Identification of regulatory SNPs and epistatic SNP pairs using deep learning and information theory

Dissertation

to attain the doctoral degree Dr. rer. nat.

of the Faculty of Agricultural Sciences

Georg-August-Universität Göttingen

Submitted by

Felix Heinrich

born on the 14.03.1994 in Herzberg

Göttingen, April 2022

1. Referee: Prof. Dr. Armin Schmitt,
Breeding Informatics Group, Department of Animal Sciences,
Georg-August-Universität Göttingen.

2. Referee: Prof. Dr. Stephan Waack,
Institute of Computer Science, Georg-August-Universität Göttingen.

3. Referee: Prof. Dr. Murtaza Özgür Yeniay,
Department of Statistics, Faculty of Science, Hacettepe University Ankara.

Date of oral examination: 12.07.2022

Abstract

In the last two decades, new technologies have made DNA genotyping and sequencing far more time and cost efficient. The resulting tremendous increase in the amount of available genomic data allows for a deeper understanding of the relationship between the genotype and the phenotype. In this thesis, I present two novel frameworks which analyze specific aspects of the relationship between the genotype and the phenotype, namely the identification of regulatory SNPs (rSNPs) as well as the detection of epistatic SNP pairs.

In my first framework, I utilized deep learning to train a convolutional neural network for the prediction of promoter sequences in the species *Vicia faba*. By exploiting the conservation of promoter signatures across closely related species, I avoided the need for the expensive and time-consuming task of assembling and annotating a reference genome for the species under study. With the detected promoter regions, I was then able to analyze putative rSNPs in terms of their effects on the binding of transcription factors. Finally, my results revealed two rSNPs which were highly associated with the trait under study, namely the vicine and convicine content (V+C) of the plants. These markers could then be further used in plant breeding programs that target a low V+C content. Furthermore, I thereby demonstrated that an annotated reference genome is not always necessary for this type of analysis.

For my second framework, I developed a method named MIDESP for the detection of epistatic interactions between SNP pairs based on mutual information. This method extends the existing information theory-based approaches for epistasis detection in two key areas. First, by adopting a kth-nearest neighbor-based approach for estimating mutual information, it is the first mutual information-based method which can be applied to detect epistasis for qualitative as well as quantitative phenotypes. Secondly, the method incorporates the average product correction (APC) to deal with possible complications in a genotype-phenotype dataset, which may otherwise give rise to the detection of false-positive interactions. I showcase the performance of MIDESP and its different aspects by means of simulated as well as real datasets, which were related to bovine tuberculosis and the weight of chicken eggs, respectively. Comparing the results with and without the application of the APC showed that the correction is necessary to reduce the prediction of false-positive interactions.

Overall, both of my frameworks provide novel insights into specific mechanisms underlying the relationship between the genotype and the phenotype and identify important SNPs that are participating in these mechanisms.

Zusammenfassung

In den letzten zwei Jahrzehnten haben neue Technologien die DNA-Genotypisierung und -Sequenzierung wesentlich zeit- und kosteneffizienter gemacht. Die dadurch erzielte enorme Zunahme der verfügbaren Genomdaten ermöglicht ein tieferes Verständnis der Beziehung zwischen dem Genotyp und dem Phänotyp. In dieser Arbeit stelle ich zwei neuartige Verfahren vor, die spezifische Aspekte der Beziehung zwischen Genotyp und Phänotyp analysieren, nämlich die Identifizierung von regulatorischen SNPs (rSNPs) und die Erkennung epistatischer SNP-Paare.

Bei meiner ersten Methode setze ich maschinelles Lernen ein, um ein neuronales Faltungsnetzwerk für die Vorhersage von Promotersequenzen in der Art *Vicia faba* zu trainieren. Durch die Ausnutzung der Konservierung von Promotersignaturen bei eng verwandten Arten, konnte ich die teure und zeitaufwändige Aufgabe der Assemblierung und Annotation eines Referenzgenomes für die untersuchte Art vermeiden. Anhand der entdeckten Promoterregionen konnte ich dann mutmaßliche rSNPs in Bezug auf ihre Auswirkungen auf die Bindung von Transkriptionsfaktoren analysieren. Schlussendlich ergaben meine Ergebnisse zwei rSNPs, die in hohem Maße mit dem untersuchten Merkmal assoziiert waren, nämlich dem Vicin- und Convicin-Gehalt (V+C) der Pflanzen. Diese Marker könnten dann in Pflanzenzuchtprogrammen verwendet werden, die auf einen niedrigen V+C-Gehalt abzielen. Ich habe damit ebenfalls gezeigt, dass für diese Art der Analyse nicht immer ein annotiertes Referenzgenom erforderlich ist.

Für meinen zweiten Ansatz habe ich eine Methode namens MIDESP zur Erkennung epistatischer Interaktionen zwischen SNP-Paaren auf Grundlage der wechselseitigen Information entwickelt. Diese Methode erweitert die bestehenden auf der Informationstheorie basierenden Ansätze zur Epistasisdetektion in zwei Schlüsselbereichen. Erstens ist sie durch die Anwendung eines *k*-nächsten Nachbarn-basierten Ansatzes zur Schätzung der wechselseitigen Information die erste auf wechselseitiger Information basierende Methode, die zur Erkennung von Epistasis sowohl für qualitative als auch für quantitative Phänotypen angewendet werden kann. Zweitens beinhaltet die Methode die sogenannte “average product correction” (APC), um mit möglichen Komplikationen in einem Genotyp-Phänotyp-Datensatz umzugehen, die andernfalls zur Erkennung von falsch-positiven Interaktionen führen könnten. Ich demonstriere die Leistung von MIDESP und seiner verschiedenen Aspekte anhand von simulierten und realen Datensätzen, die sich auf Rindertuberkulose bzw. das Gewicht von Hühnereiern beziehen. Der Vergleich der Ergebnisse mit und ohne Anwendung der APC zeigte, dass die Korrektur notwendig ist, um die Anzahl von falsch-positiven Interaktionen zu reduzieren.

Insgesamt liefern meine beiden Methoden neue Einblicke in spezifische Mechanismen, die der Beziehung zwischen Genotyp und Phänotyp zugrunde liegen, und identifizieren wichtige SNPs, die an diesen Mechanismen beteiligt sind.

Acknowledgements

During my time at the university I was accompanied and supported by many people, and I thank them all for their help.

First of all, I would like to thank Prof. Armin Schmitt, who offered me the opportunity to first do my Master's thesis and then my PhD in his group. Prof. Schmitt was always available to me for questions and discussions. In doing so, he provided a warm and welcoming atmosphere that contributed to my desire to further pursue my work in the field of academic research. I look forward to continuing my work with him in the future.

I would also like to thank my second supervisor Prof. Stephan Waack for his support and guidance throughout my PhD, and whose lectures on information theory provided valuable insights for my work.

Since my time as an undergraduate student in the Institute of Medical Bioinformatics, I was supported and mentored by the then Dr. and now Prof. Mehmet Gültas. Without his help, I would not have found my way into the Breeding Informatics group and would almost certainly not have continued my studies. His ideas, feedback and guidance as my third supervisor contributed significantly to my projects and it was a pleasure to work with him.

Further, I would like to thank Prof. Murtaza Özgür Yeniay for becoming my referee.

I would also like to thank my fellow PhD students as well as all present and former colleagues in the Breeding Informatics group. Many of them have provided invaluable support to my work or allowed me to participate in their own projects. In particular, I want to thank my co-authors Pronaya Prosun Das, Miriam Kamp, Selina Klees, Abirami Rajavel, Faisal Ramzan and Martin Wutke. Furthermore, I am grateful for the many opportunities my colleagues have given me to learn how to fix computer problems. Another thank you goes to Monika Siebert for her help with the organizational part of my work as well as to Hendrik Betram, Ata ul Haleem, Thomas Lange and Johanna Schlüter for their support during my PhD thesis.

Additional thanks go to the group of Prof. Wolfgang Link whose collaboration helped to facilitate my Master's thesis and first publication.

Finally, I would like to thank my parents. They supported me during my studies and in my daily life and they have made it possible for me to concentrate fully on my studies. I dedicate this thesis to them.

Last but not least, I thank my sister Jasmin who, despite her own work and studies, found the time to proofread my publications and this thesis.

Contents

1. Introduction **1**
- 1.1. Structure of the thesis . . . 4
- 1.2. Impact . . . 4

2. Biological background **7**
- 2.1. Genotype and phenotype . . . 7
- 2.2. Gene expression . . . 7
  - 2.2.1. DNA . . . 7
  - 2.2.2. Genes . . . 9
  - 2.2.3. Regulation of gene expression . . . 10
  - 2.2.4. Regulation of transcription by transcription factors . . . 10
- 2.3. Single Nucleotide Polymorphisms . . . 13
  - 2.3.1. Regulatory SNPs . . . 13
  - 2.3.2. Genotyping methods . . . 14
- 2.4. Epistasis . . . 15
- 2.5. Bioinformatic resources . . . 17
  - 2.5.1. Bioinformatic databases . . . 17
  - 2.5.2. Bioinformatic tools . . . 19

3. Theoretical background **23**
- 3.1. Prediction of promoter sequences . . . 23
  - 3.1.1. Detection of promoters using CNNs . . . 24
- 3.2. Information theory . . . 26
  - 3.2.1. Entropy . . . 26
  - 3.2.2. Mutual Information . . . 28
  - 3.2.3. Multivariate Mutual Information . . . 31
  - 3.2.4. Information theory for continuous variables . . . 35
  - 3.2.5. Normalized Mutual Information . . . 39
- 3.3. Genotype-phenotype association studies . . . 40
  - 3.3.1. Linkage disequilibrium . . . 40
  - 3.3.2. Detection of epistatic interactions . . . 41

4. Material and methods **45**
4.1. Datasets . . . 45
4.1.1. Genotyping-by-Sequencing Data of *Vicia faba* . . . 45
4.1.2. Partial genome and SNPs for *Vicia faba* . . . 46
4.1.3. Promoter and non-promoter sequences of several plant species . . . 47
4.1.4. Bovine tuberculosis (BT) dataset . . . 47
4.1.5. Egg weight (EW) dataset . . . 48
4.2. Identification of regulatory SNPs . . . 49
4.2.1. Identification of promoter regions . . . 49
4.2.2. Identification of putative regulatory SNPs with association to V+C . 53
4.3. Mutual information based detection of epistatic SNP pairs . . . 53
4.3.1. Pre-processing . . . 55
4.3.2. Application of mutual information for SNP×phenotype associations 55
4.3.3. Detection of SNPs with strong association signals . . . 60
4.3.4. Reduction of the background associations between SNPs and phenotype . . . 62
4.3.5. Validation of the epistatic interactions . . . 63
4.3.6. Implementation . . . 63

5. Results **65**
5.1. Identification of regulatory SNPs . . . 65
5.1.1. Intra- and inter-species promoter prediction . . . 65
5.1.2. Effect of additional features on model prediction . . . 68
5.1.3. Prediction of *Vicia faba* promoters . . . 69
5.1.4. SNPs in putative promoter regions and their association with V+C . 70
5.1.5. Systematic identification of regulatory SNPs associated with V+C in *Vicia faba* . . . 70
5.1.6. Functional analysis of the candidate gene and transcription factors . 71
5.2. Mutual information based detection of epistatic SNP pairs . . . 75
5.2.1. Analysis of simulated datasets for parameter setting . . . 75
5.2.2. Single SNP association using mutual information . . . 76
5.2.3. Illustration of background associations and their correction using APC . . . 78
5.2.4. Results from the analysis of the bovine tuberculosis dataset . . . 80
5.2.5. Results from the analysis of the egg weight dataset . . . 81
5.2.6. Comparisons with existing methods . . . 84
5.3. Detection of epistatic SNP pairs associated with V+C in *Vicia faba* . . . 86

6. Discussion **91**
6.1. Identification of regulatory SNPs . . . 91
6.2. Mutual information based detection of epistatic SNP pairs . . . 95

7. Conclusion **99**
7.1. Summary 99
7.2. Outlook 100

Bibliography **103**

A. Appendix **129**
A.1. Identification of Regulatory SNPs Associated with Vicine and Convicine Content of *Vicia faba* Based on Genotyping by Sequencing Data Using Deep Learning 129
A.2. Genotyping by Sequencing Reads of 20 *Vicia faba* Lines 146
A.3. MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs . . . 151
A.4. agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species 174
A.5. Master thesis 192

B. Curriculum Vitae **263**

List of Figures

2.1. Structure of the DNA . . . 9
2.2. Structure of the core promoter . . . 11
2.3. Construction of a PWM . . . 12
2.4. Example for a logo plot . . . 12
2.5. Consequences for a SNP-TFBS pair . . . 14
2.6. Ensembl overview for *Bos taurus* . . . 18
2.7. PCD page for *Vicia faba* . . . 19
2.8. Example of a MATCH^{TM} result . . . 21
2.9. GeneXplain overview . . . 21

3.1. Example of a convolutional operation . . . 25
3.2. One-hot encoding of a DNA sequence . . . 26
3.3. Relationship between entropy and mutual information for two variables . . . 29
3.4. Relationship between entropy and information theory measures for three variables . . . 33
3.5. Binning of a continuous random variable . . . 36

4.1. CNN model for promoter prediction . . . 49
4.2. Flowchart of the MIDESP method. . . . 54
4.3. Example for the problem of duplicate values when estimating $\mathbb{MII}$ . . . 58
4.4. Example for the adjusted $\mathbb{MII}$ estimator to handle duplicate values . . . 58
4.5. Special case for the $\mathbb{MII}$ estimator . . . 59
4.6. β distribution of $\mathbb{NMII}$ values . . . 62

5.1. Analysis of simulated datasets for varying parameter settings of k . . . 76
5.2. Epistasis example for two SNPs and a quantitative phenotype . . . 77
5.3. Effect of APC on the BT dataset . . . 79
5.4. GO treemap for genes associated with immunity to bovine tuberculosis . . . 80
5.5. GO treemap for genes associated with egg weight . . . 82
5.6. Comparison of results between original and adjusted MIDESP . . . 83
5.7. Comparison of epistasis results on the BT dataset . . . 85
5.8. Comparison of epistasis results on the EW dataset . . . 86
5.9. Effect of APC on the *Vicia faba* dataset . . . 88
5.10. Association of epistatic SNP pairs on the *Vicia faba* dataset . . . 89

List of Tables

4.1. Information about the 20 *Vicia faba* lines 46
4.2. Information about sequence data used for promoter prediction 47

5.1. Cross-Species Promoter Prediction - ACC 66
5.2. Cross-Species Promoter Prediction - Sensitivity 66
5.3. Cross-Species Promoter Prediction - Specificity 67
5.4. Cross-Species Promoter Prediction - MCC 67
5.5. Contribution of additional features in the promoter prediction 69
5.6. SNPs found in *Vicia faba* promoters and mapped to *Medicago truncatula* . 71
5.7. Alleles of the 20 *Vicia faba* lines for the two significantly associated SNPs 72
5.8. Consequences of SNPs associated with V+C 72
5.9. Comparison of mutual information and linear regression for GWAS 78
5.10. Number of SNP pairs found as epistatic interactions 84
5.11. Number of epistatic SNP pairs in *Vicia faba* with rSNPs 87

Acronyms

APC	Average Product Correction
CMI	Conditional Mutual Information
CNN	Convolutional Neural Network
CPE	Core Promoter Element
CSS	Core Similarity Score
DNA	Deoxyribonucleic Acid
FDR	False Discovery Rate
GBS	Genotyping by Sequencing
GTE	Generalized Topological Entropy
GTF	General Transcription Factor
GWAS	Genome Wide Association Studies
HMI	Horizontal Mutual Information
IG	Information Gain
ML	Machine Learning
MSS	Matrix Similarity Score
MMI	Multivariate Mutual Information
MI	Mutual Information
NGS	Next Generation Sequencing
PWM	Position Weight Matrix
rSNP	Regulatory Single Nucleotide Polymorphism
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site
V+C	Vicine and Convicine

1. Introduction

In the last two decades, new methods in the field of genome sequencing, collectively referred to as next generation sequencing (NGS), have been developed, which offer cost-effective strategies to produce massive amounts of sequencing data [1]. One of these methods is genotyping-by-sequencing (GBS), which is an efficient method to obtain genome-wide genotype data for any species [2, 3]. The characteristic feature of GBS is the reproducible generation of short genomic fragments using known restriction enzymes. Due to its easy applicability, GBS is currently the method of choice in the field of plant sciences, since it makes plants without reference genome amenable to genomic analysis. Several groups have applied GBS to obtain high-quality genome-wide single nucleotide polymorphism (SNP) markers. These markers have often been used for applications like genome wide association studies (GWASs), marker-assisted selection, breeding value estimation in genomic prediction, analysis of high density genetic maps, or assessment of population dynamics in plant genomics and plant breeding [4, 5, 6, 7, 8, 9, 10, 11]. At the same time, the development of new high-density SNP arrays for a multitude of species allows for the simultaneous genotyping of up to several hundred thousand SNPs in a cost-effective and time-efficient manner [12, 13, 14, 15]. Altogether, the amount of genomic data has increased tremendously, which enables a more detailed research into the genetic background of qualitative and quantitative traits.

One crop species that can greatly benefit from these advancements is the faba bean [16, 17]. *Vicia faba*, or faba bean, is an old world grain legume, which is globally grown for both human consumption as well as feed for livestock [18]. The species is diploid with $2x = 12$ very large chromosomes. Due to its large size of 13 Gbp [16], there is thus far no sequenced and annotated reference genome available for this plant. Despite its agro-ecological importance (nitrogen symbiosis, rotation hygiene, and pollinator support) [19] and high protein content of approximately 30% [20] it is a crop of limited importance in many countries. This is mainly caused by its anti-nutritive seed-compounds vicine and convicine (in the following termed V+C), which are co-occurring pyrimidine glycosides that are assumed to be formed in the seed coat of *Vicia faba* [21, 22]. These anti-nutrients have negative effects on animals such as laying hens, broilers and piglets, but also on 400 million humans suffering from glucose-6-phosphate-dehydrogenase deficiency [23, 24]. Therefore, the V+C content is a factor that severely limits the wider usage of *Vicia faba* as feed for animals and food for humans. Breeding new V+C-poor varieties and production and marketing of their fruits could have a number of positive effects, such as reducing environmentally critical soybean imports to Europe and North America, fostering of regional production methods, and avoid-

ing energy-intensive transports. This was for example the goal of the project Abo-Vici [25], which was supported by the German Federal Ministry of Food and Agriculture as part of its Protein Crop Strategy to raise the importance of domestic protein crops in Germany and Europe [26].
Despite ongoing research efforts and the discovery of a robust marker for the V+C content, the responsible genes and mechanisms remained unknown for a long time, and the location of the responsible locus could only be restricted to an interval on chromosome 1 of *Vicia faba*, that shows conserved synteny with a 900 kb region on chromosome 2 of the related species *Medicago truncatula*, which is located between the *Medicago truncatula* genes Medtr2g008210 and Medtr2g010180 [21, 23]. Only recently could a putative key enzyme of the V+C pathway be identified [27]. This enzyme encodes a GTP cyclohydrolase II which participates in the biosynthesis of riboflavin from GTP [28]. Moreover, the authors showed through feeding studies that GTP is a precursor for V+C and thereby further validated the role of the identified enzyme [27].
In this thesis, my aim is to present two novel frameworks that analyze specific aspects of the relationship between genotype and phenotype, namely the identification of regulatory SNPs (rSNPs) as well as the detection of epistatic SNP pairs.
rSNPs are located in the promoter regions of genes and can influence the gene regulation by affecting transcription factor binding sites (TFBSs). Today, it is well known that these rSNPs may be causal for the phenotype and could therefore possibly provide prime candidates useful for breeding programs or marker-assisted selection [29, 30, 31, 32, 33, 34, 35]. Nevertheless, the identification of regulatory SNPs requires the location of the promoter regions in the genome of the species under study. Despite the rich literature on the analysis of promoters, their prediction remains a challenging task due to their complex and diverse structure. Until now, different machine learning approaches have been developed, which form the core of most computational prediction methods for promoter regions. Whereas in early works the emphasis was on the identification of specific promoter elements (such as TATA boxes, initiator elements, downstream promoter elements and others) or the extraction of k-mer distributions [36, 37, 38, 39, 40, 41, 42, 43], nowadays a more holistic approach is given preference in which whole genomic regions are examined using convolutional neural networks (CNNs), which have been successfully applied in many species [44, 45, 46, 47, 48, 49]. However, these approaches still require annotated training data for the respective species, which is missing in the case of *Vicia faba*. To address this problem, I exploited the conservation of promoter signatures among closely related plant species [50, 51] and first trained a CNN model using the known promoter sequences of seven species of the *Leguminosae* family, to which *Vicia faba* belongs. Second, using GBS sequence reads of 20 *Vicia faba* lines with known V+C content, I assembled a *de novo* draft partial genome. Thereafter, I called the genomic variants by aligning the GBS reads to the partial genome to obtain high quality SNPs for candidate gene association studies. Next, by applying the CNN model to the partial genome sequences, I predicted the potential promoter sequences of *Vicia faba*. Finally, I analyzed the SNPs in these promoter

sequences that were associated with the V+C content of *Vicia faba* regarding their effect on the binding affinity of transcription factors. Using this framework, it is thereby possible to identify rSNPs for *Vicia faba* based directly on GBS data [52].

The aforementioned development of high-density arrays for genotyping in recent years has allowed GWASs to become powerful tools for the detection of SNPs that are associated with traits of interest. However, GWAS methods are usually based on the analysis of single loci, ignoring the potential interaction between genes, and are therefore of limited applicability for complex traits. Today, it is well-known that the expressions of complex traits are often controlled by multiple genes that can interact with each other in complex manners [53, 54, 55]. These genes may have only a small effect on the phenotype and could therefore be missed in single-locus analyses, despite having a strong influence based on their interactions [56, 57, 58, 59]. While large parts of the phenotype variance are attributed to individual SNP effects, these interactions, which are commonly referred to as epistasis, have been shown to be of importance for many complex diseases in humans such as asthma [60], cancer [61] or diabetes [62], as well as for quantitative traits in animals [55, 63, 64, 65, 66, 67] and plants [68, 69, 70, 71, 72], and could help to explain the relationship between the genetic variants and the corresponding phenotype [54, 65, 73, 74]. Among the numerous methods that have been developed to detect epistasis, several use information-theory-based measures such as mutual information to quantify epistatic interactions [75, 76, 77, 78, 79, 80, 81, 82, 83, 84]. While being quite successful in general, these methods are limited in specific aspects. For one, the application of information-theory-based approaches has so far been limited to case-control phenotypes because estimating the mutual information for a continuous phenotype is a computationally far more challenging task than it is for a discrete phenotype. This limits their utility, particularly in the field of animal and plant sciences where quantitative traits are common. Secondly, these methods do not necessarily take into account different types of complications resulting from sample structure, relatedness between the genotyped individuals or marginal effects of single SNPs on the phenotype [71, 85, 86]. Such types of complications can lead to background associations between the SNP pairs and the phenotype, and thus the importance of some SNPs in the epistatic interactions could be overestimated. Consequently, the prediction of existing methods could be biased, potentially impeding the identification of correct epistatic signals. Hence, the elimination of the bias inherent in the genotype-phenotype datasets is needed to separate the signal caused by functional interactions from the background associations between SNPs [71, 85]. To overcome these limitations, I developed a novel method called Mutual Information-based Detection of Epistatic SNP Pairs (MIDESP) for the detection of pairwise epistatic interactions, which extends the previously mentioned mutual information-based approaches by additionally enabling the identification of epistatic interactions between SNP pairs and quantitative phenotypes. For this purpose I adopt, in the context of epistasis for the first time, the mutual information estimator developed by Ross [87], which accurately estimates the level of epistasis using a kth-nearest neighbor-based approach. Moreover, to deal with possible complications within a genotype-phenotype

dataset (as mentioned above), my method incorporates the average product correction (APC) theorem [88] as an additional step to estimate the expected level of background association for each SNP pair. Finally, the removal of the estimated background from the measured epistasis values leads to the detection of true epistatic signals arising from functional interactions. In order to demonstrate its performance and functionality, I applied MIDESP to simulated datasets and two real datasets with a qualitative and quantitative phenotype, respectively, as well as to the *Vicia faba* dataset.

Overall, my results suggest that, by applying my two frameworks novel insights into the relationship between genotype and phenotype can be obtained. This knowledge could then in turn be used to provide new markers and strategies for breeding programs such as the aforementioned Abo-Vici project.

1.1. Structure of the thesis

The organization of the thesis is as follows. In Chapter 2, I introduce the most relevant biological concepts that are required to understand this thesis, in particular transcriptional gene regulation and epistasis. I further give an overview of the different bioinformatics databases and tools used in this thesis. In Chapter 3, I give a brief overview into the theoretical concepts underlying this thesis. To start, I describe the existing techniques of promoter detection followed by a primer into information theory. In the last part, I introduce the idea of genotype-phenotype association analyses with a particular focus on detecting epistastic interactions. Followed by the biological and theoretical background, I describe the used datasets and present the two analysis frameworks established in this thesis in Chapter 4. Thereby, I first introduce the framework for the identification of regulatory SNPs based on genotyping-by-sequencing data in Section 4.2. In the following Section 4.3, I describe the MIDESP algorithm for the detection of epistatic SNP pairs using mutual information. Afterwards, I present the results of both methods on simulated and real datasets in Chapter 5. The application of the two methods is discussed in Chapter 6 and finally, I complete this work in Chapter 7 by summarizing the thesis and providing an outlook for future work.

1.2. Impact

Journal articles: I have published the two frameworks described in this thesis in the following articles:

[1] **Heinrich, F.**; Wutke, M.; Das, P.P.; Kamp, M.; Gültas, M.; Link, W.; Schmitt, A.O. (2020). *Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning.* Genes, 11(6), 614.

[2] **Heinrich, F.**; Ramzan, F.; Rajavel, A.; Schmitt, A.O.; Gültas, M. (2021). *MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes.* Biology, 10(9), 921.

Detailed author contribution of Felix Heinrich to both journal articles: Participated in the design of the studies. Prepared the datasets. Participated in the model development. Conducted the bioinformatics analyses and implemented the MIDESP tool. Participated in the interpretation of the results. Participated in writing the final version of the manuscripts.

The sequence data of *Vicia faba* used in this thesis has been submitted to the European Nucleotide Archive under the accession number PRJEB38838 and published in the following article:

[3] **Heinrich, F.**; Gültas, M.; Link, W.; Schmitt, A.O. (2020). *Genotyping by Sequencing Reads of 20 Vicia faba Lines with High and Low Vicine and Convicine Content.* Data, 5(3), 63.

Further, I contributed to the following publication that is related to the topic of this thesis:

[4] Klees S.*; **Heinrich F.***; Schmitt A.O.; Gültas M. (2021). *agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species.* Biology, 5(3), 790 (*These authors contributed equally to this work.).

Conferences, Workshops, Meetings and Student's thesis
I presented topics of this thesis at the following workshops and conferences:

- Oral presentation titled "agReg-SNPdb: A database of regulatory SNPs for agricultural species" at 2019 CiBreed Workshop organized at Georg-August University, Göttingen, Germany, 2019.
- Oral presentation titled "Identification of regulatory SNPs based upon genotyping by sequencing data in Vicia faba using deep learning" at THETA seminar series, Statistical Genetics Group, Institute of Animal Genetics, Wrocław University of Environmental and Life Sciences, Wrocław, Poland, 2020.

In collaboration with Mehmet Gültas and Armin O. Schmitt I supervised the following student work:

- Miriam Kamp: *Analyse von Kandidaten-SNPs und ihren assoziierten Genen in Vicia faba mittels bioinformatischer Methoden.* Bachelor's Thesis, 2020

In collaboration with Selina Klees and Mehmet Gültas, I further provide an online database for regulatory SNPs in several animal species (see Appendix A.4) that is available via https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb.

2. Biological background

In this chapter, I briefly introduce the biological concepts that are necessary to understand this thesis and the motivation behind the two analysis frameworks I developed. Additionally, I give an overview of the existing databases and tools that are used over the course of this thesis in the final part.

2.1. Genotype and phenotype

An individual can be characterized by using its observable and measurable attributes. These attributes are commonly referred to as traits while the specific expression of an attribute in an individual is called the phenotype [89]. As an example, one can consider immunity to a specific disease as a trait with resistance and susceptibility being the two possible phenotypes that an individual can have. Traits can be divided into qualitative and quantitative traits with the former having discrete, categorical expressions while the phenotypes of the latter are continuous values. The phenotype of an individual is a result of the combination of the individual's genotype as well as environmental effects. In this regard, the genotype refers to the total of all genes that determine the phenotype [89]. An important characteristic of the relationship between genotype and phenotype is the so-called heritability. It indicates how much of the phenotypical variance can be explained by the variance of the genotype [90]. A higher heritability would mean that the phenotype could be more accurately predicted using the genotype.
For this thesis, I use trait and phenotype interchangeably to refer to an observable characteristic.

2.2. Gene expression

In this section, I describe how the genetic information of an individual is stored and how its translation to functional proteins is regulated by transcription factors. If not indicated otherwise, the content of this section is based on [91, 92].

2.2.1. DNA

Deoxyribonucleic acid (DNA) is the conveyor of genetic information in an individual. It is built from nucleotides, which in turn are each composed of a deoxyribose sugar, a phos-

phate group and one of four possible nucleic acids. These acids, which are also referred to as bases, are adenine (A), cytosine (C), guanine (G), and thymine (T). Based on the structure of the nucleic acids, they are divided into two groups. Adenine and guanine belong to the purines while cytosine and thymine are pyrimidines [93]. Through phosphodiester bonds, which are created between the phosphate group attached to the 5'-hydroxyl of one nucleotide and the 3'-hydroxyl of another nucleotide, the nucleotides are joined to long chains. The orientation of the 5'-hydroxyl and the 3'-hydroxyl also allows one to define a direction for the nucleotide chain with the general convention being to write the sequence from the 5' end to the 3' end. Considering a specific position in the chain *upstream* refers to the region from the position towards the 5' end while *downstream* refers to the region towards the 3' end. DNA is formed by two of these polynucleotide chains, which are arranged in form of a double helix so that the backbones are on the outside while the nucleic bases face each other. Hydrogen bonds, which are formed between opposing bases, grant this structure its stability. Furthermore, these bonds are specific so that adenine binds to thymine and cytosine to guanine. The two chains are therefore complementary to each other, which enables the replication of the DNA from a single chain [94]. The structure of the DNA is shown in Figure 2.1.

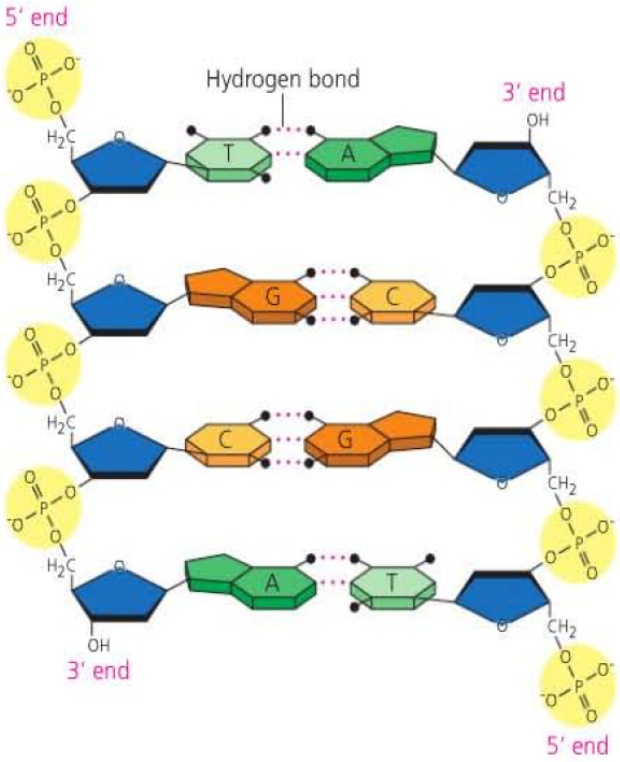


Figure 2.1.: The DNA consists of two complementary nucleotide chains. The nucleotides are linked by the phosphate groups (yellow), which are attached to the deoxyribose molecule (blue) of the nucleotides. Hydrogen bonds between the nucleic acids adenine and thymine as well as between cytosine and guanine connect the two chains. This figure is taken from [93], page 309.

2.2.2. Genes

Segments on the DNA sequence that contain the information on a functional molecule are termed genes. These genes contain the majority of the genetic information of an organism. The site of a gene in the DNA sequence is referred to as the locus of the gene [95]. To create a protein from the gene, the corresponding DNA sequence is first transcribed to a ribonucleic acid (RNA) sequence. In the second step, the RNA sequence is translated to an amino acid sequence, which folds into the protein. Similar to the DNA, the RNA is built from nucleotides. However, there are some differences between the two. RNA only has a single nucleotide chain, the sugar molecule is ribose and the nucleic acid thymine is replaced by the base uracil (U). Besides offering a template for the amino sequence, RNA can fulfill several other functions in the cell. These include for example acting as adaptors during the translation process, being part of complexes like ribosomes as well as having a

regulatory effect on the gene expression.
In eukaryotes, the genes are often constituted of several coding regions that are interrupted by non-coding regions (introns). Therefore, before translation, it is necessary to remove the introns and to join the coding regions together. This process is called splicing.
Due to mutations in the DNA sequence, alternative forms of a gene may be created if the corresponding amino sequence is changed. These different forms are known as the alleles of a gene [95]. An individual has two alleles for each gene locus with one of them being inherited from the father while the other is obtained from the mother. This specific allele set is termed the genotype of this gene for the individual. Based on the occurring alleles of a gene in an individual, a distinction can be made between a homozygous genotype (two copies of the same allele) and a heterozygous genotype (two different alleles).

2.2.3. Regulation of gene expression

Whether a gene is expressed in a specific cell at a specific time depends on multiple factors such as cell type, environmental responses as well as the function of the corresponding protein. Complex regulatory mechanisms are necessary to control the gene expression. These mechanisms encompass all steps of the gene expression starting with preventing the transcription itself by restricting the access to the corresponding region of the DNA sequence to modifications of the protein after it has been created through translation [95]. However, in this thesis, I focus on how transcription factors influence the process of transcription.

2.2.4. Regulation of transcription by transcription factors

The transcription starts with the binding of the RNA polymerase to the DNA close to the transcription start site (TSS) of a gene. This initial binding of the polymerase to the DNA requires the presence of several proteins termed general transcription factors (GTFs), which first bind to the DNA in a region called core promoter. The core promoter surrounds the TSS and is usually 40-60 nucleotides long. It contains several sequence elements that allow for the binding of GTFs to the DNA, however, its exact composition can differ from gene to gene [96]. The elements which occur relatively frequently in the core promoter are, among others, the TFIIB recognition element (BRE), the TATA element, the initiator element (Inr) as well as some elements downstream of the TSS such as the downstream promoter element (DPE), the downstream core element (DCE) and the motif ten element (MTE) [97]. It is not necessary for all elements to be present to initialize the transcription, and there are core promoters that do not contain any of those elements [98]. Figure 2.2 shows how a core promoter could exemplarily be structured.

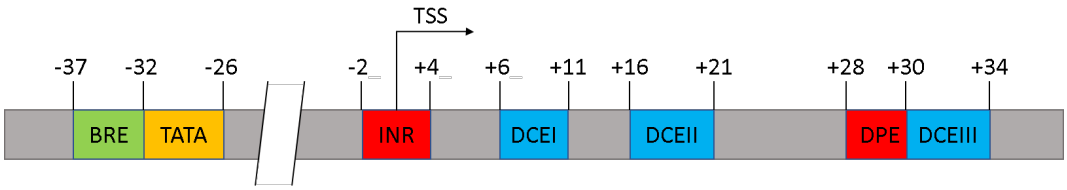


Figure 2.2.: The core promoter can contain several sequence elements, some of which are indicated here with their respective approximate start and end position relative to the TSS. This figure is based on [91], Figure 13-15.

In addition to this core promoter, there may be other regulatory elements around the TSS that are involved in the regulation of the gene expression. This greater region is termed the promoter of the gene and can encompass several 100 nucleotides upstream of the TSS, but also even more distant regions [41, 93, 99, 100].
Transcription factors (TFs) are a class of regulatory proteins that recognize and bind to these regulatory elements and, through their binding, influence the expression level of the gene. The binding of TFs can be necessary to start the transcription of a gene, but it can also completely repress the expression or control it at a more granular level [93]. Only a small fraction of TFs interact directly with the RNA polymerase. Instead, most TFs are involved only indirectly, by interacting with other proteins that act as bridges to the polymerase, or by modifying the structure of the DNA and thereby changing its accessibility [92]. The sequences to which the TFs bind (termed transcription factor binding sites (TFBSs)) are specific to the particular TF and are generally short, being in the range of 6 to 12 nucleotides [101]. It is often the case that specific positions in the TFBS are more constrained than others and require a certain nucleotide base at the corresponding position. Mutations in the DNA sequence at such positions could therefore prevent a TF from binding to the site [101]. Other positions still could be changed without affecting the ability of the TF to bind to the TFBS. To account for this variability in the analysis of TFBSs, position weight matrices (PWMs) can be used to represent them [102]. These matrices contain for each position of the respective TFBS the probability that one of the four possible DNA bases occurs at this position. A PWM can be built from an alignment of known DNA sequences for the TFBS. The process is exemplarily shown in Figure 2.3. A useful method for visualization is the logo plot where the size of the letters indicates how conserved a position is (see Figure 2.4) [103].

```
Sequence 1:     CTAATAACC
Sequence 2:     CTATAACTC
Sequence 3:     CTATAACTC
Sequence 4:     TTATAACTT
Sequence 5:     GTATAACCC
```

Pos	A	C	G	T
1	0	3	1	1
2	0	0	0	5
3	5	0	0	0
4	1	0	0	4
5	4	0	0	1
6	5	0	0	0
7	1	4	0	0
8	0	2	0	3
9	0	4	0	1

Figure 2.3.: Construction of a PWM for a TFBS based on an alignment of five nucleotide sequences of length nine bp. The numbers in the PWM give the counts for the frequency of each base at the corresponding position in the sequence alignment.

Pos	A	C	G	T
1	13	4	4	3
2	14	3	5	2
3	2	7	1	14
4	23	0	0	1
5	24	0	0	0
6	1	23	0	0
7	0	1	17	6
8	1	1	21	1
9	11	5	4	4
10	12	5	2	5

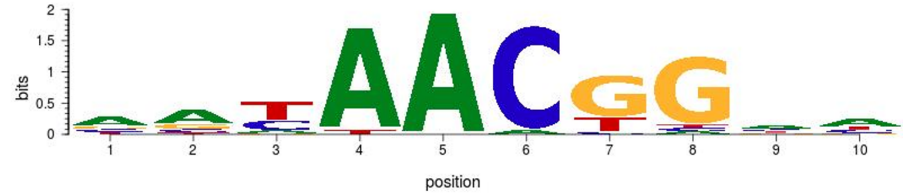


Figure 2.4.: PWM and logo plot (created using motifStack [104]) for the TFBS V$VMYB_01 [105]. The y-axis shows the conservation of each position measured in bits. The visualization indicates that the positions 4, 5 and 6 are strongly conserved while the start and end of the TFBS can be variable.

2.3. Single Nucleotide Polymorphisms

A multitude of different DNA-based molecular markers have been utilized in the last few decades to analyze the genome (see [106] for an overview). Single nucleotide polymorphisms (SNPs) have emerged as one of the most popular types of markers based on their abundance in the genome and the relative ease of their detection [107, 108]. A SNP refers to a change of the nucleotide base at a specific position in the genome, where the surrounding sequence is, however, conserved across individuals [109]. These variants are created through mutations, which are then passed on to the offspring. Similar to a gene locus, the alternative variants are termed the alleles of the SNP. Likewise, the allele set is the genotype of this SNP for the individual and one can define heterozygous and homozygous genotypes for the SNP. To distinguish SNPs from mutations which are specific to a single or few individuals, it is usually required that the least abundant allele of a SNP in a given population has a frequency of 1% or more among all individuals in that population [110]. While it is possible for a SNP to have more than two different alleles in a population, most commonly used SNPs in genome analyses only have two different alleles and are therefore also referred to as bi-allelic markers [109]. For a SNP, the allele, which corresponds to the base in the reference genome sequence of the species, is usually termed the reference allele, while the other one is the alternate allele.

2.3.1. Regulatory SNPs

Regulatory SNPs (rSNPs) are a subset of all SNPs in a genome and refer to those SNPs, that are located in the regulatory regions. These rSNPs can affect the gene expression by altering regulatory elements such as the previously introduced TFBSs [111, 112, 113]. For example, a change in the nucleotide base at a highly conserved position of a TFBS can be sufficient to disrupt it and prevent TFs from binding there. Likewise, however, it is also possible that a new binding site could be created if the nucleotide is substituted. rSNPs can also interact with other types of regulatory elements, such as binding sites for non-coding RNA sequences [114, 115]. However, for this thesis I only consider the impact of rSNPs on TFBSs. Following previous studies [35, 52, 111], I differentiate four types of consequences that a SNP may have on a TFBS: (i) no effect, (ii) change in binding affinity, (iii) loss of TFBS (a TFBS appears only for the reference allele) and (iv) gain of TFBS (a TFBS appears only for the alternate allele). These different consequences are visualized in Figure 2.5. Furthermore, a single SNP can effect multiple TFBSs with different consequences for each.

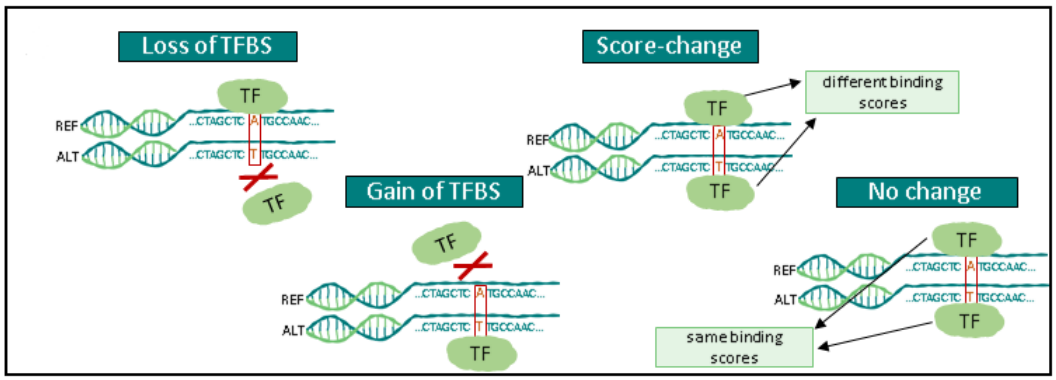


Figure 2.5.: Depiction of the four possible consequences that a SNP can have on a TFBS. This figure is taken from [35].

2.3.2. Genotyping methods

The process of determining the genotype of a SNP for an individual is termed genotyping. If the SNPs and their surrounding sequences are already known, one commonly used approach are microarrays (also known as DNA chips). These arrays are based on DNA hybridization and allow the simultaneous genotyping of several hundred thousand SNPs [108]. For each allele of a SNP, the array contains a probe which is reverse complementary to the DNA sequence around the SNP. The probes for a SNP are identical except for the location of the SNP itself, which corresponds to the specific allele. Thereby, only sequences with the corresponding allele are able to bind to the probe [108]. By using a suitable length for the probes, the probability of another DNA sequence in the genome being able to bind to them is negligible [116]. Furthermore, the probes are labelled with an allele-specific fluorescent dye. After the hybridization step, the resulting color can be analyzed to determine whether the individual has for this SNP one of the two homozygous genotypes (only a single color is visible) or a heterozygous genotype (both colors are visible) [117]. Finally, automated algorithms are utilized to determine the genotype of the SNPs based on the dye intensities [116].

An alternative strategy for SNP genotyping is the genotyping-by-sequencing (GBS) approach [3]. In this approach, the discovery of novel SNPs and genotyping occur at the same time [118]. As a first step, the genome of an individual is digested using a restriction enzyme. This enzyme cleaves the DNA at specific sites, which results in a multitude of short fragments. After that, the ends of these fragments are sequenced. Thereby, sequence reads of the genome can be obtained, which are in general a few hundred nucleotides long. The reads of each individual are then aligned against the reference genome sequence of the species to find SNPs and their genotypes [3, 119]. For species without a known reference genome, the reads are first used to assemble a partial genome, which can be used for the

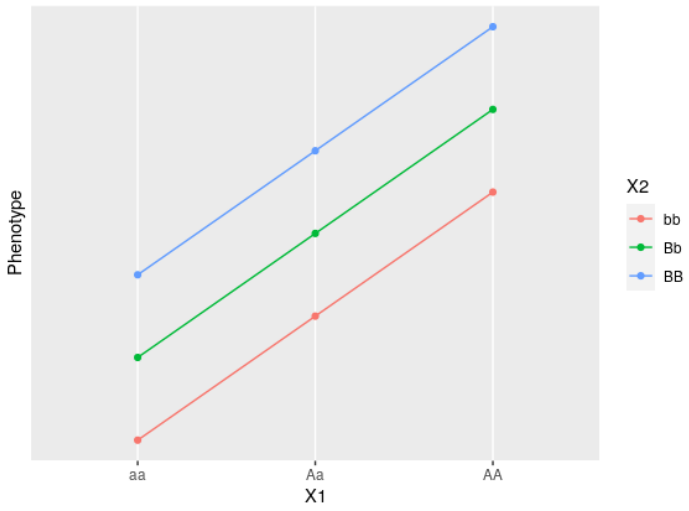
alignment [118, 120]. A particular advantage of GBS is its ability to obtain sequence data from regulatory regions [3]. Correspondingly, it is thereby possible to discover regulatory SNPs.

2.4. Epistasis

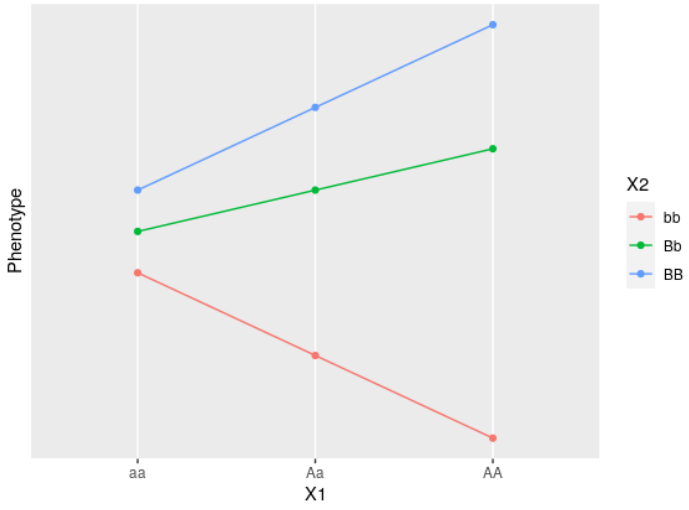
Epistasis, at its core, refers to the interactions between two or more genes (represented, for example, by SNPs), which jointly influence a phenotype. These interactions can contribute to the main/additive effects of the genes (i.e. association between a single gene and the phenotype independently of other genes) but they can also exist without any additive effects [121]. Therefore, by considering epistasis, it is possible to find genes associated with the phenotype that would otherwise be missed due to an insufficient main effect [58, 59]. Furthermore, the knowledge of these interactions can help us to understand the complex biological systems that are responsible for the expression of a trait [65]. One possible biological explanation for the existence of epistasis is the concept of canalization, which was first proposed by Waddington [122]. The idea is that evolution seeks the development of robust systems, which are buffered against genetic and environmental variation [57]. This can be observed in gene networks with redundant pathways, where a change of the phenotype requires the accumulation of several mutations. Each individual variant would then have only a small effect on the phenotype by itself [121].
The term epistasis was first introduced by Bateson in 1909 to refer to the phenomenon in which the effect of one allele at locus A is masked by the effect of another allele at locus B [121, 123]. A few years later, Fisher appropriated the term to describe statistical deviations from an additive model of multiple loci [124], which is nowadays the common usage and the basis for the computational detection of epistasis [53, 65, 121]. In this sense, the effect of one locus may depend on the genotype of another locus [53, 65], which is shown in the following example.

Example: Epistasis between two genes

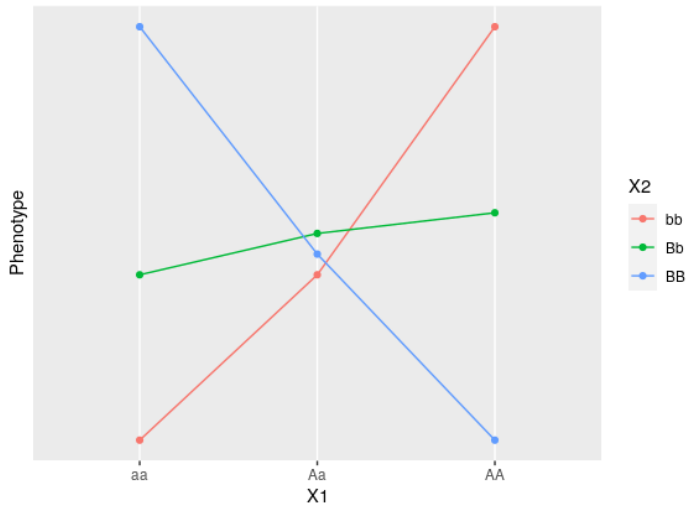
X1 and X2 are two genes with three possible genotypes each. X1 has the genotypes aa, Aa and AA, while the genotypes of X2 are bb, Bb and BB. Together, the genes control the expression of a phenotype. Based on Figure 1 in [65], I present the following visualisations for three different scenarios of how the interaction of the two genes influences the phenotype.



In the first scenario, there is no epistasis between the genes. Both of them influence the phenotype independently of each other. Regardless of the genotype of X1, the effect of X2 is the same and vice versa.



For the second scenario, however, it can be observed that the difference in the phenotype between the genotypes of X2 depends on the genotype of X1. While the difference between bb and BB is relatively small for aa, their difference for AA is much greater. In this case, the genotype of X1 increases the effect of X2.



The last scenario also shows epistasis between the genes. Here, the direction of the effect of bb and BB depends on the genotype of X1. If X1 has the genotype aa, then bb results in a small and BB in a large phenotype. However, if X1 is instead AA then this effect is reversed and bb results in a large phenotype while BB has a small phenotype.

2.5. Bioinformatic resources

In the last decades, the available amount of genomic data has increased tremendously. To facilitate the storage and exchange of data in an organized manner, numerous databases have been established. Similarly, novel tools and algorithms have been published for the analysis. In this section, I shortly present the bioinformatics databases and tools that I utilized for this thesis.

2.5.1. Bioinformatic databases

2.5.1.1. Ensembl database

The Ensembl Project database (`https://www.ensembl.org`) is one of the most known and comprehensive resources for publicly available genomics data [125, 126]. It combines primary genomic information such as genome sequences with several automated pipelines and tools for annotation and analysis. The various annotations include, for example, gene models, sequence variations (SNPs, insertions, deletions and structural variants) and regulatory features. Furthermore, the website offers the users different possibilities to query the available data, for instance by doing cross-species comparison and search of sequences as well as the prediction of variant effects [127, 128]. Figure 2.6 gives an overview of the information available for a species (here *Bos taurus*). Since the project was started

in 2000 to annotate the human genome, it has expanded to include more than 300 vertebrates as well as species belonging to different domains of life, for example plants (https://plants.ensembl.org) or bacteria (https://bacteria.ensembl.org) [129].

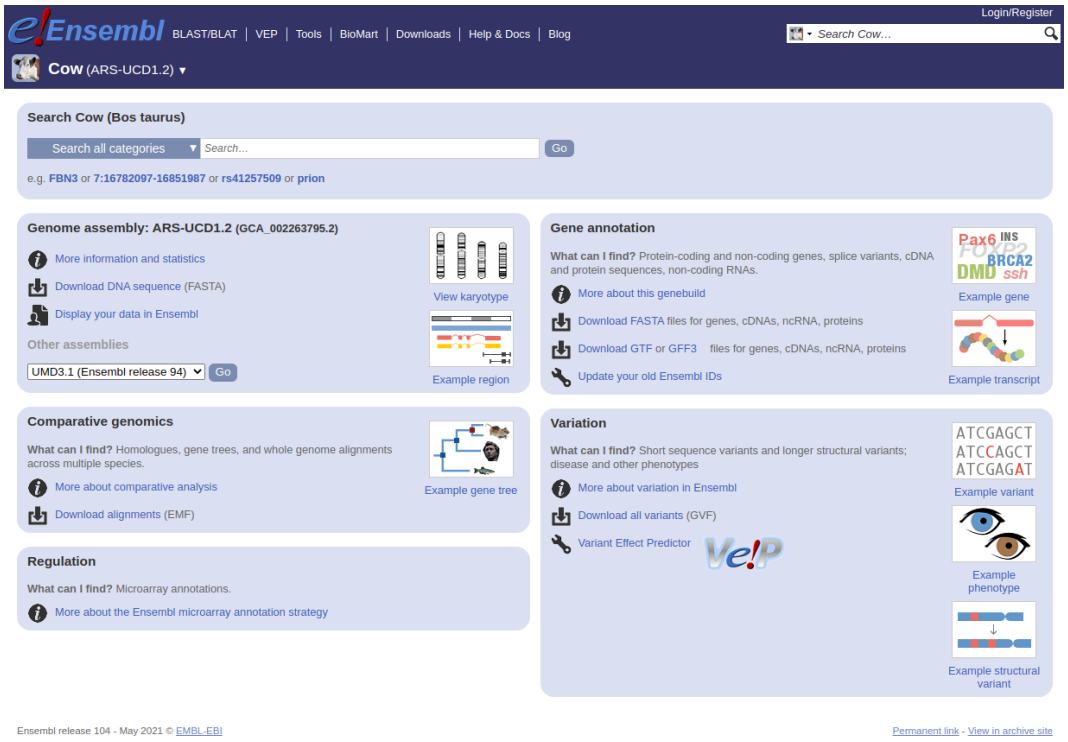


Figure 2.6.: Overview page for the species *Bos taurus* in the Ensembl database. This page includes links to the various types of information available for the species. (Source: https://www.ensembl.org/Bos_taurus/Info/Index, 11.11.2021)

2.5.1.2. Pulse Crop Database

The Pulse Crop Database (PCD) (https://www.pulsedb.org) is a public database that collects genomic, genetic and breeding resources for more than 40 species belonging to the group of pulse crops [130, 131]. It connects annotated reference genomes and transcriptomes with marker and phenotype data, thereby allowing users, for example, to identify genes related to traits of interest. The data can be downloaded as well as analyzed and visualized directly on the website using various tools.

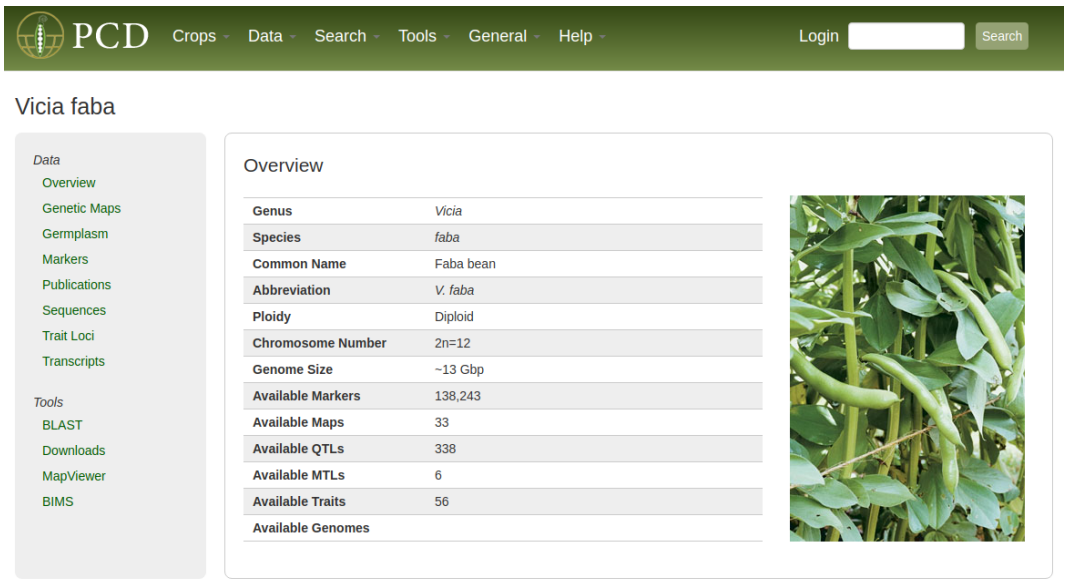


Figure 2.7.: Overview page for the species *Vicia faba* in the Pulse Crop Database. It contains basic information and a summary of the available data. (Source: https://www.pulsedb.org/bio_data/642, 11.11.2021)

2.5.1.3. TRANSFAC®

TRANSFAC®, which was firstly published in 1988 [132], is a database for eukaryotic transcription factors (TFs). The database stores information about the TFs, their binding sites and the genes that are regulated by them. Experimentally verified DNA binding sites are used to build position weight matrices (PWMs), which can then be used for computational prediction [133]. It further provides a hierarchical classification structure consisting of six levels for the TFs, which is based on their DNA-binding domains [105].

2.5.2. Bioinformatic tools

2.5.2.1. PLINK

PLINK is a popular and widely used tool set for the analysis of genomic data in the form of SNP markers [134, 135]. It uses a binary file format as an efficient way to represent SNP data and offers various possibilities to reorder, recode and filter the information according to different criteria. Furthermore, it can convert the data to different text-based formats, which are often used by other programs to read in SNP marker data. In this thesis, I mainly utilize these data management aspects of PLINK to filter and prepare the datasets for analysis. Additionally, I apply PLINK to perform tests for the analysis of association between SNP

markers and the phenotype. This includes tests for interactions between two SNPs and the phenotype (the previously introduced epistasis).

2.5.2.2. MATCH™

MATCH™ is a tool for the computational prediction of putative transcription factor binding sites (TFBSs) [136]. The program takes DNA sequences as input and screens them for TFBSs using a library of PWMs. At each possible position of the sequences, two scores are calculated, that indicate how good the match with a PWM is: (i) the matrix similarity score (MSS) and (ii) the core similarity score (CSS). Whereas the MSS uses the complete length L of the PWM, the CSS compares only the first five most conserved consecutive positions of the PWM which define its core. The MSS and CSS are calculated in a similar way for a DNA sequence as

$$SS = \frac{\text{Current} - \text{Min}}{\text{Max} - \text{Min}} \tag{2.5.1}$$

where SS stands for MSS or CSS. Equation 2.5.1 uses the following values:

$$\text{Current} = \sum_{i=1}^{L} I(i) \cdot f_{i,B_i} \tag{2.5.2}$$

$$\text{Min} = \sum_{i=1}^{L} I(i) \cdot f_i^{\min} \tag{2.5.3}$$

$$\text{Max} = \sum_{i=1}^{L} I(i) \cdot f_i^{\max}. \tag{2.5.4}$$

In above equations f_{i,B_i} is the frequency of the nucleotide B_i at position i of the PWM, $f_i^{\min}$ is the lowest frequency at position i of the PWM and $f_i^{\max}$ is the highest frequency at position i of the PWM. Furthermore, $I(i)$ is the information vector defined as

$$I(i) = \sum_{B \in \{A,C,G,T\}} f_{i,B} \cdot \ln 4 \cdot f_{i,B} \text{ , i = 1,2,...,L.} \tag{2.5.5}$$

The information vector describes the conservation of position i in the PWM [137]. By multiplying by $I(i)$, the mismatches in highly conserved regions of the PWM will inflict a stronger punishment on the score compared to regions that are less conserved. This results in an improved prediction accuracy [138].
Both MSS and CSS range between 0 and 1 with the latter indicating a perfect match between sequence and PWM. In order to decide if a TFBS exists, the program uses pre-specified cut-off values that are specific for each PWM. Such cut-off profiles are supplied, for example, by TRANSFAC® [105]. Finally, if the scores for a match exceed the cut-off values of the PWM, the corresponding information is added to the output of MATCH™ (see Figure 2.8).

```
Search for sites by WeightMatrix library: Match_original/matrix.dat
Sequence file: Match_original/Breast_cancer_gene_set_HGNC.fasta
Site selection profile: Match_original/Match/minFP.prf


Inspecting sequence ID   NM_001079874_VAV3_chr1_-_range=107688504:107689504_(5'-3')TSS_107688504

 I$HSF_01              |        2 (+) |  1.000 |  1.000 | AGAAA
 I$HSF_01              |        7 (+) |  1.000 |  1.000 | AGAAA
 I$HSF_01              |      142 (+) |  1.000 |  1.000 | AGAAA
 I$HSF_01              |      178 (-) |  1.000 |  1.000 | TTTCT
 I$HSF_01              |      500 (-) |  1.000 |  1.000 | TTTCT
 I$HSF_01              |      526 (+) |  1.000 |  1.000 | AGAAA
```

Figure 2.8.: The output of MATCH^{TM} consists of five columns. The first column contains the identifier of the PWM for which the match was found. This is followed by the start position of the match on the DNA sequence as well as a plus or minus indicating the strand. Columns three and four contain the CSS and MSS, respectively. The last column contains the matching sequence.

2.5.2.3. GeneXplain platform

GeneXplain (https://genexplain.com/) is an online platform for the computational analysis of biological data, in particular transcriptomic data [139, 140]. It provides a wide range of statistical, bioinformatics as well as systems biology functions and combines them with several biological databases including the previously mentioned TRANSFAC® database. In addition to several predefined workflows, such as gene set enrichment analysis or master regulator search, users have the option to create their own custom analysis pipelines.

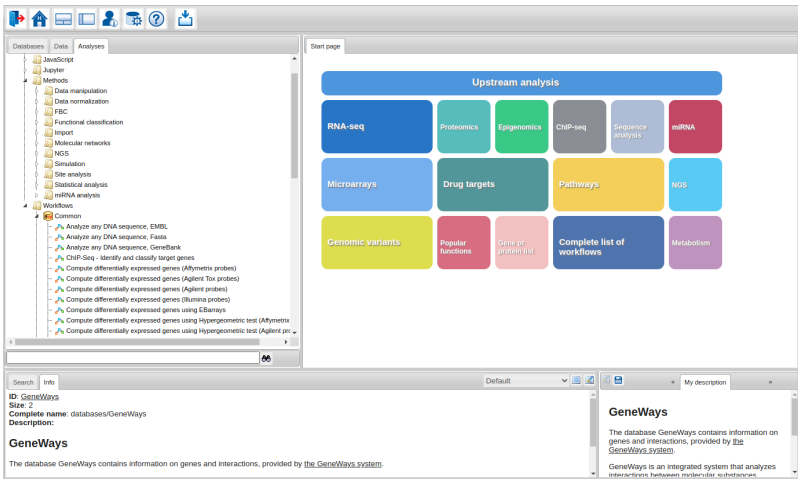


Figure 2.9.: Start page of the GeneXplain platform. (Source: https://platform.genexplain.com/bioumlweb/, 11.11.2021)

3. Theoretical background

In this section, I introduce the concepts and methods that were utilized in this thesis. To begin, I provide an overview of the techniques for promoter detection with a more detailed description of convolutional neural networks. In the second part, I give a short introduction to information theory, in particular mutual information and other related measures. Finally, I present the basic concepts of genotype-phenotype association studies, with an emphasis on the detection of epistatic interactions.

3.1. Prediction of promoter sequences

This section is mainly based on my publication [52] (see Appendix A.2) and the method overview given in [48]. The subchapter dealing with convolutional neural networks includes content from [141]. Due to their varied and complex structure, the computational identification of promoter regions is still a challenging task. The methods that are used for their prediction are mostly binary classification-based machine learning (ML) algorithms. By training them on sequence data from known promoter and non-promoter sequences, these methods are able to decide whether a new sequence is a promoter or not. However, there exists no single approach for choosing the positive and negative datasets for training. While for promoters an interval around the transcription start site (TSS) is chosen, which often differs only in the exact length and the position, the choice of the negative dataset is more diverse. The negative set can be built from sequences that correspond to a certain promoter sequence in the positive set (for example a specific region from the first exon of the gene) [46], sequences randomly taken from coding regions [40, 43, 44] as well as intergenic regions [42, 49] or by randomizing promoter sequences [48]. Moreover, the classifiers mostly differ from each other by the ML architecture that is used (support vector machines, random forests, neural networks and others) and the specific sequence features that are taken as input for the classifier. These features range from the existence and position of highly conserved sequence motifs (such as TATA boxes, initiator elements (Inrs), downstream promoter elements (DPE), transcription factor binding sites (TFBSs) and others) to more content-based features like k-mer distributions or a combination of both. PromoterScan [142], for example, utilizes a combination of TATA box and TFBS detection as input for a linear discriminator to detect promoter sequences. To address the issue of variable spacing between sequence motifs, NNPP [143] applies a time-delay neural network, but only considers the TATA box and Inr as features. PromFind [144], on the other hand, selects the 6-mers that

distinguish most strongly between promoter and non-coding as well as coding sequences from the training data to classify novel sequences. Similarly, PromMachine [43] and PromoBot [40] both use a selection of 4-mers and 6-mers, respectively, as input for an SVM to distinguish between promoter and non-promoter sequences. Finally, both TSSP-TCM [37] and TSSPlant [51] use a multitude of motif- and content-based features together as input for an SVM respectively neural network to classify sequences.
However, with the ever increasing knowledge about the diversity of promoter structures, the methods have moved on from the identification of specific promoter elements. Methods that rely on the assumed (near) universal existence of a few specific elements in the promoter regions are no longer suitable, since this assumption has been disproved [45, 48]. For example, it has been shown that the TATA box does not exist in many promoter sequences, so that algorithms are sometimes even trained separately for promoters containing a TATA box or not [37, 44, 48, 51]. On the other hand, the TATA box is not exclusive to promoter regions and can be found in abundance in non-promoter datasets [145]. Therefore, a classifier that relies solely on the presence of the TATA box to differentiate the sequences is easily prone to make false classifications [48].
Instead, the focus has changed to more holistic approaches that are able to take the whole genomic region and its spatial structure into account. These include, for example, methods that utilize differences in the free energy patterns of the DNA sequence between promoter and non-promoter regions, such as PromPredict [39]. Similarly, Li et al. [146] apply a generalized topological entropy to measure the complexity of DNA sequences and observed differences between promoter, exon and intron regions. In a similar vein, iProEP [41] considers the composition of pseudo k-mer nucleotides with a position-correlation scoring function as input for an SVM to differentiate promoter and non-promoter sequences.
Nowadays, convolutional neural networks constitute a major fraction of promoter prediction methods [44, 45, 46, 47, 48, 49].

3.1.1. Detection of promoters using CNNs

Convolutional neural networks (CNNs) are a special type of neural network, which are especially suited for analysing data with grid-like structures such as time-series or image data [141]. In these areas, they achieve state-of-the-art performances [147, 148, 149]. They have also been successfully used for various biological tasks, for example, the prediction of genes [150], sequence binding sites [151] or the effects of noncoding variants [152]. CNNs are characterized by having at least one so called convolutional layer, which is used to process the input. The convolution can be understood as calculating a weighted average. Within a convolutional layer, an array of stacked weight matrices (also referred to as filters) of dimension $W \times H \times D$, where W and H correspond to the width and height of the matrices, respectively, while D refers to the depth of the array, is moved spatially across the input data [44, 47]. The length of each step of the weight matrices is called stride. A stride of two, for example, would mean that the weight matrix moves two positions over the input

with each step. At every possible position, the summed element-wise product between the weight matrices and a subset of the input is calculated and a corresponding feature map is computed. The feature maps of all weight matrices are then taken as input for the next layer. Figure 3.1 visualizes this process for a single weight matrix.

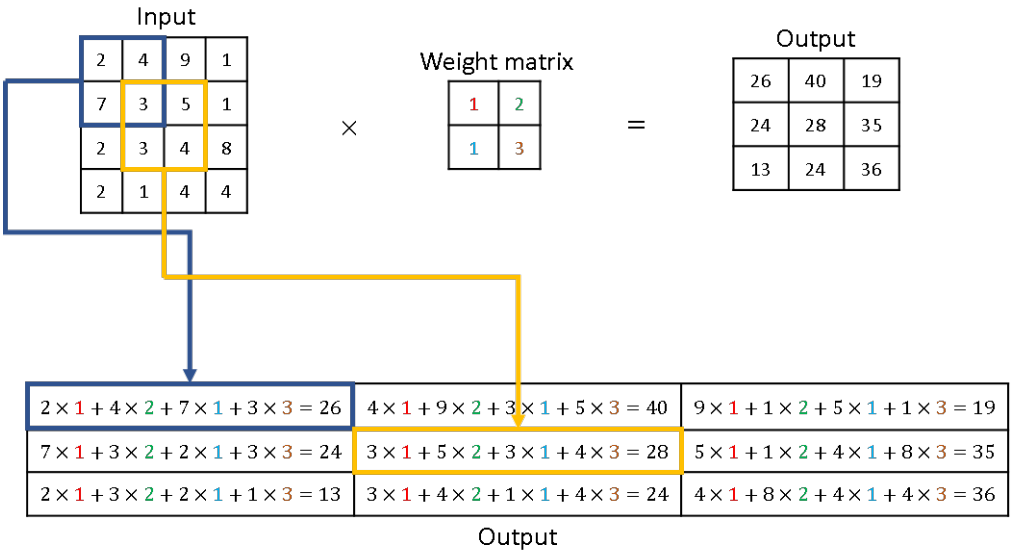


Figure 3.1.: An example for a convolution with a single 2×2 weight matrix. Input is a 4×4 2-dimensional array and output is a 3×3 2-dimensional array. As an example, the colored borders indicate which part of the input array resulted in the corresponding value of the output.

Another important aspect of convolutional layers is *zero padding*. As shown in Figure 3.1, the output of the convolution will be smaller than the input matrix if the weight matrix is not a simple scalar. If the input has a width of m and the weight matrix a width of k, then the output matrix will have a width of $m-k+1$. Zero padding prevents this shrinkage by adding additional cells with 0 to the borders of the input such that the output will be of the same dimensions as the original input matrix.

To process DNA sequences with a CNN, the 1-dimensional sequence is first transformed into a 2-dimensional array, where each nucleotide is encoded into a one-hot representation and expressed by a four-dimensional vector (see Figure 3.2).

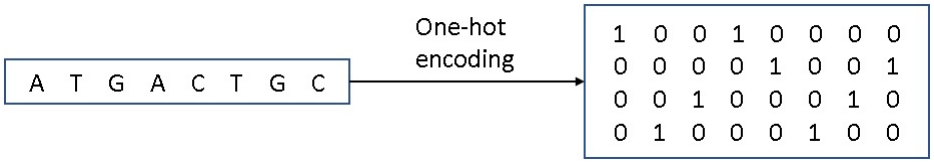


Figure 3.2.: One-hot encoding of a DNA sequence. Each nucleotide is represented by a one-hot vector: A = $(1,0,0,0)$, C = $(0,1,0,0)$, G = $(0,0,1,0)$, and T = $(0,0,0,1)$.

A major advantage of CNNs is their ability to automatically learn local as well as global features from the input data [153]. Therefore, no previous domain-knowledge or assumptions regarding the importance of features is necessary.
Umarov et al. used a CNN in their CNNProm program for promoter prediction and achieved a significantly higher accuracy than previous methods [44]. With DeeRecCT-PromID, a promoter predictor, which builds upon CNNProm and includes an iterative training scheme, Umarov et al. found that features learned by the CNN correspond to well-known characteristics of promoter sequences [45]. Nevertheless, it can still be possible to improve the performance of a CNN using specific sequence features. For example, Triska et al. showed that including certain additional features can improve the performance of the CNN compared to a baseline model, which utilizes only the sequence as input [46]. In a similar vein, Qian et al. have demonstrated that the separation of the sequence into important elements and unimportant sections followed by analyzing the former in more depth can improve the performance [47].

3.2. Information theory

The information theory was first developed by Claude Shannon in the 1940s for the analysis of communication systems [154]. He discovered that there exists a lower bound for the length that is necessary to describe a random variable on average. This measure, which he termed as entropy, can be interpreted as the amount of information that is inherent to the variable. Based on this basic quantity, a great number of different measures have been developed to analyze the relationship between multiple variables in different ways, several of which I present here. The content and notations for this section are mostly taken from [155] and [75].

3.2.1. Entropy

The entropy of a discrete random variable X, which takes on values from an alphabet $\mathcal{X}$, depends only on its probability distribution $p(x)$ and measures the average uncertainty of X.

Definition 3.1 (Entropy) *The entropy $\mathbb{H}(X)$ of a discrete random variable X with a probability distribution $p(x) = Pr\{X = x\}$, where the distribution satisfies $0 \leq p(x) \leq 1$ and $\sum_{x\in\mathfrak{X}} p(x) = 1$, is defined as*

$$\mathbb{H}(X) = -\sum_{x\in\mathfrak{X}} p(x)\log p(x). \tag{3.2.1}$$

The base of the logarithm used in this study is 2 and, therefore, log stands for $\log_2$. In this case, the unit of the entropy is the bit. It is convention that $0\log 0 = 0$. Thus, additional terms with zero probability do not change the entropy. The entropy of X is 0 if and only if X has a fixed value x with $p(x) = 1$ and, consequently, no uncertainty exists. On the other hand, the maximum entropy is obtained only if $p(x)$ is a uniform distribution over $\mathfrak{X}$ and is equal to $\log|\mathfrak{X}|$.

Example: Calculation of entropy

Let X be a discrete random variable with the following outcomes:

$$AA \;\; AA \;\; AC \;\; CC \;\; CC \;\; AA \;\; AA \;\; AC \;\; AA \;\; AC$$

For the calculation of the entropy, it is assumed that the variable X has an alphabet $\mathfrak{X} = \{\mathrm{AA}, \mathrm{AC}, \mathrm{CC}\}$. As a first step, the marginal probabilities of the possible outcomes are estimated using their relative frequencies among the observations:

$$p(AA) = \tfrac{5}{10} \qquad p(AC) = \tfrac{3}{10} \qquad p(CC) = \tfrac{2}{10}$$

The entropy $\mathbb{H}(X)$ of X is then calculated as:

$$\begin{aligned}
\mathbb{H}(X) &= -\textstyle\sum_{i=1}^{3} p(x_i)\log(p(x_i)) \\
&= -(p(AA)\log(p(AA)) + p(AC)\log(p(AC)) + p(CC)\log(p(CC))) \\
&= -(\tfrac{5}{10}\cdot\log(\tfrac{5}{10}) + \tfrac{3}{10}\cdot\log(\tfrac{3}{10}) + \tfrac{2}{10}\cdot\log(\tfrac{2}{10})) \\
&\approx -((-0.5) + (-0.5210897) + (-0.4643856)) \\
&\approx -(-1.485475) \\
&\approx 1.485475 \text{ bits}
\end{aligned}$$

By extending the notion of entropy to two or more random variables, their joint entropy can be defined. Given discrete random variables X and Y with alphabets $\mathfrak{X}$ and $\mathfrak{Y}$, respectively,

their joint entropy depends only on their joint probability distribution $p(x,y)$ with $\mathfrak{X} \times \mathfrak{Y}$ as the alphabet of the possible value pairs.

Definition 3.2 (Joint Entropy) *The joint entropy $\mathbb{H}(X,Y)$ of two discrete random variables X and Y with a joint probability distribution $p(x,y)$ is defined as*

$$\mathbb{H}(X,Y) = -\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}} p(x,y)\log p(x,y). \tag{3.2.2}$$

A related measure is the conditional entropy of Y given X, which describes the average uncertainty of Y that remains when the value of X is known.

Definition 3.3 (Conditional Entropy) *The conditional entropy $\mathbb{H}(Y|X)$ of two discrete random variables X and Y with a joint probability distribution $p(x,y)$ and a conditional probability distribution $p(y|x)$ is defined as*

$$\mathbb{H}(Y|X) = -\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}} p(x,y)\log p(y|x). \tag{3.2.3}$$

The relation between the three introduced measures is shown in the following chain rule.

Theorem 3.1 (Chain rule for entropy) *The joint entropy $\mathbb{H}(X,Y)$ of two discrete random variables X and Y can be defined in terms of the entropy and the conditional entropy of these variables as follows*

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X) \;\; or \;\; \mathbb{H}(Y,X) = \mathbb{H}(Y) + \mathbb{H}(X|Y). \tag{3.2.4}$$

This relationship is also visualized in the Venn diagram depicted in Figure 3.3. The following properties hold for the introduced measures:

- $\mathbb{H}(X) \geq 0$
- $\mathbb{H}(X) \leq \log|\mathfrak{X}|$
- $\mathbb{H}(X,Y) = \mathbb{H}(Y,X)$
- $\mathbb{H}(X,Y) \leq \mathbb{H}(X) + \mathbb{H}(Y)$ with equality if and only if X and Y are independent
- $\mathbb{H}(X,Y) \geq \max\{\mathbb{H}(X), \mathbb{H}(Y)\}$ with equality if and only if one variable is a function of the other variable
- $\mathbb{H}(X|Y) \leq \mathbb{H}(X)$ with equality if and only if X and Y are independent

3.2.2. Mutual Information

The mutual information of two discrete random variables X and Y is a measure for the amount of information that is shared between these two variables, i.e., how much information X contains about Y and vice versa.

Definition 3.4 (Mutual Information) *The mutual information between two discrete random variables X and Y with marginal probability distributions $p(x)$ and $p(y)$, respectively, as well as a joint probability distribution $p(x,y)$ is defined as*

$$\mathbb{MI}(X;Y) = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{3.2.5}$$

The relation between the previously introduced entropy measures and the mutual information is shown in Figure 3.3, where the mutual information $\mathbb{MI}(X;Y)$ corresponds to the intersection of the entropy of X with the entropy of Y. The mutual information can be rewritten in terms of entropy as

$$\mathbb{MI}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) = \mathbb{MI}(Y;X). \tag{3.2.6}$$

Based on these formulas, it can also be interpreted as the reduction of the uncertainty of X due to the knowledge of Y and vice versa.

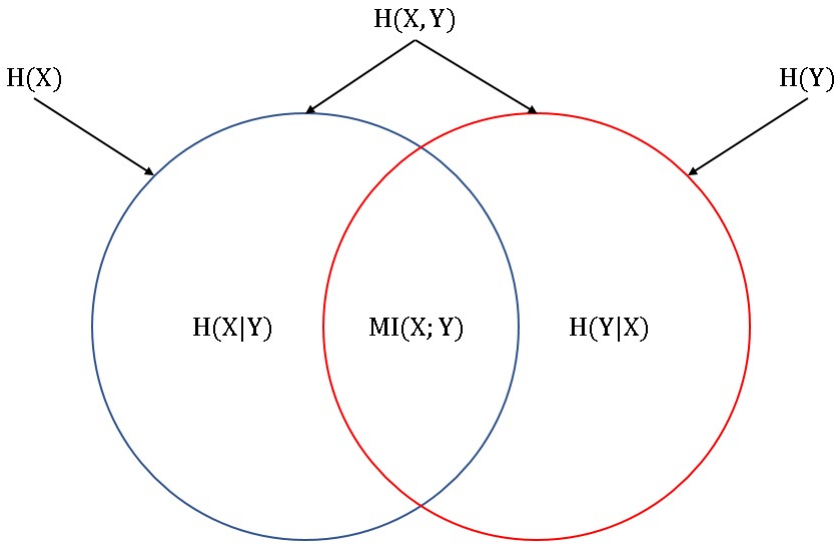


Figure 3.3.: Relationship between entropy and mutual information for two variables X and Y.

The following properties hold for the mutual information:

- $\mathbb{MI}(X;Y) \geq 0$ with equality if and only if X and Y are independent
- $\mathbb{MI}(X;Y) \leq \min\{\mathbb{H}(X),\mathbb{H}(Y)\}$ with equality if and only if one variable is a function of the other variable

- $\mathbb{MI}(X;X) = \mathbb{H}(X)$
- $\mathbb{MI}(X;Y) = \mathbb{MI}(Y;X)$
- $\mathbb{MI}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$

Example: Calculation of mutual information

Let X and Y be two discrete random variables with the following outcomes:

X:	AA	AA	AC	CC	CC	AA	AA	AC	AA	AC
Y:	1	1	1	0	0	1	1	1	1	0

To calculate the mutual information, first the marginal probabilities as well as the joint probabilities of the pair occurrences are determined based on the relative frequencies among the observations:

Marginal probabilities of X: Joint probabilities:

$$p(\text{AA}) = \tfrac{5}{10} \qquad p(\text{AA},0) = 0 \qquad p(\text{AA},1) = \tfrac{5}{10}$$

$$p(\text{AC}) = \tfrac{3}{10} \qquad p(\text{AC},0) = \tfrac{1}{10} \qquad p(\text{AC},1) = \tfrac{2}{10}$$

$$p(\text{CC}) = \tfrac{2}{10} \qquad p(\text{CC},0) = \tfrac{2}{10} \qquad p(\text{CC},1) = 0$$

Marginal probabilities of Y:

$$p(0) = \tfrac{3}{10}$$

$$p(1) = \tfrac{7}{10}$$

The mutual information $\mathbb{MI}(X,Y)$ between X and Y is then calculated as follows:

$$
\begin{aligned}
\mathbb{MI}(X;Y) &= \textstyle\sum_{i=1}^{3}\sum_{j=1}^{2} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)} \\
&= p(AA,0)\log\tfrac{p(AA,0)}{p(AA)p(0)} + p(AA,1)\log\tfrac{p(AA,1)}{p(AA)p(1)} + p(AC,0)\log\tfrac{p(AC,0)}{p(AC)p(0)} \\
&\quad + p(AC,1)\log\tfrac{p(AC,1)}{p(AC)p(1)} + p(CC,0)\log\tfrac{p(CC,0)}{p(CC)p(0)} + p(CC,1)\log\tfrac{p(CC,1)}{p(CC)p(1)} \\
&= 0 + \tfrac{5}{10}\cdot\log(\tfrac{5/10}{5/10\cdot 7/10}) + \tfrac{1}{10}\cdot\log(\tfrac{1/10}{3/10\cdot 3/10}) + \tfrac{2}{10}\cdot\log(\tfrac{2/10}{3/10\cdot 7/10}) \\
&\quad + \tfrac{2}{10}\cdot\log(\tfrac{2/10}{2/10\cdot 3/10}) + 0 \\
&\approx 0.2572866 + 0.0152003 + (-0.01407787) + 0.3473931 \\
&\approx 0.6058021
\end{aligned}
$$

3.2.3. Multivariate Mutual Information

The introduced formulas can be generalized to an arbitrary number of variables, which allows the analysis of more complex relationships. For illustration, I present several different information theoretic measures for the case of three discrete random variables X, Y and Z. The joint mutual information $\mathbb{MI}(X,Y;Z)$ is equivalent to the mutual information $\mathbb{MI}(S;Z)$ where S represents the grouping of the two variables X and Y. The grouping can be interpreted as the concatenation of the respective values of X and Y. $\mathbb{MI}(X,Y;Z)$ measures how much information X and Y together contain about Z.

Definition 3.5 (Joint Mutual Information) *The joint mutual information of a pair of discrete random variables X and Y with a third discrete random variable Z is defined as*

$$\mathbb{MI}(X,Y;Z) = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}}\sum_{z\in\mathfrak{Z}} p(x,y,z)\log\frac{p(x,y,z)}{p(x,y)p(z)}. \tag{3.2.7}$$

The joint mutual information can also be expressed as

$$\mathbb{MI}(X,Y;Z) = \mathbb{H}(X,Y) + \mathbb{H}(Z) - \mathbb{H}(X,Y,Z). \tag{3.2.8}$$

It is bound by $\max\{\mathbb{MI}(X;Z),\mathbb{MI}(Y;Z)\} \leq \mathbb{MI}(X,Y;Z) \leq \min\{\mathbb{H}(X,Y),\mathbb{H}(Z)\}$. The additional knowledge of Y cannot, therefore, decrease the information that is already available about Z due to the knowledge of X (and vice versa). The Venn diagram in Figure 3.4 a) shows the relation between the joint mutual information and the entropy values of the three variables. However, this visualization is no longer accurate in all situations, since the com-

bination of multiple variables together can result in the generation of novel knowledge, as the following example shows.

Example: Comparison of mutual information and joint mutual information

Let X, Y and Z be three discrete random variables with the following outcomes:

X:	A	C	A	C	A	C	A	C
Y:	G	T	G	T	T	G	T	G
Z:	1	1	1	1	0	0	0	0

Further, the grouping S of the variables X and Y is defined by combining their respective values.

S:	AG	CT	AG	CT	AT	CG	AT	CG

It can easily be seen that the probability distributions of X and Y are identical independent of the value of Z. Hence, the mutual information between the variables is

$$\mathbb{MI}(X;Z) = 0 \quad \text{and} \quad \mathbb{MI}(Y;Z) = 0.$$

However, the probabilities of S on the other hand depend on Z.

Joint probabilities:

$$p(\text{AG},1) = \tfrac{2}{8} \qquad p(\text{CT},1) = \tfrac{2}{8} \qquad p(\text{AT},1) = 0 \qquad p(\text{CG},1) = 0$$

$$p(\text{AG},0) = 0 \qquad p(\text{CT},0) = 0 \qquad p(\text{AT},0) = \tfrac{2}{8} \qquad p(\text{CG},0) = \tfrac{2}{8}$$

This results in a joint mutual information of

$$\mathbb{MI}(X,Y;Z) = \mathbb{MI}(S;Z) = 1.$$

S is perfectly associated with Z due to the creation of new knowledge, even though neither X nor Y share any information with Z alone. This makes an accurate depiction in a Venn diagram impossible, since it would mean that on the one hand there is no overlap between the circles representing X and Y with the circle of Z, while on the other hand the area of X and Y together should encompass the whole circle of Z.

Similar to the conditional entropy, the conditional mutual information $\mathbb{CMI}(X;Y|Z)$ is defined as the amount of information shared between X and Y that remains if the value of Z is known (see Figure 3.4 b)).

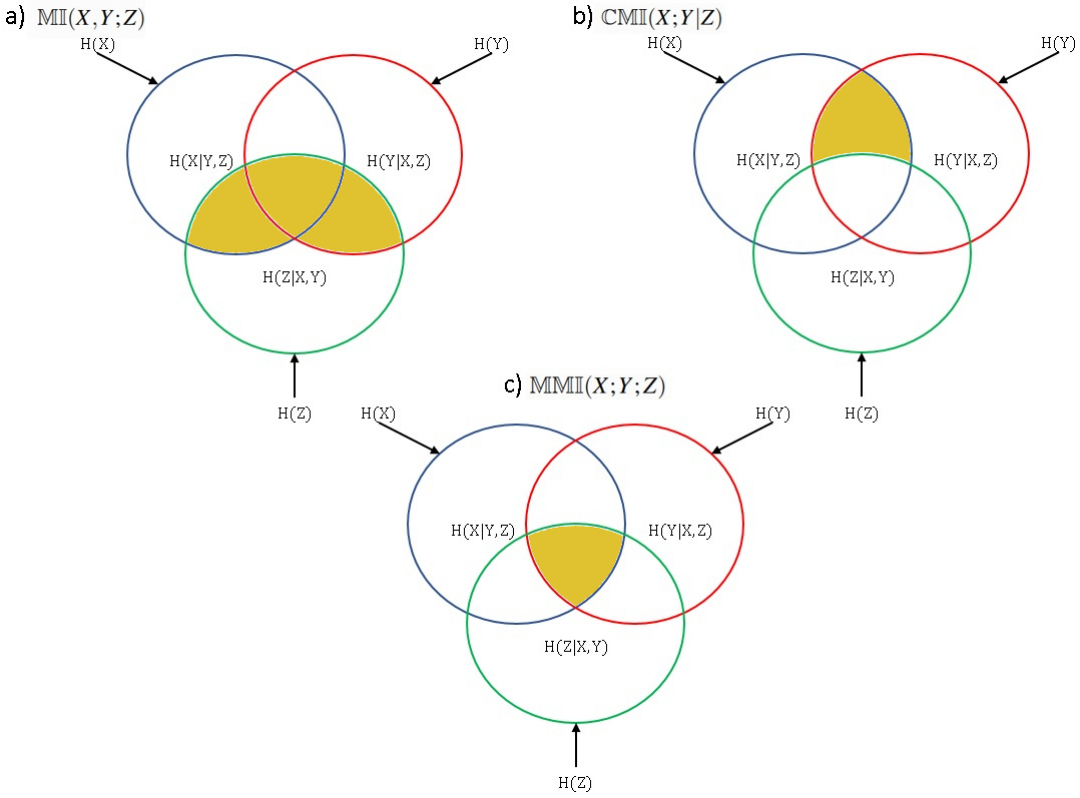


Figure 3.4.: Relationship between entropy and information theory measures for three variables X, Y and Z. The filled in area marks the corresponding measure. a) joint mutual information $\mathbb{MII}(X,Y;Z)$, b) conditional mutual information $\mathbb{CMII}(X;Y|Z)$ and c) multivariate mutual information $\mathbb{MMII}(X;Y;Z)$.

Definition 3.6 (Conditional Mutual Information) *The conditional mutual information between two discrete random variables X and Y given a third discrete random variable Z is defined as*

$$\mathbb{CMII}(X;Y|Z) = \sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}}\sum_{z\in\mathfrak{Z}} p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}. \tag{3.2.9}$$

In terms of entropies the conditional mutual information can also be expressed as

$$\mathbb{CMII}(X;Y|Z) = \mathbb{H}(X|Z) - \mathbb{H}(X|Y,Z). \tag{3.2.10}$$

$\mathbb{CMII}(X;Y|Z)$ is non-negative and it is zero if and only if X and Y are conditionally inde-

pendent given the knowledge of Z.
The multivariate mutual information $\mathbb{MMI}(X;Y;Z)$ of X, Y and Z is the amount of information that is common to all three variables. This is shown in the Venn diagram in Figure 3.4 c) as the area where all three circles overlap.

Definition 3.7 (Multivariate Mutual Information) *The multivariate mutual information between three discrete random variables X, Y and Z is defined as*

$$\begin{aligned}\mathbb{MMI}(X;Y;Z) &= \mathbb{MI}(X;Y) - \mathbb{CMI}(X;Y|Z) \\ &= \mathbb{MI}(X;Z) - \mathbb{CMI}(X;Z|Y) \\ &= \mathbb{MI}(Y;Z) - \mathbb{CMI}(Y;Z|X).\end{aligned} \tag{3.2.11}$$

The multivariate mutual information has the following properties:

- Symmetry regarding X, Y and Z
- $-\min\{\mathbb{CMI}(X;Y|Z), \mathbb{CMI}(X;Z|Y), \mathbb{CMI}(Y;Z|X)\} \leq \mathbb{MMI}(X;Y;Z)$
- $\mathbb{MMI}(X;Y;Z) \leq \min\{\mathbb{MI}(X;Y), \mathbb{MI}(X;Z), \mathbb{MI}(Y;Z)\}$

In contrast to the mutual information between two variables, the multivariate mutual information can become negative, namely if the additional knowledge of the third variable Z increases the mutual information between X and Y.
A closely related measure is the information gain $\mathbb{IG}(X;Y;Z)$, which is the difference in the mutual information between two variables due to the knowledge of the third variable.

Definition 3.8 (Information Gain) *The information gain between three discrete random variables X, Y and Z is defined as*

$$\mathbb{IG}(X;Y;Z) = \mathbb{MI}(X,Y;Z) - \mathbb{MI}(X;Y) - \mathbb{MI}(Y;Z). \tag{3.2.12}$$

The following properties apply to the information gain:

- $\mathbb{IG}(X;Y;Z) = -\mathbb{MMI}(X;Y;Z) = \mathbb{CMI}(X;Y|Z) - \mathbb{MI}(X;Y)$
- Symmetry regarding X, Y and Z
- $-\min\{\mathbb{MI}(X;Y), \mathbb{MI}(X;Z), \mathbb{MI}(Y;Z)\} \leq \mathbb{MMI}(X;Y;Z)$
- $\mathbb{MMI}(X;Y;Z) \leq \min\{\mathbb{CMI}(X;Y|Z), \mathbb{CMI}(X;Z|Y), \mathbb{CMI}(Y;Z|X)\}$

A positive value of $\mathbb{IG}(X;Y;Z)$ indicates that there is synergy between the variables which means that the *whole* ($\mathbb{MI}(X,Y;Z)$) provides additional information compared with the sum of the contributions of the single *parts* ($\mathbb{MI}(X;Z)$ and $\mathbb{MI}(Y;Z)$), while a negative value shows that redundancy or correlation exists between the variables [75, 78].

3.2.4. Information theory for continuous variables

The previously introduced measures can also be applied to continuous random variables as well as to mixtures of discrete and continuous random variables. However, there are some significant differences to the solely discrete versions.
Let X be a continuous random variable with a probability density function $f(x)$. Its differential entropy depends only on the probability density function.

Definition 3.9 (Differential Entropy) *The differential entropy $\mathbb{H}(X)$ of a continuous random variable X with a probability density function $f(x)$, which has a support set $\mathfrak{X}$ with $f(x) > 0$ for $x \in \mathfrak{X}$, is defined as*

$$\mathbb{H}(X) = -\int_{\mathfrak{X}} f(x) \log f(x)\, dx. \tag{3.2.13}$$

In contrast to the entropy of discrete variables, the differential entropy of a continuous variable can be negative. Similarly to the entropy, the mutual information between two continuous random variables can be calculated.

Definition 3.10 ((Continuous) Mutual Information) *The mutual information $\mathbb{MI}(X;Y)$ between two continuous random variables X and Y with marginal probability density functions $f(x)$ and $f(y)$ over support sets $\mathfrak{X}$ and $\mathfrak{Y}$, respectively, as well as a joint probability density function $f(x,y)$ is defined as*

$$\mathbb{MI}(X;Y) = \int_{\mathfrak{X}} \int_{\mathfrak{Y}} f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy. \tag{3.2.14}$$

However, the mutual information between continuous variables keeps the properties of its discrete counterpart [156].
The probability densities of the variables are in general unknown, which makes estimators necessary. The most basic approach is to partition the continuous random variables into bins of a finite size (as shown in Figure 3.5) and, thereby, transform them into discrete variables. Such an estimator, however, loses information through the binning process and more sophisticated methods have been developed that improve the accuracy of the estimation. There are, in particular, methods that estimate the mutual information directly without explicitly calculating the respective entropy values [156].

3.2.4.1. Estimating Mutual Information between discrete and continuous variables

Of particular relevance for this thesis is the accurate estimation of the mutual information between discrete and continuous variables. For this purpose, Ross published in 2014 an

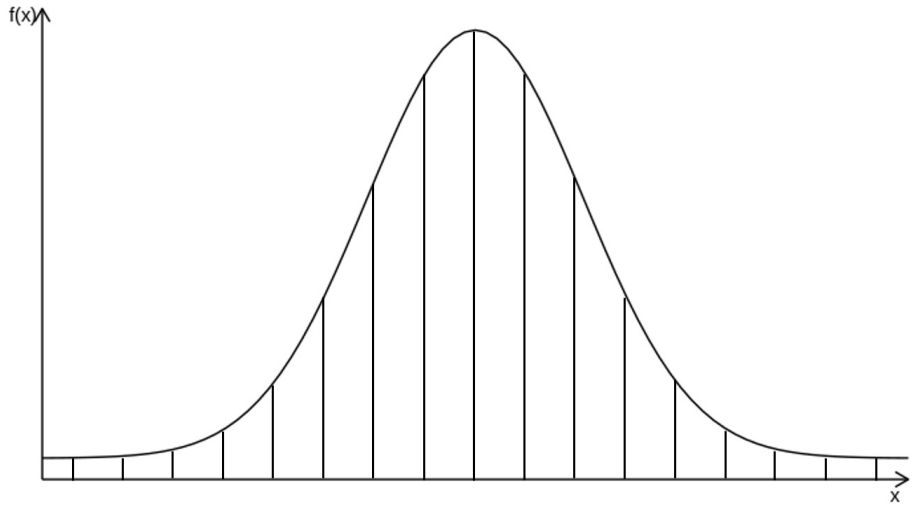


Figure 3.5.: Binning of a continuous random variable. The vertical lines mark the beginnings and ends of the discrete bins on the continuous scale.

estimator [87], the main idea of which I recapitulate here. The estimator is adapted from an earlier work from Kraskov et al. [156], who presented an estimator for the mutual information between two continuous variables. Both estimators are based on the Kozachenko and Leonenko (KL) estimator for entropy [157], which uses the distance to the kth-nearest neighbor of each observation.
Let $X = \{x_1, x_2, \ldots, x_N\}$ be a discrete random variable with N observations and an alphabet $\mathfrak{X}$ and let $Y = \{y_1, y_2, \ldots, y_N\}$ be a continuous random variable with N observations. The mutual information estimator is based on the following equation

$$\begin{aligned}\mathbb{MI}(X;Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= -\sum_{x\in\mathfrak{X}} p(x)\log p(x) - \int f(y)\log f(y)\,dy + \sum_{x\in\mathfrak{X}} \int f(x,y)\log f(x,y)\,dy \\ &= -\int f(y)\log f(y)\,dy + \sum_{x\in\mathfrak{X}} \int f(x,y)\log f(y|x)\,dy \\ &= -\langle \log f(y)\rangle + \langle \log f(y|x)\rangle,\end{aligned} \tag{3.2.15}$$

where

- $f(y)$ is the probability density of y,
- $f(y|x)$ is the probability density of y given x and
- $\langle \ldots \rangle$ indicates the average over all observations.

The probability densities $f(\ldots)$ are estimated using the aforementioned KL estimator as

$$\langle \log f(y)\rangle \approx \langle \psi(m)\rangle - \psi(N) - \langle \log V\rangle \tag{3.2.16}$$

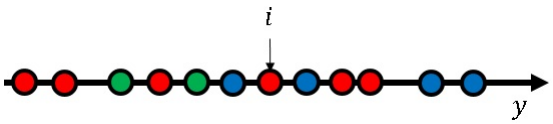
and

$$\langle \log f(y|x) \rangle \approx \psi(k) - \langle \psi(N_x) \rangle - \langle \log V \rangle. \tag{3.2.17}$$

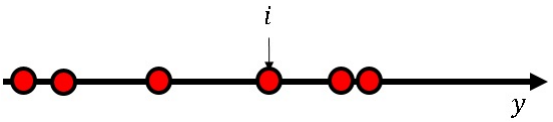
ψ indicates the digamma function [158] while k is a hyperparameter of the KL estimator. For an observation i, V is defined as the volume of observations that are closer to the observation i than its kth-nearest neighbor among those observations where the value of the discrete variable equals x_i. In this context, the distance between observations is measured as the absolute difference of the corresponding y-values. Furthermore, m refers to the number of observations that lie within V irrespective of their x-value. Finally, N_x is the total number of observations that have the same x-value as observation i.
The following example, which is based on the visualization given in [87], shows how these values are determined for a single observation in a small dataset.

Example: Estimation of mutual information

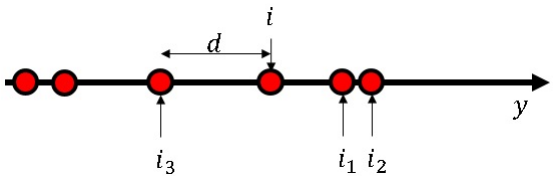
Let X be a discrete random variable which takes on values from an alphabet of size 3, and let Y be a continuous random variable. Both variables have values for 12 observations ($N = 12$). The observations are visualized with x_i representing the color and y_i the location on the y-axis of the corresponding point. I exemplarily show how the values N_{x_i} and m_i are determined for the marked observation i.



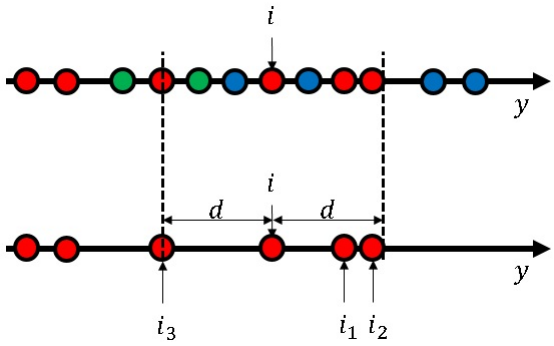
For the purpose of this example, I set $k = 3$. In a first step, only the observations with the same x-value as i are considered (here visualized as the red points). Six observations are left, which results in $N_{x_i} = 6$.



Due to $k = 3$, I determine the 3rd-nearest observation to i (marked by i_3) based on the absolute difference between their respective y-values, which is designated $d = |y_i - y_{i_3}|$.



Using the distance d, an interval $[y_i - d, y_i + d]$ is defined on the y-axis, which represents V in the equations above.



If we consider this interval in the context of the original 12 observations, we find that it contains 6 points beside i, and therefore $m_i = 6$.
It is intuitive that if m_i is close to k, then most observations in the interval around i will have the same x-value as i. In that case, x_i would be representative for this interval on y and we have an association between the two variables (at least for this interval).
Equation 3.2.16 refers to an interval over all points, while Equation 3.2.17 considers only those points with the same x-value as i in this interval. Due to defining V using the same observation (i_3) in both equations, we can cancel out the term $\langle \log V \rangle$ later.

By inserting the Equations 3.2.16 and 3.2.17 into 3.2.15 we obtain the following mutual information estimator:

$$\begin{aligned} \mathbb{MI}(X;Y) &\approx \psi(N) - \langle \psi(N_x) \rangle + \psi(k) - \langle \psi(m) \rangle \\ &= \frac{1}{N} \cdot \sum_{i=1}^{N} (\psi(N) - \psi(N_{x_i}) + \psi(k) - \psi(m_i)) \end{aligned} \tag{3.2.18}$$

This estimator can be extended to the joint mutual information of a pair of discrete random variables X and Y with a third continuous variable Z by calculating the mutual information

between the discrete grouping variable S, consisting of X and Y, and the continuous Z.

$$\mathbb{MII}(X,Y;Z) = \mathbb{MII}(S;Z) \approx \frac{1}{N} \cdot \sum_{i=1}^{N} (\psi(N) - \psi(N_{s_i}) + \psi(k) - \psi(m_i)) \tag{3.2.19}$$

3.2.5. Normalized Mutual Information

The upper bounds of the mutual information between two or more variables depend on the entropy values of the variables under study and, thus, on their alphabet sizes. These inconstant bounds can lead to misinterpretations when comparing the values of mutual information as well as when evaluating the strength of association between variables. Therefore, it is often preferred to remove these effects by normalizing the mutual information so that all values have the same range, commonly $[0,1]$. This normalized mutual information is denoted by $\mathbb{NMII}$. One common normalization strategy is to use the maximal alphabet size as shown in the following equation:

$$\mathbb{NMII}(X;Y) = \frac{\mathbb{MII}(X;Y)}{\log(\max\{|\mathfrak{X}|,|\mathfrak{Y}|\})} \tag{3.2.20}$$

An extensive overview of various normalization strategies is given in [159].

Example: Effect of normalization on mutual information

Let W, X, Y and Z be four discrete random variables with alphabets $\mathfrak{W}$, $\mathfrak{X}$, $\mathfrak{Y}$ and $\mathfrak{Z}$ and the following outcomes:

W:	1	2	1	2	1	2	1	2
X:	A	B	A	B	A	B	A	B
Y:	1	2	3	4	1	2	3	4
Z:	A	B	C	D	B	A	D	C

It can be easily seen, that there is a perfect association between W and X, with 1 corresponding to A and 2 corresponding to B, which is reflected in their mutual information of $\mathbb{MII}(W;X) = 1$. The association between Y and Z, on the other hand, is significantly weaker and there is no direct correspondence between the values of Y and Z. Despite that, their association also obtains a mutual information of $\mathbb{MII}(Y;Z) = 1$.
Therefore, to see the difference, it is necessary to normalize the mutual information.

$$\mathbb{NMII}(W;X) = \frac{\mathbb{MII}(W;X)}{\log(\max\{|\mathfrak{W}|,|\mathfrak{X}|\})} = \frac{1}{\log(\max\{2,2\})} = 1 \tag{3.2.21}$$

$$\mathbb{NMI}(Y;Z) = \frac{\mathbb{MI}(Y;Z)}{\log(\max\{|\mathfrak{Y}|,|\mathfrak{Z}|\})} = \frac{1}{\log(\max\{4,4\})} = 0.5 \tag{3.2.22}$$

Based on the normalized mutual information, it can be correctly concluded that the association between Y and Z is weaker than the association between W and X.

3.3. Genotype-phenotype association studies

Understanding the genetic causes that drive the differential expression of a phenotype among individuals in a population is of significant interest for researchers in the fields of medicine as well as animal and plant breeding [160]. Due to the complex nature of many traits, which may be controlled by intricate interactions of a multitude of genes, this is a challenging task. With the ever-increasing amount of SNPs becoming available for species, genome-wide association studies (GWASs) have become one of the premier tools used for deciphering the genetic architecture of traits [161, 162].
The basic idea of a GWAS is to test each single SNP for association with the phenotype under study [160]. SNPs which show a significant association are usually assumed to cause an effect themselves or to be in linkage disequilibrium (LD) with the actual causal gene of interest and can therefore serve as markers for it. In the end, these markers can then, for example, be utilized for prediction of the phenotype or the identification of the causal mechanisms [161].
There are different methods available for testing the association between a SNP and the phenotype. For a qualitative phenotype, such as disease immunity, one can, for example, apply a chi-square test to check for dependency between the genotypes of the SNP and the disease state. Quantitative traits, on the other hand, can be analyzed using linear regression [163].

3.3.1. Linkage disequilibrium

LD is a measure for the phenomenon of non-random association of alleles between two SNPs [164]. This association is caused by the low probability of recombination between two SNPs that are located closely together on the same chromosome [163]. Multiple measures for LD exist, but in this thesis I use r^2, which is the correlation between the genotype minor allele counts of the SNPs. This measure is implemented in PLINK [135] and, unlike many other measures, does not require knowledge about the haplotypes. As a correlation measure, r^2 ranges between 0 and 1, where the former indicates no linkage and the latter complete association, i.e., the two SNPs convey the same information [163].

3.3.2. Detection of epistatic interactions

Large parts of this chapter are taken from my publication [165] (see Appendix A.3) or are based on the reviews given in [53, 121].
The approaches used for detecting epistatic interactions can be seen as an extension of the previously introduced GWAS methods. Whereas in the latter the effect of each SNP on the phenotype is tested individually, in the former the association of two or more SNPs jointly with the phenotype is considered. In this context, the identification of interactions between SNPs that cause deviations from the sum of the single effects of the SNPs is of particular importance. This phenomenon is also referred to as statistical epistasis [121, 124, 166]. Many methods for epistasis detection are limited to a specific type of phenotype, with most of them requiring a qualitative phenotype. A major part of the methods are based on linear regression (respectively logistic regression for case-control phenotypes). For example, the software PLINK [135] implements epistasis detection by fitting the data to the following regression model

$$Y = \beta_0 + \beta_1 \cdot X^1 + \beta_2 \cdot X^2 + \beta_3 \cdot X^1 \times X^2. \tag{3.3.1}$$

In Equation 3.3.1,

- Y is the quantitative phenotype under study,
- X^1 and X^2 refer to the genotypes of the two SNPs that are tested with $X^1 \times X^2$ being their interaction,
- β_0 defines the intercept of the regression,
- β_1 and β_2 are the effect sizes of the single SNPs and
- β_3 is the effect size of the interaction between the SNPs.

For the purpose of regression, the genotypes of the SNPs are commonly encoded as $\{0,1,2\}$, where the value indicates the number of alternate alleles in the genotype [167]. By applying a Wald-test ($F_{\text{Wald}} = \frac{\widehat{\beta_3}}{SE(\beta_3)}$), the interaction effect of the SNP pair is then tested for significance. For a case-control phenotype, Y is replaced by $\log\left(\frac{P(Y=\text{case})}{P(Y=\text{control})}\right)$. A weakness of this and related approaches is, however, that they assume a specific genetic model (for example multiplicativity) underlying the phenotype, which might not reflect reality. Therefore, non-parametric methods have also become quite popular [61, 168].
Due to the large number of possible combinations of SNPs under study even if only pairwise interactions are considered, the detection of epistasis is a computational challenge, for which a large number of algorithms have been proposed. These methods can be roughly divided into different categories depending on the search strategy they use to address this issue.
Exhaustive search strategies test every possible combination of SNPs for significance, which often results in a long execution time and can become infeasible for large datasets. This strategy has been used by partitioning methods such as the Combinatorial Partitioning

Method (CPM) [169] and the Restricted Partition Method (RPM) [170], as well as several other methods [61, 171, 172].

Stochastic methods, on the other hand, use random sampling to increase their efficiency, but their results and performance can depend on parameters set by the user. Bayesian Epistasis Association Mapping (BEAM) [173], for instance, applies Markov chain Monte Carlo to compute the posterior probability for association between SNPs and a disease. Its extension epistatic MOdule DEtection (epiMODE) [174] uses Gibbs sampling with a reversible jump Markov chain Monte Carlo to find epistatic interactions.

Machine learning methods such as neural networks [175, 176, 177, 178], decision trees [179] or random forests [180, 181, 182, 183] have also been utilized for epistasis detection.

Step-wise approaches form a fourth category of algorithms, which first filter out SNPs with a very small or no association signal, and then test among the surviving SNPs for epistatic interactions. BOolean Operation-based Screening and Testing (BOOST) [166], as an example, first performs a likelihood ratio test to filter out unimportant SNPs and then performs an exhaustive search on the others. Leem et al. [76] utilized a k-means clustering of the SNPs and then searched for interactions between SNPs in different clusters. Other methods use the results of lower-order interactions to find higher-order interactions in an efficient way [77, 184].

3.3.2.1. Detection of epistasis using information-theory-based measures

Several of the aforementioned methods use information-theory-based measures such as mutual information to quantify epistatic interactions [75, 76, 77, 78, 79, 80, 81, 82, 83, 84]. These measures consider the SNPs and phenotypes as random variables, which allows them to quantify the amount of information, or uncertainty, that is inherent in a SNP or phenotype and to compute how much information is shared between them, and thus the strength of association [75]. This approach is model-free and therefore has the advantage of not requiring any prior assumptions regarding the structure of the interactions. By considering all genotype combinations of the SNPs as separate categories, this strategy also avoids the problem of choosing an appropriate encoding method for the SNPs and their interactions, which has been shown to influence the results of regression-based methods [167, 185, 186]. Nevertheless, the application of information-theory-based approaches has so far been limited to qualitative phenotypes. This is because, while the mutual information between two discrete variables can be efficiently calculated using simple contingency tables, the mutual information between a discrete and a continuous variable requires computationally more challenging approaches for an accurate estimation. Depending on the measure, these methods can be divided into two major groups. The first group utilizes the joint mutual information $\mathbb{MI}(X^1,X^2;Y)$ to measure epistasis between a SNP pair X^1 and X^2 to a phenotype Y [76, 81, 83, 84]. The second group, on the other hand, applies the information gain $\mathbb{IG}(X^1;X^2;Y)$ for this purpose [77, 78, 79, 80, 83, 84]. A major difference between these groups is that the former does not account for the main effects of the single SNPs while

the second group explicitly removes them. Thereby, $\mathbb{IG}$ largely reflects only the synergistic effect of the two SNPs on the phenotype [75, 84]. In this regard, $\mathbb{MI}$ is more general and also allows the detection of interactions where the information of the single SNPs is complementary to each other. Nevertheless, directly using $\mathbb{MI}$ can lead to the detection of spurious interactions that are caused by a single strongly associated SNP, and thereby hide the actual true epistatic interactions. The impact of this difference is shown in the following example.

Example: Comparison of joint mutual information and information gain for measuring epistasis

In this example, I present four different cases where a single binary phenotype is completely determined by the genotypes of two SNPs for 8 samples.

Case 1			**Case 2**			**Case 3**			**Case 4**		
X^1	X^2	Y	X^1	X^2	Y	X^1	X^2	Y	X^1	X^2	Y
AA	CC	0	AA	CC	0	AA	CT	0	AA	CC	0
AA	CC	0	AA	TT	0	AA	CT	0	GG	TT	0
AA	CC	0	AA	CC	0	AG	CC	0	AA	CC	0
AA	CC	0	AA	TT	0	AG	CC	0	GG	TT	0
GG	TT	1	GG	CC	1	AG	TT	1	AA	TT	1
GG	TT	1	GG	TT	1	AG	TT	1	GG	CC	1
GG	TT	1	GG	CC	1	GG	CT	1	AA	TT	1
GG	TT	1	GG	TT	1	GG	CT	1	GG	CC	1

- Case 1: Both SNPs X^1 and X^2 are perfectly associated with the phenotype Y.
- Case 2: SNP X^1 is perfectly associated with Y while SNP X^2 shows no association.
- Case 3: For both SNPs the homozygous genotypes only occur together with a specific phenotype while the heterozygous genotypes occur for both possible values of Y. However, by considering both SNPs together the phenotype can be completely determined.
- Case 4: There exists no association to the phenotype for both SNPs. The genotype frequencies are the same irrespective of the value of Y. Nevertheless, the phenotype is completely determined by taking X^1 and X^2 together.

The joint mutual information ($\mathbb{MI}$) and the information gain ($\mathbb{IG}$) behave differently in the presence of association between the single SNPs and the phenotype as the following table shows.

Case	$\mathbb{MI}(X^1;Y)$	$\mathbb{MI}(X^2;Y)$	$\mathbb{MI}(X^1,X^2;Y)$	$\mathbb{IG}(X^1;X^2;Y)$
1	1	1	1	-1
2	1	0	1	0
3	0.5	0.5	1	0
4	0	0	1	1

It can be easily observed in the table that the results of $\mathbb{MI}(X^1,X^2;Y)$ and $\mathbb{IG}(X^1;X^2;Y)$ are only identical in the total absence of association between the single SNPs X^1 and X^2 to the phenotype Y (case 4). Depending on the strength of the single associations, the information gain is non-existent (cases 2 and 3) or even negative (case 1), which represents a redundancy in the available information from the SNPs. In contrast, the joint mutual information indicates in all cases that the two SNPs together explain the phenotype completely.

4. Material and methods

In this chapter, I present the two analysis frameworks that I developed during the course of this thesis for the analysis of the genetic mechanisms underlying the V+C content in *Vicia faba*. First, I describe the *Vicia faba* sequence data, which is used as the input of the first method. In the second section, I detail the other plant sequence datasets that constitute the training data for promoter prediction. I further present two other real datasets that I used to evaluate the second framework. Afterwards, I describe the framework that I developed to identify regulatory SNPs based on genotyping-by-sequencing data and, finally, I present the MIDESP algorithm for the detection of epistatic SNP pairs using mutual information. The content in this chapter is mostly taken from my published papers [52, 165, 187] (see Appendices A.1, A.2 and A.3).

4.1. Datasets

4.1.1. Genotyping-by-Sequencing Data of *Vicia faba*

To explore the genetic background of the V+C content in *Vicia faba*, I obtained genotyping-by-sequencing (GBS) reads from 20 inbred lines of faba bean. These lines were inbred via single-seed descent from cultivars, from a gene-bank accession, from biparental crosses or from a landrace and include winter and spring types (see Table 4.1 for more information). Among those lines 6 had a low V+C content and 14 had a high V+C content. The extraction of the DNA, the sequencing and the filtering were carried out by LGC Genomics GmbH (Berlin, Germany). DNA was extracted from the grains of the plants and then sequenced using an Illumina NextSeq 500 V2 platform and the restriction enzyme MslI (NEB, recognition sequence: CAYNN^NNRTG). Sequencing adapter remnants were subsequently trimmed and reads whose 5' ends did not match the restriction enzyme site were discarded. A more detailed description of the sequencing process is given in my publication [187] (see Appendix A.2). For each sample, approximately 3 million 150 bp long paired end reads could be obtained, which results in a total sequence length of about 18 Gbp. The reads are stored as FASTQ files, which additionally to the DNA sequence contain a quality score for each base. The sequences have also been deposited at the European Nucleotide Archive (ENA) under the accession number PRJEB38838 and were published in the aforementioned work.

Table 4.1.: Vicine and convicine status of the 20 *Vicia faba* lines, name, sample ID, ENA accession number and additional notes.

ENA Accession	Sample ID	V+C	Line	Notes
ERS4652931	Sample_8	Low	Line 1268-4-1	Ancestor of low V+C content
ERS4652926	Sample_3	Low	Mélodie/2	cv. Mélodie; minor, spring bean
ERS4652927	Sample_4	Low	F7(Mélodie/2 x ILB938/2)-139-1-1	Near isogenic lines (ILB938/2 is from Ecuador)
ERS4652928	Sample_5	Low	F7(Mélodie/2 x ILB938/2)-201-3-1	
ERS4652932	Sample_9	High	F7(Mélodie/2 x ILB938/2)-139-2-1	
ERS4652933	Sample_10	High	F7(Mélodie/2 x ILB938/2)-201-4-1	
ERS4652929	Sample_6	Low	F7[VC.14.8099-843-2-1]	Near isogenic lines from a breeder's cross, spring beans
ERS4652930	Sample_7	Low	F7[VC.14.8099-848-3-1]	
ERS4652934	Sample_11	High	F7[VC.14.8099-843-3-3]	
ERS4652935	Sample_12	High	F7[VC.14.8099-848-4-1]	
ERS4652924	Sample_1	High	HediLin-1	cv. Hedin; minor, spring bean
ERS4652936	Sample_13	High	PietraLin	Major, Mediterranean bean
ERS4652937	Sample_14	High	(HediLin/1 x PietraLin)-2-4	Near isogenic lines
ERS4652938	Sample_15	High	(HediLin/1 x PietraLin)-4-4	
ERS4652939	Sample_16	High	S_281	Academic winter bean lines
ERS4652940	Sample_17	High	S_301	
ERS4652941	Sample_18	High	S_034	
ERS4652942	Sample_19	High	S_290	
ERS4652925	Sample_2	High	Hiverna/2	cv. Hiverna; minor, winter bean
ERS4652943	Sample_20	High	Côte d'Or/1	Côte d'Or; minor, winter bean

4.1.2. Partial genome and SNPs for *Vicia faba*

Using above mentioned NGS reads, I followed the strategies outlined in [188, 189] and applied the *de novo* assembler Trinity [190] to obtain a partial genome for *Vicia faba*. In total, 694,605 contigs with an average length of 236 bp were constructed. To filter out redundant contigs, I clustered the contigs with CD-HIT [191] using a threshold of 95.0 % for sequence identity. The *de novo assembly* and this filtering resulted in a partial genome consisting of 419,390 contigs with a total length of 100,037,292 bp. Through remapping of the reads to the partial genome with Bowtie2 [192] and subsequent variant calling with SAMtools [193], I derived 1,880,592 SNPs after excluding structural variants such as insertions and deletions as well as non-biallelic SNPs. The quality scores of these SNPs showed a clear bimodal distribution with a minimum at a quality score of 400. Therefore, I discarded 1,195,377 SNPs with a quality score of less than 400, leaving 685,215 SNPs with high quality. A detailed description of the process and the used program parameters can be found in Appendix A.5.

4.1.3. Promoter and non-promoter sequences of several plant species

Mainly considering members of the *Leguminosae* family, I utilized seven species (*Glycine max*, *Lupinus angustifolius*, *Medicago truncatula*, *Phaseolus vulgaris*, *Trifolium pratense*, *Vigna angularis*, and *Vigna radiata*) with a complete and annotated reference genome sequence for the promoter prediction. As a more distantly related plant, I additionally included the model species *Arabidopsis thaliana*. For each species, the core promoter sequences covering the -200 bp to +50 bp regions relative to the transcription start sites (TSSs) of protein coding genes were extracted from the Ensembl Plants database (release 45) [129] using BioMart [194]. Simultaneously, the sequences covering [TSS+751,TSS+1000] from the core gene region of the genes were extracted as non-promoter sequences. Sequences that were not assigned to a chromosome or which contained ambiguous bases were not considered. Furthermore, I included two additional datasets of non-promoter sequences. The first was randomly extracted from the *Medicago truncatula* reference genome excluding the region [TSS-1000,TSS+500] and the second set was sampled from the *Vicia faba* reference transcriptome V2, which was downloaded from the Pulse Crop Database [130]. In both cases, I sampled sequences of length 250 bp as non-promoters. The final number of sequences for each dataset is given in Table 4.2.

Table 4.2.: Number of promoter and non-promoter sequences in the datasets that were used for training.

Species	# Promoter sequences	# Non-promoter sequences
Arabidopsis thaliana	23,315	23,315
Glycine max	46,199	46,199
Lupinus angustifolius	23,463	23,463
Medicago truncatula	32,158	32,158
Phaseolus vulgaris	22,750	22,750
Trifolium pratense	14,749	14,749
Vigna angularis	19,584	19,584
Vigna radiata	15,495	15,495
Medicago truncatula (Genome-wide)	-	11,732
Vicia faba (Transcriptome)	-	57,623

4.1.4. Bovine tuberculosis (BT) dataset

To validate MIDESP for qualitative phenotypes, I analyzed a genotype×phenotype dataset with a case/control phenotype that represents the resistance of cattle towards bovine tuberculosis. This dataset was published by Bermingham et al. [195] and consists of 617,885 SNPs for 1151 cattle samples with 592 cases and 559 controls. The cattle belonged to

the Holstein-Friesian breed and were collected in Northern Ireland. Bermingham et al. performed a GWAS on this dataset to find SNPs associated with the resistance of cattle towards bovine tuberculosis. They were able to find eight significantly associated SNPs, which represent two different loci in the genome. To ensure the quality of the data, I applied several filters to the dataset following Ramzan et al. [196, 197]. I removed SNPs that

- had a minor allele frequency ≤ 0.01,
- had a genotyping call rate ≤ 0.97 or
- deviated signifcantly from the Hardy-Weinberg equilibrium (p-value $< 1 \times 10^{-6}$).

On the other hand, samples were removed if the phenotype or more than 5% of the SNPs were missing. After filtering, 616,398 SNPs and all 1151 samples remained.

4.1.5. Egg weight (EW) dataset

The final dataset relates to the egg weight of 36 weeks old chickens [198]. 1063 birds, which belong to a line of Rhode Island Red chicken, were genotyped using the Affymetrix Axiom® 600 K Chicken Genotyping Array. This resulted in an initial set of 580,961 SNPs, which were then filtered. The dataset made available by the authors consists only of the 294,705 SNPs that passed their quality filters. No further SNP or animal could be removed using the criteria given in the previous section. Although the dataset contains multiple phenotypes in the form of egg weights for different ages of the chickens, I decided to use only the phenotype data for 36 weeks old chicken because this phenotype provided the strongest signal in previous GWAS analyses [196, 198].

4.2. Identification of regulatory SNPs based on genotyping by sequencing data in *Vicia faba* using deep learning

In this section, I present the analysis workflow that I developed for the computational identification of regulatory SNPs (rSNPs), i.e. SNPs in promoter regions of genes, which are deemed to govern the V+C content of *Vicia faba*, based on genotyping by sequencing (GBS) data. Starting position for this approach is the partial genome and the SNPs for *Vicia faba*, which were described in Section 4.1.2. The workflow consists of two major parts, which I explain in detail in the following. First, I predict promoter regions in the partial *Vicia faba* genome using deep learning and annotated sequence data from related species. Second, I analyze the SNPs within these promoter regions regarding their effects on the binding affinity of transcription factors (TFs) and their association to the V+C content in order to identify important rSNPs. In a final step, I perform a functional analysis of the identified candidate gene and transcription factors. This process has been published in [52] (see Appendix A.1).

4.2.1. Identification of promoter regions

For the prediction of the promoter regions, I utilize a convolutional neural network (CNN), which, as described in Section 3.1, is nowadays one of the most popular approaches for promoter detection. The structure of the network that I used is illustrated in Figure 4.1.

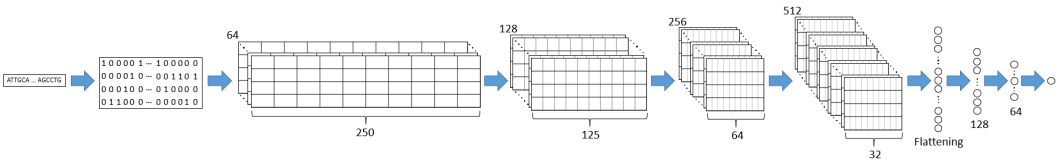


Figure 4.1.: The network architecture of the CNN used for promoter prediction consists of four 1D-convolutional layers followed by a flattening layer and two fully-connected layers. At the end, an output layer with one neuron and a sigmoid activation function computes the probability that the analyzed sequence is classified as a promoter sequence.

The input of the network is formed by a sequence of nucleotides of length 250 bp, where each nucleotide is encoded in a one-hot representation and expressed by a four-dimensional vector, with A encoded as $(1,0,0,0)$, C as $(0,1,0,0)$, G as $(0,0,1,0)$, and T as $(0,0,0,1)$. As can be seen in Figure 4.1, the network is composed of four 1D-convolutional layers followed by a flattening layer, two fully-connected layers and an output layer. All convolutional layers are implemented using a ReLU (Rectified Linear Unit) activation [199], a stride parameter of 2, zero-padding and a filter size of 21. The first layer uses 64 filters, whereas the second, third and fourth layers use 128, 256, and 512 filters, respectively. To

avoid overfitting on the training data, a dropout layer with rate = 0.2 is used after each convolution [200]. After the sequences are processed by the convolutional layers, a flattening layer transforms the output into a one-dimensional vector and passes its values to two consecutive fully-connected layers with 128 and 64 neurons, respectively. Finally, an output layer with a sigmoid activation classifies the input sequence as promoter or non-promoter. The CNN is trained using the Adam optimizer [201], L2-regularization and binary cross-entropy loss [202]. For the network, 90 % of the sequences are used for training and the remaining 10 % are used for testing. The CNN is implemented in R using Keras [203] with TensorFlow [204] as a backend.

Training the network requires sequence data from known promoter and non-promoter sequences. However, because no annotated reference genome exists for *Vicia faba*, I do not have such data available. To alleviate this lack of training data, I exploit the conservation of promoter signatures among closely related plant species [50] and use the sequences of other members of the *Leguminosae* family for training and testing the network. This data is described in Section 4.1.3.

To assess the prediction performance of the network, I identify the number of correctly predicted promoter and non-promoter sequences as True Positives (TP) and True Negatives (TN), as well as the number of true promoter sequences predicted as non-promoter sequences, False Negatives (FN), and the number of true non-promoter sequences predicted as promoter sequences, False Positives (FP). From these measures, I calculate Accuracy (ACC), Sensitivity (true positive rate), Specificity (true negative rate), and the Matthews Correlation Coefficient (MCC) as below [46, 47, 205]:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.2.1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2.2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.2.3}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4.2.4}$$

Following previous studies [46, 51], I additionally test whether explicitly adding specific sequence features can improve the performance of the network. These features are included one-by-one by concatenating their values to the flattening layer in order to explore their effects on the improvement of the classification network. I test the following six features, which have previously been used for promoter detection:

- Frequency of the dinucleotides CA and CG
- Frequency of the TATA motif

- CG-skew of sequences defined as

$$CG_{skew} = \frac{\#C - \#G}{\#C + \#G} \tag{4.2.5}$$

  where $\#C$ and $\#G$ refer to the counts of the nucleotides C and G in the sequences, respectively
- Frequency of k-mers using different values for k
- Horizontal Mutual Information (HMI), which is calculated based on a predefined distance d between two positions in a sequence and provides a measure of auto-covariation between the nucleotides of interest [206].

$$\mathbb{HMI}(d) = \sum_{m=\{A,C,G,T\}} \sum_{n=\{A,C,G,T\}} p_{mn}(d) \cdot \log \frac{p_{mn}(d)}{p_m(d)p_n(d)}, \tag{4.2.6}$$

  where $p_m(d)$, $p_n(d)$ and $p_{mn}(d)$ refer to the marginal and joint probabilities of the nucleotides being d bp apart, respectively. A high value of $\mathbb{HMI}(d)$ would indicate a strong correlation between the nucleotides regarding their distance d.
- Generalized Topological Entropy (GTE), which is a measure that reflects the complexity of the DNA sequence [207]. It has been previously used to study and compare the complexity of introns, exons and promoter regions [208]. Let ω be a DNA sequence of length $|\omega|$ and let n_ω be the unique integer such that $4^n + n - 1 \leq |\omega| < 4^{n+1} + (n+1) + 1$. Then the GTE is defined as

$$\mathbb{H}^k_{n_\omega}(\omega) = \frac{1}{k} \sum_{i=n_\omega-k+1}^{n_\omega} \frac{\log_4(p_\omega(i))}{i}, \tag{4.2.7}$$

  where $p_\omega(i)$ refers to the number of unique sub-sequences of length i that appear in ω. I set $k = n_\omega$ to consider sub-sequences of all possible lengths.

For the prediction of promoter regions in the partial genome of *Vicia faba*, it is important to note that, due to the random fragment orientation regarding the direction of the reads from GBS, the correct orientation of the sequences in the partial genome is unknown. To address this limitation, I consider four different types of the sequences for the predictions as: (i) the original obtained assembly; (ii) the complement of the obtained assembly that is gained by keeping the reading direction; (iii) the reverse of the obtained assembly that is gained by changing the reading direction; and (iv) the reverse complement of the obtained assembly. If at least one of the four sequences is predicted by the CNN as a promoter, then the corresponding original sequence is deemed a promoter. On account of the CNN requiring a DNA sequence with an exact length of 250 bp as input, I analyze the *Vicia faba* sequences using a sliding window approach. Sequences with a length of less than 250 bp are discarded. This approach and the different types of sequences are visualized in the

following example.

Example: Promoter detection of *Vicia faba* sequences using sliding windows

For the purpose of this example and an easier visualization, I assume that the CNN uses sequences of length 5 bp as input.
The exemplary genome consists of only three sequences A, B and C.

$$A = AAGTACC$$
$$B = ACGT$$
$$C = ACCTC$$

Since B has a length of less than the required 5 bp, it will not be analyzed and is instead directly marked as a non-promoter.
For each sequence that passes the required length, I create four different sequences from it as explained above.

$$A\_Original = AAGTACC \qquad C\_Original = ACCTC$$
$$A\_Complement = TTCATGG \qquad C\_Complement = TGGAG$$
$$A\_Reverse = CCATGAA \qquad C\_Reverse = CTCCA$$
$$A\_ReverseComplement = GGTACTT \qquad C\_ReverseComplement = GAGGT$$

Due to C having the exact required length, its sequences can be used directly as input for the CNN. If at least one of the four sequences (*C_Original*, *C_Complement*, *C_Reverse* and *C_ReverseComplement*) is classified as promoter, then C will be marked as a promoter. For the sequences from A, which are longer, I apply a sliding window approach to use each substring of length 5 as input for the CNN. If any of those substrings are classified as promoter, the corresponding section of A will be marked as promoter.

$$A\_Original\_0 = AAGTA \qquad A\_Original\_1 = AGTAC \qquad A\_Original\_2 = GTACC$$
$$A\_Complement\_0 = TTCAT \qquad A\_Complement\_1 = TCATG \qquad A\_Complement\_2 = CATGG$$
$$A\_Reverse\_0 = CCATG \qquad A\_Reverse\_1 = CATGA \qquad A\_Reverse\_2 = ATGAA$$
$$A\_ReverseComplement\_0 = GGTAC \qquad A\_ReverseComplement\_1 = GTACT \qquad A\_ReverseComplement\_2 = TACTT$$

The number at the end of the substring name indicates the used offset from the start of the sequence. As an example, if only *A_Original*_1 had been classified as a promoter by the CNN, then only the part of A written in red below would be classified as a promoter, while the remaining positions would be classified as non-promoters.

$$A = A\mathit{AGTAC}C$$

4.2.2. Identification of putative regulatory SNPs with association to V+C

In order to identify regulatory SNPs, I analyze the predicted promoter sequences of *Vicia faba*. For this purpose, I first select all SNPs that are located in promoter sequences and that can be mapped against the *Medicago truncatula* genome from the initial set of 685,215 SNPs (see Section 4.1.2). The mapping is performed using the BLASTN algorithm with a threshold of 0.01 for the *e-value* and of 90 for the *percent identity* [209]. Second, for each of these SNPs, I extract their flanking sequence, which covers the ± 25 bp relative to the position of the SNP. These sequences have a length of 51 bp with the SNP at position 26, which is in line with previous studies [111, 210, 211]. SNPs for which I could not obtain such flanking sequences are discarded. This may be the case, for example, if the SNP is located too close to the start or the end of the corresponding contig. Third, two copies of the extracted sequences are created: while the first sequence contains the reference allele at the SNP position, the second contains the alternate allele. Thereafter, I identify putative transcription factor binding sites (TFBSs) by applying the MATCH^{TM} program [136] together with a non-redundant plant position weight matrix (PWM) library obtained from the TRANSFAC® database [105] to the flanking sequences of the SNPs. The MATCH^{TM} program provides a matrix similarity score (MSS) for each putative TFBS, ranging from zero to one, which reflects the potential binding affinity of the related TF to it. Finally, I predict the consequence of each SNP for the TFBS by comparing their MSSs in the two sequences. As a result, I observe in my analysis four different types of consequences: (i) no effect, (ii) change in binding affinity, (iii) loss of TFBS (a TFBS appears only for the reference allele) and (iv) gain of TFBS (a TFBS appears only for the alternate allele) (see Figure 2.5). Two TFBSs are considered identical if their PWMs, positions, and their strands are equal for both alleles. If the scores computed by MATCH^{TM} are identical in both alleles, the SNP is assumed to have no effect on the TFBS. In the further analysis, I consider a SNP as an rSNP, if it has an effect on the binding affinity of at least one TF, i.e., if its type of consequence is (ii), (iii), or (iv). This follows the definition of a regulatory SNP given in Section 2.3.1. The association between candidate SNPs/rSNPs and the V+C content is tested with PLINK using a 1df chi-squared allelic test. To control the type I error rate, I set the false discovery rate (FDR) to 0.1.

4.3. Mutual information based detection of epistatic SNP pairs

In this section, I present a novel method called Mutual Information-based Detection of Epistatic SNP Pairs (MIDESP) for the detection of pairwise epistatic interactions, which extends the previously mentioned mutual information-based approaches in Section 3.3.2.1 by additionally enabling the identification of epistatic interactions between SNP pairs and quantitative phenotypes. For this purpose, in the context of epistasis, I adopt for the first time the mutual information estimator developed by Ross [87], which accurately estimates the level of epistasis using a kth-nearest neighbor-based approach. Moreover, to deal with

the possible obstacles inside a genotype×phenotype dataset, which may arise from sample structure, relatedness between the genotyped individuals or marginal effects of single SNPs on the phenotype [71, 85, 86], my method incorporates an additional step using the average product correction (APC) theorem [88] to estimate the expected level of background association for each SNP pair. Finally, the removal of the estimated background from the measured epistasis values leads to the detection of correct epistatic signals arising from functional interactions. MIDESP has been published in [165] (see Appendix A.3). The method consists of several steps, which I explain in detail in the following. An overview of the MIDESP workflow is given in Figure 4.2. Additionally, I describe in a final section how I evaluated the epistatic SNP pairs found with this method.

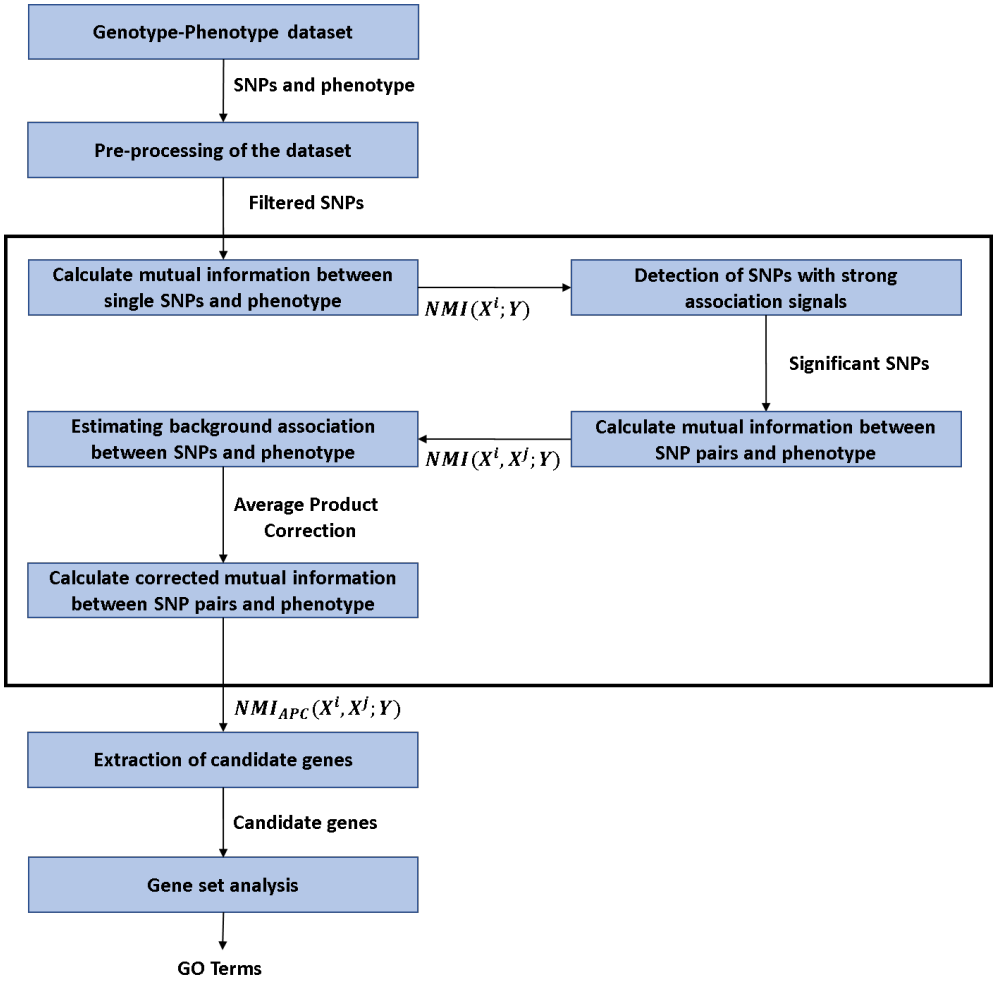


Figure 4.2.: Flowchart of the MIDESP method.

4.3.1. Pre-processing

Given a genotype×phenotype dataset, I first apply several filters to it, which are described in Section 4.1.4, to ensure the quality of the SNPs. As a second pre-processing step, I perform linkage disequilibrium (LD) pruning to remove SNPs that have redundant information about the phenotype. This pruning has two main benefits. First, it reduces the size of the input and thereby decreases the runtime of the program. Second, this pruning allows us to obtain epistasis results without confounding them through LD [212]. For pruning, I utilize PLINK [135] to remove all redundant SNPs with an LD ≥ 0.99, and thus carrying very similar information about the phenotype.

Based on the number of samples, $\mathcal{N}$, and the number of SNPs, $\mathcal{P}$, I consider a filtered and pruned genotype×phenotype dataset as a matrix, $M_{\mathcal{N}\times(\mathcal{P}+1)}$, where the rows refer to the samples and the columns refer to the phenotype and the SNPs. Furthermore, the phenotype of interest is denoted by Y^D and Y^C for qualitative (discrete) and quantitative (continuous) traits, respectively. Let S^i be a sample, let X^j be the genotypes of an SNP and let Y^i be the corresponding phenotype of S^i. The entry of M at position (i,j) is depicted by X^i_j. In the following, I also use X and Y as placeholders for any of the SNPs or phenotypes, respectively.

4.3.2. Application of mutual information for SNP×phenotype associations

MIDESP utilizes the mutual information as a measure for the strength of the association between the phenotype and a SNP. For qualitative phenotypes, the standard Equation 3.2.5 for mutual information can be applied as

$$\mathbb{MI}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y), \tag{4.3.1}$$

where X is a SNP and Y is the discrete phenotype. In the case of a quantitative phenotype, I estimate the mutual information using Ross' estimator, which I presented in Section 3.2.4.1, as

$$\mathbb{MI}(X;Y) = \frac{1}{\mathcal{N}} \cdot \sum_{i=1}^{\mathcal{N}} (\psi(\mathcal{N}) - \psi(\mathcal{N}_{x_i}) + \psi(k) - \psi(m_i)), \tag{4.3.2}$$

where:

- $\psi(\cdot)$ is the digamma function;
- $\mathcal{N}_{x_i}$ for a given sample, S^i, refers to the number of samples for which the genotype x is the same as the genotype x_i of S^i;
- d is the distance between sample S^i and its kth-nearest neighbor S^{i_k} with the same genotype as S^i, defined as the absolute difference between their phenotypes Y^i and Y^{i_k};

- m_i is assigned the number of samples where the absolute difference between their phenotypes and the phenotype Y^i is less than or equal to d, irrespective of the genotypes.

The identification of these values is shown for a toy dataset in the following example.

Example: Estimation of mutual information between a SNP and a quantitative phenotype

For the purpose of this example, I use the estimator with $k = 3$.
Let X be a SNP and Y a quantitative phenotype. Their values are given for 10 samples S1, S2, ... to S10. Based on these values, N_x is defined as the number of samples where the genotype is equal to x.

A

Sample	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
X	AA	AA	Aa	AA	aa	Aa	AA	AA	Aa	Aa
Y	7.1	0.9	8.9	4.4	1.3	9.4	7.2	4.2	9.9	4.3

B

X	AA	Aa	aa
N_X	5	4	1

For each sample S_i, a sorted list of the samples is created based on the absolute difference between Y_i and Y_j for sample S_j.

C

Sample	S1	S7	S3	S6	S4	S10	S9	S8	S5	S2
X	AA	AA	Aa	Aa	AA	Aa	Aa	AA	aa	AA
Y	7.1	7.2	8.9	9.4	4.4	4.3	9.9	4.2	1.3	0.9

Sorted based on distance to S1

Sample	S2	S5	S8	S10	S4	S1	S7	S3	S6	S9
X	AA	aa	AA	Aa	AA	AA	AA	Aa	Aa	Aa
Y	0.9	1.3	4.2	4.3	4.4	7.1	7.2	8.9	9.4	9.9

Sorted based on distance to S2

⋮

Sample	S10	S8	S4	S1	S7	S5	S2	S3	S6	S9
X	Aa	AA	AA	AA	AA	aa	AA	Aa	Aa	Aa
Y	4.3	4.2	4.4	7.1	7.2	1.3	0.9	8.9	9.4	9.9

Sorted based on distance to S10

The kth-nearest neighbor is determined for each sorted list by going along the list and counting the samples that have the same X value as the start sample. m_{S_i} can then be defined as the index of the kth-nearest neighbor in the sorted list.

D

Sample	S1	S7	S3	S6	S4	S10	S9	S8	S5	S2
X	AA	AA	Aa	Aa	AA	Aa	Aa	AA	aa	AA
Y	7.1	7.2	8.9	9.4	4.4	4.3	9.9	4.2	1.3	0.9
k	0	1			2			3		
m	0	1	2	3	4	5	6	7		

Sorted based on distance to S1

$m_{S_1} = 7$

Sample	S2	S5	S8	S10	S4	S1	S7	S3	S6	S9
X	AA	aa	AA	Aa	AA	AA	AA	Aa	Aa	Aa
Y	0.9	1.3	4.2	4.3	4.4	7.1	7.2	8.9	9.4	9.9
k	0		1		2	3				
m	0	1	2	3	4	5				

Sorted based on distance to S2

$m_{S_2} = 5$

⋮

Sample	S10	S8	S4	S1	S7	S5	S2	S3	S6	S9
X	Aa	AA	AA	AA	AA	aa	AA	Aa	Aa	Aa
Y	4.3	4.2	4.4	7.1	7.2	1.3	0.9	8.9	9.4	9.9
k	0							1	2	3
m	0	1	2	3	4	5	6	7	8	9

Sorted based on distance to S10

$m_{S_{10}} = 9$

For sample S1 which has the X value AA, the sample with the third-closest Y and the same X value is sample S8, which has the index 7 in the sorted list. Therefore, $m_{S_1} = 7$. Based on the N_x and m_{S_i} values, the mutual information can be estimated. Images are taken from [165].

As shown in the example, only the phenotype Y is a continuous variable, hence in general, the sorted tables can be reused for every SNP by only changing the values of X. This allows for an efficient calculation of m_i. Since the mutual information is only estimated, the resulting values can be outside the range of the valid interval, i.e., $[0, \mathbb{H}(X)]$. Thus, the estimated values outside of this range are set to the closest interval boundary.

If the values of the continuous phenotype Y are not unique, then a deterministic sort to find the kth-nearest neighbor is not possible. To address this problem, I initially, as suggested in [156], added very small noise ($R \sim \mathcal{N}(0, 10^{-10})$) to the phenotype Y, which guarantees the uniqueness of the values, followed by normalizing them to the interval $[0, 1]$ before applying the estimator. However, this approach is not suitable if the values of Y have a high degree of repetitiveness in which case the results of the estimator strongly depend on the random noise. Therefore, I instead use the original phenotype values and adapted the estimator accordingly. For this I exploit the property of Equation 4.3.2 that an average is calculated. This permits me to vary k for every sample S^i. An example for the problem with duplicated values is given in Figure 4.3. The phenotype values of the samples S4, S8, S9 and S10 are all the same and have the same distance to the phenotype of sample S1. Therefore, many different orders are possible, of which two are shown here. Applying the default value of $k = 3$ for sample S1 would result in different values for m_{S_1} depending on the chosen

order. To solve this problem, I increase the value of k used for this specific sample until all duplicate values are included in the interval defined by the kth-nearest neighbor and sort them so that the samples with the same genotype value as S1 among the duplicates (here S4 and S10) are at the end. In this example, increasing k to 4 would be enough to include all duplicate values, which is depicted in Figure 4.4.

Sample	S1	S7	S3	S6	**S4**	**S10**	**S9**	**S8**	S5	S2
X	AA	AA	AA	Aa	AA	AA	aa	Aa	Aa	aa
Y	0	2	5	5	7	7	7	7	8	9
k	0	1	2		3					
m	0	1	2	3	4					

Sorted based on distance to S1
$m_{S_1} = 4$
$k = 3$

Sample	S1	S7	S3	S6	**S9**	**S8**	**S10**	**S4**	S5	S2
X	AA	AA	AA	Aa	aa	Aa	AA	AA	Aa	aa
Y	0	2	5	5	7	7	7	7	8	9
k	0	1	2				3			
m	0	1	2	3	4	5	6			

Sorted based on distance to S1
$m_{S_1} = 6$
$k = 3$

Figure 4.3.: This example shows a set of 10 samples with a quantitative phenotype Y. The samples marked in **bold** have the same phenotype value. When sorting the samples according to the phenotype distance to sample S1 multiple orders are possible, of which two are shown. Depending on the chosen order, the values of m_{S_1} vary and thereby also the mutual information between SNP and phenotype.

Sample	S1	S7	S3	S6	**S9**	**S8**	**S4**	**S10**	S5	S2
X	AA	AA	AA	Aa	aa	Aa	AA	AA	Aa	aa
Y	0	2	5	5	7	7	7	7	8	9
k	0	1	2				3	4		
m	0	1	2	3	4	5	6	7		

Sorted based on distance to S1
$m_{S_1} = 7$
$k = 4$

Figure 4.4.: This example shows the same set of 10 samples with a quantitative phenotype as Figure 4.3. By using $k = 4$ and placing the samples with the same genotype value as S1 at the end, all samples with the same phenotype value of 7 are placed into the interval defined by the kth-nearest neighbor, which results in $m_{S_1} = 7$. The mutual information can now be estimated unambiguously.

Furthermore, there are two additional special cases that have to be considered for the estimator. On the one hand, it is possible that N_{x_i} is smaller than the predefined k for a specific genotype of the SNP, so that a kth-nearest neighbor for sample S^i does not exist. On the other hand, the extreme case can also occur that there is only a single sample with a specific genotype and, therefore, no neighbors exist at all. To address these two issues, I followed the implementation provided by Ross in [87]. In the case where there are simply fewer than k samples available as neighbors, I again exploit the previously mentioned property of Equation 4.3.2 that an average is calculated and k can therefore be varied for each sample S^i. If the number of samples with the current genotype, N_{x_i}, is less than the predefined k, I set k for this specific sample S^i to $N_{x_i} - 1$ and, thereby, to the maximum possible number of neighbors available in this situation. For the second case, if only a single sample S^i has the specific genotype, then this sample is uninformative for the purpose of the estimator. To signify this, k is set to 1 and m_i to $2 \times |\mathcal{X}|$ where $|\mathcal{X}|$ is the number of unique genotypes occurring in this SNP, respectively, the alphabet size of the SNP. These values simulate the situation that each existing genotype occurs two times next to this sample until the kth-nearest neighbor is reached and thereby this genotype would not be representative for the surrounding interval. This strategy is visualized in Figure 4.5.

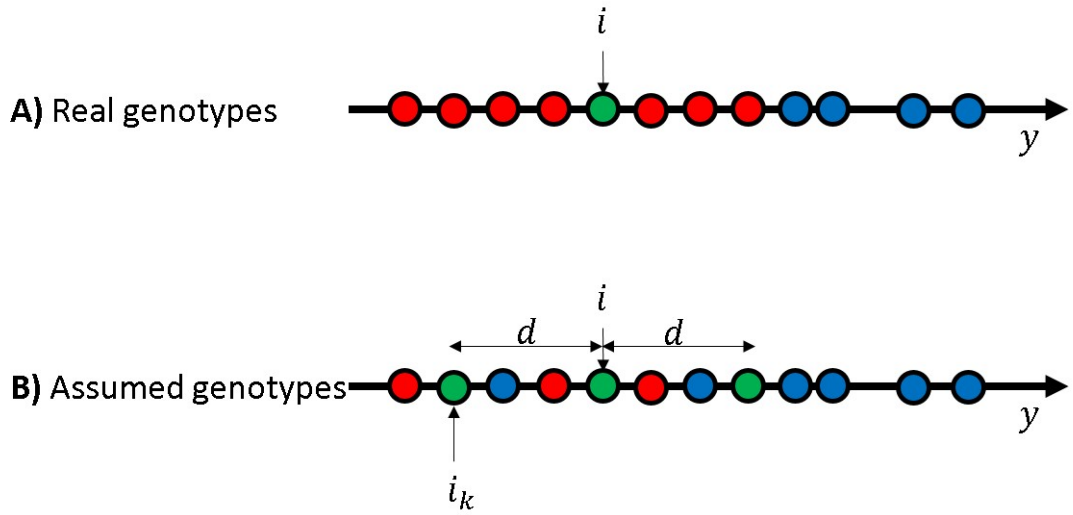


Figure 4.5.: (A) shows the visualization of the genotypes for 12 samples sorted according to a continuous phenotype Y. The samples are colored by their respective genotype. Sample S_i has a genotype which occurs only once among the samples. In this situation m_i is set to $2 \times |\mathcal{X}|$ (here 2×3) to indicate that this genotype is not representative for the surrounding interval. This means that each possible genotype occurs twice around S_i until the kth-nearest neighbor is reached. This assumed situation is represented in (B).

Similar to previous studies [76, 81, 83, 84], MIDESP applies the joint mutual information

for the detection of an epistatic interaction between the phenotype Y and a SNP pair, X^i and X^j. For this Equations 4.3.1 and 4.3.2 are extended to

$$\mathbb{MII}(X^i,X^j;Y) = \mathbb{H}(X^i,X^j) + \mathbb{H}(Y) - \mathbb{H}(X^i,X^j,Y) \tag{4.3.3}$$

for qualitative phenotypes and

$$\mathbb{MII}(X^i,X^j;Y) = \frac{1}{\mathcal{N}} \cdot \sum_{l=1}^{\mathcal{N}} (\psi(\mathcal{N}) - \psi(\mathcal{N}_{x_l^{ij}}) + \psi(k) - \psi(m_l)) \tag{4.3.4}$$

for quantitative phenotypes. In Equation 4.3.4, x_l^{ij} refers to the joint genotype of the SNP pair X^i and X^j for sample S^l.
For an increased computational performance, I use a bit-shifting-based approach to combine the genotypes of two SNPs to a value which is unique for the specific combination. This approach allows to quickly determine the frequencies of the genotype combinations and whether two samples have the same combination, which is necessary to calculate $\mathbb{MII}$. Similar approaches have been used in various other methods for epistasis detection [81, 166, 213].
As shown in [88, 159, 214], the value of the mutual information and its possible range is strongly dependent on the alphabet size and the marginal distributions of the variables. A normalization of $\mathbb{MII}$ to $\mathbb{NMII}$ as introduced in Section 3.2.5 is therefore required to address this influence and to make them comparable with each other for further analysis. While in general all participating variables are used for normalization, this is not possible if one or more variables are continuous [159]. Instead, I apply the following normalization technique, which is only based on the entropy and the maximum possible alphabet size of the SNP or SNP pair. Consequently, the $\mathbb{MII}(X;Y)$- and $\mathbb{MII}(X^i,X^j;Y)$-values are normalized as

$$\mathbb{NMII}(...;Y) = 2 \cdot \frac{\mathbb{MII}(...;Y)}{\log(\max|\mathcal{X}|) + \mathbb{H}(...)}. \tag{4.3.5}$$

4.3.3. Detection of SNPs with strong association signals

It can be easily seen that the calculation of the pairwise interactions between all SNP pairs would require a quadratic runtime. Therefore, the separation of SNPs with strong association signals from the remaining ones is necessary to reduce the number of pairs under study. MIDESP performs such a separation based on the normalized mutual information $\mathbb{NMII}(X^i;Y)$ values between each SNP X^i and the phenotype Y, which are calculated in the first step of the workflow.
For this purpose, Gültas et al. [214, 215] showed that by extending the standard multiple testing theory [216, 217], the $\mathbb{NMII}$ values can be modeled based on three different distributions:

(i) a β distribution F_0 (null distribution) representing the background signals;
(ii) a G_1 distribution referring to the unrelated associations (in this case between SNPs and phenotype);
(iii) a G_2 distribution modeling the strong association signals (in this case between SNPs and phenotype).

The β distribution for a continuous random variable Z is defined over the interval $[0,1]$ using two shape parameters α and β. For a given sample, the shape parameters of the corresponding β distribution can be estimated using the sample mean $\widehat{\mu}$ and variance $\widehat{\sigma^2}$ [218] as

$$\widehat{\alpha} = \widehat{\mu}\left(\frac{\widehat{\mu}(1-\widehat{\mu})}{\widehat{\sigma^2}} - 1\right) \tag{4.3.6}$$

and

$$\widehat{\beta} = (1-\widehat{\mu})\left(\frac{\widehat{\mu}(1-\widehat{\mu})}{\widehat{\sigma^2}} - 1\right). \tag{4.3.7}$$

Let $\mathbb{NMII}_t$ be $\mathbb{NMII}(X^t, Y)$. I estimate the β distribution for the $\mathbb{NMII}$ values $\mathbb{NMII}_1$, $\mathbb{NMII}_2$, ..., $\mathbb{NMII}_{\mathcal{P}}$ using their mean and variance and can thereby calculate the p–value for the association of a SNP X^t to the phenotype with respect to F_0 as

$$1 - F_0(\mathbb{NMII}_t) = P\{\text{random } F_0 \text{ -distributed value} \geq \mathbb{NMII}_t\}. \tag{4.3.8}$$

The p–value is uniformly distributed over $[0,1]$ if $\mathbb{NMII}_t$ is F_0-distributed. However, if X^t belongs to the G_1 distribution of unrelated SNPs, its corresponding p–value is skewed towards 1. Similarly, if X^t is G_2 distributed, its p–value is skewed towards 0. This separation is visualized in Figure 4.6. As the next step, based on two tuning parameters, λ_1 and λ_2, the fraction γ of the $\mathbb{NMII}_t$ belonging to the background is estimated using Equation 4.3.9:

$$\widehat{\gamma} = \frac{\text{number of } p\text{–values in } [\lambda_1, \lambda_2]}{\mathcal{P} \cdot (\lambda_2 - \lambda_1)} \tag{4.3.9}$$

so that the fraction of non-uniformly distributed p–values that fall into $[\lambda_1, \lambda_2]$ is negligible [216, 219]. These two parameters are dataset-dependent and are automatically tuned through a trial and error heuristic approach during the analysis [218].
Finally, a SNP X^t is deemed as significant if its p–value is less or equal to τ, where τ is a threshold depending on a user-defined false discovery rate, FDR, estimated using Equation 4.3.10.

$$\widehat{FDR}(\tau) = \frac{\widehat{\gamma} \cdot \mathcal{P} \cdot \tau}{\text{number of } p\text{–values } \leq \tau} \tag{4.3.10}$$

For the detection of epistatic interactions using the $\mathbb{NMII}(X^i, X^j; Y)$ metric, in the further analysis I only consider SNP pairs where at least one SNP is significant, which results in a

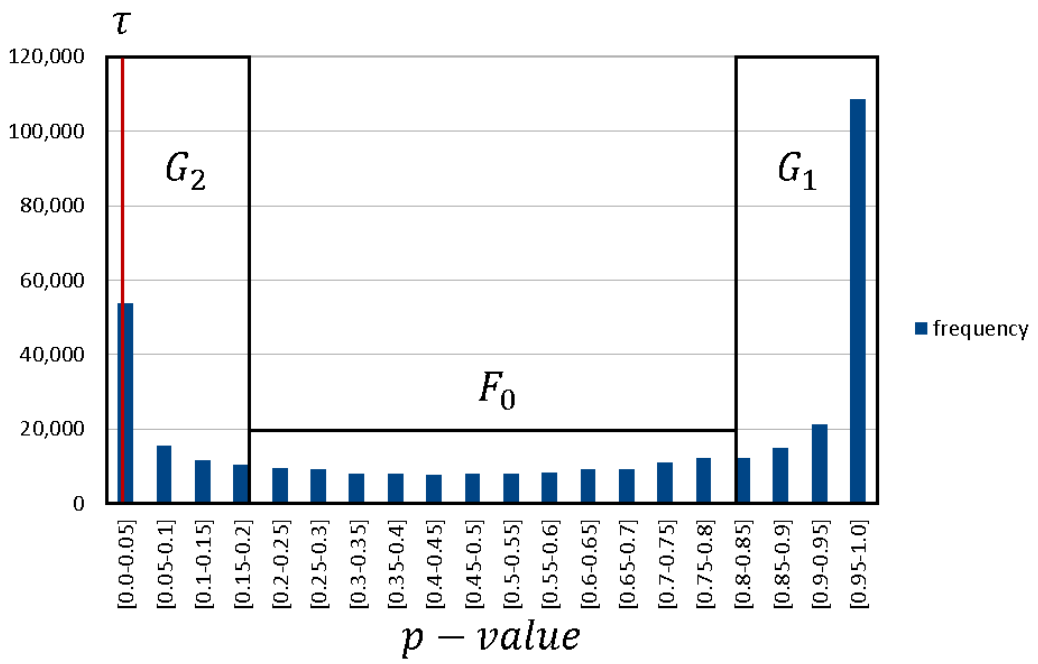


Figure 4.6.: Distribution of p–values for the 616,398 SNPs of the BT dataset: the distribution can be divided in three parts, with G_2 representing the strongly associated SNPs, G_1 the unrelated SNPs and F_0 the background. SNPs with a p–value less than or equal to the threshold τ (indicated by the red line) are deemed as significant.

significant reduction of the runtime.

4.3.4. Reduction of the background associations between SNPs and phenotype

As shown in previous studies [71, 85, 86], a dataset-dependent background association exists between the SNPs and the phenotype that may arise due to population stratification or relatedness of the individuals under study. Such phenomena could interfere with the identification of the correct epistatic signals, and thus could lead to the detection of false positive association signals. Another background association could occur in the detection of epistatic interactions using the $\mathbb{NMI}$ metric due to high levels of mutual information between a single SNP and the phenotype. I introduced this issue by way of an example in Section 3.3.2.1.

To eliminate these issues to some extent, I apply the average product correction (APC) introduced by Dunn et al. [88]. The APC theorem is a very successful information-theory-based approach to estimate the expected level of background association between the variables in a

dataset. Meckbach et al. [220] showed that this approach is universally applicable, and thus I adapted it for my method. Following this approach, I estimate the expected level of the background between the SNP pair and the phenotype in the calculation of $\mathbb{NMI}(X^i, X^j; Y)$ as

$$APC(X^i, X^j; Y) = \left(\frac{\overline{\mathbb{NMI}_{X^i}} \cdot \overline{\mathbb{NMI}_{X^j}}}{\overline{\mathbb{NMI}_{\text{SNP}}}}\right) \tag{4.3.11}$$

In Equation 4.3.11, $\overline{\mathbb{NMI}_{X^i}}$ and $\overline{\mathbb{NMI}_{X^j}}$ are the average association levels of the SNPs X^i and X^j, respectively, in the epistatic interaction:

$$\overline{\mathbb{NMI}_{X^i}} = \frac{1}{h} \cdot \sum_{l=1}^{h} \mathbb{NMI}(X^i, X^l; Y), \tag{4.3.12}$$

where h is a sufficiently large number (e.g., $h > 1000$) and the SNPs X^l are randomly chosen. Further, $\overline{\mathbb{NMI}_{\text{SNP}}}$ denotes the overall average normalized mutual information calculated using a sufficiently large number of $\mathbb{NMI}$ values.
Finally, I subtract the $\text{APC}(X^i, X^j; Y)$ value of an SNP pair and the phenotype from their initial $\mathbb{NMI}(X^i, X^j; Y)$ to obtain the corrected $\mathbb{NMI}_{\text{APC}}(X^i, X^j; Y)$.

$$\mathbb{NMI}_{\text{APC}}(X^i, X^j; Y) = \mathbb{NMI}(X^i, X^j; Y) - \text{APC}(X^i, X^j; Y) \tag{4.3.13}$$

4.3.5. Validation of the epistatic interactions

To identify the genes pertaining to epistatic SNP pairs, I only consider the p-th percentile of the pairs with an $\mathbb{NMI}_{\text{APC}}$ value > 0. For the interpretation of the interactions, I replace the SNPs with their corresponding genes based on the information provided by the Ensembl database (release 103) [125]. The data are then read into R and a gene-gene interactions network is created with the genes as nodes and their interactions as edges using the igraph package [221]. The degree of a node, thereby, gives the number of its interactions. In the final step, these degrees are transformed into z-scores and I subsequently define a gene as MIDESP-significant if its z-score is ≥ 3, as suggested in [220].
To elucidate the biological functions of these genes, I follow previous studies [197, 222] and utilize the geneXplain platform [139] to perform a gene set analysis based on the molecular functions of the genes. The results are then visualized in the form of a treemap.

4.3.6. Implementation

The MIDESP pipeline is implemented in Java and its source code as well as a JAR file are available at `https://github.com/FelixHeinrich/MIDESP`, which makes it easy to use. The calculations are completely parallelized, permitting an efficient detection of significant epistatic interactions with a multi-core CPU. Genotype and phenotype information in the form of tped and tfam files are required as input. These are transposed versions of the regular

ped/map format used by PLINK [134], with the tped containing the genotype information of a single SNP in each row and the tfam having a row for the information of each individual in the dataset.

5. Results

In this chapter, I present the results of the two analysis frameworks that I described in the previous chapter. First, I focus on the identification of promoters and regulatory SNPs using an inter-species trained neural network and the analysis of the corresponding results of the *Vicia faba* dataset. In the second part, I demonstrate different aspects of the MIDESP algorithm as well as an analysis of its results on two real datasets. In a final section, I present the results of the application of MIDESP on the *Vicia faba* data. This section is for the most part based on my published papers [52, 165] (see Appendix A.1 and Appendix A.3).

5.1. Identification of regulatory SNPs based on genotyping by sequencing data in *Vicia faba* using deep learning

In this section, I present the results for the identification of regulatory SNPs in *Vicia faba*. For the first parts, I show the performance of the promoter detection in general on the species under study as well as the impact of additional features on the detection quality. After that, I focus on the results on the *Vicia faba* dataset and display the identified regulatory SNPs along with the affected transcription factors. Finally, I present the results of a functional analysis of the found candidate gene and transcription factors.

5.1.1. Intra- and inter-species promoter prediction

In order to gain first insights into the predictability of promoters of the seven *Leguminosae* family members and *Arabidopsis thaliana*, I first trained the CNN model for each species individually. The prediction reliability of the CNN model has been examined for each species by classifying the intra- and inter-species promoters that were not used in the training process. To assess the performance of the classification, the Accuracy (ACC), Sensitivity (SEN), Specificity (SPE) and Matthews Correlation Coefficient (MCC) values were calculated. The details of these measures are given in Tables 5.1 to 5.4.

Table 5.1.: ACC values of the intra- and inter-species promoter classification using the species-specific CNNs. Off-diagonal numbers are ACC values for inter-species classification, diagonal numbers are ACC values for intra-species classification.

Evaluated / Trained	*Arabidopsis thaliana*	*Glycine max*	*Lupinus angustifolius*	*Medicago truncatula*	*Phaseolus vulgaris*	*Trifolium pratense*	*Vigna angularis*	*Vigna radiata*
Arabidopsis thaliana	0.901	0.767	0.690	0.746	0.797	0.765	0.633	0.733
Glycine max	0.837	0.864	0.915	0.847	0.863	0.724	0.914	0.856
Lupinus angustifolius	0.545	0.611	0.981	0.720	0.586	0.493	0.974	0.709
Medicago truncatula	0.755	0.797	0.959	0.876	0.789	0.715	0.951	0.841
Phaseolus vulgaris	0.845	0.842	0.888	0.834	0.898	0.748	0.880	0.853
Trifolium pratense	0.822	0.764	0.696	0.751	0.794	0.840	0.689	0.736
Vigna angularis	0.544	0.607	0.971	0.715	0.583	0.494	0.977	0.712
Vigna radiata	0.741	0.812	0.937	0.827	0.825	0.675	0.928	0.904

Table 5.2.: SEN values of the intra- and inter-species promoter classification using the species-specific CNNs. Off-diagonal numbers are sensitivity values for inter-species classification, diagonal numbers are sensitivity values for intra-species classification.

Evaluated / Trained	*Arabidopsis thaliana*	*Glycine max*	*Lupinus angustifolius*	*Medicago truncatula*	*Phaseolus vulgaris*	*Trifolium pratense*	*Vigna angularis*	*Vigna radiata*
Arabidopsis thaliana	0.88	0.76	0.602	0.778	0.769	0.838	0.459	0.672
Glycine max	0.775	0.882	0.99	0.906	0.852	0.662	0.98	0.863
Lupinus angustifolius	0.126	0.258	0.996	0.477	0.206	0.024	0.984	0.466
Medicago truncatula	0.552	0.681	0.993	0.841	0.646	0.468	0.982	0.752
Phaseolus vulgaris	0.767	0.85	0.945	0.887	0.907	0.707	0.909	0.85
Trifolium pratense	0.79	0.76	0.616	0.768	0.786	0.913	0.531	0.69
Vigna angularis	0.1203	0.246	0.977	0.46	0.196	0.022	0.978	0.454
Vigna radiata	0.57	0.758	0.975	0.794	0.758	0.5	0.971	0.885

Table 5.3.: SPE values of the intra- and inter-species promoter classification using the species-specific CNNs. Off-diagonal numbers are specificity values for inter-species and diagonal numbers are specificity values for intra-species classification.

Trained \ Evaluated	*Arabidopsis thaliana*	*Glycine max*	*Lupinus angustifolius*	*Medicago truncatula*	*Phaseolus vulgaris*	*Trifolium pratense*	*Vigna angularis*	*Vigna radiata*
Arabidopsis thaliana	0.921	0.775	0.797	0.714	0.835	0.693	0.808	0.794
Glycine max	0.899	0.845	0.837	0.789	0.873	0.786	0.847	0.848
Lupinus angustifolius	0.964	0.964	0.966	0.962	0.966	0.962	0.965	0.962
Medicago truncatula	0.946	0.923	0.926	0.928	0.94	0.907	0.93	0.932
Phaseolus vulgaris	0.91	0.836	0.839	0.788	0.895	0.779	0.851	0.852
Trifolium pratense	0.855	0.786	0.775	0.735	0.803	0.768	0.786	0.782
Vigna angularis	0.967	0.969	0.966	0.969	0.969	0.966	0.977	0.971
Vigna radiata	0.912	0.865	0.899	0.86	0.891	0.851	0.885	0.924

Table 5.4.: MCC values of the intra- and inter-species promoter classification using the species-specific CNNs. Off-diagonal numbers are MCC values for inter-species and diagonal numbers are MCC values for intra-species classification.

Trained \ Evaluated	*Arabidopsis thaliana*	*Glycine max*	*Lupinus angustifolius*	*Medicago truncatula*	*Phaseolus vulgaris*	*Trifolium pratense*	*Vigna angularis*	*Vigna radiata*
Arabidopsis thaliana	0.8	0.53	0.41	0.49	0.59	0.54	0.28	0.47
Glycine max	0.68	0.73	0.84	0.7	0.73	0.45	0.83	0.71
Lupinus angustifolius	0.16	0.31	0.96	0.5	0.26	-0.04	0.95	0.49
Medicago truncatula	0.54	0.62	0.92	0.77	0.61	0.42	0.91	0.7
Phaseolus vulgaris	0.68	0.69	0.79	0.68	0.8	0.49	0.76	0.7
Trifolium pratense	0.65	0.53	0.4	0.5	0.59	0.69	0.33	0.47
Vigna angularis	0.16	0.31	0.94	0.5	0.26	-0.04	0.96	0.5
Vigna radiata	0.51	0.63	0.88	0.65	0.66	0.37	0.86	0.81

The results presented in these tables show that, although the CNN models have been trained only using one-hot representation of sequences for each species individually, the network architecture is able to recognize specific patterns in the sequences, which leads to the predictability of promoters across different species to a high degree. These findings support the results presented in [50] and indicate that some of the promoter signatures seem to be conserved between the *Leguminosae* family members.

Furthermore, Table 5.1 demonstrates that the classification performance of some CNN models yields markedly higher ACC values for inter-species prediction than for intra-species prediction. In particular, this is the case for the species *Lupinus angustifolius* and *Vigna angularis*, whose promoters were predicted with very high accuracy by the CNN models of the other species, with the exception of the models for *Arabidopsis thaliana* and *Trifolium pratense*. This could be attributed to the underlying genome annotations of these species, since their annotations seem to be partially created based on the genome information of well-studied family members [223, 224]. Especially, this assumption appears to be true with respect to the inclusion of the *Leguminosae* family specific promoter patterns in the promoters of these species. This hypothesis has been supported by the prediction performance of the *Lupinus angustifolius* as well as the *Vigna angularis* models in the other species, which achieved very high degrees of Specificity while achieving low degrees of Sensitivity (see Tables 5.2 and 5.3). To this end, I compared the inter-species prediction ACC values of the species regarding their performance on *Lupinus angustifolius* and *Vigna angularis*. This comparison revealed that the promoters of *Lupinus angustifolius* were predicted with a slightly higher mean accuracy value than the promoters of *Vigna angularis* (0.880 vs. 0.868).

5.1.2. Effect of additional features on model prediction

The results shown so far were obtained from CNN models that were trained using only the order of the nucleotides in form of the one-hot representation of the DNA sequences. However, previous studies pointed out that the combination of one-hot encoding with additional specific features could result in a substantially improved performance in promoter identification [46, 47]. For this purpose, I followed a procedure similar to the one which was suggested by Triska et al. in [46] and systematically evaluated the combination of different sequence features with the one-hot representation for the CNN model training for each species. The results are presented for *Medicago truncatula* in Table 5.5 as an example.

In contrast to previous studies [46, 47], Table 5.5 shows that regardless of the usage of any additional feature, the performance of the CNN model could in general not be significantly improved.

Table 5.5.: Contribution of additional features in the CNN model of *Medicago truncatula*.

Features	Accuracy	Sensitivity	Specificity	MCC
DNA sequence	0.876	0.897	0.855	0.750
DNA sequence + 2-mer	0.874	0.880	0.867	0.747
DNA sequence + 2-mer + frequency of CA motif	0.862	0.828	0.897	0.726
DNA sequence + 2-mer + frequency of CG motif	0.875	0.875	0.876	0.751
DNA sequence + 2-mer + HMI	0.874	0.882	0.865	0.747
DNA sequence + 2-mer + frequency of TATAA motif	0.876	0.875	0.878	0.752
DNA sequence + 2-mer + CG skew	0.876	0.889	0.863	0.753
DNA sequence + topological entropy	0.874	0.886	0.861	0.747
DNA sequence + 2-mer + topological entropy	0.871	0.852	0.890	0.743
DNA sequence + 2-mer + HMI + frequency of TATAA motif	0.871	0.869	0.874	0.743
DNA sequence + 2-mer + HMI + frequency of CA motif + frequency of CG motif + frequency of TATAA motif + CC skew	0.873	0.859	0.888	0.747
DNA sequence + HMI + frequency of CA motif + frequency of CG motif + frequency of TATAA motif + CG skew	0.875	0.889	0.860	0.749

5.1.3. Prediction of *Vicia faba* promoters

The knowledge about the promoter signatures which are conserved between the *Leguminosae* family members provides an important clue for the precise prediction of *Vicia faba* promoters, which still remains a challenge. However, the consideration of the sequences of only one *Leguminosae* species in the CNN model could be insufficient to capture the variety of different promoter signatures that are necessary for the accurate computational identification of the *Vicia faba* promoters. To mitigate the drawback of single species models, I systematically examined different CNN models seeking to determine the preferential combination of *Leguminosae* family members by intensifying the signal from promoter sequences and thus improving the performance of the CNN model. Consequently, I trained a final CNN model based on the species *Lupinus angustifolius* and *Medicago truncatula* since the combined usage of their manually selected sequences complement each other. In the last step, I included two additional non-promoter sets (defined in Section 4.1.3) in the training of the CNN model to enhance the discriminating signals between promoter and non-promoter regions. The evaluation of this CNN model yields to clearly better ACC and MCC values of 0.98 and 0.95, respectively. A further analysis reveals that the usage of other sequence features together with one-hot encoding in my final CNN model does not affect the performance of the classifier. Finally, by applying this CNN model to the *Vicia faba* sequences of length 250bp, I classified in total 2.46 % of them as potential promoter

sequences.

5.1.4. SNPs in putative promoter regions and their association with V+C

Examination of the positions of SNPs in the contigs revealed that in total 132,399 out of 685,215 SNPs are located in the predicted *Vicia faba* promoters. A flanking sequence of $\pm$25 bp could be obtained for only 118,492 SNPs. Mapping these SNPs to the *Medicago truncatula* genome resulted in 33,846 hits for 1976 SNPs, which demonstrates the repetitiveness of the *Medicago truncatula* genome. I identified 14 SNPs that map to the predefined target region of *Medicago truncatula* that harbours orthologous genes associated with the V+C content of *Vicia faba* [21, 23]. This target region is ranging approximately from 1,300,000 bp to 2,300,000 bp of the *Medicago truncatula* chromosome 2. An overview of these SNPs and their mapped position in the *Medicago truncatula* genome along with the genes with the closest TSS is given in Table 5.6. I tested these 14 SNPs for their association with V+C content with PLINK. The adjusted p-values presented in Table 5.6 suggest that SNP_341016_236 and SNP_341016_239, which are located in the same promoter, show a highly significant association with the V+C content in *Vicia faba* while the associations of the remaining SNPs are not significant at the level $\alpha = 0.05$. For both of these SNPs, the reference allele only occurs in the low V+C lines with one exception, while the alternate allele is restricted to the high V+C lines (see Table 5.7).

5.1.5. Systematic identification of regulatory SNPs associated with V+C in *Vicia faba*

Following the studies of Xu et al. and Fu et al. [210, 211], I scanned the flanking sequences of the SNPs by applying the MATCH^{TM} program [136] to systematically identify the SNPs that are likely to affect the binding affinity of transcription factors (TFs) and, thus, influence the gene expression level. This search was done for the 1976 SNPs that were located in the predicted *Vicia faba* promoters and that could be successfully mapped onto the *Medicago truncatula* genome. I considered the results of this run with an MSS score ≥ 0.85 as putative TFBSs as suggested in [225]. 9444 putative TFBSs were identified and SNPs that were located in those TFBSs were considered as rSNPs. Their consequence types were determined by examining their predicted effects on the binding affinities of the TFs. The analysis of the 14 SNPs presented in Table 5.6 reveals that the binding affinities of 44 TFs to their 79 TFBSs were affected. Focusing on the two highly significant SNPs (SNP_341016_236 and SNP_341016_239) in the same promoter, I found that a nucleotide substitution in SNP_341016_236 is likely to entail severe consequences regarding TF binding affinity, namely loss and gain of TFBSs. The substitution in SNP_341016_239 results in only a moderate change of the binding affinities of TFs (see Table 5.8). The remarkably different consequences of both SNPs indicate their considerably different influence on the

Table 5.6.: The 14 SNPs found in the predicted promoters of *Vicia faba* that were mapped to the *Medicago truncatula* target genomic region.

SNP_ID	Genotype	FDR	Position	Medicago gene
SNP_131938_118	C/T	0.234	1,385,390	MTR_2g008290
SNP_302904_183	G/A	0.179	1,385,444	
SNP_341016_236	C/T	$1.17 \cdot 10^{-7}$	1,554,857	MTR_2g008620
SNP_341016_239	G/A	$1.17 \cdot 10^{-7}$	1,554,860	
SNP_356745_200	A/G	0.730	1,707,078	MTR_2g008960
SNP_280549_41	C/T	0.234	1,707,183	
SNP_350273_103	G/T	0.234	1,707,199	
SNP_350273_90	A/C	0.234	1,707,212	
SNP_350273_61	G/A	0.234	1,912,704	MTR_2g009430
SNP_29452_204	G/A	0.730	1,912,812	
SNP_29452_206	G/A	0.496	1,912,814	
SNP_118828_190	C/T	0.234	2,030,017	MTR_2g009690
SNP_80231_27	C/T	0.234	2,163,048	MTR_2g009940
SNP_364434_97	A/T	0.359	2,163,084	

precise and effective regulation of the corresponding gene, although their p-values are the same.

5.1.6. Functional analysis of the candidate gene and transcription factors

The *Medicago truncatula* gene MTR_2g008620 is the gene that is located closest to the two highly significant SNPs. It is a β-hydroxyacyl-ACP-dehydratase that is involved in the elongation of fatty acids as well as in the related metabolism of biotin [226, 227]. A direct association with the synthesis of V+C, which has been linked to the orotic acid pathway [22], is not obvious. This seems plausible since *Medicago truncatula* does not synthesize V+C. Of greater interest are the transcription factors for which I found putative binding sites that are affected by the two SNPs. The TF SQUA belongs to the MADS-box domain family, whose genes play vital roles in multiple aspects of plant development (for instance development of flowers, fruits and roots as well as regulation of flowering time) [228, 229]. Such genes regulate, for example, stem growth and early flowering in soybean [230] or the vernalization response in wheat [231]. SQUA itself is involved in the determination of floral meristem and organ identity [232, 233]. The MYB domain group is one of the largest families of TFs in plants. Its members are involved in the regulation of development, metabolism, the circadian rhythm, and responses to biotic and abiotic stresses in plants [229,

Table 5.7.: Alleles of the 20 *Vicia faba* lines for the two significantly associated SNPs

Line	V+C	SNP_341016_236	SNP_341016_239
Line 1268-4-1	Low	C	G
Mélodie/2	Low	C	G
F7(Mélodie/2 x ILB938/2)-139-1-1	Low	C	G
F7(Mélodie/2 x ILB938/2)-201-3-1	Low	C	G
F7[VC.14.8099-843-2-1]	Low	C	G
F7[VC.14.8099-848-3-1]	Low	C	G
F7(Mélodie/2 x ILB938/2)-139-2-1	High	T	A
F7(Mélodie/2 x ILB938/2)-201-4-1	High	T	A
F7[VC.14.8099-843-3-3]	High	T	A
F7[VC.14.8099-848-4-1]	High	T	A
HediLin-1	High	T	A
PietraLin	High	T	A
(HediLin/1 x PietraLin)-2-4	High	T	A
(HediLin/1 x PietraLin)-4-4	High	T	A
S_281	High	T	A
S_301	High	T	A
S_034	High	T	A
S_290	High	T	A
Hiverna/2	High	C	G
Côte d'Or/1	High	T	A

Table 5.8.: The two SNPs with the strongest association to the V+C content and their consequences. The column *Allele* indicates the allele of the SNP for which the binding site exists. *TFBS* refers to the name of the binding sites, which were named after their PWMs.

SNP_ID	Allele	TFBS	MSS	Consequence
SNP_341016_236	Ref	P$MYB4_01	0.945	Loss of TFBS
SNP_341016_236	Ref	P$MYB61_01	0.880	Loss of TFBS
SNP_341016_236	Ref	P$SQUA_01	0.870	Gain of TFBS
SNP_341016_239	Ref	P$MYB61_01	0.880	Score change
SNP_341016_239	Alt	P$MYB61_01	0.881	Score change

234, 235]. In *Medicago truncatula*, multiple MYB TFs, including MYB4 and MYB61, are involved in flavonoid biosynthesis during the macrosclereid cell development [236]. MYB4 in particular regulates abiotic stress responses towards UV-B light and cadmium

toxicity in *Arabidopsis thaliana* [237, 238] and cold in *Oryza sativa* [239]. It has also been shown to influence the biosynthesis of flavonoids [240]. MYB61 participates in the response to cold stress in *Medicago truncatula* [241]. In *Arabidopsis thaliana*, this TF is expressed in sink tissues, such as xylem, roots and developing seeds, and controls resource allocation influencing growth and development of the plant [242]. It has also been shown to affect trichome initiation, root development and stomatal aperture and it is necessary for the biosynthesis of gibberellin [243, 244]. Furthermore, it is required for the seed coat mucilage deposition during the development of the seed coat epidermis [245, 246]. This is a promising result considering that the seed coat is the suggested site of biosynthesis of the V+C compounds [247].

5.2. Mutual information based detection of epistatic SNP pairs

In this section, I present the results regarding my novel method MIDESP for the identification of epistatic SNP pairs. First, in order to gain insights into the influence of the prerequisite parameter k used by the mutual information estimator for quantitative phenotypes, I present the results of a systematic analysis of several simulated datasets to find the most convenient value for it. Second, I present an example to showcase that mutual information can be more powerful than linear regression in a normal GWAS. After that, I highlight the importance of the APC approach in MIDESP to account for the possible background association effects in epistatic interactions. In the following sections, using MIDESP with a false discovery rate (FDR) of 0.05, I analyze two real datasets with a qualitative and a quantitative phenotype, respectively, to demonstrate its functionality. Finally, I show a comparison of the results between MIDESP and three other well-established methods for epistasis detection on the aforementioned real datasets.

5.2.1. Analysis of simulated datasets for parameter setting

Today, it is well established that mutual information is an appropriate metric to measure the association between SNPs and qualitative (case-control) phenotypes [76, 79, 81, 82, 248, 249, 250]. However, I apply here for the first time this metric to quantitative traits. Therefore, I analyzed several simulated datasets to identify a suitable value of k, which is necessary for the $\mathbb{MII}$ estimator (see Equations 4.3.2 and 4.3.4). For this purpose, I employed the LDAK[1] software [251] to simulate several hundred genotype$\times$phenotype datasets with three heritability values: 0.05, 0.075 and 0.1. Higher values of heritability showed similar results to 0.1 and are, therefore, not depicted. Consequently, I created 500 datasets consisting of 1000 SNPs, 2000 samples and a continuous phenotype controlled by a single SNP for each heritability value, respectively. Power was calculated as the proportion of datasets where the causal SNP obtained the highest $\mathbb{MII}$ value. To establish a proper value of k for the $\mathbb{MII}$ estimator, I systematically analyzed each dataset using k-values from 1 to 60. Despite Ross [87] and Kraskov et al. [156] both recommending a low value of $k = 3$, my analyses indicate that such small values can only be considered for heritability values > 0.1. (see Figure 5.1).

[1] LDAK is a software tool for the analysis of genotype$\times$phenotype datasets and contains functions for heritability analysis and association testing, among others.

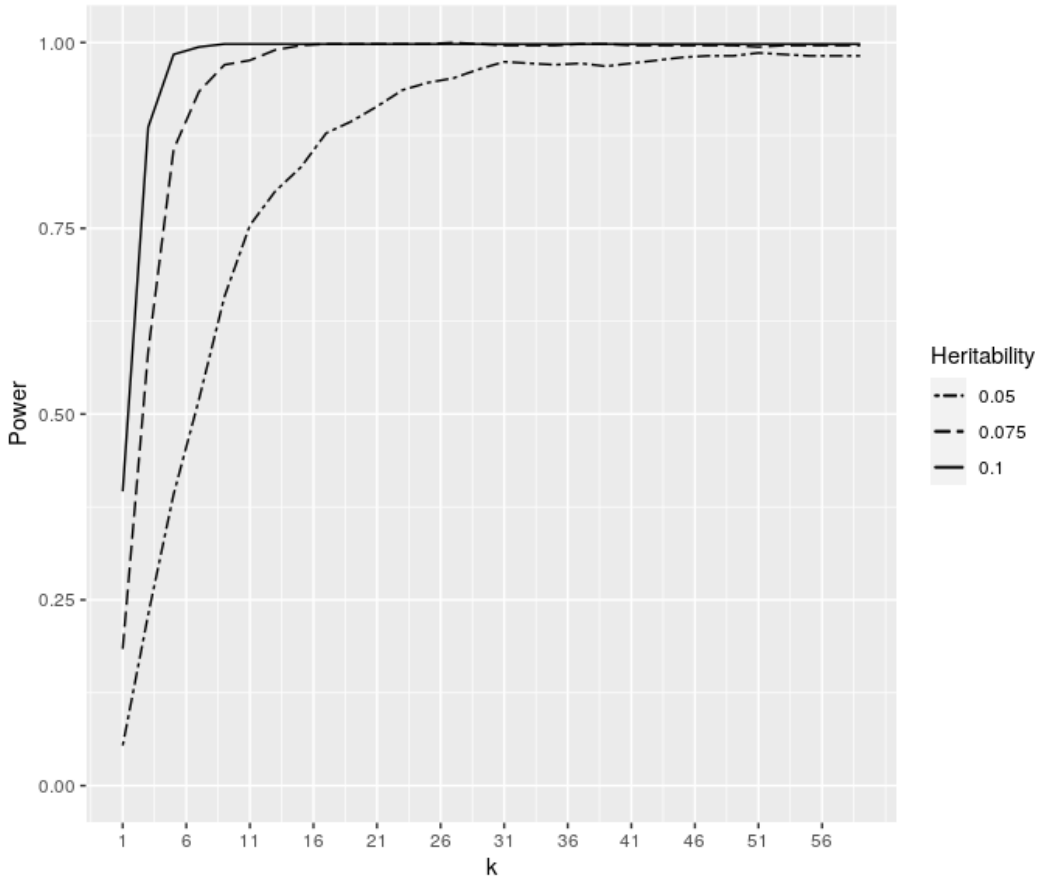


Figure 5.1.: Analysis of simulated datasets for varying parameter settings of k.

Further, Figure 5.1 suggests that simulated datasets with smaller heritability values require a much higher k-value to successfully detect the causal SNPs of interest. By systematically analyzing different k-values, I established that a value of $k = 30$ leads to the highest increase in power for the estimator based on the heritability values under study, and the power converges to nearly 1 if $k \geq 30$. I did not choose a higher value, since an increase in k results in a longer runtime for the estimator and may likewise cause problems if the sample size is not large enough.

5.2.2. Single SNP association using mutual information

Following the previous section, I present here by way of an example a comparison between the mutual information and the standard linear regression approach which is used for GWAS. I consider two SNPs X1 and X2 which together determine a quantitative phenotype

in an epistatic interaction. For this scenario, I simulated data for 1800 samples. The effect of the two SNPs on the phenotype in the simulated data is shown in Figure 5.2. This is a similar scenario to the one presented in [65], where the effect strength of one SNP (here X2) depends on the genotype of another SNP (here X1).

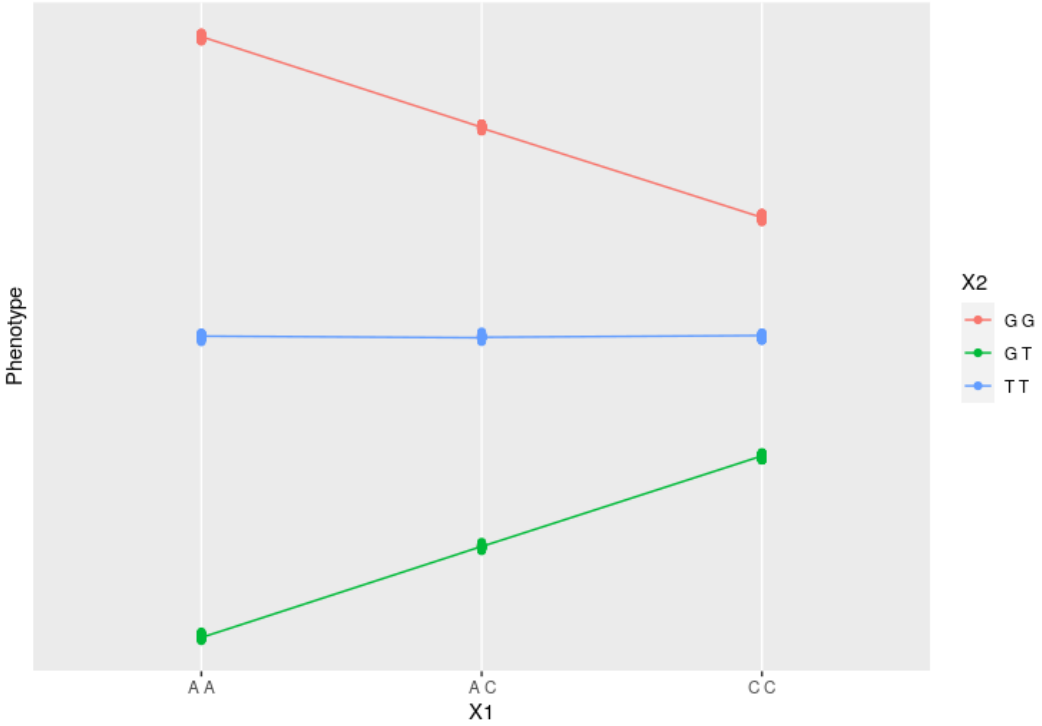


Figure 5.2.: Simulated data where a quantitative phenotype is controlled by two SNPs with three genotypes each. The genotypes of the SNPs are indicated by the x axis for X1 and by the color for X2.

As can be easily seen, the phenotypical difference between the genotypes of X2 is far greater for the genotype AA than for the genotype CC of X1. I then performed a GWAS on this dataset using PLINK and calculated the mutual information between X1 and the phenotype and X2 and the phenotype. The results are depicted in Table 5.9. PLINK uses a linear regression to test for association between the SNP and the phenotype. When considering the SNPs individually, the linear regression barely finds any difference in the phenotype values between the three genotypes of X1, which results in a p–value of nearly 1. In comparison, $\mathbb{MI}$ is not bound to a linear relationship and finds that the genotypes of X1 result in distinct phenotypes in two-thirds of the samples. This results in a very high value for the association of 0.6659. For X2, on the other hand, both approaches result in a near-perfect association of nearly 0 for the p–value and 1 for the $\mathbb{MI}$, respectively.

Table 5.9.: Association between SNPs X1 and X2 to the phenotype as determined by PLINK (p–value) and mutual information ($\mathbb{MI}$).

SNP	***p*–value**	$\mathbb{MI}$
X1	0.9987	0.6659
X2	1.875×10^{-100}	1.0

In this regard, $\mathbb{MI}$ is more suitable for detecting individual SNPs associated with the phenotype when the association does not follow a linear model. As shown here, such non-linear associations can be created by epistatic interactions.

5.2.3. Illustration of background associations and their correction using APC

In information theory, mutual information $\mathbb{MI}$ is typically measured between two variables, X^1 and Y. Additionally, based on the chain rule of information [155], it is well known that the introduction of a new variable, X^2, might affect the relationship between X^1 and Y, thus increasing the $\mathbb{MI}$ between X^1 and Y. However, if the introduction of X^2 does not result in any new information, the corresponding $\mathbb{MI}$ value will not be affected [155].
In case of SNP$\times$phenotype associations, this property of the joint mutual information needs to be considered, since only the introduction of an additional SNP^2, which increases the amount of information between SNP^1 and the phenotype Y, should be taken into account for the detection of epistatic interactions. I have exemplified in Section 3.3.2.1 that the joint mutual information can lead to a false interpretation of epistatic interactions. To deal with this problem, I apply the average product correction (APC) theorem [88], which ensures the elimination of negligible increments in the $\mathbb{MI}$ value of epistatic interactions measured using Equations 4.3.3 and 4.3.4.
Another important aspect of the usage of the $\mathbb{MI}$ metric in the context of epistatic interactions is its ability to detect the newly created relationship between a SNP pair and the phenotype, even though the single SNPs themselves might not show any association to the phenotype (see Case 4 in the example in Section 3.3.2.1).
To demonstrate the importance of the APC in the analysis of epistatic interactions, I further applied it for the correction of the $\mathbb{MI}$ values calculated using Equation 4.3.3 regarding the BT dataset. I considered the top million $\mathbb{NMI}$ values indicating the epistatic interaction between the SNP pairs and the phenotype. Afterwards, I determined for each SNP its frequency among the interactions. The frequency distribution of SNPs and their single association to the phenotype is shown in Figure 5.3A. As can be easily seen, the frequency of several SNPs is over-represented, which arises from their strong single association to the phenotype. However, the application of the APC dramatically reduces their frequencies in the epistatic interactions. This finding clearly suggests that, although these SNPs individu-

ally have a strong association with the phenotype, their epistatic interactions are negligible, as shown in Figure 5.3 with blue points.

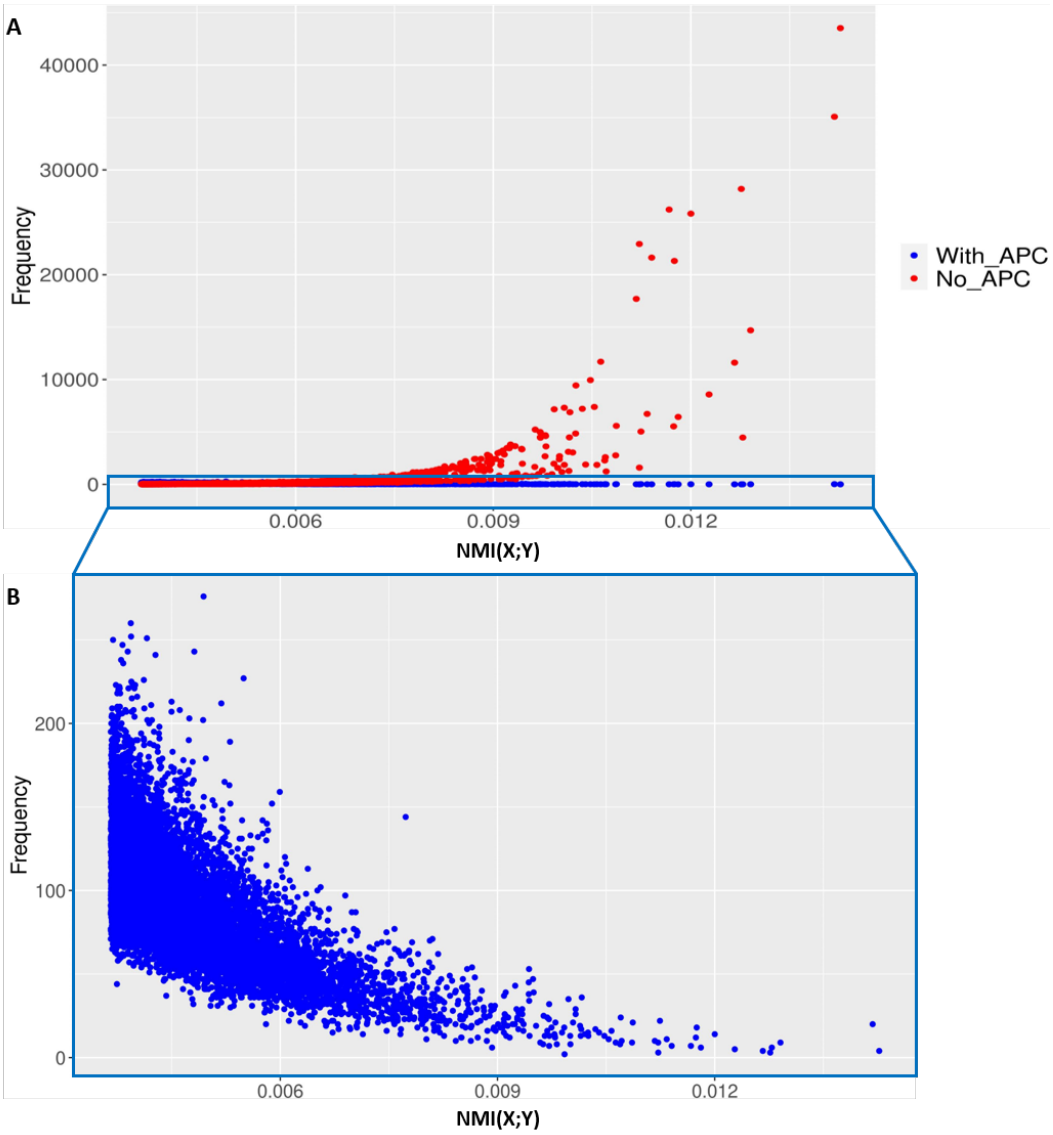


Figure 5.3.: (A) shows the distribution of the SNP frequency and their association to the phenotype. The blue and red points stand for the frequency of the SNPs based on with and without the application of the APC, respectively. (B) shows only the frequencies for the interactions with APC.

5.2.4. Results from the analysis of the bovine tuberculosis dataset

By applying MIDESP to the BT dataset, I first identified in total 10,744 single SNPs, with significant association with the phenotype. Taking all SNP pairs that contain at least one of those significant SNPs into account for the epistatic interaction analysis, I identified 3,799,984 SNP pairs, which corresponds to 0.1 % of all possible pairs under study. After that, I mapped these SNPs to their genes using the Ensembl database and created a gene-gene interaction network, as suggested in [84]. Finally, according to this network, I detected 511 genes as MIDESP-significant and investigated their roles in the bovine tuberculosis disease based on enriched Gene Ontology (GO) terms (see treemap depicted in Figure 5.4).

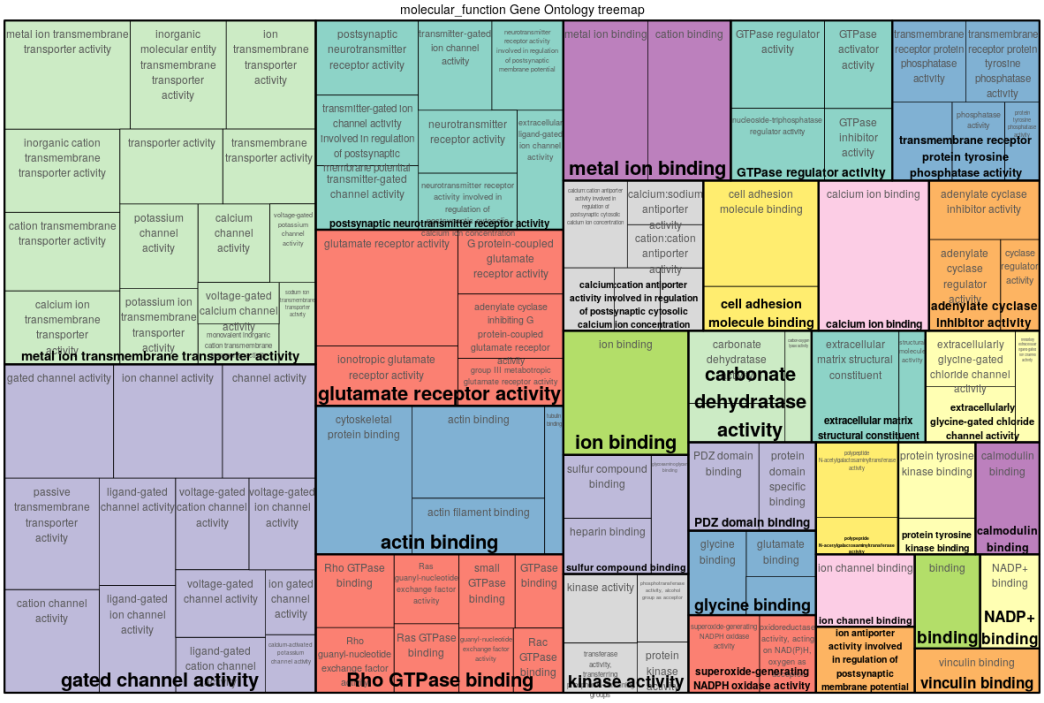


Figure 5.4.: Gene Ontology (GO) treemap for genes associated with immunity to bovine tuberculosis. The boxes are grouped together based on the upper-hierarchy GO term, which is written in bold letters.

The functional classification of these genes indicates that several of the GO categories represented in the treemap play essential roles in the immune responses towards bovine tuberculosis. Especially, metal ion transmembrane transporter activity and gated channel activity are the most significantly enriched terms, shown in the green and purple boxes in the treemap (Figure 5.4) obtained from the GO analysis, indicating the function of transmembrane proteins involved in the transportation of ions across membrane layers. In par-

ticular, ion channel blockers are known for their therapeutic implications in drug-resistant mycobacterial infection, especially voltage gated calcium channels, which are important for the regulation of immunity against pathogens [252, 253, 254, 255]. In this regard, increasing calcium influx by inhibiting the voltage gated channels in immune cells such as macrophages (immune cells engulfing the pathogens) is highly associated with protective immunity, particularly in increasing the expression of genes involved in pro-inflammatory responses [255]. Other significant GO terms including actin binding, Rho GTPase binding, glutamate receptor activity and postsynaptic neurotransmitter receptor activity were also enriched in the treemap and their roles associated with *Mycobacterium tuberculosis* are described below. Firstly, actin filament, which is an important constituent of the cytoskeleton [256], is mainly associated with pro-inflammatory responses. A primary aspect of mycobacterial infection is the manipulation of actin filaments [257], notably inside the macrophages of the host [258, 259, 260], thereby pointing out the importance of actin-binding protein regulation for enhancing the immune responses of the host. Several recent studies reported that neurotransmitters play essential roles in the activation or suppression of immune responses through the regulation of T-cell activity [261, 262]. It is well known that T-cells play an important part in the defense of the host against mycobacterial infections [263, 264, 265]. Specifically, the neurotransmitter taurine was identified in relation with the susceptibility of cattle towards bovine tuberculosis [266]. Glutamate is likewise a neurotransmitter known for its effect on the immune system for the regulation of T-cell activity [267, 268]. Finally, Ras homology GTPases (Rho GTPase) are proteins involved in the critical regulation of signaling pathways upon bacterial entry at the site of infection, and therefore are involved in innate immune responses, particularly in the multiplication of immune cells. It is essential to coordinate the immune responses at this point to prevent the neighboring tissue from taking damage from inflammation. Involved in the tight regulatory roles of multiple immune functions, these signaling proteins have been reported as targets of *Mycobacterium tuberculosis* during the host cell invasion, which might facilitate the pathogenesis of the bacteria [269, 270, 271].

5.2.5. Results from the analysis of the egg weight dataset

Similarly to the previous dataset, MIDESP was used to analyze the EW dataset, which contains a quantitative phenotype. As a first step, I detected 3,116 single SNPs that were significantly associated with the trait. Based on these SNPs, I measured the epistatic interactions between the SNP pairs and the phenotype and obtained 1,071,464 SNP pairs in total, which equates to 0.25 % of all possible pairs under study. After mapping these pairs to a gene–gene interaction network, I was able to identify 211 genes as MIDESP-significant. The analysis of their roles regarding egg weight was again carried out using their enriched GO terms (see treemap depicted in Figure 5.5).

Figure 5.5.: Gene Ontology (GO) treemap for genes associated with egg weight. The boxes are grouped together based on the upper-hierarchy GO term, which is written in bold letters.

For egg weight, one of the major GO categories that emerged as a result of the gene set analysis was the fatty acid ligase activity. Fatty acid ligases belong to the ligase family of enzymes that take part in the biosynthesis of lipids [272]. Lipids constitute a major portion of the nutrients found in the egg and are primarily contained in the egg yolk, which accounts for 31% of the total egg weight [273]. Multiple genes encoding fatty acid ligases have been reported to play important roles in the laying performance of birds [274, 275, 276]. In this regard, I was able to discover many genes with molecular functions associated with acyl-CoA ligases, a group of enzymes, which are known to play important roles in the lipid synthesis by converting the chemically inert fatty acids to active acyl-CoA [277]. This activation comprises an ATP-dependent reaction catalyzed by ligase enzymes in the presence of Mg^{2+} and CoA [278]. The usage of ATP and Mg^{2+} in this process can also explain the role of adenyl nucleotide binding and magnesium ion binding, two other categories identified in my analysis. Gated channel activity is another important GO term found in this analysis. These genes ensure the transportation of nutrients and minerals, which are required for the development of the egg. More importantly, for the synthesis of the eggshell, which contributes around 9% to the total egg weight [273], large amounts of calcium ions are supplied to the uterine fluid by transepithelial transport [273, 279]. This transepithelial transport oc-

curs with the help of ion channels, ion pumps and ion exchangers in the reproductive tract of the birds and the energy required for these processes is provided by the metabolisms of ATP molecules [279]. Both nucleotide binding and gated channel activity have been reported in association with egg weight and eggshell development in chicken [196, 197]. Furthermore, genes related to protein transmembrane transport activity were also identified, which can regulate the transportation of the large number of proteins found in an egg [273, 280]. The gene set analysis further reveals other activities pertaining to molecular bindings at different levels, which can play crucial roles for the development of the egg.
In Section 4.3.2, I argued that an adjusted $\mathbb{MII}$ estimator might be preferable if the values of the continuous phenotype are not unique. This is the case for the EW dataset, which has many repeated values in its phenotype. Therefore, I additionally applied MIDESP with the adjusted estimator on this dataset. In comparison with above results from the original method, the adjusted method reports a higher number of genes as MIDESP-significant. Compared to the original number of 211 genes, the adjusted version finds 328 genes. While a large part of the genes (112 genes) are shared between the two results, the remaining genes are specific to the respective method (see Figure 5.6). In terms of their molecular function, the newly found genes are more strongly associated with the gated channel activity than with the fatty acid ligase activity.

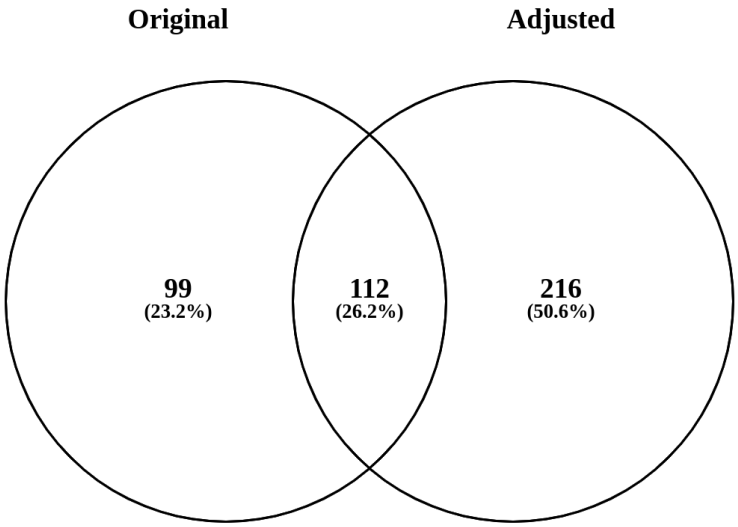


Figure 5.6.: Venn diagram which shows the overlap between the MIDESP-significant genes, which were found by the original and the adjusted method, respectively.

5.2.6. Comparisons with existing methods

To investigate the performance of my novel method, I was further interested in making pairwise comparisons between the results of my MIDESP, PLINK [135], GBOOST [281], epiGPU [282] and MatrixEpistasis [283]. Although all these methods take a genotype–phenotype dataset as input and report epistatic SNP pairs as a result, their applicability differs based on the phenotypes. While MIDESP and PLINK can be applied to qualitative as well as quantitative phenotypes, the other methods are restricted to one type. GBOOST only deals with qualitative phenotypes, while epiGPU and MatrixEpistasis only analyze quantitative phenotypes. I chose these tools since they have previously been used for pairwise epistasis detection on real datasets, as well as for comparison studies [77, 121, 284, 285, 286, 287, 288, 289], and ran them with their default parameters. It is important to note that for this comparison study, I applied MIDESP with and without APC correction. While MIDESP without APC is in line with the conventional mutual information (MI)-based methods for epistasis detection [76, 81, 83, 84], the incorporation of the APC approach is completely novel and necessary to separate the correct epistatic signals from the background. The results of this comparison are twofold. First, I compared the results of my method using the BT dataset with those of PLINK, GBOOST and the conventional MI-based metric, since the existing MI-based approaches are only applicable to qualitative phenotypes [76, 81, 83, 84]. Second, I compared the predictions of MIDESP on the quantitative EW dataset with those of PLINK, epiGPU and MatrixEpistasis. However, my attempt to apply MatrixEpistasis to this dataset was unsuccessful due to its very high memory consumption (700 GB of memory was not enough). The application of these methods results in the detection of strongly varying numbers of SNP pairs as epistatic interactions, which are given in Table 5.10.

Table 5.10.: Number of SNP pairs that were found to be an epistatic interaction by the different methods. BT and EW stand for bovine tuberculosis and egg weight, respectively.

Dataset	#MIDESP	#MIDESP_NoAPC	#PLINK	#GBOOST	#epiGPU
BT	3,799,984	3,799,984	4,982,695	346,632	-
EW	1,071,463	1,071,463	1,817,817	-	572,914

To make the predictions of the methods comparable, for both types of the traits, I considered 346,632 and 572,914 epistatic SNP pairs, which corresponds to the minimum numbers of SNP pairs found by GBOOST and epiGPU for the BT and EW datasets, respectively (see Table 5.10). For this purpose, I sorted the identified SNP pairs according to the respective score assigned by the method (e.g. p–value for PLINK) and took the best pairs from each result. Based on these top SNP pairs, I further performed an overlap comparison between the methods and visualized the results using UpSet plots in Figures 5.7 and 5.8, respectively.

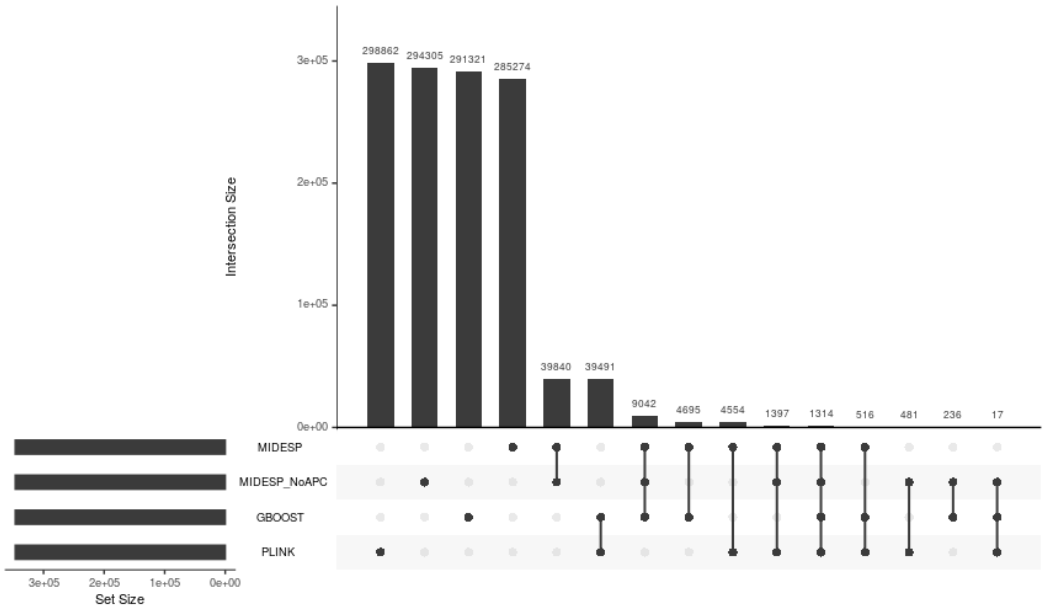


Figure 5.7.: The top 346,632 epistatic SNP pairs detected for the BT dataset and their overlap between four methods represented in matrix layouts using the UpSet technique [290]. Black circles in the matrix layout indicate which methods are part of the intersection.

Although all of these methods perform a search for epistatic SNP pairs, Figures 5.7 and 5.8 clearly show that they provide quite distinct results with only little overlap. This finding is in line with the comparison study performed in [284], which also reported divergent results between different methods for epistasis detection. The reason for that may be explained by the differences in the underlying algorithms, even though the three other methods are ultimately based on logistic and linear regression, respectively. While PLINK performs a regression with an interaction term and tests whether the coefficient for the interaction is significant, GBOOST considers the difference in the likelihood of a linear model with interaction compared to that of a model without as a sign for epistasis using approximations to speed up the process and filter out SNP pairs. On the other hand, epiGPU treats the genotype combinations as different classes and calculates differences between the class means and the population mean. Consequently, the results of this overlap analysis clearly demonstrate that these methods carry quite distinct information about epistatic interactions due to the different measures they use. The finding of this comparison analysis is also in agreement with the previous study [284] and indicates that each method takes a different type of epistatic interactions into account and, therefore, they can work complementarily with each other.

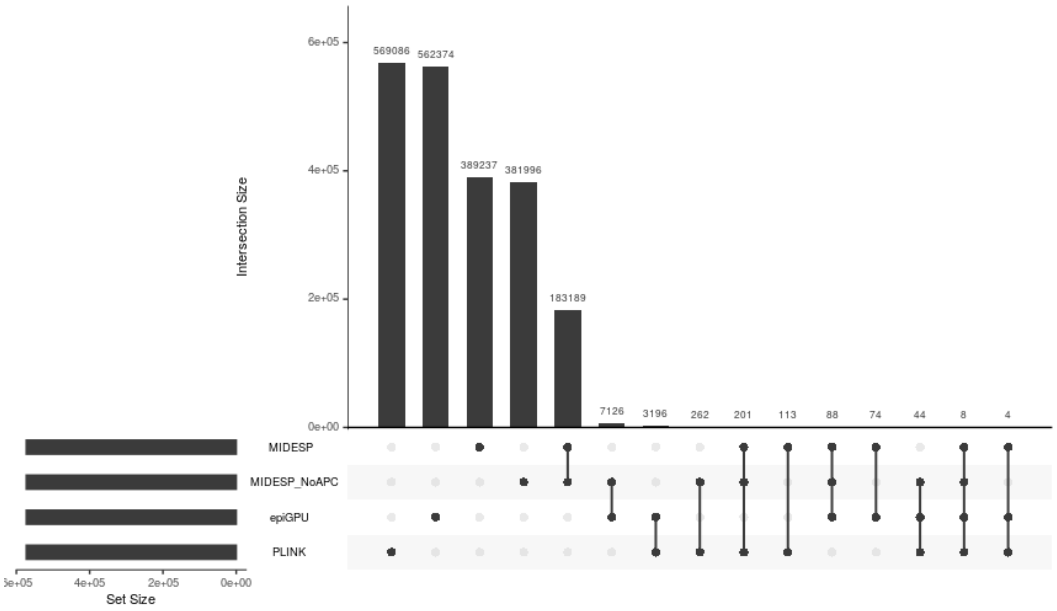


Figure 5.8.: The top 572,914 epistatic SNP pairs detected for the EW dataset and their overlap between four methods represented in matrix layouts using the UpSet technique [290]. Black circles in the matrix layout indicate which methods are part of the intersection.

5.3. Detection of epistatic SNP pairs associated with V+C in *Vicia faba*

In this section, I present the results of the MIDESP analysis for the *Vicia faba* dataset. By applying MIDESP to this dataset, I first identified 34,690 single SNPs in total with significant association with the V+C content. Considering the results of the promoter detection, 6838 (19.71 %) of these SNPs are located in putative promoter regions, which corresponds approximately to the total ratio of such SNPs in the dataset (132,399 out of 685,215 SNPs, resp. 19.32 %). However, only two of the previously identified fourteen SNPs in the target region (see Table 5.6) were found as significant by MIDESP. These are the top two markers SNP_341016_236 and SNP_341016_239, which already showed a significant association according to a GWAS. Overall, I obtained 23,168,427 SNP pairs in total, which equates to 0.1 % of all possible pairs under study. I again considered the percentage of regulatory SNPs (rSNPs) among these SNP pairs and present the results in Table 5.11. The distribution of rSNPs among the epistatic pairs follows approximately the expected distribution based on the overall percentage of rSNPs in the dataset. However, a minor increase towards pairs with 1 or 2 rSNPs can be observed.

Table 5.11.: Number of epistatic SNP pairs that were found for the *Vicia faba* dataset partitioned by the number of rSNPs participating in a SNP pair. The second row contains the expected percentages for the partitions based on the proportion of rSNPs in the dataset.

0 rSNPs	**1 rSNP**	**2 rSNPs**
14,081,339 (60.78 %)	7,992,847 (34.5 %)	1,094,241 (4.72 %)
65.09 %	31.18 %	3.73 %

Due to the lack of an annotated reference genome for *Vicia faba*, I could not perform a gene set analysis like I did for the BT and EW datasets. Therefore, I focused on the impact of the APC correction in MIDESP.

As in the example shown in Section 5.2.3, I compared the frequency of the significant SNPs among the detected SNP pairs with and without application of the APC. These frequencies are visualized in Figure 5.9 in dependency to the $\mathbb{NMI}$-value between the single SNP and the phenotype. The results show a pattern that is similar to the one in Figure 5.3. Without application of the APC, the frequency of SNPs with strong association to the phenotype is over-represented among the SNP pairs. Consequently, the APC reduces this effect of strong single associations on the final results. This behaviour is also visualized in Figure 5.10, which depicts for several of the found epistatic SNP pairs the strength of association between the single SNPs and the phenotype as well as between the SNP pair and the phenotype. The height of the bar represents the $\mathbb{NMI}$-value between the corresponding SNP respectively SNP pair and the phenotype. Additionally, the orange colored part of the bars for the SNP pairs shows how much the association of the first SNP is increased by the addition of the second SNP.
Considering the two depicted SNP pairs that were found using the APC, Figure 5.10 (A) shows a classic example of an epistatic interaction. Both single SNPs show a moderate association with the phenotype. However, when both SNPs are considered together as a pair, the association is significantly higher, which represents their epistatic interaction. Interestingly, a similar gain of association can observed in Figure 5.10 (B) by adding a second SNP, which by itself is only minimally associated with the phenotype. In contrast, without applying the APC in MIDESP, many SNP pairs are found that exhibit no true epistasis and where the second SNP provides no novel information about the phenotype. For example, in Figure 5.10 (C), the association of the SNP pair is only due to the strong association of the first SNP while the second SNP contributes nothing. Similarly, there are also SNP pairs where the second SNP carries the same information about the phenotype as the first SNP and therefore is redundant (see Figure 5.10 (D)). Overall, these examples visualize the need for the APC when it comes to the detection of epistasis using mutual information.

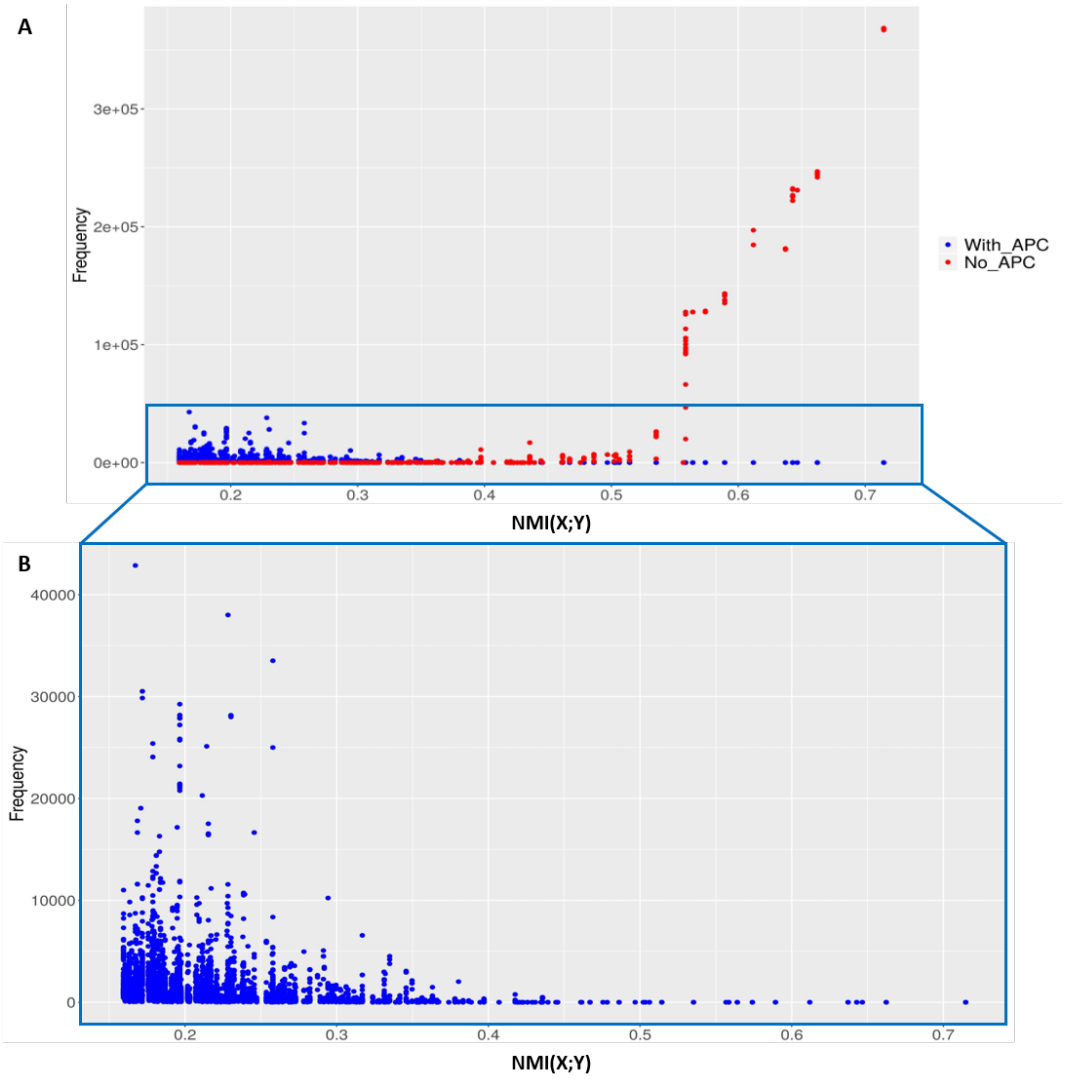


Figure 5.9.: (A) shows the distribution of the SNP frequency and their association to the phenotype. The blue and red points stand for the frequency of the SNPs with and without APC, respectively. (B) shows only the frequencies for the interactions with APC.

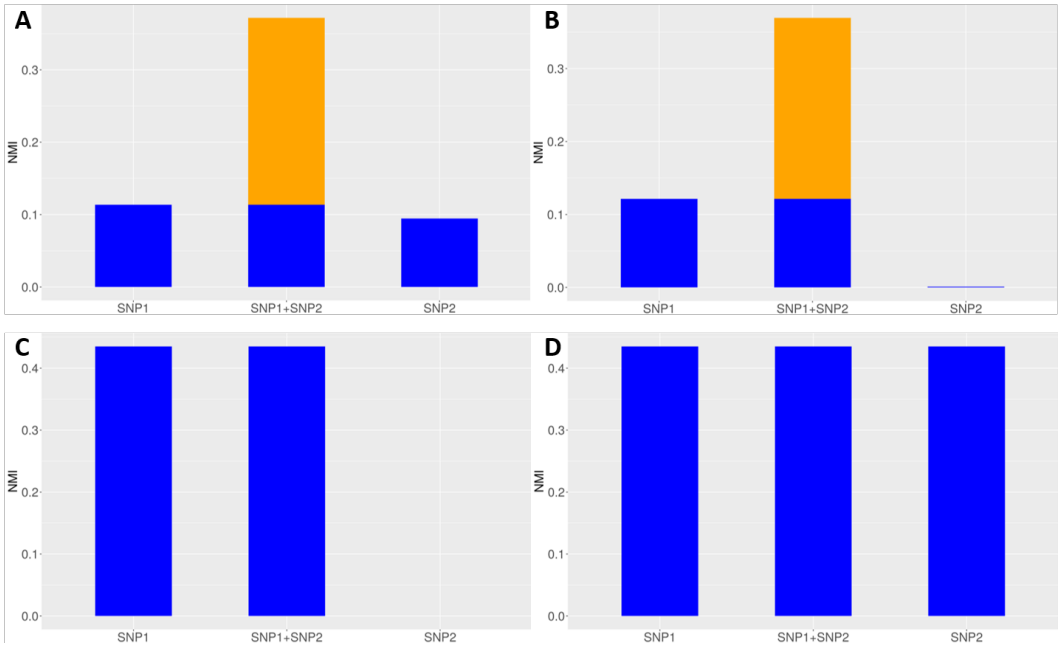


Figure 5.10.: Examples of four epistatic SNP pairs from the *Vicia faba* dataset. (A) and (B) were found as significant using the APC correction while (C) and (D) were only found without APC. The height of the bar represents the $\mathbb{NMI}$-value between the corresponding SNP respectively SNP pair and the phenotype. The orange colored part of the bars representing the association of the SNP pairs is defined as the difference between the association of the SNP pair SNP1+SNP2 and the association of the single SNP SNP1 with the phenotype. Therefore, it can be interpreted as the gain of association due to the epistatic interaction.

6. Discussion

In this chapter, I discuss the two analysis frameworks, which I presented in this thesis. The discussion presented here is partly based on my two publications [52, 165] (see Appendix A.1 and Appendix A.3).

6.1. Identification of regulatory SNPs based on genotyping by sequencing data in *Vicia faba* using deep learning

Today, it is well known that most SNPs, which are linked to a certain trait, are not part of the coding sequence of a gene but instead are located within non-coding regions [113, 291, 292]. Such SNPs can still affect the gene expression by altering the binding sites of transcription factors (TFs), which in extreme cases could cause a completely different TF to bind [113]. These so called regulatory SNPs (rSNPs) can be computationally identified by comparing the predicted transcription factor binding sites (TFBSs) between the alleles [35, 111], and lead to an increased understanding of the regulatory mechanisms controlling the trait. However, the identification process presupposes that it is known whether a SNP lies within a promoter region. Since so far no annotated reference genome exists, this information is missing for *Vicia faba*. To address this issue, I utilize the ability of genotyping-by-sequencing to access and sequence genomic regions including regulatory regions [3]. The sequences obtained in this way are then classified into promoter and non-promoter sequences using deep learning.

In the last few years, methods based on convolutional neural networks (CNNs) have become the preferred tool for the detection of promoter sequences [44, 45, 46, 47, 48, 49]. Compared to earlier approaches, they have the advantage of not requiring any prior knowledge regarding the selection of specific sequence features. Instead, these networks are able to automatically learn important features that distinguish between promoter and non-promoter sequences by themselves. However, CNNs come with the disadvantage of being more difficult to interpret [45]. While mutation or saliency maps are able to show at which positions a specific base might be required, these single positions do not necessarily reflect the complete structure of the promoter, which is taken into account by the network.

A significant challenge during the classification of the *Vicia faba* sequences was the lack of an annotated reference genome. To address this shortcoming of training data, I exploited the conservation of promoter structures across different closely related species. This conservation has been reported by Kumari et al. [50], who performed a large scale genome-wide

key study for the prediction and analysis of core promoter elements. They observed that promoter signatures are strongly conserved within larger groups of plants like monocots or dicots. Based on these findings, Shamuradov et al. [51] developed a model, namely TSS-Plant, for the prediction of plant promoters across species boundaries.
In line with these two studies, I trained multiple CNN models using sequences from seven different plant species with annotated reference genomes available, which belong to the same *Leguminosae* family as *Vicia faba*. Additionally, I used the model organism *Arabidopsis thaliana*, which is more distantly related but still belongs to the group of dicots like the other plants. An initial comparison of the classification of promoters within and between species using CNNs, each trained on sequences from a single species, showed that the prediction of promoters across the different species is, to a certain degree, possible (see Table 5.1). Remarkably, the species *Lupinus angustifolius* and *Vigna angularis* appear to have largely promoter signatures that are shared by most of the other plants under study and can therefore be predicted with a high accuracy by most tested models.
Considering that Triska et al. [46] reported a markedly improved performance by incorporating additional commonly used sequence features in their CNN model and Qian et al. [47] achieved a better recognition of promoters by separating the sequence in elements and non-elements, I also evaluated the effect of additional features on the model prediction (see Section 5.1.2). However, as exemplarily shown in Table 5.5 for the network based on *Medicago truncatula*, no such gain in performance could be observed. Similarly, the performance of the final model could not be further improved by using additional features. These results are in agreement with the findings presented in [44, 45, 48, 49] and indicate that the CNN architecture is able to learn specific patterns inherent in the sequences that differentiate between promoter and non-promoters automatically. Hence, these patterns carry information which is obviously redundant to these widely used features. Consequently, it turns out that the consideration of additional features does not lead to an improvement in the performance of the CNN model and may, on the contrary, increase the noise during training. Nevertheless, this may not be the case for every dataset, which could explain the observed improvements in [46, 47].
To better reflect the variety of promoter signatures occurring in the *Leguminosae* family, I tested different CNN models trained on combinations of datasets from multiple species. To this end, I chose to train the final model based on sequences from the species *Lupinus angustifolius* and *Medicago truncatula*. Furthermore, it has been shown that the choice of the non-promoter sequences used for training is important to minimize the number of false positive predictions in unseen sequences [48]. For this purpose, Umarov et al. [45] applied an adaptive construction method for the negative set, while Oubounyt et al. [48] used an approach based on the shuffling of the positive set. Following this, I supplemented the training data for the final model with additional non-promoter sequences taken from the reference transcriptome of *Vicia faba* [130] as well as randomly selected sequences from the *Medicago truncatula* genome excluding the promoter regions to enhance the ability of the CNN to distinguish between promoter and non-promoter.

For the analysis of the SNPs found in the predicted *Vicia faba* promoters, I focused on those SNPs, which could be successfully mapped to the target region of *Medicago truncatula* that is known to show conserved synteny with the V+C associated region of *Vicia faba*. A preliminary association analysis of these 14 SNPs revealed two markers with a significant and near-perfect association with the trait. The two markers are located only 3 bp apart and thus belong to the promoter of the same gene. Despite both of them having the same *p*-value for the association, comparing their effects on the binding of TFs revealed different consequences. Whereas the allele change of the first marker entails severe consequences in the form of the loss of two binding sites and the gain of one binding site for a different TF, the second marker causes only a minor change in binding affinity for a TFBS (see Table 5.8). In addition, functional analysis of the gene of interest and the affected transcription factors shows that, although the gene itself does not appear to be associated with the synthesis of V+C, the TF MYB61 has a potential link by being involved in the development of the seed coat epidermis [245, 246], which is the presumed site of V+C biosynthesis [247]. This exemplifies how the study of regulatory SNPs and their functional consequences on TF binding can lead to new insights into the molecular mechanisms underlying the trait under investigation.

6.2. Mutual information based detection of epistatic SNP pairs

It has previously been shown that information theoretic methods based on mutual information ($\mathbb{MII}$) are powerful approaches for the detection of epistatic interactions [76, 77, 78, 79, 80, 81]. Not only there, but also in many other fields, mutual information has been used as an effective measure for the association between variables including linear as well as non-linear relationships [88, 214, 220, 293, 294, 295, 296, 297]. However, the general applicability of a method, particularly in the field of animal and plant breeding, requires it to be suitable for qualitative as well as quantitative phenotypes. For this reason, an extension of the previous $\mathbb{MII}$ methods, which are only suitable for qualitative traits, is required, and thus I adapted the estimator developed by Ross [87] for the case of $\mathbb{MII}$ between discrete and continuous variables. As shown in Section 5.2.1, the estimator can be successfully used to detect associations between SNPs and quantitative phenotypes. Surprisingly, I found that a higher k value improves the power of the measure when it comes to the detection of associations involving traits of a low heritability (see Figure 5.1), although previous studies recommended a small value of k for this purpose [87, 156]. Furthermore, I have shown by means of an example that $\mathbb{MII}$ may be more suitable than the standard linear regression method for the detection of non-linear association between a single SNP and the phenotype (see Section 5.2.2). As shown in Figure 5.2, such non-linearity could easily be caused by an epistatic interaction.

The progress over the last decade in the field of genome sequencing and genotyping arrays has increased the available genotype data tremendously. However, with the ever-increasing amount of data, comes the challenge of providing tools that can handle such datasets in a feasible computation time. To overcome this challenge, redundant SNPs can be removed by LD pruning with a high threshold [212] (see Section 4.3.1), but there are still too many SNPs in a dataset to analyze all possible pairs. A commonly used approach to reduce the computational effort is to preselect sets of SNPs that are deemed as important and analyze only those, as is done by BOOST and other methods [166, 298, 299]. Such an approach can potentially eliminate some SNPs which, nevertheless, influence the phenotype in interaction with another SNP. To overcome this problem, in my proposed method, I consider all SNP pairs where at least one SNP shows a strong association signal to the phenotype, which ensures a tractable computational time for MIDESP. A similar filter was used by Slim et al. [300] in their epiGWAS program. For this step, I followed the approach outlined by Gültas et al. [214, 215] to separate the SNPs with strong association signals from the remaining SNPs (see Section 4.3.3).

However, the sole consideration of SNPs with strong association signals could lead to a wrong interpretation in epistasis analysis since the joint mutual information values are influenced by the association of the single SNPs with the phenotype, as I demonstrated by means of an example in Section 3.3.2.1. This can result in the detection of false positive interactions that are only found due to the effect of one SNP. To minimize this influence, the application of the average product correction (APC) is essential, which was developed

by Dunn et al. [88]. Moreover, Meckbach et al. [220] showed that the APC is universally applicable to $\mathbb{MI}$-based methods to estimate the expected (background) association level of a variable. Although the concept of the APC theorem seems to be suitable for the purpose of MIDESP, its application would require a huge additional computational overhead. Therefore, I followed a strategy based on the three different distributions of the SNPs (see Section 4.3.3) for the efficient estimation of the expected level of background association of SNPs. In particular, in Equation 4.3.12 I randomly choose the SNP X^l from the set of SNPs that follow the G_2 distribution. This process ensures that the expected background level of SNP X^i is clearly higher than it would be if estimated based on the whole set of SNPs. Consequently, the removal of the estimated background associations (APC values) from the obtained $\mathbb{NMI}$ values results in the separation of correct epistatic signals caused by SNP pair and phenotype interactions from background signals. Being of particular interest, in my analysis, I illustrated the effectiveness of the APC based on the BT dataset in Figure 5.3. This analysis reveals that the over-representation of SNPs with large single effects among the pairs with the highest $\mathbb{NMI}$ values can be considerably reduced based on the application of the APC, which in turn results in the detection of further associated genes.
As an alternative to the joint mutual information, several methods utilize the information gain ($\mathbb{IG}$) to detect epistasis [77, 78, 79, 80]. This approach has the advantage of automatically removing the main effects of the single SNPs and thereby largely reflecting only the synergistic effect of the SNP pair (see also Section 3.3.2.1). However, $\mathbb{IG}$ is more difficult to interpret due to the possibility of negative values, which also prevents the application of the APC theorem. Furthermore, the previously mentioned estimated background associations include the main effect of the SNPs in the APC correction, and thereby negate the advantage of $\mathbb{IG}$. Therefore, I decided to use the joint mutual information instead.
One limitation of $\mathbb{MI}$-based methods, which is shared by MIDESP, is that they do not return a p-value for the corresponding $\mathbb{MI}$-value. Instead, these methods and MIDESP simply report a certain percentage or absolute number of the top results with the highest $\mathbb{MI}$-values [76, 80, 301]. As an alternative approach, several methods suggest to apply permutation testing in order to obtain p-values [76, 78, 79, 213]. However, in order to apply permutation testing to MIDESP, it would become necessary to re-estimate the expected background level of each SNP for each round of permutation. The resulting computation time would be infeasible for larger datasets. For that reason, I did not include such an approach in MIDESP.
The results I presented in this thesis for the two different genotype-phenotype datasets show that the functional analysis of the detected genes provides essential information to decipher the genetic background of the traits under consideration. Surprisingly, I found a higher number of associated genes for the bovine tuberculosis dataset with a qualitative trait than for the egg weight dataset with a quantitative trait. This can be explained due to the large difference in the initial number of SNPs in both datasets. In comparison to the large numbers of associated genes detected by MIDESP, both original studies [195, 198], in which the datasets were published, were only able to find two significantly associated genes for the respective dataset using standard GWAS approaches.

To further investigate the impact of the APC theorem in the epistasis analysis and to gain more insight into its influence on the detection of genes, I analyzed both datasets with and without the application of the APC (see Figure 5.3). It can be assumed that without the APC, the results of MIDESP are in line with previous methods that utilized $\mathbb{MI}$ for the detection of epistatic interactions for qualitative phenotypes [76, 81, 83, 84]. The analysis reveals that the application of the ACP leads to a considerable increase in the number of associated genes for both datasets. For example, only 135 and 177 significant genes were found for the BT and EW datasets without using the APC, respectively. However, the correction of the background association using the APC results in the detection of 511 and 221 associated genes, respectively. The comparison of these genes showed that while 59 genes overlap for the BT dataset, 51 overlapping genes are found to be significant for the EW dataset.
Additionally, I compared the results of MIDESP with and without APC correction with three other existing methods for epistasis detection on the BT and EW datasets. It turned out that for the most part all methods identified different sets of SNP pairs as epistatic interactions with only little overlap (see Figures 5.7 and 5.8). Such divergent results have previously been noted [284] and can result from the different approaches that the methods use. Therefore, complementary approaches, which combine the results of several methods, might be suitable to account for different types of epistatic interactions.
The functional analysis of these genes based on their GO categories reveals that many of the identified genes are involved in the regulation of the immune responses regarding bovine tuberculosis, with several of the functions having a reported association with mycobacterial infections. The genes that were detected for the egg weight dataset, on the other hand, are mainly related to the production of important components of the egg and the transportation of these components to the uterine fluid. Overall, my results indicated that MIDESP is an effective method for the detection of epistatic interactions that, for the first time, enables the analysis of quantitative phenotypes using $\mathbb{MI}$ and further extends the existing information theoretic methods by correcting the influence of background associations of the SNPs through the application of the APC theorem.

7. Conclusion

In this last chapter, I first summarize the methods I presented in this thesis as well as their results. At the end, I give an outlook in which I provide some ideas for possible extensions of the methods and show some potential fields of applications for future research projects.

7.1. Summary

In this thesis, I presented two frameworks that I developed with the goal of elucidating different aspects of the relationship between the genome and a trait of interest. With the first framework, I aimed at the identification of rSNPs and their putative consequences on the binding of transcription factors (TF) (see Section 4.2). By using known promoter and non-promoter sequences from species closely related to *Vicia faba*, which have annotated reference genomes available, I was able to train a convolutional neural network for the purpose of detecting promoter sequences in a partial genome of *Vicia faba* that I assembled from genotyping-by-sequencing data obtained from 20 plants with known V+C content. In this case, the knowledge of conserved promoter signatures across related species thereby invalidates the need for the expensive and time-consuming process of assembling and annotating a reference genome for the species under study. Based on the location of the putative promoters, I was then able to analyze the SNPs in these promoters regarding their effects on the binding affinity of TFs. My analysis revealed two rSNPs that are highly associated with the V+C content of *Vicia faba*, which could be useful candidates for marker-assisted selection. Of particular note is that one rSNP disrupts a binding site of the TF MYB61, which has a potential link to the presumed site of the V+C biosynthesis (see Section 5.1.6). Finally, this application demonstrates the possible utility of using cross-species promoter prediction for species without annotated reference genomes.

In my second framework, I developed a method named MIDESP for the detection of epistatic interactions between SNP pairs based on mutual information (see Section 4.3). This method extends the existing mutual information-based approaches for detecting epistasis by additionally enabling the identification of epistatic interactions between SNP pairs and quantitative phenotypes. For this purpose, I adopt a kth-nearest neighbor-based approach to accurately estimate the level of epistasis. Furthermore, in MIDESP, I consider the existence of background associations between the SNPs and phenotype and incorporate the average product correction theorem to reduce their possible effect on the results. To demonstrate the performance of my method, I applied it on simulated as well as real datasets,

which were related to bovine tuberculosis and the weight of chicken eggs, respectively. A comparison of the results with and without application of the average product correction has shown that it has a significant impact on the identified interactions and is necessary to prevent the detection of non-epistatic pairs that arise due to the strong effect of a single SNP. For the biological evaluation of the results, I performed a gene ontology analysis of the genes involved in epistatic interactions. These findings revealed that the identified genes are strongly related to the respective trait under study for both datasets (see Section 5.2).

Overall, both methods are able to provide new insights into the mechanisms underlying the expression of the phenotype by identifying SNPs that participate in specific aspects of the association between genotype and phenotype.

7.2. Outlook

With regard to the identification of rSNPs in *Vicia faba*, the next step would be to verify the found associated SNPs on a larger set of plants with respect to their suitability as markers for the V+C content. Verified markers could then be used in marker-assisted breeding programs to create new *Vicia faba* lines with a low V+C content and, thereby, improve the use of this plant for human and animal consumption. Although the conservation of promoter signatures across related species, on which this framework is based, has been explored, a more detailed and comprehensive examination would be interesting, especially given the different prediction performances of the respective species. Furthermore, while the prediction of promoters with convolutional neural networks achieves high performances, innovations in the field of deep learning could provide further improvements. One potential new approach, for example, are variational autoencoders [302]. These are a special type of neural network that could be trained on a dataset, which contains only promoter sequences and no non-promoter sequences. The idea is that the network would learn the structures of the promoter sequences and if applied on a non-promoter sequence would recognize it as an anomaly. This would eliminate the need for a negative dataset for training, whereas the current approaches require a careful selection of non-promoter sequences for training to achieve a high performance.

Regarding the detection of epistatic SNPs, there are several possible extensions for MIDESP that might be worthwhile. So far, MIDESP is limited to the detection of epistatic SNP pairs. However, there also exists so-called higher order epistasis, which is based on groups of three or more SNPs that jointly interact with a phenotype. While the formula for the mutual information can be easily extended in this way, such a change would also require a non-trivial change in the average product correction, which is designed only for pairs of variables. Furthermore, there is the ensuing increase in the runtime to consider, which might require new filtering steps to be kept in check. An additional consideration that could be made is the explicit inclusion of confounding factors such as age, sex or population structure

for example. Linear mixed models include these factors as covariates in the field of genome-wide association studies as well as epistasis detection. Information theory offers as a related notion the formula of the conditional mutual information, which allows us to measure the association between two variables given that the information from other variables is known. Nevertheless, for a quantitative phenotype this necessitates an extensive redesign of the mutual information estimator. Finally, it would be interesting to see how the SNP pairs identified by MIDESP could affect the performance of genomic prediction resp. selection methods. Initial experiments done by others have shown promising results, however, a more in-depth analysis is still required.

Bibliography

[1] Metzker ML: **Sequencing technologies—the next generation**. *Nature reviews genetics* 2010, **11**:31–46.

[2] Deschamps S, Llaca V, May GD: **Genotyping-by-sequencing in plants**. *Biology* 2012, **1**(3):460–483.

[3] Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species**. *PloS one* 2011, **6**(5):e19379.

[4] Muktar MS, Teshome A, Hanson J, Negawo AT, Habte E, Entfellner JBD, Lee KW, Jones CS: **Genotyping by sequencing provides new insights into the diversity of Napier grass (Cenchrus purpureus) and reveals variation in genome-wide LD patterns between collections**. *Scientific reports* 2019, **9**:1–15.

[5] He J, Zhao X, Laroche A, Lu ZX, Liu H, Li Z: **Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding**. *Frontiers in plant science* 2014, **5**:484.

[6] Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N: **Marker density and read depth for genotyping populations using genotyping-by-sequencing**. *Genetics* 2013, **193**(4):1073–1081.

[7] Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, et al.: **Population genomic and genome-wide association studies of agroclimatic traits in sorghum**. *Proceedings of the National Academy of Sciences* 2013, **110**(2):453–458.

[8] (ICGMC) ICGMC: **High-resolution linkage map and chromosome-scale genome assembly for cassava (Manihot esculenta Crantz) from 10 populations**. *G3: Genes, Genomes, Genetics* 2015, **5**:133–144.

[9] Soto JC, Ortiz JF, Perlaza-Jiménez L, Vásquez AX, Lopez-Lavalle LAB, Mathew B, Léon J, Bernal AJ, Ballvora A, López CE: **A genetic map of cassava (Manihot esculenta Crantz) with integrated physical mapping of immunity-related genes**. *BMC genomics* 2015, **16**:1–16.

[10] Poland JA, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, et al.: **Genomic selection in wheat breeding using genotyping-by-sequencing**. *Plant Genome* 2012, **5**(3):103–113.

[11] Liu W, Xie J, Zhou H, Kong H, Hao G, Fritsch PW, Gong W: **Population dynamics linked to glacial cycles in Cercis chuniana FP Metcalf (Fabaceae) endemic to the montane regions of subtropical China**. *Evolutionary applications* 2021.

[12] Chen H, Xie W, He H, Yu H, Chen W, Li J, Yu R, Yao Y, Zhang W, He Y, et al.: **A high-density SNP genotyping array for rice biology and molecular breeding**. *Molecular plant* 2014, **7**(3):541–553.

[13] Winfield MO, Allen AM, Burridge AJ, Barker GL, Benbow HR, Wilkinson PA, Coghill J, Waterfall C, Davassi A, Scopes G, et al.: **High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool**. *Plant biotechnology journal* 2016, **14**(5):1195–1206.

[14] Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, Talbot R, Bishop SC, Archibald AL, Bron JE, et al.: **Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar)**. *BMC genomics* 2014, **15**:1–13.

[15] Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P, et al.: **Development of a high density 600K SNP genotyping array for chicken**. *BMC genomics* 2013, **14**:1–13.

[16] Cooper JW, Wilson MH, Derks MF, Smit S, Kunert KJ, Cullis C, Foyer CH: **Enhancing faba bean (Vicia faba L.) genome resources**. *Journal of Experimental Botany* 2017, **68**(8):1941–1953.

[17] Khazaei H, O'Sullivan DM, Stoddard FL, Adhikari KN, Paull JG, Schulman AH, Andersen SU, Vandenberg A: **Recent advances in faba bean genetic and genomic tools for crop improvement**. *Legume Science* 2021, **3**(3):e75.

[18] Carrillo-Perdomo E, Vidal A, Kreplak J, Duborjal H, Leveugle M, Duarte J, Desmetz C, Deulvot C, Raffiot B, Marget P, et al.: **Development of new genetic resources for faba bean (Vicia faba L.) breeding through the discovery of gene-based SNP markers and the construction of a high-density consensus map**. *Scientific reports* 2020, **10**:1–14.

[19] Köpke U, Nemecek T: **Ecological services of faba bean**. *Field crops research* 2010, **115**(3):217–233.

[20] Crépon K, Marget P, Peyronnet C, Carrouee B, Arese P, Duc G: **Nutritional value of faba bean (Vicia faba L.) seeds for feed and food**. *Field Crops Research* 2010, **115**(3):329–339.

[21] Duc G, Sixdenier G, Lila M, Furstoss V: **Search of genetic variability for vicine and convicine content in Vicia faba L.: a first report of a gene which codes for nearly zero-vicine and zero-convicine contents**. In *1. International Workshop on'Antinutritional Factors (ANF) in Legume Seeds', Wageningen (Netherlands), 23-25 Nov 1988*, Pudoc 1989.

[22] Brown E, Roberts F: **Formation of vicine and convicine by Vicia faba**. *Phytochemistry* 1972, **11**(11):3203–3206.

[23] Khazaei H, Purves RW, Hughes J, Link W, O'Sullivan DM, Schulman AH, Björnsdotter E, Geu-Flores F, Nadzieja M, Andersen SU, et al.: **Eliminating vicine and convicine, the main anti-nutritional factors restricting faba bean usage**. *Trends in Food Science & Technology* 2019, **91**:549–556.

[24] Arese P, Gallo V, Pantaleo A, Turrini F: **Life and death of glucose-6-phosphate dehydrogenase (G6PD) deficient erythrocytes–role of redox stress and band 3 modifications**. *Transfusion Medicine and Hemotherapy* 2012, **39**(5):328–334.

[25] **Abo-Vici-Projekt**. *https://www.uni-goettingen.de/de/abo-vici-projekt/559637.html*. [(accessed on 07 December 2021)].

[26] **Protein Crop Strategy**. *https://www.bmel.de/EN/topics/farming/plant-production/protein-crop-strategy.html*. [(accessed on 07 December 2021)].

[27] Björnsdotter E, Nadzieja M, Chang W, Escobar-Herrera L, Mancinotti D, Angra D, Xia X, Tacke R, Khazaei H, Crocoll C, et al.: **VC1 catalyses a key step in the biosynthesis of vicine in faba bean**. *Nature plants* 2021, **7**(7):923–931.

[28] Ren J, Kotaka M, Lockyer M, Lamb HK, Hawkins AR, Stammers DK: **GTP cyclohydrolase II structure and mechanism**. *Journal of Biological Chemistry* 2005, **280**(44):36912–36919.

[29] Fang L, Ahn JK, Wodziak D, Sibley E: **The human lactase persistence-associated SNP- 13910* T enables in vivo functional persistence of lactase promoter–reporter transgene expression**. *Human genetics* 2012, **131**(7):1153–1159.

[30] De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, De Jong P, et al.: **A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter**. *Science* 2006, **312**(5777):1215–1217.

[31] Ryan MT, Hamill RM, O'Halloran AM, Davey GC, McBryan J, Mullen AM, McGee C, Gispert M, Southwood OI, Sweeney T: **SNP variation in the promoter of the PRKAG3 gene and association with meat quality traits in pig**. *BMC genetics* 2012, **13**:1–17.

[32] Barkova OY, Sazanova KA, Fomichev KA, Malewski T, Parada R, Kawka M, Jaszczak K, Sazanov AA, et al.: **Associations of new rSNPs with eggshell thickness in Rhode Island layers**. *Animal Science Papers and Reports* 2013, **31**(2):165–172.

[33] Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M: **An SNP caused loss of seed shattering during rice domestication**. *Science* 2006, **312**(5778):1392–1396.

[34] Ordovas L, Roy R, Pampín S, Zaragoza P, Osta R, Rodriguez-Rey JC, Rodellar C: **The g. 763G> C SNP of the bovine FASN gene affects its promoter activity via Sp-mediated regulation: implications for the bovine lactating mammary gland**. *Physiological genomics* 2008, **34**(2):144–148.

[35] Klees S, Heinrich F, Schmitt AO, Gültas M: **agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species**. *Biology* 2021, **10**(8):790.

[36] Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition**. *Genome research* 1997, **7**(9):861–878.

[37] Shahmuradov IA, Solovyev VV, Gammerman A: **Plant promoter prediction with confidence estimation**. *Nucleic acids research* 2005, **33**(3):1069–1076.

[38] Ohler U: **Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction**. *Nucleic acids research* 2006, **34**(20):5943–5950.

[39] Morey C, Mookherjee S, Rajasekaran G, Bansal M: **DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes**. *Plant physiology* 2011, **156**(3):1300–1315.

[40] Azad A, Shahid S, Noman N, Lee H: **Prediction of plant promoters based on hexamers and random triplet pair analysis**. *Algorithms for Molecular Biology* 2011, **6**:1–10.

[41] Lai HY, Zhang ZY, Su ZD, Su W, Ding H, Chen W, Lin H: **iProEP: a computational predictor for predicting promoter**. *Molecular Therapy-Nucleic Acids* 2019, **17**:337–346.

[42] Abeel T, Saeys Y, Rouzé P, Van de Peer Y: **ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles**. *Bioinformatics* 2008, **24**(13):i24–i31.

[43] Anwar F, Baker SM, Jabid T, Hasan MM, Shoyaib M, Khan H, Walshe R: **Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach**. *BMC bioinformatics* 2008, **9**:1–8.

[44] Umarov RK, Solovyev VV: **Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks**. *PloS one* 2017, **12**(2):e0171410.

[45] Umarov R, Kuwahara H, Li Y, Gao X, Solovyev V: **Promoter analysis and prediction in the human genome using sequence-based deep learning models**. *Bioinformatics* 2019, **35**(16):2730–2737.

[46] Triska M, Solovyev V, Baranova A, Kel A, Tatarinova TV: **Nucleotide patterns aiding in prediction of eukaryotic promoters**. *PloS one* 2017, **12**(11):e0187243.

[47] Qian Y, Zhang Y, Guo B, Ye S, Wu Y, Zhang J: **An improved promoter recognition model using convolutional neural network**. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Volume 1*, IEEE 2018:471–476.

[48] Oubounyt M, Louadi Z, Tayara H, Chong KT: **DeePromoter: robust promoter predictor using deep learning**. *Frontiers in genetics* 2019, **10**:286.

[49] Pachganov S, Murtazalieva K, Zarubin A, Sokolov D, Chartier DR, Tatarinova TV: **TransPrise: a novel machine learning approach for eukaryotic promoter prediction**. *PeerJ* 2019, **7**:e7990.

[50] Kumari S, Ware D: **Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots**. *PloS one* 2013, **8**(10):e79011.

[51] Shahmuradov IA, Umarov RK, Solovyev VV: **TSSPlant: a new tool for prediction of plant Pol II promoters**. *Nucleic acids research* 2017, **45**(8):e65–e65.

[52] Heinrich F, Wutke M, Das PP, Kamp M, Gültas M, Link W, Schmitt AO: **Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning**. *Genes* 2020, **11**(6):614.

[53] Wei WH, Hemani G, Haley CS: **Detecting epistasis in human complex traits**. *Nature Reviews Genetics* 2014, **15**(11):722–733.

[54] Phillips PC: **Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems**. *Nature Reviews Genetics* 2008, **9**(11):855–867.

[55] Huang W, Richards S, Carbone MA, Zhu D, Anholt RR, Ayroles JF, Duncan L, Jordan KW, Lawrence F, Magwire MM, et al.: **Epistasis dominates the genetic architecture of Drosophila quantitative traits**. *Proceedings of the National Academy of Sciences* 2012, **109**(39):15553–15559.

[56] Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies**. *Bioinformatics* 2010, **26**(4):445–455.

[57] Moore JH, Williams SM: **Epistasis and its implications for personal genetics**. *The American Journal of Human Genetics* 2009, **85**(3):309–320.

[58] Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases**. *Nature genetics* 2005, **37**(4):413–417.

[59] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al.: **Common SNPs explain a large proportion of the heritability for human height**. *Nature genetics* 2010, **42**(7):565–569.

[60] Yoshikawa T, Kanazawa H, Fujimoto S, Hirata K: **Epistatic effects of multiple receptor genes on pathophysiology of asthma–its limits and potential for clinical application**. *Medical science monitor: international medical journal of experimental and clinical research* 2014, **20**:64.

[61] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer**. *The American Journal of Human Genetics* 2001, **69**:138–147.

[62] Cho Y, Ritchie M, Moore J, Park J, Lee KU, Shin H, Lee H, Park K: **Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus**. *Diabetologia* 2004, **47**(3):549–554.

[63] Carlborg Ö, Hocking PM, Burt DW, Haley CS: **Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth**. *Genetics Research* 2004, **83**(3):197–209.

[64] Le Rouzic A, Álvarez-Castro JM, Carlborg O: **Dissection of the genetic architecture of body weight in chicken reveals the impact of epistasis on domestication traits**. *Genetics* 2008, **179**(3):1591–1599.

[65] Mackay TF: **Epistasis and quantitative traits: using model organisms to study gene–gene interactions**. *Nature Reviews Genetics* 2014, **15**:22–33.

[66] Knaust J, Hadlich F, Weikard R, Kuehn C: **Epistatic interactions between at least three loci determine the "rat-tail" phenotype in cattle**. *Genetics Selection Evolution* 2016, **48**:1–12.

[67] Kramer LM, Ghaffar MA, Koltes J, Fritz-Waters E, Mayes M, Sewell A, Weeks N, Garrick D, Fernando R, Ma L, et al.: **Epistatic interactions associated with fatty acid concentrations of beef from angus sired beef cattle**. *BMC genomics* 2016, **17**:1–12.

[68] Würschum T, Maurer HP, Schulz B, Möhring J, Reif JC: **Genome-wide association mapping reveals epistasis and genetic interaction networks in sugar beet**. *Theoretical and applied genetics* 2011, **123**:109–118.

[69] Hu Z, Li Y, Song X, Han Y, Cai X, Xu S, Li W: **Genomic value prediction for quantitative traits under the epistatic model**. *BMC genetics* 2011, **12**:1–11.

[70] Huang A, Xu S, Cai X: **Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice**. *PloS one* 2014, **9**:e87330.

[71] Ahsan A, Monir M, Meng X, Rahaman M, Chen H, Chen M: **Identification of epistasis loci underlying rice flowering time by controlling population stratification and polygenic effect**. *DNA Research* 2019, **26**(2):119–130.

[72] Mathew B, Léon J, Sannemann W, Sillanpää MJ: **Detection of epistasis for flowering time using Bayesian multilocus estimation in a barley MAGIC population**. *Genetics* 2018, **208**(2):525–536.

[73] Carlborg Ö, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Nature Reviews Genetics* 2004, **5**(8):618–625.

[74] Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans**. *Human molecular genetics* 2002, **11**(20):2463–2468.

[75] Anastassiou D: **Computational analysis of the synergy among multiple interacting genes**. *Molecular systems biology* 2007, **3**:83.

[76] Leem S, Jeong Hh, Lee J, Wee K, Sohn KA: **Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure**. *Computational biology and chemistry* 2014, **50**:19–28.

[77] Tuo S: **FDHE-IW: A fast approach for detecting high-order epistasis in genome-wide case-control studies**. *Genes* 2018, **9**(9):435.

[78] Hu T, Chen Y, Kiralis JW, Collins RL, Wejse C, Sirugo G, Williams SM, Moore JH: **An information-gain approach to detecting three-way epistatic interactions in**

genetic association studies. *Journal of the American Medical Informatics Association* 2013, **20**(4):630–636.

[79] Anunciação O, Vinga S, Oliveira AL: **Using information interaction to discover epistatic effects in complex diseases**. *PLoS One* 2013, **8**(10):e76300.

[80] Wienbrandt L, Kassens JC, Hübenthal M, Ellinghaus D: **Fast genome-wide third-order snp interaction tests with information gain on a low-cost heterogeneous parallel fpga-gpu computing architecture**. *Procedia computer science* 2017, **108**:596–605.

[81] Ponte-Fernández C, González-Domínguez J, Martín MJ: **Fast search of third-order epistatic interactions on cpu and gpu clusters**. *The International Journal of High Performance Computing Applications* 2020, **34**:20–29.

[82] Cao X, Yu G, Liu J, Jia L, Wang J: **Clustermi: Detecting high-order snp interactions based on clustering and mutual information**. *International journal of molecular sciences* 2018, **19**(8):2267.

[83] González-Domínguez J, Schmidt B: **GPU-accelerated exhaustive search for third-order epistatic interactions in case–control studies**. *Journal of Computational Science* 2015, **8**:93–100.

[84] Wang S, Jeong Hh, Kim D, Wee K, Park HS, Kim SH, Sohn KA: **Integrative information theoretic network analysis for genome-wide association study of aspirin exacerbated respiratory disease in Korean population**. *BMC medical genomics* 2017, **10**:33–44.

[85] Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C: **Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness**. *Heredity* 2012, **108**(3):285–291.

[86] Mezmouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S, Mangin B: **Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels**. *Theoretical and applied genetics* 2011, **122**(6):1149–1160.

[87] Ross BC: **Mutual information between discrete and continuous data sets**. *PloS one* 2014, **9**(2):e87357.

[88] Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction**. *Bioinformatics* 2008, **24**(3):333–340.

[89] Bourdon RM, Bourbon RM: *Understanding animal breeding*. Prentice Hall Upper Saddle River, NJ 2000.

[90] Falconer DS: *Introduction to quantitative genetics*. Burnt Mill, Harlow, Essex, England: Longman, Scientific & Technical, 3rd edition. edition 1989.

[91] Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R: *Molecular Biology of the Gene - International Edition*. New York: Pearson Education, Limited, 6th edition. edition 2008.

[92] Berg J, Guglielmi J, Tymoczko J, Held A, Stryer L, Kuhlmann-Krieg S, Pfeiffer-Guglielmi B, Seidler L, Vogel S, von der Saal K, et al.: *Biochemie*. Spektrum Akademischer Verlag 2003, [https://books.google.de/books?id=LPTdAAAACAAJ].

[93] Campbell NA, Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, Jackson RB: *Biology*. New York: Pearson Education, Limited, 8th edition. edition 2008.

[94] Slack J: *Genes: a very short introduction*. Oxford University Press, 1st edition. edition 2014.

[95] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell*. New York: Garland Science, 5th edition. edition 2008.

[96] Mora A, Sandve GK, Gabrielsen OS, Eskeland R: **In the loop: promoter–enhancer interactions and bioinformatics**. *Briefings in bioinformatics* 2016, **17**(6):980–995.

[97] Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT: **The RNA polymerase II core promoter—the gateway to transcription**. *Current opinion in cell biology* 2008, **20**(3):253–259.

[98] Haberle V, Lenhard B: **Promoter architectures and developmental gene regulation**. In *Seminars in cell & developmental biology*, *Volume 57*, Elsevier 2016:11–23.

[99] Stepanova M, Tiazhelova T, Skoblov M, Baranova A: **A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas**. *Bioinformatics* 2005, **21**(9):1789–1796.

[100] Hernandez-Garcia CM, Finer JJ: **Identification and validation of promoters and cis-acting regulatory elements**. *Plant Science* 2014, **217**:109–119.

[101] Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome**. *Annu. Rev. Genomics Hum. Genet.* 2006, **7**:29–59.

[102] Stormo GD: **Modeling the specificity of protein-DNA interactions**. *Quantitative biology* 2013, **1**(2):115–130.

[103] Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic acids research* 1990, **18**(20):6097–6100.

[104] Ou J, Wolfe SA, Brodsky MH, Zhu LJ: **motifStack for the analysis of transcription factor binding site evolution**. *Nature methods* 2018, **15**:8–9.

[105] Wingender E: **The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation**. *Briefings in bioinformatics* 2008, **9**(4):326–332.

[106] Adhikari S, Saha S, Biswas A, Rana T, Bandyopadhyay TK, Ghosh P: **Application of molecular markers in plant genome analysis: a review**. *The Nucleus* 2017, **60**(3):283–297.

[107] Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S: **SNP markers and their impact on plant breeding**. *International journal of plant genomics* 2012, **2012**.

[108] Jehan T, Lakhanpaul S: **Single nucleotide polymorphism (SNP)–methods and applications in plant genetics: a review**. *Indian Journal of Biotechnology* 2006, :435–459.

[109] Brookes AJ: **The essence of SNPs**. *Gene* 1999, **234**(2):177–186.

[110] Vignal A, Milan D, SanCristobal M, Eggen A: **A review on SNP and other types of molecular markers and their use in animal genetics**. *Genetics selection evolution* 2002, **34**(3):275–305.

[111] Klees S, Lange TM, Bertram H, Rajavel A, Schlüter JS, Lu K, Schmitt AO, Gültas M: **In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in Brassica napus L. Using Multi-Omics Data**. *International Journal of Molecular Sciences* 2021, **22**(2):789.

[112] Riva A: **Large-scale computational identification of regulatory SNPs with rSNP-MAPPER**. *BMC Genomics* 2012, **13**(4):1–8.

[113] Degtyareva AO, Antontseva EV, Merkulova TI: **Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases**. *International Journal of Molecular Sciences* 2021, **22**(12):6454.

[114] Guo L, Du Y, Qu S, Wang J: **rVarBase: an updated database for regulatory features of human variants**. *Nucleic acids research* 2016, **44**(D1):D888–D893.

[115] Xu Z, Taylor JA: **SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies**. *Nucleic acids research* 2009, **37**(suppl_2):W600–W605.

[116] Stram DO: *Design, analysis, and interpretation of genome-wide association scans*, *Volume 15*. Springer 2014.

[117] Fan JB, Oliphant A, Shen R, Kermani B, Garcia F, Gunderson K, Hansen M, Steemers F, Butler S, Deloukas P, et al.: **Highly parallel SNP genotyping**. In *Cold Spring Harbor symposia on quantitative biology*, *Volume 68*, Cold Spring Harbor Laboratory Press 2003:69–78.

[118] Poland JA, Rife TW: **Genotyping-by-sequencing for plant breeding and genetics**. *The Plant Genome* 2012, **5**(3).

[119] Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nature Reviews Genetics* 2011, **12**(6):443–451.

[120] Russell J, Hackett C, Hedley P, Liu H, Milne L, Bayer M, Marshall D, Jorgensen L, Gordon S, Brennan R: **The use of genotyping by sequencing in blackcurrant (Ribes nigrum): developing high-resolution linkage maps in species without reference genome sequences**. *Molecular breeding* 2014, **33**(4):835–849.

[121] Niel C, Sinoquet C, Dina C, Rocheleau G: **A survey about methods dedicated to epistasis detection**. *Frontiers in genetics* 2015, **6**:285.

[122] Waddington CH: **Canalization of development and the inheritance of acquired characters**. *Nature* 1942, **150**(3811):563–565.

[123] Bateson W, Mendel G: *Mendel's principles of heredity*. Courier Corporation 2013.

[124] Fisher RA: **XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.** *Transactions of the Royal Society of Edinburgh* 1919, **52**(2):399–433.

[125] Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al.: **Ensembl 2020**. *Nucleic acids research* 2020, **48**(D1):D682–D688.

[126] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al.: **Ensembl 2021**. *Nucleic acids research* 2021, **49**(D1):D884–D891.

[127] Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S, et al.: **Ensembl comparative genomics resources**. *Database* 2016, **2016**.

[128] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The ensembl variant effect predictor**. *Genome biology* 2016, **17**:1–14.

[129] Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, Alvarez-Jarreta J, Barba M, Bolser DM, Cambell L, et al.: **Ensembl Genomes 2020—enabling non-vertebrate genomic research**. *Nucleic acids research* 2020, **48**(D1):D689–D695.

[130] Humann JL, Jung S, Cheng CH, Lee T, Zheng P, Frank M, McGaughey D, Scott K, Buble K, Yu J, et al.: **Cool Season Food Legume Genome Database: A resource for pea, lentil, faba bean and chickpea genetics, genomics and breeding**. In *Plant and Animal Genome XXVII Conference (January 12-16, 2019)*, PAG 2019.

[131] Sahruzaini NA, Rejab NA, Harikrishna JA, Khairul Ikram NK, Ismail I, Kugan HM, Cheng A: **Pulse crop genetics for a sustainable future: Where we are now and where we should be heading**. *Frontiers in plant science* 2020, **11**:531.

[132] Wingender E: **Compilation of transcription regulating proteins.** *Nucleic acids research* 1988, **16**(5 Pt B):1879.

[133] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al.: **TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes**. *Nucleic acids research* 2006, **34**(suppl_1):D108–D110.

[134] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al.: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *The American journal of human genetics* 2007, **81**(3):559–575.

[135] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**:s13742–015.

[136] Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCHTM: a tool for searching transcription factor binding sites in DNA sequences**. *Nucleic acids research* 2003, **31**(13):3576–3579.

[137] Quandt K, Frech K, Karas H, Wingender E, Werner T: **Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data**. *Nucleic acids research* 1995, **23**(23):4878–4884.

[138] Kel A, Kel-Margoulis O, Babenko V, Wingender E: **Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells**. *Journal of molecular biology* 1999, **288**(3):353–376.

[139] Wingender E, Kel A: **geneXplain—eine integrierte Bioinformatik-Plattform**. *BIOspektrum* 2012, **18**(5):554–556.

[140] Kolpakov F, Poroikov V, Selivanova G, Kel A: **GeneXplain—identification of causal biomarkers and drug targets in personalized cancer pathways**. *Journal of biomolecular techniques: JBT* 2011, **22**(Suppl):S16.

[141] Goodfellow I, Bengio Y, Courville A: *Deep learning*. MIT press 2016.

[142] Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites**. *Journal of molecular biology* 1995, **249**(5):923–932.

[143] Reese MG: **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome**. *Computers & chemistry* 2001, **26**:51–56.

[144] Hutchinson G: **The prediction of vertebrate promoter regions using differential hexamer frequency analysis**. *Bioinformatics* 1996, **12**(5):391–398.

[145] Prestrldge DS, Burks C: **The density of transcriptional elements in promoter and non-promoter sequences**. *Human molecular genetics* 1993, **2**(9):1449–1453.

[146] Li J, Zhang L, Li H, Ping Y, Xu Q, Wang R, Tan R, Wang Z, Liu B, Wang Y: **Integrated entropy-based approach for analyzing exons and introns in DNA sequences**. *BMC bioinformatics* 2019, **20**(8):1–7.

[147] Schmidhuber J: **Deep learning in neural networks: An overview**. *Neural networks* 2015, **61**:85–117.

[148] LeCun Y, Bengio Y, Hinton G: **Deep learning**. *nature* 2015, **521**(7553):436–444.

[149] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: **Going deeper with convolutions**. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2015:1–9.

[150] Al-Ajlan A, El Allali A: **CNN-MGP: Convolutional neural networks for metagenomics gene prediction**. *Interdisciplinary Sciences: Computational Life Sciences* 2019, **11**(4):628–635.

[151] Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning**. *Nature biotechnology* 2015, **33**(8):831–838.

[152] Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning–based sequence model**. *Nature methods* 2015, **12**(10):931–934.

[153] Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M: **Medical image classification with convolutional neural network**. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, IEEE 2014:844–848.

[154] Shannon CE: **A mathematical theory of communication**. *The Bell system technical journal* 1948, **27**(3):379–423.

[155] Cover, TM; Thomas, JA: *Elements of Information Theory*. John Wiley, New York. 2006.

[156] Kraskov A, Stögbauer H, Grassberger P: **Estimating mutual information**. *Physical review E* 2004, **69**(6):066138.

[157] Kozachenko L, Leonenko NN: **Sample estimate of the entropy of a random vector**. *Problemy Peredachi Informatsii* 1987, **23**(2):9–16.

[158] Abramowitz M, Stegun IA: *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, *Volume 55*. US Government printing office 1964.

[159] Kvålseth TO: **On normalized mutual information: Measure derivations and properties**. *Entropy* 2017, **19**(11):631.

[160] Korte A, Farlow A: **The advantages and limitations of trait analysis with GWAS: a review**. *Plant methods* 2013, **9**:1–9.

[161] Schmid M, Bennewitz J: **Invited review: Genome-wide association analysis for quantitative traits in livestock–a selective review of statistical models and experimental designs**. *Archives Animal Breeding* 2017, **60**(3):335–346.

[162] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J: **10 years of GWAS discovery: biology, function, and translation**. *The American Journal of Human Genetics* 2017, **101**:5–22.

[163] Bush WS, Moore JH: **Chapter 11: Genome-wide association studies**. *PLoS computational biology* 2012, **8**(12):e1002822.

[164] Emily M, Friguet C: **A Mutual Information-based method to select informative pairs of variables in case-control genetic association studies to improve the power of detecting interaction between genetic variants**. *Journal de la société française de statistique* 2018, **159**(2):84–110.

[165] Heinrich F, Ramzan F, Rajavel A, Schmitt AO, Gültas M: **MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes**. *Biology* 2021, **10**(9):921.

[166] Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W: **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies**. *The American Journal of Human Genetics* 2010, **87**(3):325–340.

[167] He D, Parida L: **Does encoding matter? a novel view on the quantitative genetic trait prediction problem**. *BMC bioinformatics* 2016, **17**(9):1–9.

[168] Li M, Lou XY, Lu Q: **On epistasis: a methodological review for detecting gene-gene interactions underlying various types of phenotypic traits**. *Recent patents on biotechnology* 2012, **6**(3):230–236.

[169] Nelson M, Kardia S, Ferrell R, Sing C: **A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation**. *Genome research* 2001, **11**(3):458–470.

[170] Culverhouse R: **The use of the restricted partition method with case-control data**. *Human heredity* 2007, **63**(2):93–100.

[171] Li X: **A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization**. *Bioinformatics* 2017, **33**(18):2829–2836.

[172] Zhang X, Huang S, Zou F, Wang W: **TEAM: efficient two-locus epistasis tests in human genome-wide association study**. *Bioinformatics* 2010, **26**(12):i217–i227.

[173] Zhang Y, Liu JS: **Bayesian inference of epistatic interactions in case-control studies**. *Nature genetics* 2007, **39**(9):1167–1173.

[174] Tang W, Wu X, Jiang R, Li Y: **Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy**. *PLoS genetics* 2009, **5**(5):e1000464.

[175] Serretti A, Smeraldi E: **Neural network analysis in pharmacogenetics of mood disorders**. *BMC Medical Genetics* 2004, **5**:1–6.

[176] Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD: **Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology**. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2008, **32**(4):325–340.

[177] Uppu S, Krishna A, Gopalan RP: **Towards Deep Learning in genome-Wide Association Interaction studies.** In *PACIS*, Taiwan 2016:20.

[178] Wang H, Yue T, Yang J, Wu W, Xing EP: **Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies**. *BMC bioinformatics* 2019, **20**(23):1–11.

[179] Xie Q, Ratnasinghe LD, Hong H, Perkins R, Tang ZZ, Hu N, Taylor PR, Tong W: **Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method**. *BMC bioinformatics* 2005, **6**(2):1–9.

[180] Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, Biernacka JM: **SNP interaction detection with random forests in high-dimensional genetic data**. *BMC bioinformatics* 2012, **13**:1–13.

[181] Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL: **Performance of random forest when SNPs are in linkage disequilibrium**. *BMC bioinformatics* 2009, **10**:1–17.

[182] Schwarz DF, König IR, Ziegler A: **On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data**. *Bioinformatics* 2010, **26**(14):1752–1758.

[183] Yoshida M, Koike A: **SNPInterForest: a new method for detecting epistatic interactions**. *BMC bioinformatics* 2011, **12**:1–10.

[184] He D, Parida L: **Muse: A Multi-Locus Sampling-Based Epistasis Algorithm for Quantitative Genetic Trait Prediction**. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, World Scientific 2017:426–437.

[185] Martini JW, Gao N, Cardoso DF, Wimmer V, Erbe M, Cantet RJ, Simianer H: **Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)**. *BMC bioinformatics* 2017, **18**:1–16.

[186] Martini JW, Rosales F, Ha NT, Heise J, Wimmer V, Kneib T: **Lost in translation: On the problem of data coding in penalized whole genome regression with interactions**. *G3: Genes, Genomes, Genetics* 2019, **9**(4):1117–1129.

[187] Heinrich F, Gültas M, Link W, Schmitt AO: **Genotyping by Sequencing Reads of 20 Vicia faba Lines with High and Low Vicine and Convicine Content**. *Data* 2020, **5**(3):63.

[188] Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M: **De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti)**. *Genome biology and evolution* 2015, **7**(4):1192–1205.

[189] Yuan S, Xia Y, Zheng Y, Zeng X: **Next-generation sequencing of mixed genomic DNA allows efficient assembly of rearranged mitochondrial genomes in Amolops chunganensis and Quasipaa boulengeri**. *PeerJ* 2016, **4**:e2786.

[190] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al.: **Full-length transcriptome assem-**

bly from RNA-Seq data without a reference genome. *Nature biotechnology* 2011, **29**(7):644–652.

[191] Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-generation sequencing data**. *Bioinformatics* 2012, **28**(23):3150–3152.

[192] Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357–359.

[193] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078–2079.

[194] Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al.: **Ensembl BioMarts: a hub for data retrieval across taxonomic space**. *Database* 2011, **2011**.

[195] Bermingham M, Bishop S, Woolliams J, Pong-Wong R, Allen A, McBride S, Ryder J, Wright D, Skuce R, McDowell SW, et al.: **Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis**. *Heredity* 2014, **112**(5):543–551.

[196] Ramzan F, Gültas M, Bertram H, Cavero D, Schmitt AO: **Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations**. *Genes* 2020, **11**(8):892.

[197] Ramzan F, Klees S, Schmitt AO, Cavero D, Gültas M: **Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forests**. *Genes* 2020, **11**(4):464.

[198] Liu Z, Sun C, Yan Y, Li G, Wu G, Liu A, Yang N: **Genome-wide association analysis of age-dependent egg weights in chickens**. *Frontiers in genetics* 2018, **9**:128.

[199] Schmidt-Hieber J: **Nonparametric regression using deep neural networks with ReLU activation function**. *The Annals of Statistics* 2020, **48**(4):1875–1897.

[200] Redmon J, Divvala S, Girshick R, Farhadi A: **You only look once: Unified, real-time object detection**. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016:779–788.

[201] Kingma DP, Ba J: **Adam: A method for stochastic optimization**. *arXiv preprint arXiv:1412.6980* 2014.

[202] Yu K, Xu W, Gong Y: **Deep learning with kernel regularization for visual recognition**. *Advances in neural information processing systems* 2008, **21**:1889–1896.

[203] Chollet F, et al.: **Keras**. *https://keras.io* 2015.

[204] Abadi, M; Agarwal, A; Barham, P; Brevdo, E; Chen, Z; Citro, C; Corrado, GS; Davis, A; Dean, J; Devin, M; et al: **TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems** 2015, [https://www.tensorflow.org/]. [Software available from tensorflow.org].

[205] Boughorbel S, Jarray F, El-Anbari M: **Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric**. *PloS one* 2017, **12**(6):e0177678.

[206] Lichtenstein F, Antoneli F, Briones MR: **MIA: Mutual Information Analyzer, a graphic user interface program that calculates entropy, vertical and horizontal mutual information of molecular sequence sets**. *BMC bioinformatics* 2015, **16**:1–19.

[207] Jin S, Tan R, Jiang Q, Xu L, Peng J, Wang Y, Wang Y: **A generalized topological entropy for analyzing the complexity of DNA sequences**. *PLoS One* 2014, **9**(2):e88519.

[208] Li J, Zhang L, Li H, Ping Y, Xu Q, Wang R, Tan R, Wang Z, Liu B, Wang Y: **Integrated entropy-based approach for analyzing exons and introns in DNA sequences**. *BMC bioinformatics* 2019, **20**(8):1–7.

[209] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC bioinformatics* 2009, **10**:1–9.

[210] Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M: **FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer**. *Genome biology* 2014, **15**(10):1–15.

[211] Xu Z, Taylor JA: **SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies**. *Nucleic acids research* 2009, **37**(suppl_2):W600–W605.

[212] Joiret M, John JMM, Gusareva ES, Van Steen K: **Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies**. *BioData mining* 2019, **12**:1–23.

[213] Bayat A, Hosking B, Jain Y, Hosking C, Kodikara M, Reti D, Twine NA, Bauer DC: **Fast and accurate exhaustive higher-order epistasis search with BitEpi**. *Scientific reports* 2021, **11**:1–12.

[214] Gültas M, Haubrock M, Tüysüz N, Waack S: **Coupled mutation finder: a new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations**. *BMC bioinformatics* 2012, **13**:1–12.

[215] Gültas M, Düzgün G, Herzog S, Jäger SJ, Meckbach C, Wingender E, Waack S: **Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming**. *BMC bioinformatics* 2014, **15**:1–17.

[216] Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proceedings of the National Academy of Sciences* 2003, **100**(16):9440–9445.

[217] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal statistical society: series B (Methodological)* 1995, **57**:289–300.

[218] Gültas, M: **Development of novel Classical and Quantum Information Theory Based Methods for the Detection of Compensatory Mutations in MSAs**. *PhD thesis*, Georg-August University Göttingen 2014.

[219] Walsh B: **Multiple comparisons: Bonferroni corrections and false discovery rates**. *Lecture Notes EEB* 2004, **581**.

[220] Meckbach C, Tacke R, Hua X, Waack S, Wingender E, Gültas M: **PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information**. *BMC bioinformatics* 2015, **16**:1–21.

[221] Csardi G, Nepusz T, et al.: **The igraph software package for complex network research**. *InterJournal, complex systems* 2006, **1695**(5):1–9.

[222] Mekonnen YA, Gültas M, Effa K, Hanotte O, Schmitt AO: **Identification of candidate signature genes and key regulators associated with Trypanotolerance in the Sheko Breed**. *Frontiers in genetics* 2019, **10**:1095.

[223] Hane JK, Ming Y, Kamphuis LG, Nelson MN, Garg G, Atkins CA, Bayer PE, Bravo A, Bringans S, Cannon S, et al.: **A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant–microbe interactions and legume evolution**. *Plant biotechnology journal* 2017, **15**(3):318–330.

[224] Kang YJ, Satyawan D, Shim S, Lee T, Lee J, Hwang WJ, Kim SK, Lestari P, Laosatit K, Kim KH, et al.: **Draft genome sequence of adzuki bean, Vigna angularis**. *Scientific reports* 2015, **5**:1–8.

[225] Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, Gould JA, Forster SC, Hertzog PJ: **CiiiDER: A tool for predicting and analysing transcription factor binding sites**. *PLoS One* 2019, **14**(9):e0215495.

[226] Heath RJ, Rock CO: **Roles of the FabA and FabZ β-hydroxyacyl-acyl carrier protein dehydratases in Escherichia coli fatty acid biosynthesis**. *Journal of Biological Chemistry* 1996, **271**(44):27795–27801.

[227] Lin S, Hanson RE, Cronan JE: **Biotin synthesis begins by hijacking the fatty acid synthetic pathway**. *Nature chemical biology* 2010, **6**(9):682–688.

[228] Smaczniak C, Immink RG, Angenent GC, Kaufmann K: **Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies**. *Development* 2012, **139**(17):3081–3098.

[229] Riechmann JL, Ratcliffe OJ: **A genomic perspective on plant transcription factors**. *Current opinion in plant biology* 2000, **3**(5):423–434.

[230] Ping J, Liu Y, Sun L, Zhao M, Li Y, She M, Sui Y, Lin F, Liu X, Tang Z, et al.: **Dt2 is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean**. *The Plant Cell* 2014, **26**(7):2831–2842.

[231] Danyluk J, Kane NA, Breton G, Limin AE, Fowler DB, Sarhan F: **TaVRT-1, a putative transcription factor associated with vegetative to reproductive transition in cereals**. *Plant Physiology* 2003, **132**(4):1849–1860.

[232] West AG, Sharrocks AD, Causier BE, Davies B: **DNA binding and dimerisation determinants of Antirrhinum majus MADS-box transcription factors**. *Nucleic Acids Research* 1998, **26**(23):5277–5287.

[233] Theißen G, Melzer R, Rümpler F: **MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution**. *Development* 2016, **143**(18):3259–3271.

[234] Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L: **MYB transcription factors in Arabidopsis**. *Trends in plant science* 2010, **15**(10):573–581.

[235] Roy S: **Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome**. *Plant signaling & behavior* 2016, **11**:e1117723.

[236] Fu F, Zhang W, Li YY, Wang HL: **Establishment of the model system between phytochemicals and gene expression profiles in Macrosclereid cells of Medicago truncatula**. *Scientific reports* 2017, **7**:1–16.

[237] Jin H, Cominelli E, Bailey P, Parr A, Mehrtens F, Jones J, Tonelli C, Weisshaar B, Martin C: **Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis**. *The EMBO journal* 2000, **19**(22):6150–6161.

[238] Agarwal P, Mitra M, Banerjee S, Roy S: **MYB4 transcription factor, a member of R2R3-subfamily of MYB domain protein, regulates cadmium tolerance via enhanced protection against oxidative damage and increases expression of PCS1 and MT1C in Arabidopsis**. *Plant Science* 2020, **297**:110501.

[239] Vannini C, Locatelli F, Bracale M, Magnani E, Marsoni M, Osnato M, Mattana M, Baldoni E, Coraggio I: **Overexpression of the rice Osmyb4 gene increases chilling and freezing tolerance of Arabidopsis thaliana plants**. *The Plant Journal* 2004, **37**:115–127.

[240] Wang XC, Wu J, Guan ML, Zhao CH, Geng P, Zhao Q: **Arabidopsis MYB4 plays dual roles in flavonoid biosynthesis**. *The Plant Journal* 2020, **101**(3):637–652.

[241] Zhang Z, Hu X, Zhang Y, Miao Z, Xie C, Meng X, Deng J, Wen J, Mysore KS, Frugier F, et al.: **Opposing control by transcription factors MYB61 and MYB3 increases freezing tolerance by relieving C-repeat binding factor suppression**. *Plant physiology* 2016, **172**(2):1306–1323.

[242] Romano JM, Dubos C, Prouse MB, Wilkins O, Hong H, Poole M, Kang KY, Li E, Douglas CJ, Western TL, et al.: **AtMYB61, an R2R3-MYB transcription factor, functions as a pleiotropic regulator via a small gene network**. *New Phytologist* 2012, **195**(4):774–786.

[243] Matías-Hernández L, Jiang W, Yang K, Tang K, Brodelius PE, Pelaz S: **Aa MYB 1 and its orthologue At MYB 61 affect terpene metabolism and trichome development in Artemisia annua and Arabidopsis thaliana**. *The Plant Journal* 2017, **90**(3):520–534.

[244] Liang YK, Dubos C, Dodd IC, Holroyd GH, Hetherington AM, Campbell MM: **AtMYB61, an R2R3-MYB transcription factor controlling stomatal aperture in Arabidopsis thaliana**. *Current Biology* 2005, **15**(13):1201–1206.

[245] Arsovski AA, Villota MM, Rowland O, Subramaniam R, Western TL: **MUM ENHANCERS are important for seed coat mucilage production and mucilage secretory cell differentiation in Arabidopsis thaliana**. *Journal of experimental botany* 2009, **60**(9):2601–2612.

[246] Penfield S, Meissner RC, Shoue DA, Carpita NC, Bevan MW: **MYB61 is required for mucilage deposition and extrusion in the Arabidopsis seed coat**. *The Plant Cell* 2001, **13**(12):2777–2791.

[247] Ramsay G, Griffiths DW: **Accumulation of vicine and convicine in Vicia faba and V. narbonensis**. *Phytochemistry* 1996, **42**:63–67.

[248] Guo H, Yu Z, An J, Han G, Ma Y, Tang R: **A two-stage mutual information based Bayesian Lasso algorithm for multi-locus genome-wide association studies**. *Entropy* 2020, **22**(3):329.

[249] Sun L, Wang C, Hu YQ: **Utilizing mutual information for detecting rare and common variants associated with a categorical trait**. *PeerJ* 2016, **4**:e2139.

[250] Yuan X, Zhang J, Wang Y: **Mutual information and linkage disequilibrium based SNP association study by grouping case-control**. *Genes & Genomics* 2011, **33**:65–73.

[251] Speed D, Hemani G, Johnson MR, Balding DJ: **Improved heritability estimation from genome-wide SNPs**. *The American Journal of Human Genetics* 2012, **91**(6):1011–1021.

[252] Machado D, Pires D, Perdigão J, Couto I, Portugal I, Martins M, Amaral L, Anes E, Viveiros M: **Ion channel blockers as antimicrobial agents, efflux inhibitors, and enhancers of macrophage killing activity against drug resistant Mycobacterium tuberculosis**. *PloS one* 2016, **11**(2):e0149326.

[253] Viveiros M, Martins M, Rodrigues L, Machado D, Couto I, Ainsa J, Amaral L: **Inhibitors of mycobacterial efflux pumps as potential boosters for anti-tubercular drugs**. *Expert review of anti-infective therapy* 2012, **10**(9):983–998.

[254] Martins M, Viveiros M, Couto I, Amaral L: **Targeting human macrophages for enhanced killing of intracellular XDR-TB and MDR-TB**. *The International journal of tuberculosis and lung disease* 2009, **13**(5):569–573.

[255] Gupta S, Salam N, Srivastava V, Singla R, Behera D, Khayyam KU, Korde R, Malhotra P, Saxena R, Natarajan K: **Voltage gated calcium channels negatively regulate protective immunity to Mycobacterium tuberculosis**. *PLoS One* 2009, **4**(4):e5305.

[256] Anes E: **Acting on Actin During Bacterial Infection**. In *Cytoskeleton-Structure, Dynamics, Function and Disease*, IntechOpen 2017.

[257] Hestvik ALK, Hmama Z, Av-Gay Y: **Mycobacterial manipulation of the host cell**. *FEMS microbiology reviews* 2005, **29**(5):1041–1050.

[258] Guérin I, de Chastellier C: **Pathogenic mycobacteria disrupt the macrophage actin filament network**. *Infection and immunity* 2000, **68**(5):2655–2662.

[259] Bettencourt P, Marion S, Pires D, Santos L, Lastrucci C, Carmo N, Blake J, Benes V, Griffiths G, Neyrolles O, et al.: **Actin-binding protein regulation by microRNAs as a novel microbial strategy to modulate phagocytosis by host cells: the case of N-Wasp and miR-142-3p**. *Frontiers in cellular and infection microbiology* 2013, **3**:19.

[260] Wang J, Yao Y, Wu J, Deng Z, Gu T, Tang X, Cheng Y, Li G: **The mechanism of cytoskeleton protein β-actin and cofilin-1 of macrophages infected by Mycobacterium avium**. *American journal of translational research* 2016, **8**(2):1055.

[261] Levite M: **Neurotransmitters activate T-cells and elicit crucial functions via neurotransmitter receptors**. *Current opinion in pharmacology* 2008, **8**(4):460–471.

[262] Pacheco R, Riquelme E, Kalergis AM: **Emerging evidence for the role of neurotransmitters in the modulation of T cell responses to cognate ligands**. *Central Nervous System Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Central Nervous System Agents)* 2010, **10**:65–83.

[263] Skinner MA, Parlane N, McCarthy A, Buddle BM: **Cytotoxic T-cell responses to Mycobacterium bovis during experimental infection of cattle with bovine tuberculosis**. *Immunology* 2003, **110**(2):234–241.

[264] Villarreal-Ramos B, McAulay M, Chance V, Martin M, Morgan J, Howard C: **Investigation of the role of CD8+ T cells in bovine tuberculosis in vivo**. *Infection and immunity* 2003, **71**(8):4297–4303.

[265] Pollock J, Neill S: **Mycobacterium boviss infection and tuberculosis in cattle**. *The Veterinary Journal* 2002, **163**(2):115–127.

[266] Finlay EK, Berry DP, Wickham B, Gormley EP, Bradley DG: **A genome wide association scan of bovine tuberculosis susceptibility in Holstein-Friesian dairy cattle**. *PloS one* 2012, **7**(2):e30545.

[267] Pacheco R, Gallart T, Lluis C, Franco R: **Role of glutamate on T-cell mediated immunity**. *Journal of neuroimmunology* 2007, **185**(1-2):9–19.

[268] Ganor Y, Levite M: **The neurotransmitter glutamate and human T cells: glutamate receptors and glutamate-induced direct and potent effects on normal human T cells, cancerous human leukemia and lymphoma T cells, and autoimmune human T cells**. *Journal of neural transmission* 2014, **121**(8):983–1006.

[269] El Masri R, Delon J: **RHO GTPases: from new partners to complex immune syndromes**. *Nature Reviews Immunology* 2021, :1–15.

[270] Bokoch GM: **Regulation of innate immunity by Rho GTPases**. *Trends in cell biology* 2005, **15**(3):163–171.

[271] Chopra P, Koduri H, Singh R, Koul A, Ghildiyal M, Sharma K, Tyagi AK, Singh Y: **Nucleoside diphosphate kinase of Mycobacterium tuberculosis acts as GTPase-activating protein for Rho-GTPases**. *FEBS letters* 2004, **571**(1-3):212–216.

[272] Soupene E, Kuypers FA: **Mammalian long-chain acyl-CoA synthetases**. *Experimental biology and medicine* 2008, **233**(5):507–521.

[273] Nys Y, Bain M, Van Immerseel F: *Improving the safety and quality of eggs and egg products: volume 1: egg chemistry, production and consumption*. Elsevier 2011.

[274] Li H, Wang T, Xu C, Wang D, Ren J, Li Y, Tian Y, Wang Y, Jiao Y, Kang X, et al.: **Transcriptome profile of liver at different physiological stages reveals potential mode for lipid metabolism in laying hens**. *BMC Genomics* 2015, **16**:1–13.

[275] Yu S, Wei W, Xia M, Jiang Z, He D, Li Z, Han H, Chu W, Liu H, Chen J: **Molecular characterization, alternative splicing and expression analysis of ACSF 2 and its correlation with egg-laying performance in geese**. *Animal genetics* 2016, **47**(4):451–462.

[276] Tian W, Zheng H, Yang L, Li H, Tian Y, Wang Y, Lyu S, Brockmann GA, Kang X, Liu X: **Dynamic expression profile, regulatory mechanism and correlation with egg-laying performance of ACSF gene family in chicken (Gallus gallus)**. *Scientific reports* 2018, **8**:1–10.

[277] Lopes-Marques M, Cunha I, Reis-Henriques MA, Santos MM, Castro LFC: **Diversity and history of the long-chain acyl-CoA synthetase (Acsl) gene family in vertebrates**. *BMC evolutionary biology* 2013, **13**:1–12.

[278] Ellis JM, Frahm JL, Li LO, Coleman RA: **Acyl-coenzyme A synthetases in metabolic control**. *Current opinion in lipidology* 2010, **21**(3):212.

[279] Brionne A, Nys Y, Hennequet-Antier C, Gautron J: **Hen uterine gene expression profiling during eggshell formation reveals putative proteins involved in the supply of minerals or in the shell mineralization process**. *BMC Genomics* 2014, **15**:1–17.

[280] Jonchère V, Réhault-Godbert S, Hennequet-Antier C, Cabau C, Sibut V, Cogburn LA, Nys Y, Gautron J: **Gene expression profiling to identify eggshell proteins involved in physical defense of the chicken egg**. *BMC Genomics* 2010, **11**:1–19.

[281] Yung LS, Yang C, Wan X, Yu W: **GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies**. *Bioinformatics* 2011, **27**(9):1309–1310.

[282] Hemani G, Theocharidis A, Wei W, Haley C: **EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards**. *Bioinformatics* 2011, **27**(11):1462–1465.

[283] Zhu S, Fang G: **MatrixEpistasis: ultrafast, exhaustive epistasis scan for quantitative traits with covariate adjustment**. *Bioinformatics* 2018, **34**(14):2341–2348.

[284] Chatelain C, Durand G, Thuillier V, Augé F: **Performance of epistasis detection methods in semi-simulated GWAS**. *BMC bioinformatics* 2018, **19**:1–17.

[285] Jing PJ, Shen HB: **MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies**. *Bioinformatics* 2015, **31**(5):634–641.

[286] Kim KH, Kim JY, Lim WJ, Jeong S, Lee HY, Cho Y, Moon JK, Kim N: **Genome-wide association and epistatic interactions of flowering time in soybean cultivar**. *PloS one* 2020, **15**:e0228114.

[287] Cui ZJ, Yang QY, Zhang HY, Zhu Q, Zhang QY: **Bioinformatics identification of drug resistance-associated gene pairs in Mycobacterium tuberculosis**. *International journal of molecular sciences* 2016, **17**(9):1417.

[288] Shen J, Li Z, Song Z, Chen J, Shi Y: **Genome-wide two-locus interaction analysis identifies multiple epistatic SNP pairs that confer risk of prostate cancer: A cross-population study**. *International journal of cancer* 2017, **140**(9):2075–2084.

[289] Egli T, Vukojevic V, Sengstag T, Jacquot M, Cabezón R, Coynel D, Freytag V, Heck A, Vogler C, Dominique JF, et al.: **Exhaustive search for epistatic effects on the human methylome**. *Scientific reports* 2017, **7**:1–10.

[290] Conway JR, Lex A, Gehlenborg N: **UpSetR: an R package for the visualization of intersecting sets and their properties**. *Bioinformatics* 2017.

[291] Rojano E, Seoane P, Ranea JA, Perkins JR: **Regulatory variants: from detection to predicting impact**. *Briefings in bioinformatics* 2019, **20**(5):1639–1654.

[292] Günther T, Schmitt AO, Bortfeldt RH, Hinney A, Hebebrand J, Brockmann GA: **Where in the genome are significant single nucleotide polymorphisms from genome-wide association studies located?** *Omics: a journal of integrative biology* 2011, **15**(7-8):507–512.

[293] Dionisio A, Menezes R, Mendes DA: **Mutual information: a measure of dependency for nonlinear time series**. *Physica A: Statistical Mechanics and its Applications* 2004, **344**(1-2):326–329.

[294] Zhang X, Zhao XM, He K, Lu L, Cao Y, Liu J, Hao JK, Liu ZP, Chen L: **Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information**. *Bioinformatics* 2012, **28**:98–104.

[295] Guo X, Zhang H, Tian T: **Development of stock correlation networks using mutual information and financial big data**. *PloS one* 2018, **13**(4):e0195941.

[296] Mohammadi S, Desai V, Karimipour H: **Multivariate mutual information-based feature selection for cyber intrusion detection**. In *2018 IEEE electrical power and energy Conference (EPEC)*, IEEE 2018:1–6.

[297] Vergara JR, Estévez PA: **A review of feature selection methods based on mutual information**. *Neural computing and applications* 2014, **24**:175–186.

[298] Wu J, Devlin B, Ringquist S, Trucco M, Roeder K: **Screen and clean: a tool for identifying interactions in genome-wide association studies**. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2010, **34**(3):275–285.

[299] Wang D, El-Basyoni IS, Baenziger PS, Crossa J, Eskridge K, Dweikat I: **Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations**. *Heredity* 2012, **109**(5):313–319.

[300] Slim L, Chatelain C, Azencott CA, Vert JP: **Novel methods for epistasis detection in genome-wide association studies**. *PloS one* 2020, **15**(11):e0242927.

[301] Pensar J, Puranen S, Arnold B, MacAlasdair N, Kuronen J, Tonkin-Hill G, Pesonen M, Xu Y, Sipola A, Sánchez-Busó L, et al.: **Genome-wide epistasis and co-selection study using mutual information**. *Nucleic acids research* 2019, **47**(18):e112–e112.

[302] Chen X, Konukoglu E: **Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders**. *arXiv preprint arXiv:1806.04972* 2018.

A. Appendix

A.1. Identification of Regulatory SNPs Associated with Vicine and Convicine Content of *Vicia faba* Based on Genotyping by Sequencing Data Using Deep Learning

genes

MDPI

Article

Identification of Regulatory SNPs Associated with Vicine and Convicine Content of *Vicia faba* Based on Genotyping by Sequencing Data Using Deep Learning

Felix Heinrich [1], Martin Wutke [1], Pronaya Prosun Das [1], Miriam Kamp [1], Mehmet Gültas [1,2], Wolfgang Link [3] and Armin Otto Schmitt [1,2,*]

1 Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; felix.heinrich@uni-goettingen.de (F.H.); martin.wutke@uni-goettingen.de (M.W.); pronayaprosun.das@stud.uni-goettingen.de (P.P.D.); miriam-kamp@gmx.net (M.K.); gueltas@cs.uni-goettingen.de (M.G.)

2 Center for Integrated Breeding Research (CiBreed), Albrecht-Thaer-Weg 3, Georg-August University, 37075 Göttingen, Germany

3 Department of Crop Sciences, Georg-August University, Von-Siebold-Str. 8, 37075 Göttingen, Germany; wlink@gwdg.de

* Correspondence: armin.schmitt@uni-goettingen.de

Received: 29 April 2020; Accepted: 28 May 2020; Published: 5 June 2020

Abstract: Faba bean (*Vicia faba*) is a grain legume, which is globally grown for both human consumption as well as feed for livestock. Despite its agro-ecological importance the usage of *Vicia faba* is severely hampered by its anti-nutritive seed-compounds vicine and convicine (V+C). The genes responsible for a low V+C content have not yet been identified. In this study, we aim to computationally identify regulatory SNPs (rSNPs), i.e., SNPs in promoter regions of genes that are deemed to govern the V+C content of *Vicia faba*. For this purpose we first trained a deep learning model with the gene annotations of seven related species of the Leguminosae family. Applying our model, we predicted putative promoters in a partial genome of *Vicia faba* that we assembled from genotyping-by-sequencing (GBS) data. Exploiting the synteny between *Medicago truncatula* and *Vicia faba*, we identified two rSNPs which are statistically significantly associated with V+C content. In particular, the allele substitutions regarding these rSNPs result in dramatic changes of the binding sites of the transcription factors (TFs) MYB4, MYB61, and SQUA. The knowledge about TFs and their rSNPs may enhance our understanding of the regulatory programs controlling V+C content of *Vicia faba* and could provide new hypotheses for future breeding programs.

Keywords: promoter; rSNP; convolutional neural network; *Vicia faba*; GBS; vicin/convicin

1. Introduction

New methods in the field of genome sequencing—commonly summarized as next generation sequencing (NGS)—offer cost-effective strategies to produce massive amounts of sequencing data. One of these methods is genotyping-by-sequencing (GBS), which is an efficient method to obtain genome-wide genotype data for any species [1]. The characteristic feature of GBS is the reproducible generation of short genomic fragments using known restriction enzymes. Thanks to its easy applicability, GBS is currently the method of choice in the field of plant sciences since it makes plants without reference genome amenable to genomic analysis. Several groups have applied GBS to obtain high-quality genome-wide SNP markers. These markers have often been used for applications

like GWAS, marker-assisted selection, breeding value estimation in genomic prediction, analysis of high density genetic maps, or assessment of population dynamics in plant genomics and plant breeding [2–10]. Another remarkable feature of GBS has been highlighted by Elshire et al. [11], namely that it made the analysis of regulatory regions of genes such as promoters feasible. Despite the growing interest in the analysis of GBS sequenced reads, this capacity of GBS has been poorly studied. This sequencing approach is particularly important for crop species which still often lack a reference genome sequence such as *Vicia faba*. The capacity of GBS to generate sequences of regulatory regions could be used for the prediction of such regions, e.g., promoters, in *Vicia faba*.

The faba bean is an Old World grain legume, which is grown both for combine harvested feed and as vegetable crop for human consumption. It is diploid with $2x = 12$ very large chromosomes. Due to its large size of 13 Gbp [12], there is thus far no sequenced and annotated reference genome available for this plant. Despite its agro-ecological importance (N-symbiosis, rotation hygiene, and pollinator support) [13] it is a crop of limited importance in many countries. This is mainly caused by its anti-nutritive seed-compounds vicine and convicine, which are co-occurring pyrimidine glycosides (in the following termed V+C) and have negative effects to animals such as laying hens, broilers and piglets, but also to 400 million humans suffering from G6PD deficiency [14,15]. The V+C content is a factor that severely limits the wider usage of *Vicia faba* as feed for animals and food for humans. Breeding V+C-poor varieties and production and marketing of their fruits could have a range of positive effects including e.g., reduction of environmentally critical soya bean imports into Europe and Northern America, fostering of regional production methods, and avoidance of energy intensive transports. To date, the location of the gene controlling the V+C content could only be restricted to a region on chromosome 1 of *Vicia faba* that exhibits conserved synteny with chromosome 2 of the related species *Medicago truncatula* between the *Medicago truncatula* genes Medtr2g008210 and Medtr2g010180 [14,16].

The promoter of a gene is the region immediately around its transcription start site (TSS) as well as further upstream of it. The promoter contains multiple elements that allow the binding of the RNA polymerase II (Pol II) along with transcription factors (TFs), thus controlling the transcription of the associated gene. Due to their impact on the gene regulation the SNPs located in the promoter regions that affect the transcription factor binding sites (TFBSs) are commonly called regulatory SNPs (rSNPs). Today it is well known that these rSNPs may be causal for the phenotype and could therefore possibly provide prime candidates useful for breeding programs or marker-assisted selection [17–22]. Despite the rich literature on the analysis of promoters, their prediction remains a challenging task due to their complex and diverse structure. Until now, different machine learning approaches have been developed, which form the core of most computational prediction methods for promoter regions. Whereas in early works the emphasis was on the identification of specific promoter elements (such as TATA boxes, initiator elements (Inrs), downstream promoter elements (DPE) and others) or extraction of k-mer distributions [23–30], nowadays a more holistic approach is given preference in that whole genomic regions are examined in Convolutional Neural Networks (CNNs), which have been successfully applied in many species [31–36].

A large scale genome-wide key study has been conducted by Kumari et al. for the prediction and analysis of core promoter elements (CPEs) across plant monocots and dicots [37]. For this purpose, CPEs of four monocots and four dicots were comprehensively analyzed and compared to establish the common as well as the specific properties of CPEs in promoter sequences. The results obtained in [37] are, on the one hand, promising to enhance the limited knowledge available about the differences between dicots and monocots with respect to their CPEs. On the other hand, they contributed to gain novel insight into the plant promoter sequence architectures and showed that some promoter signatures are strongly conserved within larger groups of plants like monocots or dicots. Based on these findings, Shahmuradov et al. developed a model, namely TSSPlant, for the prediction of plant Pol II promoters across species boundaries [38].

In line with the studies of Kumari et al. [37] and Shahmuradov et al. [38] we designed an analysis workflow for the prediction of promoter sequences as well as rSNPs of *Vicia faba* in this study. For this purpose, we first trained a CNN model using the known promoter sequences of seven plants of the Leguminosae family. Second, using GBS sequence reads of 20 *Vicia faba* lines with known V+C content, we assembled a *de novo* draft partial genome. Thereafter, we called the genomic variants by aligning the GBS reads to the partial genome to obtain high quality SNPs for candidate gene association studies. Next, applying our CNN model to the partial genome sequences, we have predicted the potential promoter sequences of *Vicia faba*. Finally, we analyzed the SNPs in these promoter sequences that were associated with the V+C content of *Vicia faba* regarding their effect on the binding affinity of TFs. Our results show that 2.46% of the assembled sequences were predicted to be promoters. We found 14 regulatory SNPs that could be mapped to the syntenic *Medicago truncatula* region harbouring the major gene for low V+C content [14,16]. These findings could be of use to increase our understanding of the regulation of the V+C content and could provide novel genomic targets for future breeding strategies of V+C poor *Vicia faba* varieties.

2. Materials and Methods

2.1. Plant Material and Sequencing

In total, 20 inbred lines of *Vicia faba* were selected of which six had low V+C content and 14 had high V+C content (see Supplementary Table S5). The lines were inbred via single-seed descent from cultivars, from a gene-bank accession, from biparental crosses or from a landrace and include winter and spring types. DNA was extracted from the grains of the plants. Two pooled grains were used per line. DNA extraction was done with LGC's sbeadex livestock kit following the lysis protocol L for plant tissue. Genotyping-by-sequencing was carried out on the Illumina NextSeq 500 V2 platform. The DNA was digested with the restriction enzyme MslI (recognition sequence: CAYNN^NNRTG). Per sample ~3 million 150 bp paired end reads were obtained. Then sequencing adapter remnants were clipped and reads whose 5′ ends did not match the restriction enzyme site were discarded. The sequencing and filtering was performed by LGC Genomics GmbH (Berlin, Germany).

2.2. Assembly of a Partial Vicia faba Genome

Following the *de novo assembly* strategies used in [39,40], we applied the *de novo* assembler Trinity [41] to the GBS sequence reads for the construction of a partial genome for *Vicia faba*. In total, 694,605 contigs with an average length of 236 bp were constructed. To filter out redundant contigs, we clustered the contigs with `CD-HIT` [42] using a threshold of 95.0% for sequence identity.

2.3. Variant Calling and Association Testing

Following the variant calling pipeline outlined in [43], we mapped the sequence reads onto the partial genome using `Bowtie2` [44]. The variant calling was done with `SAMtools mpileup` [45]. We excluded structural variants such as insertions and deletions as well as non-biallelic SNPs yielding 1,880,592 SNPs. Low quality SNPs with a quality score of lower than 400 were excluded using `PLINK 1.9` [46]. The association between candidate SNPs and the V+C content was tested with `PLINK` using a 1df chi-squared allelic test. To control the type I error rate we set the false discovery rate (FDR) to 0.1.

2.4. Data Sets for Training the Neural Network

Mainly considering the members of the Leguminosae family, we used in our analysis seven species (*Glycine max*, *Lupinus angustifolius*, *Medicago truncatula*, *Phaseolus vulgaris*, *Trifolium pratense*, *Vigna angularis*, and *Vigna radiata*) that have a complete and annotated reference genome sequence available. To further establish the cross-species promoter prediction performance of our CNN model with a more distant plant, we also chose to include in our analysis the model species *Arabidopsis thaliana*. Following [31,33], we extracted for each species their core promoter sequences covering

the −200 bp to +50 bp regions relative to the transcription start sites (TSSs) of protein coding genes from the Ensembl Plants database (release 45) [47] using BioMart [48]. Simultaneously, the sequences covering [TSS+751,TSS+1000] from the core gene region of the genes were extracted, as non-promoter sequences. Sequences that were not assigned to a chromosome or which contained ambiguous bases were not considered.

Currently, due to the absence of an annotated reference genome, there is only scarce knowledge about the promoter sequence architecture in *Vicia faba*. Hence, it is still challenging to determine characteristic signatures of *Vicia faba* promoters that distinguish them from non-promoter regions. To eliminate this lack of knowledge to some extent and to enhance the distinction of promoters vs. non-promoters in *Vicia faba*, the consideration of additional non-promoter sets is important. Consequently, we included two further sets of sequences of length 250 bp as non-promoters in our analysis. While the first set was randomly extracted from the *Medicago truncatula* reference genome by excluding the region [TSS-1000,TSS+500], the second set was sampled from the *Vicia faba* reference transcriptome V2 which was downloaded from the Pulse Crop Database [49]. The final number of sequences for each data set can be found in Table 1.

Table 1. Number of promoter and non-promoter sequences in the sets that were used as training sets.

Species	# Promoter Sequences	# Non-Promoter Sequences
Arabidopsis thaliana	23,315	23,315
Glycine max	46,199	46,199
Lupinus angustifolius	23,463	23,463
Medicago truncatula	32,158	32,158
Phaseolus vulgaris	22,750	22,750
Trifolium pratense	14,749	14,749
Vigna angularis	19,584	19,584
Vigna radiata	15,495	15,495
Medicago truncatula (Genome-wide)	-	∼12,000
Vicia faba (Transcriptome)	-	∼60,000

2.5. Sequence Features Used to Predict Promoters

In line with previous studies [33,38], we used a variety of additional features to characterize promoter sequences as precisely as possible. We have determined the distribution of the following features for the sequence sets listed in Table 1:

Feature 1: Frequency of the dinucleotides CA and CG

Feature 2: Frequency of the TATA motif

Feature 3: CG-skew of sequences ($CG_{\text{skew}} = \frac{\#C-\#G}{\#C+\#G}$ where $\#C$ and $\#G$ refer to the counts of nucleotides C and G in the sequences)

Feature 4: Frequency of k-mers using different values for k

Information theory based features: we included in our analysis two additional features, namely the Horizontal Mutual Information (HMI) and the Generalized Topological Entropy (GTE).

The HMI is calculated based on a predefined distance d between two positions in a sequence and provides a measure of auto-covariation between the nucleotides of interest [50].

$$\text{HMI}(d) = \sum_{m=\{A,C,G,T\}} \sum_{n=\{A,C,G,T\}} p_{mn}(d) \cdot \log \frac{p_{mn}(d)}{p_m(d)p_n(d)}, \quad (1)$$

where $p_m(d)$, $p_n(d)$ and $p_{mn}(d)$ refer to the marginal and joint probabilities of the nucleotides being d bp apart. A high value of HMI(d) indicates a strong correlation between the nucleotides regarding their distance d.

Entropy is a measure to reflect the complexity of sequences. It has been used to characterize the randomness of DNA sequences [51]. More specific varieties of entropies such as the GTE have been successfully applied in [52,53] to explore and to compare the complexity of introns, exons,

and promoter regions. Based on the findings of these studies, we included GTE as an additional feature in our analysis which could provide an important information.

Let ω be a DNA sequence of length $|\omega|$ and let n_ω be the unique integer such that $4^n + n - 1 \leq |\omega| < 4^{n+1} + (n+1) + 1$. Then the GTE is defined as

$$H^k_{n_\omega}(\omega) = \frac{1}{k} \sum_{i=n_\omega-k+1}^{n_\omega} \frac{\log_4(p_\omega(i))}{i} \tag{2}$$

where $p_\omega(i)$ refers to the number of unique sub-sequences of length i that appear in ω. We set $k = n_\omega$ to consider sub-sequences of all possible lengths.

2.6. Convolutional Neural Networks

Our proposed model follows a CNN architecture, which is nowadays one of the most popular neural network architectures [54]. Using convolutional layers as its core elements, a CNN is able to automatically learn local as well as global features from the data layer-wise by applying a convolution operation and by encoding specific aspects of the data [55–57]. Within a layer, an array of stacked weight matrices of dimension $W \times H \times D$, where W, H, and D correspond to the width, height, and depth of the array, respectively, is moved spatially across the input data [31,34]. At every possible position, the summed element-wise product between the weight matrices and a subset of the input is calculated and a corresponding feature map is computed.

The structure of the network that was used is illustrated in Figure 1. The input of the network is formed by a sequence of nucleotides of length 250 bp where each nucleotide is encoded into a one-hot representation and expressed by a four-dimensional vector, with A encoded as $(1,0,0,0)$, C as $(0,1,0,0)$, G as $(0,0,1,0)$, and T as $(0,0,0,1)$. As can be seen in Figure 1, the network is composed of four 1D-convolutional layers followed by a flattening layer, two fully-connected layers and an output layer. All convolutional layers are implemented using a ReLU activation, a stride parameter of 2, zero-padding and a filter size of 21. The first layer uses 64 filters, whereas the second, third and fourth layers use 128, 256, and 512 filters, respectively. To avoid overfitting the training data, a dropout layer with rate $= 0.2$ is used after each convolution [58]. After processing of the sequences by the convolutional layers, a flattening layer transforms the output to a one-dimensional vector and passes its values to two consecutive fully-connected layers with 128 and 64 neurons, respectively. Finally, an output layer with a sigmoid activation classifies the input sequence as promoter or non-promoter. Additional features were included one-by-one by concatenating their values to the flattening layer in order to explore their effect on the improvement of the classification performance.

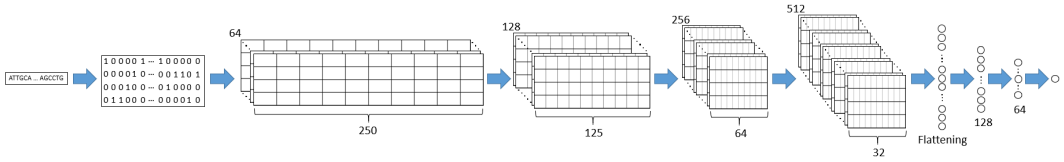


Figure 1. The network architecture of the CNN promoter prediction consists of four 1D-convolutional layers followed by a flattening layer and two fully-connected layers. At the end, an output layer with one neuron and a sigmoid activation function computes the probability that the analyzed sequence is classified as a promoter sequence.

Before training the final model, a separate network for each species was trained individually using the Adam optimizer [59], L2-regularization and binary cross-entropy loss [60]. For each network, 90% of the sequences were used for model training and 10% for testing. The CNN was implemented in R using Keras [61] with TensorFlow [62] as a backend.

To assess the prediction performance we identified the number of correctly predicted promoter and non-promoter sequences as True Positives (TP) and True Negatives (TN), as well as the number of

true promoter sequences predicted as non-promoter sequences, False Negatives (FN), and the number of true non-promoter sequences predicted as promoter sequences, False Positives (FP). From these measures, we calculated Accuracy (ACC), Sensitivity (true positive rate), Specificity (true negative rate), and the Matthews Correlation Coefficient (MCC) as below [33,34,63]:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

2.7. Identification of Putative Regulatory SNPs

In order to identify regulatory SNPs (rSNPs), we analyzed the predicted promoter sequences of *Vicia faba*. For this purpose, we first selected all SNPs that are located in promoters and that we could successfully map against the *Medicago truncatula* genome from the initial 685,215 SNPs (see Section 2.3). Second, we extracted for each SNP its flanking sequence covering ±25 bp relative to the position of the SNP. Third, two copies of the extracted sequences were created: while the first sequence contained at the SNP position the reference allele, the second contained the alternate allele. Thereafter, we identified putative transcription factor binding sites (TFBSs) by applying the MATCH™ program [64] together with a non-redundant plant position weight matrix (PWM) library obtained from the TRANSFAC database [65] to the flanking sequences of each SNP. The MATCH™ program provides for each putative TFBS a matrix similarity score (MSS) ranging from zero to one, which reflects the potential binding affinity of the related TF to it. Finally, we predicted the consequence of each SNP on the TFBS by comparing their MSSs in the two sequences. As a result we observed in our analysis four different types of consequences: (i) no effect, (ii) change in binding affinity, (iii) loss of TFBS (a TFBS appears only for the reference allele) and (iv) gain of TFBS (a TFBS appears only for the alternate allele). Two TFBSs are considered as identical if their PWMs, positions and their strands are equal for both alleles. If the scores computed by MATCH™ are identical in both alleles, the SNP is assumed to have no effect on the TFBS. In our further analysis, we define a SNP as a rSNP, if it has an effect on the binding affinity of at least one TF, i.e., if its type of consequence is (ii), (iii), or (iv).

3. Results and Discussion

Classical application of GBS includes the identification and genotyping of large numbers of genomic variants. This provides several possibilities in plant breeding like the discovery of important markers by GWAS even in the absence of the reference genome. In this study, however, we focused on another important property of the GBS approach, namely its capacity to access regulatory regions (especially promoters) which serves as a basis for the identification of rSNPs in *Vicia faba*.

3.1. Processing the GBS Data

Sequencing of the 20 *Vicia faba* samples yielded 51 GB of GBS data. The *de novo assembly* and filtering resulted in a partial genome consisting of 419,390 contigs with a total length of 100,037,292 bp. Considering that the proposed size of the *Vicia faba* genome is about 13 Gbp [12] our partial genome covered 0.77% of the total genome. Through remapping of the reads to the partial genome with `Bowtie2` and subsequent variant calling with `SAMtools` 1,880,592 SNPs could be derived. The quality scores of these SNPs as given in the vcf file showed a clear bimodal distribution with a minimum

at a quality score of 400. 1,195,377 SNPs having a quality score of lower than 400 were discarded, such that 685,215 high quality SNPs remained.

3.2. Prediction of Promoter Sequences

3.2.1. Intra- and Inter-Species Promoter Prediction

In order to gain first insights into the predictability of promoters of the seven Leguminosae family members and *Arabidopsis thaliana*, we trained our CNN model for each species individually. The prediction reliability of the CNN model has been examined for each species by classifying the intra- and inter-species promoters that were not used in the training process. To assess the performance of the classification, the ACC, Sensitivity, Specificity and MCC values were calculated. The details based on the ACC values are presented in Table 2 and the results based on the remaining measures are given in the Supplementary Tables S1–S3.

Table 2. ACC values of the intra- and inter-species promoter classification using the species-specific trained CNNs. Off-diagonal numbers are ACC values for inter-species classification, diagonal numbers are ACC values for intra-species classification. For instance, a CNN trained on *Lupinus angustifolius* and used for classification of *Vigna angularis* promoters has an accuracy of 0.974.

Evaluated / Trained	*Arabidopsis thaliana*	*Glycine max*	*Lupinus angustifolius*	*Medicago truncatula*	*Phaseolus vulgaris*	*Trifolium pratense*	*Vigna angularis*	*Vigna radiata*
Arabidopsis thaliana	0.901	0.767	0.690	0.746	0.797	0.765	0.633	0.733
Glycine max	0.837	0.864	0.915	0.847	0.863	0.724	0.914	0.856
Lupinus angustifolius	0.545	0.611	0.981	0.720	0.586	0.493	0.974	0.709
Medicago truncatula	0.755	0.797	0.959	0.876	0.789	0.715	0.951	0.841
Phaseolus vulgaris	0.845	0.842	0.888	0.834	0.898	0.748	0.880	0.853
Trifolium pratense	0.822	0.764	0.696	0.751	0.794	0.840	0.689	0.736
Vigna angularis	0.510	0.607	0.971	0.715	0.583	0.494	0.977	0.712
Vigna radiata	0.741	0.812	0.937	0.827	0.825	0.675	0.928	0.904

The results presented in Table 2 show that although the CNN models have been trained only using one-hot representation of sequences for each species individually, the network architecture is able to recognize certain patterns in the sequences which leads to the predictability of promoters across different species to a certain degree. These findings support the results presented in [37] and indicate that some of the promoter signatures seem to be conserved between the *Leguminosae* family members.

Further, Table 2 demonstrates that the classification performance of some CNN models remarkably results in higher ACC values for inter-species prediction than for intra-species prediction. In particular, this is the case for the species *Lupinus angustifolius* and *Vigna angularis* whose promoters have been predicted with very high accuracy by the CNN models of the other species except for the *Arabidopsis thaliana* and *Trifolium pratense* models. This could be attributed to the underlying genome annotations of these species since their annotations seem to be created based on the genome information of well-studied family members. Especially, this assumption is true regarding the inclusion of the Leguminosae family specific promoter patterns in the promoters of these species. This hypothesis has been supported by the prediction performance of the *Lupinus angustifolius* model on the other species, which reached very high degrees of specificity while only achieving low degrees of sensitivity. To this end, we compared the inter-species prediction ACC values of the species regarding their performance on *Lupinus angustifolius* and *Vigna angularis*. This comparison revealed that the promoters of *Lupinus*

angustifolius were predicted with a slightly higher mean accuracy value than the promoters of *Vigna angularis* (0.880 compared to 0.868).

So far, we trained our CNN model only using the order of the nucleotides in DNA sequences (one-hot encoding). However, previous studies pointed out that the combination of one-hot encoding with additional widely used features could lead to a substantially improved performance in promoter identification [33,34]. For this purpose, we followed a similar procedure as suggested by Triska et al. in [33] and systematically evaluated the combination of sequence features (see Section 2.5) in the CNN model training of each species. The results are presented only for *Medicago truncatula* as an example in Table 3.

Table 3. Contribution of additional features in the CNN model of *Medicago truncatula*.

Features	Accuracy	Sensitivity	Specificity	MCC
DNA sequence	0.876	0.897	0.855	0.750
DNA sequence + 2-mer	0.874	0.880	0.867	0.747
DNA sequence + 2-mer + frequency of CA motif	0.862	0.828	0.897	0.726
DNA sequence + 2-mer + frequency of CG motif	0.875	0.875	0.876	0.751
DNA sequence + 2-mer + HMI	0.874	0.882	0.865	0.747
DNA sequence + 2-mer + frequency of TATAA motif	0.876	0.875	0.878	0.752
DNA sequence + 2-mer + CG skew	0.876	0.889	0.863	0.753
DNA sequence + topological entropy	0.874	0.886	0.861	0.747
DNA sequence + 2-mer + topological entropy	0.871	0.852	0.890	0.743
DNA sequence + 2-mer + HMI + frequency of TATAA motif	0.871	0.869	0.874	0.743
DNA sequence + 2-mer + HMI + frequency of CA motif + frequency of CG motif + frequency of TATAA motif + CC skew	0.873	0.859	0.888	0.747
DNA sequence + HMI + frequency of CA motif + frequency of CG motif + frequency of TATAA motif + CG skew	0.875	0.889	0.860	0.749

In contrast to previous studies [33,34], Table 3 shows that regardless of the usage of any additional feature, the performance of the CNN model could in general not be significantly improved. However, these results are in agreement with findings presented in [31,32,35,36] and indicate that the CNN architecture is able to learn specific patterns inherent in the sequences automatically. Hence, these patterns carry information which is obviously redundant to these widely used features. Consequently, it turns out that the consideration of additional features does not lead to an improvement of the CNN model performance and may, on the contrary, increase the noise during training.

3.2.2. Prediction of *Vicia faba* Promoters

The knowledge about the promoter signatures which are conserved between the *Leguminosae* family members provides an important clue for the precise prediction of *Vicia faba* promoters, which still remains a challenge. However, the consideration of the sequences of only one *Leguminosae* family member in the CNN model could be insufficient to capture the variety of different promoter signatures for the accurate computational identification of the *Vicia faba* promoters. To mitigate the drawback of single species models we systematically examined different CNN models seeking to determine the preferential combination of *Leguminosae* family members by intensifying the signal of promoter sequences and thus to improve the performance of the CNN model. Consequently, we trained a CNN model based on the species *Lupinus angustifolius* and *Medicago truncatula* since the combined usage of their manually selected sequences perfectly complement each other. In the last step, we included in the CNN model training two additional non-promoter sets (defined in Section 2.4) to enhance the

distinction signals between promoter and non-promoter regions. The training sequences are given in the Supplementary Files S6 and S7. The evaluation of this CNN model yields to clearly better ACC and MCC values of 0.98 and 0.95, respectively. A further analysis reveals that the usage of other sequence features together with one-hot encoding in our final CNN model does not affect the performance of the classifier.

Finally, by applying the CNN model to the *Vicia faba* sequences of length 250 bp, we classified in total 2.46% of them as potential promoter sequences. It is important to note that, due to the random fragment orientation regarding the direction of the reads from GBS, the correct direction of the sequences in the *de novo* assembly draft partial genome of *Vicia faba* is unknown. To address this limitation, we considered in our predictions four different types of the sequences as: (i) the original obtained assembly; (ii) the complement of the obtained assembly that is gained by keeping the reading direction; (iii) the reverse of the obtained assembly that is gained by changing the reading direction; and (iv) the reverse complement of the obtained assembly.

Checking the positions of the SNPs in the contigs disclosed that in total 132,399 out of 685,215 SNPs were located in the predicted *Vicia faba* promoters. A flanking sequence of ± 25 bp could only be obtained for 118,492 SNPs. These SNPs with a complete flanking sequence were mapped against the *Medicago truncatula* genome using the BLASTN algorithm with a threshold of 0.01 for the *e-value* and of 0.9 for the *percent identity* [66]. Overall, we found 33,846 hits for 1976 SNPs showcasing the repetitiveness of the *Medicago truncatula* genome. We identified 14 SNPs that map to the predefined target region of *Medicago truncatula* that harbours orthologous genes associated with the V+C content of *Vicia faba* [14,16]. This target region is ranging approximately from 1,300,000 bp to 2,300,000 bp of the *Medicago truncatula* chromosome 2. An overview of these SNPs and their mapped position in the *Medicago truncatula* genome along with the genes with the closest TSS is given in Table 4. We tested these 14 SNPs for their association with the V+C content with `PLINK`. The adjusted *p*-values presented in Table 4 suggest that SNP_341016_236 and SNP_341016_239, which are located in the same promoter, show a highly significant association with the V+C content in *Vicia faba* while the associations of the remaining SNPs are not significant at the level $\alpha = 0.05$. For both of these SNPs the reference allele only occurs in the low V+C lines with one exception while the alternate allele is restricted to the high V+C lines (see Supplementary Table S8).

Table 4. The 14 SNPs found in the predicted promoters of *Vicia faba* that were mapped to the *Medicago truncatula* target genomic region.

SNP_ID	Genotype	FDR	Position	Medicago Gene
SNP_131938_118	C/T	0.234	1,385,390	MTR_2g008290
SNP_302904_183	G/A	0.179	1,385,444	
SNP_341016_236	C/T	$1.17 \cdot 10^{-7}$	1,554,857	MTR_2g008620
SNP_341016_239	G/A	$1.17 \cdot 10^{-7}$	1,554,860	
SNP_356745_200	A/G	0.730	1,707,078	MTR_2g008960
SNP_280549_41	C/T	0.234	1,707,183	
SNP_350273_103	G/T	0.234	1,707,199	
SNP_350273_90	A/C	0.234	1,707,212	
SNP_350273_61	G/A	0.234	1,912,704	MTR_2g009430
SNP_29452_204	G/A	0.730	1,912,812	
SNP_29452_206	G/A	0.496	1,912,814	
SNP_118828_190	C/T	0.234	2,030,017	MTR_2g009690
SNP_80231_27	C/T	0.234	2,163,048	MTR_2g009940
SNP_364434_97	A/T	0.359	2,163,084	

Genotype refers to the reference and alternative alleles; FDR is the false discovery rate obtained in an association test with the V+C content of 20 Vicia faba lines; Position is the position in bp on the Medicago truncatula chromosome 2.

3.2.3. Systematic Identification of Regulatory SNPs Associated with the V+C Content of *Vicia faba*

Following the studies of Xu et al. and Fu et al. in [67,68], we scanned the flanking sequences of the SNPs by applying the MATCH^{TM} program [64] to systematically identify the SNPs that are likely to affect the binding affinity of transcription factors (TFs) and, thus, influence the gene expression level. This search was done for the 1976 SNPs that were located in the predicted *Vicia faba* promoters and which could be successfully mapped onto the *Medicago truncatula* genome. We considered results of this run with an MSS score $\geq$0.85 as putative TFBSs (as suggested in [69]). 9444 putative TFBSs were identified. SNPs that were located in putative TFBSs were considered as rSNPs. Their consequence types were determined by examining their predicted effects on the binding affinities of the TFs. The rSNPs, their consequence and related TFs with corresponding PWM names are given in Supplementary Table S4. The analysis of the 14 SNPs presented in Table 4 reveals that the binding affinities of 44 TFs to their 79 TFBSs were affected. Focusing on the two highly significant SNPs (SNP_341016_236 and SNP_341016_239) in the same promoter, we found that a nucleotide substitution in SNP_341016_236 is likely to entail severe consequences regarding TF binding affinity, namely loss and gain of TFBSs.

The substitution in SNP_341016_239 results in only a moderate change of the binding affinities of TFs (see Table 5). The remarkably different consequences of both SNPs indicate their considerably different influence for the precise and effective regulation of the corresponding gene, although their *p*-values are the same.

Table 5. The two SNPs with the strongest association to the V+C content and their consequences. The column **Allele** indicates for which allele of the SNP the binding site was found. **TFBS** refers to the name of the binding sites, which were named after their PWMs. The structure of the PWM names is given as: P$TFname_version, where "P$" stands for the PWMs used for the prediction of the TFBSs of plant TFs. "TFname" refers to the name of the transcription factor, and "_version" refers to the version of the PWM.

SNP_ID	Allele	TFBS	MSS	Consequence
SNP_341016_236	Ref	P$MYB4_01	0.945	Loss of TFBS
SNP_341016_236	Ref	P$MYB61_01	0.880	Loss of TFBS
SNP_341016_236	Alt	P$SQUA_01	0.870	Gain of TFBS
SNP_341016_239	Ref	P$MYB61_01	0.880	Score change
SNP_341016_239	Alt	P$MYB61_01	0.881	Score change

3.3. Functional Analysis of the Candidate Gene and Transcription Factors

The *Medicago truncatula* gene MTR_2g008620 is the gene which is located closest to the two highly significant SNPs. It is a beta-hydroxyacyl-ACP-dehydratase that is involved in the elongation of fatty acids as well as in the related metabolism of biotin [70,71]. A direct association with the creation of V+C which has been linked to the orotic acid pathway [72] is not obvious. This seems plausible since *Medicago truncatula* does not synthesize V+C. Of more interest are the transcription factors for which we found putative binding sites that are affected by the two SNPs. The TF SQUA belongs to the MADS-box domain group whose genes play vital roles in multiple aspects of plant development (for instance development of flowers, fruits and roots as well as regulating flowering time) [73,74]. Such genes regulate, for example, stem growth and early flowering in soybean [75] or vernalization response in wheat [76]. SQUA itself is involved in the determination of floral meristem and organ identity [77,78]. The MYB domain group is one of the largest families of TFs in plants. Its members are involved in the regulation of development, metabolism, the circadian rhythm, and responses to biotic and abiotic stresses in plants [74,79,80]. In *Medicago truncatula* multiple MYB TFs including MYB4 and MYB61 are involved in flavonoid biosynthesis during macrosclereid cell development [81]. MYB4 in particular regulates abiotic stress responses towards UV-B light and cadmium toxicity in *Arabidopsis*

thaliana [82,83] and cold in *Oryza sativa* [84]. It has also been shown to influence the biosynthesis of flavonoids [85]. MYB61 participates in the response to cold stress in *Medicago truncatula* [86]. In *Arabidopsis thaliana*, this TF is expressed in sink tissues, such as xylem, roots and developing seeds, and controls resource allocation influencing growth and development of the plant [87]. It has also been shown to affect trichome initiation, root development and stomatal aperture and it is necessary for the biosynthesis of gibberellin [88,89]. Furthermore, it is required for the seed coat mucilage deposition during the development of the seed coat epidermis [90,91]. This is a promising result considering that the seed coat is the suggested site of biosynthesis of the V+C compounds [92].

4. Conclusions

With their anti-nutritive effects the high vicine and convicine content has so far restricted the usage of *Vicia faba* as feed for livestock or as crop for human consumption. Identifying causal markers and understanding the mechanisms of regulation of the V+C content are important steps for breeding new cultivars with lower V+C content. This task is even more challenging since a complete and annotated reference genome for *Vicia faba* is still missing. In this work we harnessed the knowledge about regulatory regions in related species to train a convolutional neural network, which allowed us to predict those regions in *Vicia faba*. This model permitted us to classify DNA sequences as promoter or non-promoter without undertaking the considerable effort of assembling and annotating a reference genome. We applied this model to GBS data of *Vicia faba* for the identification of its putative promoter regions as well as regulatory SNPs therein. Our results show that we were able to detect two rSNPs significantly associated with the V+C production in *Vicia faba*. We suggest these rSNPs as promising candidates for marker-assisted selection. In particular, the associated transcription factor MYB61 could provide new insights into the molecular mechanisms underlying V+C. To the best of our knowledge, this is the first study which uses the gene annotations of related species of the *Leguminosae* family to predict promoters and rSNPs of *Vicia faba* based on the GBS data. The analysis approach that we presented here could potentially also be applied to other species that lack a reference genome which is still the case for many crop species.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/6/614/s1, Table S1: Sensitivity values of the interspecies promoter classification using the intraspecies trained CNNs. Table S2: Specificity values of the interspecies promoter classification using the intraspecies trained CNNs. Table S3: MCC values of the interspecies promoter classification using the intraspecies trained CNNs. Table S4: Results of the TFBS analysis. Table S5: Overview of the 20 *Vicia faba* lines used in the analysis. File S6: Promoter sequences used for training the model. File S7: Non-promoter sequences used for training the model. Table S8: Alleles of the 20 *Vicia faba* lines for the two significantly associated rSNPs.

Author Contributions: M.G. designed and supervised the research. F.H. participated in the design of the study, prepared the data sets, conducted the bioinformatics analysis and developed the model together with M.W., P.P.D., and M.G. F.H. and M.K. performed the functional analysis of gene and transcription factors. A.O.S. and W.L. secured the funding for data acquisition. W.L. provided seed of the inbred lines, expertise with the crop plant and contributed to the training strategy. F.H., M.W., M.K., M.G. and A.O.S. interpreted the results and wrote the final version of the manuscript. M.G. and A.O.S. supervised the writing of the manuscript, conceived as well as managed the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Lower Saxony Ministry of Science and Culture, grant number MWK 11-76251-99-30/16.

Acknowledgments: We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the University of Göttingen. We are grateful to Rebecca Tacke, Thomas Lange, and Wolfgang Ecke for providing valuable advice on some biological aspects. We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SNP	Single Nucleotide Polymorphism
rSNP	Regulatory Single Nucleotide Polymorphism
CNN	Convolutional Neural Network
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
V+C	Vicin and Convicine
FDR	False Discovery Rate
HMI	Horizontal Mutual Information
GTE	Generalized Topological Entropy
CPE	Core Promoter Element
PWM	Position Weight Matrix
MSS	Matrix Similarity Score

References

1. Deschamps, S.; Llaca, V.; May, G.D. Genotyping-by-Sequencing in Plants. *Biology* **2012**, *1*, 460–483. [CrossRef] [PubMed]
2. Muktar, M.S.; Teshome, A.; Hanson, J.; Negawo, A.T.; Habte, E.; Entfellner, J.D.; Lee, K.; Jones, C.S. Genotyping by sequencing provides new insights into the diversity of Napier grass (Cenchrus purpureus) and reveals variation in genome-wide LD patterns between collections. *Sci. Rep.* **2019**, *9*, 6936. [CrossRef] [PubMed]
3. Raman, H.; Raman, R.; Nelson, M.N.; Aslam, M.N.; Rajasekaran, R.; Wratten, N.; Cowling, W.A.; Kilian, A.; Sharpe, A.G.; Schondelmaier, J. Diversity array technology markers: genetic diversity analyses and linkage map construction in rapeseed (Brassica napus L.). *DNA Res.* **2011**, *19*, 51–65. [CrossRef] [PubMed]
4. Wenzl, P.; Raman, H.; Wang, J.; Zhou, M.; Huttner, E.; Kilian, A. A DArT platform for quantitative bulked segregant analysis. *BMC Genom.* **2007**, *8*, 196. [CrossRef]
5. He, J.; Zhao, X.; Laroche, A.; Lu, Z.; Liu, H.; Li, Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* **2014**, *5*, 484. [CrossRef]
6. Nguyen, N.H.; Premachandra, H.K.A.; Kilian, A.; Knibb, W. Genomic prediction using DArT-Seq technology for yellowtail kingfish Seriola lalandi. *BMC Genom.* **2018**, *19*, 107. [CrossRef]
7. Von Mark, V.C.; Kilian, A.; Dierig, D.A. Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. *PLoS ONE* **2013**, *8*, e64062.
8. Morris, G.P.; Ramu, P.; Deshpande, S.P.; Hash, C.T.; Shah, T.; Upadhyaya, H.D.; Riera-Lizarazu, O.; Brown, P.J.; Acharya, C.B.; Mitchell, S.E.; et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 453–458. [CrossRef]
9. International Cassava Genetic Map Consortium. High-resolution linkage map and chromosome-scale genome assembly for cassava (Manihot esculenta Crantz) from 10 populations. *G3 Genes Genomes Genet.* **2015**, *5*, 133–144.
10. Soto, J.C.; Ortiz, J.F.; Perlaza-Jiménez, L.; Vásquez, A.X.; Lopez-Lavalle, L.A.B.; Mathew, B.; Léon, J.; Bernal, A.J.; Ballvora, A.; López, C.E. A genetic map of cassava (Manihot esculenta Crantz) with integrated physical mapping of immunity-related genes. *BMC Genom.* **2015**, *16*, 190. [CrossRef]
11. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef] [PubMed]
12. Cooper, J.W.; Wilson, M.H.; Derks, M.F.L.; Smit, S.; Kunert, K.J.; Cullis, C.; Foyer, C.H. Enhancing faba bean (Vicia faba L.) genome resources. *J. Exp. Bot.* **2017**, *68*, 1941–1953. [CrossRef] [PubMed]
13. Köpke, U.; Nemecek, T. Ecological services of faba bean. *Field Crop. Res.* **2010**, *115*, 217–233. [CrossRef]

14. Khazaei, H.; Purves, R.W.; Hughes, J.; Link, W.; O'Sullivan, D.M.; Schulman, A.H.; Björnsdotter, E.; Geu-Flores, F.; Nadzieja, M.; Andersen, S.U.; et al. Eliminating vicine and convicine, the main anti-nutritional factors restricting faba bean usage. *Trends Food Sci. Technol.* **2019**, *91*, 549–556. [CrossRef]
15. Arese, P.; Gallo, V.; Pantaleo, A.; Turrini, F. Life and Death of Glucose-6-Phosphate Dehydrogenase (G6PD) Deficient Erythrocytes - Role of Redox Stress and Band 3 Modifications. *Transfus. Med. Hemotherapy* **2012**, *39*, 328–334. [CrossRef] [PubMed]
16. Duc, G.; Sixdenier, G.; Lila, M.; Furstoss, V. Search of Genetic Variability for Vicine and Convicine Content in *Vicia faba* L.: A First Report of a Gene Which Codes for Nearly Zero-Vicine and Zero-Convicine Contents. In *Recent Advances of Research in Antinutritional Factors in Legume Seeds*; Huisman, J., van der Poel, A.F.B., Liener, I.E., Eds.; Wageningen Academic Publishers: Wageningen, The Netherlands, 1989; pp. 305–313.
17. Fang, L.; Ahn, J.K.; Wodziak, D.; Sibley, E. The human lactase persistence-associated SNP -13910*T enables in vivo functional persistence of lactase promoter-reporter transgene expression. *Hum. Genet.* **2012**, *131*, 1153–1159. [CrossRef]
18. De Gobbi, M.; Viprakasit, V.; Hughes, J.R.; Fisher, C.; Buckle, V.J.; Ayyub, H.; Gibbons, R.J.; Vernimmen, D.; Yoshinaga, Y.; de Jong, P.; et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **2006**, *312*, 1215–1217. [CrossRef] [PubMed]
19. Ordovás, L.; Roy, R.; Pampín, S.; Zaragoza, P.; Osta, R.; Rodríguez-Rey, J.C.; Rodellar, C. The g.763G>C SNP of the bovine FASN gene affects its promoter activity via Sp-mediated regulation: implications for the bovine lactating mammary gland. *Physiol. Genom.* **2008**, *34*, 144–148. [CrossRef] [PubMed]
20. Ryan, M.T.; Hamill, R.M.; O'Halloran, A.M.; Davey, G.C.; McBryan, J.; Mullen, A.M.; McGee, C.; Gispert, M.; Southwood, O.I.; Sweeney, T. SNP variation in the promoter of the PRKAG3 gene and association with meat quality traits in pig. *BMC Genet.* **2012**, *13*, 66. [CrossRef]
21. Barkova, O.Y.; Sazanova, K.A.; Fomichev, K.A.; Malewski, T.; Parada, R.; Kawka, M.; Jaszczak, K.; Sazanov, A.A. Associations of new rSNPs with eggshell thickness in Rhode Island layers. *Anim. Sci. Pap. Rep.* **2013**, *31*, 165–172.
22. Konishi, S.; Izawa, T.; Lin, S.Y.; Ebana, K.; Fukuta, Y.; Sasaki, T.; Yano, M. An SNP caused loss of seed shattering during rice domestication. *Science* **2006**, *312*, 1392–1396. [CrossRef]
23. Fickett, J.W.; Hatzigeorgiou, A.G. Eukaryotic Promoter Recognition. *Genome Res.* **1997**, *7*, 861–878. [CrossRef]
24. Shahmuradov, I.A.; Solovyev, V.V.; Gammerman, A.J. Plant promoter prediction with confidence estimation. *Nucleic Acids Res.* **2005**, *33*, 1069–1076. [CrossRef]
25. Ohler, U. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res.* **2006**, *34*, 5943–5950. [CrossRef]
26. Morey, C.; Mookherjee, S.; Rajasekaran, G.; Bansal, M. DNA Free Energy-Based Promoter Prediction and Comparative Analysis of Arabidopsis and Rice Genomes. *Plant Physiol.* **2011**, *156*, 1300–1315. [CrossRef]
27. Azad, A.K.M.; Shahid, S.; Noman, N.; Lee, H. Prediction of plant promoters based on hexamers and random triplet pair analysis. *Algorithms Mol. Biol.* **2011**, *6*, 19. [CrossRef]
28. Lai, H.; Zhang, Z.; Su, Z.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: A Computational Predictor for Predicting Promoter. *Mol. Ther. Nucleic Acids* **2019**, *17*, 337–346. [CrossRef]
29. Abeel, T.; Saeys, Y.; Rouzé, P.; Van de Peer, Y. ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics* **2008**, *24*, i24–i31. [CrossRef]
30. Anwar, F.; Baker, S.M.; Jabid, T.; Mehedi Hasan, M.; Shoyaib, M.; Khan, H.; Walshe, R. Pol II promoter prediction using characteristic 4-mer motifs: A machine learning approach. *BMC Bioinform.* **2008**, *9*, 414. [CrossRef]
31. Umarov, R.K.; Solovyev, V.V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* **2017**, *12*, e171410. [CrossRef]
32. Umarov, R.; Kuwahara, H.; Li, Y.; Gao, X.; Solovyev, V. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics* **2019**, *35*, 2730–2737. [CrossRef]
33. Triska, M.; Solovyev, V.; Baranova, A.; Kel, A.; Tatarinova, T.V. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PLoS ONE* **2017**, *12*, 1–28. [CrossRef]
34. Qian, Y.; Zhang, Y.; Guo, B.; Ye, S.; Wu, Y.; Zhang, J. An Improved Promoter Recognition Model Using Convolutional Neural Network. In Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 23–27 July 2018; Volume 1, pp. 471–476. [CrossRef]

35. Oubounyt, M.; Louadi, Z.; Tayara, H.; Chong, K.T. DeePromoter: Robust Promoter Predictor Using Deep Learning. *Front. Genet.* **2019**, *10*, 286. [CrossRef]
36. Pachganov, S.; Murtazalieva, K.; Zarubin, A.; Sokolov, D.; Chartier, D.R.; Tatarinova, T.V. TransPrise: A novel machine learning approach for eukaryotic promoter prediction. *PeerJ* **2019**, *7*, e7990. [CrossRef]
37. Kumari, S.; Ware, D. Genome-Wide Computational Prediction and Analysis of Core Promoter Elements across Plant Monocots and Dicots. *PLoS ONE* **2013**, *8*, e79011. [CrossRef]
38. Shahmuradov, I.A.; Umarov, R.K.; Solovyev, V.V. TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic Acids Res.* **2017**, *45*, e65. [CrossRef]
39. Goubert, C.; Modolo, L.; Vieira, C.; ValienteMoro, C.; Mavingui, P.; Boulesteix, M. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biol. Evol.* **2015**, *7*, 1192–1205. [CrossRef]
40. Yuan, S.; Xia, Y.; Zheng, Y.; Zeng, X. Next-generation sequencing of mixed genomic DNA allows efficient assembly of rearranged mitochondrial genomes in *Amolops chunganensis* and *Quasipaa boulengeri*. *PeerJ* **2016**, *4*, e2786. [CrossRef]
41. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
42. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
43. Hwang, S.; Kim, E.; Lee, I.; Marcotte, E.M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **2015**, *5*, 17875. [CrossRef] [PubMed]
44. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
45. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
46. Chang, C.C.; Chow, C.C.; Tellier, L.CAM.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **2015**, *4*. [CrossRef]
47. Howe, K.L.; Contreras-Moreira, B.; De Silva, N.; Maslen, G.; Akanni, W.; Allen, J.; Alvarez-Jarreta, J.; Barba, M.; Bolser, D.M.; Cambell, L.; et al. Ensembl Genomes 2020—Enabling non-vertebrate genomic research. *Nucleic Acids Res.* **2019**, gkz890. [CrossRef] [PubMed]
48. Kinsella, R.J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011**, *2011*, bar030. [CrossRef]
49. Humann, J.L.; Jung, S.; Cheng, C-H.; Lee, T.; Zheng, P.; Frank, M.; McGaughey, D.; Scott, K.; Buble, K.; Yu, J.; et al. Cool Season Food Legume Genome Database: A resource for pea, lentil, faba bean and chickpea genetics, genomics and breeding. In Proceedings of the International Plant and Animal Genome Conference, San Diego, CA, USA, 12–16 January 2019.
50. Lichtenstein, F.; Antoneli, F.; Briones, M.R.S. MIA: Mutual Information Analyzer, a graphic user interface program that calculates entropy, vertical and horizontal mutual information of molecular sequence sets. *BMC Bioinform.* **2015**, *16*, 409. [CrossRef]
51. Schmitt, A.O.; Herzel, H. Estimating the entropy of DNA sequences. *J. Theor. Biol.* **1997**, *188*, 369–377. [CrossRef]
52. Jin, S.; Tan, R.; Jiang, Q.; Xu, L.; Peng, J.; Wang, Y.; Wang, Y. A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences. *PLoS ONE* **2014**, *9*, e88519. [CrossRef]
53. Li, J.; Zhang, L.; Li, H.; Ping, Y.; Xu, Q.; Wang, R.; Tan, R.; Wang, Z.; Liu, B.; Wang, Y. Integrated entropy-based approach for analyzing exons and introns in DNA sequences. *BMC Bioinform.* **2019**, *20*, 283. [CrossRef]
54. Al-Ajlan, A.; El Allali, A. CNN-MGP: Convolutional neural networks for metagenomics gene prediction. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11*, 628–635. [CrossRef] [PubMed]
55. Chollet, F.; Allaire, J.J. *Deep Learning with R*; Manning Publications: Shelter Island, NY, USA, 2018.

56. Li, Q.; Cai, W.; Wang, X.; Zhou, Y.; Feng, D.D.; Chen, M. Medical image classification with convolutional neural network. In Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 10–12 December 2014; pp. 844–848.
57. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
58. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
59. Kingma, D.P; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:cs.LG/1412.6980.
60. Yu, K.; Xu, W.; Gong, Y. Deep learning with kernel regularization for visual recognition. In *Advances in Neural Information Processing Systems*; Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2009; pp. 1889–1896.
61. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 28 May 2020).
62. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org (accessed on 28 May 2020).
63. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678. [CrossRef]
64. Kel, A.E.; Gößling, E.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**, *31*, 3576–3579. [CrossRef]
65. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **2008**, *9*, 326–332. [CrossRef]
66. Camacho, C.; Coulouris, G.A.V.M.N.P.J.B.K.M.T. BLAST+: architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
67. Xu, Z.; Taylor, J.A. SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* **2009**, *37*, W600–W605. [CrossRef]
68. Fu, Y.; Liu, Z.; Lou, S.; Bedford, J.; Mu, X.J.; Yip, K.Y.; Khurana, E.; Gerstein, M. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **2014**, *15*, 480. [CrossRef]
69. Gearing, L.J.; Cumming, H.E.; Chapman, R.; Finkel, A.M.; Woodhouse, I.B.; Luu, K.; Gould, J.A.; Forster, S.C.; Hertzog, P.J. CiiiDER: A tool for predicting and analysing transcription factor binding sites. *PLoS ONE* **2019**, *14*, e0215495. [CrossRef]
70. Heath, R.J.; Rock, C.O. Roles of the FabA and FabZ β-Hydroxyacyl-Acyl Carrier Protein Dehydratases in Escherichia coli Fatty Acid Biosynthesis. *J. Biol. Chem.* **1996**, *271*, 27795–27801. [CrossRef] [PubMed]
71. Lin, S.; Hanson, R.E.; Cronan, J.E. Biotin synthesis begins by hijacking the fatty acid synthetic pathway. *Nat. Chem. Biol.* **2010**, *6*, 682–688. [CrossRef] [PubMed]
72. Brown, E.G.; Roberts, F.M. Formation of vicine and convicine by *Vicia faba*. *Phytochemistry* **1972**, *11*, 3203–3206. [CrossRef]
73. Smaczniak, C.; Immink, R.G.H.; Angenent, G.C.; Kaufmann, K. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* **2012**, *139*, 3081–3098. [CrossRef]
74. Riechmann, J.L.; Ratcliffe, O.J. A genomic perspective on plant transcription factors. *Curr. Opin. Plant Biol.* **2000**, *3*, 423–434. [CrossRef]
75. Ping, J.; Liu, Y.; Sun, L.; Zhao, M.; Li, Y.; She, M.; Sui, Y.; Lin, F.; Liu, X.; Tang, Z.; et al. Dt2 Is a Gain-of-Function MADS-Domain Factor Gene That Specifies Semideterminacy in Soybean. *Plant Cell* **2014**, *26*, 2831–2842. [CrossRef]
76. Danyluk, J.; Kane, N.A.; Breton, G.; Limin, A.E.; Fowler, D.B.; Sarhan, F. TaVRT-1, a Putative Transcription Factor Associated with Vegetative to Reproductive Transition in Cereals. *Plant Physiol.* **2003**, *132*, 1849–1860. [CrossRef]
77. West, A.G.; Sharrocks, A.D.; Causier, B.E.; Davies, B. DNA binding and dimerisation determinants of Antirrhinum majus MADS-box transcription factors. *Nucleic Acids Res.* **1998**, *26*, 5277–5287. [CrossRef]
78. Theißen, G.; Melzer, R.; Rümpler, F. MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development* **2016**, *143*, 3259–3271. [CrossRef]
79. Dubos, C.; Stracke, R.; Grotewold, E.; Weisshaar, B.; Martin, C.; Lepiniec, L. MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* **2010**, *15*, 573–581. [CrossRef]

80. Roy, S. Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome. *Plant Signal. Behav.* **2016**, *11*, e1117723. [CrossRef] [PubMed]
81. Fu, F.; Zhang, W.; Li, Y.; Wang, H.L. Establishment of the model system between phytochemicals and gene expression profiles in Macrosclereid cells of Medicago truncatula. *Sci. Rep.* **2017**, *7*, 2580. [CrossRef] [PubMed]
82. Jin, H.; Cominelli, E.; Bailey, P.; Parr, A.; Mehrtens, F.; Jones, J.; Tonelli, C.; Weisshaar, B.; Martin, C. Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. *EMBO J.* **2000**, *19*, 6150–6161. [CrossRef] [PubMed]
83. Agarwal, P.; Banerjee, S.; Mitra, M.; Roy, S. MYB4 transcription factor, A member of R2R3-type MYB family protein regulates Cd tolerance via activation of antioxidant defense and glutathione (GSH) dependent pathway in Arabidopsis thaliana. In Proceedings of the XIV International Geographical Union (IGU)-India Conference, Burdwan, India, 6–8 March 2020.
84. Vannini, C.; Locatelli, F.; Bracale, M.; Magnani, E.; Marsoni, M.; Osnato, M.; Mattana, M.; Baldoni, E.; Coraggio, I. Overexpression of the rice Osmyb4 gene increases chilling and freezing tolerance of Arabidopsis thaliana plants. *Plant J.* **2004**, *37*, 115–127. [CrossRef] [PubMed]
85. Wang, X.; Wu, J.; Guan, M.; Zhao, C.; Geng, P.; Zhao, Q. Arabidopsis MYB4 plays dual roles in flavonoid biosynthesis. *Plant J.* **2020**, *101*, 637–652. [CrossRef]
86. Zhang, Z.; Hu, X.; Zhang, Y.; Miao, Z.; Xie, C.; Meng, X.; Deng, J.; Wen, J.; Mysore, K.S.; Frugier, F.; et al. Opposing Control by Transcription Factors MYB61 and MYB3 Increases Freezing Tolerance by Relieving C-Repeat Binding Factor Suppression. *Plant Physiol.* **2016**, *172*, 1306–1323. [CrossRef]
87. Romano, J.M.; Dubos, C.; Prouse, M.B.; Wilkins, O.; Hong, H.; Poole, M.; Kang, K.; Li, E.; Douglas, C.J.; Western, T.L.; et al. AtMYB61, an R2R3-MYB transcription factor, functions as a pleiotropic regulator via a small gene network. *New Phytol.* **2012**, *195*, 774–786. [CrossRef]
88. Matías-Hernández, L.; Jiang, W.; Yang, K.; Tang, K.; Brodelius, P.E.; Pelaz, S. AaMYB1 and its orthologue AtMYB61 affect terpene metabolism and trichome development in Artemisia annua and Arabidopsis thaliana. *Plant J.* **2017**, *90*, 520–534. [CrossRef]
89. Liang, Y.; Dubos, C.; Dodd, I.C.; Holroyd, G.H.; Hetherington, A.M.; Campbell, M.M. AtMYB61, an R2R3-MYB Transcription Factor Controlling Stomatal Aperture in Arabidopsis thaliana. *Curr. Biol.* **2005**, *15*, 1201–1206. [CrossRef]
90. Arsovski, A.A.; Villota, M.M.; Rowland, O.; Subramaniam, R.; Western, T.L. MUM ENHANCERS are important for seed coat mucilage production and mucilage secretory cell differentiation in Arabidopsis thaliana. *J. Exp. Bot.* **2009**, *60*, 2601–2612. [CrossRef]
91. Penfield, S.; Meissner, R.C.; Shoue, D.A.; Carpita, N.C.; Bevan, M.W. MYB61 Is Required for Mucilage Deposition and Extrusion in the Arabidopsis Seed Coat. *Plant Cell* **2001**, *13*, 2777–2791. [CrossRef] [PubMed]
92. Ramsay, G.; Griffiths, D.W. Accumulation of vicine and convicine in *Vicia faba* and *V. narbonensis*. *Phytochemistry* **1996**, *42*, 63–67. [CrossRef]

A.2. Genotyping by Sequencing Reads of 20 *Vicia faba* Lines with High and Low Vicine and Convicine Content

data

MDPI

Data Descriptor

Genotyping by Sequencing Reads of 20 *Vicia faba* Lines with High and Low Vicine and Convicine Content

Felix Heinrich [1], Mehmet Gültas [1,2], Wolfgang Link [3] and Armin Otto Schmitt [1,2,*]

[1] Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; felix.heinrich@uni-goettingen.de (F.H.); gueltas@cs.uni-goettingen.de (M.G.)
[2] Center for Integrated Breeding Research (CiBreed), Albrecht-Thaer-Weg 3, Georg-August University, 37075 Göttingen, Germany
[3] Department of Crop Sciences, Georg-August University, Von-Siebold-Str. 8, 37075 Göttingen, Germany; wlink@gwdg.de
* Correspondence: armin.schmitt@uni-goettingen.de

Received: 26 June 2020; Accepted: 16 July 2020; Published: 20 July 2020

Abstract: The grain faba bean (*Vicia faba*) which belongs to the family of the Leguminosae, is a crop that is grown worldwide for consumption by humans and livestock. Despite being a rich source of plant-based protein and various agro-ecological advantages its usage is limited due to its anti-nutrients in the form of the seed-compounds vicine and convicine (V+C). While markers for a low V+C content exist the underlying pathway and the responsible genes have remained unknown for a long time and only recently a possible pathway and enzyme were found. Genetic research into *Vicia faba* is difficult due to the lack of a reference genome and the near exclusivity of V+C to the species. Here, we present sequence reads obtained through genotyping-by-sequencing of 20 *Vicia faba* lines with varying V+C contents. For each line, ∼3 million 150 bp paired end reads are available. This data can be useful in the genomic research of *Vicia faba* in general and its V+C content in particular.

Dataset: The reads have been submitted to the European Nucleotide Archive (ENA) under the accession PRJEB38838.

Keywords: *Vicia faba*; GBS; vicine/convicine

1. Summary

The Protein Crop Strategy of the German Federal Ministry of Food and Agriculture has the goal of raising the importance of domestic protein crops e.g., legumes in Germany and Europe in order to improve ecosystem services and resource conservation as well as to reduce the dependency on imported crops [1]. The faba bean (*Vicia faba*) is a prime candidate for this strategy being a globally grown legume that has several agro-ecological advantages (N-symbiosis, rotation hygiene, and pollinator support) and serving as food for humans and livestock [2]. Regardless of these benefits its usage is limited due to the anti-nutrients vicine and convicine (V+C) that occur in their seeds. These compounds have negative effects to animals as well as to humans suffering from G6PD deficiency [3,4]. Despite ongoing research efforts and the discovery of a robust marker for the V+C content, the responsible genes and mechanisms remained unknown for a long time and the location of

the locus could only be restricted to an interval on chromosome 1 of *Vicia faba* that shows conserved synteny with a region on chromosome 2 of the related species *Medicago truncatula* that is about 900,000 bp long [3]. Research has been exacerbated due to the lack of an annotated reference genome for *Vicia faba*, which is assumed to be about 13 Gbp [5]. In a recent preprint the authors have found an enzyme associated with V+C biosynthesis and identified it as a guanosine triphosphate (GTP) cyclohydrolase II, proposing the purine GTP as a precursor for vicine [6]. The breeding of novel low V+C varieties to improve the usage of *Vicia faba* as feed is the goal of the project Abo-Vici, which is supported by the German Federal Ministry of Food and Agriculture [7]. As part of this project we obtained reads from 20 *Vicia faba* lines with known V+C content through genotyping-by-sequencing (GBS). We offer this data here for the benefit of researches into *Vicia faba* and its V+C content in particular. We have so far successfully used this data for the prediction of regulatory regions in *Vicia faba* and the identification of regulatory single nucleotide polymorphisms (SNPs) that are associated with V+C content [8]. For this we built a partial genome for *Vicia faba* from the GBS reads that spanned ∼1 % of the total genome and performed variant calling with it, which resulted in more than 600,000 high quality SNPs. This partial genome is available upon request from the corresponding author. Finally, while the data itself are not enough alone, the data can support the eventual creation of an annotated reference genome for *Vicia faba*.

2. Data Description

The sequence reads for 20 *Vicia faba* lines obtained through GBS are stored as paired end reads in two FASTQ files per *Vicia faba* line. These FASTQ files contain both the nucleotide sequence and its corresponding quality scores as text. Per sample ∼3 million 150 bp paired end reads are available, such that the total amount of sequence amounted to 18 Gbp. The uncompressed data required 51 GB of disk space. The sequences have been deposited at the European Nucleotide Archive (ENA) under the accession number PRJEB38838.

3. Methods

Plant Material and Sequencing

We obtained GBS data from 20 inbred lines of *Vicia faba*. The lines were inbred via single-seed descent from cultivars, from a gene-bank accession, from biparental crosses or from a landrace and include winter and spring types (see Table 1 for more information). Six of the lines had a low V+C content and 14 had high V+C content. DNA extraction, sequencing and filtering were carried out by LGC Genomics GmbH (Berlin, Germany). The DNA was extracted via LGC's sbeadex livestock kit following the lysis protocol L for plant tissue from the grains of the plants. From each line two pooled grains were used. An extraction using the sbeadex plant kit was tested but provided poorer results than the livestock kit. For each library construction 100–200 ng of genomic DNA were used, which was quantified with a NanoDrop. The DNA was digested with 2 units of the restriction enzyme MslI (NEB, recognition sequence: CAYNN^NNRTG) in NEB4 buffer in 20 µL volume for 2 h at 37 °C. The restriction enzyme was heat inactivated by incubation at 80 °C for 20 min. The TrueSeq adapter sequences used were:

- adapter_prefix_R1 'AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATC'
- adapter_prefix_R2 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAG'

For the ligation 10 µL of each restriction digest were transferred to a new 96well PCR plate, mixed on ice first with 1.5 µL of one of 96 inline-barcoded forward blunt adaptors (pre-hybridized, concentration 5 pmol/µL), followed by addition of 20 µL Ligation master mix (contains: 15 µL NEB Quick ligation buffer, 0.4 µL NEB Quick Ligase, 7.5 pmol pre-hybridized common reverse blunt adaptor). Ligation reactions were incubated for 1 h at room temperature, followed by heat inactivation for 10 min at 65 °C. After the ligation reactions the libraries were purified using

Agencourt XP beads. Following that the libraries were size selected by a size selection on a LMP-agarose gel, removing fragments smaller than 300 bp or larger than 400 bp. For the final quality control, Fragment Analyzer and Qubit were used. The libraries were then amplified in 20 µL PCR reactions using MyTaq (Bioline) and standard Illumina TrueSeq amplification primers. The number of cycles was limited to 14. Having each line uniquely barcoded the samples were pooled and ran on the same sequencing run. An Illumina NextSeq 500 V2 platform was then used for genotyping-by-sequencing. Demultiplexing of the libraries was done using the Illumina `bcl2fastq 2.17.1.14` software. Finally sequencing adapter remnants were clipped using `cutadapt 1.13+18` [9] and reads whose 5′ ends did not match the restriction enzyme site were discarded. As a last step FastQC reports (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were generated for the FASTQ files.

Table 1. Vicine and convicine status of the 20 lines, name, sample ID, additional notes and European Nucleotide Archive (ENA) accession number under which the sample data is available.

ENA Accession	Sample ID	V+C	Line	Notes
ERS4652931	Sample_8	Low	Line 1268-4-1	Ancestor of low V+C content
ERS4652926	Sample_3	Low	Mélodie/2	cv. Mélodie; minor, spring bean
ERS4652927	Sample_4	Low	F7(Mélodie/2 x ILB938/2)-139-1-1	Near isogenic lines (ILB938/2 is from Ecuador)
ERS4652928	Sample_5	Low	F7(Mélodie/2 x ILB938/2)-201-3-1	
ERS4652932	Sample_9	High	F7(Mélodie/2 x ILB938/2)-139-2-1	
ERS4652933	Sample_10	High	F7(Mélodie/2 x ILB938/2)-201-4-1	
ERS4652929	Sample_6	Low	F7[VC.14.8099-843-2-1]	Near isogenic lines from a breeder's cross, spring beans
ERS4652930	Sample_7	Low	F7[VC.14.8099-848-3-1]	
ERS4652934	Sample_11	High	F7[VC.14.8099-843-3-3]	
ERS4652935	Sample_12	High	F7[VC.14.8099-848-4-1]	
ERS4652924	Sample_1	High	HediLin-1	cv. Hedin; minor, spring bean
ERS4652936	Sample_13	High	PietraLin	Major, Mediterranean bean
ERS4652937	Sample_14	High	(HediLin/1 x PietraLin)-2-4	Near isogenic lines
ERS4652938	Sample_15	High	(HediLin/1 x PietraLin)-4-4	
ERS4652939	Sample_16	High	S_281	Academic winter bean lines
ERS4652940	Sample_17	High	S_301	
ERS4652941	Sample_18	High	S_034	
ERS4652942	Sample_19	High	S_290	
ERS4652925	Sample_2	High	Hiverna/2	cv. Hiverna; minor, winter bean
ERS4652943	Sample_20	High	Côte d'Or/1	Côte d'Or; minor, winter bean

Author Contributions: M.G. designed and supervised the research. F.H. participated in the design of the study, prepared the data sets and conducted the bioinformatics analysis together with M.G. A.O.S. and W.L. secured the funding for data acquisition. W.L. provided seed of the inbred lines and expertise with the crop plant. F.H., M.G. and A.O.S. wrote the final version of the manuscript. M.G. and A.O.S. supervised the writing of the manuscript, conceived as well as managed the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Lower Saxony Ministry of Science and Culture, grant number MWK 11-76251-99-30/16.

Acknowledgments: We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the University of Göttingen.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GBS Genotyping by Sequencing
SNP Single Nucleotide Polymorphism
V+C Vicine and Convicine

References

1. Protein Crop Strategy. Available online: https://www.bmel.de/EN/topics/farming/plant-production/protein-crop-strategy.html (accessed on 14 June 2020).
2. Köpke, U.; Nemecek, T. Ecological services of faba bean. *Field Crops Res.* **2010**, *115*, 217–233. [CrossRef]
3. Khazaei, H.; Purves, R.W.; Hughes, J.; Link, W.; O'Sullivan, D.M.; Schulman, A.H.; Björnsdotter, E.; Geu-Flores, F.; Nadzieja, M.; Andersen, S.U.; et al. Eliminating vicine and convicine, the main anti-nutritional factors restricting faba bean usage. *Trends Food Sci. Technol.* **2019**, *91*, 549–556. [CrossRef]
4. Arese, P.; Gallo, V.; Pantaleo, A.; Turrini, F. Life and Death of Glucose-6-Phosphate Dehydrogenase (G6PD) Deficient Erythrocytes—Role of Redox Stress and Band 3 Modifications. *Transfus. Med. Hemother.* **2012**, *39*, 328–334. [CrossRef] [PubMed]
5. Cooper, J.W.; Wilson, M.H.; Derks, M.F.L.; Smit, S.; Kunert, K.J.; Cullis, C.; Foyer, C.H. Enhancing faba bean (*Vicia faba* L.) genome resources. *J. Exp. Bot.* **2017**, *68*, 1941–1953. [CrossRef] [PubMed]
6. Björnsdotter, E.; Nadzieja, M.; Chang, W.; Escobar-Herrera, L.; Mancinotti, D.; Angra, D.; Khazaei, H.; Crocoll, C.; Vandenberg, A.; Stoddard, F.L.; et al. VC1 catalyzes a key step in the biosynthesis of vicine from GTP in faba bean. *bioRxiv* **2020**. [CrossRef]
7. Abo-Vici-Projekt. Available online: https://www.uni-goettingen.de/de/abo-vici-projekt/559637.html (accessed on 14 June 2020).
8. Heinrich, F.; Wutke, M.; Das, P.P.; Kamp, M.; Gültas, M.; Link, W.; Schmitt, A.O. Identification of Regulatory SNPs Associated with Vicine and Convicine Content of Vicia faba Based on Genotyping by Sequencing Data Using Deep Learning. *Genes* **2020**, *11*, 614. [CrossRef] [PubMed]
9. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **2011**, *17*, 10–12. [CrossRef]

A.3. MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes

biology

MDPI

Article

MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes

Felix Heinrich [1,*], Faisal Ramzan [1], Abirami Rajavel [1], Armin Otto Schmitt [1,2] and Mehmet Gültas [2,3,*]

[1] Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; faisal.ramzan@uni-goettingen.de (F.R.); abirami.rajavel@uni-goettingen.de (A.R.); armin.schmitt@uni-goettingen.de (A.O.S.)
[2] Center for Integrated Breeding Research (CiBreed), Albrecht-Thaer-Weg 3, Georg-August University, 37075 Göttingen, Germany
[3] Faculty of Agriculture, South Westphalia University of Applied Sciences, Lübecker Ring 2, 59494 Soest, Germany
* Correspondence: felix.heinrich@uni-goettingen.de (F.H.); gueltas.mehmet@fh-swf.de (M.G.)

Simple Summary: The interactions between SNPs, which are known as epistasis, can strongly influence the phenotype. Their detection is still a challenge, which is made even more difficult through the existence of background associations that can hide correct epistatic interactions. To address the limitations of existing methods, we present in this study our novel method MIDESP for the detection of epistatic SNP pairs. It is the first mutual information-based method that can be applied to both qualitative and quantitative phenotypes and which explicitly accounts for background associations in the dataset.

Abstract: The interactions between SNPs result in a complex interplay with the phenotype, known as epistasis. The knowledge of epistasis is a crucial part of understanding genetic causes of complex traits. However, due to the enormous number of SNP pairs and their complex relationship to the phenotype, identification still remains a challenging problem. Many approaches for the detection of epistasis have been developed using mutual information (MI) as an association measure. However, these methods have mainly been restricted to case–control phenotypes and are therefore of limited applicability for quantitative traits. To overcome this limitation of MI-based methods, here, we present an MI-based novel algorithm, MIDESP, to detect epistasis between SNPs for qualitative as well as quantitative phenotypes. Moreover, by incorporating a dataset-dependent correction technique, we deal with the effect of background associations in a genotypic dataset to separate correct epistatic interaction signals from those of false positive interactions resulting from the effect of single SNP×phenotype associations. To demonstrate the effectiveness of MIDESP, we apply it on two real datasets with qualitative and quantitative phenotypes, respectively. Our results suggest that by eliminating the background associations, MIDESP can identify important genes, which play essential roles for bovine tuberculosis or the egg weight of chickens.

Keywords: mutual information; epistatic interactions; genome-wide association studies; single-nucleotide polymorphism

Citation: Heinrich, F.; Ramzan, F.; Rajavel, A.; Schmitt, A.O.; Gültas, M. MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes. *Biology* **2021**, *10*, 921. https://doi.org/10.3390/biology10090921

Academic Editor: Wenzhong Xiao

Received: 2 August 2021
Accepted: 13 September 2021
Published: 16 September 2021

1. Introduction

The development of high-density arrays for genotyping in recent years has allowed genome-wide association studies (GWAS) to become powerful tools for the detection of single-nucleotide polymorphisms (SNPs) that are associated with traits of interest. However, GWAS methods are usually based on the analysis of single loci, ignoring the potential interaction between genes, and are therefore of limited applicability for traits that are controlled by multiple genes with possibly complex interactions [1–3]. These genes may have only a small effect on the phenotype and could therefore be missed by single-locus

analyses despite having a strong influence based on their interactions [4–7]. While large parts of phenotype variance are attributed to individual SNP effects, these interactions, which are commonly referred to as epistasis, have been shown to be of importance for many complex diseases in humans such as asthma [8], cancer [9] or diabetes [10], as well as for quantitative traits in animals [3,11–15] and plants [16–20], and could help to explain the relationship between the genetic variants and the corresponding phenotype [2,13,21,22].

Due to the large number of possible combinations of SNPs even if only pairwise interactions are considered, the detection of epistasis is a computational challenge, for which a large number of algorithms have been proposed. These methods can be divided into different categories depending on their search strategy. Exhaustive search strategies test every possible combination of SNPs for significance, which often results in a long execution time and can become infeasible for large datasets. This strategy has been used by partitioning methods such as the Combinatorial Partitioning Method (CPM) [23] and the Restricted Partition Method (RPM) [24], as well as several other methods [9,25,26]. Stochastic methods, on the other hand, use random sampling to increase their efficiency, but their results and performance can depend on variables determined by the user. Bayesian Epistasis Association Mapping (BEAM) [27], for instance, applies Markov chain Monte Carlo to compute the posterior probability for association between SNPs and a disease. Its extension epistatic MOdule DEtection (epiMODE) [28] uses Gibbs sampling with a reversible jump Markov chain Monte Carlo to find epistatic interactions. Machine learning methods such as neural networks [29–32], decision trees [33] or random forest [34–37] have also been utilized for epistasis detection. Step-wise approaches form a fourth category of algorithms, which first filter out SNPs with a very small or no association signal, and then test among the surviving SNPs for epistatic interactions. BOolean Operation-based Screening and Testing (BOOST) [38], as an example, first performs a likelihood ratio test to filter out unimportant SNPs and then performs an exhaustive search on the others. Leem et al. [39] utilized a k-means clustering of the SNPs and then searched for interactions between SNPs in different clusters. Other methods still use the results of lower-order interactions to find higher-order interactions in an efficient way [40,41].

Several of these methods use information-theory-based measures such as mutual information to quantify epistatic interactions [39,41–46]. These measures consider the SNPs and phenotypes as random variables, which allows them to quantify the amount of information, or uncertainty, that is inherent to an SNP or a phenotype and to compute how much information is shared between them, and thereby the strength of association [42]. This approach is model-free and therefore has the advantage of not requiring any prior assumptions regarding the structure of the interactions. By considering all genotype combinations of the SNPs as separate categories, this strategy also avoids the problem of choosing an appropriate encoding method for the SNPs and their interactions, which has been shown to influence the result of regression-based methods [47–49]. Nevertheless, the application of information-theory-based approaches has so far been limited to case–control phenotypes. This is because, while the mutual information between two discrete variables can be efficiently calculated using simple contingency tables, the mutual information between a discrete and a continuous variable requires computationally more challenging approaches for an accurate estimation.

Furthermore, the methods mentioned above do not take into account different types of obstacles resulting from sample structure, relatedness between the genotyped individuals or marginal effects of single SNPs on the phenotype [19,50,51]. Such types of obstacles can lead to background associations between SNP pairs and the phenotype, and thus the importance of some SNPs in the epistatic interactions could be overestimated. Consequently, the prediction of most existing methods could be biased, potentially impeding the identification of correct epistatic signals. Hence, elimination of the bias inherent in the genotype–phenotype datasets is needed to separate the signal caused by functional interactions from the background associations between SNPs [19,50].

In this paper, we propose a novel method called Mutual Information-based Detection of Epistatic SNP Pairs (MIDESP) for the detection of pairwise epistatic interactions, which extends the previously mentioned mutual information-based approaches by additionally enabling the identification of epistatic interactions between SNP pairs and quantitative phenotypes. For this purpose we adopt, in the context of epistasis for the first time, the mutual information estimator developed by Ross [52], which accurately estimates the level of epistasis using a *k*th-nearest neighbor-based approach. Moreover, to deal with the possible obstacles inside a genotype–phenotype dataset (as mentioned above), our method incorporates an additional step using the average product correction (APC) theorem [53] to estimate the expected level of background association for each SNP pair. Finally, the removal of the estimated background from the measured epistasis values leads to the detection of correct epistatic signals arising from functional interactions.

In order to demonstrate the performance and functionality of MIDESP, we applied it on two different types of genotype–phenotype datasets. The first type contains several hundred simulated datasets, which we analyzed to optimize the parameters used in the mutual information estimator. On the other hand, the second type contains two further datasets with a qualitative and a quantitative phenotype, respectively. While the dataset with the qualitative phenotype is related to bovine tuberculosis, the other one contains the egg weight of chicken eggs. Our findings show that we are able to successfully reduce the influence of background associations in the prediction of epistatic interactions, which leads to the identification of novel markers/genes that are important to the phenotype of interest.

2. Materials and Methods

2.1. Data

We analyzed two different datasets, one of which had a qualitative (discrete) case–control phenotype, and the other one had a quantitative (continuous) phenotype. To ensure the data quality, we applied several filters to the datasets following Ramzan et al. [54,55]. We removed SNPs with a minor allele frequency ≤ 0.01, a genotyping call rate ≤ 0.97, as well as SNPs significantly deviating from the Hardy–Weinberg equilibrium (p–value $< 1 \times 10^{-6}$). A sample was removed if a phenotype was unavailable for it or if more than 5% of SNPs were missing. Further, we performed linkage disequilibrium (LD) pruning to obtain epistasis results without confounding them through LD [56]. Using PLINK [57], we removed all redundant SNPs with an LD ≥ 0.99, and thus carrying very similar information about the phenotype. Table 1 gives a short overview of the datasets and their respective sizes.

In the following section, we briefly describe the datasets. Researchers interested in more details about the bovine tuberculosis data are referred to [58] and about the egg weight data to [59].

2.1.1. Bovine Tuberculosis (BT)

This dataset was published by Bermingham et al. [58] and consists of 617,885 SNPs for 1151 cattle. The estimated SNP-based heritability attributable to additive effects for this phenotype is 21% [58]. The cattle belonged to the Holstein–Friesian breed and were collected in Northern Ireland. Genotyping was performed using the BovineHD Genotyping BeadChip. The supplied phenotype is qualitative (case–control) and represents the resistance of the animals towards bovine tuberculosis with 592 cases and 559 controls. Bermingham et al. performed a GWAS on the data to find SNPs associated with resistance to bovine tuberculosis. Overall, they found eight significantly associated SNPs representing two different loci in the genome. After applying our filters 616,398 SNPs remained.

2.1.2. Egg Weight (EW)

The dataset relates to the egg weight (EW) in 36-week-old chickens belonging to a line of Rhode Island Red chicken [59]. While the dataset contains the egg weights for multiple different ages of the chickens, we decided to only use the data for 36-week-old chickens,

since this phenotype contains the strongest signal found in previous GWAS [54,59]. For this trait, the estimated SNP-based heritability is 36% [59]. A total of 1063 birds were genotyped using the Affymetrix Axiom® 600 K Chicken Genotyping Array, resulting in an initial set of 580,961 SNPs, which were then filtered. The dataset which was provided by the authors only consists of the 294,705 SNPs that passed their quality filters. We could not remove any further SNPs using our filters.

Table 1. Overview of the datasets used in our study.

Dataset	Phenotype	#Samples	#SNPs	#SNPs after Filtering	#SNPs after LD Pruning
Bovine Tuberculosis	Qualitative	1151	617,885	616,398	358,086
Egg weight	Quantitative	1063	580,961	294,705	139,101

2.2. Method

Based on the number of samples, $\mathcal{N}$, and the number of SNPs, $\mathcal{P}$, we consider a genotype $\times$ phenotype dataset as a matrix, $M_{\mathcal{N}\times(\mathcal{P}+1)}$, where the rows refer to the samples and the columns refer to the phenotype and the SNPs. Furthermore, the phenotype of interest is denoted by Y^D and Y^C for qualitative (discrete) and quantitative (continuous) traits. Let S^i be a sample, let X^j be the genotype of an SNP and let Y^i be the corresponding phenotype of S^i. The entry of M at position (i, j) is depicted by X^i_j. In the following, we also use X and Y as placeholders for any of the SNPs or phenotypes, respectively.

An overview of the MIDESP pipeline is shown in Figure 1.

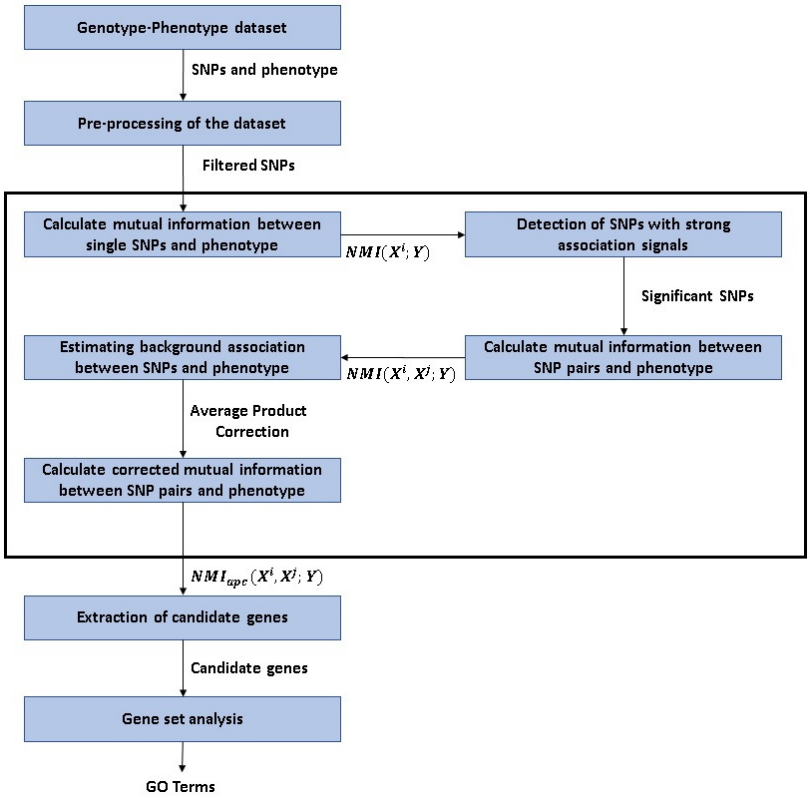


Figure 1. Flowchart of the analysis applied in this study.

2.2.1. Background on Information Theoretic Measures

In information theory, the entropy, $H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$, is a measure for the uncertainty of a discrete random variable, X, with alphabet $\mathcal{X}$, which depends solely on its probability function, $p(x) = Pr\{X = x\}$, $x \in \mathcal{X}$. It can be interpreted as the amount of information that is necessary to describe the variable on average. By considering the joint probability function, $p(x, y)$, of two discrete random variables X and Y with alphabets

$\mathcal{X}$ and $\mathcal{Y}$, this concept can be extended to the joint entropy, $H(X, Y)$, of a pair of variables. Based on these entropies, the mutual information between X and Y is defined as

$$MI(X;Y) = H(X) + H(Y) - H(X,Y), \tag{1}$$

which gives the amount of information that is shared between the variables [60]. The mutual information can be seen as a measure for the association between two variables, which includes linear as well as non-linear dependencies [61].

However, Equation (1) is not applicable if one of the variables is continuous instead of discrete. For a discrete variable X and a continuous variable Y the $MI(X;Y)$ can be estimated using the kth-nearest neighbor-based method developed by Ross [52], which has been shown to be more robust than the commonly used binning-based approaches. The mutual information is estimated as

$$MI(X;Y) = \frac{1}{\mathcal{N}} \cdot \sum_{i=1}^{\mathcal{N}} (\psi(\mathcal{N}) - \psi(\mathcal{N}_{x_i}) + \psi(k) - \psi(m_i)), \tag{2}$$

where:

- $\psi(\cdot)$ is the digamma function;
- $\mathcal{N}_{x_i}$ for a given sample, S^i, refers to the number of samples for which the genotype x is the same as the genotype x_i of S^i;
- d is the distance between sample S^i and its kth-nearest neighbor S^{i_k} with the same genotype as S^i, defined as the absolute difference between their phenotypes Y^i and Y^{i_k};
- m_i is assigned the number of samples where the absolute difference between their phenotypes and the phenotype Y^i is less than or equal to d, irrespective of the genotypes.

The identification of these values is shown for a small exemplary dataset in Figure 2.

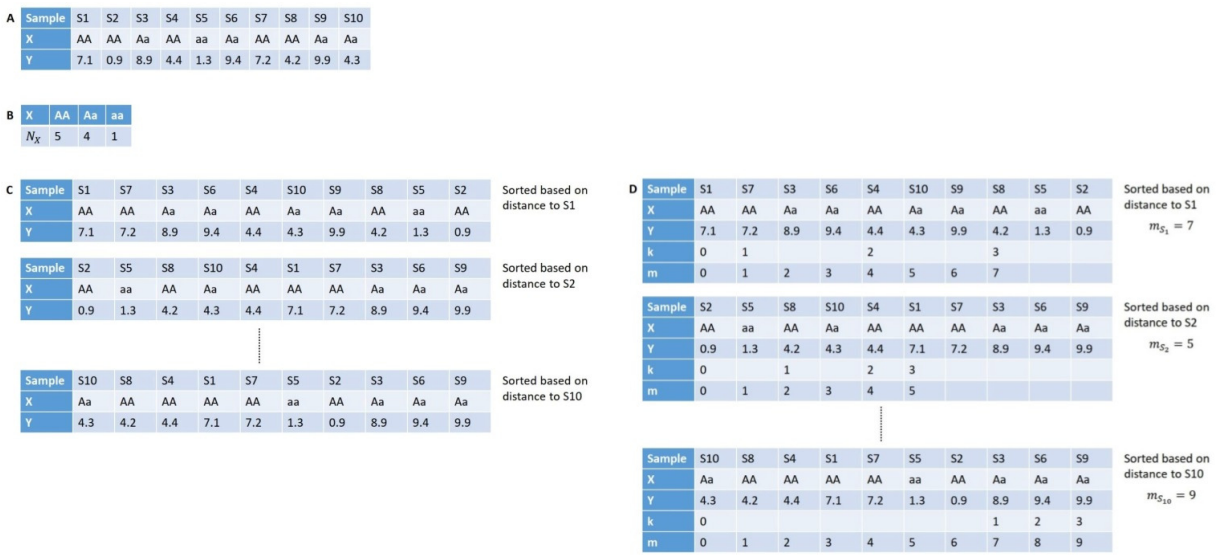


Figure 2. Procedures for estimating the mutual information between a discrete variable X as an SNP and a continuous variable Y representing a quantitative phenotype using $k = 3$: (**A**) Genotype and phenotype values are given for ten samples S1, S2, ... to S10. (**B**) N_x is defined as the number of samples where the genotype is equal to x. (**C**) For each sample, S_i, a sorted list of the samples is created based on the absolute difference between Y_i and Y_j for sample S_j. (**D**) The kth-nearest neighbor is determined for each sorted list by going along the list and counting the samples that have the same X value as the start sample. m_{S_i} can then be defined as the index of the kth-nearest neighbor in the sorted list. For sample S1 which has the X value AA, the sample with the third-closest Y value and the same X value is sample S8, which has the index 7 in the sorted list. Therefore, $m_{S_1} = 7$. Based on the $\mathcal{N}_x$ and m_{S_i} values, the mutual information can be estimated.

As shown in Figure 2, only the phenotype Y is a continuous variable, hence in general, we can reuse the sorted tables for every SNP by only changing the values of X. This allows

for an efficient calculation of m_i. Since MI is only estimated, the resulting values can be outside the range of the valid interval, i.e., $[0, H(X)]$. Thus, the estimated values outside of this range are set to the closest interval boundary.

2.2.2. Identification of Epistatic Interactions between SNP Pairs

In previous studies [39,42,46], the epistatic interaction between an SNP pair, X^i and X^j, and a qualitative phenotype, Y, has been successfully identified by employing the MI metric for which Equation (1) is extended based on the joint entropy $H(X^i, X^j)$ as:

$$MI(X^i, X^j; Y) = H(X^i, X^j) + H(Y) - H(X^i, X^j, Y), \tag{3}$$

where $H(X^i, X^j, Y)$ is the joint entropy of the SNPs X^i and X^j as well as the phenotype Y. However, the concept of MI has not yet been applied to measure the epistatic interaction between an SNP pair and a quantitative phenotype. To the best of our knowledge, we apply for the first time the MI metric for this aim using the following equation:

$$MI(X^i, X^j; Y) = \frac{1}{\mathcal{N}} \cdot \sum_{l=1}^{\mathcal{N}} (\psi(\mathcal{N}) - \psi(\mathcal{N}_{x_l^{ij}}) + \psi(k) - \psi(m_l)) \tag{4}$$

In Equation (4), x_l^{ij} refers to the joint genotype of the SNP pair X^i and X^j of sample S^l.

As shown in [53,62,63], the value of the mutual information and its possible range is strongly dependent on the alphabet size and the marginal distributions of the variables. A normalization of the values is therefore required to address this influence and to make them comparable with each other for further analysis. We apply the following normalization technique based on the entropy and the maximal possible alphabet size of the SNP and SNP pair. Consequently, the $MI(X; Y)$—and $MI(X^i, X^j; Y)$—values are normalized as

$$NMI(...; Y) = 2 \cdot \frac{MI(...; Y)}{\log(\max|\mathcal{X}|) + H(...)} \tag{5}$$

2.2.3. Detection of SNPs with Strong Association Signals

As it can be easily seen, the calculation of the pairwise interactions between all SNP pairs requires a quadratic runtime. Therefore, the separation of SNPs with strong association signals from the remaining ones is necessary to reduce the number of pairs under study.

For this purpose, Gültas et al. [63,64] showed that by extending the standard multiple testing theory [65,66], the NMI values can be modeled based on three different distributions: (i) a β distribution F_0 (null distribution) representing the background signals; (ii) a G_1 distribution referring to the unrelated associations (in our case between SNPs and phenotype); (iii) a G_2 distribution modeling the strong association signals (in our case between SNPs and phenotype).

From this follows that $1 - F_0(NMI_X)$ is the corresponding p–value for the association of a SNP X to the phenotype. The p–value is uniformly distributed over $[0, 1]$ if NMI_X is F_0-distributed. However, if X belongs to the G_1 distribution of unrelated SNPs, its corresponding p–value is skewed towards 1. On the other hand, if X is G_2 distributed, its p–value is skewed towards 0 (see Figure 3).

As the next step, based on two tuning parameters, λ_1 and λ_2, the fraction γ of the NMI_X which belong to the background is estimated using Equation (6):

$$\widehat{\gamma} = \frac{\text{number of } p\text{–values in } [\lambda_1, \lambda_2]}{\mathcal{P} \cdot (\lambda_2 - \lambda_1)} \tag{6}$$

so that the fraction of non-uniformly distributed p–values that fall into $[\lambda_1, \lambda_2]$ is negligible [65,67]. These two parameters are dataset-dependent and are automatically tuned through a trial and error heuristic approach during the analysis [68].

Finally, an SNP X is deemed as significant if its p–value is less or equal to τ, where τ is a threshold depending on a user-defined false discovery rate, FDR, estimated using Equation (7).

$$\widehat{FDR}(\tau) = \frac{\widehat{\gamma} \cdot \mathcal{P} \cdot \tau}{\text{number of } p\text{–values } \leq \tau} \tag{7}$$

For the detection of epistatic interactions using the $NMI(X^i, X^j; Y)$ metric, for our further analysis, we only consider SNP pairs where at least one SNP is significant, which results in a reduction in the runtime.

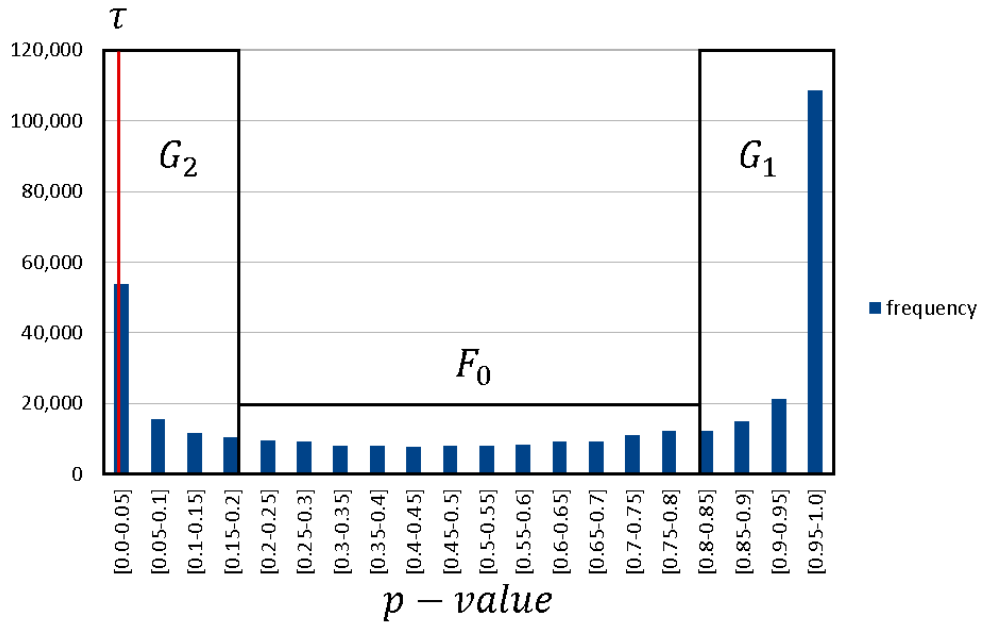


Figure 3. Distribution of p–values: the distribution can be divided in three parts, with G_2 representing the strongly associated SNPs, G_1 the unrelated SNPs and F_0 the background. SNPs with a p–value less or equal to τ are deemed as significant.

2.2.4. Reduction of the Background Associations between SNPs and Phenotype

As shown in previous studies [19,50,51], a dataset-dependent background association exists between the SNPs and the phenotype that may arise due to population stratification or relatedness of the individuals under study. Such obstacles could interfere with the identification of the correct epistatic signals, and thus could lead to the detection of false positive association signals. Another background association could occur in the detection of epistatic interactions using the NMI metric due to the high level of mutual information between a single SNP and the phenotype. We introduce this issue by way of an example in Section 3.2.

To eliminate these issues to some extent, in our study, we applied the average product correction (APC) introduced by Dunn et al. [53]. The APC theorem is a very successful information-theory-based approach to estimate the expected level of background association between the variables in a dataset. Meckbach et al. [69] showed that this approach is universally applicable, and thus we adopted it for our method. Following this approach, we estimated the expected level of the background between the SNP pair and the phenotype in the calculation of $NMI(X^i, X^j; Y)$ as

$$APC(X^i, X^j; Y) = \left(\frac{\overline{NMI_{X^i}} \cdot \overline{NMI_{X^j}}}{\overline{NMI_{SNP}}}\right) \tag{8}$$

In Equation (8), $\overline{NMI_{X^i}}$ and $\overline{NMI_{X^j}}$ are the average association levels of SNPs X^i and X^j, respectively, in the epistatic interaction:

$$\overline{NMI_{X^i}} = \frac{1}{h} \cdot \sum_{l=1}^{h} NMI(X^i, X^l; Y), \quad (9)$$

where h is a sufficiently large number (e.g., $h > 1000$) and the SNPs X^l are randomly chosen. Further, $\overline{NMI_{SNP}}$ denotes the overall average normalized mutual information calculated using a sufficiently large number of NMI values.

Finally, we subtracted the $APC(X^i, X^j; Y)$ value of an SNP pair and the phenotype from their initial $NMI(X^i, X^j; Y)$ to obtain the corrected $NMI_{apc}(X^i, X^j; Y)$.

2.2.5. Validation of the Epistatic Interactions

To identify the genes pertaining to epistatic SNP pairs, in our analysis, we only considered the p-th percentile of the pairs with an NMI_{apc} value > 0. For the interpretation of the interactions, we mapped the SNPs to their corresponding genes based on the mappings provided by the Ensembl database (release 103) [70]. The data were then read into R and a gene–gene interaction network was created with the genes as nodes and their interactions as edges using the igraph package [71]. The number of interactions of a node was termed its degree. In the final step, these degrees were transformed into z-scores and we consequently defined a gene as MIDESP-significant if its z-score was ≥ 3, as suggested in [69].

To elucidate the biological functions of these genes, we followed previous studies [55,72] and utilized the geneXplain platform [73] to perform a gene set analysis based on the molecular functions of the genes. The results were then visualized in the form of a treemap.

2.2.6. Implementation

The MIDESP pipeline was implemented in Java and is available as a JAR file from https://github.com/FelixHeinrich/MIDESP (accessed on 14 September 2021), allowing for easy usage. The calculations were completely parallelized, allowing for an efficient detection of significant epistatic interactions with a multi-core CPU. Genotype and phenotype information in the form of tped and tfam files were required as input.

3. Results

In this paper, we introduce a novel information-theory-based method, MIDESP, for the detection of epistatic interactions using genotype–phenotype datasets. MIDESP is able to analyze both qualitative as well as quantitative phenotypes, unlike previous information theoretical methods [39,41–46], which are only applicable to datasets with qualitative phenotypes. Furthermore, our method takes into account the effect of dataset-dependent background associations and eliminates them to some extent using the average product correction (APC) technique [53] to separate correct/functional epistatic signals from those of false positives.

This section consists of four major parts. First, in order to gain insights into the influence of the prerequisite parameter k used in Equation (4), we systematically analyzed several simulated datasets to find the most convenient value for it. Second, we introduced, by way of an example, the possible background association effects in epistatic interactions to highlight the importance of the APC approach in our method. In the following sections, we analyzed, by applying MIDESP with a false discovery rate (FDR) of 0.05, two different datasets with qualitative and quantitative phenotypes to demonstrate its functionality.

3.1. Analysis of Simulated Datasets for Parameter Setting

Today, it is well established that mutual information is an appropriate metric to measure the association between SNPs and qualitative (case–control) phenotypes [39,44,46,74–77]. However, we apply here for the first time this metric to quantitative traits. Therefore, we analyzed several simulated datasets to identify a proper value of k, which is necessary for the MI estimator (see Equations (2) and (4)). For this purpose, we employed the LDAK software [78] to simulate several hundred genotype and phenotype datasets with three

different heritability values: 0.05, 0.075 and 0.1. Consequently, we created 500 datasets consisting of 1000 SNPs, 2000 samples and a continuous phenotype controlled by a single SNP for each heritability value, respectively. Power was calculated as the proportion of datasets where the causal SNP obtained the highest MI value. To establish a proper value of k for the MI estimator, we systematically analyzed each dataset using k-values from 1 to 60. Despite Ross [52] and Kraskov et al. [79] both recommending a low value of $k = 3$, our analyses indicate that such small values can be only considered for heritability values > 0.1 (see Figure 4). Further, Figure 4 suggests that simulation datasets with smaller heritability values require a much higher k-value to successfully detect the causal SNPs of interest. By systematically analyzing different k-values, we established that a value of $k = 30$ leads to the highest increase in power for the estimator based on the heritability values under study. We did not choose a higher value, since an increase in k results in a longer runtime for the estimator and may likewise cause problems if the sample size is not large enough.

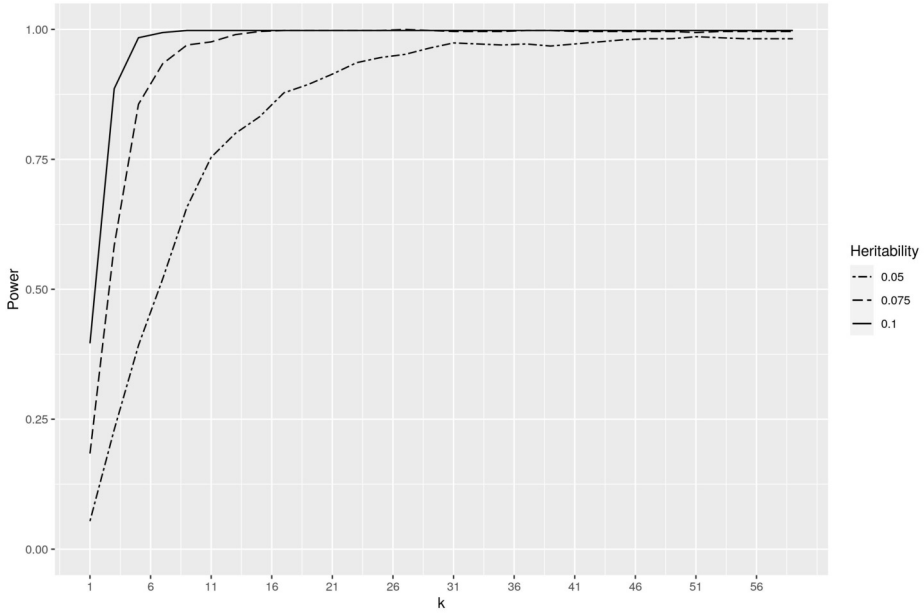


Figure 4. Analysis of simulated datasets for parameter setting of k.

3.2. Illustration of Background Associations and Its Correction Using APC

In information theory, mutual information (MI) is typically measured between two variables, X^1 and Y. Additionally, based on the chain rule of information [60], it is well known that the introduction of a new variable, X^2, might affect the relationship between X^1 and Y, thus increasing the MI between X^1 and Y. However, if the introduction of X^2 does not result in any new information, the corresponding MI value will not be affected [60].

In case of SNP×phenotype associations, this property of the MI needs to be considered since only the introduction of an additional SNP^2 which increases the amount of information between SNP^1 and the phenotype Y should be taken into account for the detection of epistatic interactions. The reason for this is exemplified in Figure 5. It can be seen in Figure 5 that SNP^1 and Y have the maximum MI value of 1, indicating their perfect association. On the other hand, SNP^2 as well as SNP^3 have an association value of 0 to Y. Applying Equation (3) clearly shows that the introduction of SNP^2 or SNP^3 does not affect the amount of association between SNP^1 and Y, but on the other hand leads to a false interpretation of epistatic interactions. To deal with this problem, we apply the average product correction (APC) theorem [53], which ensures the elimination of negligible increments in the MI value of epistatic interactions measured using Equations (3) and (4).

Another important aspect of the usage of the MI metric in the context of epistatic interactions is its ability to detect the newly created relationship between a SNP pair and the phenotype, even though the single SNPs themselves might not show any association to

the phenotype. This property of MI can be considered for measuring the level of association between $SNP^2 - SNP^3$ and Y (see Figure 5).

Y	SNP^1	SNP^2	SNP^3
1	AA	AA	AA
1	AA	Aa	Aa
1	AA	aa	aa
1	Aa	AA	AA
1	Aa	Aa	Aa
1	Aa	aa	aa
2	aa	AA	aa
2	aa	Aa	AA
2	aa	aa	Aa
2	aa	AA	aa
2	aa	Aa	AA
2	aa	aa	Aa

$MI(SNP^1;Y) = 1$
$MI(SNP^2;Y) = 0$
$MI(SNP^3;Y) = 0$
$MI(SNP^1,SNP^2;Y) = 1$
$MI(SNP^1,SNP^3;Y) = 1$
$MI(SNP^2,SNP^3;Y) = 1$

Figure 5. Example of MI values calculated from genotype data for three SNPs and twelve samples with a binary phenotype. The table cells are colored based on the genotype value of the SNP for the corresponding sample.

To demonstrate the importance of the APC in the analysis of epistatic interactions, we further applied it for the correction of the MI values calculated using Equation (3) regarding the BT dataset. We considered the top million MI values indicating the epistatic interaction between the SNP pairs and the phenotype. Afterwards, for each SNP, we determined its frequency among the interactions. The frequency distribution of SNPs and their single association to the phenotype is shown in Figure 6A. As mentioned above, the frequency of several SNPs is over-represented, which arises from their single association to the phenotype. However, the application of APC dramatically reduces their frequencies in the epistatic interactions. This finding clearly suggests that, although these SNPs individually have a strong association to the phenotype, their epistatic interactions are negligible, as shown with blue points in Figure 6.

3.3. Bovine Tuberculosis Dataset

By applying MIDESP to the BT dataset, we first identified 10,774 single SNPs in total, with significant association to the phenotype. Taking all SNP pairs that contain at least one of those significant SNPs into account, for the epistatic interaction analysis, we identified 3,799,984 SNP pairs, which corresponds to 0.1% of all possible pairs under study. After that, we mapped these SNPs to their corresponding genes using the Ensembl database and a gene–gene interaction network was created, as suggested in [80]. Finally, according to this network, we detected 511 genes as MIDESP-significant and investigated their roles in bovine tuberculosis disease based on enriched Gene Ontology (GO) terms (see treemap depicted in Figure 7 and Supplementary Table S1).

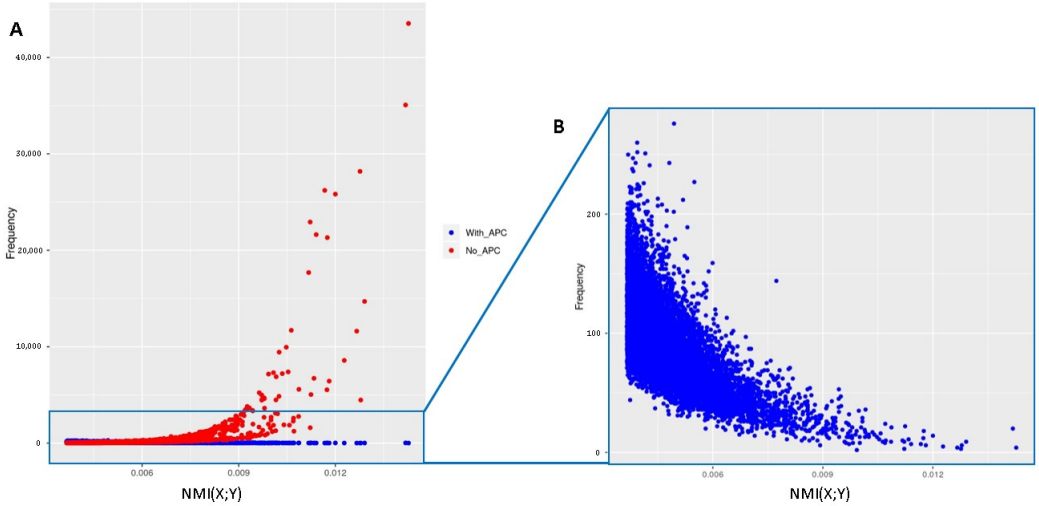


Figure 6. (**A**) shows the distribution of the SNP frequency and their association to the phenotype. The blue and red points stand for the frequency of the SNPs based on with and without the application of the APC, respectively. (**B**) only shows the frequencies for the interactions with APC.

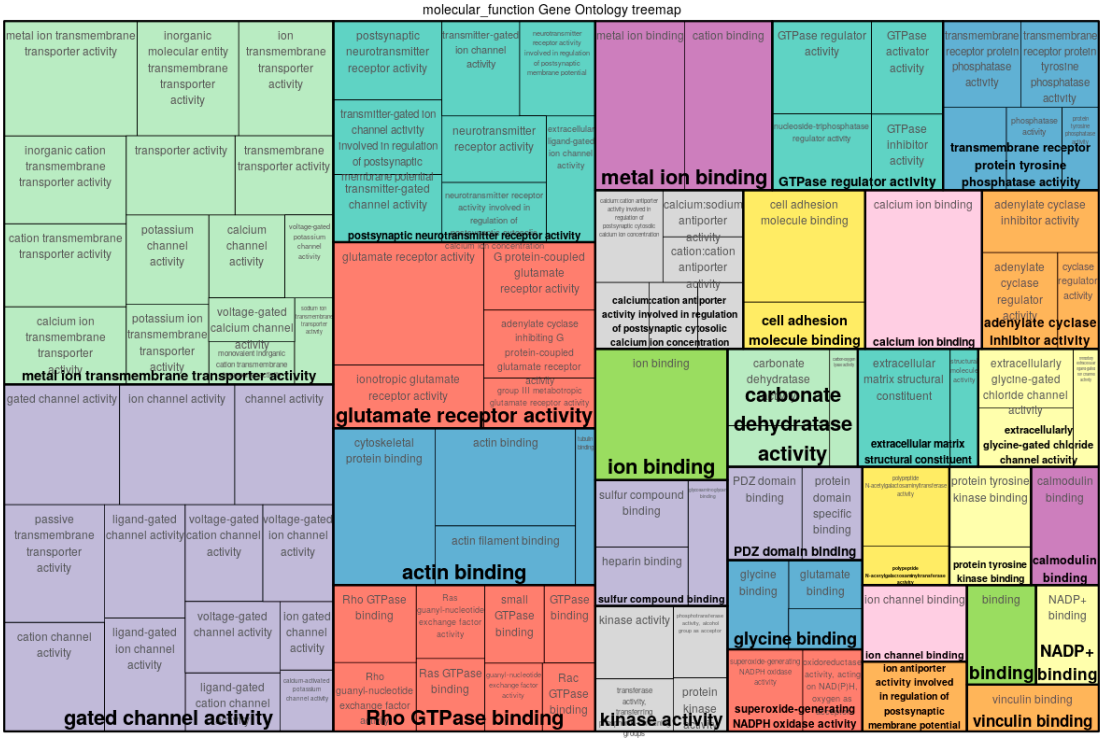


Figure 7. Gene Ontology (GO) treemap for genes associated with immunity to bovine tuberculosis. The boxes are grouped together based on the upper-hierarchy GO term, which is written in bold letters.

The functional classification of these genes indicates that several of the GO categories represented in the treemap play essential roles in the immune responses towards bovine tuberculosis. Especially, metal ion transmembrane transporter activity and gated channel activity are the most significantly enriched terms, shown in the green and purple boxes in the treemap (Figure 7) obtained from the GO analysis, indicating the function of transmembrane proteins involved in the transportation of ions across membrane layers. Particularly, ion channel blockers are known for their therapeutic implications in drug-resistant mycobacterial infection, especially voltage gated calcium channels, which are important for the regulation of immunity against pathogens [81–84]. In this regard, increasing calcium influx by inhibiting the voltage gated channels in immune cells such as macrophages is highly

associated with protective immunity, particularly in increasing the expression of genes involved in pro-inflammatory responses [84]. Other significant GO terms including actin binding, Rho GTPase binding, glutamate receptor activity and postsynaptic neurotransmitter receptor activity were also enriched in the treemap and their roles associated with *Mycobacterium tuberculosis* are described below. Firstly, actin filament, which is an important constituent of the cytoskeleton [85], is mainly associated with pro-inflammatory responses. A primary aspect of mycobacterial infection is the manipulation of actin filaments [86], notably inside the macrophages (immune cells engulfing the pathogens) of the host [87–89], thereby pointing out the importance of actin-binding protein regulation for enhancing the immune responses of the host. Several recent studies reported that neurotransmitters play essential roles in the activation or suppression of immune responses through the regulation of T-cell activity [90,91]. It is well known that T-cells play an important part in the defense of the host against mycobacterial infections [92–94]. Specifically, the neurotransmitter taurine was identified in relation with the susceptibility of cattle towards bovine tuberculosis [95]. Glutamate is likewise a neurotransmitter known for its effect on the immune system for the regulation of T-cell activity [96,97]. Finally, Ras homology GTPases (Rho GTPase) are proteins involved in the critical regulation of signaling pathways upon bacterial entry at the site of infection, and therefore are involved in innate immune responses, particularly in the multiplication of immune cells. It is essential to coordinate the immune responses at this point to prevent the neighboring tissue from taking damage from inflammation. Involved in the tight regulatory roles of multiple immune functions, these signaling proteins have been reported as targets of *Mycobacterium tuberculosis* during the host cell invasion, which might facilitate the pathogenesis of the bacteria [98–100].

3.4. Egg Weight Dataset

Similarly to the previous dataset, MIDESP was used to analyze the EW dataset, which contains a quantitative phenotype. As a first step, we detected 3116 single SNPs that were significantly associated with the trait. Based on these SNPs, we measured the epistatic interactions between the SNP pairs and the phenotype and obtained 1,071,464 SNP pairs in total that equate to 0.25% of all possible pairs under study. After mapping these pairs to a gene–gene interaction network, we were able to identify 211 genes as MIDESP-significant. The analysis of their roles regarding egg weight was again carried out using their enriched GO terms (see treemap depicted in Figure 8 and Supplementary Table S2).

For egg weight, one of the major GO categories that emerged as a result of the gene set analysis was the fatty acid ligase activity. Fatty acid ligases belong to the ligase family of enzymes that take part in the biosynthesis of lipids [101]. Lipids constitute a major portion of the nutrients found in egg and are primarily contained in egg yolk, which constitutes 31% of the total egg weight [102]. Multiple genes encoding fatty acid ligases have been reported to play important roles in the laying performance of birds [103–105]. In this regard, we were able to discover many genes with molecular functions associated with acyl-CoA ligases, a group of enzymes, which are known to play important roles in the lipid synthesis by making the chemically inert fatty acids undergo activation into acyl-CoA [106]. This activation comprises an ATP-dependent reaction catalyzed by ligase enzymes in the presence of Mg^{2+} and CoA [107]. The usage of ATP and Mg^{2+} in this process can also explain the role of adenyl nucleotide binding and magnesium ion binding, two other categories identified in our analysis. Gated channel activity is another important GO term found in this analysis. These genes ensure the transportation of nutrients and minerals, which are required for the development of the egg. More importantly, for the synthesis of the eggshell, which contributes around 9% to the total egg weight [102], large amounts of calcium ions are supplied to the uterine fluid by transepithelial transport [102,108]. This transepithelial transport occurs with the help of ion channels, ion pumps and ion exchangers in the reproductive tract of birds and the energy required for these processes is provided by the metabolisms of ATP molecules [108]. Both nucleotide binding and gated channel activity have been reported in association with egg weight and eggshell

development in chicken [54,55]. Furthermore, genes related to protein transmembrane transport activity were also identified in our analysis, which can regulate the transportation of the large number of proteins found in an egg [102,109]. The gene set analysis further reveals other activities pertaining to molecular bindings at different levels, which can play roles crucial for the development of egg.

Figure 8. Gene Ontology (GO) treemap for genes associated with egg weight. The boxes are grouped together based on the upper-hierarchy GO term, which is written in bold letters.

3.5. Comparisons with Existing Methods

To investigate the performance of our new method, we were further interested in making pairwise comparisons between the results of our MIDESP, PLINK [57], GBOOST [110], epiGPU [111] and MatrixEpistasis [112]. Although all these methods take a genotype–phenotype dataset as input and report epistatic SNP pairs as result, their applicability differs based on the phenotypes. While MIDESP and PLINK can be applied to qualitative as well as quantitative phenotypes, the other methods are restricted to one type. GBOOST only deals with qualitative phenotypes, while epiGPU and MatrixEpistasis only analyze quantitative phenotypes. We chose these tools since they have previously been used for pairwise epistasis detection on real datasets, as well as for comparison studies [41,113–119], and ran them with their default parameters. It is important to note that for this comparison study, we applied MIDESP with and without APC correction. While the MIDESP without APC is in line with the conventional mutual information (MI)-based methods for epistasis detection [39,46,80,120], the incorporation of the APC approach is completely novel and necessary to separate the correct epistatic signals from the background.

The results of this comparison are twofold. First, we compared the results of our method using the BT dataset with those of PLINK, GBOOST and the conventional MI-based metric, since the existing MI-based approaches are only applicable to qualitative phenotypes [39,46,80,120]. Second, we compared the predictions of MIDESP on the quantitative EW dataset with those of PLINK, epiGPU and MatrixEpistasis. However, our attempt to apply MatrixEpistasis to this dataset was not successful due to its very high memory consumption (700 GB of memory was not enough).

The application of these methods results in the detection of strongly varying numbers of SNP pairs as epistatic interactions, which are given in Table 2.

Table 2. Number of SNP pairs that were found to be an epistatic interaction by the different methods. BT and EW stand for bovine tuberculosis and egg weight, respectively.

Dataset	#MIDESP	#MIDESP_NoAPC	#PLINK	#GBOOST	#epiGPU
BT	3,799,984	3,799,984	4,982,695	346,632	-
EW	1,071,463	1,071,463	1,817,817	-	572,914

To make the predictions of the methods comparable, in this comparison analysis for both types of the traits, we considered 346,632 and 572,914 epistatic SNP pairs, which corresponds to the minimum numbers of SNP pairs found by GBOOST and epiGPU for the BT and EW datasets, respectively (see Table 2). Based on these top SNP pairs, we further performed an overlap comparison between the methods and visualized the results using UpSet plots in Figures 9 and 10, respectively. Although all of these methods perform a search for epistatic SNP pairs, Figures 9 and 10 clearly show that they provide quite distinct results, with only little overlap between them. This finding is in line with the comparison study performed in [113], which also reported divergent results between different methods for epistasis detection. The reason for that may be explained due to differences in the underlying algorithms, even though the three other methods are ultimately based on logistic and linear regression, respectively. While PLINK performs a regression with an interaction term and tests whether the coefficient for the interaction is significant, GBOOST considers the difference in the likelihood of a linear model with interaction compared to that of a model without as a sign for epistasis using approximations to speed up the process and filter out SNP pairs. On the other hand, epiGPU treats the different genotype combinations as different classes and calculates differences in the class means compared to the population mean.

Consequently, the results of this overlap analysis clearly demonstrate that these methods carry quite distinct information about epistatic interactions, due to the different measures they use. The finding of this comparison analysis is also in agreement with the previous study [113] and indicates that each of these methods takes into account a different manner of epistatic interactions, and thus they can work complementarily with each other.

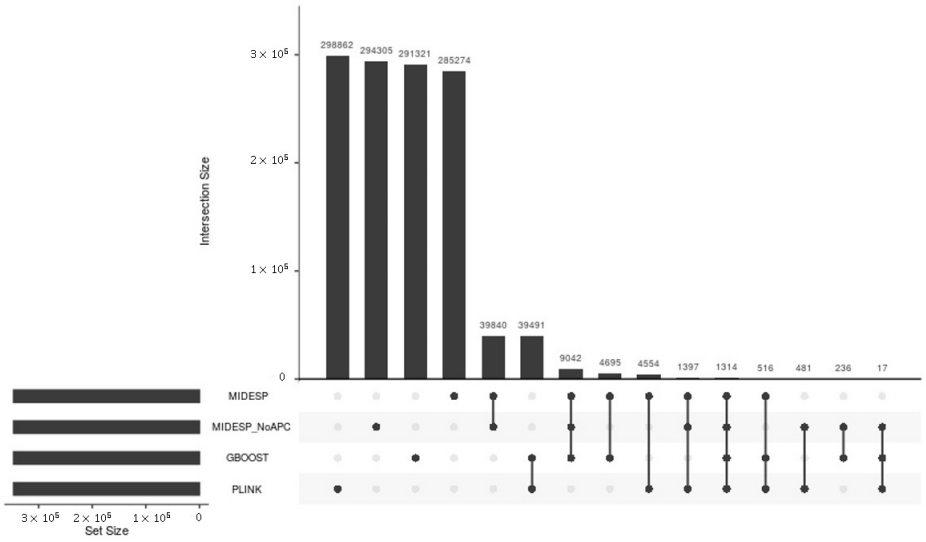


Figure 9. Number of epistatic SNP pairs detected for the BT dataset and their overlap between four methods represented in matrix layouts using the UpSet technique [121]. Black circles in the matrix layout indicate which methods are part of the intersection.

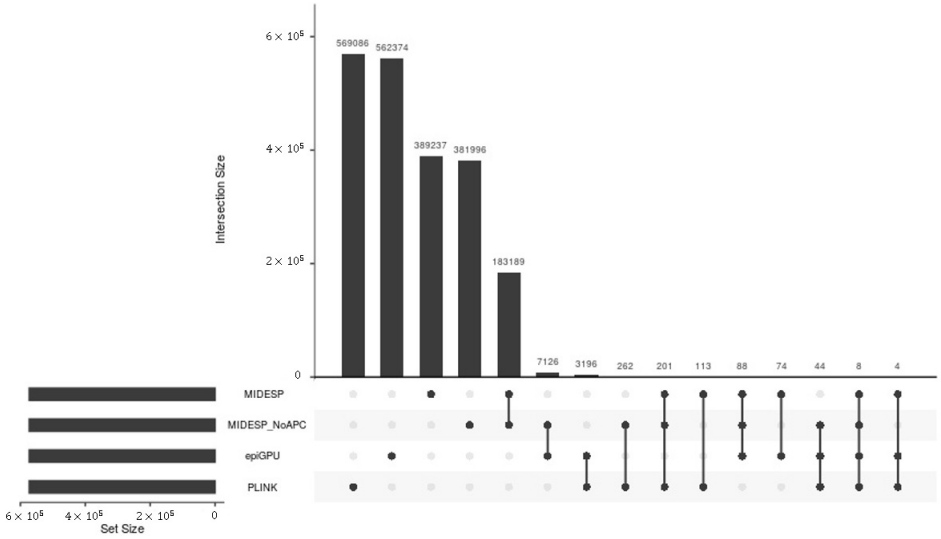


Figure 10. Number of epistatic SNP pairs detected for the EW dataset and their overlap between four methods represented in matrix layouts using the UpSet technique [121]. Black circles in the matrix layout indicate which methods are part of the intersection.

4. Discussion

It has previously been shown that information theoretical methods based on mutual information (MI) are powerful approaches for the detection of epistatic interactions [39,41,43–46]. Not only here, but also in many other fields, mutual information has been used as an effective measure for the association between variables including linear as well as non-linear relationships [53,61,63,69,122–125]. However, the general applicability of a method, particularly in the field of animal and plant breeding, requires it to be usable for qualitative as well as quantitative phenotypes. For this reason, an extension of the previous MI methods, which are only suitable for qualitative traits, is required, and thus we adapted the estimator developed by Ross [52] for the case of MI between discrete and continuous variables. As shown in Section 3.1, the estimator can be successfully used to detect associations between SNPs and quantitative phenotypes. Surprisingly, we found that a higher k value improves the power of the measure when it comes to the detection of associations involving traits of a low heritability (see Figure 4), although previous studies recommended a small value of k for this purpose [52,79].

The progress over the last decade in the field of genome sequencing and genotyping arrays has increased the amount of available genotype data tremendously. With the ever-increasing amount of data, however, comes the challenge to provide tools that can handle such datasets in a feasible computation time. To address this challenge, redundant SNPs can be removed through LD pruning with a high threshold [56] (see Section 2.1) but there are still very high numbers of SNPs in a dataset to analyze all possible pairs. A commonly used approach to reduce the computational effort is to preselect sets of SNPs that are deemed as important and only analyze those, as is performed by BOOST and other methods [38,126,127]. Such an approach can potentially eliminate some SNPs which nevertheless influence the phenotype in interaction with another SNP. To overcome this problem, in our proposed method, we consider all SNP pairs where at least one SNP shows a strong association signal to the phenotype, which ensures a tractable computational time for MIDESP. For this step, we followed the approach outlined by Gültas et al. [63,64] to separate the SNPs with strong association signals from the remaining SNPs (see Section 2.2.3).

However, the sole consideration of SNPs with strong association signals could lead to a wrong interpretation in epistasis analysis since the NMI values are influenced by the association of the single SNPs with the phenotype, as we demonstrated by means of an example in Figure 5. This can result in the detection of false positive interactions that are only found due to the effect of one SNP. To minimize this influence, the application of the

average product correction (APC) is essential, which was developed by Dunn et al. [53]. Moreover, Meckbach et al. [69] showed that the APC is universally applicable to MI-based methods to estimate the expected (background) association level of a variable. Although the concept of the APC theorem seems to be suitable for our purposes, its application would require a huge additional computational overhead. Therefore, we followed a strategy based on the three different distributions of the SNPs (see Section 2.2.3) for the efficient estimation of the expected level of background associations of SNPs. In particular, in Equation (9) we randomly choose the SNP X^l from the set of SNPs that follow the G_2 distribution. This process ensures that the expected background level of SNP X^i is clearly higher than it would be if estimated based on the whole set of SNPs. Consequently, the removal of the estimated background associations (APC values) from the obtained NMI values results in the separation of correct epistatic signals caused by SNP pair and phenotype interactions from background signals. Being of particular interest, in our analysis, we illustrated the effectiveness of the APC based on the BT dataset in Figure 6. This analysis reveals that the over-representation of SNPs with a large single effect among the pairs with the highest NMI values can be considerably reduced based on the application of APC, which in turn results in the detection of further associated genes.

The results we present in this study for the two different genotype–phenotype datasets show that the functional analysis of the detected genes provides essential information to decipher the genetic background of the traits under consideration. Surprisingly, we were able to clearly identify higher numbers of associated genes for the bovine tuberculosis dataset with a qualitative trait than for the egg weight dataset with a quantitative trait. The reason for this can be explained due to the large difference in the initial numbers of SNPs in both datasets (see Table 1). In comparison to the large numbers of associated genes detected by MIDESP, both original studies [58,59], in which the datasets were published, were only able to find two significantly associated genes for the respective dataset using standard GWAS approaches.

To further investigate the impact of the APC theorem in the epistasis analysis and to gain more insight into its influence on the detection of genes, we analyzed both datasets with and without the application of the APC (see Figure 6). It can be assumed that without the APC, the results of MIDESP are in line with previous methods that utilized MI for the detection of epistatic interactions for qualitative phenotypes [39,46,80,120]. The analysis reveals that the application of the APC leads to a considerable increase in the number of associated genes for both datasets. For example, only 135 and 177 significant genes were found for the BT and EW datasets without using the APC, respectively. However, the correction of the background association using the APC results in the detection of 511 and 211 associated genes, respectively. The comparison of these genes showed that while 59 genes overlap for the BT dataset, 51 overlapping genes are found to be significant for the EW dataset. The functional analysis of these genes based on their GO categories reveals that many of the identified genes are involved in the regulation of the immune system regarding bovine tuberculosis, with several of the functions having a reported association with mycobacterial infections. The genes that were detected for the egg weight dataset, on the other hand, are mainly related to the production of important components of the egg and the transportation of these components to the uterine fluid. Overall, our results indicate that MIDESP is an effective method for the detection of epistatic interactions that for the first time enables the analysis of quantitative phenotypes using MI and further extends the existing information theoretical methods by correcting the influence of background associations of the SNPs through the application of the APC theorem.

5. Conclusions

Today, it is well established that MI-based methods are suitable and effective approaches for the detection of epistatic interactions for qualitative phenotypes. However, these approaches are not directly applicable for quantitative phenotypes, although epistatic interactions for quantitative traits are of great interest in life sciences. To address this

limitation of the existing MI-based methods, we extend their applicability for the first time in this regard to quantitative phenotypes using a *k*th-nearest neighbor-based estimation technique. Another important challenge for the detection of epistatic interactions is the control of the effect of background associations in the genotype–phenotype datasets, which lead to false interpretation and thus the overestimation of the role of some SNPs in the epistasis. To deal with this issue, in our proposed method, MIDESP, we additionally modeled these background associations by adopting the APC theorem, which we extended for the multivariate mutual information. Our findings show that the MIDESP algorithm is applicable to genotype–phenotype datasets with qualitative as well as quantitative phenotypes in a tractable computational time. For example, the analysis of the BT dataset took only 36 minutes, while the analysis of the EW dataset was completed in 105 minutes. These runtimes were achieved on a dual Intel® Xeon® Gold 6138 Processor using 70 threads. Our results further indicate that the biological processes of the identified genes in the BT and EW datasets are strongly related to both bovine tuberculosis and the egg weight of chickens, respectively. To the best of our knowledge, MIDESP is the first method that models epistatic interactions using the MI metric for both qualitative and quantitative phenotypes and explicitly corrects for background associations. The program is written in Java and is freely available as a JAR file from https://github.com/FelixHeinrich/MIDESP, accessed on 14 September 2021.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/biology10090921/s1, Table S1: results of the gene set analysis for the BT dataset, Table S2: results of the gene set analysis for the EW dataset.

Author Contributions: M.G. developed the model underlying MIDESP, and designed and supervised the research. F.H. developed the model together with M.G. and participated in the design of the study, prepared the data sets, conducted the bioinformatics analyses and implemented the tool. F.R., A.R. and A.O.S. were involved in the interpretation of the results, together with F.H. and M.G. F.H. and M.G. wrote the final version of the manuscript. M.G. conceived and managed the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets that we analyzed as well as the source code for MIDESP can found in the repository https://github.com/FelixHeinrich/MIDESP, accessed on 14 September 2021.

Acknowledgments: We thank Stephan Waack from Göttingen for his helpful advice and insights at early stages of this project. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the University of Göttingen.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SNP	Single-nucleotide polymorphism
GWAS	Genome-wide association studies
MI	Mutual information
NMI	Normalized mutual information
BT	Bovine tuberculosis
EW	Egg weight
APC	Average product correction
FDR	False discovery rate
GO	Gene Ontology

References

1. Wei, W.H.; Hemani, G.; Haley, C. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **2014**, *15*, 722–733. [CrossRef] [PubMed]
2. Phillips, P.C. Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **2008**, *9*, 855–867. [CrossRef]
3. Huang, W.; Richards, S.; Carbone, M.A.; Zhu, D.; Anholt, R.R.H.; Ayroles, J.F.; Duncan, L.; Jordan, K.W.; Lawrence, F.; Magwire, M.M.; et al. Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 15553–15559. [CrossRef]
4. Moore, J.H.; Asselbergs, F.W.; Williams, S.M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **2010**, *26*, 445–455. [CrossRef]
5. Moore, J.H.; Williams, S.M. Epistasis and Its Implications for Personal Genetics. *Am. J. Hum. Genet.* **2009**, *85*, 309–320. [CrossRef]
6. Marchini, J.; Donnelly, P.; Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **2005**, *37*, 413–417. [CrossRef] [PubMed]
7. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [CrossRef]
8. Yoshikawa, T.; Kanazawa, H.; Fujimoto, S.; Hirata, K. Epistatic effects of multiple receptor genes on pathophysiology of asthma—Its limits and potential for clinical application. *Med. Sci. Monit.* **2014**, *20*, 64–71.
9. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am. J. Hum. Genet.* **2001**, *69*, 138–147. [CrossRef] [PubMed]
10. Cho, Y.M.; Ritchie, M.D.; Moore, J.H.; Park, J.Y.; Lee, K.-U.; Shin, H.D.; Lee, H.K.; Park, K.S. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* **2004**, *47*, 549–554. [CrossRef]
11. Carlborg, O.; Hocking, P.; Burt, D.; Haley, C. Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. *Genet. Res.* **2004**, *83 3*, 197–209. [CrossRef]
12. Le Rouzic, A.; Alvarez-Castro, J.M.; Carlborg, O. Dissection of the genetic architecture of body weight in chicken reveals the impact of epistasis on domestication traits. *Genetics* **2008**, *179*, 1591–1599. [CrossRef]
13. Mackay, T.F.C. Epistasis and quantitative traits: Using model organisms to study gene–gene interactions. *Nat. Rev. Genet.* **2014**, *15*, 22–33. [CrossRef] [PubMed]
14. Knaust, J.; Hadlich, F.; Weikard, R.; Kuehn, C. Epistatic interactions between at least three loci determine the "rat-tail" phenotype in cattle. *Genet. Sel. Evol.* **2016**, *48*, 26. [CrossRef]
15. Kramer, L.M.; Ghaffar, M.A.A.; Koltes, J.E.; Fritz-Waters, E.R.; Mayes, M.S.; Sewell, A.D.; Weeks, N.T.; Garrick, D.J.; Fernando, R.L.; Ma, L.; et al. Epistatic interactions associated with fatty acid concentrations of beef from angus sired beef cattle. *BMC Genom.* **2016**, *17*, 891. [CrossRef] [PubMed]
16. Würschum, T.; Maurer, H.P.; Schulz, B.; Möhring, J.; Reif, J.C. Genome-wide association mapping reveals epistasis and genetic interaction networks in sugar beet. *Theor. Appl. Genet.* **2011**, *123*, 109–118. [CrossRef]
17. Hu, Z.; Li, Y.; Song, X.; Han, Y.; Cai, X.; Xu, S.; Li, W. Genomic value prediction for quantitative traits under the epistatic model. *BMC Genet.* **2011**, *12*, 15. [CrossRef]
18. Huang, A.; Xu, S.; Cai, X. Whole-Genome Quantitative Trait Locus Mapping Reveals Major Role of Epistasis on Yield of Rice. *PLoS ONE* **2014**, *9*, e87330. [CrossRef]
19. Ahsan, A.; Monir, M.; Meng, X.; Rahaman, M.; Chen, H.; Chen, M. Identification of epistasis loci underlying rice flowering time by controlling population stratification and polygenic effect. *DNA Res.* **2018**, *26*, 119–130. [CrossRef]
20. Mathew, B.; Léon, J.; Sannemann, W.; Sillanpää, M.J. Detection of Epistasis for Flowering Time Using Bayesian Multilocus Estimation in a Barley MAGIC Population. *Genetics* **2018**, *208*, 525–536. [CrossRef] [PubMed]
21. Carlborg, Ö.; Haley, C.S. Epistasis: Too often neglected in complex trait studies? *Nat. Rev. Genet.* **2004**, *5*, 618–625. [CrossRef]
22. Cordell, H.J. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **2002**, *11*, 2463–2468. [CrossRef] [PubMed]
23. Nelson, M.R.; Kardia, S.L.; Ferrell, R.E.; Sing, C.F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **2001**, *11*, 458–470. [CrossRef] [PubMed]
24. Culverhouse, R. The Use of the Restricted Partition Method with Case-Control Data. *Hum. Hered.* **2007**, *63*, 93–100. [CrossRef]
25. Li, X. A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization. *Bioinformatics* **2017**, *33*, 2829–2836. [CrossRef]
26. Zhang, X.; Huang, S.; Zou, F.; Wang, W. TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* **2010**, *26*, i217–i227. [CrossRef] [PubMed]
27. Zhang, Y.; Liu, J.S. Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **2007**, *39*, 1167–1173. [CrossRef]
28. Tang, W.; Wu, X.; Jiang, R.; Li, Y. Epistatic Module Detection for Case-Control Studies: A Bayesian Model with a Gibbs Sampling Strategy. *PLoS Genet.* **2009**, *5*, e1000464. [CrossRef] [PubMed]
29. Serretti, A.; Smeraldi, E. Neural network analysis in pharmacogenetics of mood disorders. *BMC Med. Genet.* **2004**, *5*, 27. [CrossRef]

30. Motsinger-Reif, A.A.; Dudek, S.M.; Hahn, L.W.; Ritchie, M.D. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet. Epidemiol.* **2008**, *32*, 325–340. [CrossRef]
31. Uppu, S.; Krishna A.; Gopalan, R. Towards Deep Learning in genome-Wide Association Interaction studies. In Proceedings of the 20th Pacific Asia Conference on Information Systems, PACIS 2016, Chiayi, Taiwan, 27 June–1 July; Volume 20.
32. Wang, H.; Yue, T.; Yang, J.; Wu, W.; Xing, E.P. Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies. *BMC Bioinform.* **2019**, *20*, 656. [CrossRef]
33. Xie, Q.; Ratnasinghe, L.D.; Hong, H.; Perkins, R.; Tang, Z.; Hu, N.; Taylor, P.R.; Tong, W. Decision Forest Analysis of 61 Single Nucleotide Polymorphisms in a Case-Control Study of Esophageal Cancer: A novel method. *BMC Bioinform.* **2005**, *6*, S4. [CrossRef]
34. Winham, S.J.; Colby, C.L.; Freimuth, R.R.; Wang, X.; de Andrade, M.; Huebner, M.; Biernacka, J.M. SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinform.* **2012**, *13*, 164. [CrossRef]
35. Meng, Y.A.; Yu, Y.; Cupples, L.A.; Farrer, L.A.; Lunetta, K.L. Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinform.* **2009**, *10*, 78. [CrossRef]
36. Schwarz, D.F.; König, I.R.; Ziegler, A. On safari to Random Jungle: A fast implementation of Random Forests for high-dimensional data. *Bioinformatics* **2010**, *26*, 1752–1758. [CrossRef]
37. Yoshida, M.; Koike, A. SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinform.* **2011**, *12*, 469. [CrossRef] [PubMed]
38. Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N.L.S.; Yu, W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **2010**, *87*, 325–340. [CrossRef] [PubMed]
39. Leem, S.; Jeong, H.; Lee, J.; Wee, K.; Sohn, K. Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. *Comput. Biol. Chem.* **2014**, *50*, 19–28. [CrossRef] [PubMed]
40. He, D.; Parida,L. Muse: A multi-locus sampling-based epistasis algorithm for quantitative genetic trait prediction. In *Pacific Symposium on Biocomputing 2017*; World Scientific: Singapore, 2017; pp. 426–437.
41. Tuo, S. FDHE-IW: A Fast Approach for Detecting High-Order Epistasis in Genome-Wide Case-Control Studies. *Genes* **2018**, *9*, 435. [CrossRef] [PubMed]
42. Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* **2007**, *3*, 83. [CrossRef] [PubMed]
43. Hu, T.; Chen, Y.; Kiralis, J.W.; Collins, R.L.; Wejse, C.; Sirugo, G.; Williams, S.M.; Moore, J.H. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 630–636. [CrossRef] [PubMed]
44. Anunciação, O.; Vinga, S.; Oliveira, A.L. Using Information Interaction to Discover Epistatic Effects in Complex Diseases. *PLoS ONE* **2013**, *8*, e76300. [CrossRef]
45. Wienbrandt, L.; Kassens, J.C.; Hübenthal, M.; Ellinghaus, D. Fast Genome-Wide Third-order SNP Interaction Tests with Information Gain on a Low-cost Heterogeneous Parallel FPGA-GPU Computing Architecture. *Procedia Comput. Sci.* **2017**, *108*, 596–605. [CrossRef]
46. Ponte-Fernández, C.; González-Domínguez, J.; Martín, M.J. Fast search of third-order epistatic interactions on CPU and GPU clusters. *Int. J. High Perform. Comput. Appl.* **2020**, *34*, 20–29. [CrossRef]
47. He, D.; Parida, L. Does encoding matter? A novel view on the quantitative genetic trait prediction problem. *BMC Bioinform.* **2016**, *17*, 272. [CrossRef] [PubMed]
48. Martini, J.W.R.; Gao, N.; Cardoso, D.F.; Wimmer, V.; Erbe, M.; Cantet, R.J.C.; Simianer,H. Genomic prediction with epistasis models: On the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinform.* **2017**, *18*, 3. [CrossRef]
49. Martini, J.W.R.; Rosales, F.; Ha, N.; Heise, J.; Wimmer, V.; Kneib, T. Lost in Translation: On the Problem of Data Coding in Penalized Whole Genome Regression with Interactions. *G3 Genes Genomes Genet.* **2019**, *9*, 1117–1129. [CrossRef]
50. Mangin, B.; Siberchicot, A.; Nicolas, S.; Doligez, A.; This, P.; Cierco-Ayrolles, C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* **2012**, *108*, 285–291. [CrossRef]
51. Mezmouk, S.; Dubreuil, P.; Bosio, M.; Décousset, L.; Charcosset, A.; Praud, S.; Mangin, B. Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. *Theor. Appl. Genet.* **2011**, *122*, 1149–1160. [CrossRef]
52. Ross, B.C. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef]
53. Dunn, S.D.; Wahl, L.M.; Gloor, G.B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **2007**, *24*, 333–340. [CrossRef]
54. Ramzan, F.; Gültas, M.; Bertram, H.; Cavero, D.; Schmitt, A.O. Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations. *Genes* **2020**, *11*, 892. [CrossRef]
55. Ramzan, F.; Klees, S.; Schmitt, A.O.; Cavero, D.; Gültas, M. Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forests. *Genes* **2020**, *11*, 464. [CrossRef] [PubMed]
56. Joiret, M.; Mahachie John, J.M.; Gusareva, E.S.; Van Steen, K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Min.* **2019**, *12*, 11. [CrossRef] [PubMed]
57. Chang, C.C.; Chow, C.C.; Tellier, L.CAM.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **2015**, *4*, s13742-015. [CrossRef] [PubMed]

58. Bermingham, M.L.; Bishop, S.C.; Woolliams, J.A; Pong-Wong, R.; Allen, A.R.; McBride, S.H.; Ryder, J.J.; Wright, D.M.; Skuce, R.A.; McDowell, S.W.J.; et al. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. *Heredity* **2014**, *112*, 543–551. [CrossRef] [PubMed]
59. Liu, Z.; Sun, C.; Yan, Y.; Li, G.; Wu, G.; Liu, A.; Yang, N. Genome-Wide Association Analysis of Age-Dependent Egg Weights in Chickens. *Front. Genet.* **2018**, *9*, 128–128. [CrossRef]
60. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley: New York, NY, USA, 1991.
61. Dionisio,A.; Menezes,R.; Mendes,D.A. Mutual information: A measure of dependency for nonlinear time series. *Phys. A Stat. Mech. Its Appl.* **2004**, *344*, 326–329. [CrossRef]
62. Kvålseth, T.O. On Normalized Mutual Information: Measure Derivations and Properties. *Entropy* **2017**, *19*, 631. [CrossRef]
63. Gültas, M.; Haubrock, M.; Tüysüz, N.; Waack, S. Coupled mutation finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *BMC Bioinform.* **2012**, *13*, 225. [CrossRef]
64. Gültas, M.; Düzgün, G.; Herzog, S.; Jäger, S.J.; Meckbach, C.; Wingender, E.; Waack, S. Quantum coupled mutation finder: Predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. *BMC Bioinform.* **2014**, *15*, 96. [CrossRef]
65. Storey, J.D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 9440–9445. [CrossRef]
66. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
67. Walsh, B. *Multiple comparisons: Bonferroni Corrections and False Discovery Rates*; Lecture Notes for EEB 581 ; Department of Ecology and Evolutionary Biology, University of Arizona: Tucson, AZ, USA, 2004; pp. 1–17.
68. Gültas, M. Development of novel Classical and Quantum Information Theory Based Methods for the Detection of Compensatory Mutations in MSAs **2014**. Available online: https://hdl.handle.net/11858/00-1735-0000-0022-5EB0-1 (accessed on 14 September 2021).
69. Meckbach, C.; Tacke, R.; Hua, X.; Waack, S.; Wingender, E.; Gültas, M. PC-TraFF: Identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinform.* **2015**, *16*, 400. [CrossRef] [PubMed]
70. Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; et al. Ensembl 2020. *Nucleic Acids Res.* **2019**, *48*, D682–D688. [CrossRef]
71. Csardi, G.; Nepusz,T. The igraph software package for complex network research. *InterJournal Complex Syst.* **2006**, *1695*, 1–9.
72. Mekonnen, Y.A.; Gültas, M.; Effa, K.; Hanotte, O.; Schmitt, A.O. Identification of candidate signature genes and key regulators associated with Trypanotolerance in the Sheko Breed. *Front. Genet.* **2019**, *10*, 1095. [CrossRef] [PubMed]
73. Wingender, E.; Kel, A. geneXplain–eine integrierte Bioinformatik-Plattform. *BIOspektrum* **2012**, *18*, 554–556. [CrossRef]
74. Cao, X.; Yu, G.; Liu, J.; Jia, L.; Wang, J. Clustermi: Detecting high-order snp interactions based on clustering and mutual information. *Int. J. Mol. Sci.* **2018**, *19*, 2267. [CrossRef]
75. Guo, H.; Yu, Z.; An, J.; Han, G.; Ma, Y.; Tang, R. A two-stage mutual information based Bayesian Lasso algorithm for multi-locus genome-wide association studies. *Entropy* **2020**, *22*, 329. [CrossRef]
76. Sun, L.; Wang, C.; Hu, Y. Utilizing mutual information for detecting rare and common variants associated with a categorical trait. *PeerJ* **2016**, *4*, e2139. [CrossRef]
77. Yuan, X.; Zhang, J.; Wang, Y. Mutual information and linkage disequilibrium based SNP association study by grouping case-control. *Genes Genom.* **2011**, *33*, 65–73. [CrossRef]
78. Speed, D.; Hemani, G.; Johnson, M.R.; Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **2012**, *91*, 1011–1021. [CrossRef]
79. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, [CrossRef]
80. Wang, S.; Jeong, H.; Kim, D.; Wee, K.; Park, H.; Kim, S.; Sohn, K. Integrative information theoretic network analysis for genome-wide association study of aspirin exacerbated respiratory disease in Korean population. *BMC Med. Genom.* **2017**, *10*, 33–44. [CrossRef] [PubMed]
81. Machado, D.; Pires, D.; Perdigão, J.; Couto, I.; Portugal, I.; Martins, M.; Amaral, L.; Anes, E.; Viveiros, M. Ion channel blockers as antimicrobial agents, efflux inhibitors, and enhancers of macrophage killing activity against drug resistant Mycobacterium tuberculosis. *PLoS ONE* **2016**, *11*, e0149326. [CrossRef]
82. Viveiros, M.; Martins, M.; Rodrigues, L.; Machado, D.; Couto, I.; Ainsa, J.; Amaral, L. Inhibitors of mycobacterial efflux pumps as potential boosters for anti-tubercular drugs. *Expert Rev. Anti-Infect. Ther.* **2012**, *10*, 983–998. [CrossRef] [PubMed]
83. Martins, M.; Viveiros, M.; Couto, I.; Amaral, L. Targeting human macrophages for enhanced killing of intracellular XDR-TB and MDR-TB. *Int. J. Tuberc. Lung Dis.* **2009**, *13*, 569–573.
84. Gupta, S.; Salam, N.; Srivastava, V.; Singla, R.; Behera, D.; Khayyam, K.U.; Korde, R.; Malhotra, P.; Saxena, R.; Natarajan, K. Voltage gated calcium channels negatively regulate protective immunity to Mycobacterium tuberculosis. *PLoS ONE* **2009**, *4*, e5305. [CrossRef] [PubMed]
85. Anes, E. Acting on actin during bacterial infection. In *Cytoskeleton Structure, Dynamics, Function and Disease*; Jimenez-Lopez J.C., Ed.; IntechOpen: London, UK, 2017; Chapter 13, pp. 257–278. [CrossRef]
86. Hestvik, A.L.K.; Hmama, Z.; Av-Gay, Y. Mycobacterial manipulation of the host cell. *FEMS Microbiol. Rev.* **2005**, *29*, 1041–1050. [CrossRef] [PubMed]

87. Guérin, I.; de Chastellier, C. Pathogenic mycobacteria disrupt the macrophage actin filament network. *Infect. Immun.* **2000**, *68*, 2655–2662. [CrossRef]
88. Bettencourt, P.; Marion, S.; Pires, D.; Santos, L.; Lastrucci, C.; Carmo, N.; Blake, J.; Benes, V.; Griffiths, G.; Neyrolles, O.; et al. Actin-binding protein regulation by microRNAs as a novel microbial strategy to modulate phagocytosis by host cells: The case of N-Wasp and miR-142-3p. *Front. Cell. Infect. Microbiol.* **2013**, *3*, 19. [CrossRef]
89. Wang, J.; Yao, Y.; Wu, J.; Deng, Z.; Gu, T.; Tang, X.; Cheng, Y.; Li, G. The mechanism of cytoskeleton protein β-actin and cofilin-1 of macrophages infected by Mycobacterium avium. *Am. J. Transl. Res.* **2016**, *8*, 1055.
90. Levite, Mia. Neurotransmitters activate T-cells and elicit crucial functions via neurotransmitter receptors. *Curr. Opin. Pharmacol.* **2008**, *8*, 460–471. [CrossRef]
91. Pacheco, R.; Riquelme, E.; Kalergis, A.M. Emerging evidence for the role of neurotransmitters in the modulation of T cell responses to cognate ligands. In *Central Nervous System Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Central Nervous System Agents)*; Bentham Science Publishers: Sharjah, United Arab Emirates, 2010; Volume 10, pp. 65–83.
92. Skinner, M.A.; Parlane, N.; McCarthy, A.; Buddle, B.M. Cytotoxic T-cell responses to Mycobacterium bovis during experimental infection of cattle with bovine tuberculosis. *Immunology* **2003**, *110*, 234–241. [CrossRef]
93. Villarreal-Ramos, B.; McAulay, M.; Chance, V.; Martin, M.; Morgan, J.; Howard, C.J. Investigation of the role of CD8+ T cells in bovine tuberculosis in vivo. *Infect. Immun.* **2003**, *71*, 4297–4303. [CrossRef]
94. Pollock, J.M.; Neill, S.D. Mycobacterium boviss infection and tuberculosis in cattle. *Vet. J.* **2002**, *163*, 115–127. [CrossRef] [PubMed]
95. Finlay, E.K.; Berry, D.P.; Wickham, B.; Gormley, E.P.; Bradley, D.G. A genome wide association scan of bovine tuberculosis susceptibility in Holstein-Friesian dairy cattle. *PLoS ONE* **2012**, *7*, e30545.
96. Pacheco, R.; Gallart, T.; Lluis, C.; Franco, R. Role of glutamate on T-cell mediated immunity. *J. Neuroimmunol.* **2007**, *185*, 9–19. [CrossRef] [PubMed]
97. Ganor, Y.; Levite, M. The neurotransmitter glutamate and human T cells: Glutamate receptors and glutamate-induced direct and potent effects on normal human T cells, cancerous human leukemia and lymphoma T cells, and autoimmune human T cells. *J. Neural Transm.* **2014**, *121*, 983–1006. [CrossRef] [PubMed]
98. El Masri, R.; Delon, J. RHO GTPases: From new partners to complex immune syndromes. *Nat. Rev. Immunol.* **2021**, *21*, 499–513. [CrossRef]
99. Bokoch, G.M. Regulation of innate immunity by Rho GTPases. *Trends Cell Biol.* **2005**, *15*, 163–171. [CrossRef]
100. Chopra, P.; Koduri, H.; Singh, R.; Koul, A.; Ghildiyal, M.; Sharma, K.; Tyagi, A.K.; Singh, Y. Nucleoside diphosphate kinase of Mycobacterium tuberculosis acts as GTPase-activating protein for Rho-GTPases. *FEBS Lett.* **2004**, *571*, 212–216. [CrossRef] [PubMed]
101. Soupene, E.; Kuypers, F.A. Mammalian long-chain acyl-CoA synthetases. *Exp. Biol. Med.* **2008**, *233*, 507–521. [CrossRef] [PubMed]
102. Nys, Y.; Bain, M.; Van Immerseel, F. *Improving the Safety and Quality of Eggs and Egg Products: Volume 1: Egg Chemistry, Production and Consumption*; Elsevier: Amsterdam, The Netherlands, 2011.
103. Li, H.; Wang, T.; Xu, C.; Wang, D.; Ren, J.; Li, Y.; Tian, Y.; Wang, Y.; Jiao, Y.; Kang, X.; et al. Transcriptome profile of liver at different physiological stages reveals potential mode for lipid metabolism in laying hens. *BMC Genom.* **2015**, *16*, 763. [CrossRef] [PubMed]
104. Yu, S.; Wei, W.; Xia, M.; Jiang, Z.; He, D.; Li, Z.; Han, H.; Chu, W.; Liu, H.; Chen, J. Molecular characterization, alternative splicing and expression analysis of ACSF 2 and its correlation with egg-laying performance in geese. *Anim. Genet.* **2016**, *47*, 451–462. [CrossRef] [PubMed]
105. Tian, W.; Zheng, H.; Yang, L.; Li, H.; Tian, Y.; Wang, Y.; Lyu, S.; Brockmann, G.A.; Kang, X.; Liu, X. Dynamic expression profile, regulatory mechanism and correlation with egg-laying performance of ACSF gene family in chicken (*Gallus gallus*). *Sci. Rep.* **2018**, *8*, 8457. [CrossRef] [PubMed]
106. Lopes-Marques, M.; Cunha, I.; Reis-Henriques, M.A.; Santos, M.M.; Castro, L.F.C. Diversity and history of the long-chain acyl-CoA synthetase (Acsl) gene family in vertebrates. *BMC Evol. Biol.* **2013**, *13*, 271. [CrossRef] [PubMed]
107. Ellis, J.M.; Frahm, J.L.; Li, L.O.; Coleman, R.A. Acyl-coenzyme A synthetases in metabolic control. *Curr. Opin. Lipidol.* **2010**, *21*, 212. [CrossRef] [PubMed]
108. Brionne, A.; Nys, Y.; Hennequet-Antier, C.; Gautron, J. Hen uterine gene expression profiling during eggshell formation reveals putative proteins involved in the supply of minerals or in the shell mineralization process. *BMC Genom.* **2014**, *15*, 220. [CrossRef]
109. Jonchère, V.; Réhault-Godbert, S.; Hennequet-Antier, C.; Cabau, C.; Sibut, V.; Cogburn, L.A.; Nys, Y.; Gautron, J. Gene expression profiling to identify eggshell proteins involved in physical defense of the chicken egg. *BMC Genom.* **2010**, *11*, 57. [CrossRef]
110. Yung, L.S.; Yang, C.; Wan, X.; Yu, W. GBOOST: A GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics* **2011**, *27*, 1309–1310. [CrossRef]
111. Hemani, G.; Theocharidis, A.; Wei, W.; Haley, C. EpiGPU: Exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* **2011**, *27*, 1462–1465. [CrossRef]
112. Zhu, S.; Fang, G. MatrixEpistasis: Ultrafast, exhaustive epistasis scan for quantitative traits with covariate adjustment. *Bioinformatics* **2018**, *34*, 2341–2348. [CrossRef]

113. Chatelain, C.; Durand, G.; Thuillier, V.; Augé, F. Performance of epistasis detection methods in semi-simulated GWAS. *BMC Bioinform.* **2018**, *19*, 231. [CrossRef]
114. Niel, C.; Sinoquet, C.; Dina, C.; Rocheleau, G. A survey about methods dedicated to epistasis detection. *Front. Genet.* **2015**, *6*, 285. [CrossRef] [PubMed]
115. Jing, P.; Shen, H. MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics* **2014**, *31*, 634–641. [CrossRef] [PubMed]
116. Kim, K.H.; Kim, J.; Lim, W.; Jeong, S.; Lee, H.; Cho, Y.; Moon, J.; Kim, N. Genome-wide association and epistatic interactions of flowering time in soybean cultivar. *PLoS ONE* **2020**, *15*, e0228114. [CrossRef]
117. Cui, Z.; Yang, Q.; Zhang, H.; Zhu, Q.; Zhang, Q. Bioinformatics identification of drug resistance-associated gene pairs in Mycobacterium tuberculosis. *Int. J. Mol. Sci.* **2016**, *17*, 1417. [CrossRef]
118. Shen, J.; Li, Z.; Song, Z.; Chen, J.; Shi, Y. Genome-wide two-locus interaction analysis identifies multiple epistatic SNP pairs that confer risk of prostate cancer: A cross-population study. *Int. J. Cancer* **2017**, *140*, 2075–2084. [CrossRef] [PubMed]
119. Egli, T.; Vukojevic, V.; Sengstag, T.; Jacquot, M.; Cabezón, R.; Coynel, D.; Freytag, V.; Heck, A.; Vogler, C.; Dominique, J.; et al. Exhaustive search for epistatic effects on the human methylome. *Sci. Rep.* **2017**, *7*, 13669. [CrossRef] [PubMed]
120. González-Domínguez, J.; Schmidt, Bertil. GPU-accelerated exhaustive search for third-order epistatic interactions in case–control studies. *J. Comput. Sci.* **2015**, *8*, 93–100. [CrossRef]
121. Conway, J.R.; Lex, A.; Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **2017**, *33*, 2938–2940. [CrossRef]
122. Zhang, X.; Zhao, X.; He, K.; Lu, L.; Cao, Y.; Liu, J.; Hao, J.; Liu, Z.; Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **2012**, *28*, 98–104. [CrossRef] [PubMed]
123. Guo, X.; Zhang, H.; Tian, T. Development of stock correlation networks using mutual information and financial big data. *PLoS ONE* **2018**, *13*, e0195941. [CrossRef]
124. Mohammadi, S.; Desai, V.; Karimipour, H. Multivariate mutual information-based feature selection for cyber intrusion detection. In Proceedings of the 2018 IEEE Electrical Power and Energy Conference (EPEC), Toronto, ON, Canada, 10–11 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
125. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
126. Wu, J.; Devlin, B.; Ringquist, S.; Trucco, M.; Roeder, K. Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genet Epidemiol.* **2010**, *34*, 275–285. [CrossRef] [PubMed]
127. Wang, D.; Salah El-Basyoni, I.; Stephen Baenziger, P.; Crossa, J.; Eskridge, K.M.; Dweikat, I. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity* **2012**, *109*, 313–319. [CrossRef]

A.4. agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species

biology

MDPI

Article

agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species

Selina Klees [1,2,*,†], Felix Heinrich [1,†], Armin Otto Schmitt [1,2] and Mehmet Gültas [2,3,*]

[1] Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; felix.heinrich@uni-goettingen.de (F.H.); armin.schmitt@uni-goettingen.de (A.O.S.)
[2] Center for Integrated Breeding Research (CiBreed), Georg-August University, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany
[3] Faculty of Agriculture, South Westphalia University of Applied Sciences, Lübecker Ring 2, 59494 Soest, Germany
* Correspondence: selina.klees@uni-goettingen.de (S.K.); gueltas.mehmet@fh-swf.de (M.G.)
† These authors contributed equally to this work.

Simple Summary: Regulatory SNPs (rSNPs) are SNPs located within promoter regions that have a high potential to alter gene expression by changing the binding affinity of transcription factors to their binding sites. Such rSNPs are gaining importance in the life sciences due to their causality for specific traits and diseases. In this study, we present agReg-SNPdb, the first database comprising rSNP data of seven agricultural and domestic animal species: cattle, pig, chicken, sheep, horse, goat, and dog, and made it usable via a web interface.

Abstract: Transcription factors (TFs) govern transcriptional gene regulation by specifically binding to short DNA motifs, known as transcription factor binding sites (TFBSs), in regulatory regions, such as promoters. Today, it is well known that single nucleotide polymorphisms (SNPs) in TFBSs can dramatically affect the level of gene expression, since they can cause a change in the binding affinity of TFs. Such SNPs, referred to as regulatory SNPs (rSNPs), have gained attention in the life sciences due to their causality for specific traits or diseases. In this study, we present agReg-SNPdb, a database comprising rSNP data of seven agricultural and domestic animal species: cattle, pig, chicken, sheep, horse, goat, and dog. To identify the rSNPs, we constructed a bioinformatics pipeline and identified a total of 10,623,512 rSNPs, which are located within TFBSs and affect the binding affinity of putative TFs. Altogether, we implemented the first systematic analysis of SNPs in promoter regions and their impact on the binding affinity of TFs for livestock and made it usable via a web interface.

Keywords: single nucleotide polymorphism; regulatory SNP; transcription factor; transcription factor binding site; gene regulation; database; agricultural animal species; livestock

Citation: Klees, S.; Heinrich, F.; Schmitt, A.O.; Gültas, M. agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species. *Biology* **2021**, *10*, 790. https://doi.org/10.3390/biology10080790

Academic Editor: W. Brad Barbazuk

Received: 12 July 2021
Accepted: 12 August 2021
Published: 17 August 2021

1. Introduction

The transcriptional regulation of gene expression in higher organisms is essential for various biological processes. In contrast to the process of translation, the transcriptional machinery and its regulatory mechanisms are far from being deciphered [1]. These mechanisms are mainly governed by a special class of regulatory proteins, the transcription factors (TFs), and their combinatorial interplay [2,3]. TFs regulate the transcription as a response to specific environmental conditions by binding to short degenerate sequence motifs known as transcription factor binding sites (TFBSs) in promoter regions of their target genes and, thereby, enhance or repress gene transcription. Genomic variations, such as single nucleotide polymorphisms (SNPs), define and characterize specific populations or phenotypes and are, hence, used as markers in animal and plant breeding.

Due to the decreasing costs for whole genome sequencing, an increasing number of variants is detected followed by association studies statistically linking SNPs to specific

traits or diseases. However, the identification of causal variants and the elucidation of their regulatory roles is proceeding at a slow rate [4,5]. Today, it is well known that most disease- and trait-associated SNPs are not located within the coding regions of genes but in non-coding regions [6–9]. SNPs that are located in regulatory regions can alter TFBSs leading to a change in the binding affinity of TFs and, in extreme cases, even result in the disruption of a TFBS or the creation of a new TFBS (Figure 1) and, thus, affect gene expression. Such SNPs are referred to as regulatory SNPs (rSNPs) [10–12].

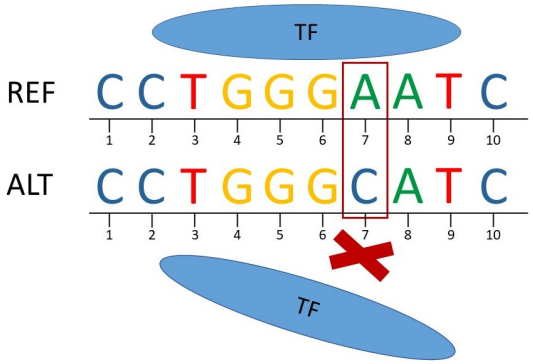


Figure 1. Scheme of the disruption of transcription factor (TF) binding due to a regulatory SNP. The TF can bind to the reference (REF) sequence while it does not bind to the alternate (ALT) sequence (C instead of A at position 7).

The importance of rSNPs has been studied extensively in humans and they are found to have a causal role for numerous traits and diseases [13–16]. A recent review on human rSNPs summarizes different rSNP studies [6]. Due to the great interest in rSNPs, several tools and databases for the analysis of the effects of SNPs on regulatory elements, e.g., TFBSs, have been developed for humans or certain model organisms. Five recent studies are summarized in Table 1, and a comprehensive overview is given in Table S1.

Recently, rSNPs are gaining attention in life sciences and animal breeding since they can be causal for specific traits and diseases and could, hence, serve as new targets for breeding. For this reason, several studies investigated the critical role of rSNPs in agriculturally important species, such as cattle [17–23], pig [24–26], and chicken [27–29]. As these studies were focused on the regulatory role of SNPs for a single trait of interest, they were highly case-specific. Thus, there still exists a lack of systematic analyses of the effects of rSNPs in agricultural species, and, until now, only a few existing tools and databases (DBs) are available for livestock.

MotifbreakR [30] and atSNP [11] are both R packages that principally include all organisms stored in the Bioconductor BSGenome package [31]; however, they require the user to supply the SNP and TFBS data (represented by position weight matrices (PWMs)), and experience in R programming is essential. The Ensembl Variant Effect Predictor (VEP) [32] stores data from experimentally supported and published rSNPs. Due to the lack of experimentally supported data of regulatory elements in livestock, the VEP mainly contains data of regulatory elements and variants for human and mouse. Therefore, the information for livestock stored in the Ensembl VEP is limited to annotations based on the position of the SNP with respect to a gene, e.g., in the upstream region or in the 5′ UTR, excluding effects on TF binding.

In order to address the limited knowledge and information available regarding the crucial functions of rSNPs and their associations with TFBSs in livestock, we systematically carried out an analysis to detect rSNPs and predicted their effects on TF binding for seven agricultural and domestic species (cattle, pig, chicken, sheep, horse, goat, and dog). In particular, we first analyzed the promoter regions (ranging from −7.5 kb to +2.5 kb) of all annotated genes and obtained the SNPs within these regions. Secondly, we extracted the flanking sequences for these SNPs and performed a TFBS prediction on the reference as well

as alternate sequences. Finally, we assigned the identified SNPs to different categories based on their consequences on TF binding (Figure 2) as suggested in [33,34]. To demonstrate our results in a proper way, we developed a database, namely agReg-SNPdb, which stores all predicted regulatory SNPs and their consequences on TF binding for each gene, and we made it accessible via a web interface (https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb, (accessed on 16 August 2021)). Furthermore, we performed a literature survey to show that our results are in agreement with previous experimental and in silico studies.

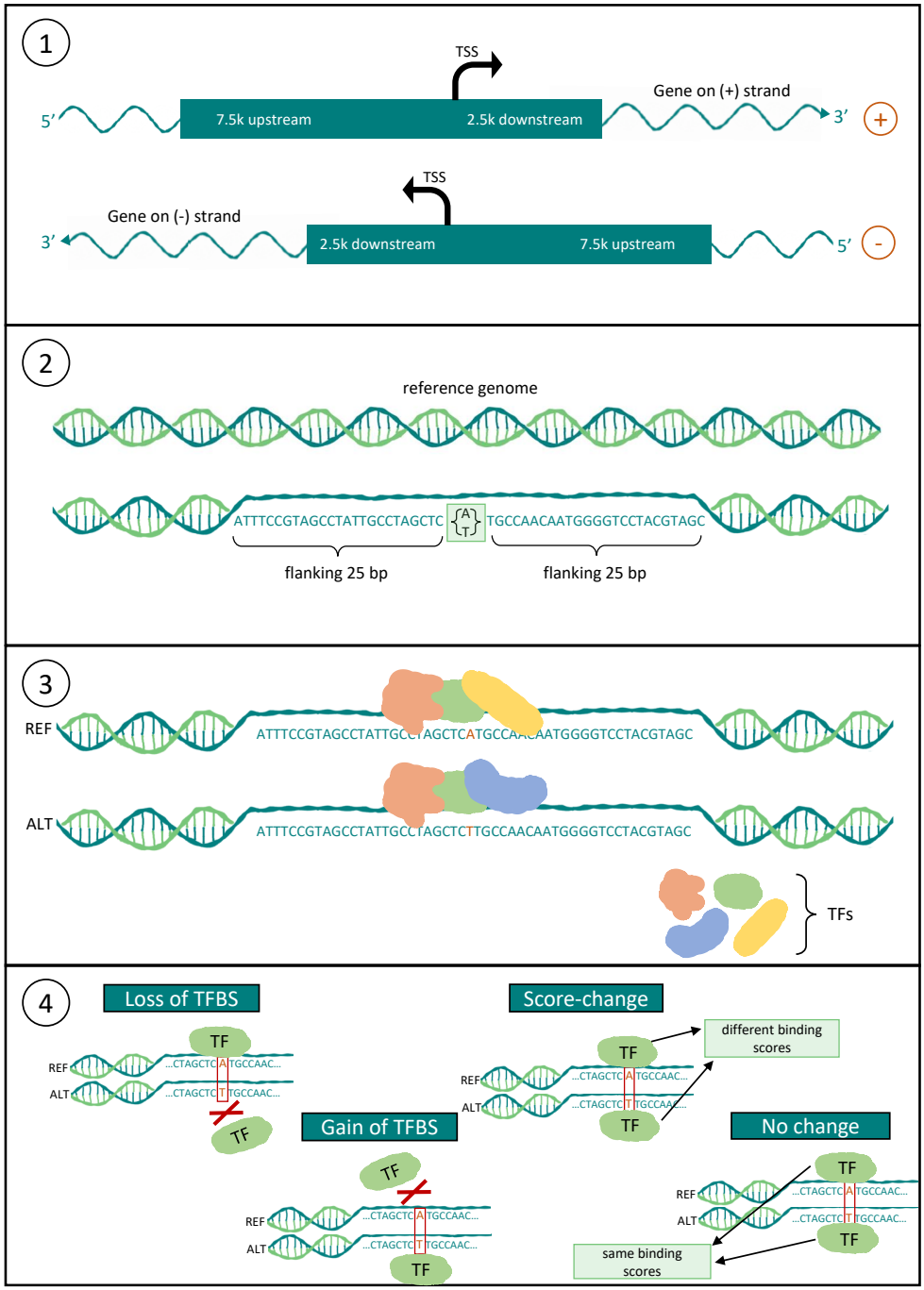


Figure 2. Scheme of the workflow applied for the detection of rSNPs. (**1**) Definition of the promoter region as 7.5 kb upstream (5′ direction) and 2.5 kb downstream (3′ direction) of the TSS, and extraction of SNPs within this region; (**2**) extraction of the flanking 25 bp around the SNPs from the reference genome; (**3**) prediction of the TFBSs for both the reference and alternate sequences; and (**4**) deriving the consequences for each SNP-TFBS pair.

Table 1. A summary of five recent studies that systematically investigated the effects of SNPs on regulatory elements, such as TFBSs. The analyses were done by either collecting experimentally supported and published data or by predicting the SNP impact on TF binding using prediction tools.

Name	Species	DB/Tool	Website	Characteristics	Experimentally Supported Data or Prediction
QBiC-Pred [35]	Human	Tool	http://qbic.genome.duke.edu (accessed on 16 August 2021)	• TFBS prediction with regression models • Prediction of changes in TF binding using ordinary least squares and evaluation of correlation between the predicted binding changes and changes in gene expression	TFBS prediction
atSNP [11] atSNP-Search [36]	Human (atSNP: organisms from Bioconductor `BSGenome` package [31])	Tool, DB	http://atsnp.biostat.wisc.edu (accessed on 16 August 2021)	• atSNP: R package for TF binding affinity testing for rSNPs (needs a SNP and motif set as input) • atSNP Search: DB for human SNP-motif pairs and the respective significance	TFBS prediction
INFERNO [37]	Human	Tool	http://inferno.lisanwanglab.org (accessed on 16 August 2021)	• Inferring causal variants from genome-wide association studies (GWAS) within annotated regulatory regions as enhancers including tissue context • TFBS prediction with HOMER	TFBS prediction
rSNPBASE [38], rSNPBASE 3.0 [10]	Human	DB	http://rsnp.psych.ac.cn (accessed on 16 August 2021) http://rsnp3.psych.ac.cn (accessed on 16 August 2021)	• DB of rSNPs with references to regulatory elements • Includes proximal and distal regulatory regions, post-transcriptional regulation, linkage disequilibrium (LD), and expression quantitative trait locus (eQTL) information • rSNPBASE 3.0 includes regulatory element-target gene pairs for regulatory networks	experimentally supported regulatory elements
SNP2TFBS [39]	Human	DB	https://ccg.epfl.ch//snp2tfbs (accessed on 16 August 2021)	• DB of human SNPs that affect TFBSs and the prediction of a consequence • DB can be downloaded as text files or accessed via the website	TFBS prediction

2. Materials and Methods

2.1. Input Data

The construction of agReg-SNPdb requires: (i) a library of PWMs representing the TFBSs and, for each animal, (ii) a reference genome, (iii) a SNP catalog, and (iv) gene annotations. As a PWM library, we used the non-redundant vertebrate matrices provided by TRANSFAC [40]. The reference genomes, SNP catalogs, and gene annotation files are downloaded from Ensembl [41]. The respective assembly versions are listed in Table 2. The SNP catalog was filtered by discarding all insertions and deletions, keeping only the SNPs. For most genes, more than one transcript isoform was annotated [32], e.g., due to different splicing variants. This ambiguity was kept during the analysis if the positions of the transcription start sites (TSSs) and, hence, the derived promoter regions were different.

Table 2. Assembly versions of the input data, including the reference genome, SNP catalog, and gene annotations. All files were downloaded from Ensembl (release 103).

Animal	Assembly Version	Download Date
Cattle	ARS-UCD1.2	1 March 2021
Pig	Sscrofa11.1	9 March 2021
Chicken	GRCg6a	25 February 2021
Sheep	Oar_rambouillet_v1.0	1 March 2021
Horse	EquCab3.0	1 March 2021
Goat	ARS1	1 March 2021
Dog	CanFam3.1	8 March 2021

2.2. Pipeline

A general workflow of the detection pipeline is shown in Figure 2. In our previous studies on faba beans [34] and rapeseed [33], we established similar pipelines for the prediction of rSNPs.

2.2.1. Detection of SNPs within the Promoter Region

The first step of this analysis was to extract SNPs, which are located within the pre-defined promoter regions. Since there exists no experimentally verified information regarding the exact location of the promoters and in order to overcome inaccuracies in TSS prediction, we chose a large promoter region of 7.5 kb upstream and 2.5 kb downstream of the TSS. Similarly large promoter regions were used in previous studies [10,37,42–48]. This promoter region can be narrowed by the user during a database search on our website. For all annotated genes, we extracted the SNPs within this region for further analysis by using the function `foverlaps` of the package `data.table` in R [49].

2.2.2. Prediction of TFBSs

For each SNP lying within a promoter region, we extracted the respective flanking sequence of 25 bp on each side of the SNP resulting in sequences with a total length of 51 bp and the SNP at position 26 (similar flanking sequences were used in [33,34,43,50]). Sequences with a length of less than 51 bp or sequences with gaps were discarded. After extracting the flanking sequences, we created two sequences per SNP, one with the reference and one with the alternate allele at the SNP position. Both were used as input for the TFBS prediction tool MATCH™ [51], which scanned the sequences to predict TFBSs using a PWM library from TRANSFAC with specific cutoff values to minimize the false positive rates. If a PWM matched a segment of genomic DNA, this sequence motif was referred to as a (potential) TFBS. As a result, the algorithm provided two scores for each predicted TFBS [40,51]: the matrix similarity score (MSS), measuring the quality of the match regarding the whole PWM sequence, and the core similarity score (CSS), measuring the quality of the match regarding the first five most-conserved consecutive positions of the PWM. Both scores were within the range [0, 1], where a score of 1 denoted an exact match of the sequence with the

PWM [51] measuring the quality of the match and indicating the binding affinity of a TF to the site.

In TRANSFAC, a PWM identifier follows a certain terminology with the structure V$*factorname_version*. In our case, each PWM starts with "V$", which indicates that the PWM originated from a vertebrate TF. The *factorname* specifies the name of the TF that is binding to the DNA motif. Since there can be several PWMs representing the sequence motif of a specific TF, the *version* was specified for unique identification [3,40].

2.2.3. Annotation of Consequences

For each SNP, we obtained two sets of predicted TFBSs—one for the reference and one for the alternate allele. By comparing these two sets, we manually determined the consequence of a SNP on a TFBS as in our previous studies [33,34]. We differentiated four different consequences: (i) no effect, (ii) change in binding affinity, (iii) loss of TFBS, and (iv) gain of TFBS. We defined two TFBS predictions as the same if their PWMs, positions, and the strand on which they were found were equal for both alleles.

A SNP was considered to have no effect on a TFBS if both scores computed by MATCH™ were equal for both alleles. A SNP was considered to cause a change in the binding affinity of a TF if the matrix similarity score computed by MATCH™ differed for the two alleles. A SNP caused a loss or gain of TFBS if the considered TFBS was only predicted for the reference or alternate sequence, respectively. In this study, we defined an rSNP as a SNP that caused a loss or gain of TFBS or a score-change for at least one TFBS.

3. Results

3.1. Database

We created the mysql database [52] agReg-SNPdb, which stores (i) general information about the SNPs, such as the ID, chromosomal position and the alleles (table *snp_info*); (ii) general information about the genes, such as the gene name and chromosomal position (table *gene_info*); (iii) the table *snp_region* connecting the tables *snp_info* and *gene_info* by storing SNPs and their corresponding target genes together with their genomic position within the promoter region based on the distance to the TSS; and, most importantly, (iv) for each SNP within a promoter region (i.e., for each SNP in table *snp_region*), we store its consequences based on the predicted TFBS binding potential (table *TFBS_results*). A summary of the number of entries for each table and animal stored in our database is shown in Table 3.

Table 3. The number of records stored in the database tables *snp_info*, *gene_info*, *snp_region*, and *TFBS_results*.

	snp_Info	gene_Info	snp_Region	TFBS_Results
Cattle	88,109,946	21,656	9,335,814	9,074,371
Pig	58,145,647	20,267	4,385,724	4,432,047
Chicken	20,917,836	16,659	3,810,524	3,901,905
Sheep	50,164,898	20,359	3,216,474	3,205,279
Horse	20,331,427	20,499	1,585,207	1,713,395
Goat	31,331,447	19,658	1,987,914	2,015,588
Dog	4,725,021	19,960	494,691	489,292
Total	273,726,222	139,058	24,816,348	24,831,877

3.2. Web Interface

The web interface (https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb, accessed on 16 August 2021) allows users to query the agReg-SNPdb without SQL knowledge and to obtain the requested results either on our website directly or by downloading them as `CSV` files. The database can be searched by (i) SNP identifiers in the form of rs numbers, (ii)

SNP positions, (iii) SNP regions in a specified chromosome, or (iv) gene identifiers, i.e., the Ensembl gene stable ID or gene name (Figure 3).

The search results will contain, at maximum, four tables: (1) a table showing general SNP information (table *snp_info*); (2) a table showing general gene information (table *gene_info*); (3) a table linking the SNPs to the genes, more specifically to the promoter regions, if they are positioned within a promoter region (table *snp_region*); and (4) for all rSNPs, a table with the predicted TFBSs overlapping each rSNP, the MATCH™ scores, and the respective consequence (table *TFBS_results*) for both alleles. An example output can be seen in Figure 4. In all tables, we provide links to sites with additional information for the SNPs and genes, and, for each PWM, we display the respective sequence logo if desired. Apart from the search site, the complete database tables can be downloaded chromosome-wise on the summary page of the respective animal.

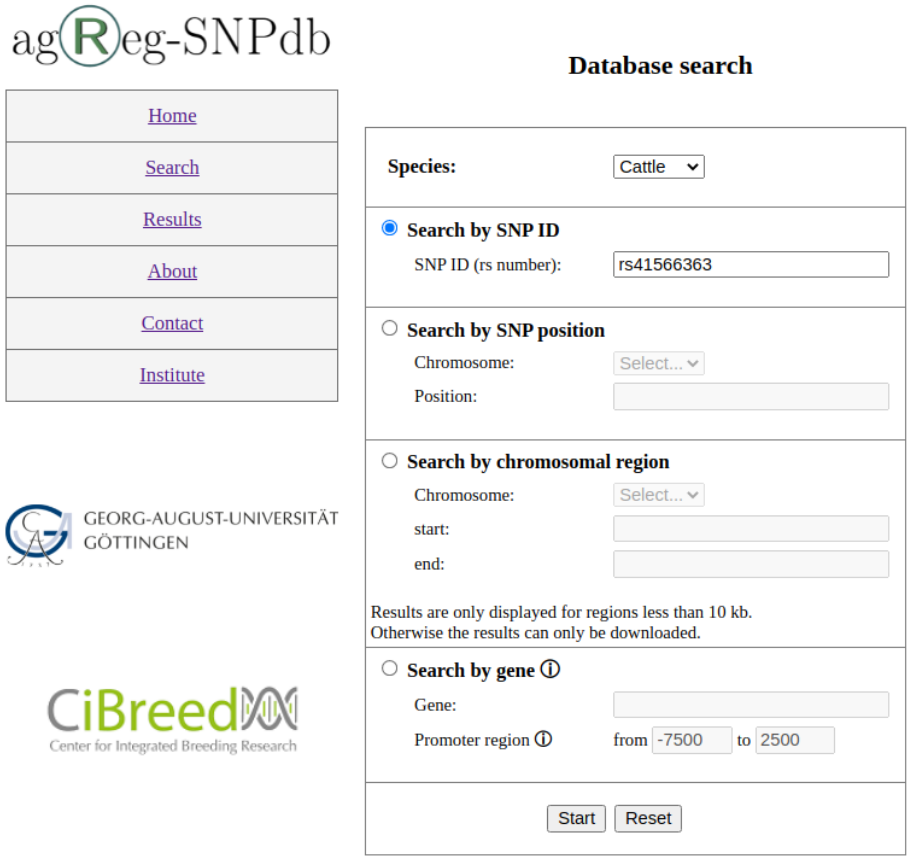


Figure 3. Search page of agReg-SNPdb. Search options are (1) by SNP ID, (2) by SNP position, (3) by chromosomal region, and (4) by gene.

3.3. Statistical Analysis of the Data

To give a brief overview of the data stored in agReg-SNPdb, we show the distribution of SNPs, genes, and rSNPs in the promoter regions along the chromosomes in an exemplary manner for the species chicken. The distributions for the remaining animals can be found in Figures S2 and S3. The distributions of SNPs and genes along the chromosomes are shown in Figure 5. As expected, the number of SNPs and genes decreased largely with increasing chromosome number and, hence, with decreasing chromosome size.

SNP information

Show 10 entries Search:

SNP_ID	Chromosome	Position	REF	ALT	Quality	Filter	INFO
rs41566363	23	23277585	G	C	.	.	ID=51111850;Variant_seq=C;evidence_values=Multiple_observations,Frequency;Dbxref=dbSNP_150:rs41566363;Reference_seq=G

Showing 1 to 1 of 1 entries Previous 1 Next

Gene information

Show 10 entries Search:

Name	Chromosome	Strand	txStart	txEnd	Name2
ENSBTAG00000020425	23	+	23277603	23337345	TFAP2D

Showing 1 to 1 of 1 entries Previous 1 Next

SNP region information

Show 10 entries Search:

SNP_ID	Gene_Name	Chromosome	Strand	txStart	txEnd	Label	Distance to TSS (bp)
rs41566363	ENSBTAG00000020425	23	+	23277603	23337345	inUpstreamPromoterRegion	-18

Showing 1 to 1 of 1 entries Previous 1 Next

Found TFBSs

Explanation of Consequences

Gain of TFBS	The TFBS exists only for the 1 (alternative) allele of the SNP
Loss of TFBS	The TFBS exists only for the 0 (reference) allele of the SNP
Score-Change	The TFBS exists for both alleles but the binding affinity differs as measured by the Core_Similarity_Score and Matrix_Similarity_Score calculated by MATCH™
No Change	The TFBS exists for both alleles with the same binding affinity

Show 10 entries Search:

SNP_ID	Allele	PWM	Position	Strand	Core_Similarity_Score	Matrix_Similarity_Score	Sequence	Consequence
rs41566363	0	V$PLZF_02	22	-	1	0.862	gcaggctagatCTTTAtcttcacaataa	Score-Change
rs41566363	1	V$PLZF_02	22	-	1	0.864	gcagcgctagatCTTTAtcttcacaataa	Score-Change
rs41566363	0	V$ZIC1_05	15	+	1	0.99	acacaCAGCAgggct	Loss of TFBS

Showing 1 to 3 of 3 entries Previous 1 Next

Figure 4. Example of a search result from agReg-SNPdb. The search was performed by the SNP id rs41566363 of cattle. The result tables contain, first, general SNP information; secondly, general gene information; thirdly, information about the SNP region, in particular the promoter region and distance to the TSS; and lastly, the overlapping TFBSs (represented by PWMs) for the SNP with predicted consequences.

Regarding the promoter regions, the number of SNPs in promoters is dependent on the number of genes (Figure 5B) for each chromosome. To overcome this dependency, we calculate the average number of rSNPs per gene in the upstream as well as the downstream promoter region. The average numbers of rSNPs for each chromosome in chicken revealed that most chromosomes had approximately 120 rSNPs per gene, while, on some chromosomes, only very few rSNPs per gene were found (Figure 6). Overall, by dividing the total number of rSNPs by the total number of genes, we identified on average 95.04 rSNPs within the promoter region (10 kb) of one gene in chicken.

To obtain further insight into the distribution of rSNPs in the promoter regions, we investigated their genomic positions relative to the TSS for the whole promoter region (−7.5 kb to +2.5 kb) and for a smaller section (−750 bp to +250 bp) for chicken (see Figure 7A,B, respectively; the figures for the remaining species are given in Figures S4). For chicken, we observed a similar finding as in our previous study on rapeseed [33] and as previously shown in rice [53]. While there are few rSNPs in close proximity to the TSS, the number of rSNPs increases with increasing distance to the TSS. Interestingly, in cattle (as well as in dogs), we observed the opposite tendency. Many rSNPs were found around, and especially directly downstream, of the TSS, while the number decreased with the distance to the TSS (the distribution of cattle rSNPs is shown in Figure 8).

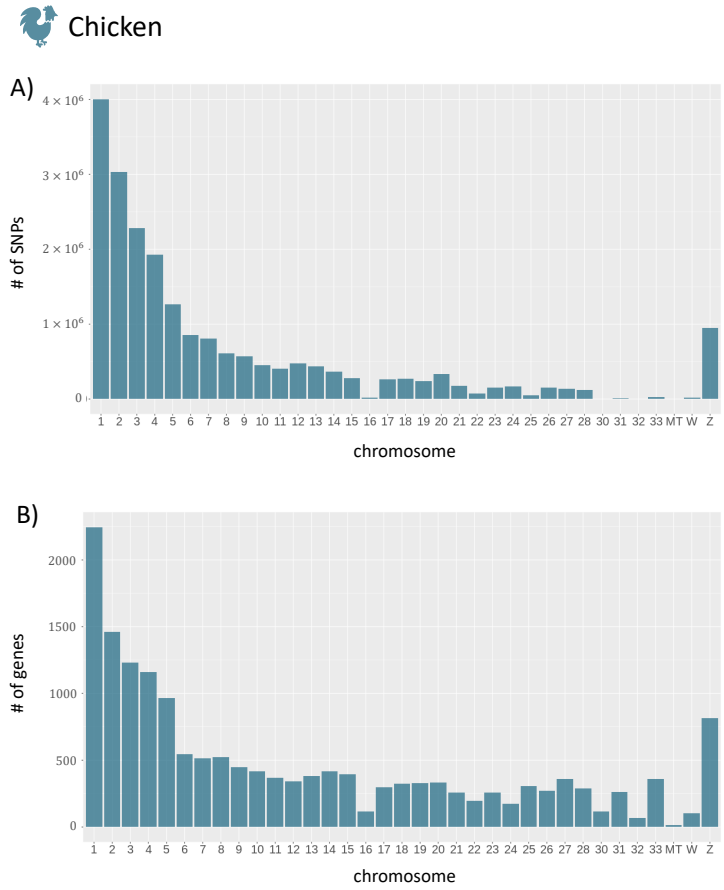


Figure 5. The total number of SNPs and genes for each chromosome of chicken. (**A**) The number of SNPs per chromosome. (**B**) The number of genes per chromosome. In total, 20,917,836 SNPs and 16,659 genes were reported. For plotting, the `R` package `ggplot2` [54] was used.

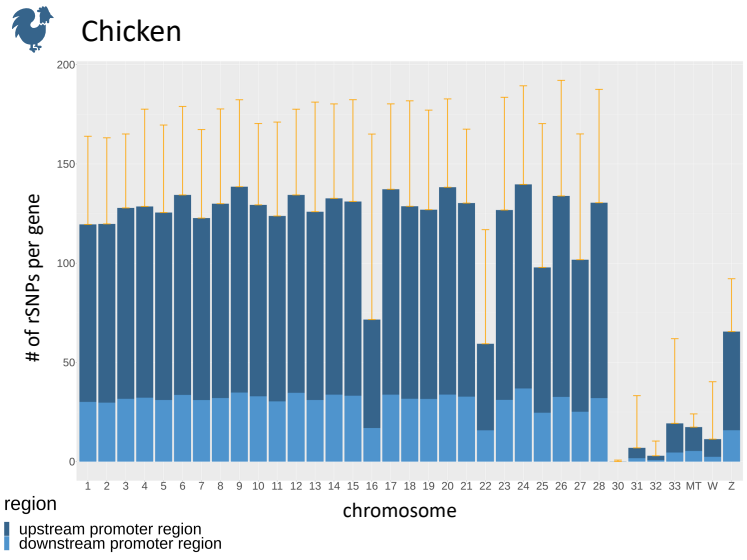


Figure 6. The average number of rSNPs in promoter regions per gene for each chromosome of chicken, divided into upstream and downstream promoters. The orange whiskers denote the mean plus one standard deviation.

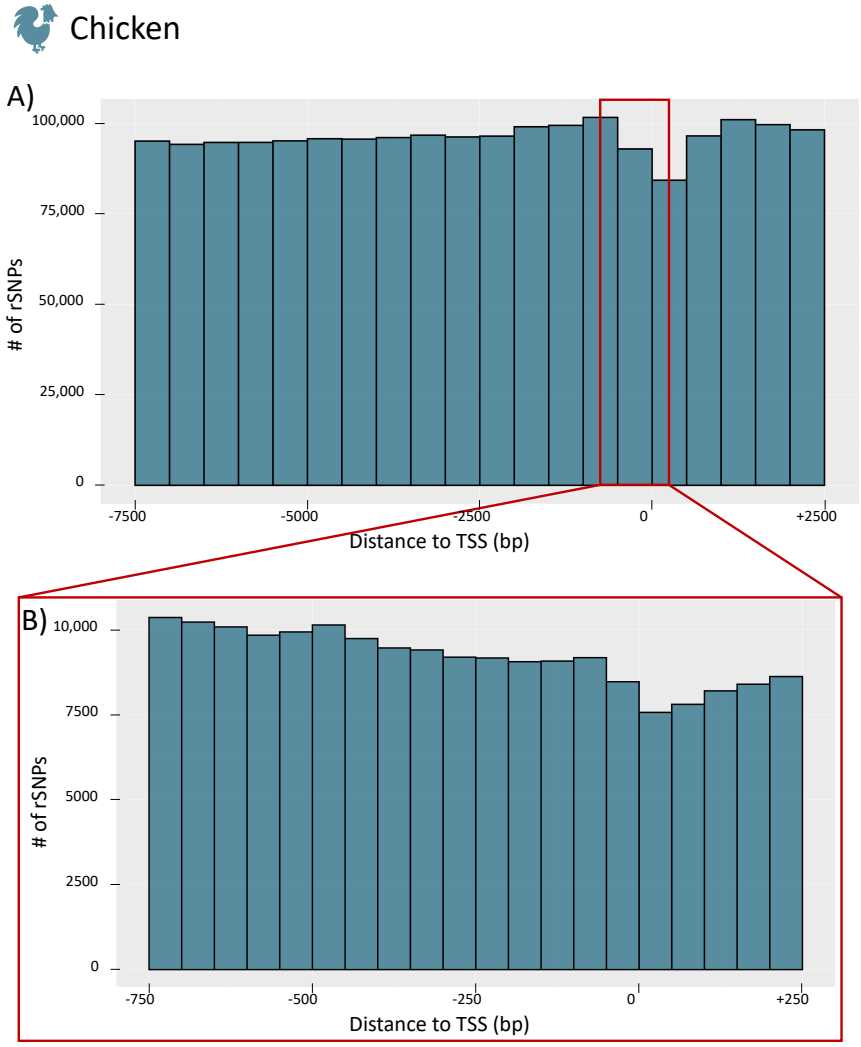


Figure 7. Distribution of the distances between rSNPs and the TSS of chicken. (**A**) The counts for the whole promoter region (−7.5 kb to +2.5 kb) in 500 bp intervals. The enlargement in (**B**) shows the proximal promoter region (−750 bp to +250 bp) in 50 bp intervals.

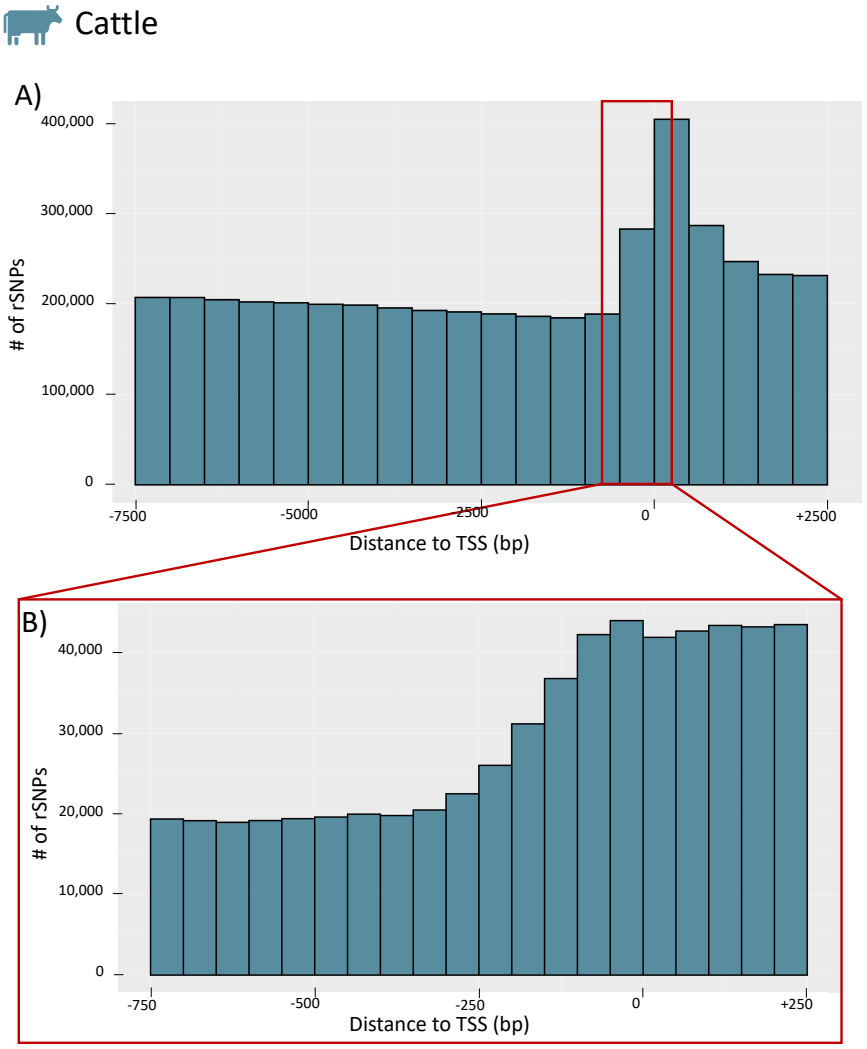


Figure 8. Distribution of the distances between rSNPs and the TSS of cattle. (**A**) The counts for the whole promoter region (−7.5 kb to +2.5 kb) in 500 bp intervals. The enlargement in (**B**) shows the proximal promoter region (−750 bp to +250 bp) in 50 bp intervals.

4. Biological Validation Based on Case-Studies

In order to validate the data stored in agReg-SNPdb, we performed literature research and assessed the importance of our findings based on selected published studies, which identified putative rSNPs that are associated with a trait under study and affect TF binding, either by prediction or as evaluated in a biological experiment.

4.1. Milk Protein and Fat Content in Dairy Cattle

Lum et al. [23] studied the molecular mechanism of different expression levels of the ß-Lactoglobulin (LGB) gene (also known as *MBLG* or *PAEP*), which plays an important role in the milk casein, protein, and fat content in dairy cattle. They described one rSNP in the *LGB* promoter with a G to C conversion 450 bp upstream of the TSS that was found within an activator protein-2 (AP-2) binding site. Measuring the different AP-2 binding affinities with DNase-I footprinting, they measured increased protein binding in the A promoter (G allele).

In our database, we identified the same rSNP (rs41255679, C/G), which was located in the proximal upstream promoter region of *PAEP* and caused a gain of the AP-2 binding

site with the G allele (Table 4) [55]. This supports the findings of different studies reporting that AP-2 binding as well as *LGB* gene expression is enhanced by the G allele and that rs41255679 could be an important regulator of *LGB* expression [23,55–57].

Table 4. Consequences of SNP rs41255679 (C/G), located upstream of the TSS of the bovine *LGB* gene. Allele 0 refers to a predicted TFBS in the reference sequence, while allele 1 stands for the alternate allele. A SNP causes a loss of TFBS if the considered TFBS (represented by a PWM) is only predicted for the reference allele. Consequently, a SNP causes a gain of TFBS if the TFBS is only predicted for the alternate allele.

SNP ID	Allele	PWM	Consequence
rs41255679	0	V$CTCF_01	Loss of TFBS
rs41255679	1	V$AP2ALPHA_03	Gain of TFBS

4.2. Fat-Related Beef Quality Traits in Cattle

Matsumoto et al. [19] investigated the role of different bovine fat-related genes, including the gene encoding the fatty acid-binding protein 4 (FABP4). Within the *FABP4* upstream promoter, they identified two SNPs in linkage disequilibrium (*FABP4* g.-295A>G and *FABP4* g.-287A>G) that were associated with several fat-related traits, such as the carcass weight and beef marbling score. Using TFSEARCH [58], they predicted TFBSs overlapping the SNPs and altering their binding sites. In agReg-SNPdb, we identified two SNPs within the *FABP4* promoter region at a distance of 8 bp to each other and A to G conversions (respectively, T to C conversions, due to the gene's location on the minus strand).

For the first SNP rs110055647, located 123 bp upstream of the TSS, we predicted a loss of TFBS for the Sex-Determining Region Y Protein (SRY) binding site, which is in line with the results of Matsumoto et al. [19]. For the neighboring rs109682576 (-115 bp from the TSS), we did not observe the CCAAT/enhancer-binding protein beta (cEBP/β) binding site predicted in their study; however, the TFBSs for Zinc finger proteins 333 (ZNF333) and 105 (ZFP105) were lost with the alternate allele, which can be seen as an extension to the results of Matsumoto et al. (Table 5) [19].

Table 5. Consequences of the SNPs rs110055647 and rs109682576 in the bovine *FABP4* upstream promoter with a T to C conversion. Allele 0 refers to a predicted TFBS in the reference sequence, while allele 1 stands for the alternate allele. A SNP causes a loss or gain of TFBS if the considered TFBS is only predicted for the reference or alternate allele, respectively. A SNP is considered to cause a score-change if the TFBS is predicted on both alleles (0,1) with a difference in the matrix similarity score computed by MATCH™ .

SNP ID	Allele	PWM	Consequence
rs110055647	0,1	V$RHOX11_01	Score-Change
rs110055647	0	V$SRY_Q6	Loss of TFBS
rs109682576	0	V$ZNF333_01	Loss of TFBS
rs109682576	0	V$ZFP105_04	Loss of TFBS

4.3. Chicken Egg Production

The prolactin (*PRL*) gene product is considered as an important reproductive hormone involved in diverse biological functions in vertebrates. In laying hens, it is an important regulator of egg production since an increased PRL secretion induces broodiness behaviour [28]. Liang et al. [29] examined the *PRL* 5′ promoter region and, using several populations of Chinese native Yuehuang, Taihe Silkie, and White Leghorn Layer chickens, they identified different rSNPs overlapping the predicted binding sites, including GATA-binding factor 1 (GATA-1), nuclear factor 1 (NF-1), and activator protein 1 (AP-1). Particularly for SNP rs313497646 (A/G conversion, 2048 bp upstream of the TSS), we

observed the same pattern with respect to TF binding in agReg-SNPdb: only the A allele allows the binding of the NF-1 factor.

Furthermore, it has been shown that the pituitary transcription factor 1 (PIT-1) is an important activator of the *PRL* gene expression [28,29,59]. In agReg-SNPdb, we store a SNP (rs731078272, G/T), located -3086 bp from the TSS and causing a loss of the PIT-1 binding site in the T allele. This result suggests that this SNP might be an important regulator of *PRL* expression where the T variant could repress *PRL* expression, which is an important indication for further studies.

4.4. Fatty-Acid Composition Related Traits in Pigs

Ballester et al. [24] studied the expression of apolipoprotein (apo-) A-II (APOA2), a protein involved in the triglyceride, fatty acid, and glucose metabolisms, and identified several SNPs associated with *APOA2* gene expression and fatty acid composition traits. Four SNPs were located in the promoter region (rs322246820, rs335066625, rs339777757, and rs333406887), among which they only found one (rs333406887, C/G) influencing a predicted TFBS—in this case, a NF-1 binding site.

Similar to their result, in agReg-SNPdb, we found the SNP rs333406887 overlapping TFBSs, such as the NF-1 binding site. Furthermore, in addition to the reported change in the binding score for NF-1, we can predict several other TFBSs that are affected by this SNP. It causes, for instance, a loss of TFBS for the kruppel-like factor 6 (also called CPBP) and a gain of TFBS for zinc finger protein X-linked (ZFX) (Table 6).

Table 6. Consequences of the SNP rs333406887 (C/G) located -238 bp from the porcine *APOA2* TSS. Allele 0 refers to a predicted TFBS in the reference sequence, while allele 1 stands for the alternate allele. A SNP causes a loss or gain of TFBS if the considered TFBS is only predicted for the reference or alternate allele, respectively. A SNP is considered to cause a score-change if the TFBS is predicted on both alleles (0,1) with a difference in the matrix similarity score computed by MATCH™ .

SNP ID	Allele	PWM	Consequence
rs333406887	0,1	V$NF1_Q6	Score-Change
rs333406887	0,1	V$AP2ALPHA_03	Score-Change
rs333406887	0	V$CPBP_Q6	Loss of TFBS
rs333406887	1	V$ZFX_01	Gain of TFBS

5. Discussion

Today, it is widely known that protein–DNA interactions govern the level of gene expression in all higher organisms to a great extent. The binding of TFs to the DNA mainly occurs in the regulatory regions, such as promoters, which are found close to the transcription start of genes [60]. The effect of rSNPs on the binding of TFs has been studied extensively in single case studies in different species, and, for humans, many tools and databases exist to facilitate these analyses (see Tables 1 and S1).

However, there is limited information available for livestock, and, to the best of our knowledge, there is no comparable data source for evaluating the effect of rSNPs. To address this lack of information, we systematically carried out a genome-wide analysis to detect rSNPs and to evaluate their consequences for TF-binding in seven animal species, which can be accessed via a web server. We showed that, by substituting a single base in a predicted TFBS, a SNP can lead to a major change in the binding affinity of the TF and, in an extreme case, even result in the disruption of the TFBS or the creation of a new TFBS.

These predictions can be of great use for scientists who have conducted: (i) an association analysis and want to reveal the underlying mechanisms caused by a SNP being significantly associated with a trait (e.g., in [19,23,33,34]); (ii) a gene expression experiment and want to identify candidate SNPs influencing the expression rate of a specific gene or a set of genes (e.g., in [24,29,33]); or (iii) a combination of both, i.e., an expression quantitative trait locus (eQTL) analysis (e.g., in [17]).

Even though our predictions are in line with many biologically tested results, as shown in the biological validation in Section 4, we note that the binding affinity of the TFs to the DNA sequence is one of the most important factors for TF binding but might not be sufficient for in vivo binding in higher organisms. Other influencing factors might include the chromatin accessibility, TF concentration, or other enhancing or repressing protein-DNA interactions, such as competitive or cooperative TF binding [3,39,61], which could not be considered in the prediction pipeline.

TF binding often occurs in a complex interplay and also includes cooperation between proximal and distal regulatory elements (promoters and enhancers) [2]. Thus, in addition to the binding of TFs in the proximal promoter regions, regulatory processes via TF-DNA interactions are also controlled by distal enhancer regions. Due to the limited knowledge of enhancer regions in livestock species, we could not incorporate these distal regulatory regions.

For our analysis pipeline, we defined a relatively wide promoter region of 7.5 kb upstream to 2.5 kb downstream of the TSS. Similarly large promoter regions were defined in previous studies ranging from 10 kb upstream to 10 kb downstream of the TSS [10,37,42–48] in order to overcome inaccuracies in the TSS prediction [53] and to ensure the inclusion of the biological promoter. The user has to be aware that the biological promoter region is usually smaller [53], and our website gives the opportunity to filter for smaller, user-defined promoter regions for each single gene. These considered promoter regions and the definition of rSNPs in our study (see Section 2.2.3) led to a relatively large number of rSNPs per gene—for instance, an average of 95.04 rSNPs per gene in chicken.

Interestingly, our results regarding the distribution of genome-wide rSNPs relative to the TSS showed two different patterns. In chicken, pig, sheep, horse, and goat, we observed that the region around the TSS was rather protected from sequence variations (Figure 7) as it was found in previous studies [33,53]. However, the data for cattle and dogs revealed a different picture, and we found an accumulation of SNPs and rSNPs around the TSS (Figure 8). This observation shows that the data stored in public databases, such as Ensembl, can show completely different patterns for different species, which could create biases for specific analyses.

6. Conclusions

To the best of our knowledge, agReg-SNPdb is the first database of regulatory SNPs for animal species of agricultural importance. It allows the users to investigate the predicted effect of an allele change on TF binding. The release of the database is an important step toward the understanding of gene regulation in the life sciences. Knowing whether a SNP causes a change in the binding affinity or even disrupts a TFBS or creates a new TFBS can be of predominant importance in order to interpret the results, from, e.g., GWAS experiments, gene expression experiments, or population studies.

The newly gained information can be used to help in genomic selection and marker establishment by identifying possibly causal rSNPs and revealing the underlying regulatory mechanisms of specific traits or diseases. Due to the regular updates of genomes as well as gene and SNP annotations, the database will be updated regularly, and, as future work, we will include several plant species with agricultural importance in agReg-SNPdb.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/biology10080790/s1, Table S1: A comprehensive overview of recent studies that investigated the effects of SNPs on regulatory elements (extension of Table 1), Figure S2: Number of SNPs and genes per chromosomes for all species, Figure S3: The average numbers of rSNPs per gene for each chromosome for all species, Figure S4: Distribution of the distances between rSNPs and the TSS for all species.

Author Contributions: M.G. designed and supervised the research. S.K. and F.H. participated in the design of the study. S.K., F.H., M.G. and A.O.S. conducted the computational and statistical analyses. S.K. performed the biological validation. F.H. created the website. S.K. and F.H. created the database.

S.K., F.H., and M.G. wrote the final version of the manuscript. M.G. conceived and managed the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb (accessed on 16 August 2021).

Acknowledgments: We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SNP	single nucleotide polymorphism
rSNP	regulatory SNP
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site
bp	base pair
eQTL	expression quantitative trait locus
LD	linkage disequilibrium
GWAS	genome-wide association study
PWM	position weight matrix
SQL	Structured Query Language

References

1. Franco-Zorrilla, J.M.; López-Vidriero, I.; Carrasco, J.L.; Godoy, M.; Vera, P.; Solano, R. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 2367–2372.
2. Steuernagel, L.; Meckbach, C.; Heinrich, F.; Zeidler, S.; Schmitt, A.O.; Gültas, M. Computational identification of tissue-specific transcription factor cooperation in ten cattle tissues. *PLoS ONE* **2019**, *14*, e0216475.
3. Meckbach, C.; Wingender, E.; Gültas, M. Removing background co-occurrences of transcription factor binding sites greatly improves the prediction of specific transcription factor cooperations. *Front. Genet.* **2018**, *9*, 189.
4. Hayes, B.J.; Daetwyler, H.D. 1000 Bull Genomes project to map simple and complex genetic traits in cattle: Applications and outcomes. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 89–102.
5. Schmitt, A.O.; Aßmus, J.; Bortfeldt, R.H.; Brockmann, G.A. CandiSNPer: A web tool for the identification of candidate SNPs for causal variants. *Bioinformatics* **2010**, *26*, 969–970.
6. Degtyareva, A.O.; Antontseva, E.V.; Merkulova, T.I. Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *Int. J. Mol. Sci.* **2021**, *22*, 6454.
7. Rojano, E.; Seoane, P.; Ranea, J.A.; Perkins, J.R. Regulatory variants: From detection to predicting impact. *Briefings Bioinform.* **2018**, *20*, 1639–1654.
8. Goodswen, S.J.; Gondro, C.; Watson-Haigh, N.S.; Kadarmideen, H.N. FunctSNP: An R package to link SNPs to functional knowledge and dbAutoMaker: A suite of Perl scripts to build SNP databases. *BMC Bioinform.* **2010**, *11*, 311.
9. Günther, T.; Schmitt, A.O.; Bortfeldt, R.H.; Hinney, A.; Hebebrand, J.; Brockmann, G.A. Where in the genome are significant single nucleotide polymorphisms from genome-wide association studies located? *Omics J. Integr. Biol.* **2011**, *15*, 507–512.
10. Guo, L.; Wang, J. rSNPBase 3.0: An updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks. *Nucleic Acids Res.* **2017**, *46*, D1111–D1116.
11. Zuo, C.; Shin, S.; Keleş, S. atSNP: Transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **2015**, *31*, 3353–3355.
12. Macintyre, G.; Bailey, J.; Haviv, I.; Kowalczyk, A. is-rSNP: A novel technique for in silico regulatory SNP detection. *Bioinformatics* **2010**, *26*, i524–i530.
13. Buroker, N.E. VEGFA rSNPs, transcriptional factor binding sites and human disease. *J. Physiol. Sci.* **2014**, *64*, 73–76.
14. Fang, L.; Ahn, J.K.; Wodziak, D.; Sibley, E. The human lactase persistence-associated SNP- 13910* T enables in vivo functional persistence of lactase promoter–reporter transgene expression. *Hum. Genet.* **2012**, *131*, 1153–1159.

15. De Gobbi, M.; Viprakasit, V.; Hughes, J.R.; Fisher, C.; Buckle, V.J.; Ayyub, H.; Gibbons, R.J.; Vernimmen, D.; Yoshinaga, Y.; De Jong, P.; et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **2006**, *312*, 1215–1217.
16. Grant, S.F.; Reid, D.M.; Blake, G.; Herd, R.; Fogelman, I.; Ralston, S.H. Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I α 1 gene. *Nat. Genet.* **1996**, *14*, 203.
17. Littlejohn, M.D.; Tiplady, K.; Fink, T.A.; Lehnert, K.; Lopdell, T.; Johnson, T.; Couldrey, C.; Keehan, M.; Sherlock, R.G.; Harland, C.; et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. *Sci. Rep.* **2016**, *6*, 1–14.
18. Muhaghegh-Dolatabady, M. Single Nucleotide Polymorphism in the Promoter Region of Bovine Interleukin 8 Gene and its Association with Milk Production Traits and Somatic Cell Score of Holstein Cattle in Iran. *Iran. J. Biotechnol.* **2014**, *12*, 36–41.
19. Matsumoto, H.; Nogi, T.; Tabuchi, I.; Oyama, K.; Mannen, H.; Sasazaki, S. The SNPs in the promoter regions of the bovine FADS2 and FABP4 genes are associated with beef quality traits. *Livest. Sci.* **2014**, *163*, 34–40.
20. Alexandre, P.A.; Gomes, R.C.; Santana, M.H.; Silva, S.L.; Leme, P.R.; Mudadu, M.A.; Regitano, L.C.; Meirelles, F.V.; Ferraz, J.B.; Fukumasu, H. Bovine NR1I3 gene polymorphisms and its association with feed efficiency traits in Nellore cattle. *Meta Gene* **2014**, *2*, 206–217.
21. Kühn, C.; Thaller, G.; Winter, A.; Bininda-Emonds, O.R.; Kaupe, B.; Erhardt, G.; Bennewitz, J.; Schwerin, M.; Fries, R. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. *Genetics* **2004**, *167*, 1873–1881.
22. Ordovas, L.; Roy, R.; Pampín, S.; Zaragoza, P.; Osta, R.; Rodriguez-Rey, J.C.; Rodellar, C. The g. 763G> C SNP of the bovine FASN gene affects its promoter activity via Sp-mediated regulation: Implications for the bovine lactating mammary gland. *Physiol. Genom.* **2008**, *34*, 144–148.
23. Lum, L.S.; Dovč, P.; Medrano, J.F. Polymorphisms of bovine β-lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor. *J. Dairy Sci.* **1997**, *80*, 1389–1397.
24. Ballester, M.; Revilla, M.; Puig-Oliveras, A.; Marchesi, J.; Castello, A.; Corominas, J.; Fernandez, A.; Folch, J. Analysis of the porcine APOA 2 gene expression in liver, polymorphism identification and association with fatty acid composition traits. *Anim. Genet.* **2016**, *47*, 552–559.
25. Ryan, M.T.; Hamill, R.M.; O'Halloran, A.M.; Davey, G.C.; McBryan, J.; Mullen, A.M.; McGee, C.; Gispert, M.; Southwood, O.I.; Sweeney, T. SNP variation in the promoter of the PRKAG3 gene and association with meat quality traits in pig. *BMC Genet.* **2012**, *13*, 66.
26. Wyszyńska-Koko, J.; Pierzchała, M.; Flisikowski, K.; Kamyczek, M.; Różycki, M.; Kurył, J. Polymorphisms in coding and regulatory regions of the porcine MYF6 and MYOG genes and expression of the MYF6 gene in m. longissimus dorsi versus productive traits in pigs. *J. Appl. Genet.* **2006**, *47*, 131–138.
27. Barkova, O.Y.; Sazanova, K.A.; Fomichev, K.A.; Malewski, T.; Parada, R.; Kawka, M.; Jaszczak, K.; Sazanov, A.A. Associations of new rSNPs with eggshell thickness in Rhode Island layers. *Anim. Sci. Pap. Rep.* **2013**, *31*, 165–172.
28. Cui, J.X.; Du, H.L.; Liang, Y.; Deng, X.M.; Li, N.; Zhang, X.Q. Association of polymorphisms in the promoter region of chicken prolactin with egg production. *Poult. Sci.* **2006**, *85*, 26–31.
29. Liang, Y.; Cui, J.; Yang, G.; Leung, F.C.; Zhang, X. Polymorphisms of 5′ flanking region of chicken prolactin gene. *Domest. Anim. Endocrinol.* **2006**, *30*, 1–16.
30. Coetzee S.G.; Coetzee, G.A.; Hazelett, D.J. motifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **2015**, *31*, 3847–3849.
31. Pagès, H. BSgenome: Infrastructure for Biostrings-based genome data packages and support for efficient SNP representation. *R Package* **2016**, *1*, 10–18129.
32. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. *Genome Biol.* **2016**, *17*, 122.
33. Klees, S.; Lange, T.M.; Bertram, H.; Rajavel, A.; Schlüter, J.S.; Lu, K.; Schmitt, A.O.; Gültas, M. In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in *Brassica napus* L. Using Multi-Omics Data. *Int. J. Mol. Sci.* **2021**, *22*, 789.
34. Heinrich, F.; Wutke, M.; Das, P.P.; Kamp, M.; Gültas, M.; Link, W.; Schmitt, A.O. Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning. *Genes* **2020**, *11*, 614.
35. Martin, V.; Zhao, J.; Afek, A.; Mielko, Z.; Gordân, R. QBiC-Pred: Quantitative predictions of transcription factor binding changes due to sequence variants. *Nucleic Acids Res.* **2019**, *47*, W127–W135.
36. Shin, S.; Hudson, R.; Harrison, C.; Craven, M.; Keleş, S. atSNP Search: A web resource for statistically evaluating influence of human genetic variation on transcription factor binding. *Bioinformatics* **2018**, *35*, 2657–2659.
37. Amlie-Wolf, A.; Tang, M.; Mlynarski, E.E.; Kuksa, P.P.; Valladares, O.; Katanic, Z.; Tsuang, D.; Brown, C.D.; Schellenberg, G.D.; Wang, L.-S. INFERNO: Inferring the molecular mechanisms of noncoding genetic variants. *Nucleic Acids Res.* **2018**, *46*, 8740–8753.
38. Guo, L.; Du, Y.; Chang, S.; Zhang, K.; Wang, J. rSNPBase: A database for curated regulatory SNPs. *Nucleic Acids Res.* **2013**, *42*, D1033–D1039.
39. Kumar, S.; Ambrosini, G.; Bucher, P. SNP2TFBS-a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **2016**, *45*, D139–D144.

40. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings Bioinform.* **2008**, *9*, 326–332.
41. Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; et al. Ensembl 2020. *Nucleic Acids Res.* **2020**, *48*, D682–D688.
42. Ryan, N.M.; Morris, S.W.; Porteous, D.J.; Taylor, M.S.; Evans, K.L. SuRFing the genomics wave: An R package for prioritising SNPs by functionality. *Genome Med.* **2014**, *6*, 79.
43. Fu, Y.; Liu, Z.; Lou, S.; Bedford, J.; Mu, X.J.; Yip, K.Y.; Khurana, E.; Gerstein, M. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **2014**, *15*, 480.
44. Riva, A. Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genom. Biomed Cent.* **2012**, *13*, S7.
45. Kwon, A.T.; Arenillas, D.J.; Hunt, R.W.; Wasserman, W.W. oPOSSUM-3: Advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 Genes Genomes Genet.* **2012**, *2*, 987–1002.
46. Coetzee, S.G.; Rhie, S.K.; Berman, B.P.; Coetzee, G.A.; Noushmehr, H. FunciSNP: An R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.* **2012**, *40*, e139.
47. Ho Sui, S.J.; Mortimer, J.R.; Arenillas, D.J.; Brumm, J.; Walsh, C.J.; Kennedy, B.P.; Wasserman, W.W. oPOSSUM: Identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.* **2005**, *33*, 3154–3164.
48. Stepanova, M.; Tiazhelova, T.; Skoblov, M.; Baranova, A. A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics* **2005**, *21*, 1789–1796.
49. Dowle, M.; Srinivasan, A.; Gorecki, J.; Chirico, M.; Stetsenko, P.; Short, T.; Lianoglou, S.; Antonyan, E.; Bonsch, M.; Parsonage, H.; et al. Package 'data. table'. *Ext. Data Fram.* **2019**, *1*.
50. Xu, Z.; Taylor, J.A. SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* **2009**, *37*, W600–W605.
51. Kel, A.E.; Gößling, E.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**, *31*, 3576–3579.
52. DuBois, P. *MySQL*; Pearson Education: London, UK, 2008.
53. Triska, M.; Solovyev, V.; Baranova, A.; Kel, A.; Tatarinova, T.V. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PLoS ONE* **2017**, *12*, e0187243.
54. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
55. Gambra, R.; Peñagaricano, F.; Kropp, J.; Khateeb, K.; Weigel, K.; Lucey, J.; Khatib, H. Genomic architecture of bovine κ-casein and β-lactoglobulin. *J. Dairy Sci.* **2013**, *96*, 5333–5343.
56. Schopen, G.; Visker, M.; Koks, P.; Mullaart, E.; Van Arendonk, J.; Bovenhuis, H. Whole-genome association study for milk protein composition in dairy cattle. *J. Dairy Sci.* **2011**, *94*, 3148–3158.
57. Kuss, A.; Gogol, J.; Geldermann, H. Associations of a polymorphic AP-2 binding site in the 5′-flanking region of the bovine β-lactoglobulin gene with milk proteins. *J. Dairy Sci.* **2003**, *86*, 2213–2218.
58. Heinemeyer, T.; Wingender, E.; Reuter, I.; Hermjakob, H.; Kel, A.E.; Kel, O.; Ignatieva, E.V.; Ananko, E.A.; Podkolodnaya, O.A.; Kolpakov, F.; et al. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.* **1998**, *26*, 362–367.
59. Nelson, C.; Albert, V.R.; Elsholtz, H.P.; Lu, L.; Rosenfeld, M.G. Activation of cell-specific expression of rat growth hormone and prolactin genes by a common transcription factor. *Science* **1988**, *239*, 1400–1405.
60. Meckbach, C.; Tacke, R.; Hua, X.; Waack, S.; Wingender, E.; Gültas, M. PC-TraFF: Identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinform.* **2015**, *16*, 400.
61. Hughes, T.R. *A Handbook of Transcription Factors*; Springer: Dordrecht, The Netherlands, 2011; Volume 52.

A.5. Master thesis: Assoziationsstudie zum Vicingehalt bei *Vicia faba* basierend auf Genotyping by Sequencing-Daten

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

ISSN 1612-6793
Nr. ZFI-BM-201x-xx

Masterarbeit

im Studiengang "Angewandte Informatik"

Assoziationsstudie zum Vicingehalt bei *Vicia faba* basierend auf Genotyping by Sequencing-Daten

Felix Heinrich

Institut für Informatik

Bachelor- und Masterarbeiten
des Zentrums für angewandte Informatik
an der Georg-August-Universität Göttingen

20. August 2018

Georg-August-Universität Göttingen
Institut für Informatik

Goldschmidtstraße 7
37077 Göttingen
Germany

☎ +49 (551) 39-172000
FAX +49 (551) 39-14403
✉ office@informatik.uni-goettingen.de
🌍 www.informatik.uni-goettingen.de

Erstbetreuer: Prof. Dr. Armin Schmitt
Zweitbetreuer: Dr. Mehmet Gültas

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Göttingen, den 20. August 2018

Zusammenfassung

Der Vicin-Gehalt von Ackerbohnen ist ein wichtiger Einflussfaktor auf ihre Eignung als Tierfutter. Bislang gibt es aber nur wenige Sorten, die einen niedrigen Vicin-Gehalt aufweisen, und es gibt bislang nur einen genetischen Marker, der eine Vorhersage des Vicin-Gehaltes in einigen Fällen erlaubt. Weitere Marker sind notwendig, um eine Feinkartierung des verantwortlichen Genes durchzuführen und eine genaue Vorhersage allgemein zu ermöglichen.

Ziel dieser Arbeit ist es, auf Basis von Genotyping-By-Sequencing-Daten von Ackerbohnen SNPs zu identifizieren und dann mit einer genomweiten Assoziationsstudie unter diesen SNPs geeignete Marker für den Vicin-Gehalt zu bestimmen.

Dazu analysieren wir Sequenzierungsdaten von 20 Pflanzen aus verschiedenen Linien mit unterschiedlichen Verfahren zur Identifikation von SNPs.

Die genomweite Assoziationsstudie wird dann einerseits mit dem klassischen Verfahren unter Verwendung des Programmes `PLINK` *durchgeführt und dann ebenfalls mit einem von uns entwickelten Verfahren auf Basis der Informationstheorie.*

Aus den auf diese Art bestimmten statistisch signifikanten SNPs werden danach die Marker ausgewählt, die sich auf den Bereich im Genom kartieren lassen, in welchem der verantwortliche Locus vermutet wird.

Inhaltsverzeichnis

1 Grundlagen **1**

1.1 *Vicia faba* . . . 1
1.2 Das Projekt Abo-Vici . . . 2
1.3 Vicin . . . 2
1.3.1 Vicin in der Tierfütterung . . . 3
1.3.2 Genetische Marker für Vicin . . . 3
1.4 Genotyping by Sequencing . . . 4
1.5 Genetische Marker . . . 5
1.6 Verwendete Daten . . . 6
1.6.1 Pflanzen . . . 6
1.6.2 Sequenzierungsdaten . . . 7

2 Identifizierung von genomischen Varianten in *Vicia faba* **9**

2.1 Variant calling mit `Bowtie2` und `Samtools` . . . 9
2.1.1 Referenzgenom . . . 9
2.1.2 `Bowtie2` . . . 12
2.1.3 `Samtools` . . . 13
2.2 `Stacks` . . . 14
2.2.1 Vorbereitung der Daten . . . 15
2.2.2 Ausführung der Pipeline . . . 16
2.3 Konvertierung der Daten zur weiteren Analyse . . . 18
2.3.1 Konversion der Daten von `Stacks` . . . 18
2.3.2 Konversion der Daten von `Bowtie2` und `Samtools` . . . 19
2.4 Ergebnisse . . . 21
2.4.1 `Bowtie2` und `Samtools` . . . 21
2.4.2 `Stacks` . . . 22

3 Überprüfung der Genotypen und Assoziationsanalyse **25**

3.1 Filterung der SNPs nach ihren Qualitätscores . . . 25

3.2 Distanzen 27
3.3 Genomweite Assoziationsstudie 28
3.3.1 PLINK 29
3.3.2 Jensen-Shannon-Divergenz 31
3.3.3 Ergebnisse der Methoden 33
3.3.4 Untersuchung der signifikanten SNPs 35
3.3.5 Vergleich zwischen PLINK und JSD 40

4 Zusammenfassung **45**

Literaturverzeichnis **47**

Appendix **53**
.1 Single-Sampling und Multi-Sampling 54

Abbildungsverzeichnis

1.1 Vicin-Strukturformel 2
1.2 Paired-End Reads - Beispiel 5
1.3 SNP - Beispiel 5

2.1 Verteilung der konkatenierten Readlängen 15
2.2 Venn-Diagramm - Single- und Multi-Sampling (Tri-cd) 22

3.1 Histogramm der Qualitätscores aller SNPs 26
3.2 Phylogenetischer Baum der Proben 28
3.3 Manhattan-Plot - `PLINK`-Ergebnisse 34
3.4 Manhattan-Plot - JSD-Ergebnisse 34
3.5 Manhattan-Plot - `PLINK`-Ergebnisse mit signifikanten, kartierten SNPs markiert . . 36
3.6 Manhattan-Plot - JSD-Ergebnisse mit signifikanten, kartierten SNPs markiert . . . 37
3.7 Plot der SNPs, die sich in der Konsensussequenz TRINITY_DN110834_c0_g1_i1 befinden 39
3.8 Venn-Diagramm - Signifikante SNPs mit `PLINK` bzw. JSD 40
3.9 Venn-Diagramm - Signifikante, kartierte SNPs mit `PLINK` bzw. JSD 40
3.10 Scatter-Plot - P-Werte gegen JSD-Werte 41
3.11 Manhattan-Plot - `PLINK`-Ergebnisse mit `PLINK`-Signifikanten markiert 41
3.12 Manhattan-Plot - `PLINK`-Ergebnisse mit gemeinsam Signifikanten markiert 42
3.13 Manhattan-Plot - JSD-Ergebnisse mit JSD-Signifikanten markiert 42
3.14 Manhattan-Plot - JSD-Ergebnisse mit gemeinsam Signifikanten markiert 43

1 Venn-Diagramm - Single- und Multi-Sampling (`Trinity`) 54
2 Venn-Diagramm - Single- und Multi-Sampling (Transkriptom) 54

Tabellenverzeichnis

1.1 *Vicia faba* - Übersicht der Proben 6

2.1 `Bowtie2` und `Samtools` - Ergebnisse 21
2.2 `Stacks` - Ergebnisse 22

3.1 Kontigenztafel der beobachteten Häufigkeiten für einen SNP mit den generischen Allelen A und a 29
3.2 Kontigenztafel der erwarteten Häufigkeiten für den SNP in Tabelle 3.1 30
3.3 `BLAST`-Ergebnisse 38
3.4 Kontigenztafel der beobachteten Häufigkeiten für SNP TRINITY_DN168610_c1_g1_i2_92 44
3.5 Kontigenztafel der beobachteten Häufigkeiten für SNP TRINITY_DN171987_c3_g5_i1_5 44

Kapitel 1

Grundlagen

1.1 *Vicia faba*

Die Ackerbohne (*Vicia faba*) ist eine Pflanzenart in der Unterfamilie Schmetterlingsblütler (Faboideae) innerhalb der Familie der Hülsenfrüchtler (Fabaceae). Die Pflanze ist diploid mit sechs Chromosomenpaaren und einem Genom von ungefähr 13.000 Megabasen [1]. Ihre Art der Fortpflanzung ist eine Mischung aus Selbstbestäubung sowie Fremdbestäubung [2]. Im Gegensatz zu anderen Leguminosen zeichnet sie sich durch verschiedene Vorteile wie einem hohen Ertragspotential [3], hohen Proteingehalt [4] sowie hohe Stickstoffakkumulierungsleistung aus [5]. Des Weiteren eignet sie sich auch durch Vorteile in pflanzenbaulicher Hinsicht, wie eine geringe Lagerneigung und hohe Platzfestigkeit der Hülsen [1]. Eine größere Bedeutung als Nahrungsmittel hat die Ackerbohne nur in Regionen Nordafrikas, des mittleren Ostens und in China [5, 6]. In Deutschland hingegen ist die Anbaufläche nur relativ gering [4], was durch hohe Ertragsschwankungen [7], geringe Toleranz gegenüber Trockenheit [5] und dem hohen Gehalt an antinutriven Inhaltstoffen wie Tannin, Vicin und Convicin begründet werden kann [8].
Gegenüber Sommersorten hat die Winterackerbohne den Vorteil, dass sie sowohl trockenheitstoleranter [6] als auch ertragreicher ist [9]. Auch im Zuge der Klimaerwärmung haben Winterackerbohnen den Vorteil, dass sie den Hitzestress des Sommers vermeiden [9, 10]. Das größte Problem besteht in Verlusten durch Auswinterung, wie Frostschäden an der Wurzel sowie geringe Winterhärte der Pflanzen [10].
Ein großes Potential hat die Ackerbohne aufgrund ihres hohen Proteingehaltes als Tierfutter, wobei sie die momentan zu diesem Zweck aus Südamerika importierten Leguminosen verdrängen könnte [11, 12]. Hierbei sind aber wieder die antinutriven Inhaltstoffe von Nachteil, welche unter anderem die Proteinverdaubarkeit verringern, sich aber auch beispielsweise negativ auf die Befruchtung bei Legehennen auswirken.

1.2 Das Projekt Abo-Vici

Das Projekt Abo-Vici hat als Ziel die "Züchtung und Agronomie neuartiger Vicin-armer Ackerbohnen zum Einsatz als einheimisches Eiweißfutter" und läuft von 2017 bis 2020 [13] . Während des Projektes soll die Genetik von Vicin in *Vicia faba* sowie die Bedeutung, die es auf die Verwendung der Ackerbohnen als Tierfutter hat, untersucht werden. Entgültiges Ziel ist es, eine Winterackerbohne zu züchten, die einen niedrigen Vicin-Gehalt aufweist. Der allgemein höhere Ertrag von Winterackerbohnen würde, in Verbindung mit der besseren Verwendung als Tierfutter für unter anderem Schweine und Geflügel auf Grund von Vicin-Armut, die Bedeutung und die Anbaufläche der Ackerbohnen erhöhen. Abo-Vici ist ein Zusammenschluss verschiedener Züchter und Gruppen an Universitäten und wird durch die Bundesanstalt für Landwirtschaft und Ernährung im Rahmen der Eiweißpflanzenstrategie gefördert.

1.3 Vicin

Vicin ist ein glykosidisch gebundenes Aminopyrimidinderivat, welches in *Vicia faba* mit dem zugehörigen Convicin in allen Pflanzenteilen vorkommt [4, 14, 15]. Abgesehen von *Vicia faba* Arten kommt Vicin nur in der Bittermelone (*Momordica charantia*) vor [16]. Der Biosyntheseweg der Verbindung ist noch nicht genau geklärt [17], aber Untersuchungen haben angedeutet, dass es innerhalb des Orotsäure-Stoffwechsels gebildet wird. Dieser stellt die Quelle des Pyrimidinringes von Vicin dar [18]. Des Weiteren lässt sich auch ein separater Biosyntheseweg in der Wurzel annehmen [15].

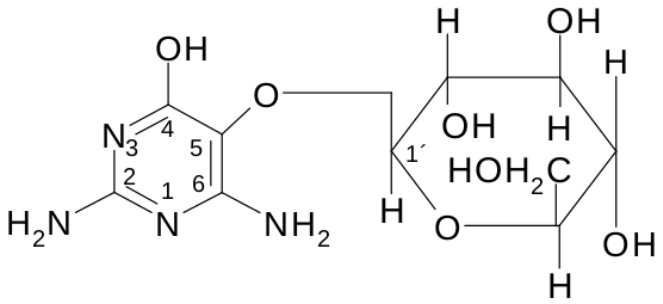


Abbildung 1.1: Strukturformel von Vicin [19]

Welche Bedeutung und Wirkung Vicin in der Pflanze hat, ist noch unklar. Man vermutet aber, dass es eine Bedeutung für die Keimung sowie eine abwehrende Wirkung gegenüber Phytopathogenen hat [20]. Zum Beispiel kommt es bei Anwesehenheit von Vicin zu einem geringeren Befall mit dem

Bohnenkäfer *Callosobruchus maculatus*, und es konnte auch eine fungizide Wirkung gegenüber den Pilzen *Botrytis cinerea* und *Aschochyta fabae* gezeigt werden [21, 22].
Das Enzym β-Glukosidase, welches in den Ackerbohnensamen vorkommt, spaltet Vicin und Convicin in die Aglycone Divicin und Isouramil. Diese sind schädlich für Menschen, die unter Glucose-6-phosphat-Dehydrogenase-Mangel (GDP6-Mangel) leiden. Bei diesen Menschen kann der Verzehr von Ackerbohnen zu Favismus führen, einer akuten hämolytischen Anämie, die potentiell tödlich sein kann. Allerdings kann der GDP6-Mangel, verstärkt durch den Verzehr von vicin- und convicinreichen Ackerbohnen, einen Schutz gegenüber der tropischen Krankheit Malaria bieten [14].

1.3.1 Vicin in der Tierfütterung

Vicin wirkt als eine antinutrive Substanz im Tierfutter monogastrischer Nutztiere wie z.B. Hühnern oder Schweinen [23]. Ein hoher Vicin-Gehalt reduziert die Energie, die das Geflügel aus dem Futter aufnimmt [14]. Aufgrund der negativen Effekte wird für die meisten Tiere ein Höchstanteil von 10-20% von Ackerbohnen am Tierfutter empfohlen [24].
Bei Zuchtsauen wurde Vicin als Ursache für negative Effekte, wie geringere Ferkelzahlen pro Wurf oder reduzierte Milchleistung, beschrieben [25]. Die Fütterung von vicinreichem Futter hat bei Küken das Wachstum verzögert und hohe Sterblichkeitsraten verursacht [26]. Bei Legehennen kommt es durch Vicin zu einem geringeren Eigewicht und Dotterindex und es treten vermehrt Blutflecken im Eidotter auf. Des Weiteren wurden die Anzahl entwicklungsfähiger Eizellen sowie Befruchtungs- und Schlupfraten reduziert [19]. Neuere Untersuchungen zeigen keinen signifikanten Einfluss des Anteils von Ackerbohnen am Futter auf Legehennen, können aber dadurch erklärt werden, dass eine Sorte mit geringem Vicin-Gehalt verwendet wurde [27]. Da Vicin im Sameninneren lokalisiert ist, können mechanische oder thermische Behandlungen des Futters den Vicin-Gehalt nicht wesentlich reduzieren [19]. Somit bleibt nur die Züchtung vicinarmer Sorten von Ackerbohnen als Lösung übrig.

1.3.2 Genetische Marker für Vicin

Vicin kann nicht komplett aus den Pflanzen entfernt werden, aber es gibt Sorten, die einen niedrigeren Vicin-Gehalt aufweisen [28]. Bei diesen liegt eine Mutation vor, die den Vicin-Gehalt um ein 10-20 faches reduziert [29]. Auf andere Inhaltstoffe der Samen zeigt die Mutation keinen Einfluss [28]. Das verantwortliche Gen konnte 5-10 Centimorgan (cM) vom Gen für das farblose Hilum kartiert werden, aber dieser morphologische Marker stellt keine Garantie für einen niedrigen Vicin-Gehalt dar.
Ein einziger Quantitative Trait Locus (QTL), der den Vicin-Gehalt der Pflanzen bestimmt, wurde auf Chromosom 1 identifiziert und seine ungefähre Position durch Marker bestimmt, welche 1,0 cM upstream und 2,6 cM downstream vom QTL liegen [17]. Diese Marker liegen auf dem Chromosom

2 der Pflanze *Medicago truncatula* (*M. truncatula*), deren 8 Chromosome Syntänie aufweisen zu den 6 Chromosomen von *Vicia faba*. Insbesondere ist Chromosom 2 kolinear zu Chromosom 1 von *Vicia faba* und hat damit ähnliche Gene in derselben Reihenfolge [30]. Der Bereich des QTLs befindet sich im Chromosom 2 von *M. truncatula* zwischen den Positionen 300.000 bp und 2.600.000 bp. Diese flankierenden Marker haben aber ebenfalls keine eindeutige Vorhersage des Vicin-Gehaltes ermöglicht, als sie auf diverse genetische Proben von *Vicia faba* angewendet wurden [31].
Letztes Jahr wurde ein KASP-Marker (Kompetitive Allele Specific PCR) entwickelt, welcher erfolgreich zwischen Proben mit niedrigem und hohen Vicin-Gehalt unterscheiden konnte [32]. Allerdings wurde festgestellt, dass der Marker nicht immer zuverlässig für bestimmte Genotypen ist. Eine Analyse mit ungefähr 1000 Pflanzen der Sorte Fabelle hat gezeigt, dass die Verteilung der beiden Allele an diesem Marker stark verzerrt war und hat zu einer nicht überzeugenden Kartierung des Markers geführt.
Dies macht es notwendig, weitere verlässlichere Marker für den Vicin-Gehalt zu entwickeln, die für mehr Genotypen verwendbar sind. Desweiteren würden mehrere Marker eine Feinkartierung des für den Vicin-Gehalt verantwortlichen Genes erlauben.

1.4 Genotyping by Sequencing

Genotyping-By-Sequencing (GBS) ist ein Verfahren zur Genom-Sequenzierung und anschließenden Suche von genetischen Markern basierend auf der Reduktion der Genomkomplexität durch Restriktionsenzyme.
Die DNA jeder Probe wird zunächst getrennt bearbeitet und mittels Restriktionsenzymen verdaut, das heißt, an Bindestellen der Restriktionsenzyme getrennt und somit in kurze Abschnitte von ungefähr 12000 Basenpaare Länge gespaltet. Daraufhin werden Adaptoren zu diesen DNA-Fragmenten hinzugefügt, welche an die Enden binden. Neben Adaptoren, die für jede Probe verwendet werden, gibt es auch proben-spezifische Barcode-Adaptoren, die später für die Zuordnung der Sequenz zur Probe genutzt werden.
Die Fragmente aller Proben werden dann zusammengetragen und mittels dem Illumina Genome Analyzer vervielfältigt. Dabei werden nur Fragmente amplifiziert, welche einen Barcode und einen gemeinsamen Adaptor aufweisen. Diese Fragmente werden dann sequenziert [33].

Paired-End Reads
Paired-End Reads sind eine Variante der DNA-Sequenzierung, bei der ein Fragment von beiden Enden aus sequenziert wird. Dies resultiert in zwei Reads pro Fragment, wobei der linke Read in Vorwärtsrichtung und der rechte Read in Richtung des reversen Komplements liegt. Abhängig von der Größe des Fragments und der Länge der Reads können die Reads überlappend sein oder nicht.

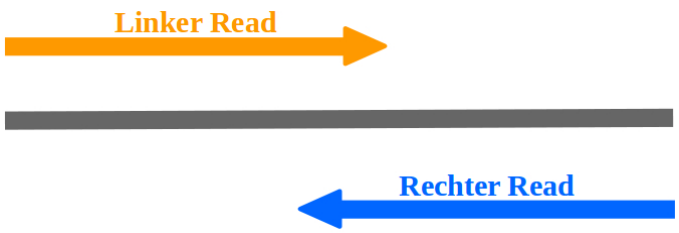


Abbildung 1.2: Vereinfachte Darstellung eines Paired-End Reads mit kurzem Überlapp

1.5 Genetische Marker

Genetische Marker sind ein wichtiger Faktor in der Pflanzenzucht und werden dort beispielsweise verwendet, um auf bestimmte Gene oder Phänotypen zu selektieren [34]. Diese Marker sind kurze DNA-Sequenzen im Genom, welche Polymorphismen wie Mutationen oder Variationen aufweisen, die es einem erlauben, zwischen verschiedenen Genotypen oder Allelen zu unterscheiden [35]. Variationen in den Allelen innerhalb von Genomen derselben Spezies können in drei Hauptgruppen geteilt werden, die Unterschiede in der Anzahl von Wiederholungen einer kurzen Sequenz an einem bestimmten Locus (Mikrosatellit, “Simple Sequence Repeats”(SSR)), Insertionen und Deletionen von Sequenzsegmenten (InDels) und “Single Nucleotide Polymorphisms”(SNPs) beinhalten.
Bei einem SNP liegt ein Unterschied an einer bestimmten Base zwischen homologen Chromosomen der selben Spezies vor. Die Abbildung 1.3 zeigt einen solchen SNP.

```
...TTGATCCGAAGGTTCCAATTTCGTAG...
...TTGATCCGAAGATTCCAATTTCGTAG...
```

Abbildung 1.3: Ausschnitt aus der Sequenz zweier homologer Chromosome der selben Spezies. Die rotmarkierte Position stellt einen SNP dar, der in einem Allel ’G’ und in dem anderen Allel ’A’ ist.

SNPs sind eine relativ neue Art von Markern, welche, trotz der Tatsache, dass sie aufgrund ihrer biallelischen Natur weniger Polymorphie als z.B. SSR-Marker aufweisen, die älteren Markerarten teilweise ersetzt haben. Dies liegt daran, dass sie reichlich und überall vorkommen sowie verwendbar für Hochdurchsatzverfahren sind [36].
Zu erwähnen sei an dieser Stelle auch der Unterschied zwischen Varianten und SNPs. Eine Variante ist eine Variation im Genom eines Individuums in Bezug auf ein Referenzgenom für diese Spezies. Diese Variante wird als ein SNP bezeichnet, wenn man sich auf die Population bezieht, zu welcher das Individuum gehört. Gegebenenfalls wird auch gefordert, dass die Variante wenigstens

in einem gewissen Anteil von Individuen der Population vorkommt (“Minor allele frequency” (MAF)). Dies wird gefordert, um die Existenz des SNP in der Population für mehrere Generationen zu garantieren.

1.6 Verwendete Daten

1.6.1 Pflanzen

In dieser Arbeit haben wir Daten von 20 verschiedenen Pflanzen analysiert. Von diesen Proben haben 6 einen niedrigen und 14 einen hohen Vicingehalt (siehe Tabelle 1.1).

ID der Probe	Vicin-Gehalt der Proben	
	Viel Vicin	Wenig Vicin
Sample_HediLin-1	+	
Sample_Hiverna-2	+	
Sample_Melodie-2		+
Sample_4		+
Sample_5		+
Sample_6		+
Sample_7		+
Sample_8		+
Sample_9	+	
Sample_10	+	
Sample_11	+	
Sample_12	+	
Sample_13	+	
Sample_14	+	
Sample_15	+	
Sample_16	+	
Sample_17	+	
Sample_18	+	
Sample_19	+	
Sample_20	+	

Tabelle 1.1: Übersicht der Proben von *Vicia faba* mit der zugewiesenen ID und ihrem Vicin-Gehalt

Die verschiedenen Sorten von *Vicia faba* liegen in Populationen vor. Um die Linien zu erhalten, aus denen die Proben genommen wurden, wurden Pflanzen aus den jeweiligen Populationen solange

selbstbefruchtet, bis man davon ausgehen konnte, dass sie sehr homozygot sind.
Sample_HediLin-1 ist eine Linie, die aus der kleinsamigen Sorte “Hedin” gezogen wurde.
Bei Sample_Hiverna-2 handelt es sich um eine Linie aus der Sorte “Hiverna”. Diese ist eine der drei in Deutschland zugelassenen Winterackerbohnensorten [37].
Sample_Melodie ist eine Linie der Sommerackerbohnensorte “Mélodie”, welche einen niedrigen Vicin-Gehalt aufweist.
Die Proben Sample_4 bis Sample_12 bestehen mit Ausnahme von Sample_8 jeweils aus Paaren. Dazu wurden vicinarme mit vicinreichen Linien gekreuzt und die entstehenden Pflanzen über mehrere Generationen lang selbstbefruchtet. Es wurden dann Pflanzen ausgewählt, die sich immer noch im Vicin-Gehalt unterschieden haben. Man kann aber davon ausgehen, dass die Genome dieses Paares ansonsten relativ ähnlich sein sollten.
Sample_4 und Sample_9 sowie Sample_5 und Sample_10 bilden jeweils Paare, die aus einer Kreuzung von der vicinarmen Linie “Melodie” mit der vicinreichen Linie “ILB938” stammen. Für die Probenpaare Sample_6 und Sample_11 sowie Sample_7 und Sample_12 gilt dasselbe, nur sind die ursprünglichen Kreuzungspartner andere gewesen.
Die Probe Sample_8 ist die Linie, aus der ursprünglich das Allel für den niedrigen Vicin-Gehalt stammt.
Sample_13 ist eine Linie aus der Sorte “Pietra Nera”, die großsamig ist. Diese Linie wurde mit der kleinsamigen Linie “Hedin” bis in die 14te Generation gekreuzt um Sample_14 und Sample_15 zu erhalten. Diese dienen dazu, in einem anderen Projekt Unterschiede zwischen groß- und kleinsamigen Ackerbohnen zu untersuchen.
Bei Sample_16 bis Sample_19 handelt es sich um Winterackerbohnen, die aus der Göttinger Winterpopulation stammen. Diese Population besteht aus elf verschiedenen Linien und wird seit 1989 auf Winterhärte selektiert [38].
Sample_20 ist eine Linie, die aus einer Pflanze der Sommerackerbohnen-Sorte “Côte d’Or” herangezogen wurde.

1.6.2 Sequenzierungsdaten

Für die 20 Proben haben wir von LGC Genomics GmbH Sequenzierungsdaten erzeugen lassen. Diese liegen in Form von 150 Basenpaare (bp) langen Paired-End Reads vor, hergestellt unter Verwendung des Illumina NextSeq 500 V2 Sequenzierers und dem Restriktionsenzym *MslI*, welches die Sequenz `5'...CAYNNNNRTG...3'` in der Mitte schneidet. Dabei gibt es für jedes Paar von Reads einen Überlapp. Im Durchschnitt haben wir ungefähr 3 Millionen Paired-End Reads pro Probe.
Wir haben drei unterschiedliche Varianten der Sequenzierungsdaten zur Verfügung gestellt bekommen:

RAW: Sequenzen aufgeteilt auf die Proben, zu denen sie gehören

AdapterClipped: Von den Sequenzen in `RAW` wurden die Überreste der Sequenzadapter entfernt sowie Reads mit einer finalen Länge kleiner als 20 bp.

RE_processed: Zusätzlich wurden hier die Bindestellen der Restriktionsenzyme entfernt

Für unsere Analyse verwenden wir nur die `RE_processed` Daten.
Die Sequenzen liegen im FASTQ-Format vor. Für jede Position in den Reads ist ein Qualitätsscore gegeben, der angibt wie wahrscheinlich es ist, dass die jeweilige Base korrekt ist.

Kapitel 2

Identifizierung von genomischen Varianten in *Vicia faba*

Für die Identifizierung von Varianten in Sequenzdaten stehen grundsätzlich zwei verschiedene Wege zur Auswahl:

Referenzbasierte Ansätze, bei denen ein Referenzgenom vorliegt, gegen welches die Reads aligniert werden, um Varianten zu bestimmen

De novo Ansätze für Organismen, bei denen es kein Referenzgenom gibt und die allein auf Basis der Reads Varianten bestimmen, indem sie diese zu einem Genom assemblieren

Da zum momentanen Zeitpunkt kein Referenzgenom für *Vicia faba* vorliegt, können wir nur de novo Verfahren zum sogenannten variant calling verwenden.
Dabei haben wir zwei unterschiedliche Ansätze verwendet, um Varianten zu bestimmen:

- Variant calling mit `Bowtie2` [39] und `Samtools` [40]
- Variant calling mit `Stacks` [41]

2.1 Variant calling mit `Bowtie2` und `Samtools`

2.1.1 Referenzgenom

Für die Verwendung von `Bowtie2` wird ein Referenzgenom benötigt. Um es für das de novo Variant calling verwenden zu können, müssen wir daher als erstes ein Äquivalent dafür haben. Dazu haben wir drei verschiedene Möglichkeiten genutzt:

- Assemblierung eines partiellen Genomes mit `Trinity` [42]

- Assemblierung eines partiellen Genomes mit `Trinity` und Clustering von diesem mit `CD-HIT` [43,44]
- Verwendung eines veröffentlichten Transkriptoms

Assemblierung eines partiellen Genomes mit `Trinity`

`Trinity` wird dazu verwendet, um aus den Sequenzierungsdaten ein partielles Genom zu assemblieren. Der verwendete Aufruf ist

```
Trinity --seqType fq --max_memory 60G --left Probe1_Links.fastq,
        ... Probe20_Links.fastq --right Probe1_Rechts.fastq,
        ... Probe20_Rechts.fastq --SS_lib_type FR --CPU 12
```

mit den Parametern

- `--seqType fq`: Format der Reads, in unserem Fall FASTQ
- `--max_memory 60G`: Begrenzung des verwendeten Arbeitsspeichers
- `--left Probe1_Links.fastq, ...`: Linke Reads jeder Probe, wobei Dateinamen mit Komma getrennt sind
- `--right Probe1_Rechts.fastq, ...`: Rechte Reads jeder Probe, wobei Dateinamen mit Komma getrennt sind
- `--SS_lib_type FR`: Orientierung der Reads, hier wird der erste Read eines Paares als Vorwärtsstrang und der zweite Read als reverser Strang betrachtet
- `--CPU 12`: Anzahl der zu verwendenden CPUs

Das finale Ergebnis ist eine Datei `Trinity.fasta`, die das erzeugte partielle Genom enthält. Dieses Genom enthält 694.605 Sequenzen, die im Idealfall je einem Locus entsprechen und nicht überlappen.

Clustering der Sequenzen mit `CD-HIT`

Eine Variante existiert, wenn es einen Unterschied in den Basen zwischen der Referenzsequenz und den alignierten Reads gibt. Bei dem Alignment wird jedem Read der Abschnitt des Referenzgenomes mit der besten Übereinstimmung zugeordnet. Da wir das Referenzgenom aus den Reads assemblieren, kann es sein, dass Reads gewissermaßen mit sich selbst aligniert werden, das heißt, mit einer Konsensussequenz, die aus diesem Read gebildet wurde, und daher keine Varianten hervorbringt.
Das von `Trinity` erzeugte partielle Genom kann Sequenzen enthalten, welche zueinander sehr ähnlich sind, aber trotzdem Abweichungen enthalten. Die Reads werden zu der jeweils am besten

passenden Referenzsequenz zugeordnet, wobei nur wenig Varianten zwischen Referenz und Read entstehen. Indem wir sehr ähnliche Sequenzen zu Clustern zusammenfassen und von jedem Cluster nur einen Repräsentanten auswählen, gibt es zwischen dem Read und seiner neuen zugeordneten Referenzsequenz, die nicht so gut übereinstimmt wie die vorherige, mehr Unterschiede und damit mehr Varianten. Diesen Vorgang kann man als Entfernung von Variabilität aus dem Referenzgenom bezeichnen. Wie man später in Abschnitt 2.4.1 sehen kann, erhöht sich damit die Anzahl der gefundenen SNPs stark.
Um die Reduktion der Variabilität zu erreichen, clustern wir die Sequenzen mit dem Programm `CD-HIT`. Stark ähnliche Sequenzen werden zu Clustern zusammengefasst und von jedem Cluster wird eine Sequenz als Repräsentant ausgewählt. Diese Repräsentanten bilden dann das reduzierte Genom.
Der verwendete Aufruf ist

```
cd-hit-est -i Trinity.fasta -o TrinityCluster -c 0.95 -n 10
           -d 0 -M 16000 - T 6
```

mit den Parametern

- `-i Trinity.fasta`: Das von `Trinity` erzeugte partielle Genom
- `-o TrinityCluster`: Ausgabedatei
- `-c 0.95`: Der Schwellenwert für die Sequenzähnlichkeit, damit zwei Sequenzen zu einem Cluster zusammengefasst werden. Die Sequenzähnlichkeit ist definiert als Anzahl der identischen Basen im Alignment geteilt durch die Länge der kürzeren Sequenz.
- `-n 10`: Wortlänge für den Algorithmus
- `-d 0`: Relevant für das Format der Ausgabe
- `-M 16000`: Größe des zu verwendenden Arbeitsspeichers
- `-T 6`: Anzahl der zu verwendenden Threads

Damit haben wir die Anzahl von Sequenzen im von `Trinity` erzeugten Genom von 694.605 auf 419.390 reduzieren können.

Veröffentlichtes Transkriptom

Für *Vicia faba* wurde online ein Transkriptom von der Cool Season Food Legume Crop Database der Washington State University veröffentlicht [45], welches durch Kombination von veröffentlichten und überprüften RNA-Seq und EST Datensätzen von *Vicia faba* erzeugt wurde.
Beide Typen von Datensätzen basieren auf der Sequenzierung von RNA und damit expremierter DNA. Bei der Transkription von DNA zur RNA können Introns entfernt und Exons auf verschiedene Arten zusammengefügt werden, was als Splicing bezeichnet wird. Da unsere Reads

durch Sequenzierung nicht gespleißter genomischer DNA erzeugt wurden, ist nicht zu erwarten, dass viele Sequenzen am Stück auf das Tranksriptom aligniert werden können. Wir wollen testen, ob die Verwendung eines Transkriptoms als Referenz eine annehmbare Alternative ist. Die Verwendung eines Transkriptomes zur Bestimmung von Varianten ist insbesondere für Spezies von Interesse, die über kein Referenzgenom verfügen und wo auch keine nah verwandte Spezies bereits sequenziert ist.
Aus 616 Millionen RNA-Seq Reads und 20.697 EST wurde ein Referenztranskriptom mit 37.378 Sequenzen konstruiert.

2.1.2 `Bowtie2`

Da wir nun unsere drei Pseudo-Referenzgenome haben, können wir `Bowtie2` starten.
`Bowtie2` ist ein sehr schnelles und speichereffizientes Programm zum Kartieren von Reads auf ein Referenzgenom. Es besteht aus zwei Schritten:

Aufbau eines Index
Im ersten Schritt wird von `Bowtie2` ein Index für das Referenzgenom erstellt. Dieser ermöglicht es, die Reads schneller zu alignieren und reduziert den erforderten Speicherbedarf.

```
bowtie2-build refGenom.fasta refGenom
```

`refGenom.fasta` ist eines der drei verwendeten Pseudo-Referenzgenome und `refGenom` ist das Präfix der erzeugten Index-Dateien, welche die Endung `.bt2` haben.

Kartieren der Reads
Im zweiten Schritt werden nun die Reads auf das Referenzgenom kartiert. Dies wird für jede Probe einzeln gemacht, daher muss der folgende Befehl in unserem Fall 20-mal ausgeführt werden, was wir in einer Schleife erledigt haben, mit den jeweiligen Daten

```
bowtie2 -x refGenom -1 Probe1_Links.fastq -2 Probe1_Rechts.fastq
        -S Probe_1.sam
```

mit Parametern:

- `-x refGenom`: Präfix der Index-Dateien für das Referenzgenom
- `-1 Probe1_Links.fastq`: Linke Reads der Probe
- `-2 Probe1_Rechts.fastq`: Rechte Reads der Probe
- `-S Probe_1.sam`: Ausgabedatei

Das Ergebnis ist ein Alignment der Reads gegen das Referenzgenom im SAM-Format.

2.1.3 Samtools

Mit `Samtools` führen wir nun die eigentliche Identifizierung der SNPs durch.
Im ersten Schritt muss wieder ein Index für das Referenzgenom erzeugt werden

```
samtools faidx refGenom.fasta
```

`refGenom.fasta` ist hierbei wieder eines der drei Pseudo-Referenzgenome.
Da die SAM-Datei eine Textdatei ist, wird diese in das binäre BAM-Format umgewandelt und anschließend sortiert und indiziert, um die Zugriffsgeschwindigkeit zu erhöhen.

```
samtools view -b -S -o Probe_1.bam Probe_1.sam
```

konvertiert die Datei, wobei:

- `-b`: Ausgabe soll im BAM-Format erfolgen
- `-S`: Eingabe ist im SAM-Format
- `-o Probe_1.bam`: Ausgabedatei

Zum Sortieren und Indizieren werden folgende Befehle verwendet

```
samtools sort Probe_1.bam Probe_1.sorted
samtools index Probe_1.sorted.bam
```

Haben wir nun alle Dateien erzeugt, können wir die SNPs bestimmen mit `samtools mpileup`.

Single- und Multi-Sample Calling

Bei diesem Schritt stehen uns zwei Optionen zur Wahl.
Erstens könnten wir jede Probe getrennt durchgehen, ihre Varianten bestimmen und diese später zu einer Datei zusammenfassen. Bei dieser Vereinigung der Ergebnisse muss darauf geachtet werden, dass Varianten an derselben Position in derselben Konsensussequenz zwischen Proben unterschiedliche Alternativallele aufweisen können. Diese müssen, weil wir nur an biallelischen SNPs interessiert sind, rausgefiltert werden. Da immer nur eine Probe untersucht wird, kann man diesen Vorgang als Single-Sampling bezeichnen.
Die zweite Möglichkeit ist das Multi-Sampling. Hierbei werden alle Proben gleichzeitig analysiert und die SNPs bestimmt. Im Vergleich zur ersten Option liegt hierbei ein Fokus auf SNPs, die in mehrere Proben der Population vorkommen. Indem die Korrelation zwischen Proben verwendet wird, können SNPs bestimmt werden, die in mehreren Proben vorkommen, aber in jeder einzelnen Probe zu schwach existieren, um sie durch das Single-Sampling zu erfassen. Andererseits verlieren wir Varianten, die nur in einer Probe vorkommen, aber dort stark genug auftreten um vom Single-Sampling gefunden zu werden.
Von den beiden Möglichkeiten ist für uns das Multi-Sampling das geeignetere Verfahren, da

uns speziell SNPs interessieren, die in einem Großteil der Proben vorkommen und die damit gegebenenfalls als Marker für den Phänotyp dienen können. Varianten hingegen, die nur in einer oder zwei Proben existieren, sind für unsere Aufgabenstellung weitaus weniger wichtig. Wir haben dennoch beide Optionen durchgeführt.
In beiden Fällen sind die verbleibenden Schritte nahezu gleich.
Der Aufruf für Single-Sampling ist

```
samtools mpileup -u -d 10000 -f refGenom.fasta Probe_1.sorted.bam > Probe_1.bcf
```

mit

- `-u`: Gibt an, dass die Ausgabe nicht komprimiert werden soll
- `-d 10000`: Gibt an, dass an einer Position maximal 10000 Reads pro Eingabedatei gelesen werden sollen
- `-f refGenom.fasta`: Das indizierte Pseudo-Referenzgenom
- `Probe_1.sorted.bam`: Vorher erzeugte BAM-Datei der Probe
- `Probe_1.bcf`: Ausgabedatei im BCF-Format

Für Multi-Sampling werden einfach die BAM-Dateien jeder Probe im Aufruf hintereinander geschrieben. Ansonsten gibt es keine Unterschiede.
BCF ist wieder ein binäres Format, welches in das menschenlesbare Format VCF umgewandelt werden muss mit

```
bcftools call -v -c Probe_1.bcf > Probe_1.vcf
```

mit Parametern:

- `-v`: Gib nur Positionen mit SNPs aus
- `-c`: Verwende die originale `samtools/bcftools` Calling-Methode
- `Probe_1.vcf`: Ausgabedatei im VCF-Format

2.2 `Stacks`

`Stacks` ist eine Software Pipeline zur Identifikation von Varianten in Sequenzierungsdaten, welche sowohl für de novo Assemblierung als auch für das variant calling mit einem Referenzgenom verwendet werden kann.
Neben den Varianten kann `Stacks` ebenfalls zur Erzeugung von Statistiken über die Genetik der Population verwendet werden, wovon wir in dieser Arbeit aber keinen Gebrauch machen.
Die von uns verwendete Version von `Stacks` (1.48) kann Paired-End Reads nicht direkt verwenden. Daher haben wir die Pipeline dreimal mit unterschiedlichen Datensätzen ausgeführt:

- Nur die linken Reads
- Nur die rechten Reads
- Linke und rechte Reads konkateniert zu jeweils einem Read. Dazu haben wir von den rechten Reads das reverse Komplement gebildet und den Teil, der bereits im linken Read vorliegt, vor dem Konkatenieren abgeschnitten. In diesem Fall konnte `Stacks` die Daten aber auch nicht verarbeiten, da die Reads unterschiedliche Längen aufwiesen. Dies hat eine Kürzung der konkatenierten Reads auf eine bestimmte Länge notwendig gemacht. Dabei verlieren wir die Reads, die eine kürzere Länge aufweisen. Wir haben die Reads auf 181 bp gekürzt, damit ungefähr 50% der Reads in jeder Probe erhalten bleiben, wie man in Abbildung 2.1 sehen kann.

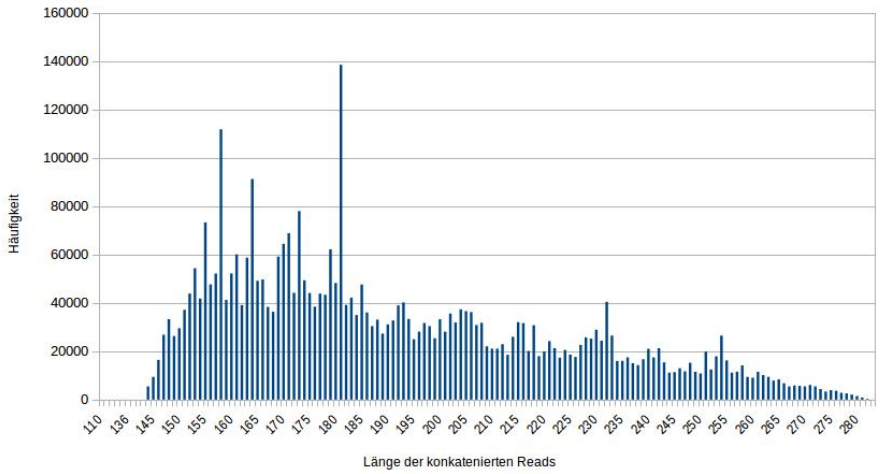


Abbildung 2.1: Verteilung der Längen der konkatenierten Reads einer Probe

Inzwischen wurde Version 2.0 von `Stacks` veröffentlicht, welches die Unterstützung für Paired-End Reads leistet. Da die Programme geändert und teilweise komplett ersetzt wurden, gilt die folgende Beschreibung des Ablaufs nicht mehr für die neue Version. Mit diesen Neuerungen ist eine neue Evaluation der Eignung von `Stacks` zum de novo variant calling mit Paired-End Reads erforderlich.

2.2.1 Vorbereitung der Daten

Bevor die eigentliche Pipeline verwendet werden kann, müssen die Sequenzierungsdaten vorbereitet werden, was mit dem zu `Stacks` gehörenden Programm `process_radtags` erfolgt. Dieses dient in erster Hinsicht dazu, die Reads zu demultiplexen, sprich sie anhand von Barcodes den Proben, zu denen sie gehören, zuzuordnen. Falls sie noch in den Reads vorhanden sein sollten,

können auch Adaptersequenzen und Bindestellen der verwendeten Restriktions-Enzyme damit entfernt werden. Da wir bereits Daten haben, bei denen diese Abschnitte aus den Reads entfernt wurden und die Reads selber den Proben zugeordnet sind, verwenden wir das Programm nur zu einer Filterung von Reads, die einen zu niedrigen Qualitätsscore aufweisen. Dies machen wir mit dem folgenden Befehl:

```
process_radtags -f Input.fastq -o Dir --disable_rad_check -q
```

wobei,

- `-f Input.fastq`: FASTQ-Datei mit den Reads
- `-o Dir`: Verzeichnis in das die Ergebnisdateien geschrieben werden
- `--disable_rad_check`: Deaktivieren der Überprüfung, ob die RAD-Bindestelle intakt ist (für unsere Daten nicht von Belang)
- `-q`: Entferne Reads mit zu niedrigen Qualitätsscores

Für die konkatenierten Reads fügen wir noch einen Parameter `-t 181` hinzu, sodass zusätzlich die Reads auf 181 bp gekürzt und kürzere Reads entfernt werden.
Zur Bestimmung der Qualität eines Reads wird dieser mit einem gleitenden Fenster untersucht. Fällt die durchschnittliche Qualität innerhalb des Fensters unterhalb von 90%, das heißt die Wahrscheinlichkeit das die Basen in diesem Abschnitt korrekt sind, wird der Read verworfen. Somit werden weiterhin isolierte Fehler zugelassen, aber Reads mit anhaltenden Sequenzen niedriger Qualität eliminiert.

2.2.2 Ausführung der Pipeline

Die `Stacks`-Pipeline ist durch ein Perl-Skript definiert, welches die von `process_radtags` erzeugten Daten entgegennimmt und die einzelnen Programme der Pipeline aufruft:

```
denovo_map.pl -m 3 -M 3 -n 2 -T 15 -B Vicia_Faba -b 1 --create_db
              -o OutputDir -O popmap --samples InputDir
              -X "populations:--vcf"
```

wobei die Parameter folgende Bedeutung haben:

- `-m 3`: Gibt die minimale Anzahl von identischen Reads an, die notwendig sind um einen Stapel (Stack) zu bilden
- `-M 3`: Gibt die Anzahl von erlaubten Mismatchen zwischen Stapel an, damit sie einen Locus bilden können, wenn eine Probe verarbeitet wird
- `-n 2`: Gibt die Anzahl von erlaubten Mismatchen zwischen Loci an, wenn der Katalog erstellt wird

- `-T 15`: Anzahl der zu verwendenden Threads
- `-o OutputDir`: Verzeichnis, in welches die Ergebnisdateien geschrieben werden sollen
- `-O popmap`: Eine Liste aller Proben mit Namen der Probe und mit der Angabe, ob sie einen hohen oder niedrigen Vicin-Gehalt hat
- `--samples InputDir`: Verzeichnis mit den von `process_radtags` erstellten Daten für alle Proben
- `-X "populations:--vcf"`: Erstellung einer VCF-Datei mit gefundenen SNPs

Nicht erwähnte Parameter beziehen sich lediglich auf das Einlesen der Ergebnisse in eine Datenbank und können weggelassen werden. Die Pipeline besteht aus den Programmen

- `USTACKS`
- `CSTACKS`
- `SSTACKS`
- `POPULATIONS`

USTACKS

`USTACKS` nimmt eine FASTQ-Datei mit den Reads einer Probe als Eingabe und bildet aus Reads mit sehr hoher Sequenzähnlichkeit multiple Alignments. Diese werden als Stapel (stacks) bezeichnet. Durch das Zulassen geringfügiger Abweichungen in der Sequenz wird sicher gestellt, dass Sequenzen, die sich nur durch ihre Allele unterscheiden, dem selben Stapel zugeordnet werden. Durch Vergleich der Stapel wird eine Menge von möglichen Loci erzeugt und Varianten werden an jedem Locus unter Verwendung eines Maximum-Likelihood-Frameworks detektiert. Des Weiteren stellt jeder Stapel eine Konsensussequenz dar. Das Programm wird für jede Probe ausgeführt.

CSTACKS

`CSTACKS` erstellt einen Katalog mit den Ergebnissen von `USTACKS`. Es erzeugt eine Menge von Konsensus-Loci, wobei Allele vereinigt werden.

SSTACKS

`SSTACKS` nimmt die Menge von Stapeln, das heißt mögliche von `USTACKS` erzeugt Loci, und sucht sie in dem Katalog, der von `CSTACKS` erstellt wurde.

POPULATIONS

POPULATIONS analysiert eine Population bestehend aus individuellen Proben, in unserem Fall 20 Proben, berechnet mehrere Statistiken der Populationsgenetik für die Daten und erlaubt den Export der Ergebnisse in mehrere Standardformate. Wir verwenden es hauptsächlich, um die gefundenen SNPs im VCF-Format zu exportieren. Diese Datei enthält, im Gegensatz zu der von Samtools erzeugten, keine Qualitätsscores für die SNPs. Neben der VCF-Datei haben wir auch für jeden SNP die dazugehörige von Stacks bestimmte Konsensussequenz.

2.3 Konvertierung der Daten zur weiteren Analyse

Jedes Verfahren liefert uns eine oder im Falle des Single-Sample Calling mehrere VCF-Dateien. Da die weitere Analyse größtenteils mit dem Programm PLINK erfolgt, konvertieren wir die VCF-Dateien in das PED/MAP-Format. Eine PED-Datei besteht aus einer Zeile pro Probe, welche neben Namen der Probe und einigen Variablen wie Phänotyp oder Geschlecht, die Allele jeder Variante für diese Probe enthält. Die Reihenfolge der Varianten ist für jede Probe gleich und ist in der MAP-Datei verzeichnet, welche neben der ID der Variante außerdem Informationen über deren Position wie Chromosom und Position in Centimorgan sowie Basenpaaren enthält. Da wir nur über Pseudo-Referenzgenome verfügen sind uns die Positionsinformationen nicht bekannt. Die entsprechenden Werte werden daher auf 0 gesetzt für Chromosom und Position in Centimorgan und auf die Position in der Referenzsequenz für die Position in Basenpaaren.
Diese Konvertierung unterscheidet sich zwischen den beiden Verfahren.

2.3.1 Konversion der Daten von `Stacks`

Stacks liefert eine VCF-Datei zurück, die direkt mit PLINK konvertiert werden kann mit

```
plink --vcf Stacks.vcf --allow-extra-chr --biallelic-only strict
      --pheno pheno.list --double-id --recode --out Stacks
```

wobei,

- `--vcf Stacks.vcf`: Ausgabedatei von Stacks
- `--allow-extra-chr`: Lässt zusätzliche Chromosom-Bezeichner zu (Stacks gibt für alle SNPs als Chromosom “un” an)
- `--biallelic-only strict`: Entfernt Varianten, die mindestens zwei Alternativallele besitzen
- `--pheno pheno.list`: Datei, welche die Phänotypen der einzelnen Proben enthält

- `--double-id`: Setzt die Familien-ID und die innere Familien-ID auf die ID der Probe
- `--recode`: Schreibt die Ergebnisse in neue PED/MAP-Dateien
- `--out Stacks`: Gibt das Präfix der erzeugten PED/MAP-Datei an

2.3.2 Konversion der Daten von `Bowtie2` und `Samtools`

Bevor wir diese Ergebnisse formatieren können, müssen wir die Spalten der VCF-Dateien ändern. Als Chromosom wird hier nämlich die ID der Konsensussequenz verwendet, während die ID der Varianten gleichgesetzt werden auf ".".
Um das Format dem von `Stacks` anzugleichen und spätere Analysen zu erleichtern, setzen wir alle Chromosome auf 0 und die ID der Variante auf *ID_Pos*, wobei *ID* die Konsensussequenz angibt und *Pos* die Position der Variante in dieser. Dazu verwenden wir das Linux-Programm `gawk` mit

```
gawk -i inplace '/#/{print;next} {t=$1;$1=0;$3=t"_"$2;print;}'
     OFS=\\t Probe1.vcf
```

Die Formatierung ist auch in der Hinsicht wichtig, dass `PLINK` nur eine begrenzte Anzahl von unterschiedlichen Chromosom-Bezeichnern zulässt. Sind es zu viele, was bei großen Anzahlen von Konsensussequenzen leicht passieren kann, bricht das Programm ab.
Für die Multi-Sampling Verfahren, die nur in einer VCF-Datei resultieren, kann der selbe Aufruf wie für `Stacks` verwendet werden, nur muss noch ein Parameter `-snps-only` hinzugefügt werden. Dieser sorgt dafür, dass Indels entfernt werden.
Verwenden wir hingegen das Single-Sampling Verfahren, müssen wir noch zusätzlich alle unsere einzelnen Dateien zusammenfügen. Dies ist problematisch, da die Varianten nur in der jeweiligen Probe vorkommen können oder aber auch triallelisch sein können, wenn man mehrere Proben betrachtet.
Dafür verwenden wir folgenden Ablauf:

1. Mit einem selber erstellten Programm fassen wir die SNP-IDs aus allen VCF-Dateien zu einer Liste zusammen, die neben der ID auch das Referenz- und Alternativallel enthält. Dabei entfernen wir die SNPs, die Indels oder nicht biallelisch über alle Proben sind.

2. Danach konvertieren wir jede VCF-Datei mit `PLINK` in das BED-Format mit

   ```
   plink --vcf Probe_x.vcf --allow-extra-chr --extract filteredSNPList
         --double-id --make-bed --out Probe_x
   ```

   Dabei werden nur die Varianten übernommen, die in unserer im ersten Schritt erstellten Liste stehen, was mit dem Parameter `--extract` gefolgt von dem Dateinamen der Liste erfolgt. Der neue Parameter `--make-bed` dient zur Ausgabe der Datei im BED-Format

3. Nun erfolgt die erste Vereinigung der einzelnen Proben mit

```
plink --bfile Probe_1 --allow-extra-chr --merge-list ProbenListe
      --out Merge
```

Mit `--merge-list` geben wir an, in welcher Datei die Proben stehen, die vereinigt werden sollen. Diese enthält für jede Probe `X` eine Zeile

```
X.bed X.bim X.fam
```

Falls unsere Liste der gefilterten SNPs nicht ganz korrekt ist, das heißt, im Fall, dass SNPs darin vorkommen, die über alle Proben betrachtet nicht biallelisch sind, kommt es zu Fehlern bei diesen SNPs, die von `PLINK` in einer Datei `Merge.missnp` geschrieben werden.
Sind solche vorhanden, müssen sie vor dem nächsten Schritt aus den einzelnen Dateien entfernt werden. Dazu wiederholt man Schritt 2 mit dem zusätzlichen Parameter `--exclude Merge.missnp`, um die fehlerhaften SNPs zu entfernen und fährt dann mit Schritt 3 fort.

4. Anschließend wandeln wir die BED-Datei ins PED-Format um und fügen gleichzeitig die Phänotypen der einzelnen Proben hinzu. Dafür brauchen wir eine Datei `phenoDatei`, die diese enthält in der Form

```
FamilienID Innere-FamilienID Phänotyp
```

Dies geschieht mit

```
plink --bfile Merge --allow-extra-chr --pheno phenoDatei
      --recode --out Merge
```

5. Da wir SNPs haben, die nicht in allen Proben gefunden werden, beispielsweise weil sie in dieser Probe homozygot bezüglich des Referezallels sind, haben die Proben für diese SNPs teilweise `0 0` als Genotyp in der PED-Datei. Diese Nullen geben an, dass keine Informationen vorliegen in den jeweiligen VCF-Dateien. Wir nehmen an, dass alle diese SNPs in den Proben, in denen die Information fehlt, homozygot bezüglich der Referenzsequenz sind. Als letzten Schritt ersetzten wir daher diese Nullen durch die jeweiligen Referenzallele, die wir aus der in Schritt 1 erzeugten Liste entnehmen können. Dazu verwenden wir wieder ein selber geschriebenes Programm.

2.4 Ergebnisse

2.4.1 `Bowtie2` und `Samtools`

Bei der Pipeline bestehend aus `Bowtie2` und `Samtools` müssen wir unterscheiden, was wir als Referenzgenom verwendet haben und ob wir während `samtools mpileup` Single- oder Multi-Sample Calling benutzt haben.

Analyse mit `Bowtie2` und `Samtools`	*Trinity*	*Tri-cd*	*Transkr.*
Anzahl von Konsensussequenzen	694.605	419.390	37.378
Anzahl von SNPs (Single-Sampling)	921.710	1.636.232	47.084
Anzahl von SNPs (Multi-Sampling)	1.335.664	1.880.592	56.173

Tabelle 2.1: Ergebnisse der Analyse mit `Bowtie2` und `Samtools`. Die Spalte *Trinity* enthält die Ergebnisse für das mit `Trinity` assemblierte partielle Genom, *Tri-cd* die für das mit `CD-HIT` reduzierte Genom und *Transkr.* die Ergebnisse für das veröffentlichte Transkriptom.

Wie Tabelle 2.1 zeigt, erhöht die Verwendung des Multi-Sampling für alle drei Datensätze die Anzahl der gefundenen SNPs. Auch hat die Reduktion des von `Trinity` erzeugtem partiellen Genom mit `CD-HIT` die Anzahl der gefunden SNPs um 77% beziehungsweise 40% erhöht in Abhängigkeit von der Sampling-Variante. Das veröffentlichte Transkriptom erzeugt im Vergleich weitaus weniger SNPs, was, wie in Abschnitt 2.1.1 erläutert, durch die RNA-Sequenzen verursacht sein dürfte, aus denen es besteht. Während kein direkter Vergleich der Ergebnisse zwischen den verschiedenen Referenzgenomen möglich ist, können wir untersuchen, wie sich die Ergebnisse von Single- und Multi-Sampling unterscheiden. Dazu zeigt Abbildung 2.2 beispielhaft ein Venn-Diagramm bestehend aus den SNPs, die mit dem geclusterten `Trinity`-Genom identifiziert wurden mit Single- bzw. Multi-Sampling.

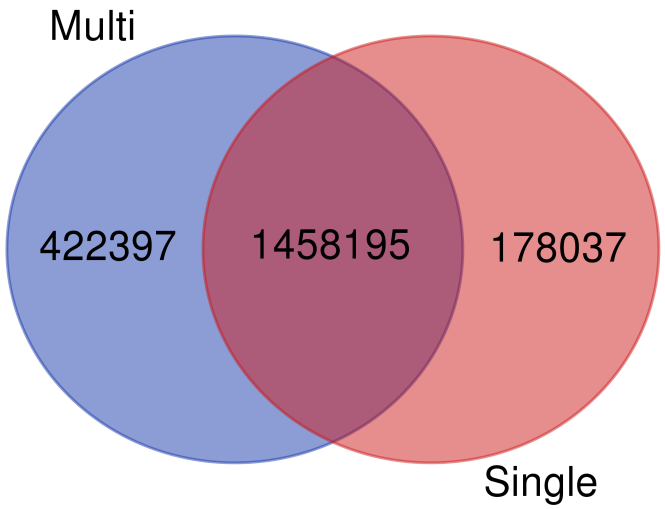


Abbildung 2.2: Venn-Diagramm der SNPs gefunden mit dem reduzierten von `Trinity` erzeugten Genom mit Single- bzw. Multi-Sampling. Der gemeinsame Anteil von SNPs beträgt 89% für Single- und 77% für Multi-Sampling.

Der größte Teil der SNPs (89% im Single-Sampling bzw. 77% im Multi-Sampling) kommt in beiden Verfahren vor. Ungefähr 180.000 SNPs sind einzigartig für das Single-Sampling, während 420.000 nur durch das Multi-Sampling gefunden werden. Letztere sind dadurch zu erklären, dass das Multi-Sampling sensitiver für SNPs ist, die in mehreren Proben vorkommen, aber dort jeweils nur schwach. Auf die Art verlieren wir aber auch die SNPs, welche nur in sehr wenigen Proben stark vorkommen, die hingegen beim Single-Sampling identifiziert werden. Ein ähnliches Bild zeigt sich auch für die beiden anderen Datensätze, die im Anhang .1 aufgeführt sind.

2.4.2 `Stacks`

`Stacks` liefert als Ergebnis die erzeugten Konsensussequenzen und SNPs, die sich auf diesen befinden. Deren Anzahlen befinden sich für die drei unterschiedlichen Daten in Tabelle 2.2.

Analyse mit `Stacks`	rechts	links	konkat
Anzahl von Konsensusseqs.	1.517.582	2.871.569	1.851.558
Anzahl von SNPs	435.906	1.263.573	625.553

Tabelle 2.2: Ergebnisse der Analyse mit `Stacks`. Die Spalte *rechts* enthält die Ergebnisse basierend auf der Verwendung von nur den rechten Reads, *links* die Ergebnisse von nur den linken Reads und *konkat* die Ergebnisse basierend auf den konkatenierten Reads.

Da jede Konsensussequenz nur eine interne Katalog-ID als Identifikator hat, ist es nicht trivial möglich die SNPs zu vergleichen, um einen eventuellen Überlapp zwischen den drei Verfahren festzustellen.
Für die weitere Analyse beschränken wir uns auf die SNPs, die wir aus dem mit `CD-HIT` reduzierten `Trinity`-Genom mit der `Bowtie2-Samtools`-Pipeline unter Verwendung von Multi-Sampling gewonnen haben. Mit diesem Verfahren erhalten wir die meisten SNPs und der Fokus dieser liegt durch das Multi-Sampling auf jene, die in mehreren Proben vorkommen und welche uns damit mehr interessieren. Eine Analyse der mit `Stacks` erzeugten SNPs und der SNPs, die auf dem Transkriptom basieren, könnte später noch durchgeführt werden, würde aber den Umfang dieser Arbeit überschreiten.

Kapitel 3

Überprüfung der Genotypen und Assoziationsanalyse

In diesem Kapitel wird beschrieben, wie wir die zuvor erzeugten PED-Dateien mit den gefundenen SNPs auswerten. Diese Auswertungen fallen in zwei Bereiche.
Einerseits wollen wir allgemein überprüfen, ob unsere Ergebnisse "richtig" sind, das heißt, ob sie in gewissen Aspekten unsere Erwartungen erfüllen, die wir aufgrund von Vorwissen haben.
Andererseits möchten wir gezielt die Assoziation zwischen den SNPs und dem Phänotyp, sprich, dem Vicin-Gehalt untersuchen, mit dem Ziel SNPs zu identifizieren, die wir als Marker zur Erkennung des Phänotypes verwenden können.

3.1 Filterung der SNPs nach ihren Qualitätscores

In unseren Ergebnissen können auch Varianten vorkommen, die nur aufgrund von Sequenzierfehlern gefunden wurden und nicht in der Realität vorkommen. Um diese zu eliminieren, haben wir die SNPs nach ihrem Qualitätscore gefiltert. Diese Werte liegen in einem Bereich von 0 bis 999, wobei 999 die bestmögliche Qualität anzeigt.
Wie Abbildung 3.1 zeigt liegen die Qualitätscores in einer stark bimodalen Verteilung vor. Von den ungefähr $1,9$ Millionen SNPs haben um die 685.000 SNPs den maximalen Qualitätscore von 999, während die anderen einen Qualitätscore niedriger als 300 haben.

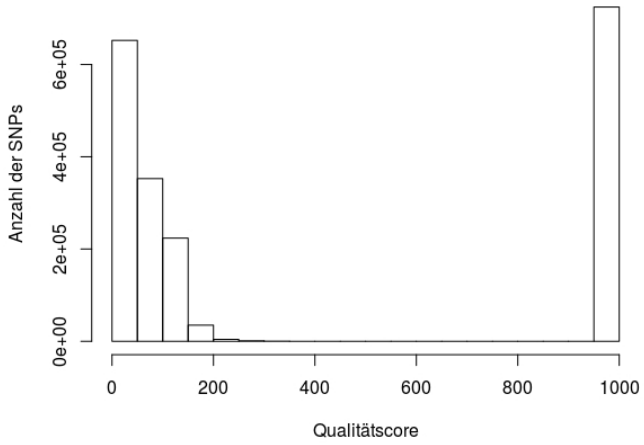


Abbildung 3.1: Histogramm der Qualitätscores aller SNPs

Um diese SNPs mit niedriger Qualität, unter denen sich sicherlich welche befinden, die nur aufgrund von Sequenzierfehlern entstanden sind, zu entfernen, filtern wir alle SNPs aus der PED-Datei, deren Qualitätscore niedriger als 999 ist.
Dies können wir ebenfalls mit `PLINK` erreichen mit

```
plink --file SNPs --qual-scores SNPs.vcf 6 3 '#' --qual-threshold 999
      --recode --out SNPs_Filtered
```

wobei

- `--qual-scores SNPs.vcf 6 3 '#'`: Gibt die Datei an, in der für die SNPs die Qualitätscores liegen. Dies ist im allgemeinen die VCF-Datei aus der die PED-Datei erzeugt wurde. 6 und 3 geben die Spalten an, in der sich Qualitätscore beziehungsweise die SNP-ID befindet und '#' gibt an, dass Zeilen, die mit # beginnen, ignoriert werden sollen.

- `--qual-threshold 999`: Gibt die untere Schranke für den Qualitätscore an. Alle SNPs mit einem niedrigeren Qualitätscore werden eliminiert.

Nach dieser Filterung haben wir nun nur noch 685.215 von den ursprünglichen $1.880.592$ SNPs.
Im folgenden verwenden wir nur diese gefilterte PED-Datei.

3.2 Distanzen

Wie in Kapitel 1.6.1 beschrieben wurde, besteht zwischen den Proben ein bestimmtes Verwandtschaftsverhältnis, sprich, sie sind unterschiedlich nah verwandt. Würde man diese Verwandschaften in einem phylogenetischen Baum darstellen, müssten zum Beispiel die Proben 4 und 9 oder 5 und 10 sehr nah verwandt sein, da sie sich eigentlich nur in dem für Vicin verantwortlichen Bereich genetisch unterscheiden. Da nah verwandte Proben einen ähnlichen Genotyp haben, sollten sich diese Verhätnisse auch in unseren SNPs wiederspiegeln. Nah verwandte Proben sollten bei den SNPs dieselben Allele haben. Mit dem Programm `treemix` [46] können wir aus diesen Informationen einen phylogenetischen Baum für die Proben erzeugen.
Dafür erstellen wir als erstes mit

```
plink --file SNPs --freq --family --out SNPsTree
```

eine Datei `SNPsTree.frq.strat`, die die Allelhäufigkeiten der Proben enthält. Diese Datei muss danach mit `gzip` komprimiert werden, was vom folgenden Programm erfordert wird.
Dann werden mit

```
python plink2treemix.py SNPsTree.frq.strat.gz Treemix2SNPsTree.frq.gz
```

die Daten für `treemix` vorbereitet und anschließend mit

```
treemix -i Treemix2Trinity-FR-cdHitTree.frq.gz -o Trinity-FR-cdHit-stem
```

in Distanzen für einen Baum umgewandelt.
Diesen Baum kann man dann beispielsweise mit `SplitsTree` [47] darstellen wie in Abbildung 3.2 zu sehen ist.
Wie zu sehen ist, sind die Proben, welche bis auf den für den Vicin verantwortlichen Teil des Genomes gleich sein sollen, so nah verwandt, dass eine genaure Trennung teilweise nicht mehr möglich ist. Dass dies für jedes dieser Paare (Sample_4 + Sample_9, Sample_5 + Sample_10, Sample_6 + Sample_11, Sample_7 + Sample_12) zutrifft, deutet auf die Richtigkeit unserer Ergebnisse hin.

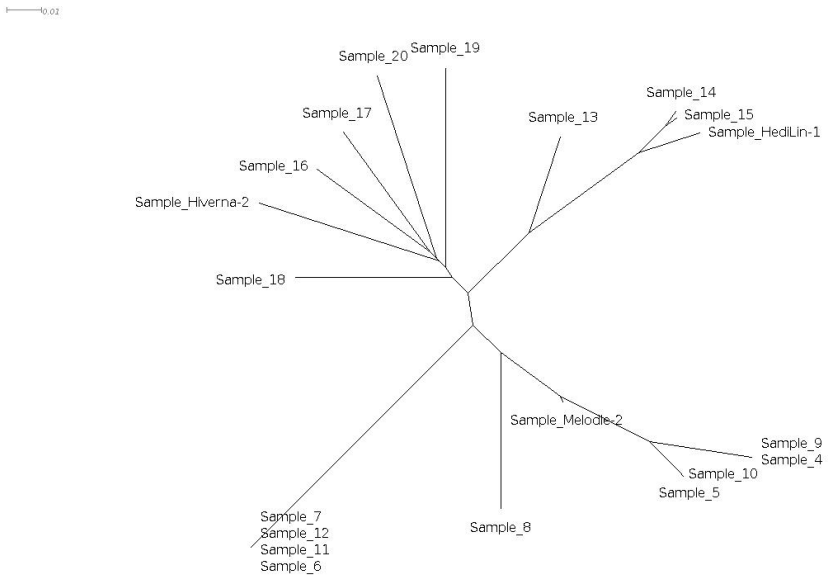


Abbildung 3.2: Phylogenetischer Baum der Proben, basierend auf der Ähnlichkeit ihrer SNPs als Distanz

3.3 Genomweite Assoziationsstudie

Bei einer genomweiten Assoziationsstudie (GWAS) wird der Zusammenhang zwischen genetischen Markern, wie z.B. SNPs, und einem Phänotyp untersucht. Der Phänotyp ist in unserem Fall der Vicin-Gehalt einer Probe in Form einer binären Variable, die angibt, ob der Vicin-Gehalt hoch oder niedrig ist. Ziel ist es, Marker zu finden, die eine möglichst eindeutige Vorhersage des Phänotypes erlauben und damit später dann zu diesem Zweck beispielsweise in der Züchtung verwendet werden können.
Allgemein gesagt werden bei einer Assoziationsstudie die Verhältnisse der Allelhäufigkeiten bzw. Genotypenhäufigkeiten der SNPs zwischen den unterschiedlichen Phänotypen verglichen und statistisch auf eine signifikante Assoziation hin getestet. Im Gegensatz zu anderen Verfahren wie Kandidatengen-Assoziationsstudien werden keine Annahmen darüber gemacht, an welcher Stelle des Genomes die signifikanten SNPs liegen, womit das Verfahren ohne Vorkenntnisse verwendet werden kann [48].

3.3.1 PLINK

PLINK bietet verschiedene Möglichkeiten, eine GWAS durchzuführen in der Hinsicht, dass es unterschiedliche Modelle zur Darstellung der SNPs gibt. Diese sind

- Allel-Test (Standardtest)
- Genotypen-Test
- Test auf dominanten Effekt
- Test auf rezessiven Effekt

Des Weiteren gibt es noch den Cochran-Armitage Trend-Test, der verwendet werden kann, um ein additives Modell zu untersuchen.
Wir beschränken uns auf den normalen Allel-Test, da wir mit Inzuchtlinien arbeiten, die theoretisch keine oder nur wenige Heterozygoten enthalten, der in PLINK mit

```
plink --file SNPs --assoc --allow-no-sex --out SNPs_Assoc
```

ausgeführt werden kann. Mit diesem Befehl wird für jeden SNP ein Chi-Quadrat-Test durchgeführt.

Chi-Quadrat-Test
Der Chi-Quadrat-Test ist ein nicht-parametrischer Test, der auf Unabhängigkeit zwischen zwei kategorische Variablen testet [49]. Dazu werden die Daten nach der ersten Variable gruppiert und dann die Verhältnisse der zweiten Variable zwischen den Gruppen verglichen. Sind die Gruppen unterschiedlicher als man durch Zufall erwarten kann, liegt eine Assoziation zwischen den Variablen vor. In unserem Fall trennen wir die Daten nach dem Phänotyp in zwei Gruppen auf und vergleichen dann die Allelhäufigkeiten zwischen den beiden Gruppen. Zur Repräsentation der Daten kann eine Kontigenztafel verwendet werden, wie Tabelle 3.1 beispielhaft für ein SNP zeigt.

	Gruppe		
Allele	Wenig Vicin	Viel Vicin	Summe
a	1	12	13
A	5	2	7
Summe	6	14	20

Tabelle 3.1: Kontigenztafel der beobachteten Häufigkeiten für einen SNP mit den generischen Allelen A und a

Zur Berechnung werden neben den beobachteten Häufigkeiten auch noch die erwarteten Häufigkeiten benötigt. Diese ergeben sich unter Annahme der Unabhängigkeit aus der Multiplikation der

Rand-Wahrscheinlichkeiten multipliziert mit der Stichprobengröße, was in Tabelle 3.2 für unser Beispiel gezeigt wird.

	Gruppe		
Allele	Wenig Vicin	Viel Vicin	Summe
a	$3,9$	$9,1$	13
A	$2,1$	$4,9$	7
Summe	6	14	20

Tabelle 3.2: Kontigenztafel der erwarteten Häufigkeiten für den SNP in Tabelle 3.1

Sind die Werte der Kontigenztafeln ähnlich, so können wir davon ausgehen, dass die Variablen unabhängig voneinander sind und damit keine Assoziation vorliegt. Diese Ähnlichkeit, bezeichnet als χ^2, wird berechnet mit

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.1}$$

Dabei sind O_{ij} die beobachteten Häufigkeiten und E_{ij} die erwarteten. Auf Basis der Chi-Quadrat-Verteilung kann dann aus χ^2 ein Signifikanzwert berechnet werden.
`PLINK` liefert uns als Ausgabe für jedes SNP diesen Signifikanzwert (P-Wert), der angibt, wie wahrscheinlich es ist, dass keine Assoziation mit dem Phänotyp vorliegt. Je niedriger er ist, desto wahrscheinlicher ist es, dass eine Assoziation besteht.

Mehrfaches Testen
Bei einfachen statistischen Tests setzt man eine obere Schwelle α für den Signifikanzwert (z.B. 0.05) und begrenzt damit die Wahrscheinlichkeit von falsch positiven Ergebnissen. Im Falle von mehrfachen Testen, wie es bei der GWAS der Fall ist, ist aber die Wahrscheinlichkeit, einen solchen Fehler zu machen, gegeben durch

$$P(\text{Falsch positive Assoziation}) = 1 - (1 - \alpha)^n \tag{3.2}$$

eine Funktion von n der Anzahl von Tests, die wir durchführen. Da wir für jeden SNP einen Test durchführen und wir mehr als eine Million SNP haben, ist die Wahrscheinlichkeit von SNPs, den fälschlicherweise eine signifikante Assoziation mit dem Phänotyp unterstellt wird, sehr hoch [50]. Um diese Gefahr zu verringern, gibt es verschiedene Korrekturmöglichkeiten für den Signifikanzwert, welche die Anzahl der Tests berücksichtigen. In `PLINK` kann mit dem zusätzlichen Parameter `--adjust` im vorherigen Befehl eine Datei erzeugt werden, welche die Signifikanzwerte nach mehreren Verfahren korrigiert enthält.
Wir verwenden die *False Discovery Rate* (FDR) nach Benjamini & Hochberg [51], wobei wir eine obere Schwelle von 0.1 festsetzen. Damit akzeptieren wir, dass bis zu 10% unserer als signifikant

gefundenen SNPs falsch Positive sein können. Die übliche Korrektur nach Bonferroni ist hier nicht verwendbar, da sie zu streng filtert und nur für Daten geeignet ist, wo es kein Linkage Disequilibrium zwischen den SNPs gibt, was hier nicht garantiert werden kann.

3.3.2 Jensen-Shannon-Divergenz

Neben der Bestimmung der signifikanten SNPs mit `PLINK` haben wir einen neuen Ansatz entwickelt und angewendet, der auf einem Maß der Informationstheorie basiert. Dieses Maß ist die Jensen-Shannon-Divergenz (JSD), welche bereits erfolgreich in anderen Gebieten der Bioinformatik angewendet wurde [52–54].

Theoretischer Hintergrund

Shannon-Entropie
Eine diskrete Zufallsvariable X nimmt Werte aus einem abzählbaren Alphabet A an nach einer Wahrscheinlichkeitsverteilung

$$p(x) = P(X = x) \text{ mit } x \in A \tag{3.3}$$

Für die Wahrscheinlichkeitsverteilung gilt $\sum_{x \in A} p(x) = 1$ und $0 \leq p(x) \leq 1$ für alle $x \in A$.
Die Shannon-Entropie dieser Variable X bzw. der Wahrscheinlichkeitsverteilung p ist definiert als

$$H(X) = -\sum_{x \in X} p_x \log_2 p_x \tag{3.4}$$

Dabei benutzen wir die Konvention, dass $0 \log_2 0 = 0$ gilt, was gerechtfertig ist, da $x \log_2 x \to 0$ gilt, wenn $x \to 0$.
Für ihren Wert gilt folgende Ungleichung

$$0 \leq H(X) \leq \log_2 n, \tag{3.5}$$

wobei n die Größe des Alphabetes angibt.
Die Shannon-Entropie ist ein Maß für die Unsicherheit über den Wert von X. Ist sie gleich oder sehr nahe an 0, können wir mit sehr großer Wahrscheinlichkeit den Wert vorhersagen, den X annimmt. Wenn sie hingegen sehr groß ist, lässt sich keine Aussage über den Wert von X machen [55, Kap. 2].

Relative Entropie
Während die Shannon-Entropie nur Aussagen über eine Wahrscheinlichkeitsverteilung macht, bezieht sich die relative Entropie auf zwei Verteilungen.

Die relative Entropie, auch Kullback-Leibler-Divergenz genannt, ist ein Maß für die Distanz zwischen zwei Verteilungen p und q definiert als

$$D(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)} \tag{3.6}$$

Dabei wird p als die wahre Wahrscheinlichkeitsverteilung angesehen, die durch q approximiert wird. Damit ist die relative Entropie ein Maß für die Genauigkeit dieser Approximation.
Es gelten die Konventionen $0 \log_2 \frac{0}{0} = 0$ sowie $0 \log_2 \frac{0}{q} = 0$ und $p \log_2 \frac{p}{0} = \infty$. Damit ist $D(p||q) = \infty$, wenn es ein $x \in A$ gibt, für das gilt $p(x) > 0$ und $q(x) = 0$.
Die relative Entropie ist immer nicht-negativ und ist gleich 0 genau dann, wenn $p = q$ gilt.
Sie ist kein wahres Distanzmaß, da sie nicht symmetrisch ist und auch nicht die Dreiecksungleichung erfüllt. Auch ist es problematisch für die Verwendung als Distanzmaß, dass sie keine obere Schranke hat [55, Kap. 2].

Jensen-Shannon-Divergenz
Die Jensen-Shannon-Divergenz (JSD) quantifiziert ebenfalls den Unterschied zwischen zwei oder mehr Wahrscheinlichkeitsverteilungen. Ihre mathematischen und statistischen Eigenschaften und Hintergründe wurden ausführlich in [56] analysiert und beschrieben.
Gegeben zwei Wahrscheinlichkeitsverteilungen $p^{(1)} \equiv (p_1^{(1)}, p_2^{(1)}, ..., p_n^{(1)})$ und $p^{(2)} \equiv (p_1^{(2)}, p_2^{(2)}, ..., p_n^{(2)})$ mit den üblichen Bedingungen $\sum_{i=1}^{n} p_i^{(j)} = 1$ und $0 \leq p_i^{(j)} \leq 1$ für alle $i = 1, 2, ..., n$ und $j = 1, 2$ und zwei Gewichte $\pi^{(1)}$ und $\pi^{(2)}$ für die Verteilungen mit den Bedingungen $\pi^{(1)} + \pi^{(2)} = 1$ und $0 \leq \pi^{(j)} \leq 1$ ist die JSD zwischen den Verteilungen definiert als

$$JSD(p^{(1)}, p^{(2)}) \equiv H(\pi^{(1)} p^{(1)} + \pi^{(2)} p^{(2)}) - (\pi^{(1)} H(p^{(1)}) + \pi^{(2)} H(p^{(2)})) \tag{3.7}$$

Dabei ist $H(p^{(j)})$ wie vorher definiert die Shannon-Entropie der Verteilung $p^{(j)}$.
Die JSD kann ebenfalls als eine symmetrische Version der relativen Entropie betrachtet werden, indem man sie definiert als

$$JSD(p^{(1)}, p^{(2)}) = \frac{1}{2} D(p^{(1)}||M) + \frac{1}{2} D(p^{(2)}||M), \tag{3.8}$$

wobei $M = \frac{1}{2}(p^{(1)} + p^{(2)})$ die mittlere Verteilung ist.
Mit $\pi^{(1)}$ und $\pi^{(2)}$ kann eine Gewichtung der Verteilungen vorgenommen werden, um Unterschiede wie z.B. die Anzahl von Daten, aus denen eine Verteilung erzeugt wurde, bei der Berechnung zu berücksichtigen.

Anwendung auf Genotypisierungsdaten

Für jeden SNP in unserer PED-Datei können wir zwei Wahrscheinlichkeitsverteilungen definieren, indem wir die Proben auf Basis ihres Vicin-Gehaltes in zwei Gruppen aufteilen. Für beide Gruppen definieren wir dann

$$P(X) = \frac{\sum_{s \in G} \delta_{sX}}{|G|} \tag{3.9}$$

als die Wahrscheinlichkeit von Allel X, wenn wir nur die Proben betrachten, die zur Gruppe G, das heißt hoher oder niedriger Vicin-Gehalt, gehören. Dabei ist δ_{sX} das Kronecker-Delta definiert als

$$\delta_{sX} = \begin{cases} 1 & \text{falls } s = X \\ 0 & \text{falls } s \neq X \end{cases}$$

Die JSD kann nun als Maß für den Unterschied zwischen diesen beiden Verteilungen sein. Ein geeigneter Marker sollte unterschiedliche Allele in Abhängigkeit vom Phänotyp haben, weswegen wir annehmen, dass signifikante SNPs eine hohe JSD bezüglich ihrer beiden Verteilungen haben sollten.
Um die unterschiedlichen Anzahlen von Proben, aus denen sich unsere Gruppen zusammensetzen zu berücksichtigen, wählen wir als Gewichte π für unsere Verteilungen $\frac{6}{20}$ für die Vicin-arme Gruppe und $\frac{14}{20}$ für die Vicin-reiche Gruppe.

Z-Score

Um zu bestimmen, wann ein JSD-Wert signifikant ist, wandeln wir diese Werte in Z-Scores um, die angeben, wie weit ein bestimmter Wert vom Mittelwert entfernt liegt. Ein Z-Score von 3 gibt zum Beispiel an, dass der ursprüngliche Wert drei Standardabweichungen größer als der Mittelwert ist. Berechnet werden sie mit

$$z = \frac{x - \mu}{\sigma} \tag{3.10}$$

dabei ist x die JSD, μ der Mittelwert der JSD-Werte über alle SNPs und σ die Standardabweichung der SNPs. Bei der Berechnung von μ und σ ignorieren wir die SNPs, die eine JSD von 0.0 haben und vermeiden es damit, die SNPs zu berücksichtigen, welche keinen Unterschied im Genotyp zwischen den Phänotypen aufweisen.
Wir verwenden eine untere Schwelle von 3 für die Z-Scores, um zu entscheiden, ob ein SNP signifikant mit dem Phänotyp assoziiert ist, welche ebenfalls in [57] und [58] verwendet wurde.

3.3.3 Ergebnisse der Methoden

Eine übliche Darstellung der Ergebnisse einer GWAS ist ein Manhattan-Plot, der auf der X-Achse die Position der SNPs in den unterschiedlichen Chromosomen und auf der Y-Achse den negativen

Logarithmus ihrer Signifikanzwerte aufträgt. Da wir aber aufgrund des selbst erzeugten Genomes keine Informationen über Positionen oder Chromosomen haben, verwenden wir für die X-Achse nur einen Index für die SNPs, der nur die Reihenfolge angibt, in der die SNPs in der VCF-Datei vorkommen. Ebenfalls ist für die Z-Scores keine Umwandlung in den negativen Logarithmus notwendig, da bereits gilt, dass je größer der Wert ist, desto stärker ist die Assoziation.

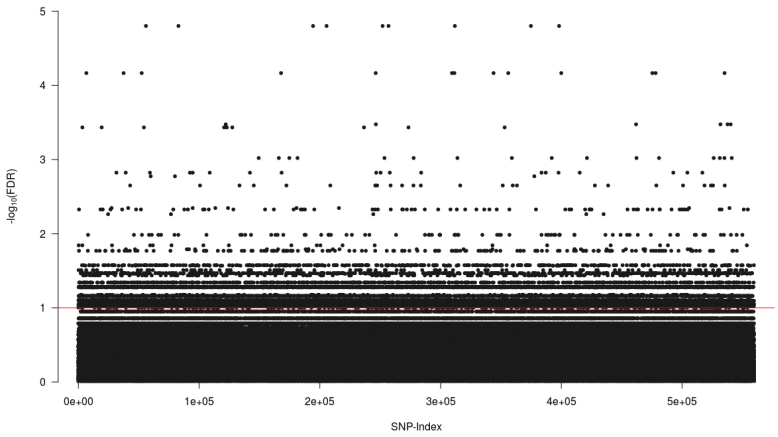


Abbildung 3.3: Manhattan-Plot der SNPs mit ihrem mit `PLINK` bestimmten Signifikanzwert. Die rote Linie markiert den unteren Schwellenwert von 0.1, der erreicht werden muss, damit ein SNP signifikant ist.

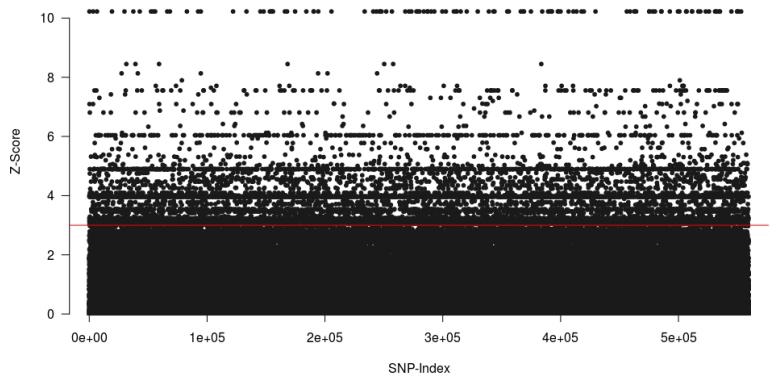


Abbildung 3.4: Manhattan-Plot der SNPs mit ihrem auf Basis von JSD bestimmten Z-Score. Die rote Linie markiert den unteren Schwellenwert von 3, der erreicht werden muss, damit ein SNP signifikant ist.

Mit PLINK finden wir 14.245 SNPs mit einer signifikanten Assoziation mit dem Phänotyp. Die Analyse der JSD hingegen resultiert nur in 10.908 signifikante SNPs. Sowohl im Manhattan-Plot für PLINK (siehe Abbildung 3.3) als auch für JSD (siehe Abbildung 3.4) könnte man versucht sein, einen stärkere Schwellenwert anzulegen, da es sehr viele gibt, die nahe an der unteren Grenze sind und die Ausreißer nach oben prozentual wenig sind. Ein hoher Signifikanzwert reicht aber nicht als alleiniges Merkmal aus, um den SNP als geeigneten Marker zu betrachten.

3.3.4 Untersuchung der signifikanten SNPs

Um geeignete Marker zu bestimmen, untersuchen wir die signifikanten SNPs weiter in Bezug auf ihre räumliche Assoziation mit dem für den Phänotyp verantwortlichen Gen.

Kartierung der SNPs auf das *Medicago truncatula* Genom

Wie in Abschnitt 1.3.2 erwähnt wurde, liegt der für den niedrigen Vicin-Gehalt verantwortliche Locus in einer Region, die Syntänie zu dem Chromosom 2 von *Medicago truncatula* (*M. truncatula*) aufweist. Um zu erfahren, welche gefundenen signifikanten SNPs überhaupt in diesem relevanten Bereich liegen, können wir die SNPs mit ihren jeweiligen flankierenden Sequenzen auf das Chromosom 2 von *M. truncatula* kartieren, da wir für dieses ein Referenzgenom vorliegen haben [59]. Für die Länge der flankierenden Sequenzen eines SNPs verwenden wir 25 Basen. Diese Flanken ergeben mit der eigentlichen SNP-Position die Sequenz der Länge 51 bp, die wir kartieren wollen. SNPs, welche nur kürzere Sequenzen haben, weil sie am Rande einer Konsensussequenz liegen, werden nicht kartiert. Diese Länge wird von der dbSNP des NCBI als Mindestlänge zur eindeutigen Identifizierung des SNP angesehen [60]. Zur Kartierung verwenden wir das Programm Blasr [61] mit dem Aufruf

```
blasr SNP.fa Medicago_chromosome.2.fa -m 0 -minPctIdentity 90 > SNP.mapping
```

mit den Parametern

- SNP.fa: Eingabedatei mit den SNP-Sequenzen im FASTA-Format
- Medicago_chromosome.2.fa: FASTA-Datei mit der Sequenz des Chromosomes 2 von *M. truncatula*
- -m 0: Spezifiziert das Format der Ausgabe
- -minPctIdentity 90: Schwellenwert von 90% Sequenzähnlichkeit
- SNP.mapping: Ausgabedatei

Als Alternative haben wir auch das Programm Exonerate [62] zum Kartieren verwendet. Mit den von uns benutzten Standardeinstellungen verwendet es ein Modell zur Alignierung ohne

Gaps. Für ein solches Modell sind sich aber *Vicia faba* und *M. truncatula* nicht ähnlich genug, was sich auch in der im Vergleich zu `Blasr` niedrigen Anzahl von Kartierungen spiegelt.
Da `Blasr` nahezu alle SNPs kartiert, die auch von `Exonerate` kartiert wurden, und da die Laufzeit wesentlich länger als bei `Blasr` ist, haben wir uns dazu entschieden, dieses wegzulassen und nur `Blasr` zu verwenden.

In den Manhattan-Plots 3.5 und 3.6 haben wir die SNPs farbig hervorgehoben, die sowohl eine signifikanten Assoziation aufweisen, als auch auf Chromosom 2 von *M. truncatula* kartiert werden konnten. Diese sind nur ein Bruchteil aller SNPs mit 17 kartierten, signifikanten SNPs in den `PLINK`-Ergebnissen und 14 SNPs in den Ergebnissen der JSD.
Dies deckt sich mit der allgemein geringen Anzahl von SNPs, die wir erfolgreich kartieren konnten. Von den 685.215 SNPs, die wir untersuchen, haben nur 565.198 vollständige flankierende Sequenzen und können damit kartiert werden. Von diesen SNPs gibt es nur 642, die auf Chromosom 2 kartiert werden können. Der prozentuale Anteil von 0.1136% ist ähnlich zu dem, der bei den signifikanten SNPs erreicht wird (`PLINK`: 0.1193%, JSD: $0,1192\%$).
Kombinieren wir die Ergebnisse von `PLINK` und JSD, so haben wir insgesamt 22 SNPs, die signifikant Assoziation aufweisen und in einem für den Vicin-Gehalt relevanten Bereich liegen.

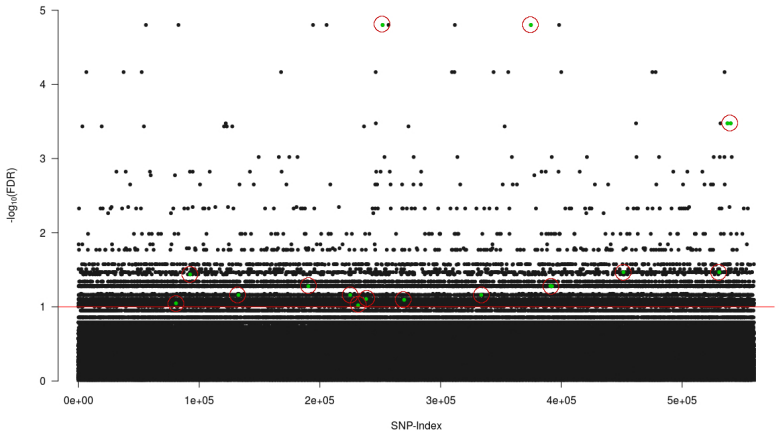


Abbildung 3.5: Manhattan-Plot der SNPs mit ihrem mit `PLINK` bestimmten Signifikanzwert. Die rote Linie markiert den unteren Schwellenwert von 0.1, der erreicht werden muss, damit ein SNP signifikant ist. Die SNPs, die grün gefärbt und rot umkreist sind, sind die signifikanten SNPs, die auf Chromosom 2 von *M. truncatula* kartiert werden konnten.

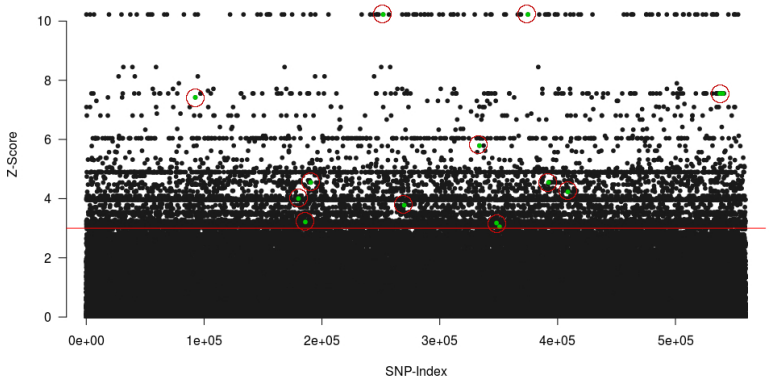


Abbildung 3.6: Manhattan-Plot der SNPs mit ihrem auf Basis von JSD bestimmten Z-Score. Die rote Linie markiert den unteren Schwellenwert von 3, der erreicht werden muss, damit ein SNP signifikant ist. Die SNPs, die grün gefärbt und rot umkreist sind, sind die signifikanten SNPs, die auf Chromosom 2 von *M. truncatula* kartiert werden konnten.

Anwendung von `BLAST` auf die signifikanten SNPs

Eine weitere Möglichkeit zur Untersuchung dieser 22 Marker stellt die Verwendung von `BLAST` [63] dar. Dieses ist ein Tool, welches es uns erlaubt, Sequenzen in den Datenbanken des NCBI zu suchen, die stark ähnlich zu Eingabesequenzen sind. Indem wir unsere Marker in Form der flankierenden Sequenzen damit untersuchen, können wir eventuelle Gene identifizieren, in denen sie liegen. Ein solcher Treffer ist für uns aber für die Eignung des SNPs als Marker nicht notwendig.
Wir verwenden die Online-Version `Nucleotide BLAST` [64] mit Standardparametern und durchsuchen die "Nucleotide collection"-Datenbank, die aus GenBank+EMBL+DDBJ+PDB+RefSeq Sequenzen besteht.

SNP_ID	*Gen*	PLINK	*JSD*
TRINITY_DN67030_c0_g1_i1_111	Medicago truncatula pentatricopeptide repeat-containing protein At4g31070, mitochondrial (LOC11428540), mRNA	+	+
TRINITY_DN67030_c0_g1_i1_157	Kein Treffer	+	+
TRINITY_DN110834_c0_g1_i1_236	Medicago truncatula 3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ (LOC11419786), mRNA	+	+
TRINITY_DN110834_c0_g1_i1_239	Medicago truncatula 3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ (LOC11419786), mRNA	+	+
TRINITY_DN73795_c0_g1_i1_46	Kein Treffer	+	+
TRINITY_DN127846_c0_g1_i1_143	Kein Treffer	+	+
TRINITY_DN106621_c0_g1_i1_87	Kein Treffer	+	+
TRINITY_DN106621_c0_g1_i1_164	Kein Treffer	+	+
TRINITY_DN42433_c0_g1_i1_118	Glycine max phosphoenolpyruvate carboxylase (PEPC4), mRNA	+	+
TRINITY_DN172780_c1_g2_i2_187	Kein Treffer	+	-
TRINITY_DN109808_c0_g1_i1_229	Durio zibethinus monothiol glutaredoxin-S10-like (LOC111307459), mRNA	+	-
TRINITY_DN142168_c1_g3_i1_164	Kein Treffer	+	-
TRINITY_DN37440_c0_g1_i1_63	Kein Treffer	+	-
TRINITY_DN162339_c2_g1_i1_101	Kein Treffer	+	-
TRINITY_DN188270_c2_g1_i4_106	Lathyrus sativus retrotransposon Ty1/copia, LTR	+	-
TRINITY_DN122434_c0_g1_i3_41	Medicago truncatula clone mth2-36c19, complete sequence	+	-
TRINITY_DN98897_c0_g1_i1_103	Kein Treffer	+	-
TRINITY_DN171223_c7_g1_i1_170	Medicago truncatula clone mth2-8m24, complete sequence	-	+
TRINITY_DN112644_c0_g2_i1_83	Kein Treffer	-	+
TRINITY_DN150963_c2_g3_i1_85	Medicago truncatula clone mth2-18j19, complete sequence	-	+
TRINITY_DN139935_c0_g1_i1_148	Kein Treffer	-	+
TRINITY_DN50517_c0_g1_i1_148	Medicago truncatula glutathione S-transferase T2 (LOC25492083), mRNA	-	+

Tabelle 3.3: Ergebnisse der Analyse der kartierten, signifikanten SNPs mit BLAST. Die Spalte *SNP_ID* ist die ID der SNPs, die wir verwenden, und die Spalte *Gen* enthält das Gen, das von BLAST für den SNP gefunden wurde bzw. "Kein Treffer", falls BLAST diese Sequenzen nicht in der Datenbank gefunden hat. Die Spalten PLINK und *JSD* enthalten ein "+", falls der SNP signifikant von der jeweiligen Methode gefunden wurde und sonst ein "-".

Wie Tabelle 3.3 zeigt, konnten wir von den 22 Markern für 12 keine Übereinstimmungen mit `BLAST` finden. Für drei weitere ist die einzige Information, dass sie im Genom von *M. truncatula* liegen, was wir durch unsere Kartierung bereits festgestellt hatten. Die verbleibenden sieben hingegen liegen in den Sequenzen bestimmter Gene. Einschränkend ist hierbei zu sagen, dass bis auf eines sämtliche Gene nur vorhergesagt sind aufgrund von Sequenzähnlichkeiten und nicht tatsächlich überprüft wurden. Zwei von den SNPs (TRINITY_DN188270_c2_g1_i4_106 und TRINITY_DN109808_c0_g1_i1_229) haben eine Übereinstimmung in einer anderen Spezies als *M. truncatula*. *Lathyrus sativus* (Saat-Platterbse) gehört zum selben Tribus *Fabeae* wie *Vicia faba*, womit eine Übereinstimmung der Sequenz nicht erstaunlich ist. Allerdings liegt der SNP nur in einem Retrotransposon und ist damit nicht direkt einem Gen zugeordnet. *Durio zibethinus* (Durianbaum) hingegen gehört zur Ordnung der Malvenartigen und ist damit nicht nahe mit *Vicia faba* verwandt. Da der Biosyntheseweg von Vicin aber noch nicht verstanden ist, können wir nicht sagen, ob diese gefundenen Gene direkt mit dem Vicin-Gehalt zusammenhängen. Dies ist insbesondere fraglich, da *M. truncatula* selber kein Vicin produziert.
Von diesen signifikanten Markern sind insbesondere TRINITY_DN110834_c0_g1_i1_236 und TRINITY_DN110834_c0_g1_i1_239 von Interesse, da sie mit einer Position von ungefähr 1.550.000 bp im Chromosom 2 von *M. truncatula* direkt in der Mitte des in Abschnitt 1.3.2 beschriebenen Bereich des QTLs liegen.

Abbildung 3.7: Plot der SNPs, die sich in der Konsensussequenz TRINITY_DN110834_c0_g1_i1 befinden, für alle 20 Proben erzeugt mit dem `Integrative Genomics Viewer` [65]. Die unterschiedlichen Farben geben an, was für einen Genotyp die jeweilige Probe für diesen SNP hat. Dabei steht grau für homozygot Referenz, türkis für homozygot Alternativ und dunkelblau für heterozygot. Die beiden SNPs am rechten Rand sind die Marker TRINITY_DN110834_c0_g1_i1_236 und TRINITY_DN110834_c0_g1_i1_239.

Wie die Abbildung 3.7 zeigt, gibt es bei diesen eine klare Trennung im Genotyp zwischen Proben mit niedrigem Vicin-Gehalt und Proben mit hohem Vicin-Gehalt. Erstere sind homozygot

bezüglich des Referenzallels, während letztere homozygot bezüglich des Alternativallels sind. Einzige Ausnahme bildet Sample_Hiverna-2, welche ebenfalls homozygot bezüglich der Referenz ist, obwohl sie einen hohen Vicin-Gehalt hat. Davon abgesehen eignen sich diese Marker, um eine Vorhersage über den Vicin-Gehalt zu treffen. Der IGV-Plot zeigt einen weiteren SNP, der die selben Genotypen aufweist wie unsere beiden Marker. Wir ziehen diesen aber nicht in Betracht, weil wir ihn nicht auf Chromosom 2 von *M. truncatula* kartieren konnten.

3.3.5 Vergleich zwischen `PLINK` und JSD

Interessant ist zu untersuchen, wie sich die Ergebnisse zwischen `PLINK` und JSD unterscheiden.

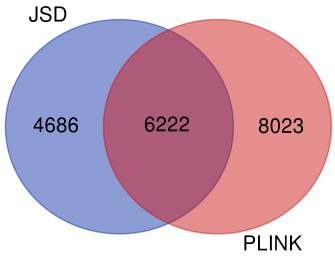


Abbildung 3.8: Venn-Diagramm der signifikanten SNPs bestimmt mit `PLINK` bzw. JSD

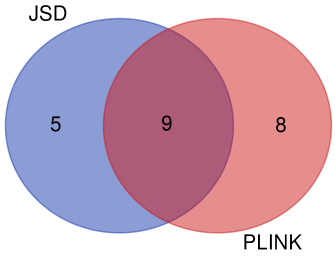


Abbildung 3.9: Venn-Diagramm der signifikanten SNPs, die nach Chromosom 2 von *M. truncatula* kartiert werden konnten, bestimmt mit `PLINK` bzw. JSD

Die Anzahl der gefundenen SNPs unterscheidet sich mit 14.245 (`PLINK`) und 10.908 (JSD) etwas, aber nicht so sehr, dass man gezielt eines der Verfahren verwerfen könnte. Wie Abbildung 3.8 zeigt, gibt es sehr viele SNPs, die nur von jeweils einer Methode als signifikant gefunden werden. Der gemeinsame Anteil beträgt nur 43.67% der *PLINK*-Ergebnisse und 57.04% der JSD-Ergebnisse. Ein ähnliches Verhältnis zeigt sich auch, wenn wir nur die signifikanten SNPs betrachten, die auf Chromosom 2 von *M. truncatula* kartiert werden konnten, wie Abbildung 3.9 zeigt.
Ein solches Verhältnis zeigt sich auch in der Korrelation nach Spearman zwischen den P- und JSD-Werten, die 0.61927 beträgt. Das Verhältnis zwischen P- und JSD-Werten wird in Abbildung

3.10 gezeigt. Die P-Werte wurden mit dem negativen Logarithmus konvertiert, sodass für beide Werte gilt, dass ein hoher Wert einer starken Assoziation entspricht.

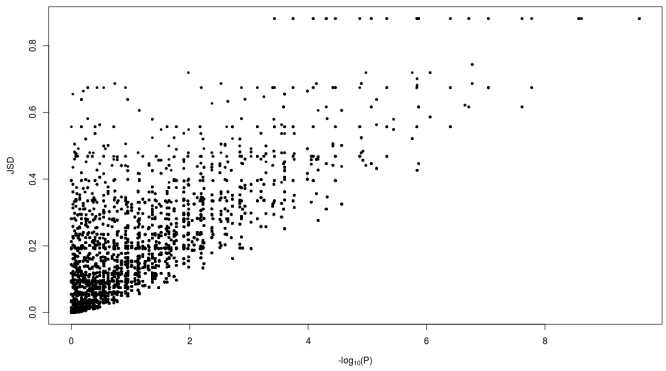


Abbildung 3.10: Scatter-Plot der P-Werte der SNPs nach Anwendung des negativen Logarithmus zur Basis 10 gegen die korrespondierenden JSD-Werte

Die Abbildungen 3.11, 3.12, 3.13 und 3.14 zeigen, wo sich die kartierten SNPs, die nur von dem jeweiligen Verfahren als signifikant gefunden wurden bzw. die von beiden Verfahren als signifikant gefunden wurden, im Manhattan-Plot der beiden Verfahren befinden. Wie man sehen kann, sind die SNPs mit sehr hohen Werten diejenigen, welche von beiden Verfahren als signifikant gefunden wurden. Die SNPs hingegen, welche nur von dem jeweiligen Verfahren als signifikant betrachtet werden, haben niedrigere Signifikanzwerte und liegen nahe am Signifikanz-Schwellenwert.

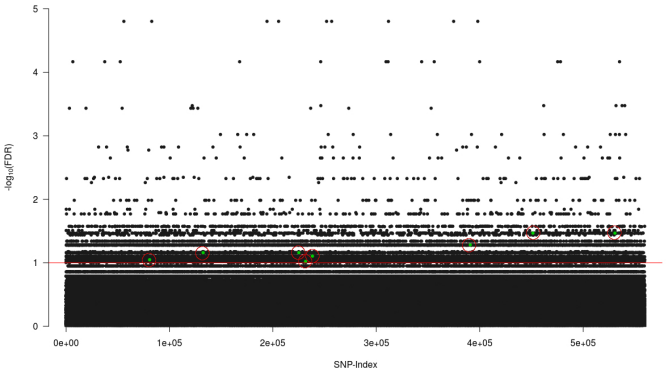


Abbildung 3.11: Manhattan-Plot der SNPs mit ihrem mit `PLINK` bestimmten Signifikanzwert. Die rote Linie markiert den unteren Schwellenwert von 0.1, der erreicht werden muss, damit ein SNP signifikant ist. Die SNPs, die grün gefärbt und rot umkreist sind, sind die kartierten `PLINK`-Signifikanten.

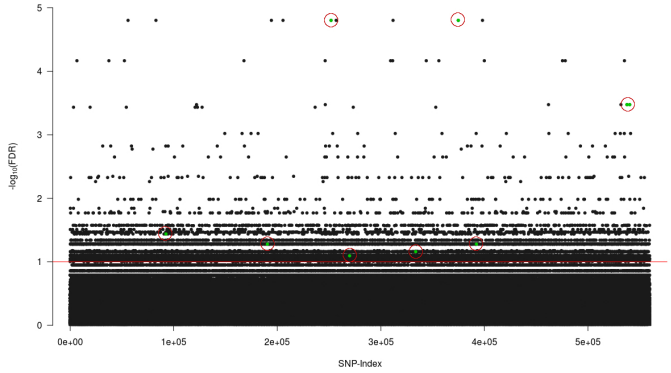


Abbildung 3.12: Manhattan-Plot der SNPs mit ihrem mit PLINK bestimmten Signifikanzwert. Die rote Linie markiert den unteren Schwellenwert von 0.1, der erreicht werden muss, damit ein SNP signifikant ist. Die SNPs, die grün gefärbt und rot umkreist sind, sind die kartierten SNPs, die von beiden Verfahren als signifikant betrachtet werden.

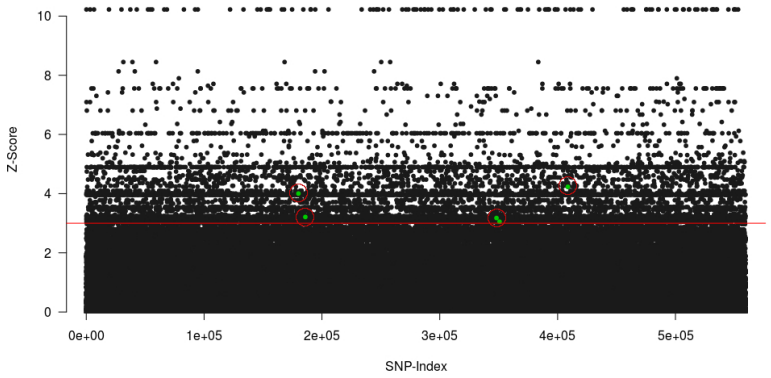


Abbildung 3.13: Manhattan-Plot der SNPs mit ihrem auf Basis von JSD bestimmten Z-Score. Die rote Linie markiert den unteren Schwellenwert von 3, der erreicht werden muss, damit ein SNP signifikant ist. Die SNPs, die grün gefärbt und rot umkreist sind, sind die kartierten JSD-Signifikanten.

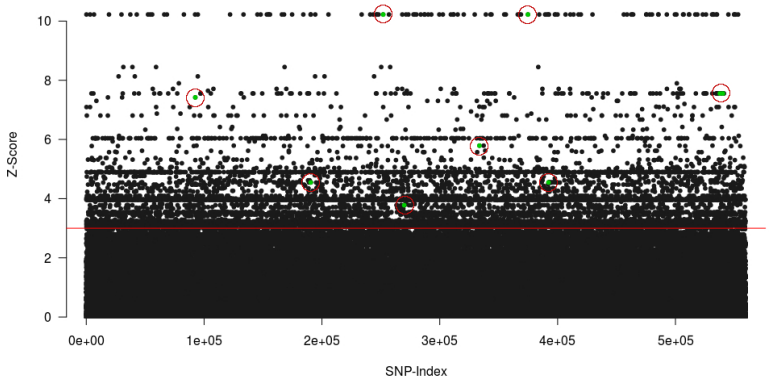


Abbildung 3.14: Manhattan-Plot der SNPs mit ihrem auf Basis von JSD bestimmten Z-Score. Die rote Linie markiert den unteren Schwellenwert von 3, der erreicht werden muss, damit ein SNP signifikant ist. Die SNPs, die grün gefärbt und rot umkreist sind, sind die kartierten SNPs, die von beiden Verfahren als signifikant betrachtet werden.

Um zu verstehen, wieso unterschiedliche SNPs als signifikant von den beiden Methoden gefunden werden, betrachten wir im Detail die Berechnung für zwei SNPs, die nur von jeweils einem Verfahren als signifikant gefunden wurden sind. Wir wählen den SNP aus, der bei dem jeweiligen Verfahren den höchsten Signifikanzwert erreicht hat, aber nicht von dem anderen Verfahren als signifikant betrachtet wird, ungeachtet, ob er kartiert werden konnte.

Untersuchung von SNPs TRINITY_DN168610_c1_g1_i2_92 und TRINITY_DN171987_c3_g5_i1_5

Der SNP TRINITY_DN168610_c1_g1_i2_92 hat einen Z-Score von 8.132 und ist damit sehr signifikant. Mit `PLINK` hingegen erhalten wir nur eine FDR von 0.2079 (unkorrigierter P-Wert = 0.01057), was doch weiter von dem Schwellenwert von 0.1 entfernt ist. Hierbei ist aber zu berücksichtigen, dass im Falle eines einzelnen Testes ein Signifikanzwert unterhalb von 0.05 ausreichen würde, um den SNP als signifikant assoziert mit dem Phänotyp zu betrachten.

	Gruppe		
Allele	Wenig Vicin	Viel Vicin	Summe
G	10	11	21
A	2	17	19
Summe	12	28	40

Tabelle 3.4: Kontigenztafel der beobachteten Häufigkeiten für SNP TRINITY_DN168610_c1_g1_i2_92

Die Tabelle 3.4 zeigt die Allelhäufigkeiten des SNP. Die Vicin-armen Proben haben, bis auf eine, die den Genotyp "AA" hat, alle "GG". Die Vicin-reichen Proben hingegen haben den Genotyp "AA" oder "AG". Somit ist bis auf eine Probe eine Unterscheidung des Phänotypes mit dem SNP perfekt möglich. Damit ist er ein SNP, den wir als signifikant finden möchten.
Die Ursache, wieso wir diesen SNP nicht mit beiden Verfahren als signifikant finden, dürfte neben der Korrektur aufgrund des mehrfachen Testens daran liegen, dass der von uns verwendete Chi-Quadrat-Test in `PLINK` auf den Allelhäufigkeiten basiert und nicht wie die JSD auf den Häufigkeiten der verschiedenen Genotypen. Wenn man nur die Allelhäufigkeiten betrachtet, so gibt es, was Guanin angeht, keinen wirklichen Unterschied zwischen den Phänotypen, wodurch der χ^2-Wert für diesen SNP niedriger ist, als wenn man ihn auf die Genotypen anwendet.
Der zweite SNP TRINITY_DN171987_c3_g5_i1_5 wurde nur von `PLINK` signifikant mit einem Wert von 0.031 gefunden. Mit JSD resultiert es nur in einem Z-Score von 2.857, der damit doch schon nahe an der Schwelle ist.

	Gruppe		
Allele	Wenig Vicin	Viel Vicin	Summe
T	6	28	34
C	6	0	6
Summe	12	28	40

Tabelle 3.5: Kontigenztafel der beobachteten Häufigkeiten für SNP TRINITY_DN171987_c3_g5_i1_5

Dieser SNP hat nur homozygote Genotypen. Die Vicin-reichen Proben haben alle den Genotyp "TT", während drei von den sechs Vicin-armen Proben "CC" haben und der Rest ebenfalls "TT". Damit kann nur der Hälfte der Fälle ein niedriger Vicin-Gehalt korrekt vorhergesagt werden, was eher für einen ungeeigneten Marker spricht.

Kapitel 4

Zusammenfassung

Ziel dieser Arbeit war es, auf Basis von Genotyping-By-Sequencing Daten von *Vicia faba* Pflanzen, SNPs für diese Pflanze zu identifizieren und unter diesen SNPs die zu bestimmen, welche sich als Marker für einen niedrigen Vicin-Gehalt eignen.
Wir haben verschiedene Verfahren durchgeführt und dann für die Ergebnisse der besten Methode eine genomweite Assoziationsstudie (GWAS) durchgeführt, um die Assoziation zwischen Genotyp und Phänotyp zu untersuchen und geeignete Marker zu identifizieren.
Um die größtmögliche Anzahl von SNPs für die weitere Analyse zu erhalten, verwenden die Ergebnisse der Methode, welche die meisten SNPs als Ergebnis hat. Hierbei ist das Multi-Sampling mit `Bowtie2` und `Samtools` des mit `CD-HIT` reduziertem Trinity-Genom deutlich besser als die anderen Verfahren. Dies zeigt sich auch an der hohen Abdeckung, die `Bowtie2` beim Kartieren der Reads auf das partielle Trinity-Genom erreicht und die mit ungefähr 74% wesentlich höher ist als die mit dem Transkriptom erreichte 10%. `Stacks` ist in dieser Hinsicht eine Blackbox und liefert diese Informationen nicht. Auf diese Art haben wir ungefähr 1.9 Millionen SNPs gewonnen, von denen nach Qualitätsfilterung noch ungefähr 685.000 SNPs für die weitere Analyse überbleiben.
Neben der klassischen GWAS mit `PLINK` haben wir dann auch ein eigenes Verfahren auf Basis der Informationstheorie entwickelt und angewendet. Nach der Bestimmung der signifikanten SNPs haben wir diese auf das Chromosom 2 von *Medicago truncatula* kartiert, um so jene zu bestimmen, welche in einem relevanten Bereich des Genomes liegen und damit als Marker geeignet sind. Unser Verfahren liefert mit 14 SNPs vergleichbare Resultate wie `PLINK` (17 SNPs). Insbesondere sind die SNPs mit einem hohen Signifikanzwert bei beiden Verfahren gleich, während sich nur die SNPs unterscheiden, die eine niedrigere Signifikanz aufweisen. Daher können beide Verfahren komplementär verwendet werden.
Insgesamt haben wir 22 SNP erhalten, die als Marker eine starke Assoziation mit dem Phänotyp aufweisen und die auch in der Region des dafür verantwortlichen Genes lokalisiert sind. Der nächste Schritt ist es nun, anhand neuer Pflanzenproben die Vorhersagegenauigkeit dieser Marker zu überprüfen. Falls sie geeignet sind, können sie dann im Rahmen des Projektes Abo-Vici

verwendet werden, um neue Vicin-arme Sorten zu züchten. Des Weiteren können sie auch zur Feinkartierung des für den niedrigen Vicin-Gehalt verantwortlichen Genes verwendet werden und somit gegebenenfalls zum Verständnis der verantwortlichen biochemischen Prozesse beitragen.

Literaturverzeichnis

[1] Torres A, Avila C, Gutierrez N, Palomino C, Moreno M T, Cubero J I. Marker-assisted selection in faba bean (*Vicia faba* L.). *Field Crops Research*, 115:243–252, 2000.

[2] Link W. Autofertility and rate of cross-fertilization: crucial characters for breeding synthetic varieties in faba beans (Vicia faba L.). *Theoretical and Applied Genetics*, 79:713–717, 1990.

[3] Jeroch H, Lipiec A, Abel H, Zentek J, Grela E R, Bellof G. *Körnerleguminosen als Futter- und Nahrungsmittel*. Frankfurt: DLG-Verlag, 2016.

[4] Link W. Züchtungsforschung bei der Ackerbohne: Fakten und Potentiale. *Journal für Kulturpflanzen*, 61:341–347, 2009.

[5] O'Sullivan D M, Angra D. Advances in faba bean genetics and genomics. *Frontiers in Genetics*, 7, 2016.

[6] Köpke U, Nemecek T. Ecological services of faba bean. *Field Crops Research*, 115:217–233, 2010.

[7] Cernay C, Ben Ari T, Pelzer E, Meynard J, Makowski D. Estimating variability in grain legume yields across Europe and the Americas. *Scientific Reports*, 5:11–171, 2015.

[8] Römer A. *Untersuchungen zu Inhaltsstoffen und zum Futterwert von Ackerbohnen (Vicia faba L.)*. 1. Auflage Göttingen: Cuvillier, 1998.

[9] Duc G, Aleksić J, Marget P, Mikić A, Paull J, Redden R et al. *Faba bean*. n: Antonio M. de Ron (Hg.): Grain legumes. New York, Heidelberg, Dordrecht, London: Springer (Handbook of plant breeding, Volume 10), 2015.

[10] Link W, Balko C, Stoddard F. Winter hardiness in faba bean: Physiology and breeding. *Field Crops Research*, 115:287–296, 2010.

[11] Anglade J, Billen G, Garnier J. Relationships for estimating N2 fixation in legumes: incidence for N balance of legume-based cropping systems in Europe. *Ecosphere*, 6:1–24, 2015.

[12] BMEL (2016). Eiweißpflanzenstrategie. Cham: Springer. Online verfügbar unter https://www.bmel.de/DE/Landwirtschaft/Pflanzenbau/Ackerbau/_Texte/Eiweisspflanzenstrategie.html, zuletzt geprüft am 24.07.2018.

[13] Link W. Züchtung und Agronomie neuartiger, Vicin-armer Ackerbohnen und Einsatz als einheimisches Eiweißfutter. Organic eprints. Online verfügbar unter http://orgprints.org/31316/, zuletzt geprüft am 31.07.2018, 2017.

[14] Crépon K, Marget P, Peyronnet C, Carrouée B, Arese P, Duc G. Nutritional value of faba bean (Vicia faba L.) seeds for feed and food. *Field Crops Research*, 115:329–339, 2010.

[15] Ramsay G, Griffiths D W. Accumulation of vicine and convicine in vicia faba and V. Narbonensis. *Phytochemistry*, 42:63–67, 1996.

[16] Gauttam V, Kalia A N. High performance thin layer chromatography method for simultaneous estimation of vicine, trigonelline and withaferin A in a polyherbal antidiabetic formulation. *International Journal of Pharmacy and Pharmaceutical Sciences*, 5:367–371, 2013.

[17] Khazaei H, O'Sullivan D M, Jones H, Pitts N, Sillanpää M J, Pärssinen P. Flanking SNP markers for vicine–convicine concentration in faba bean (Vicia faba L.). *Mol Breeding*, 35:38, 2015.

[18] Brown E G, Roberts F M. Formation of vicine and convicine by Vicia faba. *Phytochemistry*, 11:3203–3206, 1972.

[19] Halle I. *Möglichkeiten der Dekontamination von „Unerwünschten Stoffen nach Anlage 5 der Futtermittelverordnung (2006)"*. Flachowsky | Gerhard, 2006.

[20] Griffiths D W, Ramsay G. The concentration of vicine and convicine in Vicia faba and some related species and their distribution within mature seeds. *Journal of the Science of Food and Agriculture*, 59:463–468, 1992.

[21] Desroches P, El Shazly E, Mandon N, Duc G, Huignard J. Development of Callosobruchus chinensis (L.) and C. maculatus (F.) (Coleoptera: Bruchidae) in seeds of Vicia faba L. differing in their tannin, vicine and convicine contents. *Journal of Stored Products Research*, 31:83–89, 1995.

[22] Bjerg B, Heide M, Knudsen J C N, Soerensen H. Inhibitory effects of convicine, vicine and dopa from Vicia faba on the in vitro growth rates of fungal pathogens. *Journal of Plant Diseases and Protection*, pages 483–487, 1984.

[23] Jeroch H, Drochner W, Simon O, Dänicke S. *Ernährung landwirtschaftlicher Nutztiere*. UTB, Stuttgart, 2008.

[24] Jeroch H. *Futtermittelkunde*. Gustav Fischer Verlag, 1993.

[25] Bjerg B, Eggum B O, Jakobsen I, Olson O, Sorensen H. Protein quality in relation to antinutritional constituents in faba beans (Vicia faba L.). The effect of vicine, convicine and dopa added to a standard diet fed to rats. *Zeitschrift für Tierphysiologie Tierernährung und Futtermittelkunde*, 51:275–284, 1984.

[26] Arscott G H, Harper J A. The relationship of 2,5-diamino-4,6-diketopyrimidine, 2,4-diaminobutyric acid and a crude preparation of -cyano-l-alanine to the toxicity of common and hairy vetch seed fed to chicks. *Journal of Nutrition*, 80:251–254, 1963.

[27] Dänner E E. Einsatz von Vicin/Convicin-armen Ackerbohnen (Vicia faba) bei Legehennen. *Archiv für Geflügelkunde*, 67:249–252, 2003.

[28] Duc G, Marget P, Esnault R, Le Guen J, Bastianelli D. Genetic variability for feeding value of faba bean seeds (Vicia faba): Comparative chemical composition of isogenics involving zero-tannin and zero-vicine genes. *The Journal of Agricultural Science*, 133:185–196, 1999.

[29] Gutierrez N, Avila C M, Duc G, Marget P, Suso M J, Moreno M T, Torres A M. CAPs markers to assist selection for low vicine and convicine contents in faba bean (Vicia faba L.). *Theoretical and Applied Genetics*, 114:59–66, 2006.

[30] Webb A, Cottage A, Wood T, Khamassi K, Hobbs D, Gostkiewicz K, White M, Khazaei H, Ali M, Street D, Duc G, Stoddard F L, Maalouf F, Ogbonnaya F C, Link W, Thomas J, O'Sullivan D M. A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (Vicia faba L.). *Plant Biotechnology Journal*, 14:177–185, 2016.

[31] Khazaei H, Sing G, Stoddard FL, Bett KE, Vandenberg B. Faba bean (Vicia faba L.) breeding for key quality traits - past, present and future. NAPIA meeting. 4-6 November 2015. Niagara Falls, Ontario, Canada.

[32] Khazaei H, Purves R W, Song M, Stonehouse R, Bett K E, Stoddard F L, Vandenberg A. Development and validation of a robust, breeder-friendly molecular marker for the vc^- locus in faba bean. *Molecular Breeding*, 37:140, 2017.

[33] Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet.*, 12:499–510, 2011.

[34] Ortiz R R. Plant Breeding in the Omics Era. 1st ed. 2015. Cham: Springer. Online verfügbar unter http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1066318.

[35] Anderson S B ed(Ed.). *Plant breeding from laboratories to field*. Rijeka: InTech, 2013.

[36] Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla, S. SNP Markers and Their Impact on Plant Breeding. *International Journal of Plant Genomics*, 2012.

[37] Bundessortenamt. Geschützte und zugelassene Sorten. [Online; Stand 29.06.2018].

[38] Gasim S, Abel S, Link W. Ein Beitrag zur Züchtungsforschung an Winterbohnen (Vicia faba L.). Vortrag Pflanzenzüchtung (54).

[39] Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10:R25, 2009.

[40] Li H, et. al. The sequence alignment/map (sam) format and samtools. *Bioinformatics*, 25:2078–2079, 2009.

[41] Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 2013.

[42] Grabherr M, et. al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotechnol.*, 29:644–652, 2011.

[43] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658–1659, 2006.

[44] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*, 28:3150–3152, 2012.

[45] Vicia faba Transkriptom. `https://www.coolseasonfoodlegume.org/analysis/154`, 2017. [Online; accessed 14-March-2018].

[46] Pickrell J K, Pritchard J K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics*, 8:1–17, 2012.

[47] Huson D H, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23:254–267, 2006.

[48] Hirschhorn J N, Daly M J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6:95, 2005.

[49] McHugh M L. The Chi-square test of independence. *Biochemia medica*, 23:143–149, 2013.

[50] Johnson R C, Nelson G W, Troyer J L, Lautenberger J A, Kessing B D, Winkler C A, O'Brien S J. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics*, 11:724, 2010.

[51] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

[52] Capra J A, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23:1875–1882, 2007.

[53] Gültas M, Düzgün G,Herzog S, Jäger S J, Meckbach C, Wingender E, Waack S. Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. *BMC Bioinformatics*, 15:96, 2014.

[54] Dang T K L, Meckbach C, Tacke R, Waack S, Gültas M. A Novel Sequence-Based Feature for the Identification of DNA-Binding Sites in Proteins Using Jensen–Shannon Divergence. *Entropy*, 18:379, 2016.

[55] Cover T M, Thomas J A. *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[56] Grosse I, Bernaloa-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley H E. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E*, 65, 2002.

[57] Meckbach C*, Tacke R, Hua X, Waack S, Wingender E, Gültas M. PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinformatics*, 16:400, 2015.

[58] Dunn S D, Wahl L M, Gloor G B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24:333–340, 2008.

[59] Tang H, Krishnakumar V, Bidwell, S, Rosen B, Chan A, Zhou S, Gentzbittel L, Childs K L, Yandell M, Gundlach H, Mayer K FX, Schwartz D C, Town C D. An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics*, 15:312, 2014.

[60] Kitts A, Phan L, Ward M, et al. The Database of Short Genetic Variation (dbSNP) 2013 Jun 30 [Updated 2014 Apr 3]. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-. Verfügbar unter: https://www.ncbi.nlm.nih.gov/books/NBK174586/.

[61] Chaisson M, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13:238, 2012.

[62] Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005.

[63] Madden T. The BLAST Sequence Analysis Tool. 2013 Mar 15. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-. Verfügbar unter: https://www.ncbi.nlm.nih.gov/books/NBK153387/.

[64] Nucleotide BLAST. `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome`, 2018. [Online; accessed 19-July-2018].

[65] Thorvaldsdóttir H, Robinson J T, Mesirov J P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Bioinformatics*, 14:178–192, 2013.

Appendix

.1 Single-Sampling und Multi-Sampling

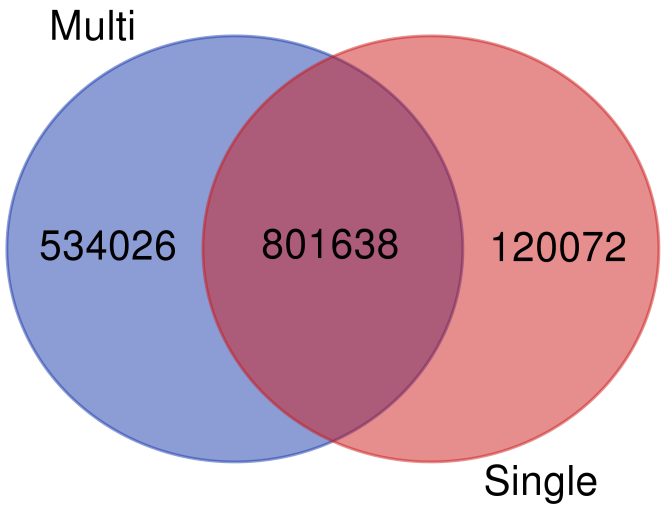


Abbildung 1: Venn-Diagramm der SNPs gefunden mit dem von `Trinity` erzeugtem Genom mit Single- bzw. Multi-Sampling

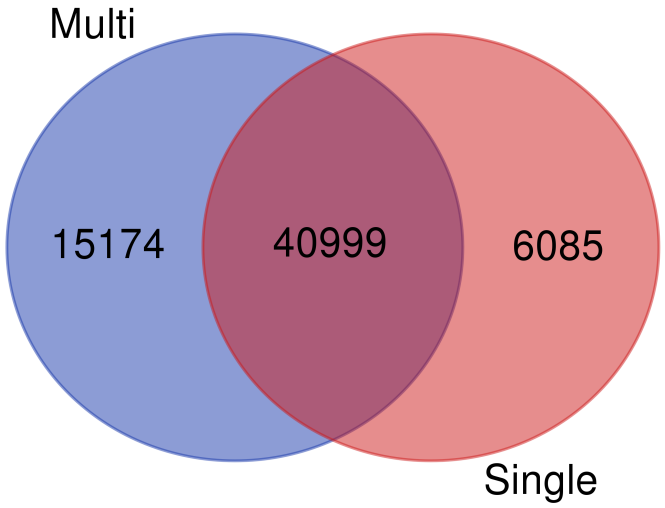


Abbildung 2: Venn-Diagramm der SNPs gefunden mit dem Transkriptom mit Single- bzw. Multi-Sampling

Danksagung

Zum Schluss möchte ich all jenen danken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Masterarbeit beigetragen haben.
Zuerst gebührt mein Dank Professor Armin Schmitt für seine Betreuung während dieser Arbeit und allgemein meiner Arbeit in der Abteilung Züchtungsinformatik.
Außerdem danke ich Professor Wolfgang Link, der uns dieses Projekt zur Verfügung gestellt und mir die Möglichkeit gegeben hat, diese Arbeit zu schreiben.
Ein großes Dankeschön geht an Mehmet Gültas, der mich seit meinem Bachelorstudium in der Bioinformatik unterstützt hat und ohne dessen Hilfe als Betreuer diese Arbeit nicht zustande gekommen wäre.
Bei Rebecca Tacke möchte ich mich bedanken für die zahlreichen Informationen, die sie mir zum Thema *Vicia faba* bereitgestellt hat und für die Beantwortung sämtlicher Fragen, die ich dazu hatte.
Ebenfalls danken möchte ich meinen Eltern, die mir mein Studium ermöglicht haben und die sich zusammen mit meiner Schwester die Zeit genommen haben, diese Arbeit Korrektur zu lesen.

B. Curriculum Vitae

Persönliche Daten

Name:	Felix Heinrich
Geburt:	14.03.1994
Geburtsort	Herzberg
Staatsangehörigkeit:	Deutsch

Wissenschaftlicher Werdegang

2000–2004	Grundschule
2004–2010	Haupt- und Realschule, Hattorf am Harz, Erweiterter Sekundarabschluss I
2010–2013	Berufliches Gymnasium mit Schwerpunkt Wirtschaft, Osterode am Harz, Allgemeine Hochschulreife
2013–2016	Studiengang der Angewandten Informatik mit Schwerpunkt Bioinformatik an der Georg-August-Universität Göttingen mit Abschluss B.Sc.
2016–2018	Studiengang der Angewandten Informatik mit Schwerpunkt Bioinformatik an der Georg-August-Universität Göttingen mit Abschluss M.Sc.
Seit 2018	Doktorand in der Arbeitsgruppe Züchtungsinformatik am Department für Nutztierwissenschaften der Georg-August-Universität Göttingen

Berufserfahrung

2015–2018	Studentische Hilfskraft am Institut für Bioinformatik, Universitätsmedizin Göttingen
2016–2018	Studentische Hilfskraft am Institut für Informatik, Georg-August-Universität Göttingen
2017–2018	Studentische Hilfskraft in der Arbeitsgruppe Züchtungsinformatik am Department für Nutztierwissenschaften der Georg-August-Universität Göttingen
Seit 2018	Wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Züchtungsinformatik am Department für Nutztierwissenschaften der Georg-August-Universität Göttingen

Publikationen

[1] Steuernagel, L.; Meckbach, C.; Heinrich, F.; Zeidler, S.; Schmitt, A.O.; Gültas, M. (2019). *Computational identification of tissue-specific transcription factor cooperation in ten cattle tissues.* PloS one, 14(5), e0216475.

[2] Lange, T.M.; Heinrich, F.; Enders, M.; Wolf, M.; Schmitt, A.O. (2020). *In silico quality assessment of SNPs—A case study on the Axiom® Wheat genotyping arrays.* Current Plant Biology, 21, 100140.

[3] Rajavel, A.; Heinrich, F.; Schmitt, A.O.; Gültas, M. (2020). *Identifying Cattle Breed-Specific Partner Choice of Transcription Factors during the African Trypanosomiasis Disease Progression Using Bioinformatics Analysis.* Vaccines, 8(2), 246.

[4] Heinrich, F.; Wutke, M.; Das, P.P.; Kamp, M.; Gültas, M.; Link, W.; Schmitt, A.O. (2020). *Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning.* Genes, 11(6), 614.

[5] Heinrich, F.; Gültas, M.; Link, W.; Schmitt, A.O. (2020). *Genotyping by Sequencing Reads of 20 Vicia faba Lines with High and Low Vicine and Convicine Content.* Data, 5(3), 63.

[6] Klees S.*; Heinrich F.*; Schmitt A.O.; Gültas M. (2021). *agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species.* Biology, 5(3), 790 (*These authors contributed equally to this work.).

[7] Heinrich, F.; Ramzan, F.; Rajavel, A.; Schmitt, A.O.; Gültas, M. (2021). *MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes.* Biology, 10(9), 921.

[8] Wutke M.; Heinrich F.; Das P.P.; Lange A.; Gentz M.; Traulsen I.; Warns F.K.; Schmitt A.O.; Gültas, M. (2021). *Detecting Animal Contacts—A Deep Learning-Based Pig Detection and Tracking Approach for the Quantification of Social Contacts.* Sensors, 21(22), 7512.