503272112).
- Stringer, Sven et al.* (2011). "Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes". In: *PloS one* 6.11, e27964.
- Sukumaran, Sivakumar et al.* (2016). "Genomic Prediction with Pedigree and Genotype Environment Interaction in Spring Wheat Grown in South and West Asia, North Africa, and Mexico". In: *G3: Genes, Genomes, Genetics* 7.2, pp. 481–495. DOI: [10.1534/g3.116.036251](https://doi.org/10.1534/g3.116.036251). URL: <https://doi.org/10.1534>.
- Tang, Chunlao et al.* (2007). "The evolution of selfing in *Arabidopsis thaliana*". In: *Science* 317.5841, pp. 1070–1072.
- Technow, Frank, Anna Bürger, and Albrecht E Melchinger* (2013). "Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups". In: *G3: Genes, Genomes, Genetics* 3.2, pp. 197–203.
- Technow, Frank et al.* (2014). "Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize". In: *Genetics* 197.4, pp. 1343–1355.
- Tetko, Igor V, David J Livingstone, and Alexander I Luik* (1995). "Neural network studies. 1. Comparison of overfitting and overtraining". In: *Journal of chemical information and computer sciences* 35.5, pp. 826–833.
- Thavamanikumar, Saravanan, Rudy Dolferus, and Bala R. Thumma* (2015). "Comparison of Genomic Selection Models to Predict Flowering Time and Spike Grain Number in Two Hexaploid Wheat Doubled Haploid Populations". In: *G3: Genes, Genomes, Genetics* 5.10, pp. 1991–1998. DOI: [10.1534/g3.115.019745](https://doi.org/10.1534/g3.115.019745). URL: <https://doi.org/10.1534>.
- Thorwarth, Patrick, Eltohamy AA Yousef, and Karl J Schmid* (2018). "Genomic Prediction and Association Mapping of Curd-Related Traits in Gene Bank Accessions of Cauliflower". In: *G3: Genes, Genomes, Genetics* 8.2, pp. 707–718.

- Timpson, Nicholas J et al. (2018). "Genetic architecture: the shape of the genetic contribution to human traits and disease". In: *Nature Reviews Genetics* 19.2, p. 110.
- Togninalli, Matteo et al. (2017). "The AraGWAS Catalog: a curated and standardized Arabidopsis thaliana GWAS catalog". In: *Nucleic acids research* 46.D1, pp. D1150–D1156.
- Twyford, Alex D. and Rob W. Ness (Sept. 1, 2017). "Strategies for complete plastid genome sequencing". In: *Molecular Ecology Resources* 17.5, pp. 858–868. ISSN: 1755-0998. DOI: [10.1111/1755-0998.12626](https://doi.org/10.1111/1755-0998.12626). URL: <http://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12626/abstract> (visited on 01/26/2018).
- Unterseer, Sandra et al. (2014). "A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array". In: *BMC genomics* 15.1, p. 823.
- Van Rossum, Guido and Fred L Drake Jr (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.
- VanRaden, Paul M (2008). "Efficient methods to compute genomic predictions". In: *Journal of dairy science* 91.11, pp. 4414–4423.
- VanRaden, PM et al. (2008). "Reliability of genomic predictions for North American dairy bulls". In: *J. Dairy Sci* 91.Suppl 1, p. 305.
- Verbyla, Klara L et al. (2009). "Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle". In: *Genetics research* 91.5, pp. 307–311.
- Vieira, IC et al. (2017). "Assessing non-additive effects in GBLUP model". In: *Genetics and molecular research: GMR* 16.2.
- Vinga, Susana et al. (2012). "Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis". In: *Algorithms for molecular biology : AMB* 7.1. 22551152[pmid], pp. 10–10. ISSN: 1748-7188. DOI: [10.1186/1748-7188-7-10](https://doi.org/10.1186/1748-7188-7-10). URL: <https://www.ncbi.nlm.nih.gov/pubmed/22551152>.
- Walsh, B and M Lynch (2018a). "Short-term Changes in the Mean: 2. Truncation and Threshold Selection". In:

- Walsh, Bruce and Michael Lynch (2018b). *Evolution and selection of quantitative traits*. Oxford University Press.
- Wang, Weiwen et al. (2018). "Assembly of chloroplast genomes with long- and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case". In: *BMC genomics* 19.1. 30594129[pmid], pp. 977–977. ISSN: 1471-2164. DOI: [10.1186/s12864-018-5348-8](https://doi.org/10.1186/s12864-018-5348-8). URL: <https://www.ncbi.nlm.nih.gov/pubmed/30594129>.
- Wang, Yu et al. (2014). "The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years". In: *BMC genomics* 15.1, p. 556.
- Warner, Brad and Manavendra Misra (1996). "Understanding neural networks as statistical tools". In: *The american statistician* 50.4, pp. 284–293.
- Watson, Amy et al. (2018). "Speed breeding is a powerful tool to accelerate crop research and breeding". In: *Nature plants* 4.1, p. 23.
- Webb, Sarah (2018). "Deep learning for biology". In: *Nature* 554.7693.
- Werner, Christian R et al. (2018). "Effective genomic selection in a narrow-genepool crop with low-density markers: Asian rapeseed as an example". In: *The plant genome* 11.2.
- Wicke, Susann et al. (July 1, 2011). "The evolution of the plastid chromosome in land plants: gene content, gene order, gene function". In: *Plant Molecular Biology* 76.3, pp. 273–297. ISSN: 1573-5028. DOI: [10.1007/s11103-011-9762-4](https://doi.org/10.1007/s11103-011-9762-4). URL: <https://doi.org/10.1007/s11103-011-9762-4> (visited on 05/16/2019).
- Windhausen, Vanessa S et al. (2012). "Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments". In: *G3: Genes, Genomes, Genetics* 2.11, pp. 1427–1436.
- Wright, Sewall (1922). "Coefficients of inbreeding and relationship". In: *The American Naturalist* 56.645, pp. 330–338.
- Würschum, Tobias (2012). "Mapping QTL for agronomic traits in breeding populations". In: *Theoretical and Applied Genetics* 125.2, pp. 201–210.



- Würschum, Tobias, Stefan Abel, and Yusheng Zhao (2014). "Potential of genomic selection in rapeseed (*B. rassica napus* L.) breeding". In: *Plant Breeding* 133.1, pp. 45–51.
- Würschum, Tobias et al. (2013). "Genomic selection in sugar beet breeding populations". In: *BMC genetics* 14.1, p. 85.
- Xavier, Alencar, William M. Muir, and Katy Martin Rainey (2016). "Assessing Predictive Properties of Genome-Wide Selection in Soybeans". In: *G3* 6.8, pp. 2611–2616. DOI: [10.1534/g3.116.032268](https://doi.org/10.1534/g3.116.032268). URL: <https://doi.org/10.1534>.
- Xiao-Ming, Zheng et al. (May 8, 2017). "Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants". In: *Scientific Reports* 7.1, p. 1555. ISSN: 2045-2322. DOI: [10.1038/s41598-017-01518-5](https://doi.org/10.1038/s41598-017-01518-5). URL: <https://www.nature.com/articles/s41598-017-01518-5> (visited on 05/16/2019).
- Xu, S (2010). "An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects". In: *Heredity* 105.5, p. 483.
- Xu, Shizhong (Aug. 2013). "Genetic Mapping and Genomic Selection Using Recombination Breakpoint Data". In: *Genetics* 195.3, pp. 1103–1115. DOI: [10.1534/genetics.113.155309](https://doi.org/10.1534/genetics.113.155309). URL: <https://doi.org/10.1534/genetics.113.155309>.
- Yang, Jian et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". In: *Nature genetics* 42.7, p. 565.
- Zapata-Valenzuela, Jaime et al. (2013). "Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine". In: *G3: Genes, Genomes, Genetics* 3.5, pp. 909–916. DOI: [10.1534/g3.113.005975](https://doi.org/10.1534/g3.113.005975). URL: <https://doi.org/10.1534>.
- Zeiler, Matthew D (2012). "ADADELTA: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701*.
- Zhang, Yan et al. (2018). "Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits". In: *Nature genetics* 50.9, p. 1318.



- Zhang, Zhiwu et al. (Mar. 2010). "Mixed linear model approach adapted for genome-wide association studies". In: *Nature Genetics* 42.4, pp. 355–360. DOI: [10.1038/ng.546](https://doi.org/10.1038/ng.546). URL: <https://doi.org/10.1038/ng.546>.
- Zheng, Xiuwen (2013). "A Tutorial for the R Package SNPRelate". In: *University of Washington, Washington, USA*.
- Zhong, Shengqiang et al. (2009). "Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study". In: *Genetics* 182.1, pp. 355–364.
- Zhou, Xiang and Matthew Stephens (June 2012). "Genome-wide efficient mixed-model analysis for association studies". In: *Nature Genetics* 44.7, pp. 821–824. DOI: [10.1038/ng.2310](https://doi.org/10.1038/ng.2310). URL: <https://doi.org/10.1038/ng.2310>.
- Žilinskas, A (2006). *Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithms*.