

# **Deciphering the genetic background of quantitative traits using machine learning and bioinformatics frameworks**

Dissertation

zur Erlangung des Doktorgrades  
der Fakultät für Agrarwissenschaften  
der Georg-August-Universität Göttingen

vorgelegt von

Faisal Ramzan  
aus Faisalabad, Pakistan

Göttingen  
2020

Supervisory committee

Prof. Dr. Armin O. Schmitt,  
Breeding Informatics Group, Department of Animal Sciences, Georg-August university  
Göttingen.

Prof. Dr. Henner Simianer,  
Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August  
University Göttingen.

Prof. Dr. Steffen Weigend,  
Federal Research Institute for Animal Health , Institute of Farm Animal Genetics,  
Mariensee.

Date of thesis defense: 26.11.2020

## Abstract

In this thesis, I developed two frameworks that can help highlight the genetic mechanisms underlying quantitative traits. In this regard, my focus was to design efficient methodologies to discover genotype-phenotype associations and then use these identified associations to describe the regulatory mechanism that affects the manifestation of phenotypic differences among the individuals. In the first framework, I investigated key regulatory mechanisms governing the development of eggshell strength. The aim was to highlight the temporal changes in the signaling cascades governing the dynamic eggshell strength during the life of birds. I considered chicken eggshell strength at two different time points during the egg production cycle and studied the genotype-phenotype associations by employing the Random Forest algorithm on genotypic data. For the analysis of corresponding genes, a well established systems biology approach was adopted to delineate gene regulatory pathways and master regulators underlying this important trait. My results indicate that, while some of the master regulators (*Slc22a1* and *Sox11*) and pathways are common at different laying stages of chicken, others (e.g., *Scn11a*, *St8sia2*, or the *TGF- $\beta$*  pathway) represent age-specific functions. Overall, my results provide: (i) significant insights into age-specific and common molecular mechanisms underlying the regulation of eggshell strength; and (ii) new breeding targets to improve the eggshell quality during the later stages of the chicken production cycle.

In my second framework, I combined the Random Forests and a signal detection strategy to identify robust genotype-phenotype associations. The objective of this framework was to improve on the efficiency of single-SNP based association analysis. Genome wide association studies (GWAS) are a well established methodology to identify genomic variants and genes that are responsible for traits of interest in all branches of the life sciences. Despite the long time this methodology has had to mature the reliable detection of genotype-phenotype associations is still a challenge for many quantitative traits mainly because of the large number of genomic loci with weak individual effects on the trait under investigation. Thus, it can be hypothesized that many genomic variants that have a small, however real, effect remain unnoticed in many GWAS approaches. Here, we propose a two-step procedure to address this problem. In a first step, cubic splines are fitted to the test statistic values and genomic regions with spline-peaks that are higher than expected by chance are considered as quantitative trait loci (QTL). Then the SNPs in these QTLs are prioritized with respect to the strength of their association with the phenotype using a Random Forests approach. As a case study, we apply our procedure to real data sets and find trustworthy numbers of, partially novel, genomic variants and genes involved in various egg quality traits.



## Zusammenfassung

In dieser Doktorarbeit habe ich zwei Ansätze verfolgt, mit denen genetische Mechanismen, welche quantitativen Merkmalen zugrunde liegen, aufgezeigt und bestimmt werden können. In diesem Zusammenhang lag mein Fokus auf der Entwicklung effizienter Methoden um Genotyp-Phänotyp Assoziationen zu identifizieren. Durch diese lassen sich im Weiteren regulatorische Mechanismen beschreiben, welche phänotypische Unterschiede zwischen Individuen verursachen. Im ersten Ansatz habe ich Schlüsselmechanismen der Genregulation untersucht, welche die Entwicklung der Bruchfestigkeit von Eierschalen steuern. Das Ziel war es zeitliche Unterschiede der Signalkaskaden, welche die Eierschalen Bruchfestigkeit im Verlauf eines Vogellebens regulieren, zu detektieren. Hierfür habe ich die Bruchfestigkeit zu zwei verschiedenen Zeitpunkten innerhalb eines Produktionszyklus betrachtet und die Genotyp-Phänotyp Assoziationen mithilfe eines Random Forest-Algorithmus bestimmt. Für die Analyse der entsprechenden Gene wurde ein etablierter systembiologischer Ansatz verfolgt, mit dem genregulatorische *Pathways* und *Master-Regulatoren* identifiziert werden konnten. Meine Ergebnisse zeigen, dass einige *Pathways* und *Master-Regulatoren* (z.B. *Slc22a1* und *Sox11*) gleichzeitig in verschiedenen Legephasen identifiziert wurden, andere (z.B. *Scn11a*, *St8sia2* oder der *TGF- $\beta$*  *Pathway*) speziell in lediglich einer Phase gefunden wurden. Sie stellen somit altersspezifische Mechanismen dar. Insgesamt liefern meine Ergebnisse (i) signifikante Einblicke in altersspezifische und allgemeine molekulare Mechanismen, welche die Eierschalen-Bruchfestigkeit regulieren und bestimmen; und (ii) neue Zuchtziele, um die Bruchstärke von Eierschalen vor allem in späteren Legephasen zu erhöhen und somit die Eierschalen Qualität zu verbessern.

In meinem zweitem Ansatz, habe ich die Methode der Random Forests mit einer Strategie zur Signaldetektierung kombiniert, um robuste Genotyp-Phänotyp-Beziehungen zu identifizieren. Ziel dieses Ansatzes war die Verbesserung der Effizienz der Einzel-SNP basierten Assoziationsanalyse. Genomweite Assoziationsstudien (GWAS) sind ein weit verbreiteter Ansatz zur Identifikation genomischer Varianten und Genen, die verantwortlich sind für Merkmale, welche von Interesse sowohl für den akademischen als auch den wirtschaftlichen Sektor sind. Trotz des langjährigen Einsatzes verschiedener GWAS-Methoden stellt die zuverlässige Identifikation von Genotyp-Phänotyp-Beziehungen noch immer eine Herausforderung für viele quantitative Merkmale dar. Dies wird hauptsächlich durch die große Anzahl genomischer Loci begründet, welche lediglich einen schwachen Effekt auf das zu untersuchende Merkmal haben. Daher lässt sich Hypothese aufstellen, dass genomische Varianten, welche zwar einen geringen, aber dennoch realen Einfluss ausüben, in vielen GWAS-Ansätzen unentdeckt bleiben. Zur Behandlung dieser Unzulänglichkeiten wird in der Arbeit ein zweistufiges Verfahren verwendet. Zunächst werden kubische Splines für Teststatistiken und genomische Regionen angepasst. Die Spline-Maxima, welche höher

als die zu erwartenden zufallsbasierten Maximalwerte ausfallen, werden als quantitative Merkmals-Loci (QTL) eingestuft. Anschließend werden die SNPs in diesen QTLs, basierend auf ihrer Assoziationsstärke mit den Phänotypen, durch einen Random Forests-Ansatz priorisiert. Im Rahmen einer Fallstudie haben wir unseren Ansatz auf reale Datensätze angewendet und eine plausible Anzahl, teilweise neuartiger, genomischer Varianten und Genen identifiziert, welche verschiedenen Qualitätsmerkmalen zugrunde liegen.

## **Acknowledgements**

During my PhD study, I was accompanied by many amazing people who paved my way for achieving this important milestone and I thank all of them for their contributions in many different ways.

Especially, I would like to thank Prof. Armin Schmitt for giving me the opportunity to do my PhD in his research group. I was really lucky to have such a helpful and patient supervisor with whom I could talk about problems and ideas at any time. Prof. Schmitt was always open for questions and inspired me to try out new ideas. Thereby, he provided me a warm and welcoming atmosphere to nourish my ideas and opened my mind to new aspects of science. I would also like to thank my second supervisor Prof. Henner Simianer for his support and guidance throughout my degree. He always provided valuable insights and critiques that has helped me grow as a scientist. I also appreciate his role as a facilitator to enroll in this PhD position and for his continuous support as a member of the supervisory committee. Further, I would like to thank Prof. Steffen Weigend who was always very kind and encouraging to me. Over the years, I had many informative discussions with him about different aspect of my research work. I would really like to thank him for being a valuable member of my supervisory committee.

Then, I would like to thank Dr. Mehmet Gültas without whom this path of my PhD would have been very difficult. I am indebted to forward my special thanks to him for providing guidance, support, and feedback at each step of the project. I also thank all of my colleagues from the breeding informatics group for accompanying me throughout my PhD. I would like to thank Ms. Siebert for her help in all the official matters. Special thanks to Yonatan who was there from the very first day of my PhD, to Selina who is co-author of my first paper and was a great help throughout, to Abirami, Hendrik and Thomas for proof reading all of my manuscripts, to Martin with whom I had very nice discussions about Random Forests and spline methods, and, last but not the least, to Felix who was always there to help me with my computational problems. Further, I would like to thank the bachelor and master students Magdalena, Johanna, and Mazharul.

In addition, I would also like to extend my thank to Prof. Beissinger, Dr. Sharifi, and Dr. Malina Erbe for providing some of their precious time to discuss the incurred issues and for providing valuable information. I also have to thank all my friends of the "Mango group" who make my stay in Göttingen memorable and who always motivated me.

Finally, I would like to use this opportunity to thank my parents for their patience and support without which I would not have succeeded to accomplish my dream. Without questioning they always gave their best to support me during my studies and my daily life. I

dedicate this thesis to them. Special thanks to my future wife Maira for her support during this time.

The chicken data used in this study were provided by the "Synbreed - Synergistic Plant and Animal Breeding" project for which I am grateful to the project team. My PhD is funded by the Government of the Punjab, Pakistan, under the project "50 overseas PhD Scholarships for the University of Agriculture, Faisalabad". I would like to acknowledge the administrative team of this project.

I once again want to thank all those who contributed anyhow in my PhD and in my life. Having a PhD has been my dream as long as I can remember so I appreciate all the help and support that is making me realize this dream.



# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Structure of the thesis . . . . .	4
1.2. Impact . . . . .	4
<b>2. Biological Background</b>	<b>7</b>
2.1. Traits, Phenotypes and Genotypes . . . . .	7
2.2. Genomics . . . . .	7
2.2.1. DNA . . . . .	8
2.2.2. Chromosomes and Genes . . . . .	9
2.2.3. Single Nucleotide Polymorphisms . . . . .	11
2.2.4. Genotyping Methods . . . . .	12
2.3. Bioinformatics Databases and Tools . . . . .	13
2.3.1. Ensembl Database . . . . .	13
2.3.2. BioMart Database . . . . .	14
2.3.3. GeneXplain platform . . . . .	14
<b>3. Theoretical background</b>	<b>19</b>
3.1. Genotype-phenotype Association Studies . . . . .	19
3.1.1. Linkage Disequilibrium Measures . . . . .	20
3.1.2. Hardy-Weinberg Equilibrium . . . . .	20
3.1.3. Population Stratification and Relatedness among Samples . . . . .	22
3.1.4. Multiple Testing Correction . . . . .	24
3.2. Cubic Smoothing Splines . . . . .	25
3.3. Random Forests based Feature Selection . . . . .	26
<b>4. Material and Methods</b>	<b>29</b>
4.1. Datasets . . . . .	29
4.1.1. Chicken Dataset 1 . . . . .	29
4.1.2. Chicken Dataset 2 . . . . .	30
4.2. Single-SNP Regression based Association Analysis . . . . .	30
4.3. Analysis Frameworks . . . . .	31
4.3.1. Framework 1: Identification of Regulatory Mechanisms Governing Quantitative Traits using Random Forests . . . . .	31

4.3.2. Framework 2: Combining Random Forests and a Signal Detection Method for the Detection of Robust Genotype-Phenotype Associations . . . . .	33
4.4. Extraction of Candidate Genes . . . . .	38
<b>5. Results</b>	<b>39</b>
5.1. Single-SNP Regression based GWAS . . . . .	39
5.2. Analysis Framework 1 . . . . .	41
5.2.1. Association Analysis Using Random Forests . . . . .	41
5.2.2. Gene Set Analysis . . . . .	42
5.2.3. Identification of Master Regulators . . . . .	44
5.2.4. Identification of Over-Represented Pathways . . . . .	49
5.3. Analysis Framework 2 . . . . .	52
5.3.1. Detection of Genotype-Phenotype Association Using the Combined Framework . . . . .	52
<b>6. Discussion</b>	<b>59</b>
6.1. Methodological Discussion . . . . .	59
6.1.1. Machine Learning Models for Association Analysis . . . . .	60
6.1.2. Combining RF and a Signal Detection Approach . . . . .	60
6.2. Biological Discussion . . . . .	61
6.2.1. Deciphering the Regulatory Mechanisms Underlying Eggshell Strength . . . . .	62
<b>7. Conclusion</b>	<b>65</b>
7.1. Summary . . . . .	65
7.2. Conclusions . . . . .	66
7.3. Outlook . . . . .	67
<b>Bibliography</b>	<b>69</b>
<b>A. Appendix</b>	<b>92</b>
A.1. Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forest . . . . .	92
A.2. Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations . . . . .	111
A.3. R-script to implement signal detection strategy. . . . .	128
A.4. R-script to extract the list of genes corresponding to the important SNPs. . . . .	134

## List of Figures

2.1. Chemical structure of DNA molecule . . . . .	8
2.2. Structure of DNA . . . . .	9
2.3. Structure of DNA . . . . .	10
2.4. Structure of Gene . . . . .	11
2.5. Ensembl database. . . . .	14
2.6. BioMart database. . . . .	15
2.7. The GeneXplain database. . . . .	16
3.1. Cubic smoothing spline . . . . .	27
4.1. Flowchart of the first analysis framework . . . . .	33
4.2. Extension of genotypic data as implemented in Boruta . . . . .	34
4.3. Distribution of Wald-test statistics. . . . .	35
4.4. Application of cubic smoothing spline on the Wald-test statistics. . . . .	36
5.1. Manhattan and Q-Q plots corresponding to eggshell strength and egg weight	40
5.2. Venn diagram depicting the number of genes associated with eggshell strength at Time Point 1 (ESS1), at Time Point 2 (ESS2), and their overlap.	41
5.3. Gene Ontology (GO) treemap for genes associated with eggshell strength at Time Point 1 . . . . .	42
5.4. Gene Ontology (GO) treemap for genes associated with eggshell strength at Time Point 2 . . . . .	43
5.5. Scheme of gene regulatory pathways revealing the top five master regulators at Time Point 1 . . . . .	47
5.6. Scheme of gene regulatory pathways revealing the top five master regulators at Time Point 2 . . . . .	48
5.7. Venn diagram of over-represented pathways . . . . .	49
5.8. Plot representing the linkage disequilibrium (LD) structure inside and around a significant peak associated with Eggshell strength . . . . .	54
5.9. Plot representing the linkage disequilibrium (LD) structure inside and around a significant peak associated with Egg weight . . . . .	56

## List of Tables

5.1. Top 15 Gene Ontology (GO) molecular function terms based on the adjusted $p$ -value for the eggshell strength at Time Point 1 (ESS1). . . . .	44
5.2. Top 15 Gene Ontology (GO) molecular function terms based on the adjusted $p$ -value for the eggshell strength at Time Point 2 (ESS2). . . . .	45
5.3. Significantly over-represented pathways for eggshell strength at ESS1 and ESS2 . . . . .	50
5.4. Significant peaks as defined in Phase 4 of our analysis framework and corresponding quantitative trait loci (QTLs) for ESS1 and ESS2. . . . .	53
5.5. Significant peaks as defined in Phase 4 of our analysis framework and corresponding QTLs for EW. . . . .	55





# 1. Introduction

The importance of genotype-phenotype association studies to understand the genetic basis of traits, either qualitative or quantitative, is well established [1]. Until now, a variety of association studies have been conducted to decipher the genetic architecture of important traits, which led to the identification of a valuable repertoire of genes controlling a range of traits (see the reviews [2, 3, 4]). Finding loci associated with a trait through genome wide association studies (GWASs) is commonly based on single-SNP models that test individual single nucleotide polymorphisms (SNPs) for their association with the phenotype ignoring their dependency on the neighboring SNPs. This statistical design of GWAS seems quite straightforward, yet it entails several challenges including those of population stratification, relationships among the samples, multiple hypothesis testing, and overestimation of SNP effects, among others, as pointed out in previous studies [5, 6, 7, 8].

Linear mixed model (LMM) based approaches that incorporate the covariance structure across individuals have been found most effective in dealing with both the kinship and the population stratification problem [9, 10, 11, 12]. Acknowledging their importance, a series of approaches have been proposed to implement the LMM in the context of GWAS [13]. Similarly, many multiple testing correction methods with a varying degree of strictness have been suggested as possible solutions and some of these have been addressed in [6, 14]. However, the hardest challenge of GWAS still persisting is the lack of power to detect the quantitative trait loci (QTLs) with medium to small effect sizes [15]. The SNPs present either inside or in the vicinity of these QTLs display association strengths which are too small to exceed the statistical significance threshold value after correcting for multiple testing. Consequently, only a small part of the overall variance is captured in a typical GWAS analysis [15]. This inability of GWAS to explain a major proportion of the heritability has been under intensive discussion. Haplotypes can capture the correlation structure of SNPs which is ignored in single-SNP based GWAS approaches. Hence, testing the haplotypes for association looks promising at least in theory. Nevertheless, haplotype based analyses are far from being simple and so far, no clear evidence is available in the literature that the haplotype based tests are more powerful than single-SNP based tests even though this topic has been investigated over the years [16, 17, 18, 19].

To address these limitations, multi-SNP GWAS models were introduced that fit all SNPs simultaneously as random effects in the model [4]. Many implementations of multi-SNP models based on Bayesian as well as LMM frameworks have been developed [20]. Numerous studies have also been conducted to show comparative performance of different

single-SNPs, haplotype and multiple-SNP models along with their different implementations [13, 21, 22, 23, 24, 25]. Recently, the growing application of machine learning approaches in different fields of science has incited their use in assessing the genotype-phenotype association as well [26, 27, 28, 29, 30]. Machine learning methods do not require prior assumptions about the distribution of the SNP effects and thus have been used for a wide variety of traits in humans [31], plants [29] and livestock [32, 33]. Nevertheless, multiple studies have revealed that machine learning algorithms surpass currently available well-known GWAS approaches in identifying genes with small effects on the phenotype [30, 31, 34]. In particular, Random Forests (RF) models have been praised for their ability to analyze a large number of loci simultaneously and to identify promising associations [30, 35]. In this regard, Briec et al. pointed out the efficiency of RF models for analyzing a large number of loci simultaneously and identifying promising associations [30].

Here, it is imperative to note that all the above mentioned methodologies have their advantages and challenges. Among other factors, the success of different association methods is heavily influenced by the genetic architectures of the trait of interest [25, 36]. Given the complexity underlying the genetics of quantitative traits, it is probably not realistic to assume that any single method can retain its statistical power for different genetic architectures [19, 37, 38]. Single-SNP based models are still popular [39, 40, 41, 42, 43] while the RF based methods are gaining importance [44]. However, an increasing number of scientists are recommending the integration of different association methods in order to improve QTL identification and interpretation [45, 46]. In this regard, to bridge the gap between single-SNP and haplotype based analysis, Zhang et al. [47] used a non-parametric spline based technique to integrate multiple single-SNP based test statistics into a single test. Furthermore, Zhang et al. [20] as well as Abed and Belzile [25] suggested the combined usage of single-SNP and multi-SNP methods together for the identification of a robust set of SNPs associated with the complex phenotypes. To combine the advantages of machine learning and parametric GWAS analysis, Nguyen et al. [27], Huang et al. [29] and Schwarz et al. [48] employed a two stage analysis integrating the Random Forests algorithm with single-SNP models. However, the selection of SNPs in one stage and the analysis of the selected SNP in the second step may not account for the hidden structure in the data and can result in inflated SNP effects in the discovery of genotype-phenotype association.

In this thesis, my aim is to describe analysis frameworks that can be used to decipher the genetic background of quantitative traits. In this respect, first I employed a single-SNP regression based GWAS approach which is commonly used to detect genotype-phenotype associations. Then I designed two frameworks to empower the genotype-phenotype association analysis to detect associations that remain undetected in a typical GWAS and to improve the utilization of the detected associations in highlighting the regulatory machinery that underlies important traits. In the first framework, I employed a RF algorithm to assess the relative importance of SNPs regarding their association level with the phenotype. For



the analysis of the genes corresponding to the important SNPs, I adopted a well established systems biology approach and identified age-specific and common key regulatory pathways and master regulators. In the second framework my focus is on the identification of robust genotype-phenotype association signals by combining the important SNPs obtained in different association analyses. For this purpose, I first perform a signal detection strategy using the test statistic values of single-SNP based GWAS analysis for the detection of QTLs. Then I prioritize the important SNPs identified by the RF based algorithm within the QTLs to discover the most robust set of markers.

In order to demonstrate the functionality of my frameworks, I have analysed two different datasets in this thesis. The first dataset contains the eggshell strength (ESS) measured at two different time points during the productive life of a chicken and the second dataset is related to egg weight (EW) in chicken. Here it is also worth mentioning that both of these traits are considered economically important. Today's poultry industry is highly invested in the development of chicken capable of producing more eggs in longer laying cycles [49]. To achieve this production goal, persistency of egg laying with sustained egg quality especially the ESS at all the production stages is crucial [49]. The calcified eggshells not only provide protection against physical damage but also play a crucial role for the development of the embryo by allowing gaseous exchange, abating moisture loss, and supplying calcium for the embryo bone development [50]. Multiple molecular actors involved in the homeostasis and transportation of minerals, especially calcium, the main constituent of the eggshell, have been identified [51, 52]. More than 500 eggshell matrix proteins have also been reported [53, 54] implicating a plethora of genes that knit together the complex protein scaffold and the mineral phase of the eggshell [52, 55]. However, most of these discoveries provide only the genes expressed in a certain segment of the chicken oviduct, the principal organ for egg development, and consequently the overall mechanisms of eggshell development remain illusive. Moreover, similar to other economically important traits, ESS remains relevant throughout the productive life and commonly deteriorates with the age of the chicken [56]. This decline in the eggshell quality remains one of the major reasons for replacing commercial flocks [49]. Hence, understanding the genetic basis of ESS at different laying stages is very important for breeders if they are to extend the laying cycle of chicken. Therefore, an analysis of this trait at different time points during the life of the bird can better delineate its genetics and its molecular mechanisms involved in this dynamic behavior [57]. The egg weight on the other hand has always been an important trait as it not only impacts the egg consumers but it is also known to affect the post-hatch chick weight, fitness and performance [58]. Eggs of extreme size can also hinder the packing and storage of eggs making this traits important for the egg industry. Therefore, the investigation of the genetic architecture of this trait is also important to extend the laying cycle of chicken [59, 60].

My results show that, using my frameworks, I am able to identify important novel markers/genes which could provide new insights into the genetic architecture of these traits. The

knowledge gained in this thesis can be utilized to design breeding strategies to improve the egg quality during the later stages of the chicken production cycle. These findings could: (i) enhance our understanding of the regulatory mechanisms underlying important traits; and (ii) provide novel targets and hypotheses for future breeding strategies.

## 1.1. Structure of the thesis

The organization of the thesis is as follows. In Chapter 2, I provide a brief introduction of the most relevant biological concepts required for any research related to genomics. I further give an overview of the bioinformatics databases used in this thesis. In Chapter 3, I give a brief overview of genotype-phenotype association analyses and the topics essential for a better comprehension of these analyses. In that chapter I also introduce a Random Forest based feature selection technique and cubic smoothing spline strategies that provide the foundation for the frameworks presented in the thesis. I describe the analysis frameworks established in this thesis to decipher the genetic background of quantitative traits in Chapter 4. Thereby, I first present the application of a Random Forest based algorithm for association analysis followed by the method for the identification of age-specific and common key regulatory mechanisms governing the eggshell strength in chicken using Random Forests in Section 4.3.1. In the following Section 4.3.2, I describe the second framework that combines Random Forests and a signal detection method to the robust detection of genotype-phenotype associations. Afterwards, I applied both frameworks to chicken datasets and present the results in Chapter 5. These results as well as the application of the suggested frameworks is discussed in Chapter 6 and finally, I complete the thesis in Chapter 7 by summarizing the thesis and give 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] **Ramzan, F.\***, Klees, S.\*, Schmitt, A. O., Cavero, D., and Gültas, M. (2020). Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forests. *Genes*, 11(4), 464. (\* These authors contributed equally to this work.)
- [2] **Ramzan, F.**, Gültas, M., Bertram, H., Cavero, D., and Schmitt, A. O. (2020). *Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations*. *Genes*, 11(8), 892.

Detailed author contribution of Faisal Ramzan to both journal articles: Participated in the design of the studies. Conducted computational and statistical analyses. Prepared and stud-

ied the GWAS datasets and interpreted the results. Participated in the development of the programming scripts required for the analysis. Prepared the manuscripts.

Further, the author contributed to the following presentations that are related to the topic of the thesis:

#### **Conferences, Workshops, Meetings and Student's thesis**

The author presented topics of this thesis at the following workshops and conferences:

- Oral presentation titled "Identification of weak associations in GWAS using methods of signal detection A case study on chicken eggshell strength" presented at DGfZ Conference 2019, at Justus Liebig University, Giessen, Germany, 2019.
- Poster presentation titled "Identification of weak associations in GWAS analysis using methods of signal detection" at CiBreed conference organized at Georg-August University, Göttingen, Germany, 2019.
- Oral presentation titled "Exploiting linkage disequilibrium in GWAS analysis" presented at the 15th European poultry conference in Dubrovnik, 2018.
- Participation in "SNPpit workshop" at Friedrich Loeffler Institut, Mariensee, Germany, 2017.
- Poster presentation titled "Identification of targets for gene editing using efficient mapping strategies" at the Bioinformatics poster day held at the Max Planck institute, Göttingen, Germany, 2017.

In collaboration with Mehmet Gültas and Armin O. Schmitt the author supervised the following student works:

- Magdalena Kircher: *Genomic Prediction of Economically Relevant Traits in Livestock using Machine Learning*. Master's Thesis, 2020.
- Magdalena Kircher: *Comparison on genomic predictions using different statistical methods in a simulated cattle population*. Project Work, 2019.
- Md Mazharul Islam: *Genome wide association (GWA) analysis for identification of markers associated with eggshell thickness using reverse regression based methods*. Master's Thesis, 2018.



## **2. Biological Background**

In this chapter, I will briefly review the basic terminology at a level appropriate for understanding the research methods that are used in this thesis. This chapter also introduces the databases and analysis programs used in the thesis. Much of the description of the basic genetics concepts was adapted from [61, 62]. For a more detailed presentation of the biological concepts, readers are referred to textbooks like [61, 62, 63, 64].

### **2.1. Traits, Phenotypes and Genotypes**

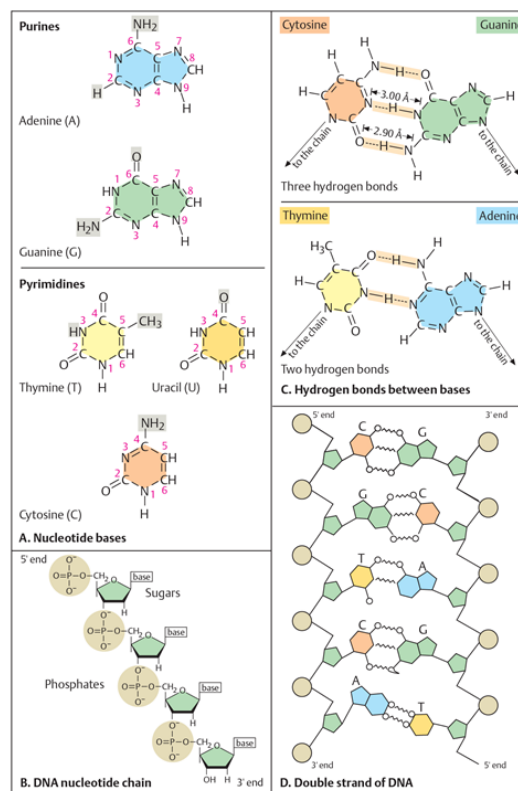
For the characterization of an individual, its appearance, performance or a combination of both is utilized. All characteristics that can be observed or measured on an individual are called traits while the observed categories and the measured levels of traits are called phenotypes [65]. The phenotypes of a qualitative trait are expressed in categories while the phenotypes of quantitative traits are continuous numbers. The genetic background of a phenotype in an individual is influenced by the genotype it carries. The genetic makeup of an individual is called its genotype. In livestock genetics, the term genotype can also be used to refer to all the genes and gene combinations that affect the traits having some economic importance in a given production system. Qualitative traits are usually affected by only a few genes. In contrast, quantitative traits are polygenic i.e. they are affected by many genes.

### **2.2. Genomics**

Genomics is the study of the genome which constitutes the entire DNA content of an individual [66]. After the successful completion of the Human Genome Project [67], the daunting task of sequencing an individual's entire genome started to ease off. Nowadays, much faster and less expensive DNA sequencing methods have been developed and are being used to study the genomes of a variety of species including those having agricultural importance. Although the terminology described in the following sections (of this chapter) is not species specific, the chicken genome being the focus in this thesis will be used as an example, wherever required.

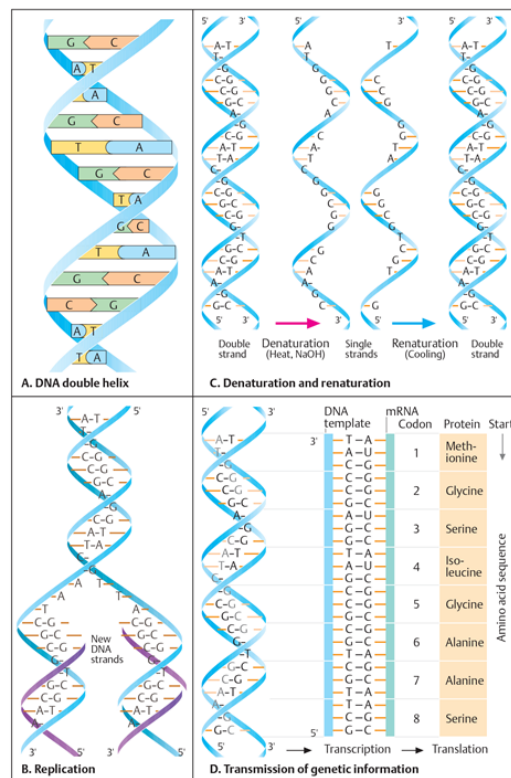
### 2.2.1. DNA

Deoxyribonucleic acid (DNA), also known as the "master molecule" of life belongs to the nucleic acid family of biological molecules. Nucleic acids are biopolymers formed from the building blocks called nucleotides. In a DNA polymer chain, nucleotides link up to form a backbone of alternating deoxyribose sugar and phosphate groups upon which nucleic acid bases are attached. The chemical structures of all constituents of DNA are shown in Figure 2.1. Each sugar molecule has a phosphate group attached to its 3' carbon while the 5' carbon is attached by the neighboring sugar molecule to give the strand of DNA its direction where the terminal sugar at one end of the DNA strand has a free 3' carbon, and the terminal sugar at the other end has a free 5' carbon. The orientation of the 3' and 5' carbons along the sugar-phosphate backbone confers a direction (sometimes called polarity) to each DNA strand. In DNA, four different nitrogen bases are found. Adenine (A) and guanine (G) which consists of two-carbon nitrogen rings are called purines while thymine (T) and cytosine (C) contain a one-carbon nitrogen ring and are known as pyrimidines [61].



**Figure 2.1.:** Chemical structure of DNA molecule and its constituents. (taken from [61]).

Each DNA molecule is made up of two DNA strands which are organized in a double helix form with intertwined sugar-phosphate backbones and nucleic acid bases pointing into the center of the helix (see Figure 2.2). These two strands of DNA are held together by hydrogen bonding between opposing bases. This bonding is specific and A always develops a double bond with T while G bonds with C by a triple hydrogen bond. This type of bond specificity makes the two strands of DNA complementary to each other, a feature that is crucial to DNA's function as the storage molecule for genetic information [63].

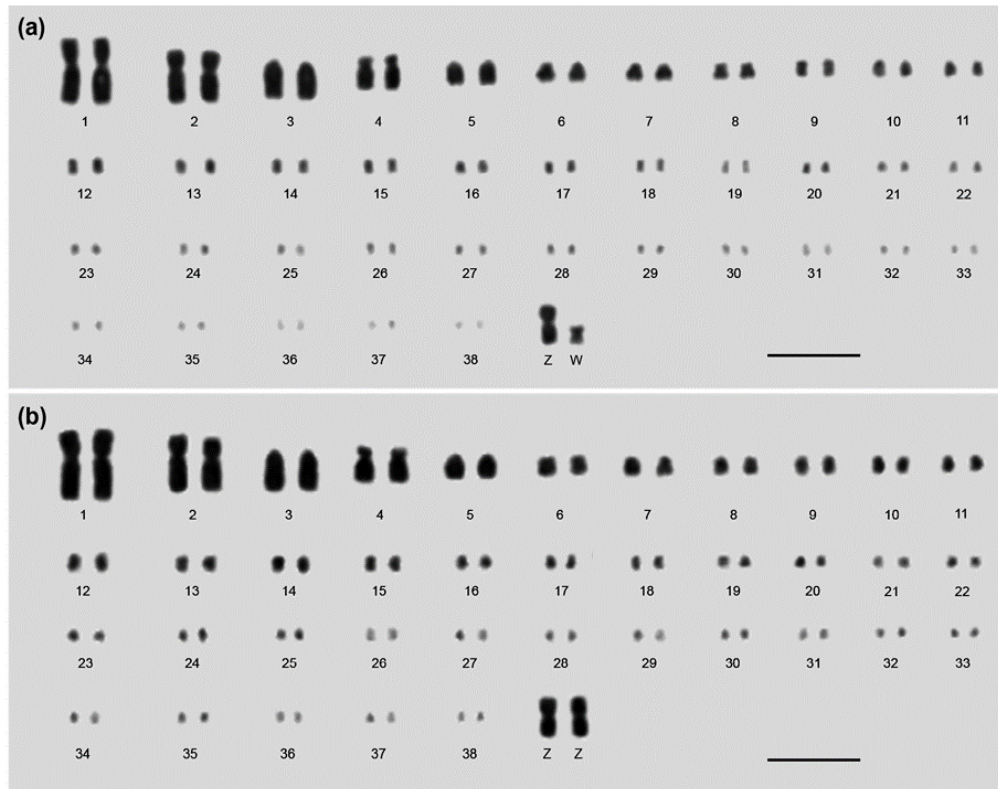


**Figure 2.2.:** Structure of DNA. Three dimensional double helical structure of DNA and transfer of genetic information from DNA to proteins (taken from [61]).

### 2.2.2. Chromosomes and Genes

The DNA molecules and their associated proteins adopt a complex configuration inside the nucleus of a cell, termed chromatin. During the metaphase stage of a cell cycle chromatin can be observed in the form of structures known as chromosomes. Each species has a certain number of chromosomes. For example, in chicken the whole genome consists of 78 ( $2n = 78$ ) chromosomes, classified as macro-chromosomes, micro-chromosomes and sexual Z and

W chromosomes [68]. The chromosomes come in pairs of homologous chromosomes, one derived from the mother, and one from the father.

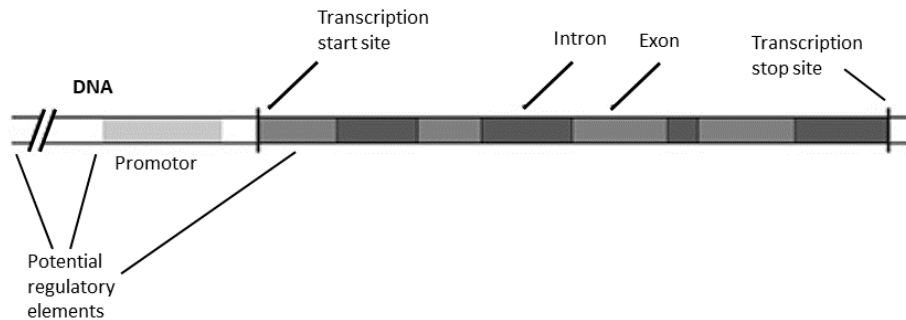


**Figure 2.3.:** Chicken chromosomes. Representative karyogram of (a) the female and (b) the male chicken. The karyograms shows 76 autosomal and ZZ (male) or ZW (female) sexual chromosomes (taken from [68]).

A DNA segment that codes for a functional molecule and consequently holds the genetic information of an organism is termed gene. Genes are arranged linearly along each chromosome with each gene having a defined position, called gene locus [61]. In complex multi-cellular organisms, each coding gene has coding regions separated by non-coding sequences (introns) and regulatory sequences that control gene expression (Figure 2.4) [63]. To produce functional molecules, the decoding of the genes is done in a process termed gene expression which is separated into transcription and translation phases. During transcription, the gene sequence is transcribed into a ribonucleic acid (RNA) sequence, which in the next phase translated into an amino acid sequence (Figure 2.2). RNA is also a nucleotide polymer but unlike DNA, it is single-stranded and contains ribose sugar as its core molecule. In RNA molecules the nitrogenous base thymine is also replaced by the single-



ring base uracil [69]. After processing RNA by removing the introns, the exons are spliced together to form messenger RNA (mRNA) which acts as a template in the transcription phase to arrange the amino acids in the sequence specified by the genetic code [61].



**Figure 2.4.:** Gene structure. The figure shows the structure of a gene in an animal or plant (adapted from [63]).

The mutation events that occur in nature can change the DNA sequence and create alternate forms of genes known as alleles. Multiple alleles of a gene can exist in a population. In an individual, a locus can be occupied by any form of that gene. Since the chromosomes occur in pairs, loci and the alleles occupying them also occur in pairs. If a gene locus contains the same allele on both of homologous chromosomes, the individual is said to be homozygous for that gene and if the alleles are different, the individual is said to be heterozygous [63].

### 2.2.3. Single Nucleotide Polymorphisms

In a genome, different types of sequence variants can be identified and utilized as genetic markers. Among these naturally occurring observable variations, single-nucleotide variants (SNVs) are the most common and ubiquitous throughout the genome [62]. The SNVs are single nucleotide changes in a DNA sequence that is otherwise conserved across individuals [66]. These SNVs originate due to a number of endogenous and exogenous sources of damage that cause the single base pair substitution mutations in DNA [70]. The occurrence of a new variant at any given chromosomal location is a rare event and the vast majority of SNVs carried in some organism's DNA today must have been originated generations ago and then passed on generation after generation to reach the current population. However, the SNVs that arose in somatic cells are not inherited through the generations and are only confined to a very small number of direct descendant cells. The variants that originate in the sex cells are inherited through generations and can also be in theory traced all the way back to the original ancestral mutation. These inherited variants are present in all the cells of a

multi cellular organism and depending upon whether the variant was inherited from one or both parents, an individual can carry either one or two copies of the inherited variant [62].

By convention, in order to classify an SNV as a single-nucleotide polymorphism (SNP), the least abundant allele of that variant is required to have a frequency of 1% or more in the population [71]. However, since the SNV frequency is generally population dependent, the distinction between SNV and SNP is somewhat arbitrary [62]. Although SNPs could in principle be bi-, tri- or tetra-allelic polymorphisms, the most commonly used SNPs in the genomic analysis are bi-allelic, having two possible alternate nucleotides. This substitution of nucleotide bases is of two types which are known as transition and transversion. In a transition either a pyrimidine (C or T) is replaced by a pyrimidine or a purine (A or G) is replaced by a purine while in a transversion a pyrimidine is substituted by a purine or vice versa [72]. Considering the two strands of DNA as equivalent and utilizing the abbreviations used in [71], the allelic nucleotides X and Y of a SNP on one DNA strand can be represented as  $X \Leftrightarrow Y$  ( $X_1 \Leftrightarrow Y_1$ ), with the complementary nucleotides  $X_1$  and  $Y_1$  shown in parenthesis. The four SNP alternates include one transition  $C \Leftrightarrow T$  ( $G \Leftrightarrow A$ ) and three transversions  $C \Leftrightarrow A$  ( $G \Leftrightarrow T$ ),  $C \Leftrightarrow G$  ( $G \Leftrightarrow C$ ), and  $T \Leftrightarrow A$  ( $A \Leftrightarrow T$ ) [62].

#### 2.2.4. Genotyping Methods

The characterization of an individual's genotype can be achieved by genotyping the several million SNPs that exist across the genome [73]. DNA hybridization based approaches are utilized for SNP interrogation using SNP arrays sold commercially by companies like Illumina and Affymetrix [74]. For this allele-specific hybridization-based genotyping, probes that are specific for the portion of the genome containing a SNP, are created. These probes are reverse complementary to DNA around that SNP. Generally, two probes, varying only at one base pair and each one complementary to one of the two possible alleles, are required. The length of the probes used in different genotyping platforms can vary. A probe of length 20 has  $4^{20}$  different possible probe sequences. For the probe length of 20, there is only approximately a 1 in 1,000 chance that the same 20 base sequence will occur more than once in a genome of length  $3 \times 10^9$  base pairs, considering equal frequency for the four bases in the genome. Each probe is labeled with a fluorescent dye of a specific color. After hybridization, the intensity of the emission of each dye indicates whether a specific sample is homozygous for one or the other SNP allele (so that only one color dye will be emitted) or whether it is heterozygous (both dyes will be emitted). An alternate strategy of utilizing two probes strategy is the single base pair extension method that uses a single probe and two different dyes are used as labels [62].

When using a large-scale SNP array, hundreds of thousands of SNPs are to be called for a study, automated algorithms are employed to call SNPs. A considerable number of algorithms for automatic genotype calling using raw dye intensities have been described [74, 75]. Generally, it is more difficult to call rare alleles than common alleles since this

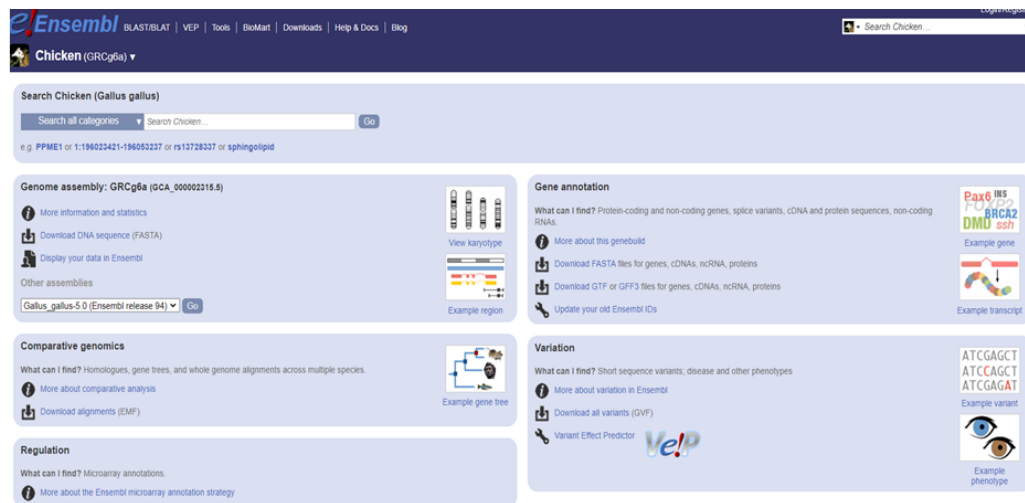
requires identification of the one or two clusters which are either very small or completely absent. Due to the fact that the two strands of DNA are complementary, it is sufficient to specify the sequence of only a single strand to depict the sequence of DNA. However, it is important to note that this complementary nature of the double-stranded structure of the DNA molecule renders the identification of the alleles of an SNP inherently ambiguous [62]. For example, if on one strand a transversion occurs that replaces G with T then due to the complementarity, C on the second strand will be replaced by A. So probes have to be designed to detect both G/T and C/A polymorphism.

## 2.3. Bioinformatics Databases and Tools

Enormous amounts of raw data generated by genomic research are one of the hallmarks of modern genomic research. To manage and share this deluge of data, sophisticated computational methodologies are required. A database is used to store and organize data in a way that makes information easily retrievable. Although the main purpose of a database is to arrange, store and manage the data in a way that makes the retrieval of information easy, biological databases also identify connections between pieces of information that were not known when the information was first entered. These features facilitate the discovery of new biological insights from raw data [76]. In this section, I will briefly describe the bioinformatics databases and tools that I have used in this thesis.

### 2.3.1. Ensembl Database

A genome can contain tens of thousands of genes. However, without the determination of locations of and relationships between individual genes, the genomic information alone is of little use. Contrary to the laborious manual annotation based on scientific journals and public databases, an automated annotation workflow can facilitate the discovery of genes and their functions greatly. Ensembl is one of the well-known systems for the management and retrieval of genomic information [77]. Launched after the first releases of the draft human genome, this database has developed into a centralized resource for geneticists that integrates genomic information for a large number of organisms including chicken (Figure 2.5) [78, 79]. The Ensembl project, through a collection of software pipelines for gene annotation, creates a set of predicted gene locations and makes these data freely accessible [80]. Originally designed to store and distribute the reference assembly produced by the Human Genome Project [67], the Ensembl databases are utilized today to store the assembly structure, genomic sequence, genome annotations, genome alignments, epigenomic data, and regulatory annotation as well as other comparative genomics information [80, 81, 82]. Furthermore, owing to the collaboration with other major databases and resources such as UniProt [83], GENCODE [84], UCSC, and NCBI [85], this database is integral to most bioinformatics analyses [86].



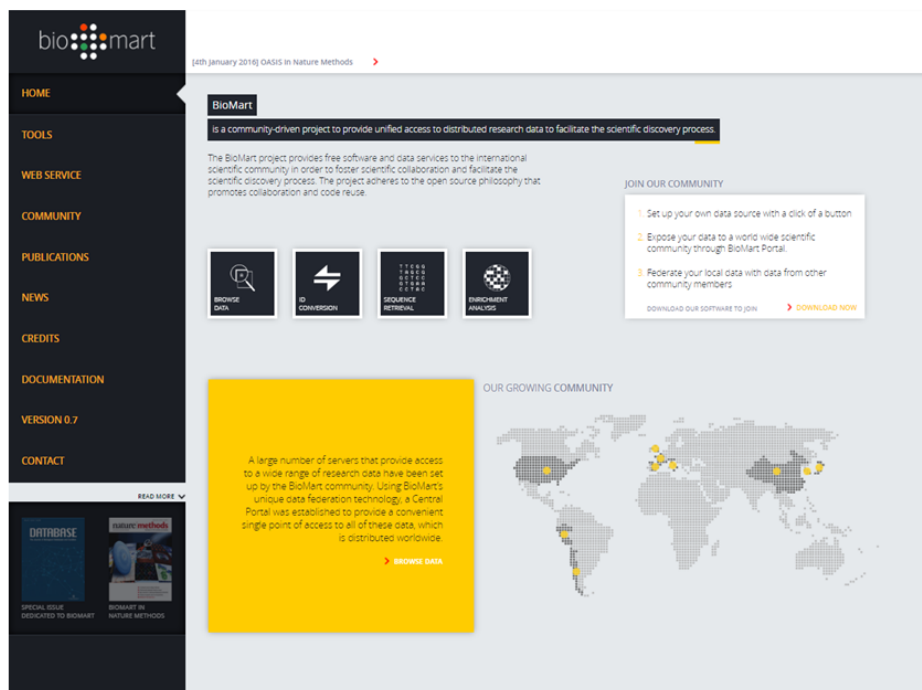
**Figure 2.5.:** Ensembl database. An overview of the Ensembl database. (Source: [http://www.ensembl.org/Gallus\\_gallus/Info/Index](http://www.ensembl.org/Gallus_gallus/Info/Index), 05-10-2020)

### 2.3.2. BioMart Database

In the current era, the volume and complexity of data that is being generated and deposited into databases are increasing. To make the best use of this knowledge, it is imperative to make complex queries to retrieve specific data. Typically, the query interfaces provided by different databases are quite specific. Getting familiar with these interfaces can be time-consuming and if more than one data source is needed to be queried, using different interfaces can be challenging. To overcome this problem, BioMart is a solution that can be used as a generic software to tap data from many databases [87]. In this regard, the Ensembl BioMart provides access to gene annotations, variation data, functional and regulatory annotation (Figure 2.6) [88]. It has been created using the database schemes and data generated under the Ensembl project. The Ensembl BioMart consists of four main databases, namely Ensembl Genes, Ensembl Variation, Ensembl Regulation, and Ensembl Vega. These databases are further supported by information including sequence data, ontology data and karyotype data from secondary databases of the PRIDE [89] and Reactome [90, 91] projects [88]. Given its importance and to make it easier to use BioMart, the external software packages have also incorporated BioMart querying capabilities into their systems [87]. These packages include Galaxy [92], Taverna [93], Cytoscape [94], and BioConductor [95].

### 2.3.3. GeneXplain platform

GeneXplain is an integrated systems biology platform [96] that provides a toolbox and workflow management system for a broad range of bioinformatics and systems biology ap-

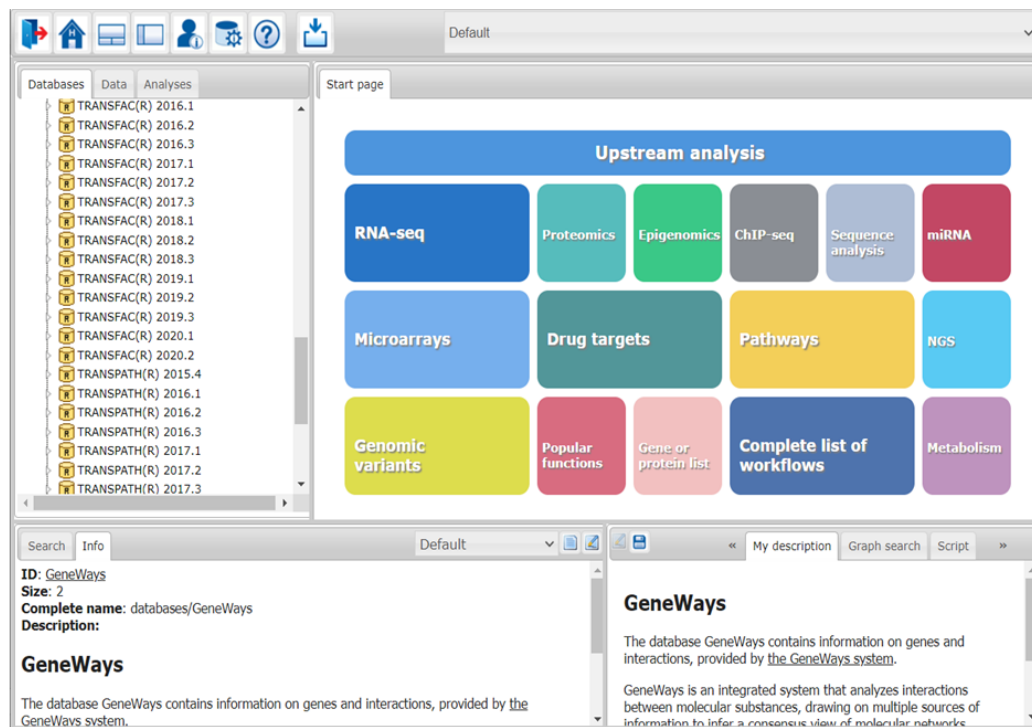


**Figure 2.6.:** An overview of BioMart database. (Source: <http://www.biomart.org>, 05-10-2020)

plications [97]. Among different analyses that can be performed with this platform, molecular network analysis or pathway enrichment, identification of network and signaling regulators, analysis of transcription factor binding sites, methods to test for enrichment of Gene Ontologies [98] or Gene Set Enrichment Analysis [99] are prominent. However, the upstream analysis [100] is considered to be the most known analysis framework implemented in this platform for the identification of causal biomarkers, playing a role in the network of gene regulation and signal transduction [96]. For these different types of analyses, GeneXplain source multiple databases and bioinformatics tools, among which the TRANSFAC<sup>®</sup>, TRANSPATH<sup>®</sup>, and Geneways databases and the master regulators search algorithm are used in this thesis.

### 2.3.3.1. TRANSFAC<sup>®</sup> database

TRANSFAC<sup>®</sup> is a database that stores data relevant for gene expression at the transcriptional level [101]. Created in 1988 [102] TRANSFAC<sup>®</sup> is now hosted by the geneXplain platform (<http://genexplain.com/>). The database is considered an encyclopedia of transcrip-



**Figure 2.7.:** An overview of GeneXplain database. (Source: <https://genexplain.gwdg.de/bioulweb/>, 05-10-2020)

tional gene regulation which along with providing vast information on transcription factors, proteins that regulate the expression of genes, also acts as a comprehensive analysis tool for the identification of potential transcription factor binding sites (TFBSs)[103]. TFBSs are DNA sequences in the vicinity of a gene that and play a role in gene regulation by binding to the transcription factor [104]. To perform their regulatory functions, TFs can either activate/repress transcription of a certain gene, or increase/decrease the level of its transcription. TFs can also alter the chromatin structure by histone or DNA modifications. In the TRANSFAC<sup>®</sup> database, structural and functional properties for each of the transcription factors are enlisted.

### 2.3.3.2. TRANSPATH<sup>®</sup> Database

TRANSPATH<sup>®</sup> is a database that provides information on gene-regulatory pathways [105]. The database is an information system harbouring protein-protein interactions as well as protein modifications involved in signal transduction. It primarily focuses on signaling molecules and the pathways that involve these molecules and regulate transcription factors

(TFs), i.e. proteins that can bind to specific DNA sequences and regulate gene expression [106]. The database contains experimentally verified information on transcription factors and provides the possibility to obtain complete signaling pathways from ligand to target genes and their gene products [107].

For a given list of genes, the TRANSPATH<sup>®</sup> database can be used to analyze the data in two directions. One can perform a so-called downstream analysis that identifies the metabolic or regulatory pathways provoked by the given set of genes [108]. Whereas the upstream analysis sets forth the regulatory pathways that activate the given genes and allows a causal analysis of co-expressed genes that are potentially under the common regulatory influences [100].

#### **2.3.3.3. GeneWays Database**

GeneWays is an open platform with an integrated system that combines several sources of biological information to infer a consensus view of molecular networks [109]. GeneWays provides a large database that contains computationally predicted as well as experimentally identified pathways.

#### **2.3.3.4. Pathway Analysis**

Biological pathways are complex networks that constitute series of interactions among a variety of molecules that occur in a well-curated manner and can initiate the assembly of new molecules or cause some specific change in the cell [90]. The most common types of biological pathways include metabolic, genetic, and signal transduction pathways [110]. The genetic regulatory networks constitute a collection of molecular regulators that govern the gene expression by interacting with each other and with other substances in the cell [111]. The regulators can be DNA or RNA segments, proteins or complexes of these [112] and the interactions between these regulators can be inductive (an increase in the expression level of one will lead to an increase in the other) or inhibitory (increase in one leads to a decrease in the other). Furthermore, the interactions can be direct or indirect through transcribed RNA or translated proteins [111]. An increasing amount of evidence is getting available that underlies the role of biological pathways and networks as a hallmark of the manifestation of complex traits [110]. Therefore, delineating these pathways can play an important role in highlighting the molecular mechanism underlying biological processes.

Following the genotype-phenotype association analysis that produces a list of (identified) genes, comes a laborious literature search for the functional importance of the genes and to put these individual genes in their correct biological context [113]. Due to this reason, the biological interpretation of the associated variants and genes remains challenging [114]. Pathway analysis which is also known as pathway enrichment analysis can be employed to meet this challenge of biological interpretation based on prior knowledge of genes and

pathways [115, 116]. Especially for the analysis of quantitative traits where the number of genes associated with a phenotype are typically large, these methods are routinely used to summarize the large gene list into a smaller list of more easily interpretable pathways consolidating the effects of multiple individual genes [113]. These pathways are then also statistically tested to identify those that are enriched in the provided gene list more than would be expected by chance [113]. The significantly enriched pathways can then be used to empower the elaboration of the genetics of complex traits [114].

### **2.3.3.5. Master Regulators**

The term master regulator was first coined over 30 years ago to denote a gene that occupies the very top of a regulatory hierarchy [117]. As the term was initially proposed for the regulation of a sex determination mechanism which is considered to be independent of prior regulatory influences, master regulatory genes were defined not to be under the regulatory influence of any other gene [118]. However, the meaning of this term has evolved over the years to now designate a gene that is regulating multiple downstream genes either directly or through a cascade of gene expression changes and can force cells to deviate from their normal tasks [117]. These master regulatory genes can code for transcription factor proteins, which alter the expression of downstream genes in the pathway [119]. Thus, in the signal transduction hierarchy, the master regulators serve as important regulatory molecules and are necessary to attain a cell state required for the orchestration of different phenotypes. Therefore, revealing such master regulators is important for the proper understanding of the genetic process underlying important traits [120]. Moreover, these genes can also be excellent candidates for modifying complex traits [121]. For the identification of master regulators in this thesis, an upstream analysis strategy is utilized that has been presented in [100]. In this analysis strategy, potential transcription factor binding sites for the provided list of genes are inferred based on a state of the art promoter analysis using the TRANSFAC<sup>®</sup> database. The corresponding TFs along with the information from TRANSPATH<sup>®</sup> as well as GeneWays databases are then used to construct upstream signal transduction networks. The convergence points of these networks are then identified as potential master regulators [100, 122].



## 3. Theoretical background

In this chapter, I will briefly introduce the concepts and the terminology related to genotype-phenotype association analysis. Here I will review the topics only briefly and for more detailed description readers are referred to textbooks like [62, 66, 123, 124, 125]. Further, I will also explain the basic concept of the smoothing spline and Random Forests based features selection methodologies which provide the foundations of the frameworks explained in the next chapter of this thesis.

### 3.1. Genotype-phenotype Association Studies

Elucidating the genetics and biological mechanisms that manifest the trait differences between individuals is of immediate interest to scientists in the field of genomics. For qualitative traits, molecular differences among the individuals that exhibit different phenotypes, are relatively simple. However, quantitative traits are multi-factorial and typically emerge from the actions of a complex molecular machinery constituting the interactions between multiple molecular elements at different biological levels. Delineating these differences between individuals at different molecular levels is challenging due to this complexity. The availability of omics data offers an opportunity to search for genomic patterns that are different among different individuals. The search for such patterns is not based on mechanistic hypothesis testing, rather it is based on the testing of the biological attributes for their association with the observed individual differences which in turn can lead to the discovery of mechanisms that create these differences [66].

Genome-wide association studies (GWASs) are based on the analysis of genomic data to identify SNP variants associated with differences between samples. A remarkable range of discoveries that has been facilitated by the GWASs are proof of the importance of this strategy in population and complex-trait genetics as well as for understanding the biology of diseases [3]. The methodology of association analysis has evolved from the techniques used to screen the genome for regions having some effect on a phenotype, exploiting the linkage and linkage disequilibrium (LD) between a small number of markers. Especially in livestock genomics, mapping quantitative trait loci (QTLs) in family based designs was the focus of many studies and resulted in thousands of identified QTLs [124]. The availability of large-scale genotyping technologies at a decreasing cost facilitated the boom of GWASs. This methodology now involves hundreds of thousands of markers distributed almost equally across the genome, hence called genome-wide association study.

The concept of LD provides the base for the success of GWASs. It is related to chromosomal linkage where SNPs present on the same chromosome remain physically joined through many generations of a family. This definition of LD also depicts the somewhat artificial difference between linkage and LD mapping, as LD between a marker and a QTL is required for both types of the analysis [126]. Linkage analysis only considers the within family LD structure, while LD mapping considers the LD between a marker and the QTL across population [6]. Nevertheless, the success of most association studies is based on LD between the functional mutations and markers in a certain region of the genome. A GWAS uses the LD at the population level and requires tens or even hundreds of thousands of markers. However, GWASs yield a much greater precision than family based QTL studies because all historical recombinations are utilized in this analysis to identify the causal mutations [124]. In the following sections of this chapter, I will present the general principles and concepts that are important while performing an association analysis.

### 3.1.1. Linkage Disequilibrium Measures

LD is a property of SNPs that describes the correlation of one SNP with another SNP within a population [6]. Multiple measures of LD are in use. The measure used in this thesis is  $r^2$ , a statistic proposed by Hill and Weir [127]. Assume A and B are the two bi-allelic markers. Marker A has alleles A and a and marker B has alleles B and b. Let the observed frequencies of these alleles be  $f_A$ ,  $f_a$ ,  $f_B$ , and  $f_b$ . The four possible haplotypes are AB, Ab, aB and ab with haplotype frequencies denoted by  $f_{AB}$ ,  $f_{Ab}$ ,  $f_{aB}$  and  $f_{ab}$ , respectively.  $r^2$  can be calculated using the following formula.

$$r^2 = \frac{(f_{AB} \times f_{ab} - f_{Ab} \times f_{aB})^2}{f_A \times f_a \times f_B \times f_b} \quad (3.1.1)$$

Values of  $r^2$  range from 0, for a pair of loci with no linkage disequilibrium between them, to 1 for a pair of loci in complete LD.

### 3.1.2. Hardy-Weinberg Equilibrium

Another concept used in this thesis in relation to the genotype-phenotype association is Hardy-Weinberg equilibrium (HWE). This principle states that under a number of assumptions that ensure the absence of evolutionary influences affecting a population, allele and genotype frequencies in that population remain constant across generations. These evolutionary influences include the followings [128].

- Genetic drift
- Mate choice
- Assortative mating
- Natural selection

- Sexual selection
- Mutation
- Gene flow
- Meiotic drive
- Genetic hitchhiking
- Population bottleneck
- founder effect
- Inbreeding

If these assumptions are satisfied, the marginal distribution of the number of copies of a given allele observed in a single individual will follow a binomial distribution having the mean parameter equal to the frequency in the population of that allele [62]. In principle, the HWE mainly implies that the probability of an allele occurring on one chromosome does not depend on which allele is present on its homologous chromosome [123].

In the context of genotype-phenotype association, it is generally assumed that deviations from HWE are the result of genotyping errors, hence testing SNPs for HWE is performed to ensure the exclusion of such SNPs from the analysis. The two tests generally applied to test SNPs for their departure from HWE are Pearson's  $\chi^2$ -test and Fisher's exact test. In this regard, being computationally advantageous, the  $\chi^2$ -test is commonly employed using the formula

$$\chi^2 = \sum_{k=1}^n \frac{(O_k - E_k)^2}{E_k} \quad (3.1.2)$$

where  $n$  denotes the number of alleles of a SNP. This test is performed independently for all SNPs to test the null hypothesis that the observed genotype frequencies do not differ significantly from those expected under HWE. After performing this test, only those SNPs are retained for further analysis for which the SNP genotype frequencies are in HWE [129]. For this purpose a p-value threshold of  $1 \times 10^{-6}$  is commonly used [130, 131, 132].

### 3.1.2.1. Minor Allele Frequency

Minor allele frequency (MAF) is the frequency of the least occurring allele at a specific position in a given population [129]. Contrary to normal statistical notation of using this term to refer to a count, here the term frequency denotes a population proportion [123]. The power of analysis to detect the genetic effects of individual SNPs is dependent on MAF especially in experiments involving SNP-arrays containing hundreds of thousands of SNPs with a wide distribution of MAFs [133]. Most association studies are underpowered to detect the effects of SNPs with a low MAF therefore SNPs having MAF lower than a set threshold are excluded from the dataset. In this regards, MAF thresholds of 0.01 and 0.05 are commonly used [129].

### 3.1.3. Population Stratification and Relatedness among Samples

An assumption implicit in conventional association analysis has been that all subjects under study are unrelated. However, in most cases this condition is not fulfilled as the datasets can have some structure among the studied individuals. This structure at the genetic level of samples is mainly caused by the presence of population stratification or relatedness among the samples due to family structure or cryptic relatedness. Population stratification implies the presence of multiple sub-populations in study samples. In other words, the individuals being studied to determine genotype-phenotype associations belong to more than one population. The relatedness among the samples, on the other hand, entails the sharing of genetic material among the samples due to their family relatedness. Both of these confounding factors have to be accounted for properly in an association analysis as the relationship structure among the samples creates confounding bias if due to stratification and leads to variance distortion if due to either family based or cryptic relatedness which results in inflated summary statistics and consequently inflated false positive rates [8, 134].

To disentangle these inflationary effects from the true effects, many approaches have been developed and are being implemented successfully. A genomic control parameter also known as the genomic inflation factor  $\lambda$  is computed to know if the confounding effects of sample structure exist in the association statistics [134]. Furthermore, the integration of principle component analysis (PCA) with linear mixed model based approaches that incorporate covariance structure across individuals thus accounting for both kinship and population stratification have been found most effective in dealing with the problem of population structure [11, 26]. All three of these concepts are briefly explained below.

#### 3.1.3.1. Genomic Inflation Factor

An approach that is widely used to evaluate the presence of confounding due to population stratification is known as genomic control. In this approach, a genomic inflation factor ( $\lambda_{GC}$ ) is calculated by dividing the median of the observed association statistic by the expected median under the null distribution [135]. A value of  $\lambda_{GC} = 1$  indicates no confounding effects, whereas  $\lambda_{GC} > 1$  points towards the presence of some stratification or other confounders in the data [134]. It is important to note that along with the population stratification, higher genomic inflation factors can also be caused by following factors [136, 137, 138].

- Strong LD between SNPs
- Strong association between SNPs and phenotypes
- Systematic bias

Therefore,  $\lambda_{GC} < 1.05$  are generally considered acceptable [134]. As this genomic inflation factor is a mere ratio of quantiles, a quantile-quantile plot is commonly used for visual depiction of this genomic inflation.

### 3.1.3.2. Principle Components Analysis

Principal components analysis (PCA) is one of the commonly used methods for detecting and visualizing the population structure [62]. This technique has been effectively used for addressing population structure in association studies [139]. For the computation of principal components, spectral or eigenvector/eigenvalue decomposition of the covariance or the correlation matrix is performed. The important characteristics of the variability of the matrix can be captured by only a few principal components which are called leading principal components [62]. However, for a massive SNP dataset, where a fairly large number of samples are genotyped for hundreds of thousands of SNPs, PCA can be performed by extracting the eigenvectors of the kinship matrix of samples [134]. Although it is a non-standard methodology for PCA, there is actually a close relationship between the principal components calculated from the correlation matrix of the genotype data and the eigenvectors of the relationship matrix [62]. In a typical association study where the number of genotyped SNPs is always much larger than the sample size, using the kinship matrix to extract eigenvectors is much more computationally efficient.

### 3.1.3.3. Linear Mixed Models

Initially, developed and utilized for the breeding value estimation in animal breeding, the linear mixed models (LMM) are finding their use in association analysis. The application of LMM in GWAS has become a standard approach especially for samples having some form of population structure [57, 131, 132]. The LMM is being implemented to test each SNP individually for its association with the phenotype in single-SNP based models or these are used to fit all SNPs simultaneously as random effects in the multi-SNP models [4]. In either case, LMM based association analysis is the method of choice for many researchers due to their ability to simultaneously incorporate population stratification and cryptic relatedness as well as the family structure while keeping the type I error at an acceptable level [140].

In LMM based approaches, a genetic relationship matrix (GRM) is constructed which is used as a random effect in the model to account for genome wide sample structure. The contribution of the sample structure to phenotypic variance is estimated and then used in the computation of association statistics. To account for population structure, markers with large allele frequency differences between populations receive a larger correction. In the case of the relatedness structure, the contribution of related individuals to the test statistics values will be reduced, preventing the overweighting of redundant information due to the correlation structure [141]. The LMM based association methods also provide an increase in power by applying a correction that is specific for the sample structure [11, 142]. Additionally, these models can also be used in studies without sample structure to increase power as has been pointed out in [141] and [143].

Furthermore, a variety of different software implementations of LMM based approaches have been developed to make these computationally efficient [13]. Different implemen-

tations of LMM differ in the way they estimate the phenotypic variance explained by the GRM. In this regard, the LMM implementations can be divided into exact and approximate methods [141]. In exact methods, the variance explained by the GRM is estimated separately for each candidate SNP [144] while the approximate methods can use the approximated variance components estimated only once based on all the SNPs [11, 142]. The approximate methods are usually considered computationally faster than the exact methods. However, efficient exact methods have also been developed [12, 145].

#### 3.1.4. Multiple Testing Correction

Multiple hypothesis testing is the other issue faced while performing GWAS which originates due to simultaneous testing of a large number of SNPs to determine their association with the phenotype [146]. The application of a typical point-wise error rate of 0.05 to classify a test result significant becomes impractical as in doing so the experiment-wise error rate will increase with the number of tests, incurring a higher than acceptable number of false positives. A correction measure is applied to limit this error probability to an acceptable level. However, the use of an overly conservative correction approach tends to overlook some of the true positives, thus reducing the power of the analysis while being too lenient would increase the number of false positives [5]. Many methods with varied strictness have been suggested as possible solutions and some of these methods have been addressed in [6] and [14]. Permutation testing is one such method to establish significance in GWAS [6]. It provides a straightforward way to generate the empirical distribution of test statistics under the null hypothesis by randomly shuffling the phenotype among the individuals in the data [147]. In doing so, any association between the genotype and phenotype is broken to satisfy the null hypothesis of no association but the LD structure of the data is retained and the false-positive rate under null is approximated. The permutation tests are considered gold standard for association analysis. However, these methods are computationally expensive and for a large number of SNPs, can be impractical [148]. Another attractive way of multiple testing correction is to calculate the effective number of independent tests. The main idea of this calculation is to filter out correlated SNPs and then use only the effective number of independent ones. After that, the Bonferroni correction method can be applied for correction by replacing the total number of SNPs with the effective number of SNPs. In this respect, a method proposed by Gao et al. [149] use principal component analysis to derive the effective number of tests. This method which is named simpleM, has been shown to provide a better estimate of independent SNPs and has been validated using the GWAS datasets. In this thesis, for multiple testing correction, the simpleM method is primarily used along with a permutation based approach.

## 3.2. Cubic Smoothing Splines

To quantify the expected change of a response variable  $y$  given an explanatory variable  $x$ , classical regression analysis is commonly applied. In which case, the expected value of  $y$  given an  $x$  value can be expressed as  $E(y|x) = f(x)$ . Here,  $f(x)$  represents the underlying function that describes the relationship between the two variables, hence specification of this function is crucial. The simplest form that  $f(x)$  can assume is linear. However, in most cases, it is unlikely for two variables to have a linear relationship over the whole measured range and the linearity is approximated sometimes as a necessity and sometimes for convenience as it makes the analysis and interpretation easier [150]. Contrary to linearity, other functional forms are also possible which are usually expressed by higher-order polynomials [151]. Higher order polynomials are quite flexible in modeling changes in means and variances along a continuous scale. But these polynomials are sensitive to the extreme values on either side of the scale which can cause a major change in the overall form of the  $f(x)$ . This is a problem that occurs frequently and can result in erratic and implausible estimates of parameters. To overcome this problem and to avoid the extensive search for the form of  $f(x)$ , non- or semi-parametric modeling methods can be applied [151]. In this regard, as an alternative to high degree polynomials, spline curves that are constructed from joining pieces of lower degree polynomials are suggested [152]. These lower degree polynomials are joined at selected points known as knots to construct the overall form of  $f(x)$ .

For the design of spline curves, we have to first define the basis functions. The choice of the ideal basis for approximating the function is very important and data-dependent. Some of the most popular spline basis include the truncated power series basis, Fourier basis, wavelet basis, cardinal spline basis, and B-spline basis [153, 154]. To estimate smooth curves of non-periodic data observed with some noise and having continuous derivatives up to certain order, B-spline basis is preferred [153]. Better numerical properties compared to other basis functions also make the B-spline basis more appropriate for smoothing and semi-parametric models [155, 156]. For a more detailed description of spline functions readers are referred to [157, 158, 159].

Genomic data obtained from microarray and sequencing experiments consist of values measured in relation to chromosomal coordinates. These values can be thought of as describing functions in a space parameterized by the chromosomal coordinates and can thus be analyzed as piecewise-polynomial curves [160]. In the case of data consisting of SNP markers, any measurement observed for individual SNPs can be considered coming from the observation of a stochastic process having an underlying functional form. For the reconstruction of this functional form, the discrete values observed for a finite set of coordinates can be used to construct sample curves which can then be joined to obtain the complete functional form. For each sample curve, the coefficients of its basis could be estimated from discrete noisy observations.

The B-spline basis functions can be utilized for this purpose by designing different types of

splines. Smoothing splines are one such type of penalized splines that are used to reduce the residual sum of squares of the fitted curve on the observed values by adding more flexibility to the regression line without allowing too much overfitting. In order to do this, in its most basic form, the smoothing spline considers each SNP position as a knot. This, however, can lead to overfitting of the model. To avoid that, curves are fitted using the so-called roughness penalties that ensures a good fit to the data (reduce the residual sum of squares) and also controls the degree of smoothness. The roughness in the curve is quantified by the integrated squared second derivative and a continuous penalty is obtained. A higher order of derivative can also be used to control the degree of smoothness of the curve [153]. The most commonly used method of smoothing spline is based on cubic B-spline functions having knots at all sampling coordinates. A cubic spline is a piecewise cubic polynomial function with the constraints of having a continuous function and continuous first and second derivatives at the knots between two adjacent segments [161] (Figure 3.1).

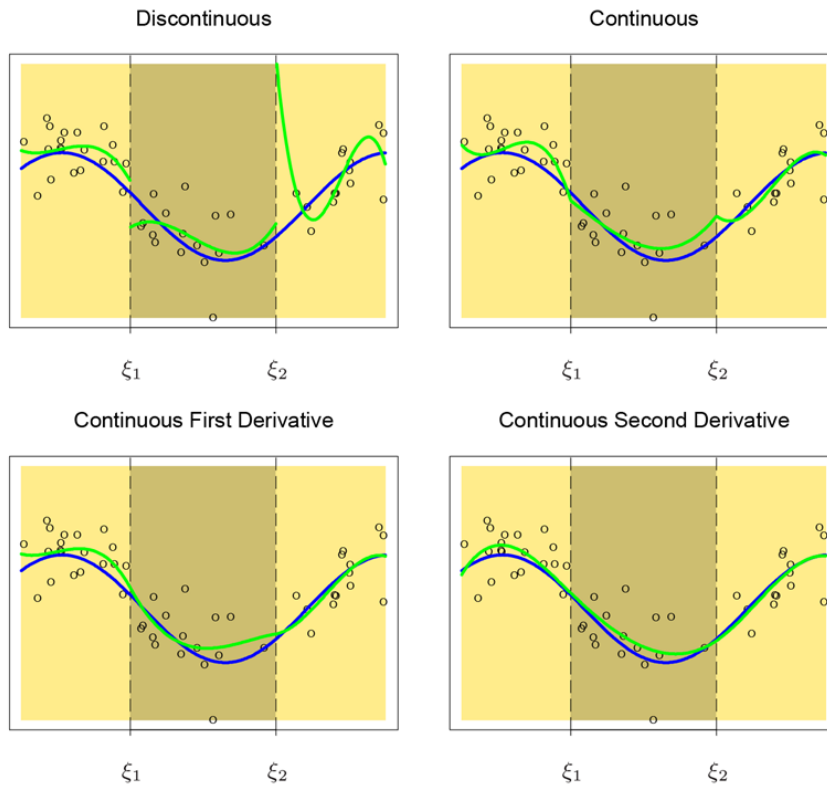
However, given the very large number of sampling points, a lower number of appropriate knots can be sufficient to smooth the curves and capture their main features [153]. It is also important to note that on the one hand having a large number of knots can provide overfitting, on the other hand, having a very small number of knots leads to underfitting. To avoid this problem, the optimum number of knots can be determined through cross-validation.

### 3.3. Random Forests based Feature Selection

For the analysis of high-dimensional omics datasets, machine learning methods are promising to perform classification, i.e. predicting qualitative traits as well as regression, i.e. predicting quantitative traits. In particular, Random Forests (RF) approaches have been proved promising for prediction based on high-dimensional omics datasets [162]. RF methods are based on decision trees and provide variable importance measures to rank predictors according to their predictive power [163]. In this context, prediction models can be built for the selection of a set of variables with good prediction performance. Moreover, selection of all relevant variables can be performed for the identification of variables involved in active networks and pathways underlying important traits [162]. In the field of genomics, these feature selection methods have gained importance for the reduction of complexity of genomic data to make it easier to analyze and translate large datasets into useful information [164]. The feature selection methods can be broadly classified into three types which are filters, wrappers, and embedded methods [165].

- Filter methods use independent techniques to evaluate the relevance of features and only consider the intrinsic properties of the data. The pre-processing of the data is independent of a subsequent learning algorithm and relevant features are chosen by setting an evaluation criterion or a score [164, 165, 166].





**Figure 3.1.:** Cubic smoothing spline. A series of piecewise-cubic polynomials with increasing order of continuity (taken from [150]).

- Wrapper methods first select a feature subset and then evaluate the subset based on the accuracy of the predictive model based on those features. This way the features are selected iteratively [167].
- Embedded methods combine the strengths of both filter and wrapper methods. For the analysis of big datasets, they make these methods as fast as filter methods and as efficient as wrapper methods [164].

There are three more types of techniques that are variants of the above mentioned three types, namely hybrid [168], ensemble [169], and integrative methods [170]. For more details of these methods, readers are referred to [164]. The implementation of filter methods is easier. However, wrapper methods are generally preferred for their better performance [164]. The Boruta algorithm that has been used in this thesis belongs to the wrapper class [171]. This algorithm has been extensively used to analyze a variety of datasets (>100 studies), including different omics datasets [172, 173].



## 4. Material and Methods

In this chapter, I present the analysis frameworks that I developed in this thesis for deciphering the genetic background of quantitative traits. First, I describe the two chicken datasets used in the thesis. Then the methodology of commonly implemented single-SNP regression based GWAS analysis is provided. Afterward, I present a framework based on a Random Forests feature selection based association analysis followed by the application of systems biology approaches to highlight the regulatory mechanism underlying quantitative traits. In the last section of this chapter, my second framework is described that combines the Random Forests based feature selection and a signal detection approach for the robust detection of genotype-phenotype associations. This chapter is mainly based on my recently published papers [35, 174] (see Appendix A.1 and Appendix A.2).

### 4.1. Datasets

In this thesis, I have analyzed two chicken datasets with the aim to detect genotype-phenotype associations underlying economically important egg quality traits, namely eggshell strength (ESS) and egg weight (EW). These two datasets and the quality control measures taken to ensure the quality of the genotypes are described in this section.

#### 4.1.1. Chicken Dataset 1

To explore the genomic background of the changes that incur to the eggshell strength during the life of laying birds, I analyzed a genotype dataset that has previously been used to investigate the accuracy of imputation as well as the prediction of genomic breeding values in chicken [175, 176, 177]. The dataset consists of a purebred commercial brown layer line with 892 animals and 580,000 SNPs generated using the Affymetrix Axiom<sup>®</sup> Chicken Genotyping Array. The genotypic data do not contain mitochondrial SNPs. The corresponding phenotypic data consist of eggshell strength (ESS) measured as the force in Newton that was required to break the eggshell for each bird at two distinct stages of its production cycle. These two stages were then regarded as Time Point 1 and Time Point 2, respectively. The first time point for ESS was recorded at the ages of 42, 45, and 48 weeks and the second time point was recorded at the ages of 64 and 68 weeks. Averages of the recorded breaking strengths at Time Point 1 (ESS1) and Time Point 2 (ESS2) were used as phenotypes in the further analysis. Extensive pedigree data consisting of, in total, 40,545 individuals from six generations, were available for these birds which were included in an

animal model for breeding value estimation of the birds. A more detailed description of the applied animal model is given in [175]. These breeding values were then de-regressed following Garrick et al. [178] to obtain the pseudo-phenotypes that were used in the further analysis. To ensure genotype quality, I filtered the genotyped data and removed the SNPs:

- that were unassigned to any chromosome or present on the sex chromosomes,
- had a minor allele frequency  $< 0.01$ ,
- had a genotyping call rate  $\leq 97\%$ ,
- significantly deviating from Hardy-Weinberg equilibrium ( $p$ -value  $< 1 \times 10^{-6}$ ),
- for animals having a SNP call rate smaller than 95%.

Finally, after filtering, I used 892 animals and 318,513 SNPs for my analyses.

#### 4.1.2. Chicken Dataset 2

The second dataset pertains to egg weight recorded in 36 weeks old adult birds. The dataset has been previously analysed to perform GWAS of age dependent egg weights (EW) in chicken [60]. The dataset provides genotypes and phenotypes of 1063 birds belonging to a pure bred line of Rhode Island Red chicken, also genotyped with the Affymetrix Axiom<sup>®</sup> 600 K Chicken Genotyping Array. From the seven age levels analysed in the original study, I re-analysed only EW at 36 weeks of age as the most significant associations were reported for this trait. The genotypic data were filtered for SNP call rates, minor allele frequencies and Hardy Weinberg equilibrium using the same threshold values as those for the first dataset. After filtering, I used 294,705 SNPs and 1036 birds in my analysis.

## 4.2. Single-SNP Regression based Association Analysis

Following the study of Liu et al. [60], I perform a GWAS to obtain the association between single-SNPs and the phenotypes. For this analysis, I first applied a principal component analysis (PCA) using the independent SNPs obtained after pruning SNPs using the indep-pair-wise option of the PLINK [179] software, with a window size of 25 SNPs, a step width of 5 SNPs and a  $r^2$  threshold of 0.2. Then I used the top five of those principal components as covariates in the association model to control for population structure. Next, I performed a GWAS analysis based on the following univariate linear mixed model implemented in the FaST-Lmm v0.2.31 software [145].

$$y = W\alpha + x\beta + u + \varepsilon \quad (4.2.1)$$

In Equation (4.2.1),

- $y$  is the vector of phenotypic values for all individuals,

- $W$  is the matrix of covariates,
- $\alpha$  is a vector of corresponding effects and the intercept,
- $x$  is the vector of genotypes for the SNPs tested,
- $\beta$  is the effect size of the marker,
- $u$  is a vector of random polygenic effects with a covariance structure as  $u \sim N(0, \mathbf{K}V_g)$ , where  $\mathbf{K}$  represents the genetic relatedness matrix derived from the SNP markers and  $V_g$  is the polygenic additive variance.
- $\varepsilon$  is the vector of random residuals with  $\varepsilon \sim N(0, \mathbf{I}V_e)$ , where  $\mathbf{I}$  is the identity matrix and  $V_e$  is the residual variance component.

To test the value of  $\beta$  for each SNP against the null hypothesis  $H_0 : \beta = 0$ , the Wald-test ( $F_{\text{Wald}} = \hat{\beta}^2 / \text{Var}(\hat{\beta})$ ) was applied. As suggested in [60], the adjusted threshold value was determined using the *simpleM* approach [148] to evaluate the significance of individual SNPs.

### 4.3. Analysis Frameworks

In this section, I present the methodology of my analysis frameworks developed in this thesis. In the design of these frameworks, the goal is to integrate different strategies in an efficient way that can help underpin the genetics underlying the quantitative traits. The main feature of this analysis framework is the application of Random Forests based feature selection and cubic smoothing splines in the context of association analysis. A detailed overview of these methods, their combined use, and their integration with some bioinformatics approaches are presented below. Both frameworks have recently been published in [35] and [174] (Appendix A.1 and Appendix A.2).

#### 4.3.1. Framework 1: Identification of Regulatory Mechanisms Governing Quantitative Traits using Random Forests

This framework utilizes a Random Forests based association analysis followed by a well established systems biology approach for the identification of regulatory mechanisms underlying quantitative traits. Here, the aim is to highlight the age-specific and common regulatory mechanisms governing the eggshell strength in chicken. For this purpose our analysis follows the structure depicted in Figure 4.1. The pre-processing of the genotypic and phenotypic data used in this framework has been explained in the Section 4.1.1 and the application of the Random Forests and other bioinformatics methods is described in the Section below.

#### 4.3.1.1. Association Analysis Using Random Forests

To identify SNPs potentially associated with eggshell strength, I used the concept of the Random Forests (RF) algorithm to estimate the relative importance of each SNP (attribute) regarding its involvement in the prediction of response variables (de-regressed breeding values). For this purpose, I employed the Boruta algorithm in our study [180], which is a specially developed powerful wrapper for the RF based feature selection approach. The main principle of the Boruta algorithm is based on the extension of the attributes by adding random attributes to the dataset which are called *shadow attributes* and created by shuffling the original values of each attribute (in our case SNPs) in the dataset (see Figure 4.2). This enlargement of the attributes means induction of randomness to the dataset, which leads to the reduction of the bias of hidden (false) signals arising from random fluctuations or correlations in the dataset [171, 180, 181]. To this end, a RF classifier is applied to the extended dataset, and SNPs are systematically and iteratively removed whose importance is significantly smaller than those of the *shadow attributes*. By repeating the process of *shadow attributes* generation and RF algorithm application, importance is assigned to all SNPs. As a result, the Boruta algorithm provides a ranked list of SNPs with a decision of whether the importance of a SNP is confirmed, rejected, or tentative. A similar idea to the Boruta algorithm was manually implemented in [182] to assess the importance of SNPs.

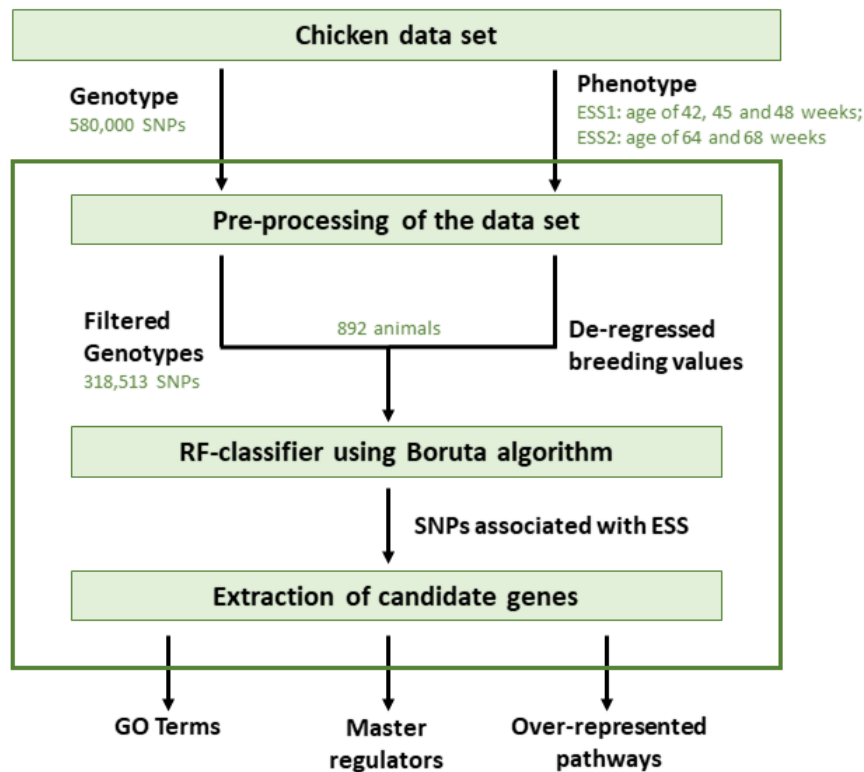
#### 4.3.1.2. Gene Set Analysis

I extracted the genes corresponding to the SNPs identified by the Boruta algorithm from Ensembl using BioMart [88]. Furthermore, I performed a gene set analysis regarding their molecular functions to obtain functional annotations of these genes.

#### 4.3.1.3. Identification of Master Regulators and Over-Represented Pathways

Following our previous studies [183, 184], I performed the upstream analysis (more details in Section 2.3.3.5) and pathway analysis using the geneXplain platform [185] to gain more insight into the functional relationships of genes. The algorithm of "upstream analysis" workflow was introduced by Koschmann et al. [186] and its main goal is to reveal the underlying key regulators that control the activity of target genes. In an "upstream analysis" first molecular pathway networks are constructed and then detects convergence points of these networks are identified. These convergence points are called master regulators and are likely to orchestrate the transcriptional regulation of several genes. In our analysis, I used the GeneWays database [187] and ran the standard "upstream analysis" workflow with a maximum radius of 10 steps upstream to identify the top five master regulators of each gene set that resulted from the previous step of the analysis.

To discover novel biological functions and to reveal the properties of the genes under study, I performed a pathway enrichment analysis as the second step of our analysis.



**Figure 4.1.:** Flowchart of the first analysis framework applied in this thesis (ESS, Eggshell strength).

To this end, I used the TRANSPATH pathway database [108], which is a regularly updated signaling pathway database and contains information about genes, molecules and reactions for the identification of age-specific and common over-represented pathways.

#### 4.3.2. Framework 2: Combining Random Forests and a Signal Detection Method for the Detection of Robust Genotype-Phenotype Associations

The second analysis framework proposed in the thesis consists of six phases to detect important SNPs associated with phenotypes under study. For the application of this framework I used both of chicken datasets given in the Sections 4.1.1 and 4.1.2.

**Phase 1:** The first phase of this framework consists of the single-SNP regression based association analysis as described in Section 4.2. The Wald-test statistics that represents the strength of association between the individual SNPs and the phenotype are recorded.

A Raw genotypic dataset						B Recoded genotypic dataset					
Sample	SNP1	SNP2	SNP3	SNP4	SNP5	Sample	SNP1	SNP2	SNP3	SNP4	SNP5
4012638	BA	AB	AA	AA	BB	4012638	1	1	2	0	2
4012591	BA	AB	BB	AA	AA	4012591	1	1	0	0	0
4010988	BA	AB	BB	AA	AB	4010988	1	1	0	0	1
4012757	AA	BB	BB	BB	AA	4012757	0	0	0	2	0
4012608	BA	AB	BB	AA	AA	4012608	1	1	0	0	0
4014050	AA	BB	AB	AA	AA	4014050	0	0	1	0	0
4012597	AA	BB	BB	AB	AB	4012597	0	0	0	1	1
4018124	AA	BB	BB	AA	AA	4018124	0	0	0	0	0
4012339	BB	AA	BB	AA	AA	4012339	2	2	0	0	0
4015259	AA	BB	BB	AA	AA	4015259	0	0	0	0	0

C Extended genotypic dataset										
Sample	SNP1	SNP2	SNP3	SNP4	SNP5	SNP1S	SNP2S	SNP3S	SNP4S	SNP5S
4012638	1	1	2	0	2	1	1	0	1	1
4012591	1	1	0	0	0	0	2	0	0	0
4010988	1	1	0	0	1	0	0	0	0	1
4012757	0	0	0	2	0	2	1	2	0	0
4012608	1	1	0	0	0	0	1	0	0	0
4014050	0	0	1	0	0	1	0	1	0	2
4012597	0	0	0	1	1	0	1	0	2	0
4018124	0	0	0	0	0	1	0	0	0	0
4012339	2	2	0	0	0	1	0	0	0	0
4015259	0	0	0	0	0	1	0	0	0	0

**Figure 4.2.:** Extension of genotypic data as implemented in Boruta. (A) A sample of raw genotypic data. (B) Recoding of the raw genotypic data as additive code. (C) Extension of the genotypic data by permutating the individual SNP genotypes among the samples and then adding them to the raw dataset.

In Figure 4.3 I exemplarily show a chromosomal region and its corresponding Wald-test statistic values.

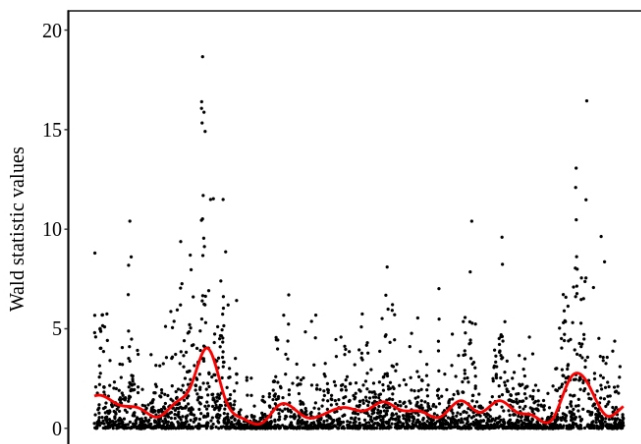
**Phase 2:** For the elaboration of association signals embedded in the Wald-test statistics, I apply a cubic smoothing spline to these values. The cubic smoothing spline is a piece-wise defined cubic function and is based on the same principle as the normal cubic regression. The assumption implicit in this approach is that the individual association values are observed with noise and that these values can be considered as estimations of some underlying function  $g$ . Given the marker positions in the genome ( $x_i$ ) and the corresponding association values ( $y_i$ ), the function  $g$  is estimated by minimizing the following expression

$$S(g) = \sum \{y_i - g(x_i)\}^2 + \lambda \int g''(x)^2 dx. \quad (4.3.1)$$

In Equation 4.3.1, the first part of the right hand side represents the residual sum of squares with the cubic spline function  $g(x_i)$  being the estimated value of the function  $g$  corresponding to SNP  $i$  at chromosomal position  $x_i$ . The integral represents a roughness penalty con-



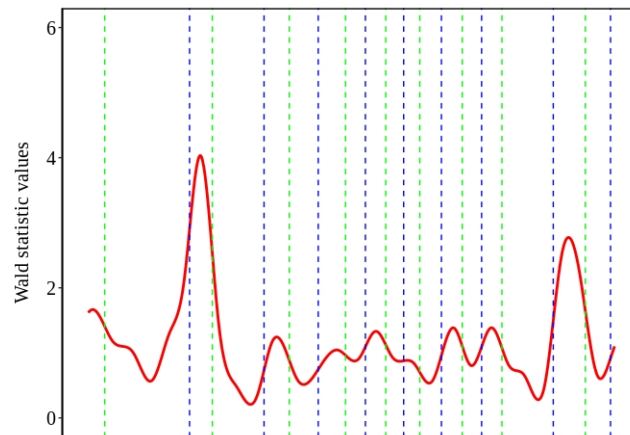
trolled by the tunable parameter  $\lambda$  whose value is determined by cross validation. The penalty controls the trade off between the conflicting goals of matching the given data and producing a smooth curve [188].  $g''$  represents the second derivative of the sought function with respect to  $x$ . The assumption of  $g$  being continuous and twice differentiable leads to its approximability via a cubic smoothing spline [189, 190]. Thus, a continuous and smooth curve, suitable for the elaboration of the association signals in the test statistic values is obtained. A similar technique has been used in [190] to define window boundaries for the general analysis of genomic data. In Figure 4.3 I exemplarily show the application of cubic smoothing spline over the Wald-test statistics in a small chromosomal region.



**Figure 4.3.:** Distribution of Wald-test statistics with the fitted cubic spline. The black dots represent the distribution of the Wald-test statistics along the length of a chromosome segment while the red line indicates the cubic spline fitted to the statistic values.

**Phase 3:** For delineation of the obtained association signals in the form of peaks, I determined the inflection points based on the smoothed values. As the smoothing curve represents a function  $g(x)$ , the inflection points indicate the positions, where  $g''(x) = 0$  and thus the curve changes its curvature. Hence, the region between two consecutive inflection points having a downward concave form is regarded as a peak. The maximum value within a peak is recorded as the height of the peak. This height of the peak is taken as a measure of association strength. In Figure 4.4, I exemplarily show the identified peak regions based on the inflection points. The schematic R-code for detecting the association peaks is given in the Listing 4.1 and the complete R-script is provided in the File A.3.

**Phase 4:** In order to separate the peak regions having association signals higher than those that would have arisen by chance, I have created a null distribution by permutating the phenotypic data. For the construction of the null distribution, Phases 1, 2 and 3 have been



**Figure 4.4.:** The inflection points of cubic spline curve. The red line depicts the same cubic spline curve as presented in Figure 4.3 and the dashed lines represent the inflection points of the curve. A pair of a left (blue) and a right (green) inflection points constitutes a peak.

applied to each permuted dataset and the maximum peak values were recorded. In our analysis, I permuted the dataset 1000 times. In the real dataset, I defined a peak region as a QTL if the corresponding peak height exceeds the 95th percentile of the null distribution.

**Phase 5:** Adopting the strategy explained in the Section 4.3.1.1 the Random Forests (RF) algorithm was used to estimate the relative importance of each SNP (attribute) for the prediction of the response variable (phenotype). For this purpose, I applied the Boruta algorithm [180] which is a powerful wrapper for the RF based feature selection approach to assess the importance of SNPs. Consequently, I obtained a decision for each SNP whether the importance of the SNP is confirmed, rejected or tentative. In our analysis I only considered SNPs with confirmed importance.

**Phase 6:** Finally, to prioritize the SNPs which are in the QTLs detected in Phase 4, I use the important SNPs from Phase 5 and define the SNPs discovered in both Phases as robust SNPs in our analysis.

```
#Step 1: Import the SNP information and the corresponding test
statistics values

#Step 2: Consider each chromosome separately to compute the cubic
spline (CP) using the test statistics values of the ordered SNP
positions
> CP = smooth.spline(SNP_Position, Wald_Statistic)

#Step 3: Compute the first and second derivatives (FD and SD) of
the smoothed test statistics values
> FD = diff(CP$Wald_Statistic) / diff(CP$SNP_Position)
> SD = diff(FD) / diff(CP$SNP_Position)

#Step 4: Identify the inflection points (IP) of the smoothed values
> IP = (SD[-1] > 0) * (SD[-nrow(SD)] < 0) |
      (SD[-1] < 0) * (SD[-nrow(SD)] > 0)

#Step 5: Identify the boundaries for all possible peaks using the
inflection points while separating the peaks from the
neighbouring fluctuations
for(j in 2 : (n - 1))
{
  if( IP[j] == TRUE &&
      ( CP$Wald_Statistic[j - 1] < CP$Wald_Statistic[j + 1]))
  {
    LeftBorder[j] <- TRUE
  }
  else if( IP[j] == TRUE &&
           ( IP$Wald_Statistic[j - 1] > IP$Wald_Statistic[j + 1]))
  {
    RightBorder[j] <- TRUE
  }
}

#Step 6: Record the peaks and their corresponding maximum Wald
statistic value as the height of the peak
```

**Listing 4.1:** The R code to perform the signal detection strategy.

#### 4.4. Extraction of Candidate Genes

I scan the genome to identify the genes corresponding to the robust SNPs using BioMart [88]. Only those genes were considered to have some association with the phenotype that were harboring at least one robust SNP within its boundaries. A schematic R-code for this analysis is given in Listing 4.2 and the complete R-script is provided in File A.4.

```
#Step 1: Load the biomart package
> library(biomart)

#Step 2: Import the SNP information

#Step 3: Connect to the Ensembl database and specify the name of
the target species
> useMart("ensembl", dataset = "ggallus_gene_ensembl")

#Step 4: Describe the gene attributes of interest
> attribute_list = c("ensembl_gene_id", "chromosome_name",
                    "strand", "start_position",
                    "end_position", "gene_biotype",
                    "external_gene_name")

#Step 5: Extract the predefined attributes of genes from ensembl
database
> getBM(attributes=attribute_list, mart = ensembl)

#Step 6: Identify the genes related to the important SNPs
```

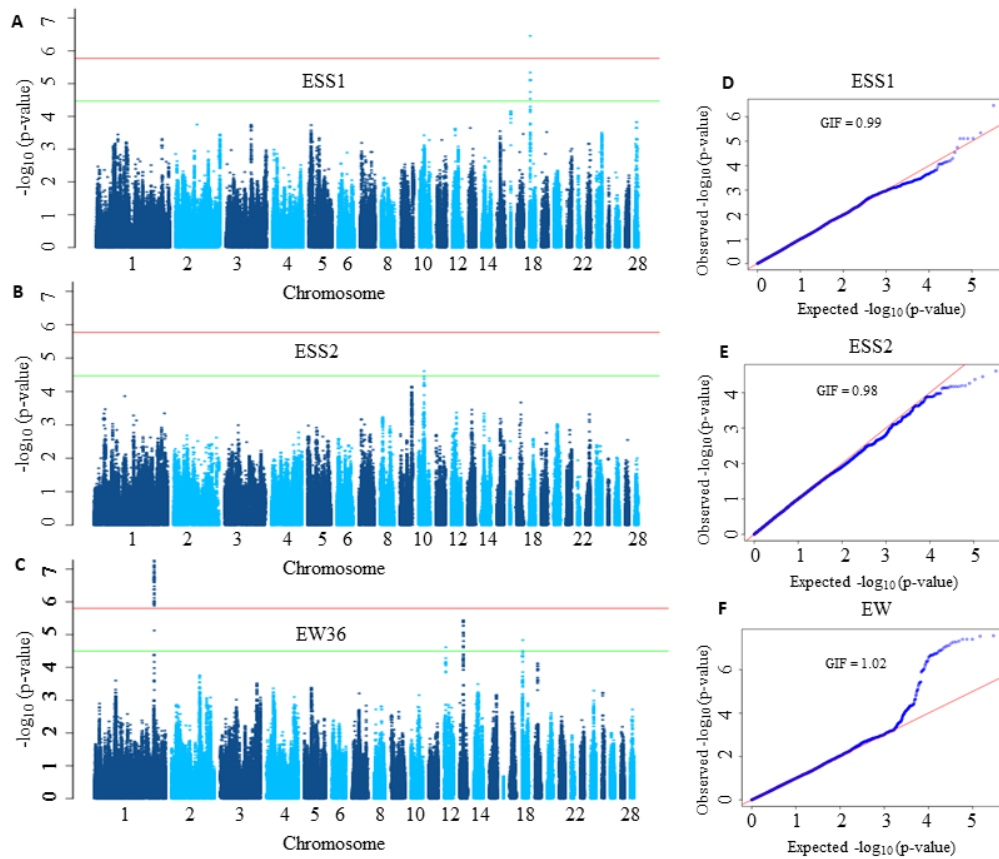
**Listing 4.2:** The R code to extract the list of genes corresponding to the important SNPs

## 5. Results

In this chapter I present the results of the application of both of the frameworks developed in this thesis. But first, to benchmark the performance of the suggested frameworks against a conventional GWAS approach, the results of single-SNP regression based GWAS are exhibited. In the second part, I focus on the application of a Random Forests based feature selection strategy to detect the genotype-phenotype associations followed by the identification of age specific and common key regulatory mechanisms governing eggshell strength in chicken. Then, I present the results of combining Random Forests and a signal detection strategy in order to detect robust associations. Both frameworks were employed in order to get new insights regarding the genetic mechanisms of biological systems underlying chicken egg quality traits. This section is mainly based on my recently published papers [35, 174] (see Appendix A.1 and Appendix A.2).

### 5.1. Single-SNP Regression based GWAS

The application of linear mixed models (LMMs) is a standard practice today to perform single-SNP regression based GWAS. I also analysed both of the datasets using a LMM as suggested for the analysis of egg weight (EW) in the study of Liu et al. [60]. In the application of LMM, I considered the correction of the population stratification and applied the *SimpleM* method [148] for multiple testing correction. The LMM model used for this analysis was successful in controlling the inflationary effects of the population and the family structure on the obtained association statistics as represented by close to one genomic inflation factor value (see Figure 5.1 D, E, F). The LMM approach for eggshell strength (ESS) at time point 1 (ESS1) and time point 2 (ESS2) led to the identification of only one significant SNP for ESS1 (see Figure 5.1A,B). Furthermore, the LMM method revealed 43 significant SNPs for EW (see Figure 5.1 C) on chromosome 1 (GGA1) which were then mapped to three genes (ITM2B, RCBTB2, RB1). Today, it is a well known fact that quantitative traits are influenced by a large number of genes mostly having small effects. But as shown in Figure 5.1, many association signals were not strong enough to reach the significance threshold, thereby their influences on the phenotype can not be considered in any post-GWAS analysis. This demonstrates the limited power of conventional single-SNP based GWAS for the analysis of quantitative traits.



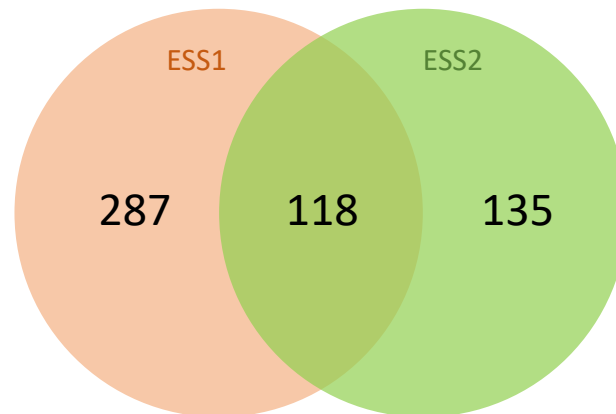
**Figure 5.1.:** Manhattan and Q-Q plots corresponding to eggshell strength at time point 1 (ESS1), time point 2 (ESS2) and egg weight at 36 weeks of age (EW36). In Manhattan plots (A-C), the horizontal red and green lines denote the genome-wide significance ( $p\text{-value} = 1.7 \times 10^{-6}$  for ESS1 and ESS2 and  $1.5 \times 10^{-6}$  for EW36) and suggestive significance thresholds ( $p\text{-value} = 3.4 \times 10^{-5}$  for ESS1 and ESS2  $3.1 \times 10^{-5}$  for EW), respectively. The  $-\log_{10}$  of the observed  $p$ -values for each single nucleotide polymorphism (SNP) is given on the y-axis while its position on a chromosome is given on the x-axis. In Q-Q plots (D-F) the observed  $-\log_{10}$  transformed  $p$ -values are plotted against the expected  $-\log_{10}$  transformed  $p$ -values. GIF stands for genomic inflation factor.

## 5.2. Analysis Framework 1

### 5.2.1. Association Analysis Using Random Forests

In this framework, I performed the RF approach using the Boruta algorithm to identify the informative SNPs associated with eggshell strength at two time points during the laying cycle of commercial brown layer chicken (Dataset 1 given in Section 4.1.1). For this purpose, the importance of each SNP was assessed separately for its association with the phenotype of interest. To this end, I obtained a list of SNPs for each time point whose importance was confirmed by the Boruta algorithm for the prediction of the phenotype. Analyzing both time points, I identified 3726 SNPs associated with eggshell strength at Time Point 1 (ESS1) and 1815 SNPs associated with eggshell strength at Time Point 2 (ESS2). These SNPs were then mapped to the genome and the genes harboring at least one of these SNPs were identified for both traits. In total, I identified 405 genes for ESS1 and 253 genes for ESS2. A closer look at these gene lists reveals that 22 % (118 genes) of them are overlapping (see Figure 5.2), which depicts the conservation of some of the underlying mechanisms involved in the synthesis of eggshell during different stages of the egg production cycle. These results also show that a considerably high number of genes that were distinct for the time points highlight the dynamic nature of this trait.

This section comprises of three parts. First, to gain a deeper insight into these gene sets, I performed a gene set analysis and clustered their functions based on the GO terms. Second, I performed the “upstream analysis” introduced by Koschmann et al. [186] for the identification of specific and common master regulators of both time points. Third, I present the over-represented pathways to further elucidate the mechanisms that control the ESS at different production stages of the chicken.



**Figure 5.2.:** Venn diagram depicting the number of genes associated with eggshell strength at Time Point 1 (ESS1), at Time Point 2 (ESS2), and their overlap.







**Table 5.1.:** Top 15 Gene Ontology (GO) molecular function terms based on the adjusted *p*-value for the eggshell strength at Time Point 1 (ESS1).

GO Term	GO Title	Number of Genes	Adjusted <i>p</i> -Value
GO:0005515	protein binding	281	$5.11 \times 10^{-8}$
GO:0005488	binding	331	$1.97 \times 10^{-7}$
GO:0043167	ion binding	155	$4.93 \times 10^{-3}$
GO:0000146	microfilament motor activity	5	$4.93 \times 10^{-3}$
GO:0003779	actin binding	20	$6.90 \times 10^{-3}$
GO:0032559	adenyl ribonucleotide binding	49	$1.47 \times 10^{-2}$
GO:0030554	adenyl nucleotide binding	49	$1.51 \times 10^{-2}$
GO:0044877	macromolecular complex binding	50	$1.54 \times 10^{-2}$
GO:0004683	calmodulin-dependent protein kinase activity	5	$1.54 \times 10^{-2}$
GO:0005524	ATP binding	47	$2.05 \times 10^{-2}$
GO:0042623	ATPase activity, coupled	16	$2.24 \times 10^{-2}$
GO:0008092	cytoskeletal protein binding	30	$3.32 \times 10^{-2}$
GO:0043168	anion binding	74	$3.93 \times 10^{-2}$
GO:0046983	protein dimerization activity	40	$4.15 \times 10^{-2}$
GO:0017016	Ras GTPase binding	12	$4.86 \times 10^{-2}$

### 5.2.3. Identification of Master Regulators

Applying the upstream analysis integrated in the geneXplain platform [185], I identified the top five age-specific and common master regulators for both traits. While the master regulators *Arx*, *Sox1*, and *Scn11a* were specific for ESS1, the master regulators *St8sia2*, *Tead2*, and *Prox1* were identified for ESS2. Additionally, *Slc22a1* and *Sox11* were identified for both time points (see Figures 5.5 and 5.6).

The ESS1 specific master regulator *Scn11a* is a gene encoding transmembrane sodium channels which control the voltage-gated sodium transport especially in the uterus [195, 196], the site of eggshell synthesis in birds. Moreover, the importance of sodium channels in the transportation of inorganic minerals deposited in the eggshell is well established [197]. Interestingly, I found the master regulator *Slc22a1* at both time points. It codes for the protein OCT1, an organic cation transporter for substrates such as putrescine [198], which plays an

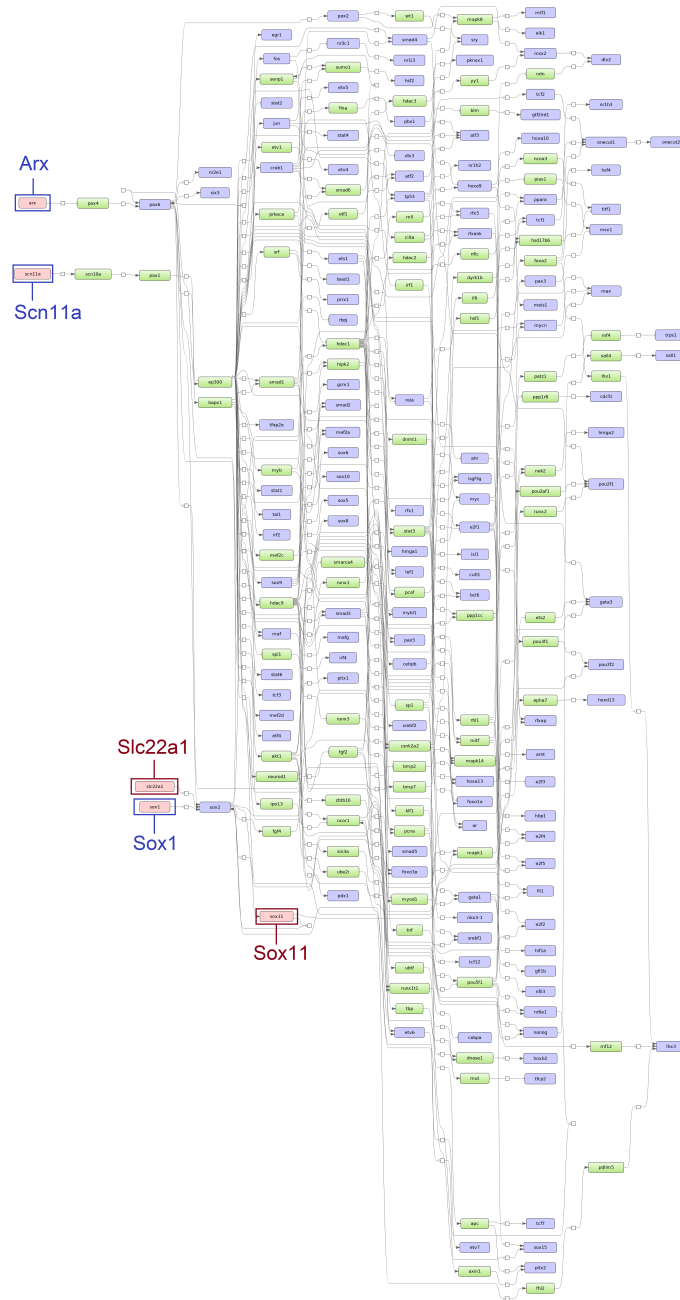
**Table 5.2.:** Top 15 Gene Ontology (GO) molecular function terms based on the adjusted  $p$ -value for the eggshell strength at Time Point 2 (ESS2).

GO Term	GO Title	Number of Genes	Adjusted $p$ -Value
GO:0005515	protein binding	168	$1.30 \times 10^{-2}$
GO:0022843	voltage-gated cation channel activity	9	$2.09 \times 10^{-2}$
GO:0005242	inward rectifier potassium channel activity	4	$2.10 \times 10^{-2}$
GO:0032549	ribonucleoside binding	40	$2.79 \times 10^{-2}$
GO:0000166	nucleotide binding	48	$2.79 \times 10^{-2}$
GO:0005524	ATP binding	34	$2.79 \times 10^{-2}$
GO:0001883	purine nucleoside binding	39	$3.66 \times 10^{-2}$
GO:0032559	adenyl ribonucleotide binding	34	$3.66 \times 10^{-2}$
GO:0005488	binding	199	$3.66 \times 10^{-2}$
GO:0030554	adenyl nucleotide binding	34	$3.66 \times 10^{-2}$
GO:0051427	hormone receptor binding	9	$3.66 \times 10^{-2}$
GO:0015276	ligand-gated ion channel activity	8	$3.66 \times 10^{-2}$
GO:0017076	purine nucleotide binding	39	$3.7 \times 10^{-2}$
GO:0022836	gated channel activity	12	$3.83 \times 10^{-2}$
GO:0036094	small molecule binding	50	$4.64 \times 10^{-2}$

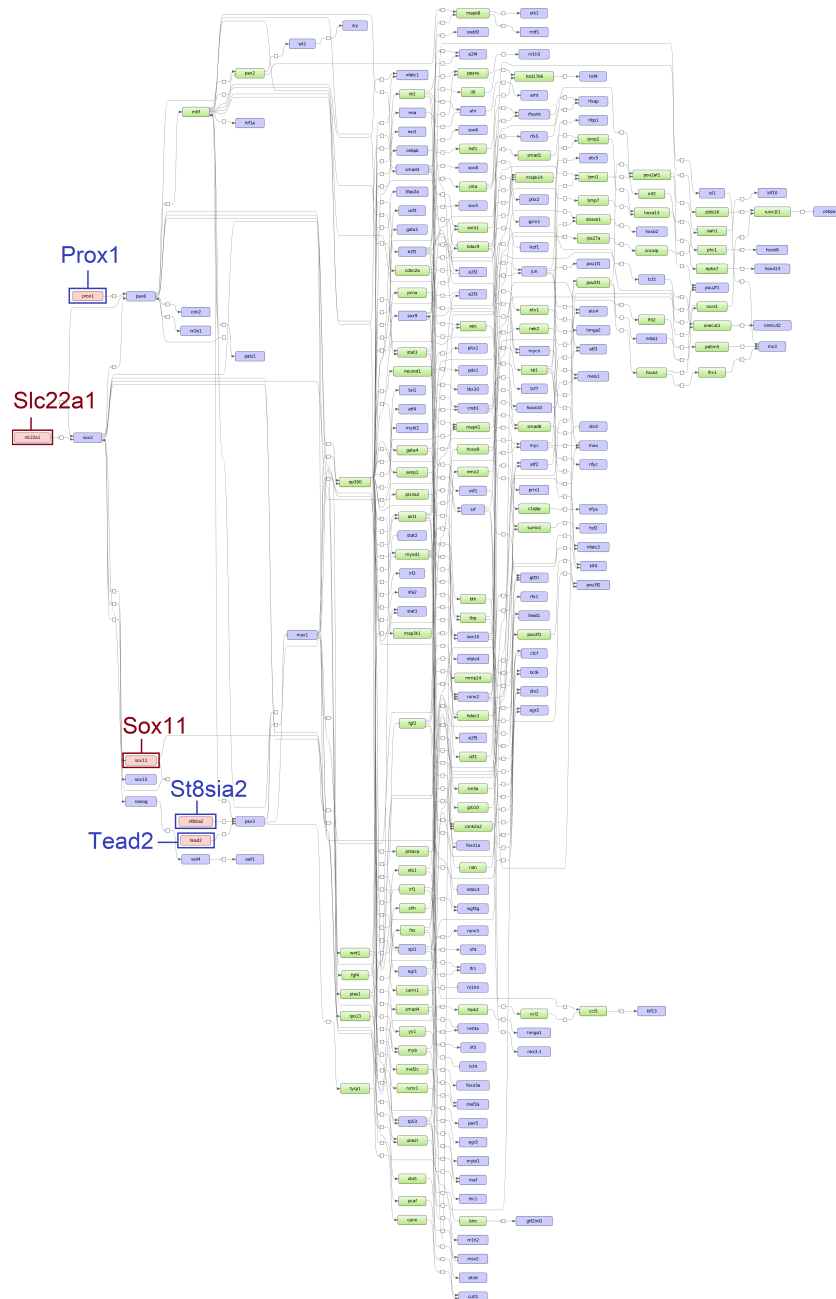
important role for eggshell thickness [199] and calcium transport in the intestine [200].

Furthermore, many other members of the super-family of transport proteins, *Slc* (solute carrier proteins), are well known to play an essential role in the homeostasis of calcium ions in a variety of tissues [201]. The *Slc* proteins have also been reported to transport magnesium ions during the egg calcification process [52]. Another interesting master regulator, *Sox11*, which encodes a member of the Sox (SRY-related HMG-box) family of transcription factors, was found at both time points. *Sox11* is known to positively regulate the process of osteogenesis (the formation of bone) [202]. This regulator gains relevance given the importance of bone as a reservoir of minerals, especially calcium [51]. In birds, the calcium homeostasis is achieved by regulating the metabolism of bone minerals as well as by controlling the absorption and excretion of calcium in the intestine and in kidneys, respectively [203]. Furthermore, the master regulator *Tead2* found for ESS2 is a regulator of osteogenesis [204] and it is also one of the direct downstream target genes of *Sox11*. This might be an indication of different regulatory mechanisms involved in the osteogenesis or bone remodeling during the later stages of the laying cycle [202].

*St8sia2*, identified as an ESS2 specific master regulator, encodes a membrane protein which catalyzes the metabolism of sialic acid [205], a carbohydrate found in the eggshell membranes [206, 207, 208]. The eggshell membranes constitute the inner layer of the eggshell and contribute to its strength. They further provide the nucleation sites for the initiation of the shell synthesis [209]. Sialic acid is also part of podocalyxin and secreted phosphoprotein 1 (SPP1), both of which are glycoproteins found in the uterus during eggshell calcification [210, 52]. Because of its high negative charge, podocalyxin is presumed to interact with calcium carbonate during the calcification of the eggshell [210]. The master regulator *Prox1* encodes the protein prospero homeobox 1 that has also been reported as part of eggshell membranes [211, 212]. However, the *Prox1* gene is mostly implicated in the regulation of the development of a variety of organs including liver, pancreas, and kidney [213]. Although the vast majority of the master regulators could be biologically characterized to be crucial for ESS, the importance and role of the two master regulators *Sox1* and *Arx* for this trait is currently biologically unconfirmed and could hence provide novel targets for future studies.



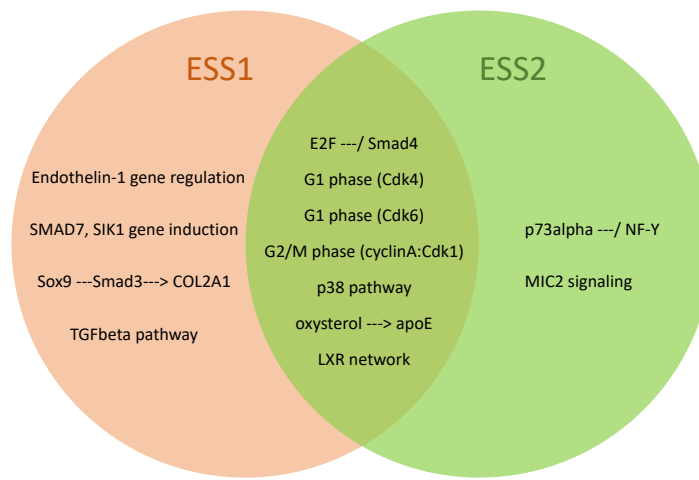
**Figure 5.5.:** Scheme of gene regulatory pathways revealing the top five master regulators (pink filled boxes) for eggshell strength at Time Point 1 (ESS1) following the upstream analysis [186]. The master regulators written in dark blue and surrounded by dark blue boxes (*Arx*, *Scn11a* and *Sox1*) were identified specifically for ESS1 while master regulators written in dark red and surrounded by dark red boxes (*Slc22a1* and *Sox11*) were identified at both time points (corresponding networks for eggshell strength at Time Point 2 (ESS2) in Figure 5.6).



**Figure 5.6.:** Scheme of gene regulatory pathways revealing the top five master regulators (pink filled boxes) for eggshell strength at Time Point 2 (ESS2) following the upstream analysis [186]. The master regulators written in dark blue and surrounded by dark blue boxes (*Prox1*, *St8sia2* and *Tead2*) were identified specifically for ESS2 while master regulators written in dark red and surrounded by dark red boxes (*Slc22a1* and *Sox11*) were identified at both time points (corresponding networks for eggshell strength at Time Point 1 (ESS1) in Figure 5.5).

### 5.2.4. Identification of Over-Represented Pathways

To further elucidate and investigate the mechanisms that control the ESS at different time points, I was interested in identifying age-specific and common over-represented pathways for both time points. Applying the pathway analysis, I identified eleven and nine significantly over-represented pathways for ESS1 and ESS2, respectively, and seven of these pathways were overlapping for both time points (see Figure 5.7 and Table 5.3).



**Figure 5.7.:** Venn diagram of over-represented pathways ( $p$  adjusted  $< 0.001$ ) of eggshell strength at Time Point 1 (ESS1), at Time Point 2 (ESS2), and their overlap. Pathways are based on the TRANSPATH pathway database [108].

Among the pathways shared by both time points, G1 phase (Cdk4), G1 phase (Cdk6), and G2/M phase (cyclinA:Cdk1) involve different members of the cyclin-dependent kinase (CDK) family which regulate transcription, mRNA processing, and, more importantly, cell cycle [214]. In the context of ESS, these pathways may influence the differentiation efficiency of osteoblasts, osteoclasts, chondrocytes [215], and uterine epithelium cells, all of which are crucial for the supply of calcium ions as well as for bone and calcium homeostasis [216, 217]. The p38 pathway is implicated in a variety of cellular responses including those related to proliferation, differentiation and apoptosis [218]. Moreover, the role of this pathway has also been reported in the egg development of *Drosophila melanogaster* [219]. The LXR (liver X receptors) network plays a central role in the transcriptional control of lipid metabolism [220]. This pathway also mediates the concentrations of oxysterols and ApoE (Apolipoprotein E), if activated in response to elevated intra-cellular cholesterol levels [221]. The oxysterols, oxygenated forms of cholesterol, are intermediates in bile acid

and steroid hormone biosynthetic pathways [222]. Among other steroid hormones, estrogen is more intimately involved in calcium homeostasis and has also been implicated in the development of osteoporosis [223]. Moreover, other forms of oxysterols are also involved in calcium metabolisms [224] and mesenchymal stem cell differentiation [225].

**Table 5.3.:** Significantly over-represented pathways for both time points ( $p$  adjusted  $< 0.001$ ) sorted by adjusted  $p$ -values (based on the smaller one of either ESS1 or ESS2). Pathways are based on the TRANSPATH pathway database [108]. (ESS1/ESS2, eggshell strength at Time Point 1/2).

Pathway name	Adjusted $p$ -Value for ESS1 / ESS2	Over-represented in
E2F —/ Smad4	$5.05 \times 10^{-5} / 7.99 \times 10^{-4}$	ESS1, ESS2
Endothelin-1 gene regulation	$5.05 \times 10^{-5} / -$	ESS1
G2/M phase (cyclin A:Cdk1)	$1.61 \times 10^{-4} / 1.65 \times 10^{-4}$	ESS1, ESS2
SMAD7, SIK1 gene induction	$1.61 \times 10^{-4} / -$	ESS1
oxysterol —>apoE	$1.61 \times 10^{-4} / 1.85 \times 10^{-4}$	ESS1, ESS2
LXR network	$1.61 \times 10^{-4} / 1.65 \times 10^{-4}$	ESS1, ESS2
p73alpha —/ NF-Y	$- / 1.65 \times 10^{-4}$	ESS2
Sox9 —Smad3—>COL2A1	$5.43 \times 10^{-4} / -$	ESS1
G1 phase (Cdk6)	$7.60 \times 10^{-4} / 7.93 \times 10^{-4}$	ESS1, ESS2
G1 phase (Cdk4)	$9.77 \times 10^{-4} / 7.99 \times 10^{-4}$	ESS1, ESS2
p38 pathway	$9.77 \times 10^{-4} / 7.99 \times 10^{-4}$	ESS1, ESS2
MIC2 signaling	$- / 7.99 \times 10^{-4}$	ESS2
TGFbeta pathway	$9.53 \times 10^{-4} / -$	ESS1

In addition to the CDKs, the Smad4 proteins, predominantly present in the nucleus of the cell, mediate the cell cycle due to their association with the E2F family of transcription factors [226]. These pathways can be upstream regulated by the transforming growth factor  $\beta$  (TGF- $\beta$ ) [227]. The transforming growth factor- $\beta$  (TGF- $\beta$ ) signaling pathway can be regarded as the most important pathway enriched for ESS1. This pathway, among its other functions, is well-known for its role in bone homeostasis [228]. Furthermore, some components of this pathway also overlap with other pathways delineated in our analysis. The Sox9 is a transcription factor that regulates the expression of the COL2A1 (collagen type II, alpha 1) gene which contributes to collagen formation [229]. During this process, Smad3,



a member of effector molecules in the signaling pathways of the TGF- $\beta$  ligand superfamily is activated [230]. Another pathway that is based on Smad7, SIK1 gene induction also regulates TGF- $\beta$  signaling [231]. Owing to this crosstalk with a variety of other pathways, the TGF- $\beta$  signaling pathway allows the bone to adapt to dynamic environments [228].

The Endothelin-1 gene (ET-1) regulation pathway includes the mechanisms regulating ET-1 gene expression. Among other functions, ET-1 is involved in osteoblast proliferation and differentiation in bone tissue as well as in the ovulation process in the uterus [232]. ET-1 gene regulation is responsive to intracellular calcium and calmodulin [233]. The MIC2 signaling pathway, which was specifically enriched for ESS2, has CD99 as the main cell surface protein and was found to be involved in apoptosis, adhesion, differentiation, and protein trafficking possibly by affecting actin cytoskeleton reorganization [234, 235]. Another ESS2 specific pathway involves the inactivation of the nuclear factor Y (NF-Y) transcription factor by p73 proteins, a process that represses the promoter of the telomerase catalytic subunit and induces replicative senescence [236, 237]. The activity of NF-Y is further linked to the parathyroid hormone, which is the main regulator of calcium and phosphorus homeostasis. Taken together, the pathways show a diversity of complex functional features in chicken in response to age-dependent changes in eggshell formation. Some pathways show a direct relevance for ESS while others seem to be indirectly linked via interactions between pathways and regulators [238, 239].

### 5.3. Analysis Framework 2

The primary objective of this analysis framework is to detect those robust SNPs with weak associations that remain undetected using conventional GWAS approaches. The overall framework comprises the following steps. First, a linear mixed model (LMM) based single-SNP GWAS is performed to obtain test statistics representing the strength of association between each SNP and the phenotype. Second, performing the signal detection strategy by fitting a cubic smoothing spline on the test statistic values, I identify QTLs. Third, I apply the RF classifier using the Boruta algorithm to assess the relative importance of SNPs regarding the level of their association with the phenotype. Finally, the important SNPs are prioritized within those QTLs to discover a robust set of SNPs associated with the phenotype. Two different GWAS (genotype and phenotype) datasets related to eggshell strength (ESS) and egg weight (EW) have been analysed using this framework to demonstrate its functionality.

#### 5.3.1. Detection of Genotype-Phenotype Association Using the Combined Framework

To identify genes showing weak association signals that remain undetected in a conventional GWAS analysis, I applied the analysis framework described in Section 4.3.2 to both of the datasets presented in Sections 4.1.1 and 4.1.2.

##### 5.3.1.1. Association Analysis of Eggshell Strength

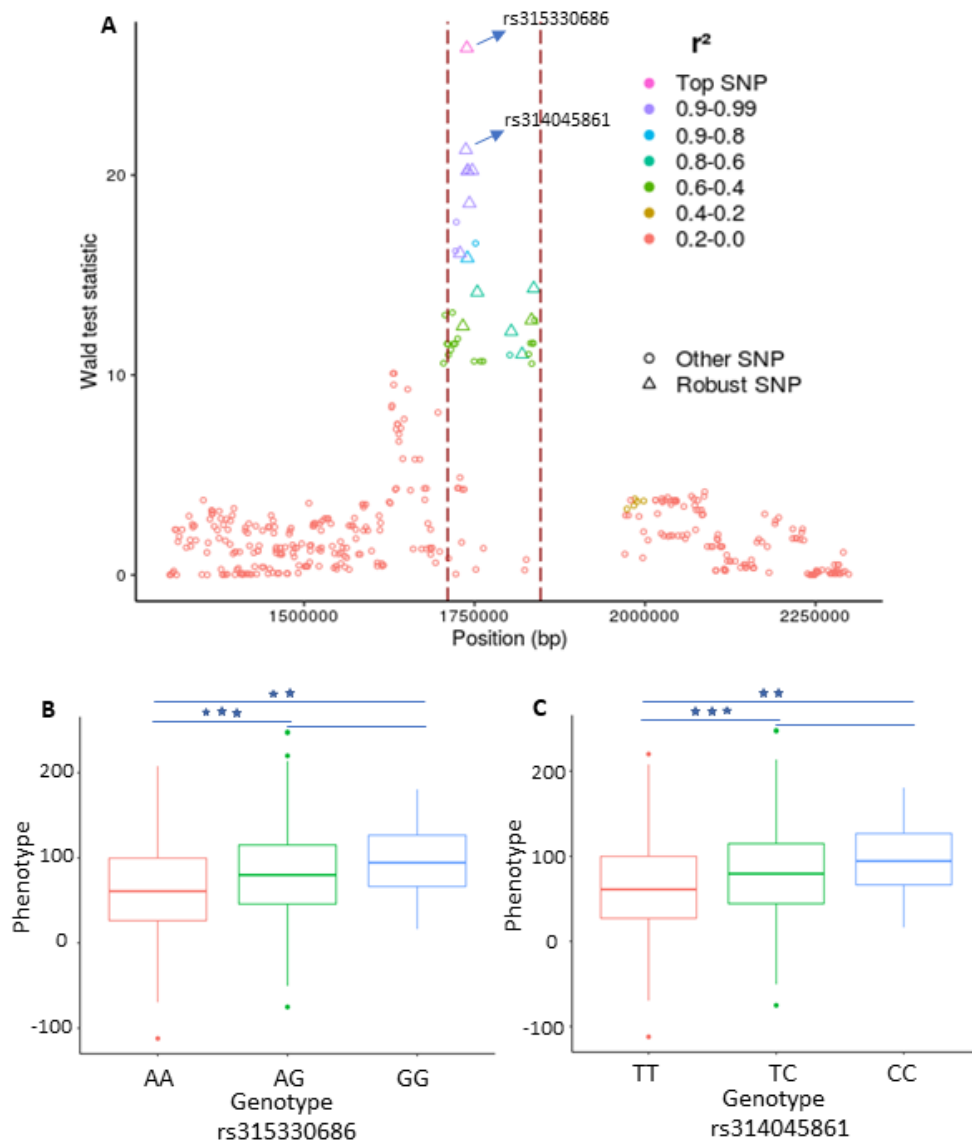
The analysis of the ESS datasets reveals eight QTLs for ESS1 and five QTLs for ESS2 based on the signal detection approach. The details of these QTLs are given in Table 5.4. Interestingly, I found chromosome 9 (GGA9), 10 (GGA10), 15 (GGA15) and 20 (GGA20) to have QTLs associated with ESS at both time points. Especially, the QTLs on GGA20 are overlapping and underpin the same genomic region as associated with ESS at both time points. In addition, the application of the RF classifier provides 3726 and 1815 SNPs which map to 405 and 253 genes associated with ESS1 and ESS2, respectively. The investigation of these SNPs in the identified QTLs reveals 158 and 14 robust SNPs (defined in Section 4.3.2) related to ESS1 and ESS2, respectively.

Of particular interest here is the linkage disequilibrium (LD) analysis that I performed based on the robust SNPs to further elaborate their makeup in the identified QTLs. The LD analysis reveals, as expected, that the robust SNPs inside the QTLs have a remarkably higher level of LD than the surrounding SNPs (see Figure 5.8A). To this end, I exemplarily compared the phenotype differences between the genotypes of the top two SNP (rs315330686, rs314045861) on GGA18. The comparison suggests that for both SNPs, the birds homozygous for the minor alleles have higher phenotypes than those of the other two genotypes (Figure 5.8B,C).

**Table 5.4.:** Significant peaks as defined in Phase 4 of our analysis framework and corresponding quantitative trait loci (QTLs) for ESS1 and ESS2.

Chromosome	No. of SNPs	Start Position	End Position	No. of Genes	Trait
2	204	147,575,318	148,273,465	3	ESS1
9	66	21,762,694	21,953,310	0	ESS1
9	82	21,777,888	22,001,729	0	ESS2
10	75	6,517,673	6,728,897	4	ESS1
10	86	9,922,422	10,054,824	2	ESS1
10	60	10,715,120	10,818,097	3	ESS2
10	61	11,245,585	11,351,799	1	ESS2
12	112	10,948,518	11,227,521	2	ESS1
15	42	4,908,007	5,006,688	7	ESS1
15	43	6,193,090	6,273,778	3	ESS2
18	38	1,722,586	1,836,741	2	ESS1
20	51	7,589,607	7,717,177	1	ESS1
20	46	7,599,368	7,711,505	1	ESS2

The extraction of the genes corresponding to the robust SNPs reveals 14 and 3 genes for ESS1 and ESS2, respectively. The functional investigation of these genes shows that the majority of them were annotated to play essential roles in the transport of minerals and organic compounds. Seven of these genes, namely ATP6V0A2 (ATPase, H<sup>+</sup> Transporting, Lysosomal V0 Subunit A2), DDX55 (DEAD-Box Helicase 55), DNAH10 (Dynein Axonemal Heavy Chain 10), GTF2H3 (General Transcription Factor IIH Subunit 3), MYO1E (Unconventional Myosin 1E), TCTN2 (Tectonic Family Member), and MYH10 (Myosin Heavy Chain 10)), have molecular functions related to the activity of the ATPase enzyme. Interestingly, in relation to eggshell formation ATPases have long been known to show intense activity in the cells of shell gland during the synthesis of eggshell [240]. Furthermore, CHRNA7 (Cholinergic Receptor Nicotinic Alpha 7 Subunit), is associated with the transport of ions, especially calcium ions. The other main function performed by the identified genes includes cell morphogenesis which ensures the homeostasis of tissues involved in the development of eggshell [241, 242]. The genes that play a role in this process include NDEL1 (NudE Neurodevelopment Protein 1 Like 1), ADGRB1 (Adhesion G Protein-Coupled Receptor B1), THSD4 (Thrombospondin Type 1 Domain Containing 4) and EIF2B1 (Eukaryotic Translation Initiation Factor 2B Subunit Alpha).



**Figure 5.8.:** Plot representing a genomic region on chromosome 18 which is in association with eggshell strength at time point 1 (ESS1). (A) Plot representing the linkage disequilibrium (LD) structure inside and around a significant peak. The dotted red lines depict the boundaries of the peak. Each point represents a single nucleotide polymorphism (SNP) and the color shows the strength of LD between the top SNP inside the peak and the SNP surrounding it. The diamond shape points inside the peak depict the robust SNPs. The X-axis contains the SNP positions on the chromosome while the y-axis depicts the Wald statistic values obtained from the single-SNP based genome wide association study (GWAS) analysis. (B,C) The effects of different genotypes of the two leading SNPs identified in the combined framework for ESS and their significance (\*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ ). The phenotype depicts the de-regressed breeding values of the eggshell strength.

Among the genes found to be associated with ESS2, TRPM7 (Transient Receptor Potential Cation Channel Subfamily M Member 7) and BNC1 (Basonuclin 1) have functions related to the homeostasis of ions in the cell. On the other hand, the CDH4 (Cadherin-4) gene that was found for both ESS1 and ESS2 encodes for R-cadherin/cadherin-4 which are single-chain integral membrane glycoproteins and mediate calcium-dependent cell—cell adhesion. Reduced levels of these cell adhesion molecules lead to the age-related decline in tissue homeostasis [243]. Along with other members of the cadherin superfamily, R-cadherins play roles in cell differentiation in a variety of tissues including bones, kidneys and uterus [244, 245, 246, 247].

### 5.3.1.2. Association Analysis of Egg Weight

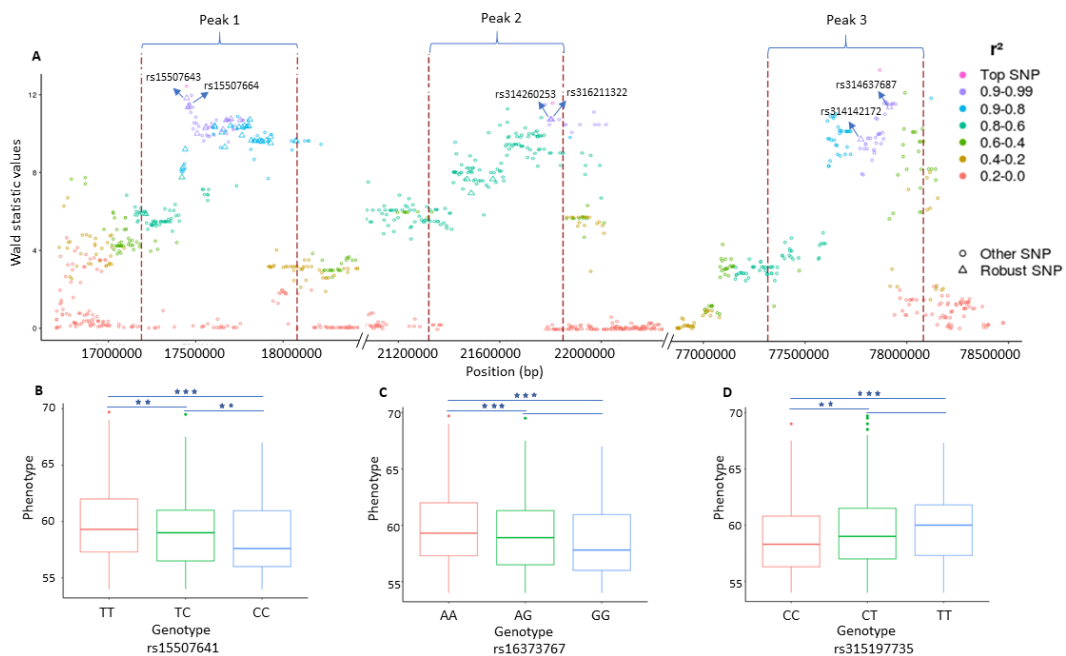
The analysis of the EW dataset resulted in the detection of eleven QTLs including the one revealed on chromosome 1 (GGA1) in the original study [60]. The additional QTLs were found on chromosomes 4 (GGA4), 12 (GGA12), 13 (GGA13), 14 (GGA14), 15 (GGA15) and 18 (GGA18). The details of these eleven QTLs are summarized in the Table 5.5. Remarkably, there is no overlap between the QTLs observed for EW and ESS. The application of the RF classifier on this dataset provides a list of 753 important SNPs. A closer look at these SNPs points out that 145 of them (including 41 SNP identified in the original study [60]) are defined to be robust SNPs due to their genomic positions within the QTLs. Similar to the analysis of the ESS dataset, LD analysis based on the EW dataset also demonstrates the presence of strong linkage between robust SNPs (Figure 5.9).

**Table 5.5.:** Significant peaks as defined in Phase 4 of our analysis framework and corresponding QTLs for EW.

Chromosome	No. of SNPs	Start Position	End Position	No. of Genes
1	304	167,931,038	169,505,140	25
4	205	17,189,770	18,080,445	9
4	143	21,319,808	21,849,558	3
4	136	77317446	78,081,369	4
12	39	2,849,562	3,010,032	7
13	49	8,495,533	8,608,578	6
14	58	7,023,793	7,188,250	4
15	41	11,193,342	11,309,808	8
15	35	11,419,957	11,514,516	3
18	30	1,057,714	1,136,220	1
18	28	1,179,899	1,238,583	0

The extraction of the genes associated with the robust SNPs related to EW results in the

determination of 16 genes. Despite no overlap between the QTLs identified for ESS and EW, a variety of genes are involved in the same biological functions. Especially, many of the genes have their functions annotated to trans-membrane transportation of minerals and proteins. In this regard, genes including SCNN1G (Sodium Channel Epithelial 1 Subunit Gamma), AFAP1L1 (Actin Filament Associated Protein 1 Like 1), CD99L2 (CD99 Molecule Like 2), GPR50 (G Protein-Coupled Receptor 50), GRIA2 (Glutamate Ionotropic Receptor AMPA Type Subunit), GRPEL2 (GGrpE Like 2, Mitochondrial), HS3ST4 (SH3 Domain And Tetratricopeptide Repeats 2), ITM2B (Integral Membrane Protein 2B), MED4 (Mediator Complex Subunit 4), MTMR1 (Myotubularin Related Protein 1) and SH3TC2 (SH3 Domain And Tetratricopeptide Repeats 2) encode proteins that are part of cell membranes.



**Figure 5.9.:** Plot representing three genomic regions on chromosome 4 in association with egg weight (EW). **(A)** Plot representing the LD structure inside and around the significant peaks. The dotted red lines depict the boundaries of the peaks. Each point represents a SNP and the color shows the strength of linkage disequilibrium (LD) between the top single nucleotide polymorphisms (SNPs) inside each peak and the surrounding SNPs. The diamond shape points inside the peak depict the robust SNPs. The X-axis contains the SNP positions on the chromosome while the y-axis depicts the Wald statistic values obtained from single-SNP based GWAS analysis. **((B–D)** The effects of different genotypes of the three leading SNPs identified for EW and their significance (\*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ )).

By regulating the transport of ingredients for the egg development they can play a role in the determination of EW. More importantly, the SCNN1G encodes a non-voltage gated sodium channel to ensure the trans-membrane transportation of sodium ions. Higher expression of this gene during egg formation has been reported to play an important role in the determination of eggshell quality [248]. Similarly, the GRIA2 gene product functions as ligand-activated cation channel that allows the trans-membrane transportation of different ions. On the other hand, genes like RCBTB2 (RCC1 and BTB Domain Containing Protein 2) and TBC1D8B (TBC1 Domain Family Member 8B) can play a role in the regulation of these transportation channels. Functional annotations of RB1 (Retinoblastoma Transcriptional Corepressor 1) and MED4 genes are related to nuclear hormone receptor binding, a process principally involved in mineral metabolism. In particular, the MED4 encoded protein is a component of the vitamin D receptor-interacting protein complex that has been shown to contribute critically for the regulation of calcium absorption in the intestine [249].





## 6. Discussion

In this chapter, I discuss the justification and important methodological aspects of the suggested frameworks. Afterwards, the results obtained by their application of these frameworks and their importance in highlighting the genetic mechanisms governing the observed phenotypic differences among the individuals are discussed.

### 6.1. Methodological Discussion

Unravelling the genetic architecture of traits has been an area of intense investigation for more than a decade. The discovery of genes associated with a trait at various organisational levels helps scientists better understand the underlying mechanisms of different traits. To uncover the associations between genetic variants and phenotypes, genome wide association studies (GWASs) have become the method of choice [3]. Despite their success in identifying a multitude of genes, the performance of GWAS strategies is limited [6, 8, 250]. Especially, deciphering genotype-phenotype associations for quantitative traits still remains challenging due to the weak contribution of many individual SNPs to the phenotype. Several approaches including single-SNP or multiple-SNP based models have been developed [4]. The worth of single-SNP models is well testified by the repertoire of genes related to a variety of traits that have been discovered using these models [3]. However, for quantitative traits where a multitude of genes may act in concert to confer a particular phenotypic value to an individual, the power of these single-SNP based models is limited [6, 8, 250].

Alternatively, multi-marker models including different Bayesian frameworks were introduced for GWAS. In these models, all SNPs are fitted simultaneously as random effects assuming a certain prior distribution of SNP effects [4]. These multi-SNP models are potentially more competent for the detection of smaller effects, but mostly require a prior distribution of SNP effects that is not known for most of the traits while for some traits they may not even follow a strict distribution [4, 251]. To overcome these limitations, combining single-SNP based statistics over a genomic region to test its association with the trait has been the method of choice for many scientists [252, 253, 254, 255]. In this regard, Beissinger et al. [190] show the superiority of cubic smoothing spline techniques over some other methods to combine single-SNP based statistics for the discovery of selection signatures. Furthermore, Zhang et al. [47] have praised the utility of spline based techniques to integrate association statistics in order to identify the causal alleles. However, these meth-

ods do not provide a clear framework that can be used to identify genomic regions with subtle effects on the phenotypes in samples with family or population structures.

### 6.1.1. Machine Learning Models for Association Analysis

With the growing application of machine learning algorithms in the field of genomics, their application to ascertain the genotype-phenotype association is gaining importance. Unlike traditional statistical models, machine learning methods do not require these prior assumptions about the genetic architecture of traits and have been applied in GWAS in humans [31] as well as in livestock [33, 32]. Especially, Romagnoni et al. [31] and Huang et al. [29] showed that machine learning based algorithms provide promising prediction power to assess genotype-phenotype associations. In particular, the Random Forests (RF) algorithm has been successfully applied for this purpose. These articles encouraged me to utilize RF in this thesis since the application of conventional GWAS methods to identify genetic variants associated with egg quality traits was futile. In this thesis I successfully applied an RF approach to two datasets to assess the importance of SNPs and identified large numbers of genes associated with eggshell strength and egg weight.

### 6.1.2. Combining RF and a Signal Detection Approach

Despite the success of the RF based approaches in association analysis, there is still a need to prioritize the identified genes to recognize the genes having robust association with the phenotype. This prioritization constitutes a means to delve deeper into the functioning of the individual genes to understand their (marginal) influences on the manifestation of the phenotype differences among the samples. For this purpose, I investigated genes within the QTLs that have association signals higher than expected by chance. The identification of QTLs is a fundamental step in my study which I have performed using a spline based strategy in several phases. Using this technique I harness the association signals, in order to detect the genomic regions harbouring genes potentially playing roles in the phenotype manifestation.

The results show that the determination of QTLs by the signal detection approach and then the prioritization of SNPs within these QTLs (called robust SNPs), can lead to the discovery of genes which despite being associated with the phenotypes, remain undetected in the typical GWAS analysis. Especially, the combined usage of both methods (RF and signal detection) not only identify the QTLs having small effects but also helps identify the SNPs in those QTLs that have a association value (peak height) higher than expected by chance. (Figures 5.8A and 5.9A). Moreover, the LD between the robust SNPs (Figures 5.8A and 5.9A) supports us, on the one hand, to monitor their strong mutual correlation which is crucial to explain the genetic makeup of the underlying QTLs. On the other hand, it further substantiates my idea regarding the presence of signals which are caused by the strong LD in the QTLs and embedded in the association statistics.

## 6.2. Biological Discussion

Applying RF, I was able to identify a remarkably high number of genes related to ESS and EW which is in agreement with the findings of Maan et al. [53, 54], Mikšík et al. [256, 257], and Brionne et al. [52], who pointed out a large number of genes/proteins involved in egg development due to the complexity of the trait. The overlap between the genes for both time points (see Figure 5.2) reflects that certain molecular functions remain relevant to eggshell development during the laying cycle of chicken. Particularly, the similarity of genes responsible for the transportation of ions is in line with the findings of Park et al. [258] and Fan et al. [197] who found that the concentration level of different ions in blood does not change with the age of the chicken. Interestingly, a closer look at the biological processes of these traits reveals that, while highly significant GO terms are involved in development for ESS1, the significant biological processes for ESS2 are rather related to different metabolic processes. The differences in biological processes at both time points could be associated with the temporal changes in the signaling cascades influencing the dynamic behavior of eggshell strength over time.

Although both traits analysed in this thesis were related to egg quality, the identified genes were found to be distinct for ESS and EW in this study. This distinction was expected as the chickens genotyped in the two datasets have different genetic backgrounds. Remarkably, however some of the identified genes are involved in the same biological function related to transmembrane transportation of elements including minerals and organic compounds. Further, the majority of the ESS1 related genes are responsible for the availability of calcium ( $\text{Ca}^{2+}$ ) and bicarbonate ( $\text{HCO}_3^{-}$ ) which are prerequisites for eggshell mineralization in the uterus part of the oviduct. These ions are supplied in large amounts via trans-epithelial transport in the uterus, for which ion channels, ion pumps and ion exchangers are required [259]. This function is mainly regulated by ATPase, an enzyme which is implicated in this process through several genes which were identified in this analysis for ESS. The ATPase enzyme decomposes ATP into ADP to release the energy required to perform energy intensive tasks by the cell. Regarding eggshell formation, ATPases have long been known to influence the microvilli of the tubular cells of the shell gland during the process of eggshell formation [240]. Similarly, inhibition of ATPase from the shell glands has been demonstrated to cause the thinning of the eggshell due to the inhibition of the calcium transport across the shell gland epithelium which is known to be an energy expensive process [260]. The hydrogen potassium ATPase maintains a certain pH level of the uterine fluid during the eggshell formation by acting as a pump to transfer the hydrogen ions ( $\text{H}^+$ ) from the uterine cell of the chicken to plasma. In this regard, two paralogs (ATP6V1B, ATP6V1C2) of the ATP6V0A2 gene found in my study have been previously reported to transfer hydrogen ions from the chicken uterine cells to blood plasma during the process of egg calcification [259, 261]. When integrated into biological membranes, the so-called transmembrane ATPases take part in the transportation of metabolites across the membranes [192]. Transmem-

brane ATPases exchange many metabolites across the membranes and provide the necessary environment for activities of the cell [262]. Similarly, genes discovered for EW encode cell membrane proteins which can act as channels for the transportation of minerals as well as proteins. Among these, one of the most important channel proteins is encoded by *SCNN1G*. This gene belongs to the sodium channel gene family. Many members of this gene family are known to affect egg weight as well as other egg quality traits [248].

The other important functional category that many of the genes related to ESS could be linked to is cell morphogenesis. Previous studies presenting the transcriptome profile of different segments of the chicken oviduct have also reported a large number of genes annotated for functions related to morphogenesis [259, 263, 264]. It is also important to note the difference in genes identified for ESS1 and ESS2. It depicts the change in the genetic and environmental components of the phenotypic variance over age which has been previously reported for other complex traits [265, 266].

### 6.2.1. Deciphering the Regulatory Mechanisms Underlying Eggshell Strength

Identification of the regulatory mechanism governing the expression of the genes underlying the important traits is considered as important as identifying the genes associated with the trait, if not more than that. In line with previous studies [113, 267, 183, 268, 186], I applied a systems biology approach and identified master regulators to investigate and unravel the transcriptional regulatory machinery of ESS associated genes. Interestingly, my results show that, similar to the genes, there are common master regulators (*Sox11* and *Slc22a1*) for both time points, which are likely to govern various eggshell related processes during the laying of the birds. In particular, being a member of the *Slc* superfamily which is involved in the transmembrane transport, the *Slc22a1* could be essential to eggshell development. For ESS1, the most promising master regulator *Scn11a* controls the sodium transport in the uterus [195, 196] to maintain a voltage difference as well as osmolarity across the uterine cell membranes to help in the calcium transportation [197]. In ESS2, the master regulator *Tead2* together with the master regulator *Sox11* underline the importance of bone remodeling during the later stages of the production cycle of the chicken.

Another fundamental step of my analysis was the identification of over-represented pathways. The results of this analysis also reinforce the findings of the gene set analysis as well as the identified master regulators. Some of the over-represented pathways were conserved at both time points while others were age-specific. Here, I specifically highlight the well-characterized TGF- $\beta$  pathway that interacts with most of the identified pathways in my analysis to regulate bone homeostasis and thus might play an important role in ESS [228]. The majority of the remaining pathways, especially those which are common to both time points, were found to be related to the cell cycle. The uterine epithelium and bone are the tissues that actively take part in the development of the eggshell, hence the renewal of the cells of both tissues is crucial for the synthesis of a strong eggshell [51]. Furthermore, mul-

tiple studies suggest that a declining ability of uterine epithelium cells to transport calcium is the main reason of the age-related deterioration of eggshells [197, 258]. In particular, the ESS2 specific p73alpha —/ NF-Y pathway that results in the inactivation of the NF-Y transcription factor by p73 proteins and consequently causes replicative senescence of cells [236] may also point towards the underlying reason for weaker eggshells during the later stages of the production cycle.

Recently, the use of systems biology based approaches to study the traits of economic importance is gaining importance in the field of agriculture [269, 267, 268, 183]. However, one of the major impediments in the use of this approach in practical animal breeding is to integrate this large amount of information into traditional genetic evaluation programs [270]. A small group of master regulators such as those identified in my analysis integrated into prediction models can possibly be a remedy and might provide novel breeding targets to improve the economically important trait of ESS. Additionally, the knowledge about the specific pathways such as TGF- $\beta$  could provide novel hypotheses for further studies. To the best of my knowledge, it is the first time that the master regulators and over-represented pathways have been revealed in the context of eggshell strength.



## 7. Conclusion

In this chapter, I first summarize the frameworks designed in this thesis along with their results. Afterwards, I give an outlook in which I provide some ideas for method extensions and list some potential applications for future research topics.

### 7.1. Summary

In this thesis, I developed two analysis frameworks with the aim to help decipher the intricate genetic background of quantitative traits. In this respect, first I performed a conventional single-SNP based GWAS which is a well established method to identify genomic variants and genes associated with the trait of interest. As evident from the results presented in the Section 5.1, this type of conventional GWAS had limited power to detect SNPs associated with eggshell strength (ESS) and eggweight (EW). To overcome this limitation, I designed two frameworks based on a Random Forests feature selection technique that can identify the genotype-phenotype associations that remain undetected in single-SNP based GWAS (Section 5.2.1). Furthermore, for the analysis of genes corresponding to the important SNPs identified by RF approach, I incorporated a systems biology approach in my first framework that can highlight the regulatory mechanism governing the expression of those genes. I used this framework to investigate the key regulatory mechanisms governing the development of eggshell strength. Moreover, to highlight the temporal changes in the signaling cascades governing the dynamic eggshell strength during the life of the birds, I considered chicken eggshell strength at two different time points during the egg production cycle. As a result I delineated the key master regulators and regulatory pathways for both time points (Sections 5.2.3 and 5.2.4). My results indicate that, while some of the master regulators (*Slc22a1* and *Sox11*) and pathways are common at different laying stages of chicken, others (e.g., *Scn11a*, *St8sia2*, or the *TGF- $\beta$*  pathway) represent age-specific functions.

In my second framework, I combined the Random Forests with a signal detection strategy to identify robust genotype-phenotype associations. This framework consists of two main steps. The first step is to fit cubic splines to the single-SNP based test statistic values to identify genomic regions with spline-peaks that are higher than expected by chance. These regions are considered as quantitative trait loci (QTL). Then the SNPs in these QTLs are prioritized with respect to the strength of their association with the phenotype using a Random Forests approach. As a case study, we applied our procedure to eggshell strength and

egg weight datasets and found trustworthy numbers of, partially novel, genomic variants and genes involved in both of the traits (Section 5.3).

Overall, the results of my thesis provide: (i) novel genes that are potentially associated with the studied traits; (ii) significant insights into age-specific and common molecular mechanisms underlying the regulation of eggshell strength; and (iii) new breeding targets to improve the egg quality at different stages of the chicken production cycle.

## 7.2. Conclusions

To decipher the genetic background of important traits using the a genomic dataset, genotype-phenotype association studies are commonly applied. In this regard I also employed a single-SNP regression based GWAS to identify potential candidate genes influencing the eggshell strength and egg weight in chicken. My results of GWAS confirmed the notion that this conventional association analysis lack power to detect the weak association signals. To overcome this limitation, I designed two analyses frameworks based on Random Forests (RF) feature selection strategy and its combination with a signal detection approach. Additionally, the first framework employs a systems biology/genetics approach using the genes corresponding to the important SNPs identified by the RF algorithm. Regulatory pathways involving the identified genes and their master regulators are delineated. This analysis was used to highlight the regulatory machinery underlying the eggshell strength at two stages of the egg production cycle. As a result, I identified master regulators and regulatory pathways which are generally associated with the functions that ensure the homeostasis of minerals as well as the tissues involved in the egg development. Furthermore, my findings also indicate that some biological processes related to eggshell development remain conserved across production stages while others are age-specific and thus changing over time. To the best of my knowledge, this is the first study revealing master regulators and over-represented pathways in the context of eggshell strength and my findings could be further utilized to design novel hypothesis for future studies. In the second framework, I designed a two step analysis that combines a RF based feature selection algorithm and a signal detection strategy for robust identification of genotype-phenotype associations. This analysis framework, on one hand, is able to identify weak association signals that are missed by conventional GWAS approaches, on the other hand, prioritizes the associations identified by the RF algorithm. Using this procedure I was able to identify a robust set of SNPs and their corresponding genes which were potentially associated with eggshell strength and egg weight in chicken. Finally, based on all of these results I can conclude that the analysis frameworks described in this thesis are well appropriate to decipher the genetic background of quantitative traits and can hence be used to study the genetics of different economically important traits.



### 7.3. Outlook

Regarding the detection of genotype-phenotype association analysis based on the Random Forests (RF) feature selection strategy it might be worthwhile to evaluate RF algorithms other than the one used in this thesis. Similarly, regarding the use of splines in association analysis, other types of the spline methods can be incorporated in our framework. To further enhance our understanding of the genetic background of the quantitative traits the usage of multi-omics data can highlight the interrelationships of the involved molecules and their functions and can hence prove very powerful and accurate to study the complex biological processes holistically. In this regard, multi-omics approaches that are able to integrate the data obtained from studies of the genome, transcriptome, proteome, epigenome, and metabolome, can be used. In this regard, systems genetic approaches that integrate genomics and transcriptomics data to identify expression quantitative trait loci (eQTLs) associated with the expression of genes can be used. This combined approach can be much more powerful to reveal heritable variations in the transcriptome as well as to study the functions of causal genes underlying QTL regions. All these different sources of information, when integrated systematically, can highlight the intricate genetic mechanisms underlying complex traits.



## Bibliography

- [1] Gallagher MD, Chen-Plotkin AS: **The post-GWAS era: from association to function.** *The American Journal of Human Genetics* 2018, **102**(5):717–730.
- [2] Zhang H, Wang Z, Wang S, Li H: **Progress of genome wide association study in domestic animals.** *Journal of animal science and biotechnology* 2012, **3**:26.
- [3] 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.
- [4] Schmid M, Bennewitz J: **Invited review: Genome-wide association analysis for quantitative traits in livestock—a selective review of statistical models and experimental designs.** *Archiv fuer Tierzucht* 2017, **60**(3):335.
- [5] Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ: **Accounting for multiple comparisons in a genome-wide association study (GWAS).** *BMC genomics* 2010, **11**:724.
- [6] Bush WS, Moore JH: **Genome-wide association studies.** *PLoS computational biology* 2012, **8**(12):e1002822.
- [7] Korte A, Farlow A: **The advantages and limitations of trait analysis with GWAS: a review.** *Plant methods* 2013, **9**:29.
- [8] Holland D, Fan CC, Frei O, Shadrin AA, Smeland OB, Sundar V, Andreassen OA, Dale AM: **Estimating inflation in GWAS summary statistics due to variance distortion from cryptic relatedness.** *BioRxiv* 2017, :164939.
- [9] Zhang YM, Mao Y, Xie C, Smith H, Luo L, Xu S: **Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.).** *Genetics* 2005, **169**(4):2267–2275.
- [10] Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al.: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nature genetics* 2006, **38**(2):203–208.

- [11] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, Eskin E, et al.: **Variance component model to account for sample structure in genome-wide association studies.** *Nature genetics* 2010, **42**(4):348–354.
- [12] Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nature genetics* 2012, **44**(7):821–824.
- [13] Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ, 2 WTCCC, et al.: **Comparison of methods to account for relatedness in genome-wide association studies with family-based data.** *PLoS genetics* 2014, **10**(7).
- [14] Balding DJ: **A tutorial on statistical methods for population association studies.** *Nature reviews genetics* 2006, **7**(10):781–791.
- [15] Young AI: **Solving the missing heritability problem.** *PLoS genetics* 2019, **15**(6):e1008222.
- [16] Long AD, Langley CH: **The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits.** *Genome research* 1999, **9**(8):720–731.
- [17] Akey J, Jin L, Xiong M: **Haplotypes vs single marker linkage disequilibrium tests: what do we gain?** *European Journal of Human Genetics* 2001, **9**(4):291.
- [18] Zhang K, Calabrese P, Nordborg M, Sun F: **Haplotype block structure and its applications to association studies: power and study designs.** *The American Journal of Human Genetics* 2002, **71**(6):1386–1394.
- [19] Lorenz AJ, Hamblin MT, Jannink JL: **Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley.** *PloS one* 2010, **5**(11):e14079.
- [20] Zhang YM, Jia Z, Dunwell JM: **The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits.** *Frontiers in plant science* 2019, **10**.
- [21] Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY, Wang SB, Dunwell JM, Zhang YM, Wu R: **Methodological implementation of mixed linear models in multi-locus genome-wide association studies.** *Briefings in bioinformatics* 2018, **19**(4):700–712.
- [22] Cui Y, Zhang F, Zhou Y: **The application of multi-Locus GWAS for the detection of salt-tolerance loci in rice.** *Frontiers in plant science* 2018, **9**:1464.

- [23] Ma L, Liu M, Yan Y, Qing C, Zhang X, Zhang Y, Long Y, Wang L, Pan L, Zou C, et al.: **Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies.** *Frontiers in plant science* 2018, **9**:561.
- [24] Xu Y, Yang T, Zhou Y, Yin S, Li P, Liu J, Xu S, Yang Z, Xu C: **Genome-wide association mapping of starch pasting properties in maize using single-locus and multi-locus models.** *Frontiers in plant science* 2018, **9**:1311.
- [25] Abed A, Belzile F: **Comparing Single-SNP, Multi-SNP, and Haplotype-Based Approaches in Association Studies for Major Traits in Barley.** *The Plant Genome* 2019, **12**(3).
- [26] Zhao Y, Chen F, Zhai R, Lin X, Wang Z, Su L, Christiani DC: **Correction for population stratification in random forest analysis.** *International journal of epidemiology* 2012, **41**(6):1798–1806.
- [27] Nguyen TT, Huang JZ, Wu Q, Nguyen TT, Li MJ: **Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests.** In *BMC genomics, Volume 16*, Springer 2015:S5.
- [28] Armero C, Cabras S, Castellanos ME, Quirós A: **Two-Stage Bayesian Approach for GWAS With Known Genealogy.** *Journal of Computational and Graphical Statistics* 2019, **28**:197–204.
- [29] Huang X, Zhou W, Bellis ES, Stubblefield J, Causey J, Qualls J, Walker K: **Minor QTLs mining through the combination of GWAS and machine learning feature selection.** *BioRxiv* 2019, :712190.
- [30] Brieuc MS, Waters CD, Drinan DP, Naish KA: **A practical introduction to Random Forest for genetic association studies in ecology and evolution.** *Molecular ecology resources* 2018, **18**(4):755–766.
- [31] Romagnoni A, Jégou S, Van Steen K, Wainrib G, Hugot JP: **Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data.** *Scientific reports* 2019, **9**:1–18.
- [32] van der Heide E, Veerkamp R, van Pelt M, Kamphuis C, Athanasiadis I, Ducro B: **Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle.** *Journal of dairy science* 2019, **102**(10):9409–9421.
- [33] Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y: **Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods.** *Frontiers in genetics* 2018, **9**:237.

- [34] Nguyen T, Le L: **Detection of SNP-SNP Interactions in Genome-wide Association Data Using Random Forests and Association Rules**. In *2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, IEEE 2018:1–7.
- [35] 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.
- [36] Hamblin MT, Jannink JL: **Factors affecting the power of haplotype markers in association studies**. *The Plant Genome* 2011, **4**(2):145–153.
- [37] Sarti F, Lasagna E, Ceccobelli S, Di Lorenzo P, Filippini F, Sbarra F, Giontella A, Pieramati C, Panella F: **Influence of single nucleotide polymorphisms in the myostatin and myogenic factor 5 muscle growth-related genes on the performance traits of Marchigiana beef cattle**. *Journal of Animal Science* 2014, **92**(9):3804–3810.
- [38] Sarti FM, Ceccobelli S, Lasagna E, Di Lorenzo P, Sbarra F, Pieramati C, Giontella A, Panella F: **Influence of single nucleotide polymorphisms in some candidate genes related to the performance traits in Italian beef cattle breeds**. *Livestock Science* 2019, **230**:103834.
- [39] Yang Y, Wu L, Wu X, Li B, Huang W, Weng Z, Lin Z, Song L, Guo Y, Meng Z, et al.: **Identification of Candidate Growth-Related SNPs and Genes Using GWAS in Brown-Marbled Grouper (*Epinephelus fuscoguttatus*)**. *Marine Biotechnology* 2020, :1–14.
- [40] Freebern E, Santos DJ, Fang L, Jiang J, Gaddis KLP, Liu GE, Vanraden PM, Maltecca C, Cole JB, Ma L: **GWAS and fine-mapping of livability and six disease traits in Holstein cattle**. *BMC genomics* 2020, **21**:41.
- [41] Sanchez MP, Guatteo R, Davergne A, Saout J, Grohs C, Deloche MC, Taussat S, Fritz S, Boussaha M, Blanquefort P, et al.: **Identification of the ABCC4, IER3, and CBFA2T2 candidate genes for resistance to paratuberculosis from sequence-based GWAS in Holstein and Normande dairy cattle**. *Genetics Selection Evolution* 2020, **52**:1–17.
- [42] Sinclair-Waters M, Ødegård J, Korsvoll SA, Moen T, Lien S, Primmer CR, Barson NJ: **Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon**. *Genetics Selection Evolution* 2020, **52**:1–11.

- [43] Horn SS, Ruyter B, Meuwissen TH, Moghadam H, Hillestad B, Sonesson AK: **GWAS identifies genetic variants associated with omega-3 fatty acid composition of Atlantic salmon fillets.** *Aquaculture* 2020, **514**:734494.
- [44] Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP: **Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci.** *Frontiers in Genetics* 2020, **11**:350.
- [45] Misra G, Badoni S, Anacleto R, Graner A, Alexandrov N, Sreenivasulu N: **Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice.** *Scientific reports* 2017, **7**:1–16.
- [46] Li C, Fu Y, Sun R, Wang Y, Wang Q: **Single-locus and multi-locus genome-wide association studies in the genetic dissection of fiber quality traits in upland cotton (*Gossypium hirsutum* L.).** *Frontiers in plant science* 2018, **9**:1083.
- [47] Zhang X, Roeder K, Wallstrom G, Devlin B: **Integration of association statistics over genomic regions using Bayesian adaptive regression splines.** *Human Genomics* 2003, **1**:20.
- [48] Schwarz DF, Szymczak S, Ziegler A, König IR: **Picking single-nucleotide polymorphisms in forests.** In *BMC proceedings, Volume 1*, Springer 2007:S59.
- [49] Bain M, Nys Y, Dunn I: **Increasing persistency in lay and stabilising egg quality in longer laying cycles. What are the challenges?** *British poultry science* 2016, **57**(3):330–338.
- [50] Chien YC, Hincke M, McKee M: **Ultrastructure of avian eggshell during resorption following egg fertilization.** *Journal of structural biology* 2009, **168**(3):527–538.
- [51] 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.
- [52] 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.
- [53] Mann K, Maček B, Olsen JV: **Proteomic analysis of the acid-soluble organic matrix of the chicken calcified eggshell layer.** *Proteomics* 2006, **6**(13):3801–3810.
- [54] Mann K, Olsen JV, Maček B, Gnad F, Mann M: **Phosphoproteins of the chicken eggshell calcified layer.** *Proteomics* 2007, **7**:106–115.

- [55] Yin Z, Lian L, Zhu F, Zhang ZH, Hincke M, Yang N, Hou ZC: **The transcriptome landscapes of ovary and three oviduct segments during chicken (*Gallus gallus*) egg formation.** *Genomics* 2019.
- [56] Crosara FSG, Pereira VJ, Lellis CG, Barra KC, Santos SKAd, Souza LCGMd, Morais TAd, Litz F, Limão VA, Braga PFS, et al.: **Is the Eggshell Quality Influenced by the Egg Weight or the Breeder Age?** *Brazilian Journal of Poultry Science* 2019, **21**(2).
- [57] Sun L, Wu R: **Mapping complex traits as a dynamic system.** *Physics of life reviews* 2015, **13**:155–185.
- [58] Nangsuay A, Ruangpanit Y, Meijerhof R, Attamangkune S: **Yolk absorption and embryo development of small and large eggs originating from young and old breeder hens.** *Poultry Science* 2011, **90**(11):2648–2655.
- [59] Yi G, Shen M, Yuan J, Sun C, Duan Z, Qu L, Dou T, Ma M, Lu J, Guo J, et al.: **Genome-wide association study dissects genetic architecture underlying longitudinal egg weights in chickens.** *BMC genomics* 2015, **16**:746.
- [60] 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.
- [61] Eberhard P: **Color atlas of genetics** 2007.
- [62] Stram DO: *Design, analysis, and interpretation of genome-wide association scans, Volume 15.* Springer 2014.
- [63] Slack J: *Genes: a very short introduction.* OUP Oxford 2014.
- [64] Patrick G: *Organic Chemistry: A Very Short Introduction.* Oxford University Press 2017.
- [65] Bourdon RM, Bourbon RM: *Understanding animal breeding, Volume 2.* Prentice Hall Upper Saddle River, NJ 2000.
- [66] González JR, Cáceres A: *Omic Association Studies with R and Bioconductor.* CRC Press 2019.
- [67] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: **Initial sequencing and analysis of the human genome** 2001.
- [68] Mendonça MAC, Carvalho CR, Clarindo WR: **DNA amount of chicken chromosomes resolved by image cytometry.** *Caryologia* 2016, **69**(3):201–206.



- [69] Watson JD: *Molecular biology of the gene*. Pearson Education India 2004.
- [70] Spencer DH, Zhang B, Pfeifer J: **Single nucleotide variant detection using next generation sequencing**. In *Clinical Genomics*, Elsevier 2015:109–127.
- [71] Brookes AJ: **The essence of SNPs**. *Gene* 1999, **234**(2):177–186.
- [72] Zhao H, Li Q, Li J, Zeng C, Hu S, Yu J: **The study of neighboring nucleotide composition and transition/transversion bias**. *Science in China Series C: Life Sciences* 2006, **49**(4):395–402.
- [73] Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations**. *Science* 2005, **307**(5712):1072–1079.
- [74] LaFramboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances**. *Nucleic acids research* 2009, **37**(13):4181–4193.
- [75] Li G, Gelernter J, Kranzler HR, Zhao H: **M3: an improved SNP calling algorithm for Illumina BeadArray data**. *Bioinformatics* 2012, **28**(3):358–365.
- [76] Xiong J: *Essential bioinformatics*. Cambridge University Press 2006.
- [77] Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser**. *Genome research* 2004, **14**(5):951–955.
- [78] Hubbard T, Andrews D, Cáccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, et al.: **Ensembl 2005**. *Nucleic acids research* 2005, **33**(suppl\_1):D447–D453.
- [79] Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, et al.: **Ensembl's 10th year**. *Nucleic acids research* 2010, **38**(suppl\_1):D557–D562.
- [80] Ruffier M, Kähäri A, Komorowska M, Keenan S, Laird M, Longden I, Proctor G, Searle S, Staines D, Taylor K, et al.: **Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation**. *Database* 2017, **2017**.
- [81] Zerbino DR, Johnson N, Juetteman T, Sheppard D, Wilder SP, Lavidas I, Nuhn M, Perry E, Raffailac-Desfosses Q, Sobral D, et al.: **Ensembl regulation resources**. *Database* 2016, **2016**.

- [82] 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**.
- [83] **UniProt: the universal protein knowledgebase**. *Nucleic acids research* 2017, **45**(D1):D158–D169.
- [84] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al.: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome research* 2012, **22**(9):1760–1774.
- [85] Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruff BJ, et al.: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes**. *Genome research* 2009, **19**(7):1316–1323.
- [86] Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al.: **Ensembl 2018**. *Nucleic acids research* 2018, **46**(D1):D754–D761.
- [87] Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart–biological queries made easy**. *BMC genomics* 2009, **10**:22.
- [88] 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**.
- [89] Vizcaíno JA, Reisinger F, Côté R, Martens L: **PRIDE and “Database on Demand” as valuable tools for computational proteomics**. In *Data Mining in Proteomics*, Springer 2011:93–105.
- [90] Croft D, O’Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, et al.: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic acids research* 2010, **39**(suppl\_1):D691–D697.
- [91] Haw RA, Croft D, Yung CK, Ndegwa N, D’Eustachio P, Hermjakob H, Stein LD: **The Reactome BioMart**. *Database* 2011, **2011**.
- [92] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al.: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome research* 2005, **15**(10):1451–1455.
- [93] Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services**. *Nucleic acids research* 2006, **34**(suppl\_2):W729–W732.

- [94] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, et al.: **Integration of biological networks and gene expression data using Cytoscape**. *Nature protocols* 2007, **2**(10):2366.
- [95] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome biology* 2004, **5**(10):R80.
- [96] 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.
- [97] Stegmaier P, Kel A, Wingender E: **geneXplainR: An R interface for the geneXplain platform**. *Journal of Open Source Software* 2017, **2**(18):412.
- [98] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.: **Gene ontology: tool for the unification of biology**. *Nature genetics* 2000, **25**:25–29.
- [99] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al.: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545–15550.
- [100] Koschmann J, Bhar A, Stegmaier P, Kel AE, Wingender E: **“Upstream analysis”: an integrated promoter-pathway analysis approach to causal interpretation of microarray data**. *Microarrays* 2015, **4**(2):270–286.
- [101] Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites**. *Nucleic acids research* 1996, **24**:238–241.
- [102] Wingender E: **Compilation of transcription regulating proteins**. *Nucleic Acids Res.* 1988, **16**(5):1879–1902.
- [103] Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al.: **Assessing computational tools for the discovery of transcription factor binding sites**. *Nature biotechnology* 2005, **23**:137–144.
- [104] Wingender E, Schoeps T, Donitz J: **TFClass: an expandable hierarchical classification of human transcription factors**. *Nucleic Acids Res.* 2013, **41**(Database issue):D165–170.
- [105] Schacherer F, Choi C, Götze U, Krull M, Pistor S, Wingender E: **The TRANSPATH signal transduction database: a knowledge base on signal transduction networks**. *Bioinformatics* 2001, **17**(11):1053–1057.

- [106] Mitsis T, Efthimiadou A, Bacopoulou F, Vlachakis D, Chrousos GP, Eliopoulos E: **Transcription factors and evolution: An integral part of gene expression.** *World Academy of Sciences Journal* 2020, **2**:3–8.
- [107] Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E: **TRANSPATH®: an integrated database on signal transduction and a tool for array analysis.** *Nucleic acids research* 2003, **31**:97–100.
- [108] Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, et al.: **TRANSPATH®: an information resource for storing and visualizing signaling pathways and their pathological aberrations.** *Nucleic acids research* 2006, **34**(suppl\_1):D546–D551.
- [109] Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboué PA, Weng W, Wilbur WJ, et al.: **GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.** *Journal of biomedical informatics* 2004, **37**:43–53.
- [110] Wang J, Zhang Y, Marian C, Resson HW: **Identification of aberrant pathways and network activities from high-throughput data.** *Briefings in bioinformatics* 2012, **13**(4):406–419.
- [111] Vijesh N, Chakrabarti SK, Sreekumar J, et al.: **Modeling of gene regulatory networks: a review.** *Journal of Biomedical Science and Engineering* 2013, **6**(02):223.
- [112] Liu E, Li L, Cheng L: **Gene Regulatory Network Review** 2019.
- [113] Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, et al.: **Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap.** *Nature protocols* 2019, **14**(2):482–517.
- [114] Cirillo E, Parnell LD, Evelo CT: **A review of pathway-based analysis tools that visualize genetic variants.** *Frontiers in genetics* 2017, **8**:174.
- [115] Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies.** *Nature Reviews Genetics* 2010, **11**(12):843–854.
- [116] García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E: **Pathway analysis: state of the art.** *Frontiers in physiology* 2015, **6**:383.
- [117] Chan SSK, Kyba M: **What is a master regulator?** *Journal of stem cell research & therapy* 2013, **3**.
- [118] Ohno S: *Major sex-determining genes, Volume 11.* Springer Science & Business Media 2013.

- [119] Mattick JS, Taft RJ, Faulkner GJ: **A global view of genomic information—moving beyond the gene and the master regulator.** *Trends in genetics* 2010, **26**:21–28.
- [120] Sikdar S, Datta S: **A novel statistical approach for identification of the master regulator transcription factor.** *BMC bioinformatics* 2017, **18**:1–11.
- [121] Century K, Reuber TL, Ratcliffe OJ: **Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products.** *Plant physiology* 2008, **147**:20–29.
- [122] Kel AE: **Search for Master Regulators in Walking Cancer Pathways.** In *Biological Networks and Pathway Analysis*, Springer 2017:161–191.
- [123] Foulkes AS: *Applied statistical genetics with R*. Springer 2009.
- [124] Khatib H: *Molecular and quantitative animal genetics*. John Wiley & Sons 2015.
- [125] Gondro C: *Primer to analysis of genomic data using R*. Springer 2015.
- [126] HAYES B: **COURSE NOTES** 2011.
- [127] Hill W, Weir B: **Maximum-likelihood estimation of gene location by linkage disequilibrium.** *American journal of human genetics* 1994, **54**(4):705.
- [128] Hartl D, Clark A: **Principles of population genetics. 2007.** *Sunderland, Massachusetts: Fourth Edition Sinauer Associates.*
- [129] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM: **A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.** *International journal of methods in psychiatric research* 2018, **27**(2):e1608.
- [130] Sun C, Qu L, Yi G, Yuan J, Duan Z, Shen M, Qu L, Xu G, Wang K, Yang N: **Genome-wide association study revealed a promising region and candidate genes for eggshell quality in an F 2 resource population.** *BMC genomics* 2015, **16**:565.
- [131] Yuan J, Wang K, Yi G, Ma M, Dou T, Sun C, Qu LJ, Shen M, Qu L, Yang N: **Genome-wide association studies for feed intake and efficiency in two laying periods of chickens.** *Genetics Selection Evolution* 2015, **47**:82.
- [132] Liu Z, Sun C, Yan Y, Li G, Shi F, Wu G, Liu A, Yang N: **Genetic variations for egg quality of chickens at late laying period revealed by genome-wide association study.** *Scientific reports* 2018, **8**:10832.
- [133] Tabangin ME, Woo JG, Martin LJ: **The effect of minor allele frequency on the likelihood of obtaining false positives.** In *BMC proceedings, Volume 3*, Springer 2009:S41.

- [134] Price AL, Zaitlen NA, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies.** *Nature Reviews Genetics* 2010, **11**(7):459–463.
- [135] van den Berg S, Vandenplas J, van Eeuwijk FA, Lopes MS, Veerkamp RF: **Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data.** *Journal of Animal Breeding and Genetics* 2019, **136**(6):418–429.
- [136] Hinrichs AL, Larkin EK, Suarez BK: **Population stratification and patterns of linkage disequilibrium.** *Genetic epidemiology* 2009, **33**(S1):S88–S92.
- [137] Reich DE, Goldstein DB: **Detecting association in a case-control study while correcting for population stratification.** *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2001, **20**:4–16.
- [138] Zheng G, Freidlin B, Gastwirth JL: **Robust genomic control for association studies.** *The American Journal of Human Genetics* 2006, **78**(2):350–356.
- [139] Stram DO: **Meta-analysis of published data using a linear mixed-effects model.** *Biometrics* 1996, :536–544.
- [140] Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z: **A SUPER powerful method for genome wide association study.** *PloS one* 2014, **9**(9):e107684.
- [141] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL: **Advantages and pitfalls in the application of mixed-model association methods.** *Nature genetics* 2014, **46**(2):100–106.
- [142] Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al.: **Mixed linear model approach adapted for genome-wide association studies.** *Nature genetics* 2010, **42**(4):355–360.
- [143] Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D: **Improved linear mixed models for genome-wide association studies.** *Nature methods* 2012, **9**(6):525–526.
- [144] Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: **Efficient control of population structure in model organism association mapping.** *Genetics* 2008, **178**(3):1709–1723.
- [145] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: **FaST linear mixed models for genome-wide association studies.** *Nature methods* 2011, **8**(10):833–835.

- [146] KARACAÖREN B, et al.: **Multiple Hypothesis Testing in a Genome Wide Association Study of Bovine Tuberculosis.** *Kafkas Universitesi Veteriner Fakültesi Dergisi* 2017, **23**:87–94.
- [147] Churchill GA, Doerge RW: **Empirical threshold values for quantitative trait mapping.** *Genetics* 1994, **138**(3):963–971.
- [148] Gao X, Becker LC, Becker DM, Starmer JD, Province MA: **Avoiding the high Bonferroni penalty in genome-wide association studies.** *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2010, **34**:100–105.
- [149] Gao X, Starmer J, Martin ER: **A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms.** *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2008, **32**(4):361–369.
- [150] Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media 2009.
- [151] Kagerer K: **A short introduction to splines in least squares regression analysis** 2013.
- [152] Meyer K: **Random regression analyses using B-splines to model growth of Australian Angus cattle.** *Genetics Selection Evolution* 2005, **37**(6):1–28.
- [153] Aguilera AM, Aguilera-Morillo M: **Comparative study of different B-spline approaches for functional data.** *Mathematical and Computer Modelling* 2013, **58**(7-8):1568–1579.
- [154] Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M: **A review of spline function procedures in R.** *BMC medical research methodology* 2019, **19**:46.
- [155] Eilers PH, Marx BD: **Splines, knots, and penalties.** *Wiley Interdisciplinary Reviews: Computational Statistics* 2010, **2**(6):637–653.
- [156] Ruppert D, Wand MP, Carroll RJ: *Semiparametric regression.* 12, Cambridge university press 2003.
- [157] De Boor C, De Boor C, Mathématicien EU, De Boor C, De Boor C: *A practical guide to splines, Volume 27.* springer-verlag New York 1978.
- [158] Wahba G: *Spline models for observational data.* SIAM 1990.
- [159] Green PJ, Silverman BW: *Nonparametric regression and generalized linear models: a roughness penalty approach.* Crc Press 1993.

- [160] Tarabichi M, Detours V, Konopka T: **Piecewise polynomial representations of genomic tracks**. *PloS one* 2012, **7**(11):e48941.
- [161] White I, Thompson R, Brotherstone S: **Genetic and environmental smoothing of lactation curves with cubic splines**. *Journal of Dairy Science* 1999, **82**(3):632–638.
- [162] Degenhardt F, Seifert S, Szymczak S: **Evaluation of variable selection methods for random forests and omics data sets**. *Briefings in bioinformatics* 2019, **20**(2):492–503.
- [163] Breiman L: **Random forests**. *Machine learning* 2001, **45**:5–32.
- [164] Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A: **Feature selection methods and genomic big data: a systematic review**. *Journal of Big Data* 2019, **6**:79.
- [165] Dong NT, Winkler L, Khosla M: **Revisiting Feature Selection with Data Complexity for Biomedicine**. *bioRxiv* 2019, :754630.
- [166] Landset S, Khoshgoftaar TM, Richter AN, Hasanin T: **A survey of open source tools for machine learning with big data in the Hadoop ecosystem**. *Journal of Big Data* 2015, **2**:24.
- [167] Kushmerick N, Weld DS, Doorenbos R: *Wrapper induction for information extraction*. University of Washington Washington 1997.
- [168] Naseriparsa M, Bidgoli AM, Varaeae T: **A hybrid feature selection method to improve performance of a group of classification algorithms**. *arXiv preprint arXiv:1403.2372* 2014.
- [169] Tsymbal A, Pechenizkiy M, Cunningham P: **Diversity in search strategies for ensemble feature selection**. *Information fusion* 2005, **6**:83–98.
- [170] Grasnack B, Perscheid C, Uflacker M: **A framework for the automatic combination and evaluation of gene selection methods**. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, Springer 2018:166–174.
- [171] Kursu MB, Jankowski A, Rudnicki WR: **Boruta—a system for feature selection**. *Fundamenta Informaticae* 2010, **101**(4):271–285.
- [172] Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, Raza S, Weidler EM, Qin X, Coarfa C, Milosavljevic A, et al.: **Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome**. *Gastroenterology* 2011, **141**(5):1782–1791.
- [173] Guo P, Luo Y, Mai G, Zhang M, Wang G, Zhao M, Gao L, Li F, Zhou F: **Gene expression profile based classification models of psoriasis**. *Genomics* 2014, **103**:48–55.



- [174] 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.
- [175] Erbe M, Cavero D, Weigend A, Weigend S, Pausch H, Preisinger R, Simianer H: **Genomic prediction in laying hens.** In *Proceedings of the 8th European Symposium on Poultry Welfare* 2013.
- [176] Ni G, Strom TM, Pausch H, Reimer C, Preisinger R, Simianer H, Erbe M: **Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken.** *BMC genomics* 2015, **16**:824.
- [177] Ni G, Cavero D, Fangmann A, Erbe M, Simianer H: **Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture.** *Genetics Selection Evolution* 2017, **49**:1–14.
- [178] Garrick DJ, Taylor JF, Fernando RL: **Deregressing estimated breeding values and weighting information for genomic regression analyses.** *Genetics Selection Evolution* 2009, **41**:55.
- [179] 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.
- [180] Kursa MB, Rudnicki WR, et al.: **Feature selection with the Boruta package.** *J Stat Softw* 2010, **36**(11):1–13.
- [181] Kursa MB, Rudnicki WR: **The all relevant feature selection using random forest.** *arXiv preprint arXiv:1106.5112* 2011.
- [182] Nguyen TT, Huang JZ, Wu Q, Nguyen TT, Li MJ: **Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests.** In *BMC genomics, Volume 16*, Springer 2015:S5.
- [183] Ayalew Y, Gültas M, Effa K, Hanotte OH, Schmitt A: **Identification of candidate signature genes and key regulators associated with Trypanotolerance in the Sheko Breed.** *Frontiers in Genetics* 2019, **10**:1095.
- [184] Wlochowitz D, Haubrock M, Arackal J, Bleckmann A, Wolff A, Beißbarth T, Winger E, Gültas M: **Computational identification of key regulators in two different colorectal cancer cell lines.** *Frontiers in genetics* 2016, **7**:42.

- [185] Wingender E, Kel A: **geneXplain—eine integrierte Bioinformatik-Plattform.** *BIOspektrum* 2012, **18**(5):554–556.
- [186] Koschmann J, Bhar A, Stegmaier P, Kel A, Wingender E: **“Upstream analysis”: an integrated promoter-pathway analysis approach to causal interpretation of microarray data.** *Microarrays* 2015, **4**(2):270–286.
- [187] Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboué PA, Weng W, Wilbur WJ, et al.: **GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.** *Journal of biomedical informatics* 2004, **37**:43–53.
- [188] Wood SN: *Generalized additive models: an introduction with R.* Chapman and Hall/CRC 2017.
- [189] Silverman BW: **Some aspects of the spline smoothing approach to non-parametric regression curve fitting.** *Journal of the Royal Statistical Society: Series B (Methodological)* 1985, **47**:1–21.
- [190] Beissinger TM, Rosa GJ, Kaeppler SM, Gianola D, De Leon N: **Defining window-boundaries for genomic analyses using smoothing spline techniques.** *Genetics Selection Evolution* 2015, **47**:30.
- [191] Jonchère V, Brionne A, Gautron J, Nys Y: **Identification of uterine ion transporters for mineralisation precursors of the avian eggshell.** *BMC physiology* 2012, **12**:10.
- [192] Chakraborti S, Dhalla NS: *Regulation of Membrane Na<sup>+</sup>-K<sup>+</sup> ATPase.* Springer 2016.
- [193] Colbran RJ: **Targeting of calcium/calmodulin-dependent protein kinase II.** *Biochemical Journal* 2004, **378**:1–16.
- [194] Meyer MB, Watanuki M, Kim S, Shevde NK, Pike JW: **The human transient receptor potential vanilloid type 6 distal promoter contains multiple vitamin D receptor binding sites that mediate activation by 1, 25-dihydroxyvitamin D3 in intestinal cells.** *Molecular endocrinology* 2006, **20**(6):1447–1461.
- [195] Ogata K, Jeong SY, Murakami H, Hashida H, Suzuki T, Masuda N, Hirai M, Isahara K, Uchiyama Y, Goto J, et al.: **Cloning and expression study of the mouse tetrodotoxin-resistant voltage-gated sodium channel  $\alpha$  subunit NaT/Scn11a.** *Biochemical and biophysical research communications* 2000, **267**:271–277.
- [196] Seda M, Pinto FM, Wray S, Cintado CG, Noheda P, Buschmann H, Candenas L: **Functional and molecular characterization of voltage-gated sodium channels in uteri from nonpregnant rats.** *Biology of reproduction* 2007, **77**(5):855–863.

- [197] Fan YF, Hou ZC, Yi GQ, Xu GY, Yang N: **The sodium channel gene family is specifically expressed in hen uterus and associated with eggshell quality traits.** *BMC genetics* 2013, **14**:90.
- [198] Koepsell H: **The SLC22 family with transporters of organic cations, anions and zwitterions.** *Molecular aspects of medicine* 2013, **34**(2-3):413–435.
- [199] Chowdhury S, Smith T: **Dietary interaction of 1, 4-diaminobutane (putrescine) and calcium on eggshell quality and performance in laying hens.** *Poultry Science* 2002, **81**:84–91.
- [200] Shinki T, Tanaka H, Takito J, Yamaguchi A, Nakamura Y, Yoshiki S, Suda T: **Putrescine is involved in the vitamin D action in chick intestine.** *Gastroenterology* 1991, **100**:113–122.
- [201] Altimimi HF, Schnetkamp PP: **Na<sup>+</sup>/Ca<sup>2+</sup>-K<sup>+</sup> exchangers (NCKX): functional properties and physiological roles.** *Channels* 2007, **1**(2):62–69.
- [202] Gadi J, Jung SH, Lee MJ, Jami A, Ruthala K, Kim KM, Cho NH, Jung HS, Kim CH, Lim SK: **The transcription factor protein Sox11 enhances early osteoblast differentiation by facilitating proliferation and the survival of mesenchymal and osteoblast progenitors.** *Journal of Biological Chemistry* 2013, **288**(35):25400–25413.
- [203] ELAROUSSI MA, FORTE LR, EBER SL, BIELLIER HV: **Calcium Homeostasis in the Laying Hen.: 1. Age and Dietary Calcium Effects.** *Poultry Science* 1994, **73**(10):1581–1589.
- [204] Häkelien AM, Bryne JC, Harstad KG, Lorenz S, Paulsen J, Sun J, Mikkelsen TS, Myklebost O, Meza-Zepeda LA: **The regulatory landscape of osteogenic differentiation.** *Stem Cells* 2014, **32**(10):2780–2793.
- [205] Scheidegger EP, Sternberg LR, Roth J, Lowe JB: **A human STX cDNA confers polysialic acid expression in mammalian cells.** *Journal of Biological Chemistry* 1995, **270**(39):22685–22688.
- [206] Itoh T, Munakata K, Adachi S, Hatta H, Nakamura T, Kato T, Kim M, et al.: **Chalaza and egg yolk membrane as excellent sources of sialic acid (N-acetylneuraminic acid) for an industrial-scale preparation.** *Japanese Journal of Zootechnical Science* 1990, **61**(3):277–282.
- [207] Nakano K, Nakano T, Ahn D, Sim J: **Sialic acid contents in chicken eggs and tissues.** *Canadian Journal of Animal Science* 1994, **74**(4):601–606.
- [208] Nakano T, Ikawa N, Ozimek L: **Chemical composition of chicken eggshell and shell membranes.** *Poultry Science* 2003, **82**(3):510–514.

- [209] Du J, Hincke MT, Rose-Martel M, Hennequet-Antier C, Brionne A, Cogburn LA, Nys Y, Gautron J: **Identifying specific proteins involved in eggshell membrane formation using gene expression analysis and bioinformatics.** *BMC genomics* 2015, **16**:792.
- [210] 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**:57.
- [211] Ahmed TA, Suso HP, Hincke MT: **Experimental datasets on processed eggshell membrane powder for wound healing.** *Data in brief* 2019, **26**:104457.
- [212] Ahmed TA, Suso HP, Hincke MT: **In-depth comparative analysis of the chicken eggshell membrane proteome.** *Journal of proteomics* 2017, **155**:49–62.
- [213] Kim Ym, Kim WY, Nam SA, Choi AR, Kim H, Kim YK, Kim HS, Kim J: **Role of PROX1 in the transforming ascending thin limb of Henle's loop during mouse kidney development.** *PloS one* 2015, **10**(5).
- [214] Malumbres M: **Cyclin-dependent kinases.** *Genome biology* 2014, **15**(6):122.
- [215] Ogasawara T, Mori Y, Abe M, Suenaga H, Kawase-Koga Y, Saijo H, Takato T: **Role of cyclin-dependent kinase (Cdk) 6 in osteoblast, osteoclast, and chondrocyte differentiation and its potential as a target of bone regenerative medicine.** *Oral Science International* 2011, **8**:2–6.
- [216] Whitehead C: **Overview of bone biology in the egg-laying hen.** *Poultry science* 2004, **83**(2):193–199.
- [217] Bar A: **Calcium transport in strongly calcifying laying birds: mechanisms and regulation.** *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 2009, **152**(4):447–469.
- [218] Ono K, Han J: **The p38 signal transduction pathway activation and function.** *Cellular signalling* 2000, **12**:1–13.
- [219] Suzanne M, Irie K, Glise B, Agnès F, Mori E, Matsumoto K, Noselli S: **The Drosophila p38 MAPK pathway is required during oogenesis for egg asymmetric development.** *Genes & Development* 1999, **13**(11):1464–1474.
- [220] Zelcer N, Tontonoz P: **Liver X receptors as integrators of metabolic and inflammatory signaling.** *The Journal of clinical investigation* 2006, **116**(3):607–614.
- [221] Vaya J, Schipper HM: **Oxysterols, cholesterol homeostasis, and Alzheimer disease.** *Journal of neurochemistry* 2007, **102**(6):1727–1737.

- [222] Griffiths WJ, Abdel-Khalik J, Crick PJ, Yutuc E, Wang Y: **New methods for analysis of oxysterols and related compounds by LC-MS.** *The Journal of steroid biochemistry and molecular biology* 2016, **162**:4–26.
- [223] Beck M, Hansen K: **Role of estrogen in avian osteoporosis.** *Poultry science* 2004, **83**(2):200–206.
- [224] Mackrill JJ: **Oxysterols and calcium signal transduction.** *Chemistry and physics of lipids* 2011, **164**(6):488–495.
- [225] Kha HT, Basseri B, Shouhed D, Richardson J, Tetradis S, Hahn TJ, Parhami F: **Oxysterols regulate differentiation of mesenchymal stem cells: pro-bone and anti-fat.** *Journal of Bone and Mineral Research* 2004, **19**(5):830–840.
- [226] Frederick JP, Liberati NT, Waddell DS, Shi Y, Wang XF: **Transforming growth factor  $\beta$ -mediated transcriptional repression of c-myc is dependent on direct binding of Smad3 to a novel repressive Smad binding element.** *Molecular and cellular biology* 2004, **24**(6):2546–2559.
- [227] Chen CR, Kang Y, Siegel PM, Massagué J: **E2F4/5 and p107 as Smad cofactors linking the TGF $\beta$  receptor to c-myc repression.** *Cell* 2002, **110**:19–32.
- [228] Tang SY, Alliston T: **Regulation of postnatal bone homeostasis by TGF $\beta$ .** *BoneKEy reports* 2013, **2**.
- [229] Bell DM, Leung KK, Wheatley SC, Ng LJ, Zhou S, Ling KW, Sham MH, Koopman P, Tam PP, Cheah KS: **SOX9 directly regulates the type-II collagen gene.** *Nature genetics* 1997, **16**(2):174–178.
- [230] Massagué J, Chen YG: **Controlling TGF- $\beta$  signaling.** *Genes & development* 2000, **14**(6):627–644.
- [231] Lönn P, Vanlandewijck M, Raja E, Kowanetz M, Watanabe Y, Kowanetz K, Vasilaki E, Heldin CH, Moustakas A: **Transcriptional induction of salt-inducible kinase 1 by transforming growth factor  $\beta$  leads to negative regulation of type I receptor signaling in cooperation with the Smurf2 ubiquitin ligase.** *Journal of Biological Chemistry* 2012, **287**(16):12867–12878.
- [232] Stow LR, Jacobs ME, Wingo CS, Cain BD: **Endothelin-1 gene regulation.** *The FASEB Journal* 2011, **25**:16–28.
- [233] Strait KA, Stricklett PK, Kohan JL, Miller MB, Kohan DE: **Calcium regulation of endothelin-1 synthesis in rat inner medullary collecting duct.** *American Journal of Physiology-Renal Physiology* 2007, **293**(2):F601–F606.

- [234] Yoon SS, Jung KI, Choi YL, Choi EY, Lee IS, Park SH, Kim TJ: **Engagement of CD99 triggers the exocytic transport of ganglioside GM1 and the reorganization of actin cytoskeleton.** *FEBS letters* 2003, **540**(1-3):217–222.
- [235] Pasello M, Manara MC, Scotlandi K: **CD99 at the crossroads of physiology and pathology.** *Journal of cell communication and signaling* 2018, **12**:55–68.
- [236] Yao Y, Bellon M, Shelton SN, Nicot C: **Tumor suppressors p53, p63TA $\alpha$ , p63TA $\gamma$ , p73 $\alpha$ , and p73 $\beta$  use distinct pathways to repress telomerase expression.** *Journal of Biological Chemistry* 2012, **287**(24):20737–20747.
- [237] Jung MS, Yun J, Chae HD, Kim JM, Kim SC, Choi TS, Shin DY: **p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of NF-Y transcription factor.** *Oncogene* 2001, **20**(41):5818–5825.
- [238] Alimov AP, Park-Sarge OK, Sarge KD, Malluche HH, Koszewski NJ: **Transactivation of the parathyroid hormone promoter by specificity proteins and the nuclear factor Y complex.** *Endocrinology* 2005, **146**(8):3409–3416.
- [239] Jääskeläinen T, Huhtakangas J, Mäenpää P: **Negative regulation of human parathyroid hormone gene promoter by vitamin D3 through nuclear factor Y.** *Biochemical and biophysical research communications* 2005, **328**(4):831–837.
- [240] Yamamoto T, Ozawa H, Nagai H: **Histochemical studies of Ca-ATPase, succinate and NAD $^{+}$ -dependent isocitrate dehydrogenases in the shell gland of laying Japanese quails: with special reference to calcium-transporting cells.** *Histochemistry* 1985, **83**(3-4):221–226.
- [241] Wang Y, Guo F, Qu H, Luo C, Wang J, Shu D: **Associations between variants of bone morphogenetic protein 7 gene and growth traits in chickens.** *British poultry science* 2018, **59**(3):264–269.
- [242] Jin S: **Bipotent stem cells support the cyclical regeneration of endometrial epithelium of the murine uterus.** *Proceedings of the National Academy of Sciences* 2019, **116**(14):6848–6857.
- [243] Boyle M, Wong C, Rocha M, Jones DL: **Decline in self-renewal factors contributes to aging of the stem cell niche in the *Drosophila* testis.** *Cell stem cell* 2007, **1**(4):470–478.
- [244] Adams CL, Chen YT, Smith SJ, James Nelson W: **Mechanisms of epithelial cell–cell adhesion and cell compaction revealed by high-resolution tracking of E-cadherin–green fluorescent protein.** *The Journal of cell biology* 1998, **142**(4):1105–1119.

- [245] Dahl U, Sjödin A, Larue L, Radice GL, Cajander S, Takeichi M, Kemler R, Semb H: **Genetic dissection of cadherin function during nephrogenesis.** *Molecular and cellular biology* 2002, **22**(5):1474–1487.
- [246] Marie PJ, Haÿ E, Modrowski D, Revollo L, Mbalaviele G, Civitelli R: **Cadherin-mediated cell–cell adhesion and signaling in the skeleton.** *Calcified tissue international* 2014, **94**:46–54.
- [247] Vazquez-Levin MH, Marín-Briggiler CI, Caballero JN, Veiga MF: **Epithelial and neural cadherin expression in the mammalian reproductive tract and gametes and their participation in fertilization-related events.** *Developmental biology* 2015, **401**:2–16.
- [248] Fan YF, Hou ZC, Yi GQ, Xu GY, Yang N: **The sodium channel gene family is specifically expressed in hen uterus and associated with eggshell quality traits.** *BMC genetics* 2013, **14**:90.
- [249] Fleet JC, Schoch RD: **Molecular mechanisms for regulation of intestinal calcium absorption by vitamin D and other factors.** *Critical reviews in clinical laboratory sciences* 2010, **47**(4):181–195.
- [250] Josephs EB, Stinchcombe JR, Wright SI: **What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits?** *New Phytologist* 2017, **214**:21–33.
- [251] Liu Y, Wang D, He F, Wang J, Joshi T, Xu D: **Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean.** *Frontiers in Genetics* 2019, **10**:1091.
- [252] Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS: **Truncated product method for combining P-values.** *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2002, **22**(2):170–185.
- [253] Dudbridge F, Koeleman BP: **Rank truncated product of P-values, with application to genomewide association scans.** *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2003, **25**(4):360–366.
- [254] Yang HC, Lin CY, Fann CS: **A sliding-window weighted linkage disequilibrium test.** *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2006, **30**(6):531–545.
- [255] Yang HC, Hsieh HY, Fann CS: **Kernel-based association test.** *Genetics* 2008, **179**(2):1057–1068.

- [256] Mikšík I, Eckhardt A, Sedláková P, Mikulíková K: **Proteins of insoluble matrix of avian (*Gallus gallus*) eggshell.** *Connective Tissue Research* 2007, **48**:1–8.
- [257] Mikšík I, Sedláková P, Lacinová K, Pataridis S, Eckhardt A: **Determination of insoluble avian eggshell matrix proteins.** *Analytical and bioanalytical chemistry* 2010, **397**:205–214.
- [258] Park JA, Sohn SH: **The Influence of Hen Aging on Eggshell Ultrastructure and Shell Mineral Components.** *Korean journal for food science of animal resources* 2018, **38**(5):1080.
- [259] 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**:220.
- [260] Khan HM, Cutkomp L: **In vitro studies of DDT, DDE, and ATPase as related to avian eggshell thinning.** *Archives of environmental contamination and toxicology* 1982, **11**(5):627–633.
- [261] Jonchère V, Brionne A, Gautron J, Nys Y: **Identification of uterine ion transporters for mineralisation precursors of the avian eggshell.** *BMC physiology* 2012, **12**:10.
- [262] Morth JP, Pedersen BP, Buch-Pedersen MJ, Andersen JP, Vilsen B, Palmgren MG, Nissen P: **A structural overview of the plasma membrane Na<sup>+</sup>, K<sup>+</sup>-ATPase and H<sup>+</sup>-ATPase ion pumps.** *Nature Reviews Molecular Cell Biology* 2011, **12**:60.
- [263] Wan Y, Jin S, Ma C, Wang Z, Fang Q, Jiang R: **RNA-Seq reveals seven promising candidate genes affecting the proportion of thick egg albumen in layer-type chickens.** *Scientific reports* 2017, **7**:1–9.
- [264] Yin Z, Lian L, Zhu F, Zhang ZH, Hincke M, Yang N, Hou ZC: **The transcriptome landscapes of ovary and three oviduct segments during chicken (*Gallus gallus*) egg formation.** *Genomics* 2020, **112**:243–251.
- [265] Elks CE, Den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJ, Ong KK: **Variability in the heritability of body mass index: a systematic review and meta-regression.** *Frontiers in endocrinology* 2012, **3**:29.
- [266] He L, Sillanpää MJ, Silventoinen K, Kaprio J, Pitkäniemi J: **Estimating modifying effect of age on genetic and environmental variance components in twin models.** *Genetics* 2016, **202**(4):1313–1328.
- [267] Neupane M, Geary TW, Kiser JN, Burns GW, Hansen PJ, Spencer TE, Neibergs HL: **Loci and pathways associated with uterine capacity for pregnancy and fertility in beef cattle.** *PloS one* 2017, **12**(12).



- 
- [268] Woldesemayat AA, Ntwasa M: **Pathways and Network Based Analysis of Candidate Genes to Reveal Cross-Talk and Specificity in the Sorghum (*Sorghum bicolor* (L.) Moench) Responses to Drought and It's Co-occurring Stresses.** *Frontiers in genetics* 2018, **9**:557.
- [269] Ramayo-Caldas Y, Renand G, Ballester M, Saintilan R, Rocha D: **Multi-breed and multi-trait co-association analysis of meat tenderness and other meat quality traits in three French beef cattle breeds.** *Genetics Selection Evolution* 2016, **48**:37.
- [270] Kadarmideen HN, von Rohr P, Janss LL: **From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding.** *Mammalian Genome* 2006, **17**(6):548–564.

## **A. Appendix**

### **A.1. Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forest**

Article

# Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forests

Faisal Ramzan <sup>1,2,†</sup> , Selina Klees <sup>1,†</sup> , Armin Otto Schmitt <sup>1,3</sup> , David Cavero <sup>4</sup>  
and Mehmet Gültas <sup>1,3,\*</sup> 

<sup>1</sup> Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; faisal.ramzan@stud.uni-goettingen.de (F.R.); selina.klees@uni-goettingen.de (S.K.); armin.schmitt@uni-goettingen.de (A.O.S.)

<sup>2</sup> Department of Animal Breeding and Genetics, University of Agriculture Faisalabad, 38000 Faisalabad, Pakistan

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

<sup>4</sup> H&N International, 27472 Cuxhaven, Germany; cavero@ltz.de

\* Correspondence: gueltas@informatik.uni-goettingen.de

† These authors contributed equally to this work.

Received: 16 March 2020; Accepted: 21 April 2020; Published: 24 April 2020



**Abstract:** In today's chicken egg industry, maintaining the strength of eggshells in longer laying cycles is pivotal for improving the persistency of egg laying. Eggshell development and mineralization underlie a complex regulatory interplay of various proteins and signaling cascades involving multiple organ systems. Understanding the regulatory mechanisms influencing this dynamic trait over time is imperative, yet scarce. To investigate the temporal changes in the signaling cascades, we considered eggshell strength at two different time points during the egg production cycle and studied the genotype–phenotype associations by employing the Random Forests algorithm on chicken genotypic data. For the analysis of corresponding genes, we adopted a well established systems biology approach to delineate gene regulatory pathways and master regulators underlying this important trait. Our results indicate that, while some of the master regulators (*Slc22a1* and *Sox11*) and pathways are common at different laying stages of chicken, others (e.g., *Scn11a*, *St8sia2*, or the TGF- $\beta$  pathway) represent age-specific functions. Overall, our results provide: (i) significant insights into age-specific and common molecular mechanisms underlying the regulation of eggshell strength; and (ii) new breeding targets to improve the eggshell quality during the later stages of the chicken production cycle.

**Keywords:** eggshell strength; chicken; Random Forests; feature selection; master regulators; over-represented pathways

## 1. Introduction

Today's poultry industry is highly invested in the development of chicken capable of producing more eggs in longer laying cycles [1]. This production goal, however, must go hand in hand with improvement in sustainability of egg quality, especially eggshell strength (ESS), during the whole laying period [1,2]. The calcified eggshells not only provide protection against physical damage but also play a crucial role for the development of the embryo by allowing gaseous exchange, abating moisture loss, and supplying calcium for the embryo bone development [3]. Multiple molecular actors involved in the homeostasis and transportation of minerals, especially calcium, the main constituent

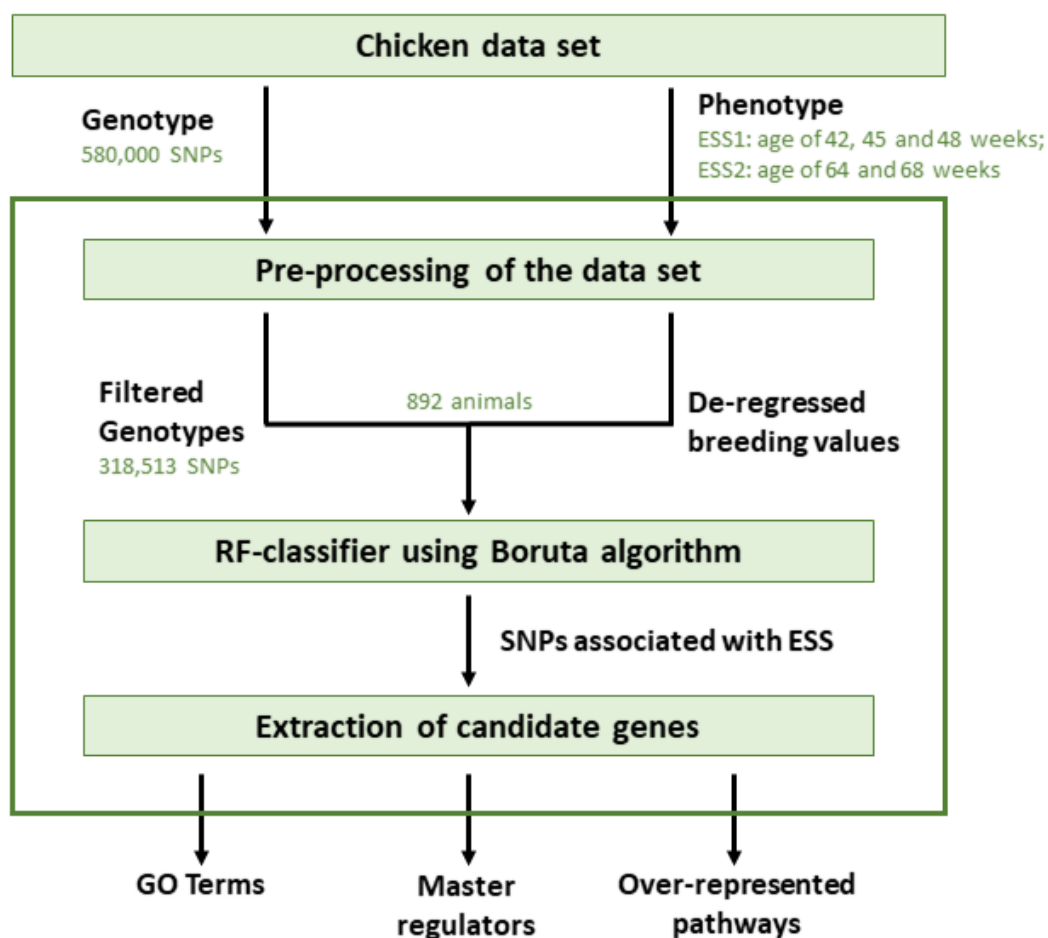
of the eggshell, have been identified [4,5]. More than 500 eggshell matrix proteins have also been reported [6,7] implicating a plethora of genes that knit together the complex protein scaffold and the mineral phase of the eggshell [5,8]. However, most of these discoveries provide only the genes expressed in a certain segment of the chicken oviduct, the principal organ for egg development, and consequently the overall mechanisms of eggshell development remain illusive. Moreover, similar to other economically important traits, ESS remains relevant throughout the productive life and commonly deteriorates with the age of the chicken [9]. This decline in the eggshell quality remains one of the major reasons for replacing commercial flocks [1]. Hence, understanding the genetic basis of ESS at different laying stages is very important for breeders if they are to extend the laying cycle of chicken. Therefore, an analysis of this trait at different time points during the life of the bird can better delineate its genetics and its molecular mechanisms involved in this dynamic behavior [10]. This knowledge can then be utilized to design breeding strategies to improve the eggshell quality during the later stages of the chicken production cycle.

Until now, a variety of association studies have been conducted to decipher the genetic architecture of quantitative traits such as ESS, which led to the identification of a valuable repertoire of genes controlling a range of traits (see the reviews [11–13]). Finding loci associated with a trait through genome wide association studies (GWAS) is commonly based on single-SNP based models that test each SNP for its association with the phenotype, ignoring its dependency on the neighboring SNPs. This statistical design of GWAS seems quite straightforward, yet entails several challenges including those of population stratification, relationships among the samples, multiple hypothesis testing, and overestimation of SNP effects, among others, as pointed out in previous studies [14–17]. Many approaches such as different multiple testing correction methods and linear mixed models have been proposed to overcome these challenges [15,18,19]. However, the most devastating challenge of GWAS still persisting is the lack of power to detect the loci having medium to small effect sizes [20]. This inability of GWAS to explain a major proportion of the heritability has been under intensive discussion.

To overcome these limitations of GWAS, application of Bayesian frameworks as well as machine learning algorithms have gained importance in the last decade [21–25]. Their comparative performance has been evaluated for a variety of traits with different genetic architectures (see the reviews [13,26,27]). Nevertheless, multiple studies have revealed that machine learning algorithms surpass currently available well-known GWAS approaches in identifying genes having small effects on the phenotype [28–30]. In particular, Briec et al. pointed out the efficiency of Random Forests (RF) models for analyzing a large number of loci simultaneously and identifying promising associations [28]. Inspired by Briec et al.'s study, we applied the RF algorithm to assess the importance of SNPs that could provide a clue of their essential roles for ESS and to characterize the differences observed in this trait at different time points. For the analysis of the corresponding genes of these SNPs, we adopted a well established systems biology approach and identified age-specific and common key regulatory pathways and master regulators. These findings could: (i) enhance our understanding of the regulatory mechanisms underlying eggshell strength; and (ii) provide novel targets and hypotheses for future breeding strategies. To the best of our knowledge, it is the first study in this field which mainly focuses on the importance of the age-specific and common key regulatory mechanisms in chicken to reveal the genetic programs influencing the eggshell strength.

## 2. Materials and Methods

In this section, we describe the chicken dataset analyzed and the methods applied. Our analysis follows the structure of Figure 1.



**Figure 1.** Flowchart of the analysis applied in this study (ESS, Eggshell strength).

### 2.1. Chicken Dataset

To explore the genomic background of the changes that incur to the eggshell strength during the life of laying birds, we analyzed a genotype dataset that has previously been used to investigate the accuracy of imputation as well as the prediction of genomic breeding values in chicken [31–33]. The dataset consists of a purebred commercial brown layer line with 892 animals and 580,000 SNPs generated using Affymetrix Axiom Chicken Genotyping Array. The genotypic data do not contain mitochondrial SNPs. The corresponding phenotypic data consist of eggshell strength (ESS) measured (as the force in Newton that was required to break the eggshell) for each bird at two distinct stages of its production cycle. These two stages were then regarded as Time Point 1 and Time Point 2, respectively. The first time point for ESS was recorded at the ages of 42, 45, and 48 weeks and the second time point was recorded at the ages of 64 and 68 weeks. Averages of the recorded breaking strengths at Time Point 1 (ESS1) and Time Point 2 (ESS2) were used as phenotypes in the further analysis. Extensive pedigree data, consisting of, in total, 40,545 individuals from six generations, were available on these birds which were included in an animal model for breeding values estimation of the birds. These breeding values were then de-regressed following Garrick et al. [34] to obtain the pseudo-phenotypes that were used in the further analysis. To ensure genotype quality, we filtered the genotyped data and removed the SNPs: (i) that were unassigned to any chromosome or present on the sex chromosomes; (ii) with a minor allele frequency  $< 0.01$ ; (iii) with a genotyping call rate  $\leq 97\%$ ; (iv) significantly deviating from Hardy–Weinberg equilibrium ( $p$ -value  $< 1 \times 10^{-6}$ ); and (v) for animals having a SNP call rate smaller than 95%. Finally, after filtering, we used 892 animals and 318,513 SNPs for our analyses.

## 2.2. Association Analysis Using Random Forests

To identify SNPs potentially associated with eggshell strength, we used the concept of the Random Forests (RF) algorithm to estimate the relative importance of each SNP (attribute) regarding its involvement in the prediction of response variables (de-regressed breeding values). For this purpose, we employed the Boruta algorithm in our study [35], which is a specially developed powerful wrapper for the RF based feature selection approach. The main principle of the Boruta algorithm is based on the extension of the attributes by adding random attributes to the dataset which are called *shadow attributes* and created by shuffling the original values of each attribute (in our case SNPs) in the dataset. The enlargement of the attributes results in apposition of the randomness to the dataset, which leads to the reduction of the bias of hidden (false) signals arising from random fluctuations or correlations in the dataset [35–37]. To this end, a RF classifier is applied to the extended dataset, and SNPs are systematically and iteratively removed whose importance are significantly smaller than those of the *shadow attributes*. By repeating the process of *shadow attributes* generation and RF algorithm application, importance is assigned to all SNPs. As a result, the Boruta algorithm provides a ranked list of SNPs with a decision of whether the importance of a SNP is confirmed, rejected, or tentative. It is important to note that a similar idea to the Boruta algorithm is manually implemented in [22] to assess the importance of SNPs.

## 2.3. Gene Set Analysis

We extracted the genes corresponding to the SNPs identified by the Boruta algorithm from Ensembl using BioMart [38] (R-script given in File S1). Furthermore, we performed a gene set analysis regarding their molecular functions to obtain functional annotations of these genes.

## 2.4. Identification of Master Regulators and Over-Represented Pathways

Following our previous studies [39,40], we performed the "upstream analysis" and pathway analysis using the geneXplain platform [41] to gain more insight into the functional relationships of genes. The algorithm of "upstream analysis" workflow was introduced by Koschmann et al. [42] and its main goal is to reveal the underlying key regulators that control the activity of target genes. For this purpose, the underlying algorithm of "upstream analysis" firstly constructs molecular pathway networks and then detects convergence points of these networks, which are called master regulators and are likely to orchestrate the transcriptional regulation of several genes. In our analysis, we used the GeneWays database [43] and ran the standard "upstream analysis" workflow with a maximum radius of 10 steps upstream to identify the top five master regulators of each gene set resulted from the previous step of the analysis.

To discover novel biological functions and to reveal the properties of the genes under study, we performed a pathway enrichment analysis as the second step of our analysis. To this end, we used the TRANSPATH pathway database [44], which is a regularly updated signaling pathway database and contains information about genes, molecules and reactions for the identification of age-specific and common over-represented pathways.

## 3. Results

In this study, we performed the RF approach using the Boruta algorithm to identify the informative SNPs associated with eggshell strength at two time points during the laying cycle of commercial brown layer chicken. For this purpose, the importance of each SNP was separately assessed for its association with the phenotype of interest. To this end, we obtained a list of SNPs for each time point whose importance was confirmed by the Boruta algorithm for the prediction of the phenotype. Analyzing both time points, we identified 3726 SNPs associated with eggshell strength at Time Point 1 (ESS1) and 1815 SNPs associated with eggshell strength at Time Point 2 (ESS2) (the lists of SNPs are given in Table S2). These SNPs were then mapped to the genome and the genes harboring at least one of these

SNPs were identified for both traits. In total, we identified 405 genes for ESS1 and 253 genes for ESS2 (the lists of genes and their Gene Ontology (GO) categories are given in Tables S2 and S3, respectively). A closer look at these gene lists reveals that 22 % (118 genes) of them are overlapping (see Figure 2), which depicts the conservation of some of the underlying mechanisms involved in the synthesis of eggshell during different stages of the egg production cycle. Our results also show that a considerably high number of genes that were distinct for the time points highlight the dynamic nature of this trait.

This section is comprised of three parts. First, to gain a deeper insight into these gene sets, we performed a gene set analysis and clustered their functions based on the GO terms. Second, we performed the “upstream analysis” introduced by Koschmann et al. [42] for the identification of specific and common master regulators of both time points. Third, we present the over-represented pathways to further elucidate the mechanisms that control the ESS at different production stages of birds.



**Figure 2.** Venn diagram depicting the number of genes associated with eggshell strength at Time Point 1 (ESS1), at Time Point 2 (ESS2), and their overlap.

### 3.1. Gene Set Analysis

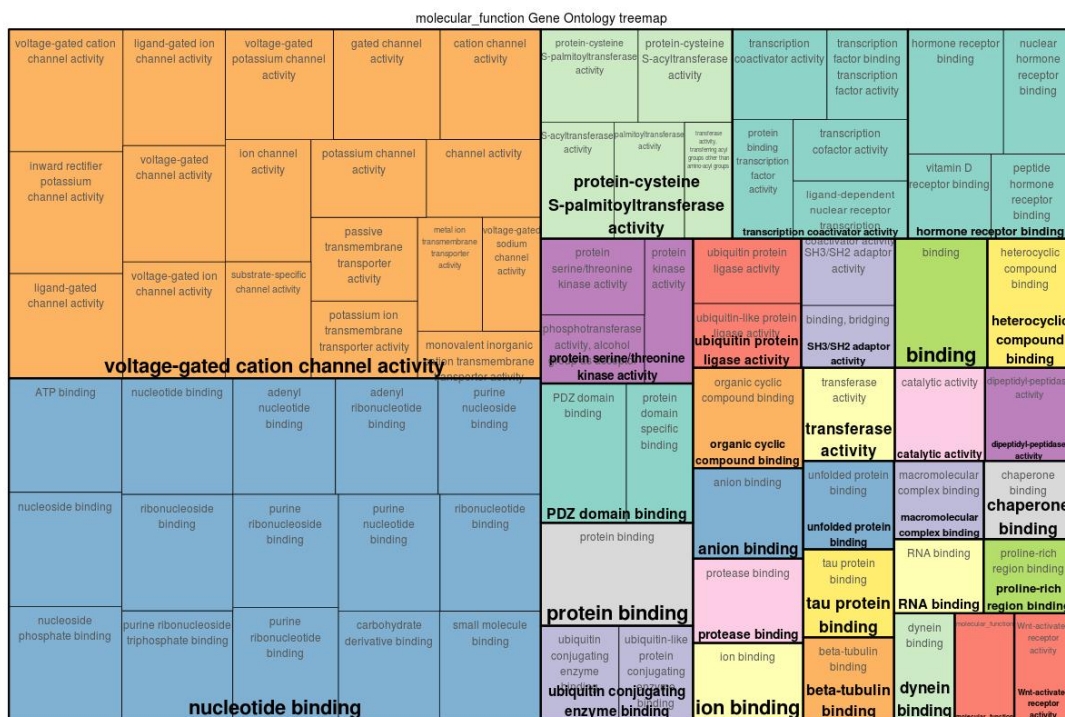
The functional classification of both gene sets indicates that there are several GO categories that were common for both time points (see the treemaps depicted in Figures 3 and 4 and the top 15 GO terms in Tables 1 and 2). In particular, the transportation of cations across membranes was the most salient function for the underlying mechanism of ESS at both time points. In this regard, calcium ions, being the main constituent of the eggshell, are supplied in large amount to the uterine fluid by transepithelial transport. In addition, other cations such as sodium, magnesium, and potassium are exchanged across the uterine endothelium to maintain the cell homeostasis [4,5]. This transmembrane transport remains important during the production cycle to ensure the development of an eggshell. The gene set analysis further reveals that the activities pertaining to ATPase, GTPase, calmodulin binding, calmodulin-dependent protein kinase, and Smad binding were specific for ESS1. Meanwhile, functions related to hormone/vitamin D receptor binding, chaperone binding, and Wnt-activated receptors were more relevant for ESS2.

Among others, the function of ATPase in eggshell formation has been well investigated in previous studies [5,45]. Along with maintaining a pH of the uterine fluid during the eggshell formation, the ATPases also provide the required energy and function as transmembrane transportation channels for ions [46]. The calmodulin binding and calmodulin-dependent protein kinase activity is known to regulate the concentration of calcium in various cells [47] and so does the vitamin D receptor binding [48]. The chaperone binding activity of the genes associated with ESS2 is another distinctive finding of this study. Chaperone proteins have been reported in the uterine fluid where they perform the folding of the eggshell matrix proteins into a rigid scaffold upon which mineralization takes place to produce the fabric of eggshell [5]. These functional classes elucidate the molecular functions that

gain more relevance depending on the age of the birds and demonstrate the key functions that remain important throughout the laying cycle of the birds.



**Figure 3.** Gene Ontology (GO) treemap for genes associated with eggshell strength at Time Point 1 (ESS1). The boxes are grouped together based on the upper-hierarchy GO-term which is written in bold letters.



**Figure 4.** Gene Ontology (GO) treemap for genes associated with eggshell strength at Time Point 2 (ESS2). The boxes are grouped together based on the upper-hierarchy GO-term which is written in bold letters.



**Table 1.** Top 15 Gene Ontology (GO) molecular function terms based on the adjusted *p*-value for the eggshell strength at Time Point 1 (ESS1).

GO Term	GO Title	Number of Genes	Adjusted <i>p</i> -Value
GO:0005515	protein binding	281	$5.11 \times 10^{-8}$
GO:0005488	binding	331	$1.97 \times 10^{-7}$
GO:0043167	ion binding	155	$4.93 \times 10^{-3}$
GO:0000146	microfilament motor activity	5	$4.93 \times 10^{-3}$
GO:0003779	actin binding	20	$6.9 \times 10^{-3}$
GO:0032559	adenyl ribonucleotide binding	49	$1.47 \times 10^{-2}$
GO:0030554	adenyl nucleotide binding	49	$1.51 \times 10^{-2}$
GO:0044877	macromolecular complex binding	50	$1.54 \times 10^{-2}$
GO:0004683	calmodulin-dependent protein kinase activity	5	$1.54 \times 10^{-2}$
GO:0005524	ATP binding	47	$2.05 \times 10^{-2}$
GO:0042623	ATPase activity, coupled	16	$2.24 \times 10^{-2}$
GO:0008092	cytoskeletal protein binding	30	$3.32 \times 10^{-2}$
GO:0043168	anion binding	74	$3.93 \times 10^{-2}$
GO:0046983	protein dimerization activity	40	$4.15 \times 10^{-2}$
GO:0017016	Ras GTPase binding	12	$4.86 \times 10^{-2}$

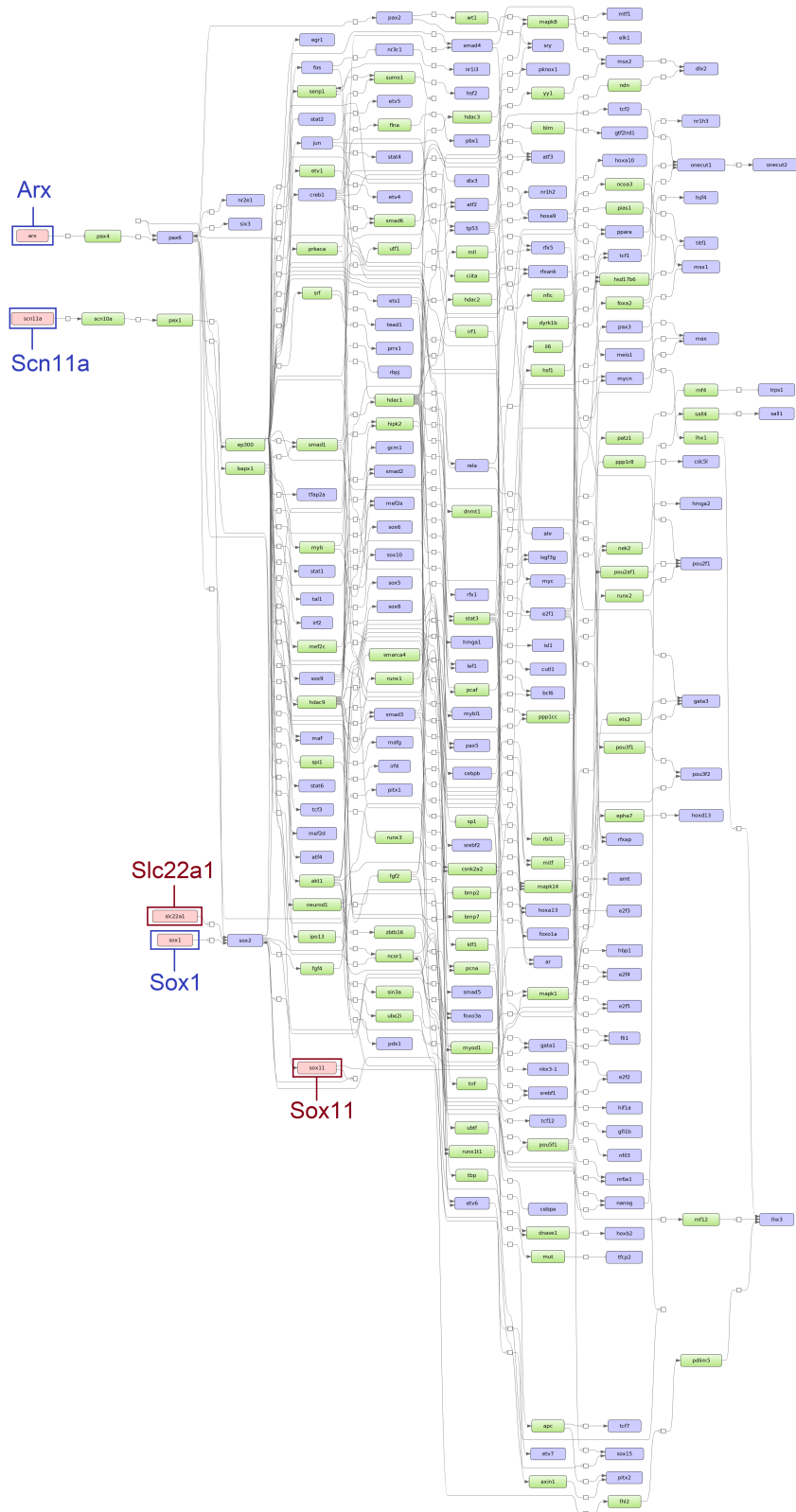
**Table 2.** Top 15 Gene Ontology (GO) molecular function terms based on the adjusted *p*-value for the eggshell strength at Time Point 2 (ESS2).

GO Term	GO Title	Number of Genes	Adjusted <i>p</i> -Value
GO:0005515	protein binding	168	$1.30 \times 10^{-2}$
GO:0022843	voltage-gated cation channel activity	9	$2.09 \times 10^{-2}$
GO:0005242	inward rectifier potassium channel activity	4	$2.10 \times 10^{-2}$
GO:0032549	ribonucleoside binding	40	$2.79 \times 10^{-2}$
GO:0000166	nucleotide binding	48	$2.79 \times 10^{-2}$
GO:0005524	ATP binding	34	$2.79 \times 10^{-2}$
GO:0001883	purine nucleoside binding	39	$3.66 \times 10^{-2}$
GO:0032559	adenyl ribonucleotide binding	34	$3.66 \times 10^{-2}$
GO:0005488	binding	199	$3.66 \times 10^{-2}$
GO:0030554	adenyl nucleotide binding	34	$3.66 \times 10^{-2}$
GO:0051427	hormone receptor binding	9	$3.66 \times 10^{-2}$
GO:0015276	ligand-gated ion channel activity	8	$3.66 \times 10^{-2}$
GO:0017076	purine nucleotide binding	39	$3.7 \times 10^{-2}$
GO:0022836	gated channel activity	12	$3.83 \times 10^{-2}$
GO:0036094	small molecule binding	50	$4.64 \times 10^{-2}$

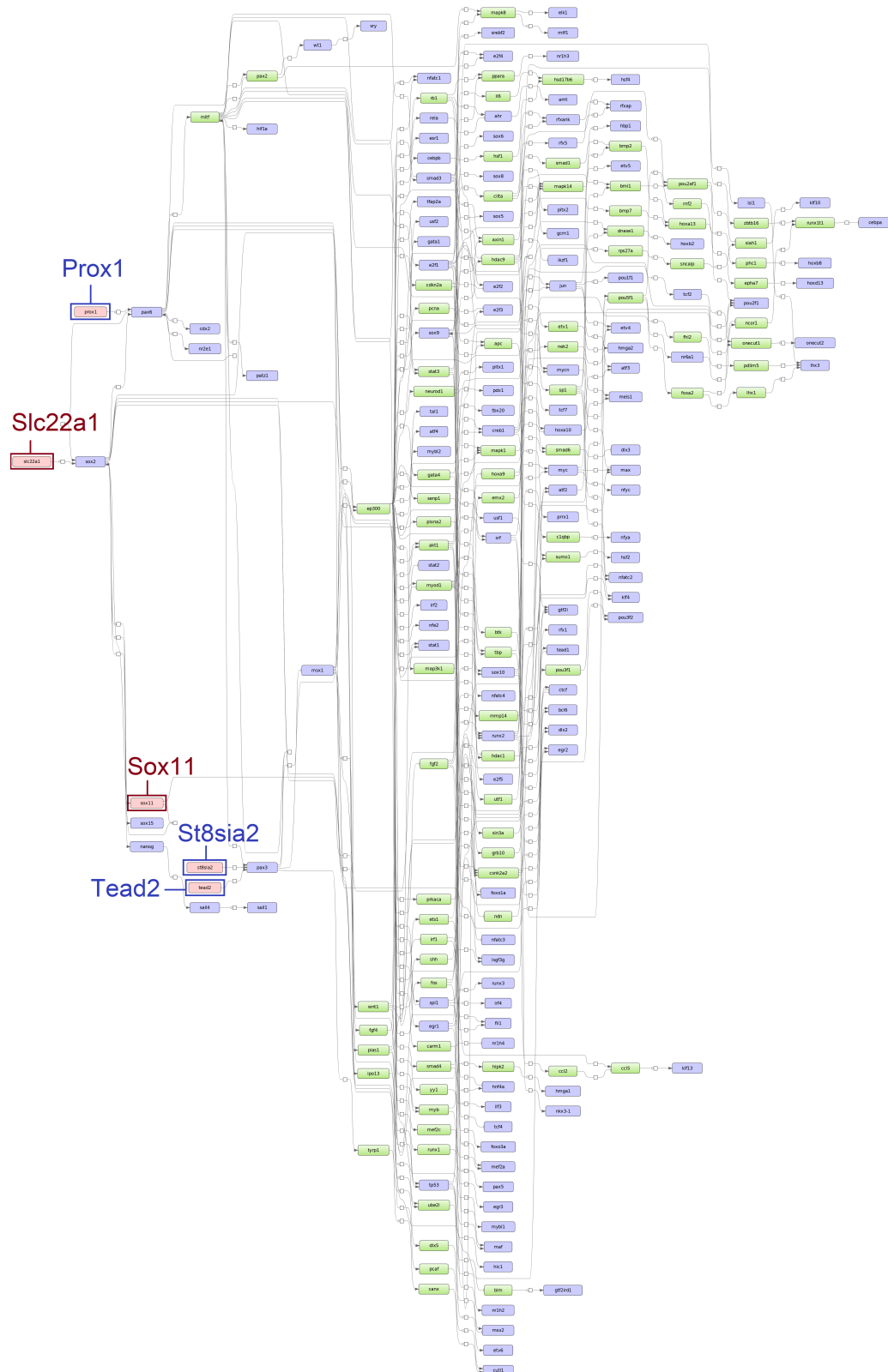
### 3.2. Identification of Master Regulators

Applying the “upstream analysis” integrated in the geneXplain platform [41], we identified the top five age-specific and common master regulators. While the master regulators *Arx*, *Sox1*, and *Scn11a* were specifically found for ESS1, the master regulators *St8sia2*, *Tead2*, and *Prox1* were identified for ESS2. Additionally, *Slc22a1* and *Sox11* were identified for both time points (see Figures 5 and 6).

The ESS1 specific master regulator *Scn11a* is a gene encoding transmembrane sodium channels which control the voltage-gated sodium transport especially in the uterus [49,50], the site of eggshell synthesis in birds. Moreover, the importance of sodium channels in the transportation of inorganic minerals deposited in the eggshell is well established [51]. Interestingly, we found the master regulator *Slc22a1* at both time points. It codes for the protein OCT1, an organic cation transporter for substrates such as putrescine [52], which plays an important role for eggshell thickness [53] and calcium transport in the intestine [54]. Furthermore, many other members of the super-family of transport proteins, *Slc* (solute carrier proteins), are well known to play an essential role in the homeostasis of calcium ions in a variety of tissues [55]. The *Slc* proteins have also been reported to transport magnesium ions during the egg calcification process [5].



**Figure 5.** Scheme of gene regulatory pathways revealing the top five master regulators (pink filled boxes) for eggshell strength at Time Point 1 (ESS1) following the “upstream analysis” [42]. The master regulators written in dark blue and surrounded by dark blue boxes (*Arx*, *Scn11a* and *Sox1*) were identified specifically for ESS1 while master regulators written in dark red and surrounded by dark red boxes (*Slc22a1* and *Sox11*) were identified at both time points (corresponding networks for eggshell strength at Time Point 2 (ESS2) in Figure 6).



**Figure 6.** Scheme of gene regulatory pathways revealing the top five master regulators (pink filled boxes) for eggshell strength at Time Point 2 (ESS2) following the “upstream analysis” [42]. The master regulators written in dark blue and surrounded by dark blue boxes (*Prox1*, *St8sia2* and *Tead2*) were identified specifically for ESS2 while master regulators written in dark red and surrounded by dark red boxes (*Slc22a1* and *Sox11*) were identified at both time points (corresponding networks for eggshell strength at Time Point 1 (ESS1) in Figure 5).

Another interesting master regulator, *Sox11*, which encodes a member of the Sox (SRY-related HMG-box) family of transcription factors, was found at both time points. *Sox11* is known to positively regulate the process of osteogenesis (the formation of bone) [56]. This regulator gains relevance given the importance of bone as a labile reservoir of minerals, especially calcium [4]. In birds, the calcium homeostasis is achieved by regulating the metabolism of bone minerals as well as by controlling the absorption and excretion of calcium in the intestine and in kidneys, respectively [57]. Furthermore, the master regulator *Tead2* found for ESS2 is a regulator of osteogenesis [58] and it is also one of the direct downstream target genes of *Sox11*. This might be an indication of different regulatory mechanisms involved in the osteogenesis or bone remodeling during the later stages of the laying cycle [56].

The *St8sia2*, identified as an ESS2 specific master regulator, encodes a membrane protein which catalyzes the metabolism of sialic acid [59], a carbohydrate found in the eggshell membranes [60–62]. The eggshell membranes constitute the inner layer of the eggshell and contribute to its strength. They further provide the nucleation sites for the initiation of the shell synthesis [63]. Sialic acid is also part of podocalyxin and secreted phosphoprotein 1 (SPP1), both of which are glycoproteins found in the uterus during eggshell calcification [5,64]. Because of its high negative charge, podocalyxin is presumed to interact with calcium carbonate during the calcification of the eggshell [64]. The master regulator *Prox1* encodes the protein prospero homeobox 1 that has also been reported as part of eggshell membranes [65,66]. However, the *Prox1* gene is mostly implicated in the regulation of the development of a variety of organs including liver, pancreas and kidney [67]. Although the vast majority of the master regulators could be biologically characterized to be crucial for ESS, the importance and role of the two master regulators *Sox1* and *Arx* for this trait is currently biologically unconfirmed and could hence provide novel targets for future studies.

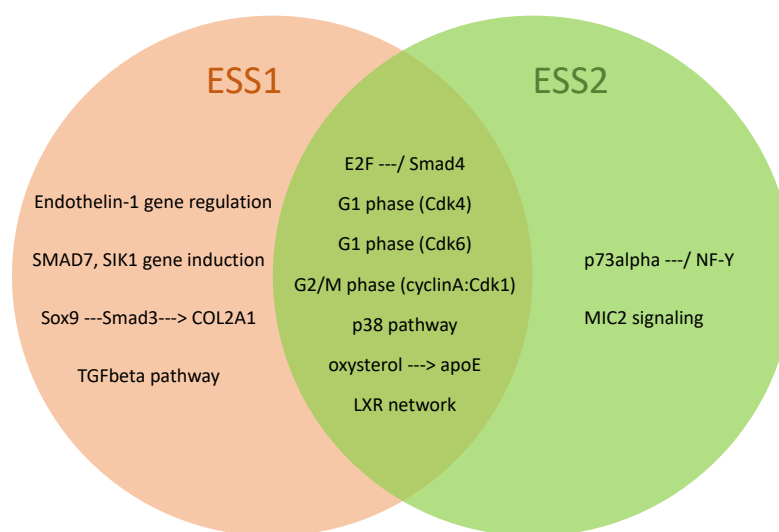
### 3.3. Identification of Over-Represented Pathways

To further elucidate and investigate the mechanisms that control the ESS at different time points, we were interested in identifying age-specific and common over-represented pathways. Applying the pathway analysis, we identified eleven and nine significantly over-represented pathways for ESS1 and ESS2, respectively, and seven of these pathways are overlapping for both time points (see Figure 7 and Table 3).

Among the pathways shared by both time points, G1 phase (Cdk4), G1 phase (Cdk6), and G2/M phase (cyclinA:Cdk1) involve different members of the cyclin-dependent kinase (CDK) family which regulate transcription, mRNA processing, and, more importantly, cell cycle [68]. In the context of ESS, these pathways may influence the differentiation efficiency of osteoblasts, osteoclasts, chondrocytes [69], and uterine epithelium cells, all of which are crucial for the supply of calcium ions as well as for bone and calcium homeostasis [70,71]. The p38 pathway is implicated in a variety of cellular responses including those related to proliferation, differentiation and apoptosis [72]. Moreover, the role of this pathway has also been reported in the egg development of *Drosophila melanogaster* [73]. The LXR (liver X receptors) network plays a central role in the transcriptional control of lipid metabolism [74]. This pathway also mediates the concentrations of oxysterols and ApoE (Apolipoprotein E) if activated in response to elevated intra-cellular cholesterol levels [75]. The oxysterols, oxygenated forms of cholesterol, are intermediates in bile acid and steroid hormone biosynthetic pathways [76]. Among other steroid hormones, estrogen is more intimately involved in calcium homeostasis and has also been implicated in the development of osteoporosis [77]. Moreover, other forms of oxysterols are also involved in calcium metabolisms [78] and mesenchymal stem cell differentiation [79]. In addition to the CDKs, the Smad4 proteins, predominantly present in the nucleus of the cell, mediate the cell cycle due to their association with the E2F family of transcription factors [80]. These pathways can be upstream regulated by the transforming growth factor  $\beta$  (TGF- $\beta$ ) [81].

The transforming growth factor- $\beta$  (TGF- $\beta$ ) signaling pathway can be regarded as the most important pathway enriched for ESS1. This pathway, among its other functions, is well-known for its role in bone homeostasis [82]. Furthermore, some components of this pathway also overlap with other pathways delineated in our analysis. The Sox9 is a transcription factor that regulates the expression of the COL2A1 (collagen type II, alpha 1) gene which contributes to collagen formation [83]. During this process, Smad3, a member of effector molecules in the signaling pathways of the TGF- $\beta$  ligand superfamily is activated [84]. Another pathway that is based on Smad7, SIK1 gene induction also regulates TGF- $\beta$  signaling [85]. Owing to this crosstalk with a variety of other pathways, the TGF- $\beta$  signaling pathway allows the bone to adapt to dynamic environments [82].

The Endothelin-1 gene (ET-1) regulation pathway includes the mechanisms regulating ET-1 gene expression. Among other functions, ET-1 is involved in osteoblast proliferation and differentiation in bone tissue as well as in the ovulation process in the uterus [86]. ET-1 gene regulation is responsive to intracellular calcium and calmodulin [87]. The MIC2 signaling pathway, which was specifically enriched for ESS2, has CD99 as the main cell surface protein and has been implicated in apoptosis, adhesion, differentiation, and protein trafficking possibly by affecting actin cytoskeleton reorganization [88,89]. Another ESS2 specific pathway involves the inactivation of the nuclear factor Y (NF-Y) transcription factor by p73 proteins, a process that represses the promoter of the telomerase catalytic subunit and induces replicative senescence [90,91]. The activity of NF-Y is further linked to the parathyroid hormone, which is the main regulator of calcium and phosphorus homeostasis. Taken together, the pathways show a diversity of complex functional features in chicken in response to age-dependent changes in eggshell formation. Some pathways show a direct relevance for ESS while others seem to be indirectly linked via interactions between pathways and regulators [92,93].



**Figure 7.** Venn diagram of over-represented pathways ( $p$  adjusted  $< 0.001$ ) of eggshell strength at Time Point 1 (ESS1), at Time Point 2 (ESS2), and their overlap. Pathways are based on the TRANSPATH pathway database [44].

**Table 3.** Significantly over-represented pathways for both time points ( $p$  adjusted < 0.001) sorted by adjusted  $p$ -values (based on the smaller one of either ESS1 or ESS2). Pathways are based on the TRANSPATH pathway database [44]. (ESS1/ESS2, eggshell strength at Time Point 1/2).

Pathway Name	Adjusted $p$ -Value for ESS1 / ESS2	Over-Represented in
E2F —/ Smad4	$5.05 \times 10^{-5} / 7.99 \times 10^{-4}$	ESS1, ESS2
Endothelin-1 gene regulation	$5.05 \times 10^{-5} / -$	ESS1
G2/M phase (cyclin A:Cdk1)	$1.61 \times 10^{-4} / 1.65 \times 10^{-4}$	ESS1, ESS2
SMAD7, SIK1 gene induction	$1.61 \times 10^{-4} / -$	ESS1
oxysterol —>apoE	$1.61 \times 10^{-4} / 1.85 \times 10^{-4}$	ESS1, ESS2
LXR network	$1.61 \times 10^{-4} / 1.65 \times 10^{-4}$	ESS1, ESS2
p73alpha —/ NF-Y	$- / 1.65 \times 10^{-4}$	ESS2
Sox9 —Smad3—>COL2A1	$5.43 \times 10^{-4} / -$	ESS1
G1 phase (Cdk6)	$7.60 \times 10^{-4} / 7.93 \times 10^{-4}$	ESS1, ESS2
G1 phase (Cdk4)	$9.77 \times 10^{-4} / 7.99 \times 10^{-4}$	ESS1, ESS2
p38 pathway	$9.77 \times 10^{-4} / 7.99 \times 10^{-4}$	ESS1, ESS2
MIC2 signaling	$- / 7.99 \times 10^{-4}$	ESS2
TGFbeta pathway	$9.53 \times 10^{-4} / -$	ESS1

#### 4. Discussion

To uncover the associations between genetic variants and phenotypes, genome wide association studies (GWAS) have become the method of choice [12]. Despite their success in identifying a multitude of genes, the prediction performance of single-SNP based GWAS strategies is limited [15,17,94]. Alternatively, multi-marker models including different Bayesian frameworks were introduced for GWAS. In these models, all SNPs are fitted simultaneously as random effects assuming a certain prior distribution of SNP effects [13]. In practice, these SNP effects are unknown and may not even strictly follow a certain distribution [25]. Unlike these traditional statistical models, machine learning methods do not require these prior assumptions about the genetic architecture of traits and have been applied in GWAS in humans [30] as well as in livestock [27,95]. Especially, Romagnoni et al. [30] and Huang et al. [24] showed that machine learning based algorithms provide promising prediction power to assess genotype–phenotype associations. In particular, the Random Forests (RF) algorithm has been successfully applied for this purpose. These articles encouraged us to utilize RF in our study since the application of GWAS to identify genetic variants associated with ESS was futile.

Applying RF, we were able to identify a remarkably high number of genes related to ESS which is in agreement with the findings of Maan et al. [6,7], Mikšík et al. [96,97], and Brionne et al. [5], who pointed out a large number of genes/proteins involved in ESS due to the complexity of this trait. The large difference in the number of genes identified for ESS1 and ESS2 reflects the change in the genetic and environmental components of the phenotypic variance over age, as has been reported before for complex traits [98,99]. The overlap between the genes for both time points (see Figure 2) reflects that certain molecular functions remain relevant to eggshell development during the laying cycle of chicken. Particularly, the similarity of genes responsible for the transportation of ions is in line with the findings of Park et al. [100] and Fan et al. [51] who found that the concentration level of different ions in blood does not change with the age of chicken. Interestingly, a closer look at the biological processes of these traits reveals that, while highly significant GO terms are involved in development for ESS1, the significant biological processes for ESS2 are rather related to different metabolic processes (Table S3). The differences in biological processes at both time points could be associated with the temporal changes in the signaling cascades influencing dynamic behavior of eggshell strength over time.

In line with previous studies [39,42,101–103], we applied a systems biology approach and identified master regulators to investigate and unravel the transcriptional regulatory machinery of ESS associated genes. Interestingly, our results show that, similar to the genes, there are common master

regulators (*Sox11* and *Slc22a1*) for both time points, which are likely to govern various eggshell related processes during the laying of the birds. In particular, being a member of the *Slc* superfamily which is involved in the transmembrane transport, the *Slc22a1* could be essential to eggshell development. For ESS1, the most promising master regulator *Scn11a* controls sodium transport in the uterus [49,50] to maintain a voltage difference as well as osmolarity across uterine cell membranes to help in the calcium transportation [51]. In ESS2, the master regulator *Tead2* together with the master regulator *Sox11* underline the importance of bone remodeling during the later stages of the production cycle of the chicken.

Another fundamental step of our analysis was the identification of the over-represented pathways. The results of this analysis also reinforce the findings of gene set analysis as well as the identified master regulators. Some of the over-represented pathways were conserved at both time points while others were age-specific. Here, we specifically highlight the well-characterized TGF- $\beta$  pathway that interacts with most of the identified pathways in our analysis to regulate bone homeostasis and thus might play an important role in ESS [82]. The majority of the remaining pathways, especially those which are common to both time points, were found to be related to the cell cycle. The uterine epithelium and bone are the tissues that actively take part in the development of eggshell, hence the renewal of the cells of both tissues is crucial for the synthesis of a strong eggshell [4]. Furthermore, multiple studies suggest that a declining ability of uterine epithelium cells to transport calcium is the main reason of the age-related deterioration of eggshells [51,100]. In particular, the ESS2 specific p73alpha—/ NF-Y pathway that results in the inactivation of the NF-Y transcription factor by p73 proteins and consequently causes replicative senescence of cells [90] may also point towards the underlying reason for weaker eggshells during the later stages of the production cycle.

Recently, the use of systems biology based approaches to study the traits of economic importance is gaining importance in the field of agriculture [39,102–104]. However, one of the major impediments in the use of this knowledge in practical animal breeding is to integrate this large amount of information into traditional genetic evaluation programs [105]. A small group of master regulators such as those identified in our analysis integrated into prediction models can possibly be a remedy and might provide novel breeding targets to improve the economically important trait of ESS. Additionally, the knowledge about the specific pathways such as TGF- $\beta$  could provide novel hypotheses for further studies.

## 5. Conclusions

In this study, we performed a systematic analysis to investigate the age-specific and common regulatory mechanisms that underlie the dynamic trait eggshell strength in chicken. For this purpose, we applied a RF feature selection algorithm to detect the age-dependent genotype–phenotype associations and then used a well established systems biology approach to highlight the master regulators and regulatory pathways that govern the underlying genetic mechanisms of eggshell development. Our results show that most of the genes identified for the ESS at both time points are in agreement with previous studies. Our findings further indicate that some biological processes related to eggshell development remain conserved across production stages while others are age-specific and thus changing over time. To the best of our knowledge, this is the first study revealing master regulators and over-represented pathways in the context of ESS and our findings should be further utilized to design novel hypothesis for future studies.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2073-4425/11/4/464/s1>, Script S1: R-script for analysis of SNPs and for the extraction of corresponding genes, Table S1: The list of important SNPs, Table S2: The list of genes, Table S3: Gene Ontology categories

**Author Contributions:** M.G. designed and supervised the research. F.R. and S.K. participated in the design of the study, and conducted computational and statistical analyses together with M.G. F.R. prepared and studied the GWAS data. S.K. performed and adjusted the bioinformatics analysis. A.O.S. was involved in the interpretation of the results together with F.R., S.K., D.C., and M.G. F.R., S.K., 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.

**Acknowledgments:** The chicken data used in this study were provided by the “Synbreed—Synergistic Plant and Animal Breeding” project for which we are grateful to the project team. This work is part of FR’s doctoral program, which is funded by the overseas scholarship program of the University of Agriculture Faisalabad, Pakistan. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University. We would like to thank Abirami Rajavel and Martin Wutke for proofreading the manuscript and Malena Erbe for providing important insights into the chicken dataset.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Bain, M.; Nys, Y.; Dunn, I. Increasing persistency in lay and stabilising egg quality in longer laying cycles. What are the challenges? *Br. Poult. Sci.* **2016**, *57*, 330–338. [[CrossRef](#)]
- Pottgüter, R. Feeding laying hens to 100 weeks of age. *Lohmann Inf.* **2016**, *50*, 18–21.
- Chien, Y.C.; Hincke, M.; McKee, M. Ultrastructure of avian eggshell during resorption following egg fertilization. *J. Struct. Biol.* **2009**, *168*, 527–538. [[CrossRef](#)] [[PubMed](#)]
- 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: Cambridge, UK, 2011.
- Brienne, 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*, 1–17. [[CrossRef](#)] [[PubMed](#)]
- Mann, K.; Maček, B.; Olsen, J.V. Proteomic analysis of the acid-soluble organic matrix of the chicken calcified eggshell layer. *Proteomics* **2006**, *6*, 3801–3810. [[CrossRef](#)]
- Mann, K.; Olsen, J.V.; Maček, B.; Gnad, F.; Mann, M. Phosphoproteins of the chicken eggshell calcified layer. *Proteomics* **2007**, *7*, 106–115. [[CrossRef](#)]
- Yin, Z.; Lian, L.; Zhu, F.; Zhang, Z.H.; Hincke, M.; Yang, N.; Hou, Z.C. The transcriptome landscapes of ovary and three oviduct segments during chicken (*Gallus gallus*) egg formation. *Genomics* **2020**, *112*, 243–251. [[CrossRef](#)]
- Crosara, F.S.G.; Pereira, V.J.; Lellis, C.G.; Barra, K.C.; Santos, S.K.A.D.; Souza, L.C.G.M.D.; Morais, T.A.D.; Litz, F.; Limão, V.A.; Braga, P.F.S.; et al. Is the Eggshell Quality Influenced by the Egg Weight or the Breeder Age? *Braz. J. Poult. Sci.* **2019**, *21*. [[CrossRef](#)] confirmed
- Sun, L.; Wu, R. Mapping complex traits as a dynamic system. *Phys. Life Rev.* **2015**, *13*, 155–185. [[CrossRef](#)]
- Zhang, H.; Wang, Z.; Wang, S.; Li, H. Progress of genome wide association study in domestic animals. *J. Anim. Sci. Biotechnol.* **2012**, *3*, 26. [[CrossRef](#)]
- Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [[CrossRef](#)]
- Schmid, M.; Bennewitz, J. Invited review: Genome-wide association analysis for quantitative traits in livestock—a selective review of statistical models and experimental designs. *Arch. Fuer Tierz.* **2017**, *60*, 335. [[CrossRef](#)]
- 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 Genom.* **2010**, *11*, 724. [[CrossRef](#)]
- Bush, W.S.; Moore, J.H. Genome-wide association studies. *PLoS Comput. Biol.* **2012**, *8*, e1002822. [[CrossRef](#)] [[PubMed](#)] Confirmed
- Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* **2013**, *9*, 29. [[CrossRef](#)] [[PubMed](#)]
- Holland, D.; Fan, C.C.; Frei, O.; Shadrin, A.A.; Smeland, O.B.; Sundar, V.; Andreassen, O.A.; Dale, A.M. Estimating inflation in GWAS summary statistics due to variance distortion from cryptic relatedness. *BioRxiv* **2017**, 164939, doi:10.1101/164939.
- Kang, H.M.; Sul, J.H.; Service, S.K.; Zaitlen, N.A.; Kong, S.Y.; Freimer, N.B.; Sabatti, C.; Eskin, E.; et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **2010**, *42*, 348. [[CrossRef](#)] [[PubMed](#)]
- Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [[CrossRef](#)] [[PubMed](#)]
- Young, A.I. Solving the missing heritability problem. *PLoS Genet.* **2019**, *15*, e1008222. [[CrossRef](#)]



21. Zhao, Y.; Chen, F.; Zhai, R.; Lin, X.; Wang, Z.; Su, L.; Christiani, D.C. Correction for population stratification in random forest analysis. *Int. J. Epidemiol.* **2012**, *41*, 1798–1806. [[CrossRef](#)]
22. Nguyen, T.T.; Huang, J.Z.; Wu, Q.; Nguyen, T.T.; Li, M.J. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genom.* **2015**, *16*, S5. [[CrossRef](#)]
23. Armero, C.; Cabras, S.; Castellanos, M.E.; Quirós, A. Two-Stage Bayesian Approach for GWAS With Known Genealogy. *J. Comput. Graph. Stat.* **2019**, *28*, 197–204. [[CrossRef](#)]
24. Huang, X.; Zhou, W.; Bellis, E.S.; Stubblefield, J.; Causey, J.; Qualls, J.; Walker, K. Minor QTLs mining through the combination of GWAS and machine learning feature selection. *BioRxiv* **2019**, 712190, doi:10.1101/712190. [[CrossRef](#)]
25. Liu, Y.; Wang, D.; He, F.; Wang, J.; Joshi, T.; Xu, D. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* **2019**, *10*, 1091. [[CrossRef](#)] [[PubMed](#)]
26. Chen, C.C.; Schwender, H.; Keith, J.; Nunkesser, R.; Mengersen, K.; Macrossan, P. Methods for identifying SNP interactions: A review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 1580–1591. [[CrossRef](#)] [[PubMed](#)]
27. van der Heide, E.; Veerkamp, R.; van Pelt, M.; Kamphuis, C.; Athanasiadis, I.; Ducro, B. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. *J. Dairy Sci.* **2019**, *102*, 9409–9421. [[CrossRef](#)] [[PubMed](#)]
28. Briec, M.S.; Waters, C.D.; Drinan, D.P.; Naish, K.A. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol. Ecol. Resour.* **2018**, *18*, 755–766. [[CrossRef](#)]
29. Nguyen, T.; Le, L. Detection of SNP-SNP Interactions in Genome-wide Association Data Using Random Forests and Association Rules. In Proceedings of the 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Phnom Penh, Cambodia, 3–5 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7. Confirmed
30. Romagnoni, A.; Jégou, S.; Van Steen, K.; Wainrib, G.; Hugot, J.P. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Sci. Rep.* **2019**, *9*, 1–18. [[CrossRef](#)]
31. Erbe, M.; Caverio, D.; Weigend, A.; Weigend, S.; Pausch, H.; Preisinger, R.; Simianer, H. Genomic prediction in laying hens. In Proceedings of the 8th European Symposium on Poultry Genetics, Venice, Italy, 25–27 September 2013.
32. Ni, G.; Strom, T.M.; Pausch, H.; Reimer, C.; Preisinger, R.; Simianer, H.; Erbe, M. Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. *BMC Genom.* **2015**, *16*, 824. [[CrossRef](#)]
33. Ni, G.; Caverio, D.; Fangmann, A.; Erbe, M.; Simianer, H. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Sel. Evol.* **2017**, *49*, 8. [[CrossRef](#)]
34. Garrick, D.J.; Taylor, J.F.; Fernando, R.L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **2009**, *41*, 55. [[CrossRef](#)]
35. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [[CrossRef](#)]
36. Kursa, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—a system for feature selection. *Fundam. Informaticae* **2010**, *101*, 271–285. [[CrossRef](#)]
37. Kursa, M.B.; Rudnicki, W.R. The all relevant feature selection using random forest. *arXiv* **2011**, arXiv:1106.5112.
38. 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*, doi:10.1093/database/bar030. [[CrossRef](#)]
39. Ayalew, Y.; Gültas, M.; Effa, K.; Hanotte, O.H.; Schmitt, A. Identification of candidate signature genes and key regulators associated with Trypanotolerance in the Sheko Breed. *Front. Genet.* **2019**, *10*, 1095.
40. Wlochowitz, D.; Haubrock, M.; Arackal, J.; Bleckmann, A.; Wolff, A.; Beißbarth, T.; Wingender, E.; Gültas, M. Computational identification of key regulators in two different colorectal cancer cell lines. *Front. Genet.* **2016**, *7*, 42. [[CrossRef](#)]
41. Wingender, E.; Kel, A. geneXplain—eine integrierte Bioinformatik-Plattform. *BIOspektrum* **2012**, *18*, 554–556. [[CrossRef](#)]

42. Koschmann, J.; Bhar, A.; Stegmaier, P.; Kel, A.; Wingender, E. “Upstream analysis”: An integrated promoter-pathway analysis approach to causal interpretation of microarray data. *Microarrays* **2015**, *4*, 270–286. [[CrossRef](#)]
43. Rzhetsky, A.; Iossifov, I.; Koike, T.; Krauthammer, M.; Kra, P.; Morris, M.; Yu, H.; Duboué, P.A.; Weng, W.; Wilbur, W.J.; et al. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* **2004**, *37*, 43–53. [[CrossRef](#)]
44. Krull, M.; Pistor, S.; Voss, N.; Kel, A.; Reuter, I.; Kronenberg, D.; Michael, H.; Schwarzer, K.; Potapov, A.; Choi, C.; et al. TRANSPATH®: An information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **2006**, *34*, D546–D551. [[CrossRef](#)]
45. Jonchère, V.; Brionne, A.; Gautron, J.; Nys, Y. Identification of uterine ion transporters for mineralisation precursors of the avian eggshell. *BMC Physiol.* **2012**, *12*, 10–51. [[CrossRef](#)]
46. Chakraborti, S.; Dhalla, N.S. *Regulation of Membrane Na<sup>+</sup>-K<sup>+</sup> ATPase*; Springer: Heidelberg, Germany, 2016.
47. Colbran, R.J. Targeting of calcium/calmodulin-dependent protein kinase II. *Biochem. J.* **2004**, *378*, 1–16. [[CrossRef](#)] [[PubMed](#)]
48. Meyer, M.B.; Watanuki, M.; Kim, S.; Shevde, N.K.; Pike, J.W. The human transient receptor potential vanilloid type 6 distal promoter contains multiple vitamin D receptor binding sites that mediate activation by 1, 25-dihydroxyvitamin D<sub>3</sub> in intestinal cells. *Mol. Endocrinol.* **2006**, *20*, 1447–1461. [[CrossRef](#)] [[PubMed](#)]
49. Ogata, K.; Jeong, S.Y.; Murakami, H.; Hashida, H.; Suzuki, T.; Masuda, N.; Hirai, M.; Isahara, K.; Uchiyama, Y.; Goto, J.; et al. Cloning and expression study of the mouse tetrodotoxin-resistant voltage-gated sodium channel  $\alpha$  subunit NaT/Scn11a. *Biochem. Biophys. Res. Commun.* **2000**, *267*, 271–277. [[CrossRef](#)] [[PubMed](#)]
50. Seda, M.; Pinto, F.M.; Wray, S.; Cintado, C.G.; Noheda, P.; Buschmann, H.; Candenias, L. Functional and molecular characterization of voltage-gated sodium channels in uteri from nonpregnant rats. *Biol. Reprod.* **2007**, *77*, 855–863. [[CrossRef](#)] [[PubMed](#)]
51. Fan, Y.F.; Hou, Z.C.; Yi, G.Q.; Xu, G.Y.; Yang, N. The sodium channel gene family is specifically expressed in hen uterus and associated with eggshell quality traits. *BMC Genet.* **2013**, *14*, 90. [[CrossRef](#)]
52. Koepsell, H. The SLC22 family with transporters of organic cations, anions and zwitterions. *Mol. Asp. Med.* **2013**, *34*, 413–435. [[CrossRef](#)]
53. Chowdhury, S.; Smith, T. Dietary interaction of 1, 4-diaminobutane (putrescine) and calcium on eggshell quality and performance in laying hens. *Poult. Sci.* **2002**, *81*, 84–91. [[CrossRef](#)]
54. Shinki, T.; Tanaka, H.; Takito, J.; Yamaguchi, A.; Nakamura, Y.; Yoshiki, S.; Suda, T. Putrescine is involved in the vitamin D action in chick intestine. *Gastroenterology* **1991**, *100*, 113–122. [[CrossRef](#)]
55. Altimimi, H.F.; Schnetkamp, P.P. Na<sup>+</sup>/Ca<sup>2+</sup>-K<sup>+</sup> exchangers (NCKX): Functional properties and physiological roles. *Channels* **2007**, *1*, 62–69. [[CrossRef](#)] [[PubMed](#)]
56. Gadi, J.; Jung, S.H.; Lee, M.J.; Jami, A.; Ruthala, K.; Kim, K.M.; Cho, N.H.; Jung, H.S.; Kim, C.H.; Lim, S.K. The transcription factor protein Sox11 enhances early osteoblast differentiation by facilitating proliferation and the survival of mesenchymal and osteoblast progenitors. *J. Biol. Chem.* **2013**, *288*, 25400–25413. [[CrossRef](#)] [[PubMed](#)]
57. ELAROUSSI, M.A.; FORTE, L.R.; EBER, S.L.; BIELLIER, H.V. Calcium Homeostasis in the Laying Hen.: 1. Age and Dietary Calcium Effects. *Poult. Sci.* **1994**, *73*, 1581–1589. [[CrossRef](#)] [[PubMed](#)]
58. Håkelién, A.M.; Bryne, J.C.; Harstad, K.G.; Lorenz, S.; Paulsen, J.; Sun, J.; Mikkelsen, T.S.; Myklebost, O.; Meza-Zepeda, L.A. The regulatory landscape of osteogenic differentiation. *Stem Cells* **2014**, *32*, 2780–2793. [[CrossRef](#)] [[PubMed](#)]
59. Scheidegger, E.P.; Sternberg, L.R.; Roth, J.; Lowe, J.B. A human STX cDNA confers polysialic acid expression in mammalian cells. *J. Biol. Chem.* **1995**, *270*, 22685–22688. [[CrossRef](#)] [[PubMed](#)]
60. Itoh, T.; Munakata, K.; Adachi, S.; Hatta, H.; Nakamura, T.; Kato, T.; Kim, M. Chalaza and egg yolk membrane as excellent sources of sialic acid (N-acetylneuraminic acid) for an industrial-scale preparation. *Jpn. J. Zootech. Sci.* **1990**, *61*, 277–282.
61. Nakano, K.; Nakano, T.; Ahn, D.; Sim, J. Sialic acid contents in chicken eggs and tissues. *Can. J. Anim. Sci.* **1994**, *74*, 601–606. [[CrossRef](#)]
62. Nakano, T.; Ikawa, N.; Ozimek, L. Chemical composition of chicken eggshell and shell membranes. *Poult. Sci.* **2003**, *82*, 510–514. [[CrossRef](#)]

63. Du, J.; Hincke, M.T.; Rose-Martel, M.; Hennequet-Antier, C.; Brionne, A.; Cogburn, L.A.; Nys, Y.; Gautron, J. Identifying specific proteins involved in eggshell membrane formation using gene expression analysis and bioinformatics. *BMC Genom.* **2015**, *16*, 792. [[CrossRef](#)]
64. 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](#)]
65. Ahmed, T.A.; Suso, H.P.; Hincke, M.T. Experimental datasets on processed eggshell membrane powder for wound healing. *Data Brief* **2019**, *26*, 104457. [[CrossRef](#)] [[PubMed](#)]
66. Ahmed, T.A.; Suso, H.P.; Hincke, M.T. In-depth comparative analysis of the chicken eggshell membrane proteome. *J. Proteom.* **2017**, *155*, 49–62. [[CrossRef](#)] [[PubMed](#)]
67. Kim, Y.m.; Kim, W.Y.; Nam, S.A.; Choi, A.R.; Kim, H.; Kim, Y.K.; Kim, H.S.; Kim, J. Role of PROX1 in the transforming ascending thin limb of Henle's loop during mouse kidney development. *PLoS ONE* **2015**, *10*, doi:10.1371/journal.pone.0127429.
68. Malumbres, M. Cyclin-dependent kinases. *Genome Biol.* **2014**, *15*, 122. [[CrossRef](#)] [[PubMed](#)]
69. Ogasawara, T.; Mori, Y.; Abe, M.; Suenaga, H.; Kawase-Koga, Y.; Saijo, H.; Takato, T. Role of cyclin-dependent kinase (Cdk) 6 in osteoblast, osteoclast, and chondrocyte differentiation and its potential as a target of bone regenerative medicine. *Oral Sci. Int.* **2011**, *8*, 2–6. [[CrossRef](#)]
70. Whitehead, C. Overview of bone biology in the egg-laying hen. *Poult. Sci.* **2004**, *83*, 193–199. [[CrossRef](#)]
71. Bar, A. Calcium transport in strongly calcifying laying birds: Mechanisms and regulation. *Comp. Biochem. Physiol. Part A Mol. Integr. Physiol.* **2009**, *152*, 447–469. [[CrossRef](#)]
72. Ono, K.; Han, J. The p38 signal transduction pathway activation and function. *Cell. Signal.* **2000**, *12*, 1–13. [[CrossRef](#)]
73. Suzanne, M.; Irie, K.; Glise, B.; Agnès, F.; Mori, E.; Matsumoto, K.; Noselli, S. The Drosophila p38 MAPK pathway is required during oogenesis for egg asymmetric development. *Genes Dev.* **1999**, *13*, 1464–1474. [[CrossRef](#)]
74. Zelcer, N.; Tontonoz, P. Liver X receptors as integrators of metabolic and inflammatory signaling. *J. Clin. Investig.* **2006**, *116*, 607–614. [[CrossRef](#)]
75. Vaya, J.; Schipper, H.M. Oxysterols, cholesterol homeostasis, and Alzheimer disease. *J. Neurochem.* **2007**, *102*, 1727–1737. [[CrossRef](#)] [[PubMed](#)]
76. Griffiths, W.J.; Abdel-Khalik, J.; Crick, P.J.; Yutuc, E.; Wang, Y. New methods for analysis of oxysterols and related compounds by LC–MS. *J. Steroid Biochem. Mol. Biol.* **2016**, *162*, 4–26. [[CrossRef](#)] [[PubMed](#)]
77. Beck, M.; Hansen, K. Role of estrogen in avian osteoporosis. *Poult. Sci.* **2004**, *83*, 200–206. [[CrossRef](#)]
78. Mackrill, J.J. Oxysterols and calcium signal transduction. *Chem. Phys. Lipids* **2011**, *164*, 488–495. [[CrossRef](#)] [[PubMed](#)]
79. Kha, H.T.; Basseri, B.; Shouhed, D.; Richardson, J.; Tetradis, S.; Hahn, T.J.; Parhami, F. Oxysterols regulate differentiation of mesenchymal stem cells: pro-bone and anti-fat. *J. Bone Miner. Res.* **2004**, *19*, 830–840. [[CrossRef](#)]
80. Frederick, J.P.; Liberati, N.T.; Waddell, D.S.; Shi, Y.; Wang, X.F. Transforming growth factor  $\beta$ -mediated transcriptional repression of c-myc is dependent on direct binding of Smad3 to a novel repressive Smad binding element. *Mol. Cell. Biol.* **2004**, *24*, 2546–2559. [[CrossRef](#)]
81. Chen, C.R.; Kang, Y.; Siegel, P.M.; Massagué, J. E2F4/5 and p107 as Smad cofactors linking the TGF $\beta$  receptor to c-myc repression. *Cell* **2002**, *110*, 19–32. [[CrossRef](#)]
82. Tang, S.Y.; Alliston, T. Regulation of postnatal bone homeostasis by TGF $\beta$ . *BoneKEy Rep.* **2013**, *2*, doi:10.1038/bonekey.2012.255. [[CrossRef](#)]
83. Bell, D.M.; Leung, K.K.; Wheatley, S.C.; Ng, L.J.; Zhou, S.; Ling, K.W.; Sham, M.H.; Koopman, P.; Tam, P.P.; Cheah, K.S. SOX9 directly regulates the type-II collagen gene. *Nat. Genet.* **1997**, *16*, 174–178. [[CrossRef](#)]
84. Massagué, J.; Chen, Y.G. Controlling TGF- $\beta$  signaling. *Genes Dev.* **2000**, *14*, 627–644.
85. Lönn, P.; Vanlandewijck, M.; Raja, E.; Kowanetz, M.; Watanabe, Y.; Kowanetz, K.; Vasilaki, E.; Heldin, C.H.; Moustakas, A. Transcriptional induction of salt-inducible kinase 1 by transforming growth factor  $\beta$  leads to negative regulation of type I receptor signaling in cooperation with the Smurf2 ubiquitin ligase. *J. Biol. Chem.* **2012**, *287*, 12867–12878. [[CrossRef](#)] [[PubMed](#)]
86. Stow, L.R.; Jacobs, M.E.; Wingo, C.S.; Cain, B.D. Endothelin-1 gene regulation. *FASEB J.* **2011**, *25*, 16–28. [[CrossRef](#)] [[PubMed](#)]




87. Strait, K.A.; Stricklett, P.K.; Kohan, J.L.; Miller, M.B.; Kohan, D.E. Calcium regulation of endothelin-1 synthesis in rat inner medullary collecting duct. *Am. J. Physiol.-Ren. Physiol.* **2007**, *293*, F601–F606. [[CrossRef](#)] [[PubMed](#)]
88. Yoon, S.S.; Jung, K.I.; Choi, Y.L.; Choi, E.Y.; Lee, I.S.; Park, S.H.; Kim, T.J. Engagement of CD99 triggers the exocytic transport of ganglioside GM1 and the reorganization of actin cytoskeleton. *FEBS Lett.* **2003**, *540*, 217–222. [[CrossRef](#)]
89. Pasello, M.; Manara, M.C.; Scotlandi, K. CD99 at the crossroads of physiology and pathology. *J. Cell Commun. Signal.* **2018**, *12*, 55–68. [[CrossRef](#)]
90. Yao, Y.; Bellon, M.; Shelton, S.N.; Nicot, C. Tumor suppressors p53, p63TA $\alpha$ , p63TA $\gamma$ , p73 $\alpha$ , and p73 $\beta$  use distinct pathways to repress telomerase expression. *J. Biol. Chem.* **2012**, *287*, 20737–20747. [[CrossRef](#)]
91. Jung, M.S.; Yun, J.; Chae, H.D.; Kim, J.M.; Kim, S.C.; Choi, T.S.; Shin, D.Y. p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of NF-Y transcription factor. *Oncogene* **2001**, *20*, 5818–5825. [[CrossRef](#)]
92. Alimov, A.P.; Park-Sarge, O.K.; Sarge, K.D.; Malluche, H.H.; Koszewski, N.J. Transactivation of the parathyroid hormone promoter by specificity proteins and the nuclear factor Y complex. *Endocrinology* **2005**, *146*, 3409–3416. [[CrossRef](#)]
93. Jääskeläinen, T.; Huhtakangas, J.; Mäenpää, P. Negative regulation of human parathyroid hormone gene promoter by vitamin D3 through nuclear factor Y. *Biochem. Biophys. Res. Commun.* **2005**, *328*, 831–837. [[CrossRef](#)]
94. Josephs, E.B.; Stinchcombe, J.R.; Wright, S.I. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytol.* **2017**, *214*, 21–33. [[CrossRef](#)]
95. Li, B.; Zhang, N.; Wang, Y.G.; George, A.W.; Reverter, A.; Li, Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* **2018**, *9*, 237. [[CrossRef](#)] [[PubMed](#)]
96. Mikšik, I.; Eckhardt, A.; Sedláková, P.; Mikulíková, K. Proteins of insoluble matrix of avian (*Gallus gallus*) eggshell. *Connect. Tissue Res.* **2007**, *48*, 1–8. [[CrossRef](#)] [[PubMed](#)]
97. Mikšik, I.; Sedláková, P.; Lacinová, K.; Pataridis, S.; Eckhardt, A. Determination of insoluble avian eggshell matrix proteins. *Anal. Bioanal. Chem.* **2010**, *397*, 205–214. [[CrossRef](#)] [[PubMed](#)]
98. He, L.; Sillanpää, M.J.; Silventoinen, K.; Kaprio, J.; Pitkaniemi, J. Estimating modifying effect of age on genetic and environmental variance components in twin models. *Genetics* **2016**, *202*, 1313–1328. [[CrossRef](#)]
99. Elks, C.E.; Den Hoed, M.; Zhao, J.H.; Sharp, S.J.; Wareham, N.J.; Loos, R.J.; Ong, K.K. Variability in the heritability of body mass index: A systematic review and meta-regression. *Front. Endocrinol.* **2012**, *3*, 29. [[CrossRef](#)]
100. Park, J.A.; Sohn, S.H. The Influence of Hen Aging on Eggshell Ultrastructure and Shell Mineral Components. *Korean J. Food Sci. Anim. Resour.* **2018**, *38*, 1080. [[CrossRef](#)]
101. Reimand, J.; Isserlin, R.; Voisin, V.; Kucera, M.; Tannus-Lopes, C.; Rostamianfar, A.; Wadi, L.; Meyer, M.; Wong, J.; Xu, C.; et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **2019**, *14*, 482–517. [[CrossRef](#)]
102. Neupane, M.; Geary, T.W.; Kiser, J.N.; Burns, G.W.; Hansen, P.J.; Spencer, T.E.; Neiberghs, H.L. Loci and pathways associated with uterine capacity for pregnancy and fertility in beef cattle. *PLoS ONE* **2017**, *12*, e0188997. [[CrossRef](#)]
103. Woldesemayat, A.A.; Ntwasa, M. Pathways and Network Based Analysis of Candidate Genes to Reveal Cross-Talk and Specificity in the Sorghum (*Sorghum bicolor* (L.) Moench) Responses to Drought and It's Co-occurring Stresses. *Front. Genet.* **2018**, *9*, 557. [[CrossRef](#)]
104. Ramayo-Caldas, Y.; Renand, G.; Ballester, M.; Saintilan, R.; Rocha, D. Multi-breed and multi-trait co-association analysis of meat tenderness and other meat quality traits in three French beef cattle breeds. *Genet. Sel. Evol.* **2016**, *48*, 37. [[CrossRef](#)]
105. Kadarmideen, H.N.; von Rohr, P.; Janss, L.L. From genetical genomics to systems genetics: Potential applications in quantitative genomics and animal breeding. *Mamm. Genome* **2006**, *17*, 548–564. [[CrossRef](#)] [[PubMed](#)]



**A.2. Combining Random Forests and a Signal Detection  
Method Leads to the Robust Detection of  
Genotype-Phenotype Associations**

Article

# Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations

Faisal Ramzan <sup>1,2</sup>, Mehmet Gültas <sup>1,3</sup>, Hendrik Bertram <sup>1</sup>, David Caverio <sup>4</sup> and Armin Otto Schmitt <sup>1,3,\*</sup>

<sup>1</sup> Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; faisal.ramzan@stud.uni-goettingen.de (F.R.); gueltas@informatik.uni-goettingen.de (M.G.); hendrik.bertram@stud.uni-goettingen.de (H.B.)

<sup>2</sup> Department of Animal Breeding and Genetics, University of Agriculture Faisalabad, 38000 Faisalabad, Pakistan

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

<sup>4</sup> H&N International, 27472 Cuxhaven, Germany; caverio@ltz.de

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

Received: 9 July 2020; Accepted: 3 August 2020; Published: 5 August 2020



**Abstract:** Genome wide association studies (GWAS) are a well established methodology to identify genomic variants and genes that are responsible for traits of interest in all branches of the life sciences. Despite the long time this methodology has had to mature the reliable detection of genotype–phenotype associations is still a challenge for many quantitative traits mainly because of the large number of genomic loci with weak individual effects on the trait under investigation. Thus, it can be hypothesized that many genomic variants that have a small, however real, effect remain unnoticed in many GWAS approaches. Here, we propose a two-step procedure to address this problem. In a first step, cubic splines are fitted to the test statistic values and genomic regions with spline-peaks that are higher than expected by chance are considered as quantitative trait loci (QTL). Then the SNPs in these QTLs are prioritized with respect to the strength of their association with the phenotype using a Random Forests approach. As a case study, we apply our procedure to real data sets and find trustworthy numbers of, partially novel, genomic variants and genes involved in various egg quality traits.

**Keywords:** Random Forests; signal detection; genome wide association studies; boruta; eggshell strength; egg weight

## 1. Introduction

The importance of genotype-phenotype association studies to understand the genetic basis of traits, either qualitative or quantitative, is well established [1]. Single-SNP based models that test individual SNPs for their association with the phenotype in a genome wide association study (GWAS) are widely used in this regard. Although this approach has been quite successful in discovering genes affecting important traits [2], some daunting aspects still persist that reduce its power at best and make it error prone at worst. These inherent features include population stratification or relatedness among the samples, multiple hypothesis testing, and overestimation of SNP effects as pointed out in previous studies [3–6]. Linear mixed model (LMM) based approaches that incorporate the covariance structure across individuals have been found most effective in dealing with both the kinship and the population stratification problem [7–10]. Acknowledging their importance, a series of approaches have

been proposed to implement the LMM in the context of GWAS [11]. Similarly, many multiple testing correction methods with varying strictness have been suggested as possible solutions and some of these have been addressed in References [4,12].

A further challenge in analyzing quantitative traits is to discover loci having moderate to small phenotypic effects. The SNPs present either inside or in the vicinity of these quantitative trait loci (QTLs) display association strengths which are too small to exceed the statistical significance threshold value. Consequently, only a small part of the overall variance is captured in a typical GWAS analysis [13]. Haplotypes can capture the correlation structure of SNPs which is ignored in single-SNP based GWAS approaches. Hence, testing the haplotypes for association looks promising at least in theory. Nevertheless, haplotype based analyses are far from being simple and so far, no clear evidence is available in the literature that the haplotype based tests are more powerful than single-SNP based tests even though this topic has been investigated over the years [14–17]. To address these limitations, multi-SNP GWAS models were introduced that fit all SNPs simultaneously as random effects in the model [18]. Many implementations of multi-SNP models based on Bayesian as well as LMM frameworks have been developed [19]. Numerous studies have also been conducted to show comparative performance of different single-SNPs, haplotype and multiple-SNP models along with their different implementations [11,20–24]. Recently, the growing application of machine learning approaches in different fields of science has incited their use in assessing the genotype-phenotype association as well [25–29]. Multiple studies have confirmed the superiority of machine learning algorithms compared to GWAS approaches by identifying genes having small effects on the phenotype [26,29,30]. Machine learning methods do not require prior assumptions about the distribution of the SNP effects, hence can be used for a wide variety of traits in humans [31], plants [28] and livestock [32,33]. In particular, Random Forests (RF) models have been praised for their ability to analyze a large number of loci simultaneously and to identify promising associations [29,30].

All the above mentioned methodologies have their advantages and challenges. Among other factors, the success of different association methods is heavily influenced by the genetic architectures of the trait of interest [24,34]. Given the complexity underlying the genetics of quantitative traits, it is probably not realistic to assume that any one method can retain its statistical power for different genetic architectures [17,35,36]. Single-SNP based models are still popular [37–41] while the RF based methods are gaining importance [42]. However, an increasing number of scientists are recommending the integration of different association methods in order to improve QTL identification and interpretation [43,44]. In this regard, to bridge the gap between single-SNP and haplotype based analysis, Zhang et al. [45] used a non-parametric spline based technique to integrate multiple single-SNP based test statistics into a single test. Furthermore, Zhang et al. [19] as well as Abed and Belzile [24] suggested the combined usage of single-SNP and multi-SNP methods together for the identification of a robust set of SNPs associated with the complex phenotypes. To combine the advantages of machine learning and parametric GWAS analysis, Nguyen et al. [26], Huang et al. [28] and Schwarz et al. [46] employed a two stage analysis integrating the Random Forests algorithm with single-SNP models. However, the selection of SNPs in one stage and the analysis of the selected SNP in the second step may not account for the hidden structure in the data and can result in inflated SNP effects in the discovery of genotype-phenotype association.

In this study, we propose a framework that mainly focuses on the identification of robust genotype-phenotype association signals by combining the important SNPs obtained in different association analyses. For this purpose, we first perform a signal detection strategy using the test statistic values of single-SNP based GWAS analysis for the detection of QTLs. Second, using a Random Forests based feature selection technique, we assess the relative importance of SNPs regarding their association level with the phenotype. Unlike the previous two stage studies [26,28,46], we finally prioritize the important SNPs within the QTLs to discover the most robust set of markers.

In order to demonstrate the functionality of our framework, we have analysed two different GWAS (genotype-phenotype) datasets in this study. The first dataset contains the eggshell strength

(ESS) measured at two different time points during the productive life of chicken and the second dataset is related to egg weight (EW) in chicken. Our results show that, using our framework, we are able to identify important novel markers/genes which could provide new insights into the genetic architecture of these traits.

## 2. Materials and Methods

### 2.1. Data Sets

In this study, we have analysed two chicken datasets to detect genotype-phenotype associations underlying economically important egg quality traits, namely eggshell strength (ESS) and egg weight (EW).

**Dataset 1:** The first dataset contains eggshell strength recorded at two time points during the lifetime of the birds. We have previously used this dataset in Reference [30] to identify the key regulatory mechanisms governing eggshell strength in chicken. The dataset consists of 892 birds from six generations of a purebred commercial brown layer line genotyped with the Affymetrix Axiom<sup>®</sup> 600 K Chicken Genotyping Array. The corresponding phenotypic data contain de-regressed breeding values of eggshell breaking strength from individual birds at two different stages of production. The eggshell strength was measured at the poles of an egg and represents the force in Newton needed to break the egg. For the first time point, ESS was recorded at the age of 42, 45, and 48 weeks, while for the second time point, recordings were made at the age of 64 and 68 weeks. Average values of the recorded breaking strengths at time point 1 (ESS1) and time point 2 (ESS2) were then used in an animal model for the breeding value estimation. In this analysis, we also used pedigree data consisting of 40,545 individuals from six generations, in total. The estimated breeding values were then de-regressed following Garrick et al. [47] to obtain the pseudo-phenotypes that were then used for the further analysis. To ensure the quality of our data, we filtered the genotypic data to remove the SNPs having minor allele frequency  $\leq 0.01$ , genotyping call rate  $\leq 97\%$  and also those deviating from Hardy–Weinberg equilibrium ( $p$ -value  $< 1 \times 10^{-6}$ ). Birds having a SNP call rate smaller than 95% were also removed. Finally, we had 892 animals and 318,513 SNPs for our analyses.

**Dataset 2:** The second dataset pertains to egg weight recorded in 36 weeks old adult birds. The dataset has been previously analysed to perform GWAS of age dependent egg weights (EW) in chicken [48]. The dataset provides genotypes and phenotypes of 1063 birds belonging to a pure bred line of Rhode Island Red chicken, also genotyped with the Affymetrix Axiom<sup>®</sup> 600 K Chicken Genotyping Array. From the seven age levels analysed in the original study, we re-analysed only EW at 36 weeks of age as the most significant associations were reported for this trait. The genotypic data were filtered for SNP call rates, minor allele frequencies and Hardy Weinberg equilibrium using the same threshold values as given for the first dataset. After filtering, we used 294,705 SNPs and 1036 birds in our analysis.

### 2.2. Analysis Framework

Our proposed analysis framework consists of six phases to detect important SNPs associated with phenotypes under study.

**Phase 1:** Following the study of Liu et al. [48], we perform a GWAS to obtain the association between single-SNPs and the phenotypes. For this analysis, we first applied a principal component analysis (PCA) using the independent SNPs obtained after pruning SNPs using the indep-pair-wise option in PLINK [49] software, with a window size of 25 SNPs, a step of 5 SNPs and a  $r^2$  threshold of 0.2. Then we used the top five of those principal components as covariates in the association model to control for population structure. Next, we performed a GWAS analysis based on the following univariate linear mixed model implemented in the FaST-Lmm v0.2.31 software [50].

$$y = W\alpha + x\beta + u + \epsilon. \quad (1)$$



In Equation (1),  $y$  is the vector of phenotypic values for all individuals;  $W$  is the matrix of covariates;  $\alpha$  is a vector of corresponding effects and the intercept;  $x$  is the vector of genotypes for the SNPs tested;  $\beta$  is the effect size of the marker;  $u$  is a vector of random polygenic effects with a covariance structure as  $u \sim N(0, KV_g)$ , where  $K$  represents the genetic relatedness matrix derived from the SNP markers and  $V_g$  is the polygenic additive variance.  $\epsilon$  is the vector of random residuals with  $\epsilon \sim N(0, IV_e)$ , where  $I$  is the identity matrix and  $V_e$  is the residual variance component. To test the value of  $\beta$  for each SNP against the null hypothesis  $H_0 : \beta = 0$ , the Wald-test ( $F_{\text{Wald}} = \hat{\beta}^2 / \text{Var}(\beta)$ ) was applied. As suggested in Reference [48], the adjusted threshold value was determined using the *simpleM* approach [51] to evaluate the significance of individual SNPs. In Figure 1A we exemplarily show a chromosomal region and its corresponding Wald statistic values.

**Phase 2:** For the elaboration of association signals embedded in the Wald test statistics, we apply a cubic smoothing spline on these values. The cubic smoothing spline is a piece-wise defined cubic function and is based on the same principle as the normal cubic regression. The assumption implicit in this approach is that the individual association values are observed with noise and that these values can be considered as estimations of some underlying function  $g$ . Given the marker positions in the genome ( $x_i$ ) and the corresponding association values ( $y_i$ ), the function  $g$  is estimated by minimizing the following expression

$$S(f) = \sum \{y_i - g(x_i)\}^2 + \lambda \int g''(x)^2 dx. \quad (2)$$

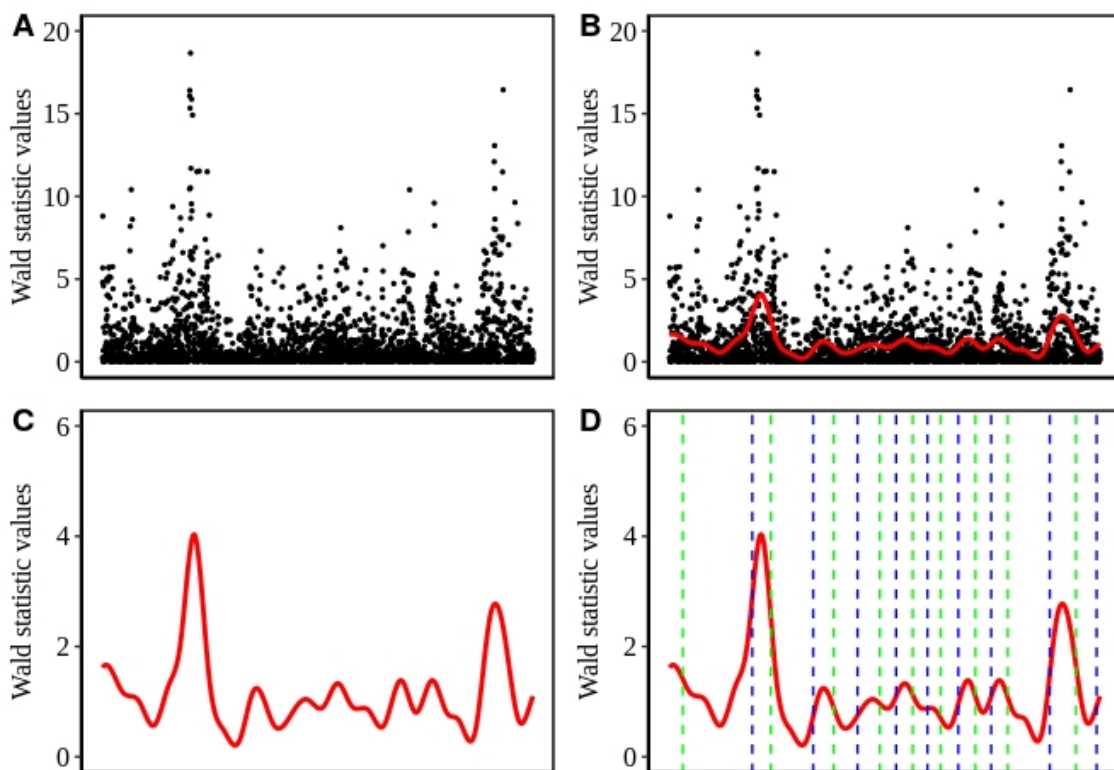
In Equation (2), the first part of right hand side represents the residual sum of squares with the cubic spline function  $g(x_i)$  being the estimated value of the function  $g$  corresponding to SNP  $i$  at chromosomal position  $x_i$ . The integral represents a roughness penalty controlled by the tunable parameter  $\lambda$  whose value is determined by cross validation. The penalty controls the trade off between the conflicting goals of matching the given data and producing a smooth curve [52].  $g''$  represents the second derivative of the sought function with respect to  $x$ . The assumption of  $g$  being continuous and twice differentiable leads to its approximability via a cubic smoothing spline [53]. Thus, a continuous and smooth curve, suitable for the elaboration of the association signals in the test statistic values is obtained. In Figure 1B,C, we exemplarily show the application of cubic smoothing spline over the Wald statistics in a small chromosomal region.

**Phase 3:** For delineation of the obtained association signals in the form of peaks, we determined the inflection points based on the smoothed values. As the smoothing curve represents a function  $g(x)$ , the inflection points indicate the positions, where  $g''(x) = 0$  and thus the curve changes its curvature. Hence, the region between two consecutive inflection points having a downward concave form is regarded as a peak. To this end, the maximum value within a peak is recorded as the height of the peak. In Figure 1D, we exemplarily show the identified peak regions based on the inflection points.

**Phase 4:** In order to separate the peak regions having association signals higher than those arose by chance, we have created a null distribution by permutating the phenotypic data. For the construction of the null distribution, Phases 1, 2 and 3 have been applied to each permuted dataset and the maximum peak values were recorded. In our analysis, we permuted the dataset 1000 times. In the real dataset, we defined a peak region as a QTL if the corresponding peak height exceeds the 95th percentile of the null distribution.

**Phase 5:** Adopting the strategy from our previous study [30], the Random Forests (RF) algorithm was used to estimate the relative importance of each SNP (attribute) for the prediction of the response variable (phenotype). For this purpose, we applied the Boruta algorithm [54] which is a powerful wrapper for the RF based feature selection approach to assess the importance of SNPs. Consequently, we obtained a decision for each SNP whether the importance of the SNP is confirmed, rejected or tentative. In our analysis we only considered SNPs with confirmed importance.

**Phase 6:** Finally, to prioritize the SNPs which are in the QTLs detected in Phase 4, we use the important SNPs from Phase 5 and define the SNPs discovered in both Phases as robust SNPs in our analysis.



**Figure 1.** Step by step representation of the peak detection method. (A) Distribution of the test statistic values along the length of a chromosome segment. (B) The red line indicates the cubic spline fitted on the test statistic values represented by the black dots. (C) The same cubic spline curve as in B without points, y-axis rescaled (D) Dashed lines represent the inflection points of the curve. A pair of a left (blue) and a right (right) inflection point constitute a peak.

### 2.3. Extraction of the Candidate Genes

We scan the genome to identify the genes corresponding to the robust SNPs using BioMart [55]. Only those genes were considered to have some association with the phenotype that were harboring at least one of the robust SNPs within its boundaries. The R-script used for this analysis is provided in Supplementary File S1.

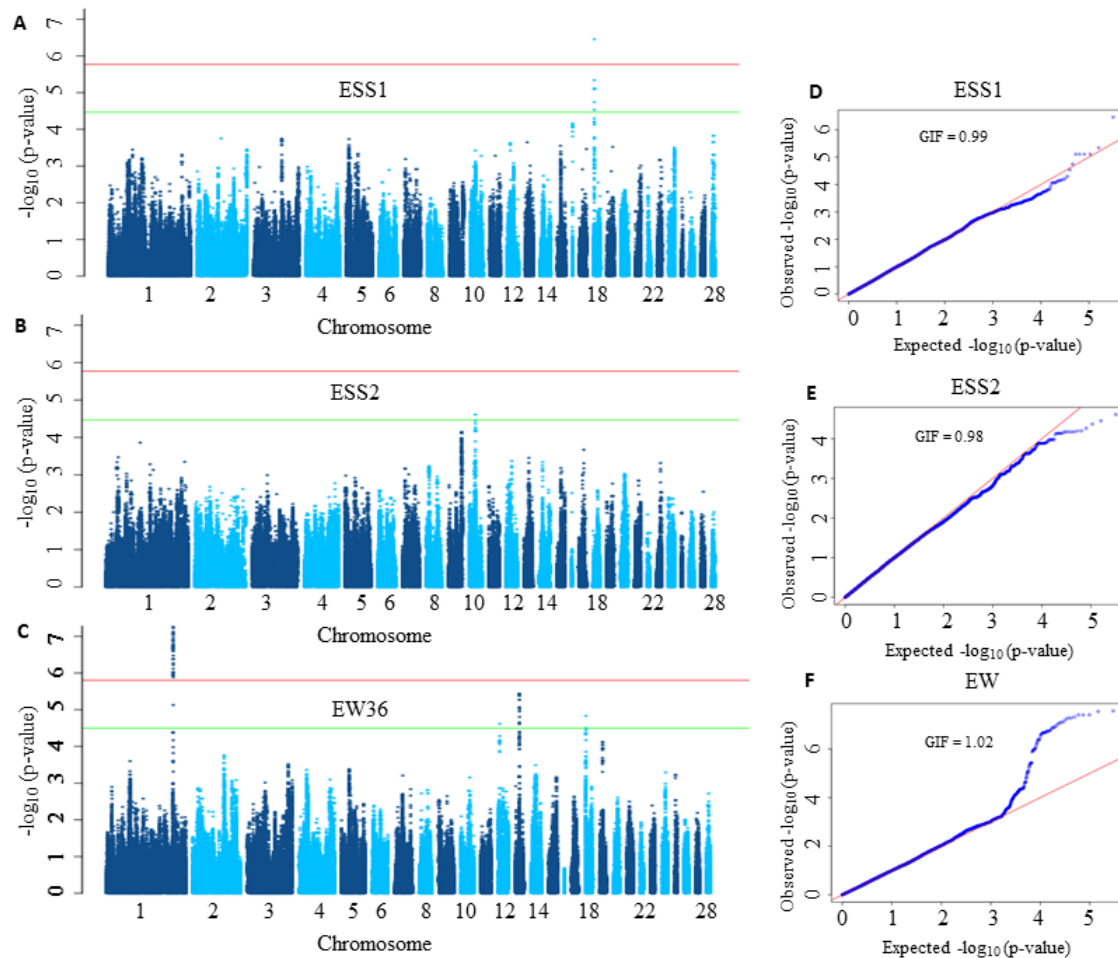
## 3. Results

In our study, we suggest an analysis framework to improve the power of commonly implemented GWAS. The overall framework comprises the following steps. First, a linear mixed model (LMM) based single-SNP GWAS is performed to obtain test statistics representing the strength of association between each SNP and the phenotype. Second, performing the signal detection strategy by fitting a cubic smoothing spline on the test statistic values, we identify QTLs. Third, we apply the RF classifier using the Boruta algorithm to assess the relative importance of SNPs regarding the level of their association with the phenotype. Finally, the important SNPs are prioritized within those QTLs to discover a robust set of SNPs associated with the phenotype. Two different GWAS (genotype and phenotype) datasets related to eggshell strength (ESS) and egg weight (EW) have been analysed using this framework to demonstrate its functionality.

### 3.1. Single-SNP Based GWAS Analysis

In order to demonstrate the limited power of conventional single-SNP based GWAS analysis, we first analysed both datasets using a LMM as suggested for the analysis of egg weight (EW) in the study of Liu et al. [48]. In the application of LMM, we considered the correction of the population

stratification and applied the *SimpleM* method [51] for multiple testing correction. The LMM approach for eggshell strength (ESS) at time point 1 (ESS1) and time point 2 (ESS2) led to the identification of only one significant SNP for ESS1 (see Figure 2A,B). Furthermore, the LMM method revealed 43 significant SNPs for EW (see Figure 2C) on chromosome 1 (GGA1) which were then mapped to three genes (ITM2B, RCBTB2, RB1).



**Figure 2.** Manhattan and Q-Q plots corresponding to eggshell strength at time point 1 (ESS1), time point 2 (ESS2) and egg weight at 36 weeks of age (EW36). In Manhattan plots (A–C), the horizontal red and green lines denote the genome-wide significance ( $p\text{-value} = 1.7 \times 10^{-6}$  for ESS1 and ESS2 and  $1.5 \times 10^{-6}$  for EW36) and suggestive significance thresholds ( $p\text{-value} = 3.4 \times 10^{-5}$  for ESS1 and ESS2  $3.1 \times 10^{-5}$  for EW), respectively. The  $-\log_{10}$  of the observed  $p$ -values for each single nucleotide polymorphism (SNP) is given on the y-axis while its position on a chromosome is given on the x-axis. In Q-Q plots (D–F) the observed  $-\log_{10}$  transformed  $p$ -values are plotted against the expected  $-\log_{10}$  transformed  $p$ -values. GIF stands for genomic inflation factor.

Today it is well known that quantitative traits are influenced by a large number of genes mostly having small effects. But as shown in Figure 2, many association signals were not strong enough to reach the significance threshold, thereby their influences on the phenotype are missed.

### 3.2. Detection of Genotype-Phenotype Association Using the Combined Framework

To identify genes showing weak association signals that remain undetected in the typical GWAS analysis, we applied our analysis framework to both datasets.

The analysis of the ESS datasets reveals eight QTLs for ESS1 and five QTLs for ESS2 based on the signal detection approach. The details of these QTLs are given in Table 1. Interestingly, we found chromosome 9 (GGA9), 10 (GGA10), 15 (GGA15) and 20 (GGA20) to have QTLs associated with ESS at both time points. Especially, the QTLs on GGA20 are overlapping and underpin the same genomic region as associated with ESS at both time points. In addition, the application of the RF classifier provides 3726 and 1815 SNPs which map to 405 and 253 genes associated with ESS1 and ESS2, respectively. The lists of these SNPs and their corresponding genes are taken from our previous study [30]. The investigation of these SNPs in the identified QTLs reveals 158 and 14 robust SNPs related to ESS1 and ESS2, respectively (the list of the SNPs is given in Table S1).

Of particular interest here is the LD analysis that we performed based on the robust SNPs to further elaborate their makeup in the identified QTLs. The LD analysis reveals, as expected, that the robust SNPs inside the QTLs have a remarkably higher level of LD than the surrounding SNPs (see Figure 3A). To this end, we exemplarily compared the phenotype differences between the genotypes of the top two SNP (rs315330686, rs314045861) on GGA18. The comparison suggests that for both SNPs, the birds homozygous for the minor alleles have higher phenotypes than those of the other two genotypes (Figure 3B,C).

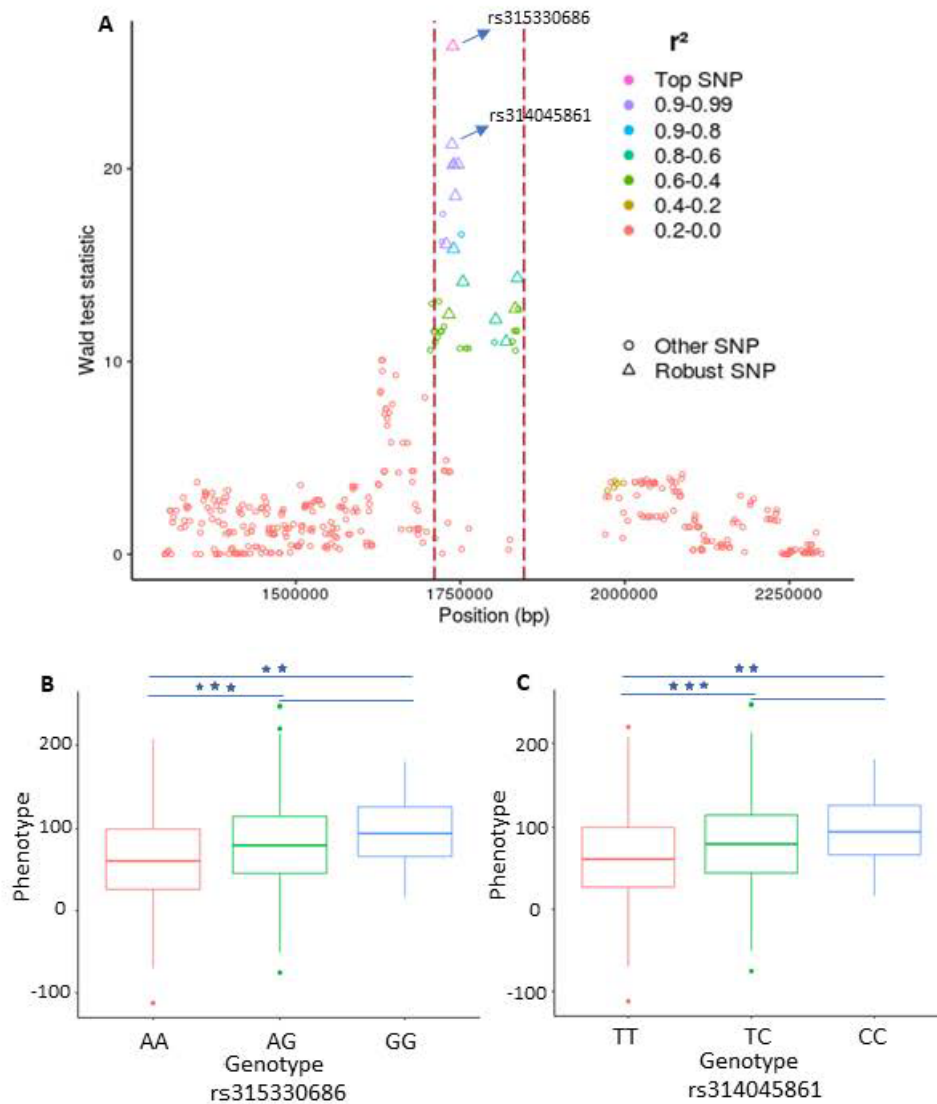
**Table 1.** Significant peaks as defined in Phase 4 of our analysis framework and corresponding quantitative trait loci (QTLs) for ESS1 and ESS2.

Chromosome	No. of SNPs	Start Position	End Position	No. of Genes	Trait
2	204	147,575,318	148,273,465	3	ESS1
9	66	21,762,694	21,953,310	0	ESS1
9	82	21,777,888	22,001,729	0	ESS2
10	75	6,517,673	6,728,897	4	ESS1
10	86	9,922,422	10,054,824	2	ESS1
10	60	10,715,120	10,818,097	3	ESS2
10	61	11,245,585	11,351,799	1	ESS2
12	112	10,948,518	11,227,521	2	ESS1
15	42	4,908,007	5,006,688	7	ESS1
15	43	6,193,090	6,273,778	3	ESS2
18	38	1,722,586	1,836,741	2	ESS1
20	51	7,589,607	7,717,177	1	ESS1
20	46	7,599,368	7,711,505	1	ESS2

The extraction of the genes corresponding to the robust SNPs reveals 14 and 3 genes for ESS1 and ESS2, respectively (the list of the genes is given in Table S1). The functional investigation of these genes shows that the majority of them were annotated to play essential roles in the transport of minerals and organic compounds. Seven of these genes, namely ATP6V0A2 (ATPase, H<sup>+</sup> Transporting, Lysosomal V0 Subunit A2), DDX55 (DEAD-Box Helicase 55), DNAH10 (Dynein Axonemal Heavy Chain 10), GTF2H3 (General Transcription Factor IIH Subunit 3), MYO1E (Unconventional Myosin 1E), TCTN2 (Tectonic Family Member), and MYH10 (Myosin Heavy Chain 10)), have molecular functions related to the activity of the ATPase enzyme. Interestingly, in relation to eggshell formation ATPases have long been known to show intense activity in the cells of shell gland during the synthesis of eggshell [56]. Furthermore, CHRNA7 (Cholinergic Receptor Nicotinic Alpha 7 Subunit), is associated with the transport of ions, especially calcium ions. The other main function performed by the identified genes includes cell morphogenesis which ensures the homeostasis of tissues involved in the development of eggshell [57,58]. The genes that play a role in this process include NDEL1 (NudE Neurodevelopment Protein 1 Like 1), ADGRB1 (Adhesion G Protein-Coupled Receptor B1), THSD4 (Thrombospondin Type 1 Domain Containing 4) and EIF2B1 (Eukaryotic Translation Initiation Factor 2B Subunit Alpha).

Among the genes found to be associated with ESS2, TRPM7 (Transient Receptor Potential Cation Channel Subfamily M Member 7) and BNC1 (Basonuclin 1) have functions related to the homeostasis

of ions in the cell. On the other hand, the CDH4 (Cadherin-4) gene that was found for both ESS1 and ESS2 encodes for R-cadherin/cadherin-4 which are single-chain integral membrane glycoproteins and mediate calcium-dependent cell–cell adhesion. Reduced levels of these cell adhesion molecules lead to the age-related decline in tissue homeostasis [59]. Along with other members of the cadherin superfamily, R-cadherins play roles in cell differentiation in a variety of tissues including bones, kidneys and uterus [60–63].



**Figure 3.** Plot representing a genomic region on chromosome 18 which is in association with eggshell strength at time point 1 (ESS1). (A) Plot representing the linkage disequilibrium (LD) structure inside and around a significant peak. The dotted red lines depict the boundaries of the peak. Each point represents a single nucleotide polymorphism (SNP) and the color shows the strength of LD between the top SNP inside the peak and the SNP surrounding it. The diamond shape points inside the peak depict the robust SNPs. The X-axis contains the SNP positions on the chromosome while the y-axis depicts the Wald statistic values obtained from the single-SNP based genome wide association study (GWAS) analysis. (B,C) The effects of different genotypes of the two leading SNPs identified in the combined framework for ESS and their significance (\*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ ).

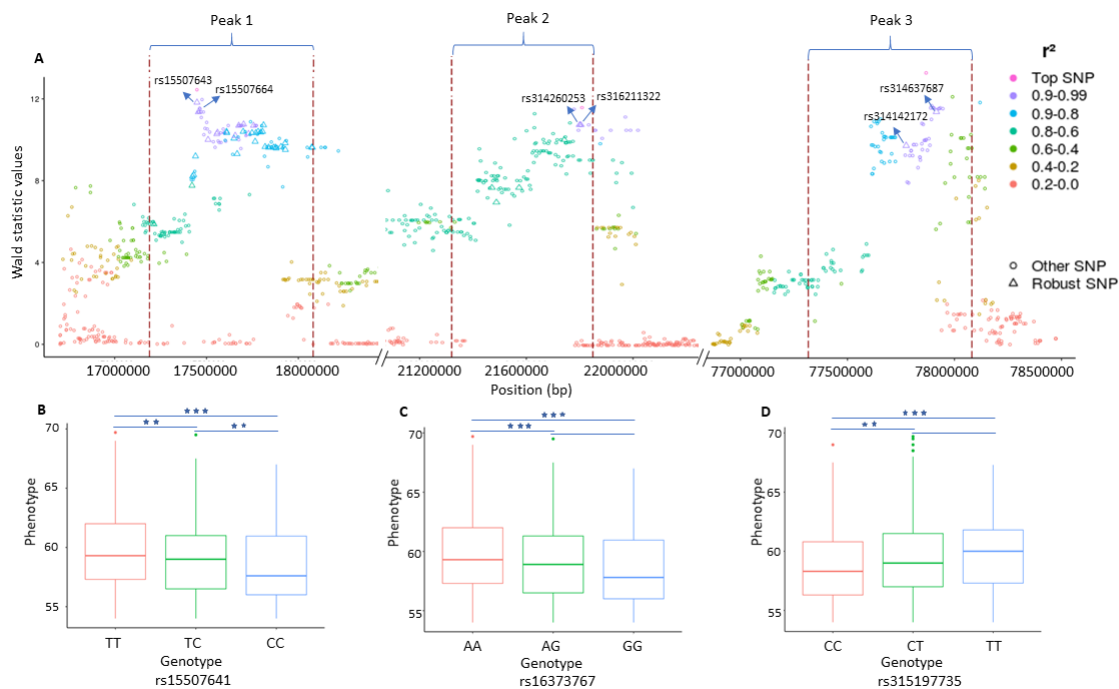
The analysis of the EW dataset resulted in the detection of eleven QTLs including the one revealed on chromosome 1 (GGA1) in the original study [48]. The additional QTLs were found on chromosomes 4 (GGA4), 12 (GGA12), 13 (GGA13), 14 (GGA14), 15 (GGA15) and 18 (GGA18). The details of these

eleven QTLs are summarized in the Table 2. Remarkably, there is no overlap between the QTLs observed for EW and ESS. The application of the RF classifier on this dataset provides a list of 753 important SNPs. A closer look at these SNPs points out that 145 of them (including 41 SNP identified in the original study [48]) are defined to be robust SNPs due to their genomic positions within the QTLs (the list of the SNPs is given in Table S1). Similar to the analysis of the ESS dataset, LD analysis based on the EW dataset also demonstrates the presence of strong linkage between robust SNPs.

**Table 2.** Significant peaks as defined in Phase 4 of our analysis framework and corresponding QTLs for EW.

Chromosome	No. of SNPs	Start Position	End Position	No. of Genes
1	304	167,931,038	169,505,140	25
4	205	17,189,770	18,080,445	9
4	143	21,319,808	21,849,558	3
4	136	77,317,446	78,081,369	4
12	39	2,849,562	3,010,032	7
13	49	8,495,533	8,608,578	6
14	58	7,023,793	7,188,250	4
15	41	11,193,342	11,309,808	8
15	35	11,419,957	11,514,516	3
18	30	1,057,714	1,136,220	1
18	28	1,179,899	1,238,583	0

The extraction of the genes associated with the robust SNPs related to EW results in the determination of 16 genes (the list is given in Table S1). Despite no overlap between the QTLs identified for ESS and EW, a variety of genes are involved in the same biological functions. Especially, many of the genes have their functions annotated to trans-membrane transportation of minerals and proteins. In this regard, genes including SCNN1G (Sodium Channel Epithelial 1 Subunit Gamma), AFAP1L1 (Actin Filament Associated Protein 1 Like 1), CD99L2 (CD99 Molecule Like 2), GPR50 (G Protein-Coupled Receptor 50), GRIA2 (Glutamate Ionotropic Receptor AMPA Type Subunit), GRPEL2 (GGrpE Like 2, Mitochondrial), HS3ST4 (SH3 Domain And Tetratricopeptide Repeats 2), ITM2B (Integral Membrane Protein 2B), MED4 (Mediator Complex Subunit 4), MTMR1 (Myotubularin Related Protein 1) and SH3TC2 (SH3 Domain And Tetratricopeptide Repeats 2) encode proteins that are part of cell membranes. By regulating the transport of ingredients for the egg development they can play a role in the determination of EW. More importantly, the SCNN1G encodes a non-voltage gated sodium channel to ensure the trans-membrane transportation of sodium ions. Higher expression of this gene during egg formation has been reported to play an important role in the determination of eggshell quality [64]. Similarly, the GRIA2 gene product functions as ligand-activated cation channel that allows the trans-membrane transportation of different ions. On the other hand, genes like RCBTB2 (RCC1 and BTB Domain Containing Protein 2) and TBC1D8B (TBC1 Domain Family Member 8B) can play a role in the regulation of these transportation channels. Functional annotations of RB1 (Retinoblastoma Transcriptional Corepressor 1) and MED4 genes are related to nuclear hormone receptor binding, a process principally involved in mineral metabolism. In particular, the MED4 encoded protein is a component of the vitamin D receptor-interacting protein complex that has been shown to contribute critically for the regulation of calcium absorption in the intestine [65]. The regulation of the intra-cellular protein transport and the cellular protein localization are biological functions performed by the ABLIM3 (Actin Binding LIM Protein Family Member 3) gene.



**Figure 4.** Plot representing three genomic regions on chromosome 4 in association with egg weight (EW). (A) Plot representing the LD structure inside and around the significant peaks. The dotted red lines depict the boundaries of the peaks. Each point represents a SNP and the color shows the strength of linkage disequilibrium (LD) between the top single nucleotide polymorphisms (SNPs) inside each peak and the surrounding SNPs. The diamond shape points inside the peak depict the robust SNPs. The X-axis contains the SNP positions on the chromosome while the y-axis depicts the Wald statistic values obtained from single-SNP based GWAS analysis. ((B–D) The effects of different genotypes of the three leading SNPs identified for EW and their significance (\*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ )).

#### 4. Discussion

Deciphering genotype-phenotype associations for quantitative traits still remains challenging due to the weak contribution of many individual SNPs to the phenotype. To address this problem, several approaches including single-SNP or multiple-SNP based models have been developed [18]. The worth of single-SNP models is well testified by the repertoire of genes related to a variety of traits that has been discovered using these models [2]. However, for quantitative traits where a multitude of genes may act in concert to confer a particular phenotypic value to an individual, the power of these single-SNP based models is limited [4,6,30,66]. Multi-SNP models are potentially more competent for the detection of smaller effects, but mostly require a prior distribution of SNP effects that is not known for most of the traits while for some traits they may not even follow a strict distribution [18,67]. To overcome these limitations, combining single-SNP based statistics over a genomic region to test its association with the trait has been the method of choice for many scientists [68–71]. In this regard, Beissinger et al. [72] show the superiority of cubic smoothing spline techniques over some other methods to combine single-SNP based statistics for the discovery of selection signatures. Furthermore, Zhang et al. [45] have praised the utility of spline based techniques to integrate association statistics in order to identify the causal alleles. However, these methods do not provide a clear framework that can be used to identify genomic regions with subtle effects on the phenotypes in samples with family or population structures.

With the growing application of machine learning algorithms in the field of genomics, their application to ascertain the genotype-phenotype association is gaining importance. Contrary to traditional multi-SNP models, machine learning methods do not require any prior assumptions about

the genetic architecture of traits. In our recent study [30], we successfully applied an RF classifier to the ESS dataset to assess the importance of SNPs and identified large numbers of genes associated with eggshell strength at two different production stages. Despite the success of the RF classifiers in association analysis, there is still a need to prioritize the identified genes to recognize the genes having most robust association with the phenotype. This prioritization constitutes a means to delve deeper into the functioning of the individual genes to understand their marginal influences on the manifestation of the phenotype differences among the samples. For this purpose, we investigated genes within the QTLs that have association signals higher than expected by chance. The identification of QTLs is a fundamental step in our study which we have performed using a splines based strategy in several phases. Unlike previous studies [45,72], using this technique we harness the association signals, in order to detect the genomic regions harbouring genes potentially playing roles in the phenotype manifestation.

Our results show that the determination of QTLs by our signal detection approach and then the prioritization of SNPs within these QTLs (called robust SNPs), can lead to the discovery of genes which despite having association to the phenotypes, remain undetected in the typical GWAS. Especially, the combined usage of both methods (RF and signal detection) not only identify the QTLs having small effects but also helps to identify the SNPs in those QTLs that had their association value higher than expected by chance. (see Figures 3A and 4A). Moreover, the LD based on the robust SNPs (Figures 3A and 4A) supports us, on the one hand, to monitor their strong mutual correlation which is crucial to explain the genetic makeup of the underlying QTLs. On the other hand, it further substantiates our idea regarding the presence of signals which are caused by the strong LD in the QTLs and embedded in the association statistics.

Although both of the traits analysed in this study were related to egg quality, the identified genes are distinct for ESS and EW in this study. This distinction was expected as the chickens genotyped in the two datasets have different genetic backgrounds. Remarkably, however some of these genes are involved in the same biological function related to transmembrane transportation of elements including minerals and organic compounds. Further, the majority of the ESS1 related genes are responsible for the availability of calcium ( $\text{Ca}^{2+}$ ) and bicarbonate ( $\text{HCO}_3^-$ ) which are prerequisites for eggshell mineralization in the uterus part of the oviduct. These ions are supplied in large amounts via trans-epithelial transport in the uterus, for which ion channels, ion pumps and ion exchangers are required [73]. This function is mainly regulated by ATPase, an enzyme which is implicated in this process through several genes which were identified in this analysis for ESS. The ATPase enzyme decomposes ATP into ADP to release the energy required to perform energy intensive tasks by the cell. Regarding eggshell formation, ATPases have long been known to influence the microvilli of the tubular cells of the shell gland during the process of eggshell formation [56]. Similarly, inhibition of ATPase from the shell glands has been demonstrated to cause the thinning of the eggshell due to the inhibition of the calcium transport across the shell gland epithelium which is known to be an energy expensive process [74]. The hydrogen potassium ATPase maintains a certain pH level of the uterine fluid during the eggshell formation by acting as a pump to transfer the hydrogen ions ( $\text{H}^+$ ) from the uterine cell of chicken to plasma. In this regard, two paralogs (ATP6V1B, ATP6V1C2) of the ATP6V0A2 gene found in our study have been previously reported to transfer hydrogen ion from chicken uterine cells to blood plasma during the process of egg calcification [73,75]. When integrated into biological membranes, the so-called transmembrane ATPases take part in the transportation of metabolites across the membranes [76]. Transmembrane ATPases exchange many metabolites across the membranes and provide the necessary environment for activities of the cell [77]. Similarly, genes discovered for EW encode cell membrane proteins which can act as channels for the transportation of minerals as well as proteins. Among these, one of the most important channel protein is encoded by the SCNN1G. This gene belongs to the sodium channel gene family. Many members of this gene family are known to affect egg weight as well as other egg quality traits [64].



The other important functional category that many of the genes related to ESS could be linked to is cell morphogenesis. Previous studies presenting the transcriptome profile of different segments of the chicken oviduct have also reported a large number of genes annotated for functions related to morphogenesis [73,78,79]. It is also important to note the difference in genes identified for ESS1 and ESS2. It depicts the change in the genetic and environmental components of the phenotypic variance over age which has been previously reported for other complex traits [80,81]. Given all these results, our suggested framework is capable of highlighting the important genes within the QTLs having moderate to small effects. The availability of larger datasets can further improve the power of this framework to detect novel QTLs. Furthermore, well established polygenic approaches can also be integrated in this framework for the discovery of even robust associations. On top of that, our strategy is complementary to our previous study in which we performed a RF based feature selection technique for genotype-phenotype association.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2073-4425/11/8/892/s1>, Script S1: R-script for analysis of SNPs and for the extraction of corresponding genes, Table S1: The list of important SNPs and genes, Figure S1: Venn diagrams showing the overlap between the SNPs identified using our signal detection approach and those identified as important by a Random Forests classifier.

**Author Contributions:** M.G. designed and supervised the research. F.R. and A.O.S. participated in the design of the study. F.R. conducted computational and statistical analyses. F.R. prepared and studied the GWAS data and interpret the results. M.G. was involved in the interpretation of the results together with F.R. H.B. and A.O.S. developed the programming scripts with F.R. D.C. constructed the ESS dataset. F.R. 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 work is part of FR's doctoral program which is funded by the of Government of the Punjab, Pakistan under the project "50 oversees PhD Scholarships for University of Agriculture, Faisalabad".

**Acknowledgments:** The chicken data used in this study were provided by the "Synbreed—Synergistic Plant and Animal Breeding" project for which we are grateful to the project team. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University. We would like to thank Selina Klees, Abirami Rajavel and Martin Wutke for proofreading the manuscript and Malena Erbe for providing important insights into the chicken dataset.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gallagher, M.D.; Chen-Plotkin, A.S. The post-GWAS era: From association to function. *Am. J. Hum. Genet.* **2018**, *102*, 717–730. [[CrossRef](#)]
2. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [[CrossRef](#)] [[PubMed](#)]
3. 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 Genom.* **2010**, *11*, 724. [[CrossRef](#)] [[PubMed](#)]
4. Bush, W.S.; Moore, J.H. Genome-wide association studies. *PLoS Comput. Biol.* **2012**, *8*, e1002822. [[CrossRef](#)] [[PubMed](#)]
5. Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* **2013**, *9*, 29. [[CrossRef](#)]
6. Holland, D.; Fan, C.C.; Frei, O.; Shadrin, A.A.; Smeland, O.B.; Sundar, V.; Andreassen, O.A.; Dale, A.M. Estimating inflation in GWAS summary statistics due to variance distortion from cryptic relatedness. *BioRxiv* **2017**, 164939. [[CrossRef](#)]
7. Zhang, Y.M.; Mao, Y.; Xie, C.; Smith, H.; Luo, L.; Xu, S. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* **2005**, *169*, 2267–2275. [[CrossRef](#)]
8. Yu, J.; Pressoir, G.; Briggs, W.H.; Bi, I.V.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B.; et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **2006**, *38*, 203–208. [[CrossRef](#)]

9. Kang, H.M.; Sul, J.H.; Service, S.K.; Zaitlen, N.A.; Kong, S.y.; Freimer, N.B.; Sabatti, C.; Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **2010**, *42*, 348. [[CrossRef](#)]
10. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821. [[CrossRef](#)]
11. Eu-Ahsunthornwattana, J.; Miller, E.N.; Fakiola, M.; Jeronimo, S.M.; Blackwell, J.M.; Cordell, H.J.; Wellcome Trust Case Control Consortium 2. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet.* **2014**, *10*, e1004445. [[CrossRef](#)] [[PubMed](#)]
12. Balding, D.J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **2006**, *7*, 781–791. [[CrossRef](#)] [[PubMed](#)]
13. Young, A.I. Solving the missing heritability problem. *PLoS Genet.* **2019**, *15*, e1008222. [[CrossRef](#)] [[PubMed](#)]
14. Long, A.D.; Langley, C.H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **1999**, *9*, 720–731.
15. Akey, J.; Jin, L.; Xiong, M. Haplotypes vs single marker linkage disequilibrium tests: What do we gain? *Eur. J. Hum. Genet.* **2001**, *9*, 291. [[CrossRef](#)]
16. Zhang, K.; Calabrese, P.; Nordborg, M.; Sun, F. Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.* **2002**, *71*, 1386–1394. [[CrossRef](#)]
17. Lorenz, A.J.; Hamblin, M.T.; Jannink, J.L. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS ONE* **2010**, *5*, e14079. [[CrossRef](#)]
18. Schmid, M.; Bennewitz, J. Invited review: Genome-wide association analysis for quantitative traits in livestock—A selective review of statistical models and experimental designs. *Arch. Tierz.* **2017**, *60*, 335. [[CrossRef](#)]
19. Zhang, Y.M.; Jia, Z.; Dunwell, J.M. The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* **2019**, *10*, 100. [[CrossRef](#)] [[PubMed](#)]
20. Wen, Y.J.; Zhang, H.; Ni, Y.L.; Huang, B.; Zhang, J.; Feng, J.Y.; Wang, S.B.; Dunwell, J.M.; Zhang, Y.M.; Wu, R. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* **2018**, *19*, 700–712. [[CrossRef](#)] [[PubMed](#)]
21. Cui, Y.; Zhang, F.; Zhou, Y. The application of multi-Locus GWAS for the detection of salt-tolerance loci in rice. *Front. Plant Sci.* **2018**, *9*, 1464. [[CrossRef](#)] [[PubMed](#)]
22. Ma, L.; Liu, M.; Yan, Y.; Qing, C.; Zhang, X.; Zhang, Y.; Long, Y.; Wang, L.; Pan, L.; Zou, C.; et al. Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies. *Front. Plant Sci.* **2018**, *9*, 561. [[CrossRef](#)]
23. Xu, Y.; Yang, T.; Zhou, Y.; Yin, S.; Li, P.; Liu, J.; Xu, S.; Yang, Z.; Xu, C. Genome-wide association mapping of starch pasting properties in maize using single-locus and multi-locus models. *Front. Plant Sci.* **2018**, *9*, 1311. [[CrossRef](#)]
24. Abed, A.; Belzile, F. Comparing Single-SNP, Multi-SNP, and Haplotype-Based Approaches in Association Studies for Major Traits in Barley. *Plant Genome* **2019**, *12*, 1–14. [[CrossRef](#)]
25. Zhao, Y.; Chen, F.; Zhai, R.; Lin, X.; Wang, Z.; Su, L.; Christiani, D.C. Correction for population stratification in random forest analysis. *Int. J. Epidemiol.* **2012**, *41*, 1798–1806. [[CrossRef](#)] [[PubMed](#)]
26. Nguyen, T.T.; Huang, J.Z.; Wu, Q.; Nguyen, T.T.; Li, M.J. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genom.* **2015**, *16*, S5.
27. Armero, C.; Cabras, S.; Castellanos, M.E.; Quirós, A. Two-Stage Bayesian Approach for GWAS With Known Genealogy. *J. Comput. Graph. Stat.* **2019**, *28*, 197–204. [[CrossRef](#)]
28. Huang, X.; Zhou, W.; Bellis, E.S.; Stubblefield, J.; Causey, J.; Qualls, J.; Walker, K. Minor QTLs mining through the combination of GWAS and machine learning feature selection. *BioRxiv* **2019**, 712190. [[CrossRef](#)]
29. Brieuc, M.S.; Waters, C.D.; Drinan, D.P.; Naish, K.A. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol. Ecol. Resour.* **2018**, *18*, 755–766. [[CrossRef](#)]
30. 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](#)]
31. Romagnoni, A.; Jégou, S.; Van Steen, K.; Wainrib, G.; Hugot, J.P. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Sci. Rep.* **2019**, *9*, 1–18. [[CrossRef](#)]

32. Van der Heide, E.; Veerkamp, R.; van Pelt, M.; Kamphuis, C.; Athanasiadis, I.; Ducro, B. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. *J. Dairy Sci.* **2019**, *102*, 9409–9421. [[CrossRef](#)] [[PubMed](#)]
33. Li, B.; Zhang, N.; Wang, Y.G.; George, A.W.; Reverter, A.; Li, Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* **2018**, *9*, 237. [[CrossRef](#)]
34. Hamblin, M.T.; Jannink, J.L. Factors affecting the power of haplotype markers in association studies. *Plant Genome* **2011**, *4*, 145–153. [[CrossRef](#)]
35. Sarti, F.; Lasagna, E.; Ceccobelli, S.; Di Lorenzo, P.; Filippini, F.; Sbarra, F.; Giontella, A.; Pieramati, C.; Panella, F. Influence of single nucleotide polymorphisms in the myostatin and myogenic factor 5 muscle growth-related genes on the performance traits of Marchigiana beef cattle. *J. Anim. Sci.* **2014**, *92*, 3804–3810. [[CrossRef](#)] [[PubMed](#)]
36. Sarti, F.M.; Ceccobelli, S.; Lasagna, E.; Di Lorenzo, P.; Sbarra, F.; Pieramati, C.; Giontella, A.; Panella, F. Influence of single nucleotide polymorphisms in some candidate genes related to the performance traits in Italian beef cattle breeds. *Livest. Sci.* **2019**, *230*, 103834. [[CrossRef](#)]
37. Yang, Y.; Wu, L.; Wu, X.; Li, B.; Huang, W.; Weng, Z.; Lin, Z.; Song, L.; Guo, Y.; Meng, Z.; et al. Identification of Candidate Growth-Related SNPs and Genes Using GWAS in Brown-Marbled Grouper (*Epinephelus fuscoguttatus*). *Mar. Biotechnol.* **2020**, *22*, 153–166. [[CrossRef](#)]
38. Freebern, E.; Santos, D.J.; Fang, L.; Jiang, J.; Gaddis, K.L.P.; Liu, G.E.; Vanraden, P.M.; Maltecca, C.; Cole, J.B.; Ma, L. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. *BMC Genom.* **2020**, *21*, 41. [[CrossRef](#)]
39. Sanchez, M.P.; Guatteo, R.; Davergne, A.; Saout, J.; Grohs, C.; Deloche, M.C.; Taussat, S.; Fritz, S.; Boussaha, M.; Blanquefort, P.; et al. Identification of the ABCC4, IER3, and CBFA2T2 candidate genes for resistance to paratuberculosis from sequence-based GWAS in Holstein and Normande dairy cattle. *Genet. Sel. Evol.* **2020**, *52*, 1–17. [[CrossRef](#)]
40. Sinclair-Waters, M.; Ødegård, J.; Korsvoll, S.A.; Moen, T.; Lien, S.; Primmer, C.R.; Barson, N.J. Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon. *Genet. Sel. Evol.* **2020**, *52*, 1–11. [[CrossRef](#)]
41. Horn, S.S.; Ruyter, B.; Meuwissen, T.H.; Moghadam, H.; Hillestad, B.; Sonesson, A.K. GWAS identifies genetic variants associated with omega-3 fatty acid composition of Atlantic salmon fillets. *Aquaculture* **2020**, *514*, 734494. [[CrossRef](#)]
42. Nicholls, H.L.; John, C.R.; Watson, D.S.; Munroe, P.B.; Barnes, M.R.; Cabrera, C.P. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. *Front. Genet.* **2020**, *11*, 350. [[CrossRef](#)] [[PubMed](#)]
43. Misra, G.; Badoni, S.; Anacleto, R.; Graner, A.; Alexandrov, N.; Sreenivasulu, N. Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* **2017**, *7*, 1–16.
44. Li, C.; Fu, Y.; Sun, R.; Wang, Y.; Wang, Q. Single-locus and multi-locus genome-wide association studies in the genetic dissection of fiber quality traits in upland cotton (*Gossypium hirsutum* L.). *Front. Plant Sci.* **2018**, *9*, 1083. [[CrossRef](#)] [[PubMed](#)]
45. Zhang, X.; Roeder, K.; Wallstrom, G.; Devlin, B. Integration of association statistics over genomic regions using Bayesian adaptive regression splines. *Hum. Genom.* **2003**, *1*, 20. [[CrossRef](#)] [[PubMed](#)]
46. Schwarz, D.F.; Szymczak, S.; Ziegler, A.; König, I.R. Picking single-nucleotide polymorphisms in forests. *BMC Proc.* **2007**, *1*, S59.
47. Garrick, D.J.; Taylor, J.F.; Fernando, R.L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **2009**, *41*, 55. [[CrossRef](#)]
48. 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. [[CrossRef](#)]
49. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; De Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [[CrossRef](#)]
50. Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C.M.; Davidson, R.I.; Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **2011**, *8*, 833. [[CrossRef](#)]

51. Gao, X.; Becker, L.C.; Becker, D.M.; Starmer, J.D.; Province, M.A. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **2010**, *34*, 100–105. [[CrossRef](#)]
52. Wood, S.N. *Generalized Additive Models: An Introduction with R*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2017.
53. Silverman, B.W. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. R. Stat. Soc. Ser. B (Methodol.)* **1985**, *47*, 1–21. [[CrossRef](#)]
54. Kursu, M.B.; Rudnicki, W.R.; et al. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [[CrossRef](#)]
55. 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. [[CrossRef](#)] [[PubMed](#)]
56. Yamamoto, T.; Ozawa, H.; Nagai, H. Histochemical studies of Ca-ATPase, succinate and NAD<sup>+</sup>-dependent isocitrate dehydrogenases in the shell gland of laying Japanese quails: with special reference to calcium-transporting cells. *Histochemistry* **1985**, *83*, 221–226. [[CrossRef](#)] [[PubMed](#)]
57. Wang, Y.; Guo, F.; Qu, H.; Luo, C.; Wang, J.; Shu, D. Associations between variants of bone morphogenetic protein 7 gene and growth traits in chickens. *Br. Poult. Sci.* **2018**, *59*, 264–269. [[CrossRef](#)]
58. Jin, S. Bipotent stem cells support the cyclical regeneration of endometrial epithelium of the murine uterus. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6848–6857. [[CrossRef](#)]
59. Boyle, M.; Wong, C.; Rocha, M.; Jones, D.L. Decline in self-renewal factors contributes to aging of the stem cell niche in the Drosophila testis. *Cell Stem Cell* **2007**, *1*, 470–478. [[CrossRef](#)]
60. Adams, C.L.; Chen, Y.T.; Smith, S.J.; James Nelson, W. Mechanisms of epithelial cell–cell adhesion and cell compaction revealed by high-resolution tracking of E-cadherin–green fluorescent protein. *J. Cell Biol.* **1998**, *142*, 1105–1119. [[CrossRef](#)]
61. Dahl, U.; Sjödin, A.; Larue, L.; Radice, G.L.; Cajander, S.; Takeichi, M.; Kemler, R.; Semb, H. Genetic dissection of cadherin function during nephrogenesis. *Mol. Cell. Biol.* **2002**, *22*, 1474–1487. [[CrossRef](#)]
62. Marie, P.J.; Haÿ, E.; Modrowski, D.; Revollo, L.; Mbalaviele, G.; Civitelli, R. Cadherin-mediated cell–cell adhesion and signaling in the skeleton. *Calcif. Tissue Int.* **2014**, *94*, 46–54. [[CrossRef](#)]
63. Vazquez-Levin, M.H.; Marín-Briggiler, C.I.; Caballero, J.N.; Veiga, M.F. Epithelial and neural cadherin expression in the mammalian reproductive tract and gametes and their participation in fertilization-related events. *Dev. Biol.* **2015**, *401*, 2–16. [[CrossRef](#)]
64. Fan, Y.F.; Hou, Z.C.; Yi, G.Q.; Xu, G.Y.; Yang, N. The sodium channel gene family is specifically expressed in hen uterus and associated with eggshell quality traits. *BMC Genet.* **2013**, *14*, 90. [[CrossRef](#)] [[PubMed](#)]
65. Fleet, J.C.; Schoch, R.D. Molecular mechanisms for regulation of intestinal calcium absorption by vitamin D and other factors. *Crit. Rev. Clin. Lab. Sci.* **2010**, *47*, 181–195. [[CrossRef](#)] [[PubMed](#)]
66. Josephs, E.B.; Stinchcombe, J.R.; Wright, S.I. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytol.* **2017**, *214*, 21–33. [[CrossRef](#)] [[PubMed](#)]
67. Liu, Y.; Wang, D.; He, F.; Wang, J.; Joshi, T.; Xu, D. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* **2019**, *10*, 1091. [[CrossRef](#)]
68. Zaykin, D.V.; Zhivotovsky, L.A.; Westfall, P.H.; Weir, B.S. Truncated product method for combining P-values. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **2002**, *22*, 170–185.
69. Dudbridge, F.; Koeleman, B.P. Rank truncated product of P-values, with application to genomewide association scans. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **2003**, *25*, 360–366. [[CrossRef](#)]
70. Yang, H.C.; Lin, C.Y.; Fann, C.S. A sliding-window weighted linkage disequilibrium test. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **2006**, *30*, 531–545. [[CrossRef](#)]
71. Yang, H.C.; Hsieh, H.Y.; Fann, C.S. Kernel-based association test. *Genetics* **2008**, *179*, 1057–1068. [[CrossRef](#)]
72. Beissinger, T.M.; Rosa, G.J.; Kaeppeler, S.M.; Gianola, D.; De Leon, N. Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genet. Sel. Evol.* **2015**, *47*, 30. [[CrossRef](#)]
73. 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](#)]

74. Khan, H.M.; Cutkomp, L. In vitro studies of DDT, DDE, and ATPase as related to avian eggshell thinning. *Arch. Environ. Contam. Toxicol.* **1982**, *11*, 627–633. [[CrossRef](#)] [[PubMed](#)]
75. Jonchère, V.; Brionne, A.; Gautron, J.; Nys, Y. Identification of uterine ion transporters for mineralisation precursors of the avian eggshell. *BMC Physiol.* **2012**, *12*, 10. [[CrossRef](#)]
76. Chakraborti, S.; Dhalla, N.S. *Regulation of Membrane Na<sup>+</sup>-K<sup>+</sup> ATPase*; Springer: Berlin, Germany, 2016.
77. Morth, J.P.; Pedersen, B.P.; Buch-Pedersen, M.J.; Andersen, J.P.; Vilsen, B.; Palmgren, M.G.; Nissen, P. A structural overview of the plasma membrane Na<sup>+</sup>, K<sup>+</sup>-ATPase and H<sup>+</sup>-ATPase ion pumps. *Nat. Rev. Mol. Cell Biol.* **2011**, *12*, 60. [[CrossRef](#)] [[PubMed](#)]
78. Wan, Y.; Jin, S.; Ma, C.; Wang, Z.; Fang, Q.; Jiang, R. RNA-Seq reveals seven promising candidate genes affecting the proportion of thick egg albumen in layer-type chickens. *Sci. Rep.* **2017**, *7*, 1–9.
79. Yin, Z.; Lian, L.; Zhu, F.; Zhang, Z.H.; Hincke, M.; Yang, N.; Hou, Z.C. The transcriptome landscapes of ovary and three oviduct segments during chicken (*Gallus gallus*) egg formation. *Genomics* **2020**, *112*, 243–251. [[CrossRef](#)]
80. Elks, C.E.; Den Hoed, M.; Zhao, J.H.; Sharp, S.J.; Wareham, N.J.; Loos, R.J.; Ong, K.K. Variability in the heritability of body mass index: a systematic review and meta-regression. *Front. Endocrinol.* **2012**, *3*, 29. [[CrossRef](#)]
81. He, L.; Sillanpää, M.J.; Silventoinen, K.; Kaprio, J.; Pitkäniemi, J. Estimating modifying effect of age on genetic and environmental variance components in twin models. *Genetics* **2016**, *202*, 1313–1328. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

### **A.3. R-script to implement signal detection strategy.**

## R-Script for signal detection analysis

```
#Specify type of data that is read in. One string without spaces

type <- "Real" # Both Real and permuted data can be used. Will be used in output file
#type <- "Permutated"

inFile <- "Realdata.assoc"
outFile <- "Peaks.txt"

cat("Reading in data set", inFile, "\n")
D <- read.table(file = inFile, header = TRUE)

cat("Dataset read in!\n")
cat("Type of data:\t", type, "\n")

#Pull out chromosomes
chromosomes <- unique(D$Chromosome)
chromosomes <- sort(chromosomes, decreasing = F)
#number of chromosomes
n.chr <- length(chromosomes)

GlobalListOfCandidatePeaks <- c()

for(i in 1 : n.chr)
{

  #Consider just one chromosome
  chr <- chromosomes[i]
  cat("Treating chromosome:\t", chr, "\n")

  D.aux <- subset(D, Chromosome == i)

  #make sure the entries are ordered with increasing BP
  o <- order(D.aux$Position, decreasing = F)
  D.aux <- D.aux[o,]

  #number of SNPs in this chromosome
  n.snp <- nrow(D.aux)

  cat("Number of SNPs on this chromosome:\t", n.snp, "\n")

  #Plot to get first impression. Change WaldStat to whatever the test statistic is called
  plot(D.aux$Position, D.aux$WaldStat, xlab = "Position (bp)", ylab = "WaldStat-value",
       main = paste("Chromosome", i, sep = " "))

  #Apply spline smoother, cv = FALSE specifies the generalized' cross-validation
```

```

splineSmoo <- smooth.spline(x = D.aux$Position, y = D.aux$WaldStat, cv = FALSE)
lines(x = splineSmoo$x, y = splineSmoo$y, col = "red")

#Spline smoothed curve consists of how many point pairs?
n.splinesmoo <- length(splineSmoo$x)

#Now make the derivates
#First derivative
firstDeriv.y <- diff(splineSmoo$y) / diff(splineSmoo$x)
firstDeriv.x <- splineSmoo$x[-length(splineSmoo$x)] #Kick out last value
#Second derivative
secondDeriv.y <- diff(firstDeriv.y) / diff(firstDeriv.x)
#Kick out first and last x-Value. One value is lost in every derivative
secondDeriv.x <- splineSmoo$x[-c(1, length(splineSmoo$x))]

SecondDerivative <- c(NA, secondDeriv.y, NA)
FirstDerivative <- c(firstDeriv.y, NA)

A <- data.frame(BP = splineSmoo$x,
               FunctionValue = splineSmoo$y,
               FirstDerivative = FirstDerivative,
               SecondDerivative = SecondDerivative)

A1 <- D.aux[D.aux$Position %in% A$BP, ]

A <- cbind(A1$SNP, A)

InflectionPoint <- (A$SecondDerivative[-1] > 0) * (A$SecondDerivative[-nrow(A)] < 0) |
  (A$SecondDerivative[-1] < 0) * (A$SecondDerivative[-nrow(A)] > 0)

InflectionPoint <- c(InflectionPoint, NA)
infl <- c(FALSE, diff(diff(splineSmoo$y)>0)!=0, FALSE)
A1 <- cbind(A, infl)
A <- cbind(A, InflectionPoint)

n <- nrow(A)
LeftBorder <- vector(length = n, mode = "logical")
RightBorder <- vector(length = n, mode = "logical")

for(j in 2 : (n - 1))
{
  if(!is.na(A$InflectionPoint[j]) && A$InflectionPoint[j] == TRUE &&
      (A$FunctionValue[j - 1] < A$FunctionValue[j + 1]))
  {
    LeftBorder[j] <- TRUE
  }
  if(!is.na(A$InflectionPoint[j]) && A$InflectionPoint[j] == TRUE &&
      (A$FunctionValue[j - 1] > A$FunctionValue[j + 1]))
  {
    RightBorder[j] <- TRUE
  }
}
}

```



```

A <- cbind(A, LeftBorder, RightBorder)

pos.l <- A[A$LeftBorder == TRUE,]$BP
abline(v = pos.l, col = "red")

pos.r <- A[A$RightBorder == TRUE,]$BP
abline(v = pos.r, col = "blue")

#Confirmation. A left border and the next right border form a peak.
#If a left border is followed by a left border it is disqualified

#Start confirmation process at left border with the smallest position (leftmost)
#First border must be left
start <- min(subset(A, LeftBorder == TRUE)$BP)

#Last border must be right
end <- max(subset(A, RightBorder == TRUE)$BP)
B <- subset(A, BP >= start & BP <= end & InflectionPoint == TRUE)

n.borders <- nrow(B)

Confirmation <- vector(length = n.borders, mode = "logical")
is.na(Confirmation) <- TRUE

for(j in 1 : n.borders)
{
  if(j == 1)
  {
    if( B[j, ]$LeftBorder == TRUE && B[j + 1, ]$LeftBorder == FALSE )
      { Confirmation[j] <- TRUE }
    else{ Confirmation[j] <- FALSE }
  }
  else if(j > 1 || j < n.borders)
  {
    if( ( B[j, ]$LeftBorder == TRUE && B[j + 1, ]$LeftBorder == FALSE ) ||
        ( B[j, ]$RightBorder == TRUE && B[j - 1, ]$RightBorder == FALSE ) )
      { Confirmation[j] <- TRUE }
    else { Confirmation[j] <- FALSE }
  }
  else
  {
    if( B[j, ]$RightBorder == TRUE && B[j + -1, ]$RightBorder == FALSE )
      { Confirmation[j] <- TRUE }
    else { Confirmation[j] <- FALSE }
  }
}

B <- cbind(B, Confirmation)

ConfirmedPeaks <- B[B$Confirmation,]

```

```

#The last border must be a right border
#If this is not the case throw out the last entry

if(ConfirmedPeaks[nrow(ConfirmedPeaks),]$RightBorder == FALSE)
{
  ConfirmedPeaks <- ConfirmedPeaks[-nrow(ConfirmedPeaks),]
}

#Assign peak numbers

ConfirmedPeaks$Peak <- ceiling((1 : nrow(ConfirmedPeaks)) / 2)
n.confirmedPeaks <- max(ConfirmedPeaks$Peak)

#Draw confirmed borders with fat lines

pos.l <- ConfirmedPeaks[ConfirmedPeaks$LeftBorder == TRUE,]$BP
abline(v = pos.l, col = "red", lwd = 2)

pos.r <- ConfirmedPeaks[ConfirmedPeaks$RightBorder == TRUE,]$BP
abline(v = pos.r, col = "blue", lwd = 2)

#Draw the peaks and determine height

Peaks <- vector(length = n.confirmedPeaks)
Height <- vector(length = n.confirmedPeaks)
Pos.left <- vector(length = n.confirmedPeaks)
Pos.right <- vector(length = n.confirmedPeaks)
Chr <- vector(length = n.confirmedPeaks)
NoSNP <- vector(length = n.confirmedPeaks)
iSNP <- vector(length = n.confirmedPeaks)
lSNP <- vector(length = n.confirmedPeaks)

for(j in 1 : n.confirmedPeaks)
{
  Peaks[j] <- j
  bp.left <- subset(ConfirmedPeaks, Peak == j)$BP[1]
  bp.right <- subset(ConfirmedPeaks, Peak == j)$BP[2]

  SNPS <- D.aux[D.aux$Position >= bp.left & D.aux$Position <= bp.right, ]
  NoSNP[j] <- nrow(SNPS)
  iSNP[j] <- as.vector(SNPS$SNP[1])
  lSNP[j] <- as.vector(SNPS$SNP[nrow(SNPS)])
  Pos.left[j] <- bp.left
  Pos.right[j] <- bp.right
  Chr[j] <- chr

  x.coord <- subset(A, BP >= bp.left & BP <= bp.right)$BP
  y.coord <- subset(A, BP >= bp.left & BP <= bp.right)$FunctionValue

  peak.height <- max(y.coord)
  Height[j] <- peak.height

  x.coord <- c(x.coord[1], x.coord, x.coord[length(x.coord)])

```

```

y.coord <- c(0, y.coord, 0)
  polygon(x = x.coord, y = y.coord, col = 'red')
}

ListOfCandidatePeaks <- data.frame(Peak = Peaks, NSNP = NoSNP, InitialSNP = iSNP,
  lastSNP = lSNP, Height = Height, Pos.left = Pos.left,
  Pos.right = Pos.right, Chr = Chr)

GlobalListOfCandidatePeaks <- rbind(GlobalListOfCandidatePeaks, ListOfCandidatePeaks)
}

Type <- rep(type, nrow(GlobalListOfCandidatePeaks))

GlobalListOfCandidatePeaks <- cbind(GlobalListOfCandidatePeaks, Type)

write.table(file = outFile, x = GlobalListOfCandidatePeaks, quote = F,
  row.names = F, sep = "\t")

```

**A.4. R-script to extract the list of genes corresponding to the important SNPs.**

## R-Script for extraction of genes corresponding to the important SNP from Ensembl database using Biomart

```
#####  
##### Script to get genes corresponding to a set of SNPs #####  
#####  
  
#### Required packages ####  
library(data.table)  
library(dplyr)  
library(biomart)  
  
#### Reading in the data frame containing SNPs and their chromosomal positions ####  
df <- read.table("/Home/faisal/PeakDetection/RealData/VarImportantES1.txt", header = T)  
  
#### Getting the information on all the chicken genes available on Ensembl ####  
  
ensembl = useMart("ensembl",dataset="ggallus_gene_ensembl")  
attribute_list = c("ensembl_gene_id","chromosome_name","strand",  
                  "start_position","end_position","gene_biotype","external_gene_name")  
gene_info = getBM(attributes=attribute_list, mart = ensembl)  
  
gene_info = data.table(gene_info)  
gene_info$chromosome_name = as.character(gene_info$chromosome_name)  
  
setkey(gene_info,chromosome_name,start_position,end_position)  
  
df = data.table(df)  
df$Chromosome = as.character(df$Chromosome)  
  
##### A left and right boundy around every SNPs can be attached  
  
df$Pos.left <- df$Position #### position of the SNPs on the chromosome  
df$Pos.right <- df$Position  
  
setkey(df,Chromosome, Pos.left, Pos.right)  
  
Finalgenes = foverlaps(gene_info,df) %>% na.omit
```