

# **Investigation of the Regulatory Processes of Gene Expression in Animal and Plant Sciences**

Dissertation

zur Erlangung des Doktorgrades  
"Doctor rerum naturalium"  
der Georg-August-Universität Göttingen

vorgelegt von

Selina Sarah Wilhelmi (geb. Klees)  
aus Oldenburg, Deutschland

Göttingen  
Dezember 2022

Supervisory committee

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

Prof. Dr. Klaus Jung,  
Genomics and Bioinformatics of Infectious Diseases, Institute for Animal Breeding and  
Genetics, University of Veterinary Medicine Hannover

Prof. Dr. Mehmet Gültas,  
Statistics and Data Science, Faculty of Agriculture, South Westphalia University of  
Applied Sciences Soest

Date of thesis defense: 16.03.2023

## Abstract

Transcriptional regulation of gene expression in higher organisms is fundamental for numerous biological processes. These processes are mainly controlled by a special class of regulatory proteins, the transcription factors (TFs), and their combinatorial interplay. Various genetic programs, such as environmental adaptation, tissue development, or disease control, are governed by the binding of TFs to short DNA motifs, called transcription factor binding sites (TFBS), in the regulatory regions of their target genes. Single nucleotide polymorphisms (SNPs) located in promoter regions can alter TFBSs leading to a change in the binding affinity of TFs and, thus, affect gene expression. Such SNPs are referred to as regulatory SNPs (rSNPs).

In recent years, rSNPs have come into the focus of research, and the underlying mechanisms resulting in a differential gene expression have been studied for many specific traits and diseases mainly in humans or model organisms, but also in agricultural species. However, these studies mostly concentrate on single regulatory variants and do not include systematic analyses. Thus, there is still a lack of such comprehensive analyses and genome-wide collections of rSNPs, and to date, only few tools and databases are available for livestock or crop species.

In this work, I developed a pipeline for the detection of rSNPs and created the databases agReg-SNPdb and agReg-SNPdb-Plants, storing genome-wide collections of rSNPs and their predicted effects on TF binding for agricultural animal and plant species, respectively. agReg-SNPdb includes seven livestock and domestic species, namely cattle, pig, chicken, sheep, horse, goat, and dog and agReg-SNPdb-Plants includes 13 crop species and sub-species, namely African rice, Asian rice (with its subspecies *Indica* and *Japonica*), barley, bread wheat, durum wheat, grape, maize, rapeseed, sorghum, sunflower, tomato, and wild rice.

Out of all species stored in agReg-SNPdb-Plants, rapeseed holds a special role. In contrast to the remaining species, where I used the data from Ensembl Plants as basis, in rapeseed, to date, there is no genome-wide collection of SNPs available. Therefore, I used a previously published data set based on different resequenced *Brassica napus* L. cultivars for the identification of rSNPs in agReg-SNPdb-Plants.

Based on this data set, I investigated the regulatory mechanisms in two cultivars, namely Zhongshuang11 (ZS11), a so-called double-low accession with low content of erucic acid and glucosinolate, which is characterized by high oil content, and Zhongyou821 (ZY821), a so-called double-high accession with high content of erucic acid and glucosinolate, which is characterized by low oil content. In this way, I demonstrate the application of rSNPs together with multi-omics data to perform a systematic analysis of the complex interplay between rSNPs, TFs, and differentially expressed genes (DEGs) in four tissues (flower, leaf, stem, and root) which are underlying the oil content and -quality in rapeseed.

Finally, I present a project in which I investigated the transcriptional gene regulation in chicken and duck following an infection with avian influenza. To date, the regulatory mechanisms underlying the susceptibility of chicken to avian influenza and the effective immune response of duck have not been fully deciphered. To address the limited knowledge regarding upstream regulators, I identified TFs, their cooperations and master regulators that may be important in triggering an effective differential gene expression in chicken to control the virus.

Overall, to the best of my knowledge, this work provides the first databases of rSNPs and their predicted consequences on TF binding in animal and plant species of agricultural importance. By making the databases accessible via a website, I enable scientists to interpret and evaluate their results from genome-wide association studies, gene expression experiments, or a combination of both to uncover mechanisms underlying a trait of interest. In the two application projects, I obtained novel insights into regulatory mechanisms underlying (i) the oil content and -quality of rapeseed and (ii) avian influenza virus control of chicken and duck and, thus, I could provide novel research objectives for future studies.



## Zusammenfassung

Die transkriptionelle Regulation der Genexpression in höheren Organismen ist für zahlreiche biologische Prozesse von grundlegender Bedeutung. Diese Prozesse werden hauptsächlich durch eine spezielle Klasse von regulatorischen Proteinen, den Transkriptionsfaktoren (TFs), und deren kombinatorischem Zusammenspiel gesteuert. Verschiedene genetische Programme, wie die Anpassung an Umweltbedingungen, die Entwicklung von Geweben oder die Kontrolle von Krankheiten, werden durch die Bindung von TFs an kurze DNA-Motive, so genannte Transkriptionsfaktorbindestellen (TFBS), in den regulatorischen Regionen ihrer Zielgene gesteuert. Einzelnukleotid-Polymorphismen (SNPs, engl. 'single nucleotide polymorphisms') in Promotorregionen können TFBSs verändern, was zu einer Änderung der Bindungsaffinität von TFs führt und somit die Genexpression beeinflusst. Solche SNPs werden als regulatorische SNPs (rSNPs) bezeichnet. In den letzten Jahren rückten rSNPs in den Mittelpunkt der Forschung, und die zugrunde liegenden Mechanismen, die zu einer differentiellen Genexpression führen, wurden für viele spezifische Merkmale und Krankheiten hauptsächlich bei Menschen oder Modellorganismen, aber auch bei landwirtschaftlichen Arten untersucht. Diese Studien konzentrierten sich jedoch meist nur auf einzelne regulatorische Varianten und umfassen keine systematischen Analysen. Daher fehlen bis heute solche umfassenden Analysen und genomweite Kollektionen von rSNPs, und es sind nur wenige Tools und Datenbanken für Nutztiere oder Nutzpflanzen verfügbar.

In dieser Arbeit habe ich eine Pipeline für die Erkennung von rSNPs entwickelt und die Datenbanken agReg-SNPdb und agReg-SNPdb-Plants erstellt, welche genomweite Sammlungen von rSNPs und deren vorhergesagte Auswirkungen auf die TF-Bindung für landwirtschaftliche Tier- bzw. Pflanzenarten enthalten. agReg-SNPdb umfasst die sieben Nutz- und Haustierarten Rind, Schwein, Huhn, Schaf, Pferd, Ziege und Hund, und agReg-SNPdb-Plants umfasst die 13 Pflanzenarten und -unterarten Afrikanischen Reis, Asiatischen Reis (mit den Unterarten Indica und Japonica), Brotweizen, Gerste, Hartweizen, Mais, Raps, Sorghum, Sonnenblumen, Tomaten, Weintrauben und Wildreis. Von allen in agReg-SNPdb-Plants gespeicherten Spezies nimmt Raps eine Sonderrolle ein. Im Gegensatz zu den übrigen Spezies, bei denen ich Daten aus Ensembl Plants als Grundlage verwendet habe, gibt es für Raps bisher keine genomweite Sammlung von SNPs. Daher verwendete ich einen bereits veröffentlichten Datensatz, der auf resequenzierten *Brassica napus* L.-Sorten basiert, für die Identifizierung von rSNPs in agReg-SNPdb-Plants.

Auf der Grundlage dieses Datensatzes untersuchte ich die regulatorischen Mechanismen in zwei Sorten, nämlich, Zhongshuang11 (ZS11), einer so genannten Doppelnull Sorte mit geringem Gehalt an Erucasäure und Glucosinolat, welche durch einen hohen Ölgehalt charakterisiert ist, und Zhongyou821 (ZY821), einer so genannten Doppelplus Sorte mit hohem

Gehalt an Erucasäure und Glucosinolat, welche durch einen niedrigen Ölgehalt charakterisiert ist. Dadurch demonstriere ich die Anwendung von rSNPs zusammen mit Multi-Omics-Daten, um eine systematische Analyse des komplexen Zusammenspiels zwischen rSNPs, TFs und differenziell exprimierten Genen (DEGs) in vier Geweben (Blüte, Blatt, Stamm und Wurzel) durchzuführen, die dem Ölgehalt und der Ölqualität von Raps zugrunde liegen.

Schließlich stelle ich ein Projekt vor, in dem ich die transkriptionelle Genregulation bei Hühnern und Enten nach einer Infektion mit der Vogelgrippe untersucht habe. Bis heute sind die Mechanismen, die die Anfälligkeit von Hühnern für die Vogelgrippe und die wirksame Immunantwort von Enten regulieren, noch nicht vollständig entschlüsselt. Um den Forschungsbedarf in Bezug auf upstream-Regulatoren auszugleichen, habe ich TFs und ihre Kooperationen und Master-Regulatoren identifiziert, die für die Aktivierung einer effektiven differentiellen Genexpression bei Hühnern zur Bekämpfung des Vogelgrippevirus von Bedeutung sein könnten.

Insgesamt bietet diese Arbeit meines Wissens die ersten Datenbanken über rSNPs und ihre Auswirkungen auf die TF-Bindung bei Tier- und Pflanzenarten von landwirtschaftlicher Bedeutung. Indem die Datenbanken über eine Webseite zugänglich sind, haben Wissenschaftler\*innen die Möglichkeit, ihre Ergebnisse aus genomweiten Assoziationsstudien, Genexpressionsexperimenten oder einer Kombination aus beiden zu interpretieren und zu bewerten, um Mechanismen aufzudecken, die einem Merkmal von Interesse zugrunde liegen. In den beiden Anwendungsprojekten habe ich neue Erkenntnisse über die Regulationsmechanismen gewonnen, die (i) dem Ölgehalt und der Qualität von Raps und (ii) der Kontrolle des Vogelgrippevirus bei Hühnern und Enten zugrunde liegen und die neue Forschungsziele für künftige Studien bieten könnten.

## Danksagung

Zunächst danke ich den Mitgliedern meines Betreuungsausschusses für ihre Unterstützung, ihren Rat und die aufschlussreichen Diskussionen. Ich danke insbesondere Prof. Dr. Armin Schmitt für die erstklassige Betreuung und Unterstützung während meiner gesamten Zeit in der Gruppe für Züchtungsinformatik. Ich hatte Glück, einen so hilfsbereiten Doktorvater zu haben, mit dem ich jederzeit über Ideen und Probleme sprechen konnte. Er hat mit der Züchtungsinformatik eine Gruppe geschaffen, die ein ausgezeichnetes wissenschaftliches Umfeld bietet, in dem neue Ideen entstehen und umgesetzt werden können.

Ich danke Prof. Dr. Klaus Jung, nicht nur als Mitglied meines Prüfungskomitees, sondern auch für die bereichernden Gespräche und Komitee-Treffen, aus denen ich immer mit neuen Ideen und voller Motivation ging.

Mein besonderer Dank gilt Prof. Dr. Mehmet Gültas, dem durch seine Unterstützung und Motivation ein erheblicher Teil dieser Arbeit gebührt. Ohne seine Anregungen, seine wertvollen Erfahrungen, sein Ideenreichtum, und vor allem die Möglichkeit jederzeit über alles reden zu können, wäre ich nicht so weit gekommen.

Der gesamten Gruppe der Züchtungsinformatik möchte ich meinen Dank für viel Hilfsbereitschaft, Offenheit und Spaß aussprechen. Ich bin sehr froh so nette Kollegen und Freunde zu haben, die mir durch zahlreiche "Strategic meetings", Grillabende und Mittagspausen immer zu neuen Motivationsschüben verholfen haben.

Danke an Monika Siebert für die Organisation und Hilfsbereitschaft bei administrativen Aufgaben und für die Versorgung mit Snacks und Schokoladennikoläusen; an Dr. Felix Heinrich für das Lösen all meiner Computerprobleme, die ständige Bereitschaft und Hilfe bei Installationen und die gnadenlos ehrlichen Kommentare zu Manuskripten und Vorträgen; an Dr. Abirami Rajavel für die sehr wertvolle Hilfe bei der Interpretation von biologischen Ergebnissen; an Prof. Dr. Faisal Ramzan, mein Bürokollege während der ersten Jahre, mit dem ich zusammen mein erstes Paper schreiben durfte; an Martin Wutke für die vielen Diskussionen, die mir andere Sichtweisen eröffnet haben, sei es im wissenschaftlichen oder philosophischen Sinne; an Thomas Martin Lange vor allem für die letzten Wochen, in denen wir "Leidensgenossen" waren und uns gegenseitig zum Schreiben motiviert haben; an Johanna-Sophie Schlüter für die gute Zusammenarbeit beim Raps-Projekt und die vielen tollen Gespräche über Pflanzen und Essen; an Hendrik Bertram für die sehr wertvolle Arbeit und Hilfe mit verschiedenen Tools und Analysen; an Ata ul Haleem für die Zusammenarbeit in seinem Mais-Paper; und Danke an Dr. Yonatan Mekonnen, Dr. Sebastian Zeidler, Dr. Rita Tonin, Maria Rotärmel und Antje Christine Kurzweg für die tolle Zeit in der Züchtungsinformatik.

Es hat mir großen Spaß bereitet, verschiedene Bachelor- und Masterarbeiten sowie Forschungsprojekte zu unterstützen und ich habe dabei sehr viel gelernt. Vielen Dank an Hendrik Bertram, Antje Christine Kurzweg, Jendrik Schellhorn und Johanna-Sophie Schlüter.

Ich möchte auch dem gesamten CiCom-Team danken. Besonders die Zeit nach den Online-Treffen hat gezeigt, dass CiCom eine echte Bereicherung nicht nur für CiBreed ist, sondern auch um Kooperationen und Freundschaften zu schaffen.

Schließlich möchte ich mich bei meiner Familie, meinen Freunden und vor allem bei meinem Mann Tim bedanken, die mich bedingungslos unterstützt sowie meine Arbeiten Korrektur gelesen haben und immer für mich da waren. Danke!

# Contents

<b>1. Preface</b>	<b>1</b>
1.1. Impact . . . . .	1
1.2. Structure of the thesis . . . . .	4
<b>2. General Introduction</b>	<b>7</b>
2.1. Systems Biology and Omics . . . . .	8
2.1.1. Genomics – the DNA Carries the Genetic Information . . . . .	8
2.1.2. Transcriptomics – the RNA Transmits the Genetic Information . . . . .	9
2.1.3. Proteomics – Proteins Form the Diversity of the Cell . . . . .	10
2.2. Single Nucleotide Polymorphisms . . . . .	12
2.2.1. Regulatory SNPs . . . . .	13
2.3. Application Projects . . . . .	15
2.3.1. Oil Content and Quality in Rapeseed . . . . .	15
2.3.2. Avian Influenza in Chicken and Duck . . . . .	16
<b>3. agReg-SNPdb</b>	<b>19</b>
3.1. Simple Summary . . . . .	19
3.2. Abstract . . . . .	20
3.3. Introduction . . . . .	20
3.4. Materials and Methods . . . . .	25
3.4.1. Input Data . . . . .	25
3.4.2. Pipeline . . . . .	25
3.5. Results . . . . .	27
3.5.1. Database . . . . .	27
3.5.2. Web Interface . . . . .	27
3.5.3. Statistical Analysis of the Data . . . . .	29
3.6. Biological Validation Based on Case-Studies . . . . .	34
3.6.1. Milk Protein and Fat Content in Dairy Cattle . . . . .	34
3.6.2. Fat-Related Beef Quality Traits in Cattle . . . . .	34
3.6.3. Chicken Egg Production . . . . .	35
3.6.4. Fatty-Acid Composition Related Traits in Pigs . . . . .	36
3.7. Discussion . . . . .	36
3.8. Conclusions . . . . .	38

---

3.9. Supplementary Materials . . . . .	38
<b>4. agReg-SNPdb-Plants</b>	<b>39</b>
4.1. Simple Summary . . . . .	39
4.2. Abstract . . . . .	40
4.3. Introduction . . . . .	40
4.4. Materials and Methods . . . . .	42
4.5. Results . . . . .	43
4.5.1. Database . . . . .	43
4.5.2. Web Interface . . . . .	43
4.5.3. Statistical Overview of the Data . . . . .	45
4.6. Discussion . . . . .	47
4.7. Conclusions . . . . .	48
4.8. Supplementary Materials . . . . .	49
<b>5. Oil Content and Quality in Rapeseed</b>	<b>51</b>
5.1. Abstract . . . . .	52
5.2. Introduction . . . . .	52
5.3. Results and Discussion . . . . .	54
5.3.1. Differentially Expressed Genes . . . . .	54
5.3.2. Transcription Factor Binding Site Enrichment Analysis . . . . .	55
5.3.3. Analysis of Regulatory SNPs . . . . .	57
5.3.4. Analysis of Important Regulatory SNPs . . . . .	58
5.3.5. DEGs Harboring <i>Important rSNPs</i> in the Promoter Region . . . . .	58
5.4. Materials and Methods . . . . .	61
5.4.1. <i>B. napus</i> Data Set and Data Preparation . . . . .	62
5.4.2. Transcription Factor Binding Site Enrichment Analysis in Promoter Sequences . . . . .	63
5.4.3. Identification of Regulatory SNPs and Their Importance . . . . .	63
5.4.4. Association Analysis Using Random Forests . . . . .	64
5.5. Conclusions . . . . .	64
5.6. Supplementary Materials . . . . .	66
<b>6. Avian Influenza in Chicken and Duck</b>	<b>67</b>
6.1. Simple Summary . . . . .	67
6.2. Abstract . . . . .	68
6.3. Introduction . . . . .	68
6.4. Materials and Methods . . . . .	70
6.4.1. Transcriptome Data . . . . .	70
6.4.2. Identification of Enriched TFs and TF-TF Cooperations . . . . .	72
6.4.3. Identification of Master Regulators . . . . .	73

6.4.4. Annotations and Ortholog Mapping . . . . .	73
6.5. Results and Discussion . . . . .	73
6.5.1. Transcription Factor Binding Site Enrichment . . . . .	75
6.5.2. TF-TF Cooperations . . . . .	77
6.5.3. Master Regulators . . . . .	82
6.6. Conclusions . . . . .	85
6.7. Supplementary Materials . . . . .	85
<b>7. Discussion</b>	<b>87</b>
7.1. Methodical Discussion . . . . .	87
7.1.1. Identification of rSNPs . . . . .	87
7.1.2. TFBS Prediction . . . . .	89
7.1.3. Random Forest-Based Feature Selection to Identify SNP-Phenotype Associations . . . . .	90
7.1.4. Upstream Analysis to Identify Master Regulators . . . . .	91
7.2. Biological Discussion . . . . .	93
7.2.1. Regulatory Impact of Transcription Factors and rSNPs . . . . .	93
7.2.2. The Distribution of rSNPs and SNPs around the TSS . . . . .	94
7.2.3. Oil Content and Quality in Rapeseed . . . . .	94
7.2.4. Avian Influenza in Chicken and Duck . . . . .	97
<b>8. Conclusion</b>	<b>99</b>
<b>9. List of Abbreviations</b>	<b>101</b>
<b>A. Appendix</b>	<b>127</b>
A.1. Curriculum vitae . . . . .	127
A.2. Erklärung . . . . .	131

## List of Figures

1.1.	Graphical abstract of the study described in Chapter 3. . . . .	4
1.2.	Graphical abstract of the study described in Chapter 4. . . . .	5
1.3.	Graphical abstract of the study described in Chapter 5. . . . .	5
1.4.	Graphical abstract of the study described in Chapter 6. . . . .	6
2.1.	Scheme of the central dogma of molecular biology. . . . .	7
2.2.	Structure of the gene coding region and the promoter. . . . .	9
2.3.	PWM and Sequence Logo of the TFBS for MEF2. . . . .	12
3.1.	Scheme of the disruption of TF binding due to an rSNP. . . . .	21
3.2.	Scheme of the workflow applied for the detection of rSNPs. . . . .	23
3.3.	Search page of agReg-SNPdb. . . . .	28
3.4.	Example of a search result from agReg-SNPdb. . . . .	29
3.5.	The total number of SNPs and genes for each chromosome of chicken. . . .	30
3.6.	The average number of rSNPs in promoter regions per gene for each chromosome of chicken, divided into upstream and downstream promoters. . . .	31
3.7.	Distribution of the distances between rSNPs and the TSS of chicken. . . . .	32
3.8.	Distribution of the distances between rSNPs and the TSS of cattle. . . . .	33
4.1.	Example of a search result from agReg-SNPdb-Plants showing table <i>TFBS_results</i> . . . . .	45
4.2.	The total number of SNPs and genes per chromosome of maize ( <i>Zea mays</i> ). . . .	46
4.3.	Distribution of the distances between rSNPs and the TSS of (A) Asian rice Japonica and (B) Asian rice Indica. . . . .	47
5.1.	Venn diagram for the enriched TFs found for the tissues flower, leaf, stem, and root of <i>B. napus</i> . . . . .	55
5.2.	Overlap of the DEGs in (A) and rSNPs in (B) for the four investigated tissues. . . .	57
5.3.	Distribution of rSNPs relative to the TSS of the corresponding genes. . . . .	58
5.4.	Distribution of <i>important rSNPs</i> relative to the TSS of the corresponding genes. . . . .	59
5.5.	Flowchart of the analysis applied in this study. . . . .	61
5.6.	Pseudo-code for the Boruta algorithm. . . . .	65



---

6.1. Flow chart of the employed analyses. . . . .	71
6.2. Venn diagrams of the DEGs ( <b>A</b> ) duck in ileum, ( <b>B</b> ) chicken in ileum, ( <b>C</b> ) duck in lung, and ( <b>D</b> ) chicken in lung with selected enriched GO terms. . . . .	79
6.3. Venn diagrams of TFBS enrichment to compare over- (OR) and underrepresented (UR) binding sites in chicken and duck. . . . .	80
6.4. Differences in TF cooperation networks found by the PC-TraFF algorithm for ( <b>A</b> ) ileum 1 dpi, ( <b>B</b> ) ileum 3 dpi, ( <b>C</b> ) lung 1 dpi and ( <b>D</b> ) lung 3 dpi with HPAIV H5N1. . . . .	81
6.5. Common and species-specific master regulators for the ( <b>A</b> ) ileum and ( <b>B</b> ) lung tissue regulating the TF-TF cooperations. . . . .	83
7.1. Distribution of rSNPs and all SNPs around the TSS of barley. . . . .	95
7.2. Distribution of rSNPs and SNPs around the TSS of cattle. . . . .	96

## List of Tables

3.1.	A summary of five recent studies that systematically investigated the effects of SNPs on regulatory elements such as TFBSs. . . . .	24
3.2.	Assembly versions of the input data, including the reference genome, SNP catalog, and gene annotations. . . . .	25
3.3.	The number of records stored in the database tables <i>snp_info</i> , <i>gene_info</i> , <i>snp_region</i> , and <i>TFBS_results</i> . . . . .	27
3.4.	Consequences of SNP rs41255679 (C/G), located upstream of the TSS of the bovine <i>LGB</i> gene. . . . .	34
3.5.	Consequences of the SNPs rs110055647 and rs109682576 in the bovine <i>FABP4</i> upstream promoter with a T to C conversion. . . . .	35
3.6.	Consequences of the SNP rs333406887 (C/G) located -238 bp from the porcine <i>APOA2</i> TSS. . . . .	36
4.1.	Assembly versions of the input data from Ensembl Plants including reference genome, SNP catalog and gene annotations. . . . .	43
4.2.	The number of records stored in the database tables <i>snp_info</i> , <i>gene_info</i> , <i>snp_region</i> , and <i>TFBS_results</i> separated by species. . . . .	44
5.1.	Numbers of DEGs in four tissues based on the comparison of the cultivars Zhongshuang11 (ZS11) against Zhongyou821 (ZY821). . . . .	54
5.2.	Meta data of the RNA-seq experiment samples which were used for differential expression analysis. . . . .	63
6.1.	Numbers of DEGs in duck and chicken for the treatments with H5N1 (HPAI) and H5N2 (LPAI) virus after 1 and 3 dpi. . . . .	74





# 1. Preface

## 1.1. Impact

### Journal articles:

I have published the following articles, which form the basis of this thesis:

- [1] **Klees, S.**, Heinrich, F., Schmitt, A. O. & Gültas, M. (2022). agReg-SNPdb-Plants: A Database of Regulatory SNPs for Agricultural Plant Species. *Biology*, 11(5), 684.
- [2] **Klees, S.**, Schlüter, J. S., Schellhorn, J., Bertram, H., Kurzweg, A. C., Ramzan, F., Schmitt, A. O. & Gültas, M. (2022). Comparative Investigation of Gene Regulatory Processes Underlying Avian Influenza Viruses in Chicken and Duck. *Biology*, 11(2), 219.
- [3] **Klees, S.\***, Heinrich, F.\*, Schmitt, A. O. & Gültas, M. (2021). agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species. *Biology*, 10(8), 790. (\* These authors contributed equally to this work.)
- [4] **Klees, S.**, Lange, T. M., Bertram, H., Rajavel, A., Schlüter, J. S., Lu, K., Schmitt, A. O. & Gültas, M. (2021). In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in *Brassica napus* L. Using Multi-Omics Data. *International Journal of Molecular Sciences*, 22(2), 789.

Detailed author contribution of Selina Wilhelmi (née Klees) to the above mentioned journal articles: Participated in the design of the studies. Prepared the data sets. Conducted computational, statistical, and bioinformatics analyses. Created and developed the databases, pipelines and websites. Interpreted the results. Performed the biological validation. Wrote the final versions of the manuscripts.

Further, I am a co-author of the following articles related to the topics of my thesis:

- [1] Rajavel, A., **Klees, S.**, Hui, Y., Schmitt, A. O. & Gültas, M. (2022). Deciphering the Molecular Mechanism Underlying African Animal Trypanosomiasis by Means of the 1000 Bull Genomes Project Genomic Dataset. *Biology*, 11(5), 742.
- [2] Haleem, A., **Klees, S.**, Schmitt, A. O. & Gültas, M. (2022). Deciphering Pleiotropic Signatures of Regulatory SNPs in *Zea mays* L. Using Multi-Omics Data and Machine Learning Algorithms. *International Journal of Molecular Sciences*, 23(9), 5121.

- [3] Rajavel, A., **Klees, S.**, Schlüter, J. S., Bertram, H., Lu, K., Schmitt, A. O. & Gültas, M. (2021). Unravelling the Complex Interplay of Transcription Factors Orchestrating Seed Oil Content in *Brassica napus* L. *International Journal of Molecular Sciences*, 22(3), 1033.
- [4] Ramzan, F.\*, **Klees, S.\***, Schmitt, A. O., Cavero, D. & 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.)

### Conferences, workshops, and poster presentations

- CiBreed Fall Workshop, October 13-14, 2022, Göttingen: As a part of the organization team, I was involved in the preparation and realization of the workshop.
- DGfZ-/GfT-Jahrestagung, September 21-22, 2022, Kiel: Oral presentation entitled "Untersuchung der transkriptionellen Genregulation während einer aviären Influenza bei Ente und Huhn"
- German Conference on Bioinformatics (GCB), September 6-8, 2022, Halle (Saale): Oral presentation entitled "Analysis of regulatory SNPs with agReg-SNPdb-Plants and its application to oil content and -quality of rapeseed (*Brassica napus* L.)"
- CiBreed Fall Workshop, October 14-15, 2021, Göttingen (online): Oral presentation entitled "In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in *Brassica napus* L. Using Multi-Omics Data"
- German Conference on Bioinformatics (GCB), September 14-17, 2020, Frankfurt am Main (online)
- CiBreed Fall Workshop, September 29, 2020, Göttingen (online): As a part of the organization team, I was involved in the preparation and realization of the workshop.
- CiBreed Fall Workshop September 9-10, 2019, Göttingen: Poster presentation entitled "Regulatory SNPs and their importance in animal and plant breeding" and winning the "People's Choice Award".

### Project works and student's theses

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

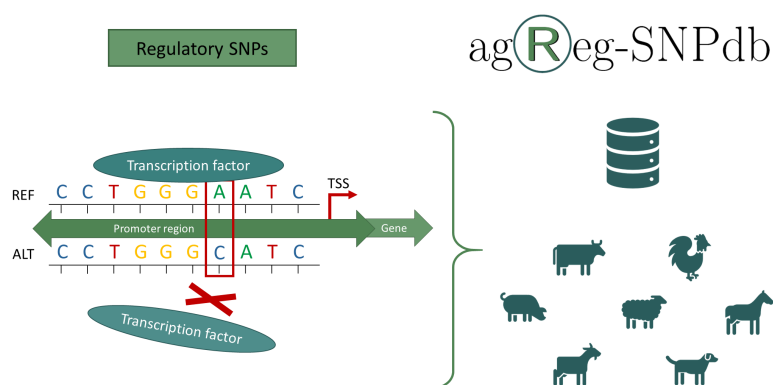
- Master's thesis: "Comparative Investigation of Coincident SNPs Underlying Avian Influenza Viruses in Chicken and Duck" (Hendrik Bertram, submission date: 20 October 2022)

- Research project Applied bioinformatics with R: "Comparative Investigation of Avian Influenza Viruses in Chicken and Duck" (Johanna-Sophie Schlüter, Jendrik Schellhorn, and Hendrik Bertram, submission date: 31 March 2021)
- Bachelor's thesis: "A multi-omics study to identify regulatory SNPs in *Brassica napus* L. with tissue specific effects on the binding ability of transcription factors" (Johanna-Sophie Schlüter, submission date: 27 January 2021)
- Bachelor's thesis: "Untersuchung der transkriptionellen Genregulation in der Lunge während einer aviären Influenza bei Ente und Huhn" (Antje Christine Kurzweg, submission date: 2 November 2020)
- Bachelor's thesis: "Vorhersage von regulatorischen SNPs und deren Einfluss auf die Bindeaffinität von TFs in Pflanzen" (Hendrik Bertram, submission date: 25 April 2020)
- Research project Biometrie mit R: "Analyse von RNA-Sequenzierungsdaten von Rennpferden" (Antje Christine Kurzweg, submission date: 13 March 2020)

## 1.2. Structure of the thesis

This thesis is structured as follows. In Chapter 2, I give an overview of the molecular processes of living organisms which are required as background and introduction for this thesis. I start with systems biology and the concept of omics technologies and provide a more detailed overview of the three main omics disciplines genomics, transcriptomics and proteomics. Then, I focus on single nucleotide polymorphisms (SNPs) and a special kind of them, the regulatory SNPs. Lastly, I introduce the two application projects of my thesis, the oil content and -quality based on the oil crop *Brassica napus* L. and the immune response of chicken and duck after an infection with avian influenza. In the following four chapters, I provide my publications relevant for this thesis [1–4].

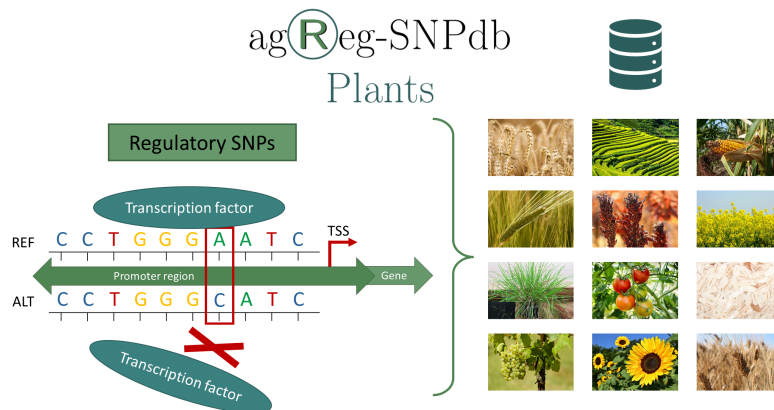
In Chapter 3 I present my study "agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species" [3], where I developed a pipeline to predict rSNPs, applied it to seven agricultural and domestic animal species, namely cattle, pig, chicken, sheep, horse, goat, and dog, and stored genome-wide collections of rSNPs and their effects on TF binding in the database agReg-SNPdb. In this study, I performed a literature survey to show that the obtained results are in agreement with previous experimental and *in silico* studies. In order to ensure a convenient database search, I have developed a website to query agReg-SNPdb by SNP IDs, chromosomal regions, or genes. The graphical abstract of this study is shown in Figure 1.1.



**Figure 1.1.: Graphical abstract for the study described in Chapter 3.**

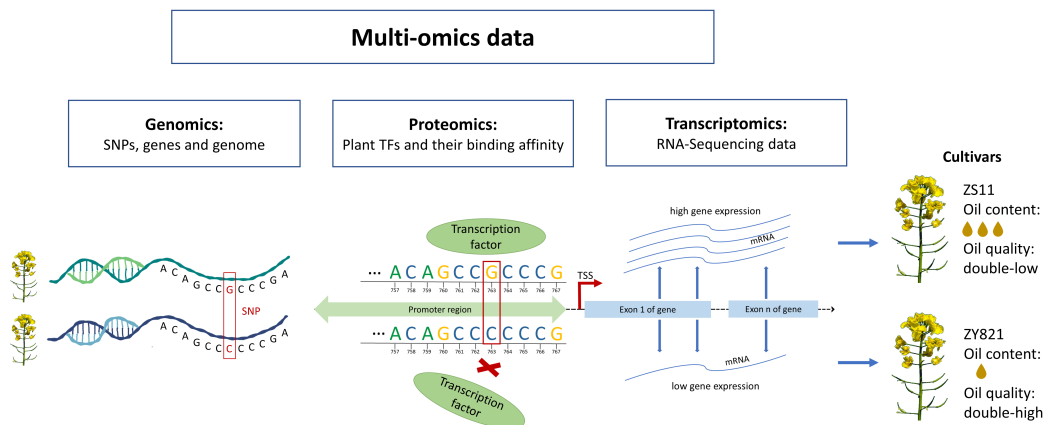
The study described in Chapter 4 can be considered as an extension of agReg-SNPdb. I have developed agReg-SNPdb-Plants, a database of regulatory SNPs for agriculturally important plant species and subspecies (African rice, Asian rice (Indica and Japonica), barley, bread wheat, durum wheat, grape, maize, rapeseed, sorghum, sunflower, tomato, and wild rice) [1]. The graphical abstract of this study is shown in Figure 1.2.





**Figure 1.2.:** Graphical abstract of the study described in Chapter 4.

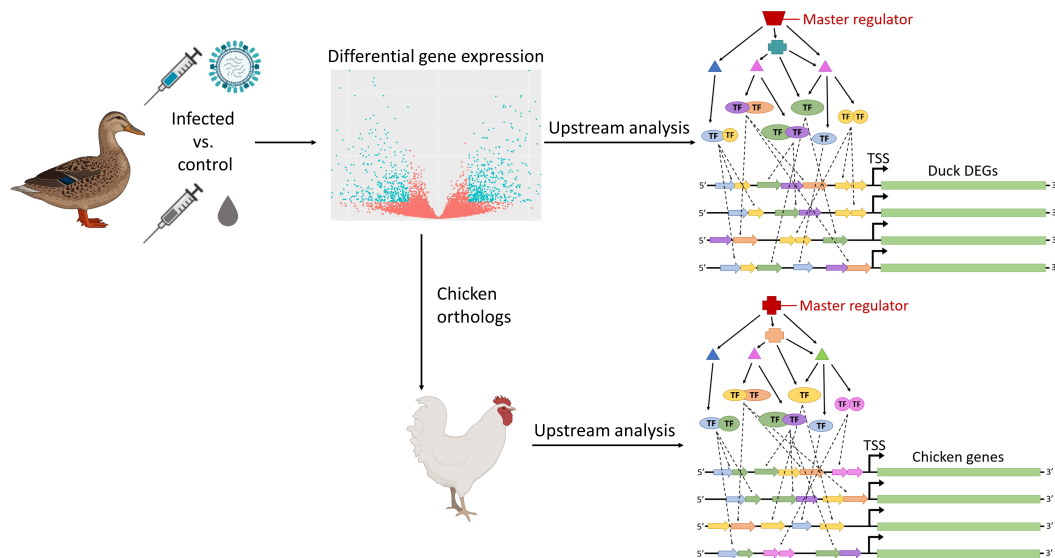
In Chapter 5, I present the first application study. In this study of different *B. napus* cultivars, I demonstrate the application of rSNPs together with multi-omics data to perform a systematic analysis of the complex interplay between rSNPs, TFs, and differentially expressed genes (DEGs) in vegetative and floral tissues underlying rapeseed oil content and -quality [4] (see Figure 1.3).



**Figure 1.3.:** Graphical abstract of the study described in Chapter 5.

In Chapter 6, I present a study in which I investigated the transcriptional gene regulation controlling the expression of genes induced by an infection with avian influenza in chicken

and duck [2]. This uncovered master regulators that could stimulate an effective immune response in ducks following viral infection, while being dysfunctional in chicken. The graphical abstract of this study is shown in Figure 1.4.

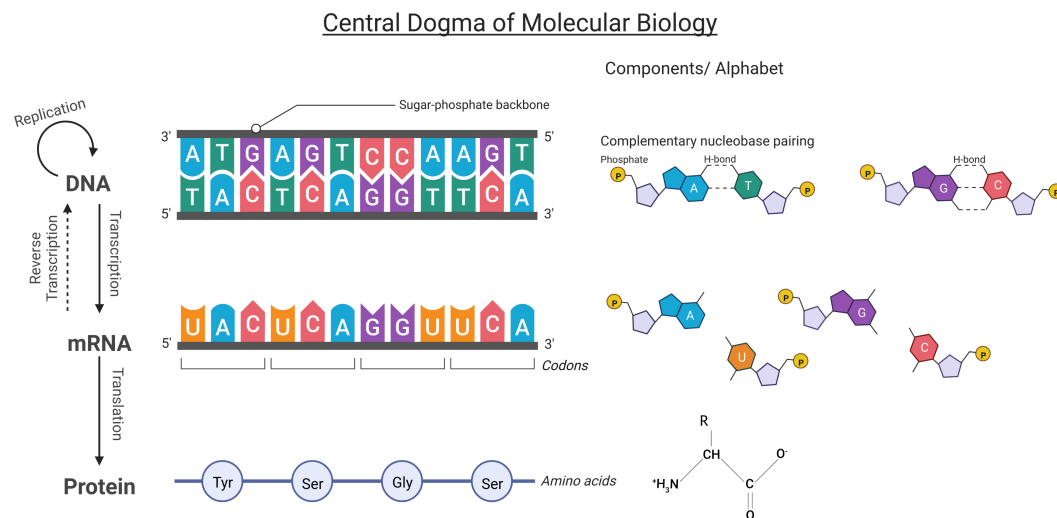


**Figure 1.4.: Graphical abstract of the study described in Chapter 6.**

The discussion (Chapter 7) is divided into two parts, a methodological discussion and a biological discussion. In the first part, the methods used in this thesis, such as the rSNP prediction pipeline or TFBS prediction, are discussed and compared with other existing methods. In the second part, I focus on the interpretation of the biological results and potential future experiments to support them. Finally, in Chapter 8, I conclude the thesis and provide an outlook for future work.

## 2. General Introduction

Within a living cell, the universal basis for the flow of genetic information is described by the so-called central dogma of molecular biology. In its basic form, it describes the process of protein synthesis from deoxyribonucleic acid (DNA), which is transcribed into messenger ribonucleic acid (mRNA) followed by the translation of mRNA into proteins (Figure 2.1). Introduced as early as 1958 by Francis Crick [5], the dogma still holds true today and lays the foundation of modern biology. Since then, it has been modified and refined by new discoveries such as the reverse transcriptase, splicing, epigenetic modifications or chaperones for protein folding. Today, new technologies enable the generation of large sets of experimental and sequencing data to study living organisms [6].



**Figure 2.1.: Scheme of the central dogma of molecular biology.** The flow of information from DNA via RNA to proteins is indicated by arrows, with dashed lines indicating rare events. On the right, the molecular components of DNA, mRNA, and proteins are shown schematically, including base pairing in the case of DNA. The figure was created with BioRender (<https://biorender.com/>).

## 2.1. Systems Biology and Omics

In biological studies, it has always been advantageous to consider a living system not only as the sum of its components, but to have a holistic view and see it in its entirety [7]. This complex view of biological systems, encompassing all components and their interactions and regulations, describes the concept of systems biology and has given rise to the suffix of "-omics". Omics can be described as different disciplines or biological entities, such as genomics, transcriptomics, proteomics, or metabolomics, with each of it being composed of a variety of different regulatory mechanisms that are in constant interplay with each other [7, 8]. Consequently, to study a system's biology, it is necessary to systematically determine the components (e.g., DNA, RNA, and proteins) and to assemble and interpret their interactions and regulations in order to obtain knowledge about the system as a whole [8]. In the following, I will address three of the main omics technologies, i.e., genomics, transcriptomics, and proteomics, as these are the technologies studied in this thesis. I will provide a definition, an overview of the structural properties, and the corresponding technologies as well as the data used to study them.

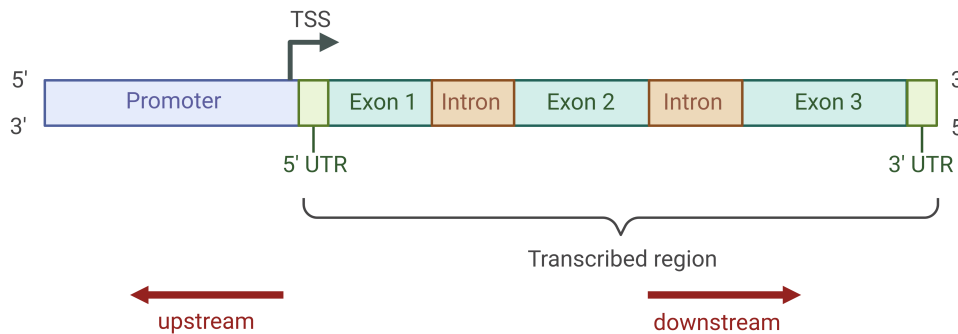
### 2.1.1. Genomics – the DNA Carries the Genetic Information

The genome of an individual is defined as the totality of an organism's genetic material, i.e., DNA, including its genes. The key to genomics studies is therefore the determination and decoding of the DNA sequence, the sequence of letters in a four-digit alphabet. On the molecular level, this alphabet is determined by the four nucleic bases adenine (A), guanine (G), thymine (T), and cytosine (C), which are attached to a sugar-phosphate backbone to build a directed DNA chain. The polarity of the chains is determined based on the phosphodiester linkage between the phosphate and the sugar, where each sugar molecule has one phosphate linked to its 3' and one to its 5' carbon, resulting in one 3' and one 5' end of the strand [9, 10]. The DNA double-helix is then built by two strands wound around each other in antiparallel direction, stabilized by hydrogen bonds between the nucleic bases in the middle of the helix, with A and T as well as C and G paired. This feature leads to the fact that both strands are complementary to each other and carry the same information [10, 11].

Today, it is possible to sequence an individual's entire genome within hours at an ever decreasing cost using next-generation sequencing (NGS). In this way, both coding and non-coding regions are taken into account, and all types of genomic variations between different genomes, such as single nucleotide polymorphisms, copy number variants, insertions, or deletions, can be detected.

One major field of study within genomics is the study of genes, those sections on the DNA coding for the synthesis of the gene product, either RNA or protein. In complex organisms, each gene has a specific start position, the transcription start site (TSS), followed by one or several coding sequences (exons), possibly interspersed by non-coding sequences (in-

trons), where the beginning and the end of the coding sequence are framed by a start and end codon, respectively. Those are always located in 3' direction (downstream) of the TSS. In the other direction, upstream of the TSS (in the 5' direction), resides the promoter region. The promoter is a regulatory region that allows the binding of transcription factors to enable the binding of the RNA polymerase, which is crucial for the initiation of transcription (Figure 2.2).



**Figure 2.2.: Structure of the gene coding region and the promoter.** TSS stands for transcription start site and is defined as the start of transcription by the RNA-polymerase. Based on the 5' → 3' strand, the transcribed region is always positioned in 3' direction (downstream) of the TSS and the promoter is mainly found in 5' direction, upstream of the TSS. The transcribed region involves the 5' untranslated region (UTR) and the 3' UTR framing the gene on the 5' and 3' end, respectively. Surrounded by the UTRs, the protein-coding sequence is located in the exons, which may be interspersed with non-coding introns. The figure was created with BioRender (<https://biorender.com/>).

### 2.1.2. Transcriptomics – the RNA Transmits the Genetic Information

Transcriptomics is known as the study of RNA, that is, everything that is transcribed in a cell. Since every somatic cell in an organism owns the exact same set of chromosomes and hence genomic sequence, gene expression and its regulation is the key to an efficient control of the time and quantity of gene product to be expressed [12]. This enables the creation of specific cells and tissues, allows an organism to adapt to different environments and stimuli and, thus, forms the basis for the control of structure, functionality, versatility, and adaptability.

In more detail, gene transcription, the first step of protein synthesis, works as follows. During the initiation phase, the promoter region of a gene is of particular importance (see Fig-

ure 2.2), as it forms the foundation for the binding of regulatory proteins, and most importantly, the enzyme RNA polymerase. Several regulatory proteins, called transcription factors (TFs), bind to specific transcription factor binding sites (TFBSs) in the promoter region to mediate RNA polymerase binding and, thus, to allow the formation of the transcription initiation complex. One of the best known TFBSs is the TATA box, which is found in most eukaryotic promoters, approximately 25-30 bp upstream of the TSS [9, 13]. After the RNA polymerase binds to the DNA, the DNA double strand is unwound and the base pairs are disrupted, forming a 'transcription bubble' with single stranded DNA to be transcribed. The RNA polymerase begins the synthesis on the template strand, forming the RNA strand in the 5' → 3' direction. During the second phase of transcription, the elongation, the RNA polymerase moves along the DNA, continuously unwinds the double helix, adds one new nucleotide to the building RNA strand at a time, dissociates the growing RNA chain from the template, and performs proofreading functions [9, 10]. Depending on the amount of protein required, a gene can be transcribed simultaneously by several consecutive RNA polymerases, creating an enzyme convoy [9]. In the last step, after the coding region is transcribed, the termination step comprises the stop of the synthesis, the release of the RNA product, and the dissociation of the enzyme from the DNA [10].

The study of the transcriptome provides many important insights into the expression of an organism's phenotype, and poses certain challenges regarding experimental design and data analysis, as measurements are highly context-, tissue-, and time-dependent. Currently, the two main methods to measure the transcriptome are microarrays and RNA-sequencing (RNA-seq), the former measuring the presence of a set of predefined sequences and the latter detecting all transcripts under a given condition using high-throughput sequencing methods. In both types of experiments, it is important to collect multiple (ideally >2) replicates per condition and tissue and to provide control measures for ideally the same number of replicates to account for biological variation and to apply reliable significance tests to identify differentially expressed genes (DEGs) between a condition and a control set.

### 2.1.3. Proteomics – Proteins Form the Diversity of the Cell

After transcription, the pre-messenger RNA (pre-mRNA) is processed and spliced into the mature mRNA molecule ready for translation. RNA-processing involves the addition of the 5'-cap and the poly-A tail which enables the export of the molecule from the nucleus and prevents premature degradation. The splicing process involves cutting the introns out of the pre-mRNA and joining the ends of the exons together. In most cases, this can be done in several ways, so that a different set of exons results in different mature mRNA molecules.

This so-called alternative splicing makes it possible, among other things, that one gene can give rise to several different proteins [10]. While the human genome codes for approximately 20,000 protein-coding genes, there are estimated to be at least 500,000 different human proteins [14]. The proteome, the totality of all proteins present in an organism or cell (type) at a given condition, includes all different types of proteins, such as e.g., en-

zymes, structural proteins, transport proteins, antibodies, or TFs. Since this study focuses on transcriptional gene regulation and, thus, on the regulatory proteins, TFs are the type of proteins studied in this work.

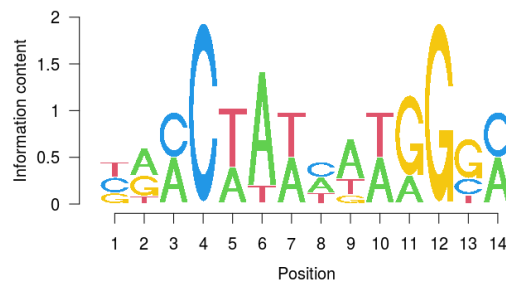
TFs govern the regulation of transcription by binding to the DNA at characteristic sequence motifs recognized by a highly specific binding domain of the protein. The sequence motifs, also called TFBSs, are typically between 5 and 15 base pairs (bp) long and are preferably found in promoters or enhancers, i.e., proximal or distal regulatory regions, respectively (Figure 2.2) [15, 16]. Most TFs have well documented sequence preferences in the form of position weight matrices (PWMs), which can be used for their prediction in a given sequence [1, 17, 18]. A PWM describing a DNA sequence is a  $4 \times l$  matrix, for a binding site of length  $l$  with one row per nucleotide, most frequently containing the log-likelihood ratio for each nucleotide and position [19, 20]. For visualization, they can be represented as sequence logos (Figure 2.3), which reveal the information content for each position via the bin height. Positions with higher information content are highly conserved among species while others are rather variable [20]. PWMs are calculated based on experimentally validated binding sites in different species using, e.g., SELEX, chromatin immunoprecipitation-sequencing (ChIP-seq), or DNA pull-down experiments [19, 21].

It is well known, that the TFBSs occur in clusters within the regulatory regions, which enable the formation of TF pairs and complexes during DNA binding [23, 24]. Thus, the interplay between TFs and their specific partner choices orchestrate the dynamic and diverse regulatory programs as a response to certain environmental conditions and determine the highly context-specific gene expression [23].

(A) Position weight matrix (PWM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	-1.94	0.8	0.8	-1.94	0.45	1.51	0.8	0.45	1.08	0.8	0	-1.94	-1.94	0.8
C	0.45	-1.94	0.8	1.69	-1.94	-1.94	-1.94	0.45	-1.94	-1.94	-1.94	-1.94	0	0.8
G	0	0.45	-1.94	-1.94	-1.94	-1.94	-1.94	-1.94	-0.67	-1.94	1.31	1.69	1.08	-1.94
T	0.45	-0.67	-1.94	-1.94	1.08	-0.67	0.8	0	0	0.8	-1.94	-1.94	-0.67	-1.94

(B) Sequence logo



**Figure 2.3.: Position weight matrix (PWM) and sequence logo of the transcription factor binding site (TFBS) for myocyte enhancer factor 2 (MEF2).** (A) shows the PWM of the MEF2 binding site in terms of log-likelihood ratios. (B) shows the respective sequence logo visually representing the TFBS. The height of each bin describes the information content in bits of the corresponding position [19]. The example and figure are based on [22].

## 2.2. Single Nucleotide Polymorphisms

An individual's genome is defined by a unique sequence characterized by a specific combination of genomic variations. Every heritable genomic variation occurring today, was introduced once as a random mutation in the germline. The most common type of genomic variations are single nucleotide polymorphisms (SNPs), nucleotide changes at a single position with a minor allele frequency of at least 1% within a population [25]. While theoretically up to four different alleles are possible, in practice SNPs usually occur bi-allelic, comprising one reference and one alternate allele [25, 26].

Especially due to their abundance in the genome, they are considered as the markers of choice for genome-wide association studies (GWAS), revealing the association of genomic markers to genetic traits or diseases [27]. However, this generally shows only a statistical association with the trait under study, with the causative SNP or its mode of action often going undetected [28]. Nevertheless, it is of great importance to identify the biologically causative variant, not only to ensure its efficient and robust use for breeding purposes, but also to decipher the mechanisms forming a particular phenotype [28]. The vast majority of trait- and disease-associated variants identified in GWAS are located in intergenic and



intronic regions and are enriched in the regulatory regions suggesting that they are likely to influence transcriptional gene regulation [23, 28, 29].

### 2.2.1. Regulatory SNPs

SNPs in promoter regions can alter regulatory elements such as TFBSs and, thus, can have an effect on the transcriptional activity of the gene [23]. Such SNPs are known as regulatory SNPs (rSNPs) and they are promising candidates in the search for causality of disease- or trait-associated SNPs [1]. Depending on the nucleotide affected within the TFBS, an rSNP may cause only a small change in binding affinity of a TF or, if a highly conserved position is substituted, a binding site may be disrupted or a new one created. In the past, several case studies have shown that rSNPs can have a major impact on the phenotype and, in extreme cases, can even be causal for a particular trait or disease.

#### 2.2.1.1. Examples of rSNPs in Humans, Animals, and Plants

A well-known example in humans is the phenotype of lactase persistence, also referred to as lactose tolerance, which can be caused by an rSNP commonly observed in European populations [30]. This rSNP (-13,910\*C/T) is located in a distal regulatory region and causes an Oct-1 binding site for the T allele, whereas the C allele does not allow the binding of Oct-1. Studies in transgenic mice have shown that the insertion of human DNA fragments with the -13,910\*T variant alone could prevent the post-weaning decline of lactase expression, whereas mice with the -13,910\*C variant were lactase non-persistent, suggesting a causal role of a single rSNP in the lactase persistence phenotype [31].

In addition, a number of rSNPs have been found to be associated with various diseases in other human studies, summarized in a review by Degtyareva et al. [23]. For example, Allen et al. [32] investigated an rSNP within an eQTL for the gene IFITM3, which is associated with severe influenza in humans. The risk allele (A) showed decreased binding affinity for the activator TF IRF3, while the inhibitor CTCF bound with higher affinity, resulting in decreased gene expression.

Another study by Y. Wang et al. (2020) [33] identified an rSNP associated with lung cancer in Chinese populations located 682 bp upstream of the DCBLD1 TSS. The T allele creates a binding site for YY1, while the risk allele (C) does not allow YY1 binding leading to decreased DCBLD1 expression.

Furthermore, Korneev et al. [34] showed in their study that an rSNP creating a PU.1 binding site enhances TLF transcription and leads to a higher risk of several diseases involving chronic inflammation.

Several studies have also examined different rSNPs in agriculturally important animal and plant species with respect to a specific trait or phenotype [1, 3, 4]. For example, Konishi et al. [35] discovered an rSNP in rice that causes a loss of TFBS for an ABI3 type TF in the promoter region of the quantitative trait locus (QTL) for seed shattering on chromosome 1

(qSH1). This rSNP is causative for the loss of seed shattering and, thus, paved the way for rice domestication [35].

In maize, several rSNPs were detected in the promoter of the maize rough dwarf disease candidate gene eukaryotic translation initiation factor 4E (eIF4E) and control its expression level [36].

Furthermore, in wheat, an rSNP associated with wheat grain weight affects the binding of a calmodulin-binding TF and hence the gene expression of the TaGW2-6A gene, a candidate gene for grain weight [37].

A previous study on the grain legume faba bean discovered two rSNPs which are significantly associated with the vicine and convicine content and affect the binding of the TFs MYB4, MYB61, and SQUA [38].

In their study on dairy cattle, Lum et al. [39] investigated the molecular mechanism underlying the  $\beta$ -Lactoglobulin (LGB) gene expression, which plays an important role in the milk casein, protein, and fat content. They found one rSNP in the LGB upstream promoter within a binding site for AP-2 that affected the protein affinity to the sequence.

Other studies [40, 41] investigated the chicken the prolactin gene, coding for the important reproductive hormone prolactin, and identified several rSNPs overlapping different TFBSs. Ballester et al. [42] identified one rSNP associated with fatty acid composition traits in pigs, which is located in the promoter region of apolipoprotein (apo-) A-II (APOA2) affecting the binding of NF-1.

### 2.2.1.2. Detection of rSNPs

In practice, the detection of rSNPs can be done in several ways. Most commonly, the first step is an *in silico* discovery of TFBSs and a prediction of the consequence caused by a nucleotide change. This first step is of utmost importance to provide prior knowledge and a starting point for experimental validation. TFBS prediction can be done using different methods and tools like MATCH™ [18], MEME [43], or ConSite [44]. To predict the effect of variants to TF binding, there exist various tools and databases, which are summarized in Table 3.1 and Table S1 of Chapter 3. However, almost all of them focus on humans or a few model organisms, and, thus, there is a great need for tools and databases addressing rSNPs in different agricultural animal and plant species [1, 3]. To the best of my knowledge, there exist currently three tools, which generally allow the detection of rSNPs in crop or livestock species. As a web-based tool, the RSAT variation-tool [45] allows the analysis of plant SNPs of user-provided inputs on the fly. However, this tool does not give any information on related genes, such as the distance to the transcription start site (TSS) or consequences such as gain- or loss of TFBS. Hence, the users need to interpret the output themselves. The RSAT variation-tool includes eight crop species and subspecies (*Hordeum vulgare*, *Oryza sativa* Indica, *Oryza sativa* Japonica, *Solanum lycopersicum*, *Sorghum bicolor*,

*Triticum turgidum*, *Vitis vinifera*, and *Zea mays*). The R packages MotifbreakR [46] and atSNP [47] principally comprise organisms stored in the Bioconductor BSGenome package [48], which includes only the crop species *Oryza sativa* and *Vitis vinifera* and the livestock species *Gallus gallus*, *Bos taurus*, and *Sus scrofa*. For both packages, the user has to provide the SNPs as well as the representation of TFBSs under study in the form of PWMs and experience in R programming is imperative.

After *in silico* analysis, an experimental validation can be done via different *in vitro* (e.g., EMSA, pull-down and reporter assays, or SNP-SELEX) and *in vivo* (e.g., ChIP-PCR, CRISPR/Cas9-mediated single nucleotide editing) experiments [23]. In addition, with genomics and transcriptomics data available for the same individuals, the impact of a genomic locus on transcription levels can be measured via expression QTL (eQTL) analysis. Similar to regular QTL analysis, which measures the association between a genomic locus and a phenotype (quantitative trait), eQTL analysis determines the association with the level of gene expression, i.e., the amount of mRNA. Hence, the detection of eQTLs is a method to determine genomic regions which have an impact on gene expression, and, thus, are likely to harbour rSNPs [23].

## 2.3. Application Projects: The Study of Regulatory Mechanisms in Agricultural Species

In this thesis, I investigate the gene regulatory mechanisms underlying a trait of interest based on omics data in both plant and animal sciences. To this end, I analyzed two case studies which I will introduce here. First, I will provide background information on the oil content and -quality of the oilseed rape. The second application project involves the investigation of immune responses of chicken and duck after infection with avian influenza.

### 2.3.1. Oil Content and Quality in Rapeseed

*Brassica napus* L. was formed around 7,500 years ago through natural hybridization between the diploid progenitors *Brassica rapa* and *Brassica oleracea*, followed by chromosome doubling. This process, known as polyploidization, gave rise to the allopolyploid crop *B. napus* ( $2n = 4x = 38$ , AACC) which is characterized by a total of 38 chromosomes, 20 of which coming from *B. rapa* ( $2n = 2x = 20$ , AA) and 18 coming from *B. oleracea* ( $2n = 2x = 18$ , CC) [49, 50]. A study by Lu et al. [49] suggested that the A subgenome evolved from a European turnip ancestor and the C subgenome from the common ancestor of kohlrabi, cauliflower, broccoli, and Chinese kale.

Today, rapeseed is one of the most important oilseed crops, cultivated worldwide not only for its high seed oil content, but also for its high protein content, which makes the rapeseed meal remaining after oil extraction a valuable animal feed [51, 52]. The oil, in addition to

its use for human consumption, is also used as lubricant and, especially in Germany and Europe, as biodiesel.

Until the early 1970s, the ability to grow rapeseed for human and animal consumption was highly limited due to its erucic acid and glucosinolate-containing oil composition. From a breeding perspective, *B. napus* is therefore a prime example of breeding improvements, because in no other crop, important quality characteristics have been changed completely in such a short time [53]. Originally, one characteristic of rapeseed oil was erucic acid, which is not found in other oil crops. This made the oil of concern for human consumption, as erucic acid is not only bitter in taste but can cause cardiac damage and other health issues in mammals [54]. An important step in breeding rapeseed with low erucic acid content was therefore a mutation that blocked a step in erucic acid synthesis so that predominantly oleic acid, a precursor of erucic acid, was formed [53, 55]. Another undesirable component found in the meal is glucosinolate, which can form toxic cleavage products during digestion, leading to adverse health effects such as liver and kidney damage and lymphatic disorders [54]. Today, it is possible to produce varieties with low erucic acid and low glucosinolate, giving rise to the so called double-low varieties, the canola as it is known today [49].

Improving the oil content is an important breeding goal today, and in this context, the resistance to several stress factors is a relevant objective [51, 53, 56]. The oil is stored within the seeds in the form of triacylglycerols (TAGs) in oil bodies, while TAG synthesis takes place in plastids through a variety of different interacting metabolic pathways and regulatory processes [57]. However, the pathways as well as the underlying transcriptional machinery controlling the oil content and -quality could vary across different *B. napus* cultivars [49, 58]. Hence, the investigation of such biological processes is an important task to assess the genetic programs of two cultivars in this study: (i) Zhongshuang11 (ZS11) characterized by a double-low accession (00, low erucic acid and low glucosinolate) and a high oil content and; (ii) Zhongyou821 (ZY821) with double-high accession (++, high erucic acid and high glucosinolate) and low oil content [49].

### 2.3.2. Avian Influenza in Chicken and Duck

The avian influenza virus (AIV) primarily infects birds such as wild waterfowl or gallinaceous poultry, but also has zoonotic potential and poses a high risk for a future pandemic [59]. After the first reports of human infections with high pathogenic avian influenza (HPAI) H5N1 in 1997, avian influenza became a globally recognized disease that was now of interest not only to veterinary medicine but also to public health [60]. Between 2003 and 2022, the World Health Organization (WHO) reported 865 cases of human infections with H5N1, 456 of which resulted in death [61].

As a type A influenza virus, AIV belongs to the family of *Orthomyxoviridae*, which are segmented negative-sense RNA viruses [60]. Their naming is based on their surface proteins neuraminidase (NA) and hemagglutinin (HA). Among 16 existing HA types, only two (H5 and H7) can cause respiratory and systemic diseases in birds [62]. Further, AIVs can

be classified into high- and low pathogenic avian influenza viruses (HPAIVs and LPAIVs, respectively) based on their pathogenicity in chicken [63]. While chicken can usually withstand an LPAI infection, they succumb to infection with HPAI within a few days with a mortality rate of up to 100% [64]. Mallard ducks, on the other hand, are known to successfully fight all LPAI and most HPAI infections, with usually only mild symptoms, and are hence considered a natural reservoir of the virus [59].

However, to date, the mechanisms underlying the susceptibility of chicken to avian influenza and the effective immune response of ducks, in particular wild mallards, have not been fully deciphered. Partially, the susceptibility of chicken can be explained by the absence of virus pattern recognition receptor RIG-I gene and the gene for the RIG-I binding protein, RNF135, both of which exist in ducks [59, 65]. The RIG-I receptor recognizes double-stranded RNA and initiates self-promoting pathways leading to the early type I interferon (IFN) response, which is important for innate immune response. In chicken, other pattern recognition receptors, such as MDA5 and TLR7, are upregulated in response to viral entry, which also leads to the induction of IFN and IFN-stimulated gene expression [62, 63, 65, 66]. However, the immediate induction of type I IFNs (IFN- $\alpha$  and IFN- $\beta$ ) seems to be much more robust and effective in ducks than in chicken or other avian species. This first checkpoint for controlling the virus is crucial to the delay and prevention of viral replication, but it is by far not the only mechanism which is responsible for the successful immune response of ducks. Evseev and Magor [62] provide a comprehensive overview of the differences in innate immune responses in chicken and duck and highlight also factors like the sialic acid receptor distribution in the trachea and intestinal tract, different mechanisms to control inflammation, rapid apoptotic response or the adaptive immunity. However, host-pathogen interactions and, in particular, their underlying transcriptional gene regulation in duck and chicken are multifactorial and highly complex, and further elucidation is needed to gain deeper insight into the effective immune response against AIV in ducks, while it proves lethal to chicken [62].



### 3. agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species

This chapter contains the article of the same name published in August 2021 in the MDPI journal *Biology* (<https://doi.org/10.3390/biology10080790>). For the sake of consistency within this thesis, the journal style is not adopted in this chapter.

This article is a joined work of Selina Klees<sup>1,2,\*†</sup>, Felix Heinrich<sup>1,†</sup>, Armin Otto Schmitt<sup>1,2</sup> and Mehmet Gültas<sup>2,3,\*</sup>

<sup>1</sup>Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany

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

<sup>3</sup>Faculty of Agriculture, South Westphalia University of Applied Sciences, Lübecker Ring 2, 59494 Soest, Germany

\* Authors to whom correspondence should be addressed.

† These authors contributed equally to this work.

**Author contributions by Selina Klees (S.K.):** S.K. participated in the design of the study, conducted the computational and statistical analyses, performed the biological validation, created the database, and wrote the final version of the manuscript.

#### 3.1. Simple Summary

Regulatory SNPs (rSNPs) are SNPs located within promoter regions that have a high potential to alter gene expression by changing the binding affinity of transcription factors to their binding sites. Such rSNPs are gaining importance in the life sciences due to their causality for specific traits and diseases. In this study, we present agReg-SNPdb, the first database comprising rSNP data of seven agricultural and domestic animal species: cattle, pig, chicken, sheep, horse, goat, and dog, and made it usable via a web interface.

## 3.2. Abstract

Transcription factors (TFs) govern transcriptional gene regulation by specifically binding to short DNA motifs, known as transcription factor binding sites (TFBSs), in regulatory regions, such as promoters. Today, it is well known that single nucleotide polymorphisms (SNPs) in TFBSs can dramatically affect the level of gene expression, since they can cause a change in the binding affinity of TFs. Such SNPs, referred to as regulatory SNPs (rSNPs), have gained attention in the life sciences due to their causality for specific traits or diseases. In this study, we present agReg-SNPdb, a database comprising rSNP data of seven agricultural and domestic animal species: cattle, pig, chicken, sheep, horse, goat, and dog. To identify the rSNPs, we constructed a bioinformatics pipeline and identified a total of 10,623,512 rSNPs, which are located within TFBSs and affect the binding affinity of putative TFs. Altogether, we implemented the first systematic analysis of SNPs in promoter regions and their impact on the binding affinity of TFs for livestock and made it usable via a web interface.

### Keywords

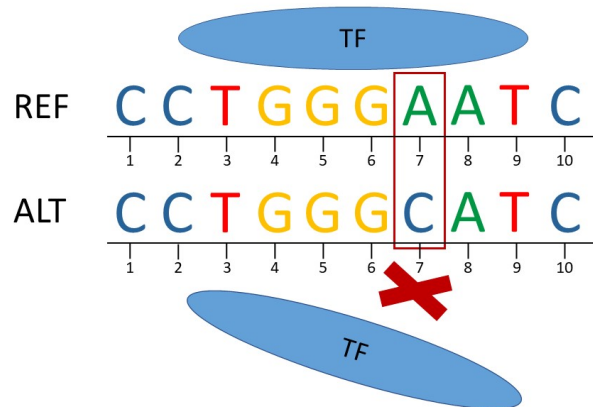
single nucleotide polymorphism; regulatory SNP; transcription factor; transcription factor binding site; gene regulation; database; agricultural animal species; livestock

## 3.3. Introduction

The transcriptional regulation of gene expression in higher organisms is essential for various biological processes. In contrast to the process of translation, the transcriptional machinery and its regulatory mechanisms are far from being deciphered [67]. These mechanisms are mainly governed by a special class of regulatory proteins, the transcription factors (TFs), and their combinatorial interplay [68, 69]. TFs regulate the transcription as a response to specific environmental conditions by binding to short degenerate sequence motifs known as transcription factor binding sites (TFBSs) in promoter regions of their target genes and, thereby, enhance or repress gene transcription. Genomic variations, such as single nucleotide polymorphisms (SNPs), define and characterize specific populations or phenotypes and are, hence, used as markers in animal and plant breeding. Due to the decreasing costs for whole genome sequencing, an increasing number of variants is detected followed by association studies statistically linking SNPs to specific traits or diseases. However, the identification of causal variants and the elucidation of their regulatory roles is proceeding at a slow rate [70, 71]. Today, it is well known that most disease- and trait-associated SNPs are not located within the coding regions of genes but in non-coding regions [23, 29, 72, 73]. SNPs that are located in regulatory regions can alter TFBSs leading to a change in the binding affinity of TFs and, in extreme cases, even result in the disruption of a TFBS or



the creation of a new TFBS (Figure 3.1) and, thus, affect gene expression. Such SNPs are referred to as regulatory SNPs (rSNPs) [47, 74, 75].



**Figure 3.1.: Scheme of the disruption of transcription factor (TF) binding due to a regulatory SNP.** The TF can bind to the reference (REF) sequence while it does not bind to the alternate (ALT) sequence (C instead of A at position 7).

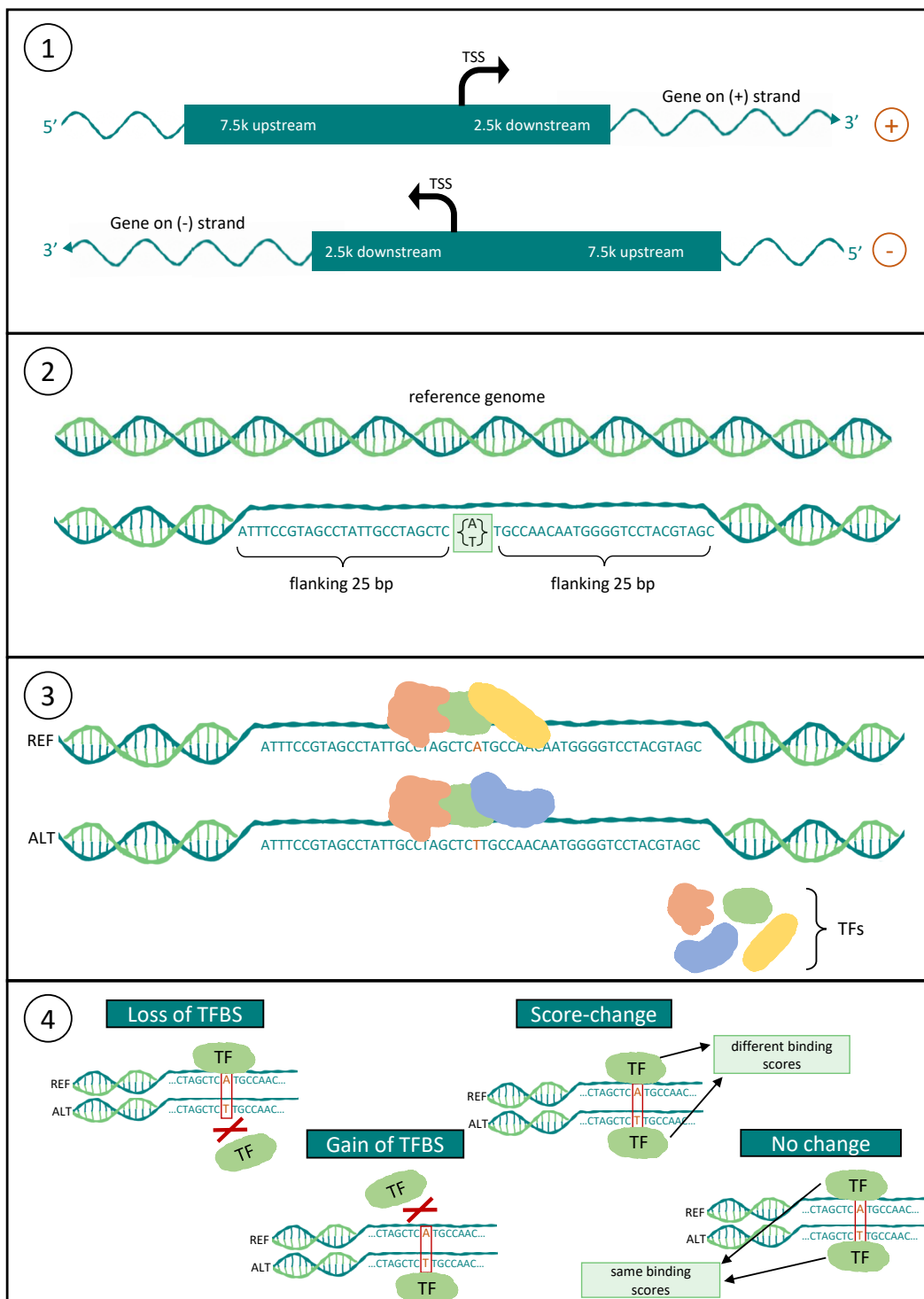
The importance of rSNPs has been studied extensively in humans and they are found to have a causal role for numerous traits and diseases [31, 76–78]. A recent review on human rSNPs summarizes different rSNP studies [23]. Due to the great interest in rSNPs, several tools and databases for the analysis of the effects of SNPs on regulatory elements, e.g., TFBSs, have been developed for humans or certain model organisms. Five recent studies are summarized in Table 3.1, and a comprehensive overview is given in Table S1.

Recently, rSNPs are gaining attention in life sciences and animal breeding since they can be causal for specific traits and diseases and could, hence, serve as new targets for breeding. For this reason, several studies investigated the critical role of rSNPs in agriculturally important species, such as cattle [39, 79–84], pig [42, 85, 86], and chicken [40, 41, 87]. As these studies were focused on the regulatory role of SNPs for a single trait of interest, they were highly case-specific. Thus, there still exists a lack of systematic analyses of the effects of rSNPs in agricultural species, and, until now, only a few existing tools and databases (DBs) are available for livestock.

MotifbreakR [46] and atSNP [47] are both R packages that principally include all organisms stored in the Bioconductor BSGenome package [48]; however, they require the user to supply the SNP and TFBS data (represented by position weight matrices (PWMs)), and experience in R programming is essential. The Ensembl Variant Effect Predictor (VEP) [88] stores data from experimentally supported and published rSNPs. Due to the lack of

experimentally supported data of regulatory elements in livestock, the VEP mainly contains data of regulatory elements and variants for human and mouse. Therefore, the information for livestock stored in the Ensembl VEP is limited to annotations based on the position of the SNP with respect to a gene, e.g., in the upstream region or in the 5' UTR, excluding effects on TF binding.

In order to address the limited knowledge and information available regarding the crucial functions of rSNPs and their associations with TFBSs in livestock, we systematically carried out an analysis to detect rSNPs and predicted their effects on TF binding for seven agricultural and domestic species (cattle, pig, chicken, sheep, horse, goat, and dog). In particular, we first analyzed the promoter regions (ranging from  $-7.5$  kb to  $+2.5$  kb) of all annotated genes and obtained the SNPs within these regions. Secondly, we extracted the flanking sequences for these SNPs and performed a TFBS prediction on the reference as well as alternate sequences. Finally, we assigned the identified SNPs to different categories based on their consequences on TF binding (Figure 3.2) as suggested in [4, 38]. To demonstrate our results in a proper way, we developed a database, namely agReg-SNPdb, which stores all predicted regulatory SNPs and their consequences on TF binding for each gene, and we made it accessible via a web interface (<https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb>). Furthermore, we performed a literature survey to show that our results are in agreement with previous experimental and in silico studies.



**Figure 3.2.: Scheme of the workflow applied for the detection of rSNPs.** (1) Definition of the promoter region as 7.5 kb upstream (5' direction) and 2.5 kb downstream (3' direction) of the TSS, and extraction of SNPs within this region; (2) extraction of the flanking 25 bp around the SNPs from the reference genome; (3) prediction of the TFBSs for both the reference and alternate sequences; and (4) deriving the consequences for each SNP-TFBS pair.

**Table 3.1.: A summary of five recent studies that systematically investigated the effects of SNPs on regulatory elements such as TFBSs.** The analyses were done by either collecting experimentally supported and published data or by predicting the SNP's impact on TF binding using prediction tools.

Name	Species	DB/ tool	Website	Characteristics	Experimentally supported data or prediction
QBiC-Pred [89]	Human	Tool	<a href="http://qbic.genome.duke.edu">http://qbic.genome.duke.edu</a> (accessed on 16 August 2021)	<ul style="list-style-type: none"> <li>TFBS prediction with regression models</li> <li>Prediction of changes in TF binding using ordinary least squares and evaluation of correlation between the predicted binding changes and changes in gene expression</li> </ul>	TFBS prediction
atSNP [47], atSNP-Search [90]	Human (atSNP: organisms from Bioconductor BSGenome package [48])	Tool, DB	<a href="http://atsnp.biostat.wisc.edu">http://atsnp.biostat.wisc.edu</a> (accessed on 16 August 2021)	<ul style="list-style-type: none"> <li>atSNP: R package for TF binding affinity testing for rSNPs (needs a SNP and motif set as input)</li> <li>atSNP Search: DB for human SNP-motif pairs and the respective significance</li> </ul>	TFBS prediction
INFERNO [91]	Human	Tool	<a href="http://inferno.hispanwaglab.org">http://inferno.hispanwaglab.org</a> (accessed on 16 August 2021)	<ul style="list-style-type: none"> <li>Infering causal variants from genome-wide association studies (GWAS) within annotated regulatory regions as enhancers including tissue context</li> <li>TFBS prediction with HOMER</li> </ul>	TFBS prediction
rSNPBASE [92], rSNPBASE 3.0 [74]	Human	DB	<a href="http://rsnp.psych.ac.cn">http://rsnp.psych.ac.cn</a> (accessed on 16 August 2021) <a href="http://rsnp3.psych.ac.cn">http://rsnp3.psych.ac.cn</a> (accessed on 16 August 2021)	<ul style="list-style-type: none"> <li>DB of rSNPs with references to regulatory elements</li> <li>Includes proximal and distal regulatory regions, post-transcriptional regulation, linkage disequilibrium (LD), and expression quantitative trait locus (eQTL) information</li> <li>rSNPBASE 3.0 includes regulatory element-target gene pairs for regulatory networks</li> </ul>	experimentally supported regulatory elements
SNP2TFBS [17]	Human	DB	<a href="https://csg.epfl.ch/snp2tfbs">https://csg.epfl.ch/snp2tfbs</a> (accessed on 16 August 2021)	<ul style="list-style-type: none"> <li>DB of human SNPs that affect TFBSs and the prediction of a consequence</li> <li>DB can be downloaded as text files or accessed via the website</li> </ul>	TFBS prediction

## 3.4. Materials and Methods

### 3.4.1. Input Data

The construction of agReg-SNPdb requires: (i) a library of PWMs representing the TFBSs and, for each animal, (ii) a reference genome, (iii) a SNP catalog, and (iv) gene annotations. As a PWM library, we used the non-redundant vertebrate matrices provided by TRANSFAC [93]. The reference genomes, SNP catalogs, and gene annotation files are downloaded from Ensembl [94]. The respective assembly versions are listed in Table 3.2. The SNP catalog was filtered by discarding all insertions and deletions, keeping only the SNPs. For most genes, more than one transcript isoform was annotated [88], e.g., due to different splicing variants. This ambiguity was kept during the analysis if the positions of the transcription start sites (TSSs) and, hence, the derived promoter regions were different.

**Table 3.2.: Assembly versions of the input data, including the reference genome, SNP catalog, and gene annotations.** All files were downloaded from Ensembl (release 103).

Animal	Assembly Version	Download Date
Cattle	ARS-UCD1.2	03/01/2021
Pig	Sscrofa11.1	03/09/2021
Chicken	GRCg6a	02/25/2021
Sheep	Oar_rambouillet_v1.0	03/01/2021
Horse	EquCab3.0	03/01/2021
Goat	ARS1	03/01/2021
Dog	CanFam3.1	03/08/2021

### 3.4.2. Pipeline

A general workflow of the detection pipeline is shown in Figure 3.2. In our previous studies on faba beans [38] and rapeseed [4], we established similar pipelines for the prediction of rSNPs.

#### 3.4.2.1. Detection of SNPs within the Promoter Region

The first step of this analysis was to extract SNPs, which are located within the pre-defined promoter regions. Since there exists no experimentally verified information regarding the exact location of the promoters and in order to overcome inaccuracies in TSS prediction, we chose a large promoter region of 7.5 kb upstream and 2.5 kb downstream of the TSS. Similarly large promoter regions were used in previous studies [15, 74, 91, 95–100]. This promoter region can be narrowed by the user during a database search on our website. For

all annotated genes, we extracted the SNPs within this region for further analysis by using the function `foverlaps` of the package `data.table` in R [101].

#### 3.4.2.2. Prediction of TFBSs

For each SNP lying within a promoter region, we extracted the respective flanking sequence of 25 bp on each side of the SNP resulting in sequences with a total length of 51 bp and the SNP at position 26 (similar flanking sequences were used in [4, 38, 96, 102]). Sequences with a length of less than 51 bp or sequences with gaps were discarded. After extracting the flanking sequences, we created two sequences per SNP, one with the reference and one with the alternate allele at the SNP position. Both were used as input for the TFBS prediction tool MATCH<sup>TM</sup> [18], which scanned the sequences to predict TFBSs using a PWM library from TRANSFAC with specific cut-off values to minimize the false positive rates. If a PWM matched a segment of genomic DNA, this sequence motif was referred to as a (potential) TFBS. As a result, the algorithm provided two scores for each predicted TFBS [18, 93]: the matrix similarity score (MSS), measuring the quality of the match regarding the whole PWM sequence, and the core similarity score (CSS), measuring the quality of the match regarding the first five most-conserved consecutive positions of the PWM. Both scores were within the range [0,1], where a score of 1 denoted an exact match of the sequence with the PWM [18] measuring the quality of the match and indicating the binding affinity of a TF to the site.

In TRANSFAC, a PWM identifier follows a certain terminology with the structure *V\$factorname\_version*. In our case, each PWM starts with “V\$”, which indicates that the PWM originated from a vertebrate TF. The *factorname* specifies the name of the TF that is binding to the DNA motif. Since there can be several PWMs representing the sequence motif of a specific TF, the *version* was specified for unique identification [69, 93].

#### 3.4.2.3. Annotation of Consequences

For each SNP, we obtained two sets of predicted TFBSs — one for the reference and one for the alternate allele. By comparing these two sets, we manually determined the consequence of a SNP on a TFBS as in our previous studies [4, 38]. We differentiated four different consequences: (i) no effect, (ii) change in binding affinity, (iii) loss of TFBS, and (iv) gain of TFBS. We defined two TFBS predictions as the same if their PWMs, positions, and the strand on which they were found were equal for both alleles.

A SNP was considered to have no effect on a TFBS if both scores computed by MATCH<sup>TM</sup> were equal for both alleles. A SNP was considered to cause a change in the binding affinity of a TF if the matrix similarity score computed by MATCH<sup>TM</sup> differed for the two alleles. A SNP caused a loss or gain of TFBS if the considered TFBS was only predicted for the reference or alternate sequence, respectively. In this study, we defined an rSNP as a SNP that caused a loss or gain of TFBS or a score-change for at least one TFBS.

## 3.5. Results

### 3.5.1. Database

We created the mysql database [103] agReg-SNPdb, which stores (i) general information about the SNPs, such as the ID, chromosomal position and the alleles (table *snp\_info*); (ii) general information about the genes, such as the gene name and chromosomal position (table *gene\_info*); (iii) the table *snp\_region* connecting the tables *snp\_info* and *gene\_info* by storing SNPs and their corresponding target genes together with their genomic position within the promoter region based on the distance to the TSS; and, most importantly, (iv) for each SNP within a promoter region (i.e., for each SNP in table *snp\_region*), we store its consequences based on the predicted TFBS binding potential (table *TFBS\_results*). A summary of the number of entries for each table and animal stored in our database is shown in Table 3.3.

**Table 3.3.: The number of records stored in the database tables *snp\_info*, *gene\_info*, *snp\_region*, and *TFBS\_results*.**


	<i>snp_info</i>	<i>gene_info</i>	<i>snp_region</i>	<i>TFBS_results</i>
<b>Cattle</b>	88,109,946	21,656	9,335,814	9,074,371
<b>Pig</b>	58,145,647	20,267	4,385,724	4,432,047
<b>Chicken</b>	20,917,836	16,659	3,810,524	3,901,905
<b>Sheep</b>	50,164,898	20,359	3,216,474	3,205,279
<b>Horse</b>	20,331,427	20,499	1,585,207	1,713,395
<b>Goat</b>	31,331,447	19,658	1,987,914	2,015,588
<b>Dog</b>	4,725,021	19,960	494,691	489,292
<b>Total</b>	273,726,222	139,058	24,816,348	24,831,877

### 3.5.2. Web Interface

The web interface (<https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb>, accessed on 16 August 2021) allows users to query the agReg-SNPdb without SQL knowledge and to obtain the requested results either on our website directly or by downloading them as CSV files. The database can be searched by (i) SNP identifiers in the form of rs numbers, (ii) SNP positions, (iii) SNP regions in a specified chromosome, or (iv) gene identifiers, i.e., the Ensembl gene stable ID or gene name (Figure 3.3).


The search results will contain, at maximum, four tables: (1) a table showing general SNP information (table *snp\_info*); (2) a table showing general gene information (table *gene\_info*); (3) a table linking the SNPs to the genes, more specifically to the promoter regions, if they are positioned within a promoter region (table *snp\_region*); and (4) for all rSNPs,


a table with the predicted TFBSs overlapping each rSNP, the MATCH™ scores, and the respective consequence (table *TFBS\_results*) for both alleles. An example output can be seen in Figure 3.4. In all tables, we provide links to sites with additional information for the SNPs and genes, and, for each PWM, we display the respective sequence logo if desired. Apart from the search site, the complete database tables can be downloaded chromosome-wise on the summary page of the respective animal.



### Database search

<a href="#">Home</a>	<b>Species:</b> <input type="text" value="Cattle"/>
<a href="#">Search</a>	<input checked="" type="radio"/> <b>Search by SNP ID</b> SNP ID (rs number): <input type="text" value="rs41566363"/>
<a href="#">Results</a>	<input type="radio"/> <b>Search by SNP position</b> Chromosome: <input type="text" value="Select..."/> Position: <input type="text"/>
<a href="#">About</a>	<input type="radio"/> <b>Search by chromosomal region</b> Chromosome: <input type="text" value="Select..."/> start: <input type="text"/> end: <input type="text"/>
<a href="#">Contact</a>	<small>Results are only displayed for regions less than 10 kb. Otherwise the results can only be downloaded.</small>
<a href="#">Institute</a>	<input type="radio"/> <b>Search by gene</b> ⓘ Gene: <input type="text"/> Promoter region ⓘ from <input type="text" value="-7500"/> to <input type="text" value="2500"/>
<input type="button" value="Start"/> <input type="button" value="Reset"/>	





Center for Integrated Breeding Research

**Figure 3.3.: Search page of agReg-SNPdb.** Search options are (1) by SNP ID, (2) by SNP position, (3) by chromosomal region, and (4) by gene.



## SNP information

Show 10 entries

Search: 

SNP_ID ▲	Chromosome	Position	REF	ALT	Quality	Filter	INFO
<a href="#">rs41566363</a>	23	23277585	G	C	.	.	ID=51111850;Variant_seq=C;evidence_values=Multiple_observations,Frequency;Dbxref=dbSNP_150:rs41566363;Reference_seq=G

Showing 1 to 1 of 1 entries

Previous  Next

## Gene information

Show 10 entries

Search: 

Name	Chromosome	Strand	txStart	txEnd	Name2
<a href="#">ENSBTAG00000020425</a>	23	+	23277603	23337345	TFAP2D

Showing 1 to 1 of 1 entries

Previous  Next

## SNP region information

Show 10 entries

Search: 

SNP_ID ▲	Gene_Name	Chromosome	Strand	txStart	txEnd	Label	Distance to TSS (bp)
<a href="#">rs41566363</a>	<a href="#">ENSBTAG00000020425</a>	23	+	23277603	23337345	inUpstreamPromoterRegion	-18

Showing 1 to 1 of 1 entries

Previous  Next

## Found TFBSs

## Explanation of Consequences

<b>Gain of TFBS</b>	The TFBS exists only for the 1 (alternative) allele of the SNP
<b>Loss of TFBS</b>	The TFBS exists only for the 0 (reference) allele of the SNP
<b>Score-Change</b>	The TFBS exists for both alleles but the binding affinity differs as measured by the Core_Similarity_Score and Matrix_Similarity_Score calculated by MATCH™
<b>No Change</b>	The TFBS exists for both alleles with the same binding affinity

Show 10 entries

Search: 

SNP_ID ▲	Allele	PWM	Position	Strand	Core_Similarity_Score	Matrix_Similarity_Score	Sequence	Consequence
<a href="#">rs41566363</a>	0	<a href="#">VSPLZF_02</a>	22	-	1	0.862	gcaggctagatCTTTAtcttcacataa	Score-Change
<a href="#">rs41566363</a>	1	<a href="#">VSPLZF_02</a>	22	-	1	0.864	gcagcgtagatCTTTAtcttcacataa	Score-Change
<a href="#">rs41566363</a>	0	<a href="#">VSZIC1_05</a>	15	+	1	0.99	acacaCAGCAgggct	Loss of TFBS

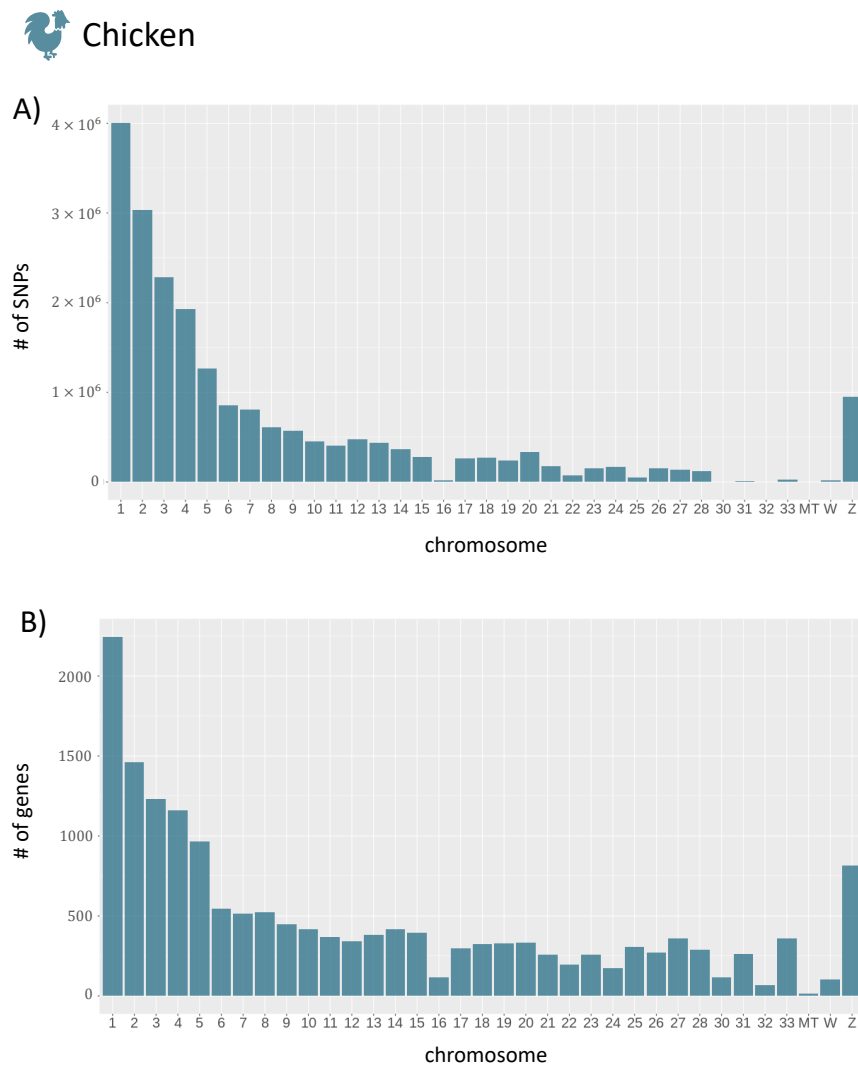
Showing 1 to 3 of 3 entries

Previous  Next

**Figure 3.4.: Example of a search result from agReg-SNPdb.** The search was performed by the SNP id rs41566363 of cattle. The result tables contain, first, general SNP information; secondly, general gene information; thirdly, information about the SNP region, in particular the promoter region and distance to the TSS; and lastly, the overlapping TFBSs (represented by PWMs) for the SNP with predicted consequences.

### 3.5.3. Statistical Analysis of the Data

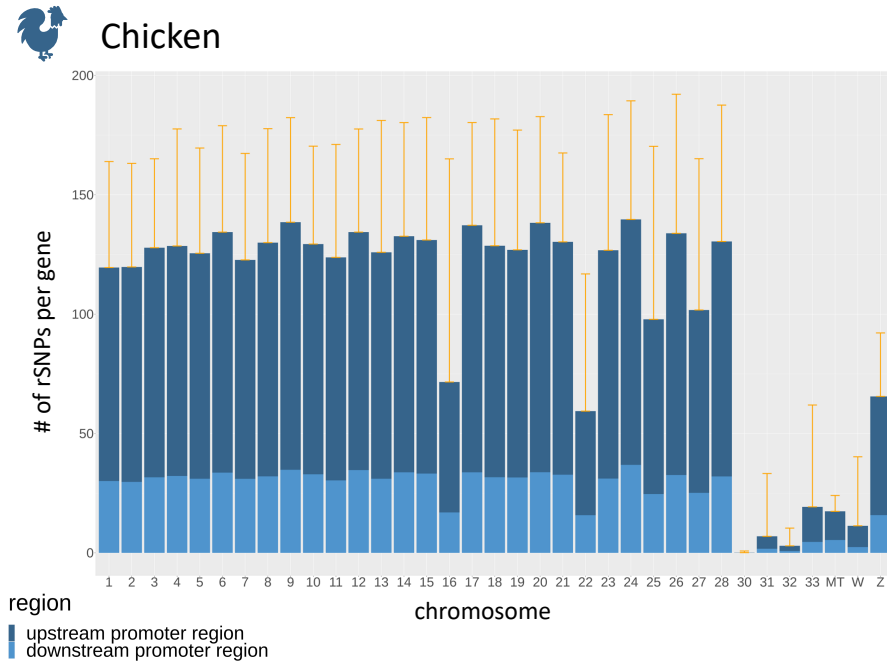
To give a brief overview of the data stored in agReg-SNPdb, we show the distribution of SNPs, genes, and rSNPs in the promoter regions along the chromosomes in an exemplary manner for the species chicken. The distributions for the remaining animals can be found in Figures S2 and S3. The distributions of SNPs and genes along the chromosomes are shown in Figure 3.5. As expected, the number of SNPs and genes decreased largely with increasing chromosome number and, hence, with decreasing chromosome size.



**Figure 3.5.: The total number of SNPs and genes for each chromosome of chicken. (A)** The number of SNPs per chromosome. **(B)** The number of genes per chromosome. In total, 20,917,836 SNPs and 16,659 genes were reported. For plotting, the R package ggplot2 [104] was used.

Regarding the promoter regions, the number of SNPs in promoters is dependent on the number of genes (Figure 3.5 B) for each chromosome. To overcome this dependency, we calculate the average number of rSNPs per gene in the upstream as well as the downstream promoter region. The average numbers of rSNPs for each chromosome in chicken revealed

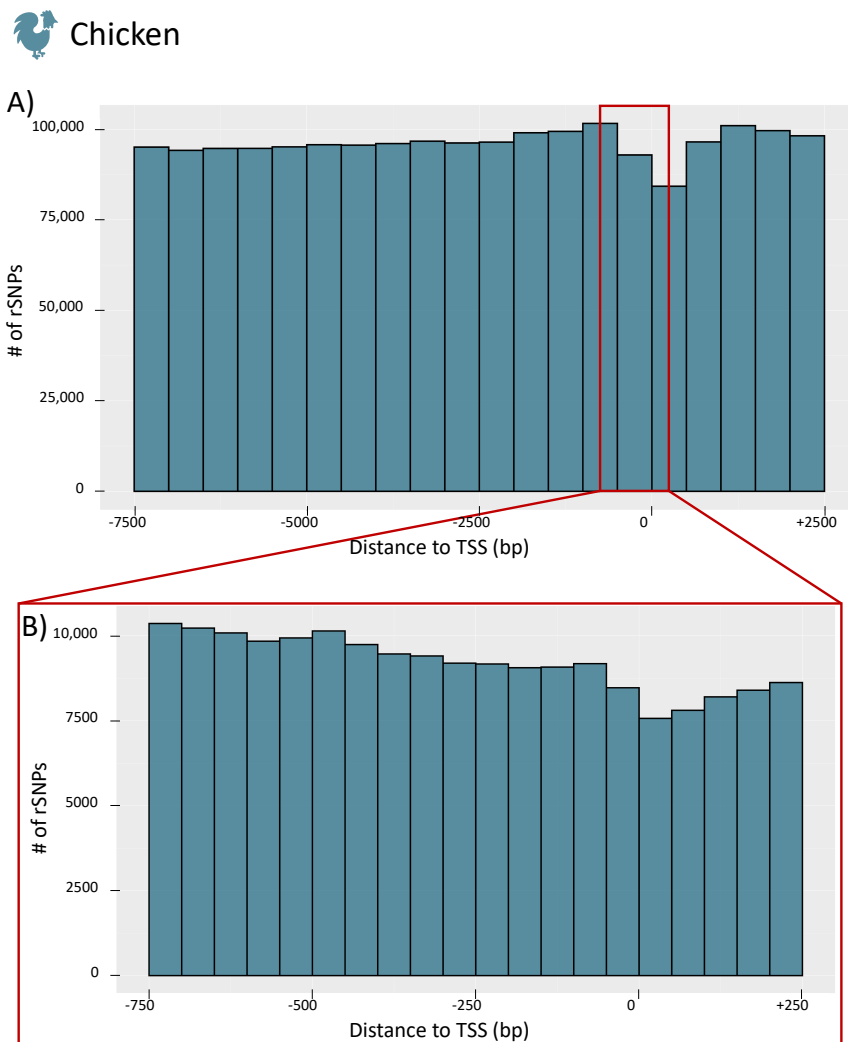
that most chromosomes had approximately 120 rSNPs per gene, while, on some chromosomes, only very few rSNPs per gene were found (Figure 3.6).



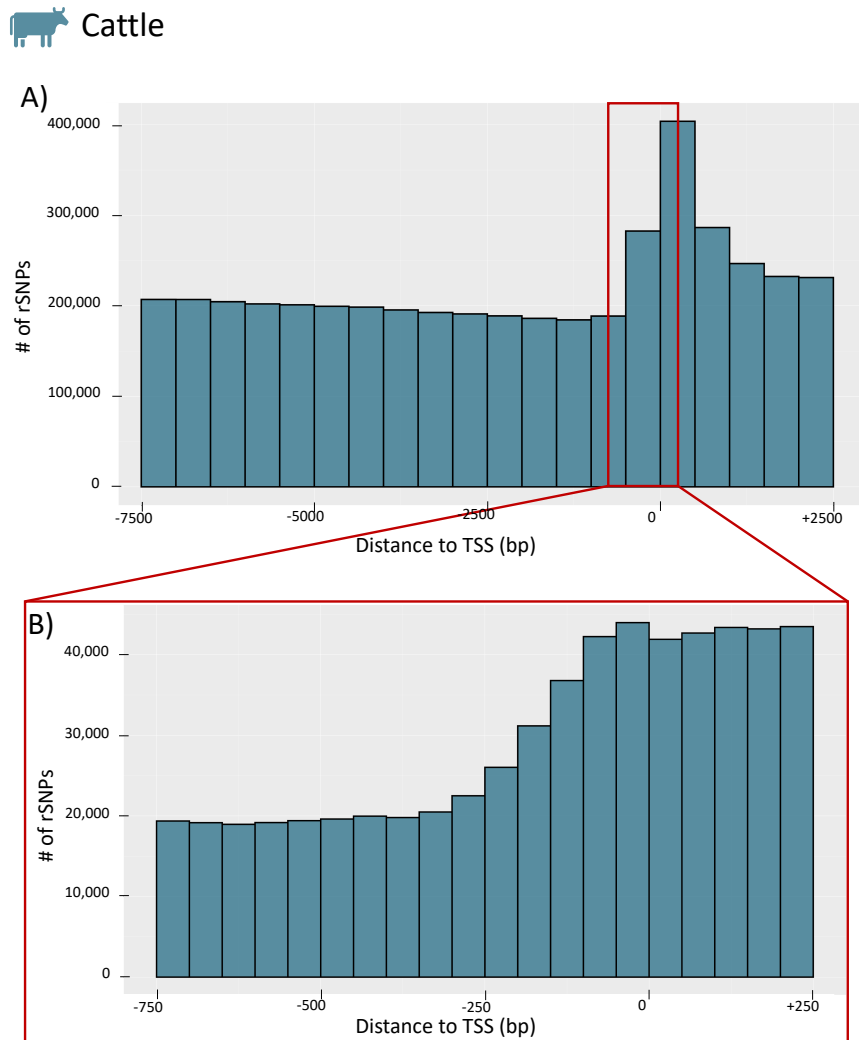
**Figure 3.6.:** The average number of rSNPs in promoter regions per gene for each chromosome of chicken, divided into upstream and downstream promoters. The orange whiskers denote the mean plus one standard deviation.

Overall, by dividing the total number of rSNPs by the total number of genes, we identified on average 95.04 rSNPs within the promoter region (10 kb) of one gene in chicken.

To obtain further insight into the distribution of rSNPs in the promoter regions, we investigated their genomic positions relative to the TSS for the whole promoter region ( $-7.5$  kb to  $+2.5$  kb) and for a smaller section ( $-750$  bp to  $+250$  bp) for chicken (see Figure 3.7 A and B, respectively; the figures for the remaining species are given in Figures S4). For chicken, we observed a similar finding as in our previous study on rapeseed [4] and as previously shown in rice [13]. While there are few rSNPs in close proximity to the TSS, the number of rSNPs increases with increasing distance to the TSS. Interestingly, in cattle (as well as in dogs), we observed the opposite tendency. Many rSNPs were found around, and especially directly downstream, of the TSS, while the number decreased with the distance to the TSS (the distribution of cattle rSNPs is shown in Figure 3.8).



**Figure 3.7.: Distribution of the distances between rSNPs and the TSS of chicken. (A)** The counts for the whole promoter region (−7.5 kb to +2.5 kb) in 500 bp intervals. The enlargement in **(B)** shows the proximal promoter region (−750 bp to +250 bp) in 50 bp intervals.



**Figure 3.8.: Distribution of the distances between rSNPs and the TSS of cattle.** (A) The counts for the whole promoter region (−7.5 kb to +2.5 kb) in 500 bp intervals. The enlargement in (B) shows the proximal promoter region (−750 bp to +250 bp) in 50 bp intervals.

### 3.6. Biological Validation Based on Case-Studies

In order to validate the data stored in agReg-SNPdb, we performed literature research and assessed the importance of our findings based on selected published studies, which identified putative rSNPs that are associated with a trait under study and affect TF binding, either by prediction or as evaluated in a biological experiment.

#### 3.6.1. Milk Protein and Fat Content in Dairy Cattle

Lum et al. [39] studied the molecular mechanism of different expression levels of the  $\beta$ -Lactoglobulin (*LGB*) gene (also known as *MBLG* or *PAEP*), which plays an important role in the milk casein, protein, and fat content in dairy cattle. They described one rSNP in the *LGB* promoter with a G to C conversion 450 bp upstream of the TSS that was found within an activator protein-2 (AP-2) binding site. Measuring the different AP-2 binding affinities with DNase-I footprinting, they measured increased protein binding in the A promoter (G allele).

In our database, we identified the same rSNP (rs41255679, C/G), which was located in the proximal upstream promoter region of *PAEP* and caused a gain of the AP-2 binding site with the G allele (Table 3.4) [105]. This supports the findings of different studies reporting that AP-2 binding as well as *LGB* gene expression is enhanced by the G allele and that rs41255679 could be an important regulator of *LGB* expression [39, 105–107].

**Table 3.4.: Consequences of SNP rs41255679 (C/G), located upstream of the TSS of the bovine *LGB* gene.** Allele 0 refers to a predicted TFBS in the reference sequence, while allele 1 stands for the alternate allele. A SNP causes a loss of TFBS if the considered TFBS (represented by a PWM) is only predicted for the reference allele. Consequently, a SNP causes a gain of TFBS if the TFBS is only predicted for the alternate allele.

SNP ID	Allele	PWM	Consequence
rs41255679	0	V\$CTCF_01	Loss of TFBS
rs41255679	1	V\$AP2ALPHA_03	Gain of TFBS

#### 3.6.2. Fat-Related Beef Quality Traits in Cattle

Matsumoto et al. (2014) [81] investigated the role of different bovine fat-related genes, including the gene encoding the fatty acid-binding protein 4 (*FABP4*). Within the *FABP4* upstream promoter, they identified two SNPs in linkage disequilibrium (*FABP4* g.-295A>G and *FABP4* g.-287A>G) that were associated with several fat-related traits, such as the carcass weight and beef marbling score. Using TFSEARCH [108], they predicted TFBSs overlapping the SNPs and altering their binding sites. In agReg-SNPdb, we identified two SNPs

within the *FABP4* promoter region at a distance of 8 bp to each other and A to G conversions (respectively, T to C conversions, due to the gene's location on the minus strand).

For the first SNP rs110055647, located 123 bp upstream of the TSS, we predicted a loss of TFBS for the Sex-Determining Region Y Protein (SRY) binding site, which is in line with the results of Matsumoto et al. (2014) [81]. For the neighboring rs109682576 (-115 bp from the TSS), we did not observe the CCAAT/enhancer-binding protein beta (cEBP/ $\beta$ ) binding site predicted in their study; however, the TFBSs for Zinc finger proteins 333 (ZNF333) and 105 (ZFP105) were lost with the alternate allele, which can be seen as an extension to the results of Matsumoto et al. (2014) (Table 3.5) [81].

**Table 3.5.: Consequences of the SNPs rs110055647 and rs109682576 in the bovine *FABP4* upstream promoter with a T to C conversion.** Allele 0 refers to a predicted TFBS in the reference sequence, while allele 1 stands for the alternate allele. A SNP causes a loss or gain of TFBS if the considered TFBS is only predicted for the reference or alternate allele, respectively. A SNP is considered to cause a score-change if the TFBS is predicted on both alleles (0,1) with a difference in the matrix similarity score computed by MATCH™.

SNP ID	Allele	PWM	Consequence
rs110055647	0,1	V\$RHOX11_01	Score-Change
rs110055647	0	V\$SRY_Q6	Loss of TFBS
rs109682576	0	V\$ZNF333_01	Loss of TFBS
rs109682576	0	V\$ZFP105_04	Loss of TFBS

### 3.6.3. Chicken Egg Production

The prolactin (*PRL*) gene product is considered as an important reproductive hormone involved in diverse biological functions in vertebrates. In laying hens, it is an important regulator of egg production since an increased PRL secretion induces broodiness behaviour [41]. Liang et al. (2006) [40] examined the *PRL* 5' promoter region and, using several populations of Chinese native Yuehuang, Taihe Silkie, and White Leghorn Layer chickens, they identified different rSNPs overlapping the predicted binding sites, including GATA-binding factor 1 (GATA-1), nuclear factor 1 (NF-1), and activator protein-1 (AP-1). Particularly for SNP rs313497646 (A/G conversion, 2048 bp upstream of the TSS), we observed the same pattern with respect to TF binding in agReg-SNPdb: only the A allele allows the binding of the NF-1 factor.

Furthermore, it has been shown that the pituitary transcription factor 1 (PIT-1) is an important activator of the *PRL* gene expression [40, 41, 109]. In agReg-SNPdb, we store a SNP (rs731078272, G/T), located -3086 bp from the TSS and causing a loss of the PIT-1 binding site in the T allele. This result suggests that this SNP might be an important regulator of

*PRL* expression where the T variant could repress *PRL* expression, which is an important indication for further studies.

#### 3.6.4. Fatty-Acid Composition Related Traits in Pigs

Ballester et al. [42] studied the expression of apolipoprotein (apo-) A-II (*APOA2*), a protein involved in the triglyceride, fatty acid, and glucose metabolisms, and identified several SNPs associated with *APOA2* gene expression and fatty acid composition traits. Four SNPs were located in the promoter region (rs322246820, rs335066625, rs339777757, and rs333406887), among which they only found one (rs333406887, C/G) influencing a predicted TFBS — in this case, a NF-1 binding site.

Similar to their result, in agReg-SNPdb, we found the SNP rs333406887 overlapping TFBSs, such as the NF-1 binding site. Furthermore, in addition to the reported change in the binding score for NF-1, we can predict several other TFBSs that are affected by this SNP. It causes, for instance, a loss of TFBS for the kruppel-like factor 6 (also called CPBP) and a gain of TFBS for zinc finger protein X-linked (*ZFX*) (Table 3.6).

**Table 3.6.: Consequences of the SNP rs333406887 (C/G) located -238 bp from the porcine *APOA2* TSS.** Allele 0 refers to a predicted TFBS in the reference sequence, while allele 1 stands for the alternate allele. A SNP causes a loss or gain of TFBS if the considered TFBS is only predicted for the reference or alternate allele, respectively. A SNP is considered to cause a score-change if the TFBS is predicted on both alleles (0,1) with a difference in the matrix similarity score computed by MATCH™.

SNP ID	Allele	PWM	Consequence
rs333406887	0,1	V\$NF1_Q6	Score-Change
rs333406887	0,1	V\$AP2ALPHA_03	Score-Change
rs333406887	0	V\$CPBP_Q6	Loss of TFBS
rs333406887	1	V\$ZFX_01	Gain of TFBS

### 3.7. Discussion

Today, it is widely known that protein–DNA interactions govern the level of gene expression in all higher organisms to a great extent. The binding of TFs to the DNA mainly occurs in the regulatory regions, such as promoters, which are found close to the transcription start of genes [110]. The effect of rSNPs on the binding of TFs has been studied extensively in single case studies in different species, and, for humans, many tools and databases exist to facilitate these analyses (see Tables 3.1 and S1).

However, there is limited information available for livestock, and, to the best of our knowledge, there is no comparable data source for evaluating the effect of rSNPs. To address this



lack of information, we systematically carried out a genome-wide analysis to detect rSNPs and to evaluate their consequences for TF-binding in seven animal species, which can be accessed via a web server. We showed that, by substituting a single base in a predicted TFBS, a SNP can lead to a major change in the binding affinity of the TF and, in an extreme case, even result in the disruption of the TFBS or the creation of a new TFBS.

These predictions can be of great use for scientists who have conducted: (i) an association analysis and want to reveal the underlying mechanisms caused by a SNP being significantly associated with a trait (e.g., in [4, 38, 39, 81]); (ii) a gene expression experiment and want to identify candidate SNPs influencing the expression rate of a specific gene or a set of genes (e.g., in [4, 40, 42]); or (iii) a combination of both, i.e., an expression quantitative trait locus (eQTL) analysis (e.g., in [79]).

Even though our predictions are in line with many biologically tested results, as shown in the biological validation in Section 3.6, we note that the binding affinity of the TFs to the DNA sequence is one of the most important factors for TF binding but might not be sufficient for *in vivo* binding in higher organisms. Other influencing factors might include the chromatin accessibility, TF concentration, or other enhancing or repressing protein-DNA interactions, such as competitive or cooperative TF binding [17, 69, 111], which could not be considered in the prediction pipeline.

TF binding often occurs in a complex interplay and also includes cooperation between proximal and distal regulatory elements (promoters and enhancers) [68]. Thus, in addition to the binding of TFs in the proximal promoter regions, regulatory processes via TF-DNA interactions are also controlled by distal enhancer regions. Due to the limited knowledge of enhancer regions in livestock species, we could not incorporate these distal regulatory regions.

For our analysis pipeline, we defined a relatively wide promoter region of 7.5 kb upstream to 2.5 kb downstream of the TSS. Similarly large promoter regions were defined in previous studies ranging from 10 kb upstream to 10 kb downstream of the TSS [15, 74, 91, 95–100] in order to overcome inaccuracies in the TSS prediction [13] and to ensure the inclusion of the biological promoter. The user has to be aware that the biological promoter region is usually smaller [13], and our website gives the opportunity to filter for smaller, user-defined promoter regions for each single gene. These considered promoter regions and the definition of rSNPs in our study (see Section 3.4.2.3) led to a relatively large number of rSNPs per gene — for instance, an average of 95.04 rSNPs per gene in chicken.

Interestingly, our results regarding the distribution of genome-wide rSNPs relative to the TSS showed two different patterns. In chicken, pig, sheep, horse, and goat, we observed that the region around the TSS was rather protected from sequence variations (Figure 3.7) as it was found in previous studies [4, 13]. However, the data for cattle and dogs revealed a different picture, and we found an accumulation of SNPs and rSNPs around the TSS (Figure 3.8). This observation shows that the data stored in public databases, such as Ensembl, can show completely different patterns for different species, which could create biases for specific analyses.

### 3.8. Conclusions

To the best of our knowledge, agReg-SNPdb is the first database of regulatory SNPs for animal species of agricultural importance. It allows the users to investigate the predicted effect of an allele change on TF binding. The release of the database is an important step toward the understanding of gene regulation in the life sciences. Knowing whether a SNP causes a change in the binding affinity or even disrupts a TFBS or creates a new TFBS can be of predominant importance in order to interpret the results, from, e.g., GWAS experiments, gene expression experiments, or population studies.

The newly gained information can be used to help in genomic selection and marker establishment by identifying possibly causal rSNPs and revealing the underlying regulatory mechanisms of specific traits or diseases. Due to the regular updates of genomes as well as gene and SNP annotations, the database will be updated regularly, and, as future work, we will include several plant species with agricultural importance in agReg-SNPdb.

### 3.9. Supplementary Materials

The following supplementary material is available via the original publication <https://doi.org/10.3390/biology10080790>. Table S1: A comprehensive overview of recent studies that investigated the effects of SNPs on regulatory elements (extension of Table 1), Figure S2: Number of SNPs and genes per chromosomes for all species, Figure S3: The average numbers of rSNPs per gene for each chromosome for all species, Figure S4: Distribution of the distances between rSNPs and the TSS for all species.

## 4. agReg-SNPdb-Plants: A Database of Regulatory SNPs for Agricultural Plant Species

This chapter contains the article of the same name published in April 2022 in the MDPI journal *Biology* (<https://doi.org/10.3390/biology11050684>). For the sake of consistency within this thesis, the journal style is not adopted in this chapter.

This article is a joined work of Selina Klees<sup>1,2,\*</sup>, Felix Heinrich<sup>1</sup>, Armin Otto Schmitt<sup>1,2</sup> and Mehmet Gültas<sup>2,3,\*</sup>

<sup>1</sup>Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany

<sup>2</sup>Center for Integrated Breeding Research (CiBreed), Carl-Sprengel-Weg 1, Georg-August University, 37075 Göttingen, Germany

<sup>3</sup>Faculty of Agriculture, South Westphalia University of Applied Sciences, Lübecker Ring 2, 59494 Soest, Germany

\*Authors to whom correspondence should be addressed.

### **Author contributions by Selina Klees (S.K.):**

S.K. participated in the design of the study, conducted computational and statistical analyses, created the database and website, and wrote the final version of the manuscript. All authors have read and agreed to the published version of the manuscript.

### **4.1. Simple Summary**

In breeding research, the investigation of regulatory SNPs (rSNPs) is becoming increasingly important due to their potential causal role for specific functional traits. Especially for crop species, there is still a lack of systematic analyses to detect rSNPs and their predicted effects on the binding of transcription factors. In this study, we present agReg-SNPdb-Plants, a database storing genome-wide collections of regulatory SNPs for agricultural plant species which can be queried via a web interface.

## 4.2. Abstract

Single nucleotide polymorphisms (SNPs) that are located in the promoter regions of genes and affect the binding of transcription factors (TFs) are called regulatory SNPs (rSNPs). Their identification can be highly valuable for the interpretation of genome-wide association studies (GWAS), since rSNPs can reveal the biologically causative variant and decipher the regulatory mechanisms behind a phenotype. In our previous work, we presented agReg-SNPdb, a database of regulatory SNPs for agriculturally important animal species. To complement this previous work, in this study we present the extension agReg-SNPdb-Plants storing rSNPs and their predicted effects on TF-binding for 13 agriculturally important plant species and subspecies (*Brassica napus*, *Helianthus annuus*, *Hordeum vulgare*, *Oryza glaberrima*, *Oryza glumipatula*, *Oryza sativa* Indica, *Oryza sativa* Japonica, *Solanum lycopersicum*, *Sorghum bicolor*, *Triticum aestivum*, *Triticum turgidum*, *Vitis vinifera*, and *Zea mays*). agReg-SNPdb-Plants can be queried via a web interface that allows users to search for SNP IDs, chromosomal regions, or genes. For a comprehensive interpretation of GWAS results or larger SNP-sets, it is possible to download the whole list of SNPs and their impact on transcription factor binding sites (TFBSs) from the website chromosome-wise.

### Keywords

regulatory SNP; transcription factor; transcription factor binding site; gene regulation; GWAS; database; agricultural plant species; crops

## 4.3. Introduction

Climate change and its anticipated consequences pose severe challenges to mankind. For agriculture, global warming means that pathogens previously restricted to warmer climates will threaten local animal and plant species as well as expose plants to drought stress due to the increasing water shortage. A rapid and effective adaptation to the new environmental conditions is of paramount importance and can only be achieved through supportive plant breeding programs [112, 113]. While breeding once used to be a relatively slow process limited by the generation interval of the species under study, the advent of molecular biology technologies, particularly large-scale genotyping at the whole-genome level, has turned the tide [4, 114]. Today, genomic predictions aid the selection process in reproduction, and genome-wide association studies (GWAS) make it possible to identify the genomic loci that are beneficial or deleterious with respect to a trait under study. However, one remaining challenge is to identify not only genomic variants that are statistically associated with a trait, but also those that are actually biologically causative, because this would ensure their efficient use for breeding purposes [28]. In the search for causality of disease- or trait-associated SNPs, one often encounters regulatory SNPs (rSNPs) that influence the amount

of genetic material, and hence play a crucial role in the expression of a phenotype. Compared to SNPs in the exonic regions, predicting the consequences of SNPs in the promoter regions is not as straightforward [3, 4, 29, 38]. Such consequences could be the disruption or creation of one or more transcription factor binding sites (TFBSs), which can have a major impact on the level of gene transcription. To date, there exist many tools and databases for the prediction of rSNPs and their impact on regulatory elements such as TFBSs. However, most of them are restricted to the human genome or a few model organisms [17, 23, 46, 74, 89–91, 115, 116].

To the best of our knowledge, there exist currently three tools, which generally allow the analysis of plant rSNPs. As a web-based tool, the RSAT variation-tool [45] allows the analysis of user-provided inputs on the fly. However, this tool does not give any information on related genes, as the distance to the transcription start site (TSS) or consequences such as gain- or loss of TFBS, hence the users need to interpret the output themselves. The RSAT variation-tool includes eight crop species and subspecies (*Hordeum vulgare*, *Oryza sativa* Indica, *Oryza sativa* Japonica, *Solanum lycopersicum*, *Sorghum bicolor*, *Triticum turgidum*, *Vitis vinifera*, and *Zea mays*). The R packages MotifbreakR [46] and atSNP [47] principally comprise organisms stored in the Bioconductor BSGenome package [48], which includes only the crop species *Oryza sativa* and *Vitis vinifera*. In both, the user has to provide the SNPs as well as TFBSs (motifs represented as position weight matrices; PWMs) and experience in R programming is imperative.

In our previous studies, we addressed this limited knowledge and created a pipeline for the systematic detection of rSNPs, which we applied to different agriculturally important species such as rapeseed [4], faba bean [38], and various animal species [3]. By creating the database agReg-SNPdb [3], we have provided genome-wide collections of rSNPs for seven different animal species (cattle, pig, chicken, sheep, horse, goat, and dog). In order to extend the available information on rSNPs to additional plant species, we present in this study the database agReg-SNPdb-Plants, which can be considered as an extension of agReg-SNPdb. To the best of our knowledge, agReg-SNPdb-Plants is the first comprehensive database of genome-wide collections of rSNPs and their impact on TFBSs for agriculturally important plant species, which can be queried in various ways: (i) search by SNP ID, (ii) search by chromosomal region, (iii) search by gene, or (iv) a chromosome-wise download of all rSNPs. agReg-SNPdb-Plants includes various important crop species, i.e., Asian rice (Indica and Japonica), barley, bread wheat, durum wheat, grape, maize, rapeseed, sorghum, sunflower, and tomato as well as species, which can serve as genetic resources for the improvement of cultivated species, i.e., African rice and wild rice [117, 118]. The availability of rSNPs in rapeseed is particularly noteworthy because to date there exists no genome-wide SNP catalog in Ensembl Plants [119] for this crop. In contrast to the remaining species, where we used the data from Ensembl Plants as basis, we employed a SNP catalog from [49] for rapeseed, which we also used for our previous studies [4, 120]. The agReg-SNPdb-Plants web interface is available under <https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb-plants/> (accessed on 28 April 2022).

## 4.4. Materials and Methods

In our previous work, we have established a pipeline for the detection of rSNPs [3], which requires as input for each species a SNP catalog (as GVF file [121, 122]), a reference genome (as fasta file), and gene annotations (as GFF3 file [123]). For all species except for rapeseed, the input data were downloaded from Ensembl Plants [119], with genome assemblies listed in Table 4.1. The SNP catalog was filtered by removing insertions and deletions as well as SNPs with more than one alternate allele. Since there is no available SNP catalog for rapeseed in Ensembl plants, we used the rapeseed input data from our previous work [4]. This includes a SNP catalog of 670,028 high-quality SNPs (MAF > 0.05) from the cultivars Zhongshuang11 and Zhongyou821 (280 and 133 samples, respectively) collected and published by Lu et al. [49]. The *Brassica napus* reference genome (version 4.1) and gene annotations were obtained from [50] and are available at <https://www.genoscope.cns.fr/brassicanapus/data/> (accessed on 3 March 2022).

In brief, the pipeline can be described in the following five steps. For a more detailed description, we refer to [3].

1. **Selection of SNPs in the promoter and surrounding region:** For each gene, we considered a promoter region of 7.5 kb upstream to 2.5 kb downstream from the transcription start site (TSS) and selected all SNPs located within that region. On the website, the user has the possibility to insert a user-defined promoter region with the default being  $-1$  kb to  $+100$  bp.
2. **Extraction of the SNP-flanking region:** Using the reference genomes under study, we extracted 25 bp on each side of a SNP to obtain 51 bp long sequences with the SNP in the central position. During this step, we discarded sequences with a total length of less than 51 bp, sequences containing N's, and sequences in which the nucleotide at position 26 differed from the reference allele of the SNP (as specified in the SNP catalog in GVF format [121]). The latter only occurred in the species tomato, Asian rice (Indica Group), and sorghum.
3. **Creation of search sequences:** For each SNP, we created an additional copy of its 51 bp long sequence by replacing the reference allele with its alternate allele.
4. **TFBS prediction:** Applying the tool MATCH<sup>TM</sup> [18] with a plant-specific PWM library containing non-redundant matrices with specific cut-offs that minimize the false positive rate, we predicted TFBSs in the sequences of each SNP. The PWM library is provided by TRANSFAC [93].
5. **Annotation of consequences:** By comparing the two sets of predicted TFBSs, we assessed the consequences of each SNP on a specific TFBS. In particular, the effect of each SNP on a TFBS was assigned to one of the following consequences:
  - Gain of TFBS: the TFBS exists only for the alternate allele of the SNP.

- Loss of TFBS: the TFBS exists only for the reference allele of the SNP.
- Score-Change: the TFBS exists for both alleles but with differing binding affinity as determined by the MATCH<sup>TM</sup> scores.
- No Change: the TFBS exists for both alleles with the same binding affinity.

**Table 4.1.: Assembly versions of the input data from Ensembl Plants including reference genome, SNP catalog and gene annotations.**

Plant	Assembly Version	Download Date
<i>Helianthus annuus</i> (sunflower)	HanXRQr1.0	11/08/2021
<i>Hordeum vulgare</i> (barley)	MorexV3_pseudomolecules_assembly	12/22/2021
<i>Oryza glaberrima</i> (African rice)	Oryza_glaberrima_V1	11/08/2021
<i>Oryza glumipatula</i> (wild rice)	Oryza_glumaepatula_v1.5	11/08/2021
<i>Oryza sativa</i> Indica (Asian rice Indica)	ASM465v1	12/22/2021
<i>Oryza sativa</i> Japonica (Asian rice Japonica)	IRGSP-1.0	11/08/2021
<i>Solanum lycopersicum</i> (tomato)	SL3.0	12/22/2021
<i>Sorghum bicolor</i> (sorghum)	Sorghum_bicolor_NCBIv3	12/22/2021
<i>Triticum aestivum</i> (bread wheat)	IWGSC	11/08/2021
<i>Triticum turgidum</i> (durum wheat)	Svevo.v1	11/08/2021
<i>Vitis vinifera</i> (grape)	12X	11/08/2021
<i>Zea mays</i> (maize)	Zm-B73-REFERENCE-NAM-5.0	11/08/2021

## 4.5. Results

### 4.5.1. Database

agReg-SNPdb-Plants is centered around four tables: (i) *snp\_info* contains general information about the SNPs, (ii) *gene\_info* stores general information about the genes, (iii) *snp\_region* connects the tables *snp\_info* and *gene\_info* for all SNPs located in the promoter region of at least one gene, and (iv) *TFBS\_results* stores the rSNPs and their consequences with respect to TF-binding. Table 4.2 shows the numbers of database entries per table and species.

### 4.5.2. Web Interface

Following the concept of Ensembl and Ensembl Plants, we created an extra web interface for agReg-SNPdb-Plants (<https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb-plants/>, accessed on 28 April 2022). The basic functionality was inherited from agReg-SNPdb, e.g., the ability to query the database by searching for (i) SNP identifiers, (ii) SNP position,

**Table 4.2.: The number of records stored in the database tables *snp\_info*, *gene\_info*, *snp\_region*, and *TFBS\_results* separated by species.**

Plant	snp_info	gene_info	snp_region	TFBS_results
African rice	7,567,669	33,164	7,341,550	8,336,778
Asian rice Indica	4,340,785	37,878	4,589,915	4,441,820
Asian rice Japonica	25,135,669	37,960	20,155,983	20,940,720
Barley	12,771,762	35,106	2,545,069	2,736,205
Bread wheat	18,093,867	107,889	13,334,911	19,733,723
Durum wheat	1,815,904	66,559	1,121,107	1,734,495
Grape	400,940	29,971	334,500	290,793
Maize	48,830,598	44,289	15,439,220	13,101,269
Rapeseed	670,028	406,325	5,110,349	506,859
Sorghum	8,081,051	34,023	6,414,543	3,118,613
Sunflower	11,834	52,191	2335	1498
Tomato	60,973,560	33,869	28,709,218	10,347,415
Wild rice	4,865,161	35,735	4,752,796	5,154,313
Total	193,558,828	954,959	109,851,496	90,444,501

(iii) chromosomal region, or (iv) gene. Additionally, we enabled the search for several SNP IDs at a time, by pasting white-space separated SNP IDs in the search field.

Furthermore, we simplified the visualization of the *TFBS\_results*, which is shown exemplarily in Figure 4.1. The first column of table *TFBS\_results* (Figure 4.1) shows the SNP ID. This SNP ID should be the ID as specified in Ensembl Plants. An exception is the naming of the rapeseed SNP IDs, as they are not available in Ensembl Plants we used an annotation as *chr-pos-ref-alt*, e.g., A01-1093-A-G. The second column 'Gene strand' refers to the strand of the gene in whose promoter region the SNP is located (the gene strand hence also defines the strand of the sequence). If a SNP occurs in the promoter of two different genes, one on the plus and one on the minus strand, there will be two different tables showing the TFBSs for the plus and minus strands separately. The column 'PWM' (position weight matrix) represents the TFBS. The names of the PWMs are defined by TRANSFAC [93] as P\$*factorname*\_version, where the P\$ indicates that the PWM originated from a plant TF and *factorname* specifies the name of the represented TF. The core and matrix similarity scores are the MATCH™ [18] output scores. The 'Core similarity score' measures the quality of the match in the first five consecutive most-conserved positions of the PWM and the 'Matrix similarity score' measures the quality of the match for the whole PWM. The 'Sequence' shows the input sequence matching the PWM with the capital letters representing the core of the PWM and the nucleotides in red representing the SNP position. In case of a loss or gain only the allele for which a TFBS is observed is displayed while in case of a score-change or no change both alleles are displayed. The column 'Binding site' is a



schematic representation of the column 'Consequence', and depicts the presence or absence of a binding site for each allele.

Show  entries

Search:

SNP ID	Gene strand	PWM	Core Similarity Score	Matrix Similarity Score	Sequence	Binding sites	Consequence
<a href="#">10105262583</a>	-	PSANAC094_01	- / 0.760	- / 0.824	gGCCGCcgaggg[a]cgcg	Ref(C) ✖ Alt(T) ✔	Gain of TFBS
<a href="#">10105262583</a>	-	PSANAC094_01	- / 0.894	- / 0.827	gccgccgagg[A]CGCGt	Ref(C) ✖ Alt(T) ✔	Gain of TFBS
<a href="#">10105262583</a>	-	PSFAR1_01	1.000 / -	0.878 / -	aggg[g]CGCGTcccga	Ref(C) ✔ Alt(T) ✖	Loss of TFBS
<a href="#">10105262583</a>	-	PSANAC094_01	0.894 / 0.894	0.843 / 0.844	[g/a]CGCGTcccgcgctg	Ref(C) ✔ Alt(T) ✔	Score-Change
<a href="#">10105262583</a>	-	PSERF73_01	1.000 / 1.000	0.882 / 0.882	gcattggcCGCCGcaggg[g/a]cgc	Ref(C) ✔ Alt(T) ✔	No Change
<a href="#">10105262583</a>	-	PSLBD23_01	0.729 / 0.729	0.795 / 0.795	cGCCGCaggg[g/a]cgcg	Ref(C) ✔ Alt(T) ✔	No Change

Showing 1 to 6 of 6 entries Previous  Next

**Figure 4.1.:** Example of a search result from agReg-SNPdb-Plants showing table *TFBS\_results*. The search was performed with the SNP ID 10105262583 from Asian rice (Japonica Group).

### 4.5.3. Statistical Overview of the Data

Similar to our previous studies [3, 4], we first provide a brief overview of the data stored in agReg-SNPdb-Plants.

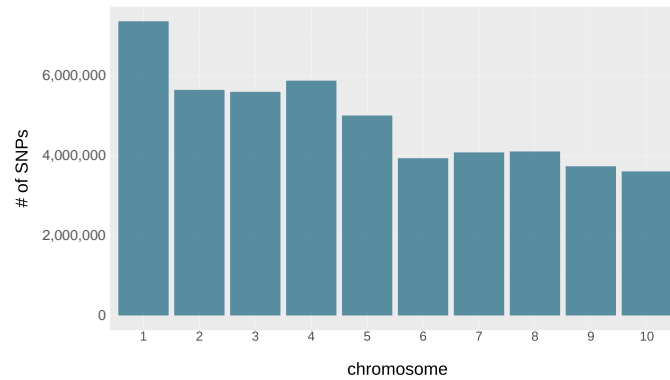
The distributions of SNPs and genes along the chromosomes are exemplary shown for maize (Figure 4.2; the remaining plots are given in Supplementary Figure S1). As expected, for maize and most other species the absolute numbers of SNPs and genes per chromosomes depend mainly on chromosome size and hence decrease in general with increasing chromosome numbers.

The average number of rSNPs (SNPs that cause a loss or gain of TFBS or a score-change for at least one TFBS) per gene differs strongly across the species. For example, in sunflower we only detected an average of 0.0015 rSNPs per promoter region ( $-1$  kb to  $+100$  bp) while we observed 28.48 rSNPs per promoter in tomato (absolute counts of SNPs and genes for each species can be seen in Table 4.2). Considering the  $-1$  kb to  $+100$  bp promoter region, on average  $\sim 4\%$  of all SNPs are predicted as rSNPs, with a minimum amount of  $0.6\%$  in sunflower and a maximum of  $13.6\%$  in rapeseed. When examining the number of TFBSs affected by an rSNP, we identified an overall average of  $\sim 2$  affected TFBSs per rSNP.

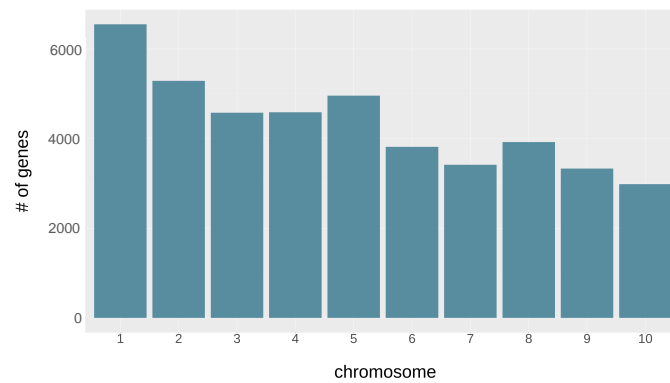
To obtain further insights into the data, we investigated the distribution of rSNPs relative to the TSS (Supplementary Figures S2). Similar to the animal species in agReg-SNPdb, we observed two different patterns for the distributions. The first pattern shows that the

## Maize

(A)

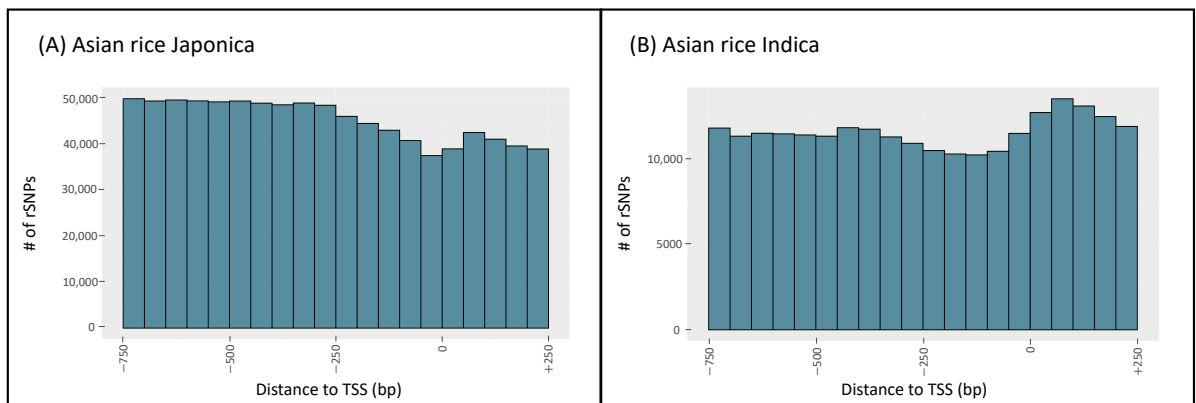


(B)



**Figure 4.2.:** The total number of SNPs and genes per chromosome of maize (*Zea mays*). (A) The number of SNPs per chromosome. (B) The number of genes per chromosome.

sequence is protected from variations in close proximity to the TSS, while the number of rSNPs increases with increasing distance in the upstream direction [3, 4, 13]. A similar pattern was observed in rapeseed, barley, Asian rice Japonica, maize, tomato, wild rice, and sorghum (Figures 4.3A and S2). The second pattern shows the opposite: The number of rSNPs increases with increasing downstream distance. This was observed in sunflower, African rice, Asian rice Indica, bread wheat, durum wheat, and grape (Figures 4.3B and S2). Figure 4.3 exemplary shows the comparison of the rSNP distance to the TSS for the two types of *Oryza sativa*, Japonica in (A) and Indica in (B).



**Figure 4.3.: Distribution of the distances between rSNPs and the TSS of (A) Asian rice Japonica and (B) Asian rice Indica.** The histograms show the number of rSNPs in the proximal promoter region ( $-750$  bp to  $+250$  bp) in 50 bp intervals.

## 4.6. Discussion

Transcription factors bind to the promoter region to fine-tune the level of gene expression in all higher organisms. A regulatory SNP within a TFBS can influence this transcriptional gene regulation to a great extent and hence could have a causative effect on the phenotype. In plants, several studies investigated (single) rSNPs with respect to a specific trait or phenotype [4, 35–38]. For example, Konishi et al. revealed an rSNP in rice that causes a loss of TFBS for an ABI3 type TF in the promoter region of the quantitative trait locus (QTL) for seed shattering on chromosome 1 (*qSH1*). This rSNP is causative for the loss of seed shattering and thus paved the way for rice domestication [35]. In maize, several rSNPs were detected in the promoter of the maize rough dwarf disease candidate gene eukaryotic translation initiation factor 4E (*eIF4E*) and control its expression level [36]. Furthermore, in wheat, an rSNP associated with wheat grain weight affects the binding of a calmodulin-binding TF and hence the gene expression of the *TaGW2-6A* gene, a candidate gene for grain weight [37]. Similar to these studies, in our previous study on the grain legume faba bean we discovered two rSNPs which are significantly associated with the vicine and convicine content and affect the binding of the TFs MYB4, MYB61, and SQUA [38]. To this end, we have investigated the seed oil content in rapeseed of the cultivars Zhongshuang11 and Zhongyou821 and obtained a genome-wide collection of rSNPs which are significantly associated with the oil content and positioned in promoter regions of genes differentially expressed between high and low oil content cultivars [4].

Due to the increasing interest in finding causative rSNPs yet limited availability of resources to detect rSNPs in crop species, we used our rSNP detection pipeline to systematically ana-

lyze 13 crop plants and provide a database of genome-wide rSNPs which can be queried via a web interface (<https://azifi.tz.agrar.uni-goettingen.de/agreg-snpdb-plants/>, accessed on 28 April 2022). This pipeline could be highly valuable for scientists to interpret their results from e.g., a GWAS or next-generation sequencing (NGS) experiments.

In our pipeline, one important step was the selection of the range of the promoter regions, since this determines if a SNP is considered for further analyses. Even though the core promoter is considered to be positioned within ~200 bp around the TSS [13], a wider promoter region can be targeted by TFs to regulate gene transcription. Previous studies defined different promoter regions for TFBS prediction, ranging from -10 kb to +10 kb [3, 15, 74, 91, 95–100] (the different promoter definitions and respective textual evidences are provided in Table S1). Therefore, we used a relatively wide promoter region ranging from -7.5 kb to +2.5 kb relative to the TSS, in order to ensure the inclusion of the regulatory regions. However, it is important to note that the biological promoter is usually smaller and, hence, our web interface provides the possibility to select a smaller user-defined promoter region. In total, we analyzed 13 species and subspecies for the construction of the agReg-SNPdb-Plants database, for twelve of which reference genome, gene annotations, and a SNP catalog were available in Ensembl Plants.

However, for some species the available information, e.g., the reference genome, might not be of the same quality compared to other, well-investigated species. Furthermore, due to the amount of repetitive sequences in some plant species such as bread wheat or maize, both the reference genome annotation as well as locating genomic variants can be challenging [124, 125]. The quality of the promoter region highly influences the quality of TFBS predictions and we want to emphasize that our predictions can only rely on the available information. For the species tomato, Asian rice (Indica), and sorghum, we observed that the alleles of several SNPs do not fit to the reference genome, in particular, their reference alleles were not present at the SNP position in the reference genome. An example for this issue, can be shown based on the tomato SNP *vcZYOCUX* (T/A), where the base at the respective position in the reference genome is G ([https://plants.ensembl.org/Solanum\\_lycopersicum/Variation/Explore?r=1:39003479-39004479;v=vcZYOCUX;vdb=variation;vf=3506065](https://plants.ensembl.org/Solanum_lycopersicum/Variation/Explore?r=1:39003479-39004479;v=vcZYOCUX;vdb=variation;vf=3506065), accessed on 28 April 2022). Such issues indicate that there is still a need for further investigation or updates to improve the genome sequences as well as SNP annotations. In our pipeline, we excluded such SNPs from further analysis to ensure the highest possible reliability of our results.

## 4.7. Conclusions

In breeding research, the knowledge about rSNPs can help to unravel the regulatory mechanisms underlying specific phenotypes and could hence lead to the identification of causal SNPs, which are of great importance for the establishment of robust markers. To the best of our knowledge, until now there exists no database storing genome-wide rSNPs and their

consequences on TF binding in plant sciences which can be queried in various ways. In order to address this lack of information, and thus complementing our previous work, we created agReg-SNPdb-Plants, a database of rSNPs for 13 agricultural plant species and subspecies with currently available SNP annotations. Its web interface is a helpful resource for scientists who are conducting association analyses such as GWAS, gene expression experiments, expression QTL (eQTL) studies, or population studies. Consequently, they can automatically investigate the candidate SNPs or specific genes to rate them by their importance or causality. In this regard, our user interface provides different search functions and delivers information on the consequences of rSNPs on TF binding such as (i) gain of TFBS, (ii) loss of TFBS, (iii) change of binding affinity, or (iv) no change. Due to regular updates of genomes, gene- and SNP-annotations, our database will be regularly updated to add new plant species when available and to update existing ones.

#### **4.8. Supplementary Materials**

The following supplementary material is available via the original publication <https://doi.org/10.3390/biology11050684>. Figure S1: Number of SNPs and genes per chromosome for all species, Figure S2: Distribution of the distances between rSNPs and the TSS for all species, Table S1: Different promoter definitions and textual evidences from previous studies.



## 5. In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in *Brassica napus* L. Using Multi-Omics Data

This chapter contains the article of the same name published in January 2021 in the MDPI *International Journal of Molecular Sciences* (<https://doi.org/10.3390/ijms22020789>). For the sake of consistency within this thesis, the journal style is not adopted in this chapter.

This article is a joined work of Selina Klees<sup>1</sup>, Thomas Martin Lange<sup>1</sup>, Hendrik Bertram<sup>1</sup>, Abirami Rajavel<sup>1</sup>, Johanna-Sophie Schlüter<sup>1</sup>, Kun Lu<sup>2,3,4</sup>, Armin Otto Schmitt<sup>1,5</sup>, and Mehmet Gültas<sup>1,5,\*</sup>

<sup>1</sup>Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany

<sup>2</sup>College of Agronomy and Biotechnology, Southwest University, Chongqing 400715, China

<sup>3</sup>Academy of Agricultural Sciences, Southwest University, Chongqing 400715, China

<sup>4</sup>State Cultivation Base of Crop Stress Biology for Southern Mountainous Land of Southwest University, Chongqing 400715, China

<sup>5</sup>Center for Integrated Breeding Research (CiBreed), Albrecht-Thaer-Weg 3, Georg-August University, 37075 Göttingen, Germany

\* Author to whom correspondence should be addressed.

### **Author contributions by Selina Klees (S.K.):**

S.K. participated in the design of the study, conducted computational and statistical analyses, was involved in the interpretation of the results, and wrote the final version of the manuscript.

## 5.1. Abstract

Regulatory SNPs (rSNPs) are a special class of SNPs which have a high potential to affect the phenotype due to their impact on DNA-binding of transcription factors (TFs). Thus, the knowledge about such rSNPs and TFs could provide essential information regarding different genetic programs, such as tissue development or environmental stress responses. In this study, we use a multi-omics approach by combining genomics, transcriptomics, and proteomics data of two different *Brassica napus* L. cultivars, namely Zhongshuang11 (ZS11) and Zhongyou821 (ZY821), with high and low oil content, respectively, to monitor the regulatory interplay between rSNPs, TFs and their corresponding genes in the tissues flower, leaf, stem, and root. By predicting the effect of rSNPs on TF-binding and by measuring their association with the cultivars, we identified a total of 41,117 rSNPs, of which 1141 are significantly associated with oil content. We revealed several enriched members of the TF families DOF, MYB, NAC, or TCP, which are important for directing transcriptional programs regulating differential expression of genes within the tissues. In this work, we provide the first genome-wide collection of rSNPs for *B. napus* and their impact on the regulation of gene expression in vegetative and floral tissues, which will be highly valuable for future studies on rSNPs and gene regulation.

### Keywords

rSNPs; transcription factor; multi-omics; gene expression; random forest; DOF

## 5.2. Introduction

With rapidly evolving genomic sequencing technologies, the number of identified single nucleotide polymorphisms (SNPs) is increasing at a remarkable pace. Due to their straightforward functional interpretation, SNPs located in the protein coding regions of the genes are mostly in the focus of research. However, results from genome-wide association studies (GWAS) reveal that the vast majority of phenotype-associated SNPs are located in intergenic and intronic regions [28, 29]. Many of these non-coding SNPs are located within the regulatory regions, such as the promoter regions, and could hence influence the gene expression by changing the binding affinity of regulatory proteins. In recent years, these so-called regulatory SNPs (rSNPs) have come into the focus of research and the underlying mechanisms resulting in a differential gene expression are closely studied for many specific traits and diseases [29, 75]. It is well known that the differential gene expression in different tissues and under certain environmental conditions is governed by the binding of transcription factors (TFs) to specific DNA-sequence motifs, the transcription factor binding sites (TFBSs). By altering the sequence within such a TFBS, an rSNP can have a severe effect on TF binding and, hence, could change a gene's expression rate [29, 36, 75]. In plant sciences, previous studies identified different putative rSNPs affecting different traits,



as e.g., seed shattering in rice [35], maize rough dwarf disease [36], grain weight in wheat [37], or vicine and convicine content of *Vicia faba* [38]. Until now, there are several tools predicting a SNP's impact on TF binding (e.g., [46, 47, 75, 89, 91, 126]), but the Regulatory Sequence Analysis Tool (RSAT) [126] is one of the few tools supporting plants. In RSAT, users have the possibility to retrieve specific genetic variants with the corresponding flanking sequences and predict their impact on TF binding in a variety of organisms [126]. However, all these studies and tools concentrate on single regulatory variants and do not cover a systematic analysis to obtain a genome-wide prediction of rSNPs. Notwithstanding that the importance of rSNPs and their regulatory power is well known, no such systematic analysis including a genome-wide prediction of rSNPs for *Brassica napus* L. exists.

As an important oilseed crop, *B. napus* is grown and used worldwide for its oil and fodder production where the oil is widely used for human consumption and biofuel production, while the rapeseed meal remaining after oil extraction can be used as high-protein animal fodder [51, 52]. *B. napus* has gained global importance due to an intensive breeding program focusing on the reduction of nutritionally undesirable components in the oil and fodder and thus, enabled the production of varieties with both low erucic acid and glucosinolate content [127]. Today, improving the oil content is an important breeding goal and in this context the resistance to several stresses is a relevant objective [51, 53, 56]. The oil is stored within the seeds as triacylglycerols (TAGs) in oil bodies, but the TAG synthesis takes place in the plastids through a variety of different interacting metabolic pathways and regulatory processes [57]. However, such pathways as well as the underlying transcriptional machinery controlling the oil content and -quality could vary across different *B. napus* cultivars [49, 58]. Hence, the investigation of such biological processes is an important task to assess the genetic programs of two cultivars: (i) Zhongshuang11 (ZS11) characterized by a double-low accession (00, low erucic acid and low glucosinolate) and a high oil content and; (ii) Zhongyou821 (ZY821) with double-high accession (++, high erucic acid and high glucosinolate) and low oil content [49].

To unravel such genetic programs in both *B. napus* cultivars, we computationally identified the regulatory processes controlling specific biological functions associated with oil content, plant growth, or responses to environmental stresses. For this purpose, we used multi-omics data including genomics and transcriptomics data of two cultivars and plant proteomics data to identify rSNPs, important genes and transcriptional regulators orchestrating specific genetic programs in different tissues and thus, leading to phenotypic differences of both cultivars. To this end, mainly focusing on the vegetative and floral tissues such as flower, leaf, stem, and root, we first identified differentially expressed genes (DEGs) between both cultivars in these four tissues. Second, by analyzing 670,028 high-quality SNPs, we obtained a genome-wide collection of rSNPs and their predicted consequences on the binding affinity of the TFs. Similar to our previous studies [128, 129], we applied a random forest (RF) feature selection approach to assess the importance of rSNPs with respect to the phenotype. Subsequently, we determined tissue-specific DEGs harboring those important rSNPs within their promoter region, whose transcription is likely to be affected by the con-

sequences of rSNPs on TF binding. By causing a disruption of a TFBS or the creation of a new TFBS, rSNPs can strongly influence the binding affinity of TFs and, thus, can lead to the differentiation of a wide range of genetic processes in both cultivars like their oil content, tissue development, or stress-resistance mechanisms [51, 130]. Our results show that the consideration and systematic analysis of multi-omics data (genomics, transcriptomics, and proteomics) of two different *B. napus* cultivars provides: (i) essential information about functions of transcription factors involved in the regulation of transcriptional activity of vegetative and floral tissues; and (ii) novel insights into the regulatory programs controlling oil content and -quality underlying both cultivars.

### 5.3. Results and Discussion

#### 5.3.1. Differentially Expressed Genes

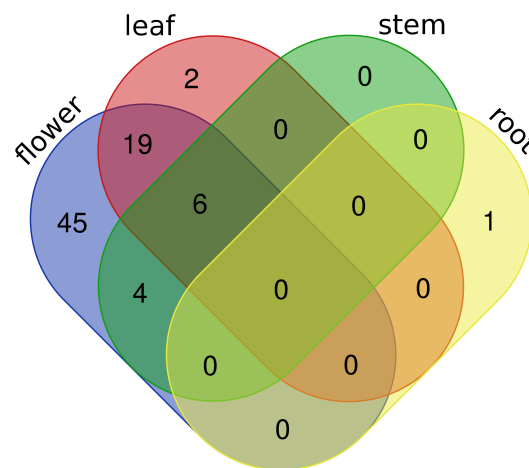
The comparison of the ZS11 (high oil content, double-low cultivar) against the ZY821 (low oil content, double-high cultivar) in the four tissues revealed several differentially expressed genes, of which the up-regulated DEGs refer to genes with a significantly higher expression in ZS11 than in ZY821, whereas down-regulated genes are significantly higher expressed in ZY821 than in ZS11 (Table 5.1, the full lists of DEGs is given in Table S1). The overlap of the four tissues showed that 171 and 252 DEGs were found up- and down-regulated in all four tissues, respectively. To assess the underlying biological processes, we provide the Gene Ontology (GO) terms and treemaps for the respective up- and down-regulated DEG sets in Table S2 and Figure S1.

**Table 5.1.: Numbers of differentially expressed genes (DEGs) in four tissues based on the comparison of the cultivars Zhongshuang11 (ZS11) against Zhongyou821 (ZY821).** Up-regulated and down-regulated DEGs are defined as  $\log_2$  fold change  $>2$  and  $\log_2$  fold change  $<-2$  and an adjusted  $p$ -value threshold of 0.05, respectively.

Tissue	No. of DEGs	No. of Up-Regulated DEGs	No. of Down-Regulated DEGs
Flower	11,442	5221	6221
Leaf	3234	1486	1748
Stem	4198	2510	1688
Root	2318	1448	870

### 5.3.2. Transcription Factor Binding Site Enrichment Analysis

For understanding the expression behavior of DEGs regarding their up- or down-regulation, the knowledge on the TFs, which are involved in controlling the regulatory programs of these genes, is important to explain gene expression changes between both *B. napus* cultivars. Applying TFBS enrichment analyses, we assessed the potential roles of TFs in the regulation of the DEGs based on the over-representation of their TFBSs in the promoter regions. In the following, we refer to a TF as enriched in a tissue, if its respective TFBS is significantly over-represented in the set of promoter sequences of the DEGs in that tissue. The results of these analyses show that the number of enriched TFs is remarkably different between tissues: While the largest number of enriched TFs was identified in the flower tissue (74), there was only one enriched TF in the root tissue. We further found 27 and 10 TFs enriched in the leaf and stem tissues, respectively (Figure 5.1; the complete list of enriched TFs is given in Table S3).



**Figure 5.1.:** Venn diagram for the enriched transcription factors (TFs) found for the tissues flower, leaf, stem, and root of *B. napus* (visualized with <http://bioinformatics.psb.ugent.be/webtools/Venn/>).

Interestingly, Figure 5.1 shows that the number of unique enriched TFs found for flower is clearly higher than those of the remaining tissues. In this regard, the transcription factor GATA19 found only for the root tissue is a member of GATA-type zinc finger proteins, which are known to be involved in light-mediated gene expression and nitrogen-dependent stress response [131, 132].

Furthermore, the TCP family members TCP16 and ARALYDRAFT\_897773 (also known as TCP4) were identified as enriched only in the leaf tissue. As shown in previous studies, TCP genes participate in the developmental control of plant form as, e.g., flower and leaf

shape or shoot branching by regulating cell proliferation and they have been shown to be highly expressed in leaf [133, 134].

Moreover, a minority of the TFs (PIF1, PIF7, bHLH74, UNE10, OJ1058\_F05.8, and BEH2) are simultaneously enriched for flower, leaf, and stem tissues. Besides the two phytochrome interacting factors (PIF1 and PIF7), the factors PIF3, PIF4, and PIF5 are enriched only in flower and leaf. The PIFs belong to one of the largest classes of plant TFs, the basic/helix-loop-helix (bHLH) proteins [58], and they are known to repress photomorphogenesis in darkness by promoting the transcription of genes which positively regulate cell elongation in *A. thaliana* [135]. In particular, while PIF1 has been reported to negatively regulate seed germination in response to light and hormone signaling [136, 137], PIF4 and PIF5 are regulators of de-etiolation [138], and PIF7 is a main regulator of stem elongation in light [139].

The factor unfertilized embryo sac10 (UNE10) is another member of the bHLH class. It is supposed to inhibit far-red light signaling by interacting with phytochromes [140] and to play an important role during the fertilization of ovules by pollen in *A. thaliana* [141].

Several BES1 (BRI1-EMS-SUPPRESSOR1) family members, in particular BEH2, BEH3, BEH4, and BZR1, are enriched in flower and leaf and/or stem tissues, and are known to regulate brassinosteroid-mediated genes. Different BES1-family members are suggested to regulate different auxin and jasmonic acid-related genes, resulting in enhanced growth and vigor in *B. napus* and *A. thaliana* [142, 143] and to be involved in stress resistance such as salt and drought stress in *B. napus* and *B. rapa* [144, 145].

Interestingly, we found members of the TF families MYB (or MYB-related; MYB46, MYB98, MYB119, MYB59, and MYB111), DOF type C2H2 zinc finger factors (DOF4.2, OBP3, AT2G28810 (DOF2.2), AT5G02460 (DOF5.1), and AT5G66940 (DOF5.8)), and NAC (NAC080, NAC028, NAC025, NAC058, NAC055, NAC043, NAC083, and T11I18.17) enriched exclusively in flower. The MYB TFs are involved in several processes as, e.g., response to biotic and abiotic stress, development, and differentiation; in particular, MYB46 is involved in secondary wall and fiber biosynthesis; MYB98 and MYB119 are important regulators of female gametophyte development; MYB59 is involved in cell cycle progression, and MYB111 plays a crucial role in flavonol biosynthesis in *A. thaliana* [146–149].

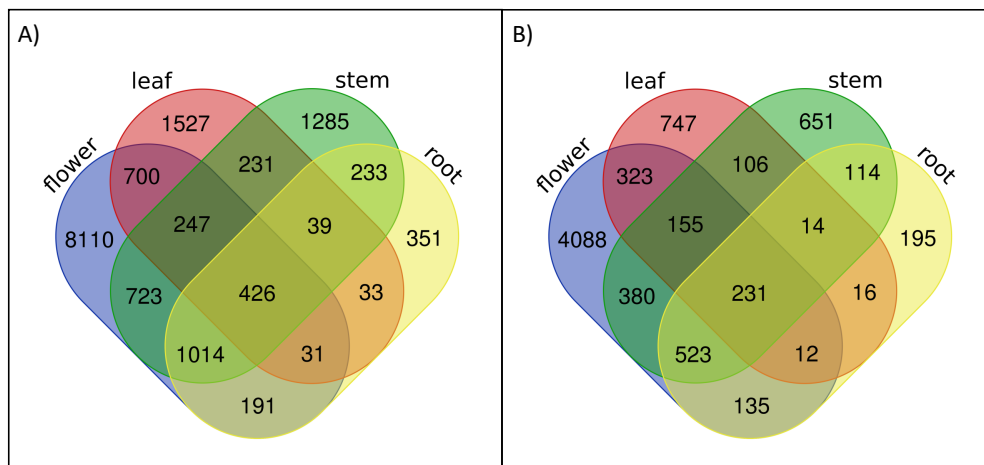
On the other hand, the DOF proteins are characterized by a highly conserved DNA binding domain (DOF domain) and are present in many different plant species [150]. DOF proteins play a role in several biological processes as, e.g., flowering time, seed development, and responses to hormones and abiotic stress [150–152]. Interestingly, He et al. (2015) [151] found the *Arabidopsis* DOF5.8 to be an upstream regulator of a gene encoding an NAC family member in response to drought and salt stress.

The NAC transcription factors make up one of the largest plant-specific TF families with specific functions regarding plant development, biotic stress response, and response to environmental stress [153]. Research performed in *B. napus* revealed upregulation of NAC genes after mechanical wounding and infection with *Sclerotinia sclerotiorum*. In the same

way, NAC genes were upregulated after the induction of a cold shock [154]. Interestingly, we have shown that members of TF families such as GATA, DOF, NAC, or MYB are important regulators of genes with a monotonic expression pattern in both cultivars in the seed tissue by forming TF co-operations [120].

### 5.3.3. Analysis of Regulatory SNPs

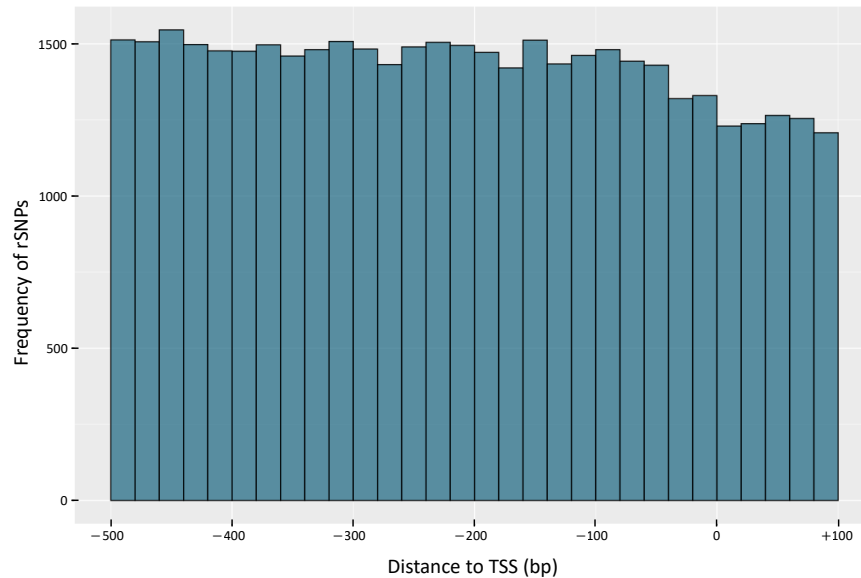
Today it is well known that the binding affinity of TFs can be affected by rSNPs to a great extent and, hence, either enable or repress the protein–DNA interaction. In order to be able to explain the observed differences in the expression of the DEGs, we investigated the role of rSNPs causing such severe effects on TF binding. Taking the initial 670,028 high-quality SNPs into account, we determined 41,117 of them as rSNPs due to their genomic positions in the promoter regions of *B. napus* genes and their consequences of either “Gain of TFBS” or “Loss of TFBS”. A closer look at these rSNPs reveals that 5847 (flower), 1604 (leaf), 2174 (stem) and 1240 (root) rSNPs are related to the DEGs (the full list of rSNP predictions can be found in Table S4). Interestingly, a direct comparison of the rSNP and DEG numbers shows that approximately 50% of DEGs contain on average one rSNP within the promoter region (Figure 5.2).



**Figure 5.2.: Overlap of the DEGs in (A) and rSNPs in (B) for the four investigated tissues** (visualized with <http://bioinformatics.psb.ugent.be/webtools/Venn/>).

To gain a better insight into the distribution of the rSNPs in the promoter regions, we investigated their genomic positions relative to the transcription start sites (TSS). The results of this analysis indicate that while there are fewer rSNPs around the TSS, we observed a tendency of increasing rSNP numbers in the remaining upstream promoter regions (Figure 5.3).

This finding goes in line with the observation of Triska et al. [13], who performed a similar analysis based on the SNP distributions in the promoters of rice.



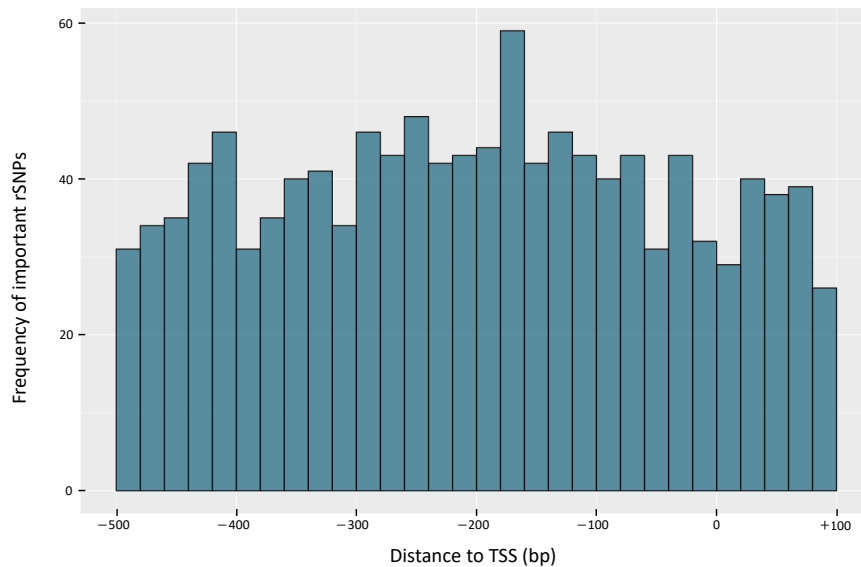
**Figure 5.3.: Distribution of rSNPs relative to the transcription start sites (TSS) of the corresponding genes.** Position 0 denotes the position of the TSS.

### 5.3.4. Analysis of Important Regulatory SNPs

Moreover, we assessed the importance of rSNPs regarding their significant association with oil content of both cultivars and identified 1141 *important rSNPs* (the complete list of *important rSNPs* is given in Table S5). The consideration of the *important rSNPs* in the DEGs of the tissues consequently results in 245 *important rSNPs* in the flower, 68 in the leaf, 142 in the stem and 82 in the root tissue. Surprisingly, the distribution of *important rSNPs* relative to the TSS (see Figure 5.4) shows a behavior in the promoter regions, that is remarkably different from that of rSNPs (Figure 5.3). This finding suggests that the *important rSNPs* do not follow a certain pattern but rather spread throughout the considered promoter regions.

### 5.3.5. DEGs Harboring *Important rSNPs* in the Promoter Region

In order to assess the regulatory impact of *important rSNPs* on the regulation of the DEGs and, hence, to explain their differential expression status, we identified the regulatory interplay between rSNPs, TFs and their corresponding DEGs of interest.



**Figure 5.4.:** Distribution of *important rSNPs* relative to the transcription start sites (TSS) of the corresponding genes. Position 0 denotes the position of the TSS.

As a result, we found 145, 44, 81, and 50 DEGs harboring *important rSNPs* in the promoter region for flower, leaf, stem, and root, respectively.

To gain a deeper insight into the functions of these genes, we identified their related enriched GO terms as well as the pathways (Table S6). Regarding enriched GO terms of biological processes, we could observe DEGs related to protein folding, alcohol, lipid or phytosteroid biosynthesis in leaf and a variety of genes related to oxidation–reduction processes in the leaf and flower tissue.

Interestingly, the gene *BnaA06g33360D* occurs in the flower and stem set of DEGs harboring *important rSNPs*, which leads to the significant enrichment of the monoterpene biosynthesis pathway [155]. Monoterpenoids are volatile secondary plant products that could play a role in olfactory cues for pollinating insects in *A. thaliana* [156]. Surprisingly, the gene *BnaA06g33360D*, which presumably codes for a monoterpene synthase, is down-regulated in flower tissue, while it is up-regulated in stem tissue.

In the gene set of the leaf tissue, several KEGG pathways [155] related to fatty acid metabolism were enriched. Especially the gene *BnaA04g26960D*, which is significantly up-regulated in the leaf tissue, is represented in the enriched pathways fatty acid metabolism, fatty acid biosynthesis, fatty acid degradation or peroxisome. *BnaA04g26960D*, also called *BnaLACS1-4*, is a member of the long-chain Acyl-CoA synthetase (*LACS*) family of genes, which have been shown to be involved in fatty acid biosynthesis in chloroplasts and seed

oil accumulation in *B. napus* [157]. Furthermore, several *LACS* genes showed differential gene expression in multiple tissues in the comparison between high and low oil content *B. napus* cultivars [157]. Within the promoter region of *BnaA04g26960D*, we identified one *important rSNP* (chromosome A04, position 19042835, C → T) which causes a “Gain of TFBS” for the binding site of MNB1A (the maize DOF1 TF) 90 bp downstream of the TSS. More specifically, this means the DOF1 binding site is not present in the reference allele (C), while DOF1 binding is likely to be enabled by the alternate allele (T). The importance of DOF-mediated gene regulation has already been shown in the results of TF enrichment (see Section 5.3.2). Interestingly, the soybean DOF proteins GmDOF4 and GmDOF11 have been shown to directly induce *LACS* genes, and also increased the fatty acid content in transgenic *Arabidopsis* seeds [150, 158]. In cotton, an overexpression of the *GhDOF1* gene led to an increase of lipid levels in the seeds [150, 159]. These results suggest that this *important rSNP* might play an important role in the DOF1-mediated expression rate of the *LACS* gene *BnaA04g26960D* and, hence, might regulate the fatty acid content in *B. napus*. The gene *BnaC08g26140D*, present in the significantly enriched pathways fatty acid metabolism, biosynthesis of unsaturated fatty acids and fatty acid elongation of the leaf gene set, encodes a Trans-2,3-enoyl-CoA reductase (ECR). This enzyme is involved in the synthesis of very-long-chain fatty acids (VLCFAs) which are essential for the synthesis of cuticular waxes, sphingolipids and Triacylglycerols (TAGs) in *B. napus* [160]. As an enzyme of VLCFA synthesis, it is also known to catalyze the fourth reaction of the elongase complex during erucic acid synthesis [160, 161]. Surprisingly, we found the *ECR* gene up-regulated in the double-low cultivar with high oil content. One possible explanation for its up-regulation in the low erucic acid cultivar might be that the synthesized VLCFAs are precursors for a variety of different lipids in higher plants, such as cuticular waxes [160]. In the promoter region of the *B. napus ECR* gene, we found three *important rSNPs* (hereinafter referred to as ECR-rSNP1, ECR-rSNP2 and ECR-rSNP3), affecting five different binding sites. ECR-rSNP1 (chromosome C08, position 27619847, G → T) is positioned –152 bp from the TSS and causes a “Loss of TFBS” for the *Arabidopsis* response regulator (ARR10) or response regulator 10 (RR10) binding site. As a cytokin response regulator, RR10 is involved in cytokinin-mediated signaling pathways and acts, e.g., as negative regulator of drought response in *A. thaliana* [162]. In *B. napus*, it has been shown to be up-regulated in leaves under salt stress [163]. The ECR-rSNP2 (chromosome C08, position 27619942, 247 bp upstream of the TSS, A → G) causes a “Loss of TFBS” for TF DOF4.5 and a “Gain of TFBS” for TF MYB56. The DOF4.5 is another member of the DOF family of TFs, which is assumed to share regulatory functions in, e.g., shoot branching and seed coat formation together with other DOF family members in *A. thaliana* [152]. MYB56 is a member of the previously described MYB family and is known to be a positive regulator of seed size and to control seed coat development in *Arabidopsis* [57, 164]. The ECR-rSNP3 (258 bp upstream of the TSS) causes a “Loss of TFBS” for TF DOF4.5 and a “Gain of TFBS” for ethylene-responsive transcription factor ERF069. Within the AP2/EREBP superfamily of TFs, ERF069 belongs to the ethylene-responsive element bind-

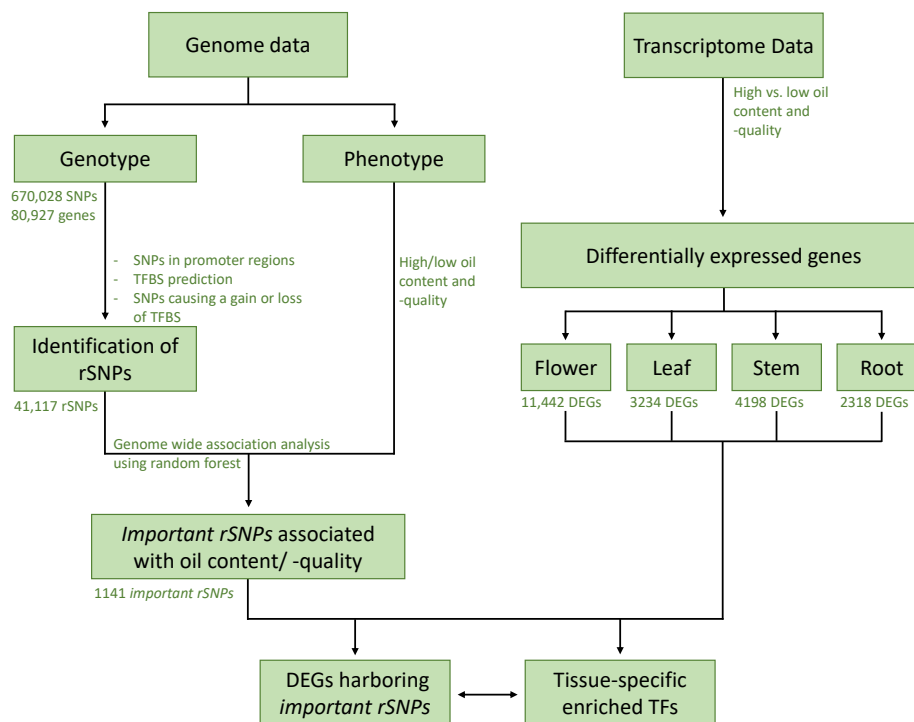


ing proteins (EREBP) subfamily, which are known to respond to abiotic stress [165]. Liu et al. (2020) [165] observed an up-regulation of ERF069 in response to chromium treatment in *A. thaliana*. In the foxtail millet, *SiAP2/ERF-069* was up-regulated under drought and salinity stress [166] and in *B. napus* and *ERF069* was up-regulated under Pi-starvation in 3- and 5-leaf stage seedlings [167].

With this analysis, we identified several interesting tissue-specific DEGs whose regulation is likely to be influenced by the “Loss-” or “Gain of a TFBS” caused by an *important rSNP* within their regulatory region. The TFs overlapping these *important rSNPs* provide a promising basis for further investigation of their regulatory roles and underlying pathways that lead to the distinction between the two cultivars.

## 5.4. Materials and Methods

Our analysis framework follows the structure shown in Figure 5.5, i.e., we start with the analysis of genomics and transcriptomics data to systemically monitor the important (tissue-specific) regulatory SNPs and TFs regulating the DEGs.



**Figure 5.5.: Flowchart of the analysis applied in this study.**

### 5.4.1. *B. napus* Data Set and Data Preparation

In this study, we use publicly available genomics and transcriptomics data sets of two *B. napus* cultivars, which are briefly explained below. Readers who are interested in learning more about these data sets are kindly referred to the original study [49].

#### 5.4.1.1. Genotype Data

To identify the rSNPs that are likely to be associated with different genetic programs in two *B. napus* cultivars, namely Zhongshuang11 (ZS11) with double-low accession (low erucic acid and glucosinolate, 00) and high oil content and Zhongyou821 (ZY821) with double-high accession (high erucic acid and glucosinolate, ++) and low oil content, we analyzed a genotype data set that has previously been used in [49]. Prof. Kun Lu from the Southwest University, China provided the genotype data set for this study. The raw sequencing data are available at the BIG Data Center under BioProject accession code PRJCA000376. The genotype data set comprises 670,028 high-quality SNPs (MAF > 0.05) for 280 Zhongshuang11 (ZS11) and 133 Zhongyou821 (ZY821) samples. The data set contains SNPs which are located on the chromosomes A01-A10 and C01-C09 (originated during hybridization of *B. rapa* (AA, 2n = 20) and *B. oleracea* (CC, 2n = 18) [49]) including 80,927 genes.

#### 5.4.1.2. Transcriptome Data

The RNA-sequencing data of four tissues (flower, leaf, stem, and root) from both cultivars (ZS11 and ZY821) with two biological replicates each were generated by Lu et al. [49]. The raw sequencing data were downloaded from the BIG Data Center under BioProject accession code PRJCA001246. In line with [49], we mapped the filtered reads to the *B. napus* reference genome version 4.1 (obtained from [50] and available at <https://wwwdev.genoscope.cns.fr/brassicnapus/data/>) using STAR 2.4.2a [168]. Finally, applying the htseq-count program [169] to the aligned sequencing reads, we identified the number of reads (gene count table).

For the identification of differentially expressed genes (DEGs), we applied the DESeq2 tool (R package version 1.24.0) with default settings in the median-of-ratios normalization method, fold change shrinkage and a significance cut-off of an absolute  $\log_2$  fold change of 2 and an adjusted p-value of 0.05 [170]. The experimental design of the differential expression analysis is shown in Table 5.2.

**Table 5.2.: Meta data of the RNA-seq experiment samples which were used for differential expression analysis.** ZS11 and ZY821 stand for Zhongshuang11 and Zhongyou821, respectively. 00 and ++ stand for low erucic acid, low glucosinolate and high erucic acid, high glucosinolate, respectively.

Cultivar	Oil Quality	Oil Content	Biological Replicates
ZS11	00	high	2
ZY821	++	low	2

#### 5.4.2. Transcription Factor Binding Site Enrichment Analysis in Promoter Sequences

In order to identify transcription factors (TFs) with significantly over-represented transcription factor binding sites (TFBSs) in the promoter sequences of the DEGs, we employed the CiiiDER algorithm [171].

However, the selection of the promoter regions is crucial: (i) to avoid the redundancy between sequences which could lead to the overestimation of some TFBSs [110] (ii) to address the inaccuracy of transcription start site (TSS) positions resulting from their imprecise prediction. To overcome these issues, we followed a similar strategy to those suggested in previous studies [13, 38, 68, 110, 172–176] and accordingly extracted two sets of promoter sequences for each tissue ranging from  $-500$  bp to  $+100$  bp relative to the TSS using the reference genome version 4.1 and gene annotation given in [50]. While the first sequence set refers to the promoter sequences of the DEGs (foreground set), the second set contains the promoter sequences of genes having the same GC-content as the foreground set (background set) [177]. For the generation of background sets, we used the oPOSSUM3.0 [98] web application ([http://opossum.cisreg.ca/GC\\_compo/](http://opossum.cisreg.ca/GC_compo/)) and selected only sequences that are not included in the foreground set. Second, following the workflow of the CiiiDER program [171], we scanned each sequence by applying the MATCH<sup>TM</sup> program [18] with a non-redundant plant position weight matrix (PWM) library from the JASPAR database [178] to detect the potential TFBSs. Finally, comparing the distribution of TFBSs predicted in the foreground as well as the background promoter sequence set, the enrichment of TFBSs was assessed (Bonferroni adjusted p-value threshold of 0.01).

#### 5.4.3. Identification of Regulatory SNPs and Their Importance

Following the regulatory SNP (rSNP) detection method of Heinrich et al. [38], we selected the SNPs from the genome data which are located in the promoter regions of *B. napus* genes and analyzed them to detect their impact on the TFBSs. For this purpose, we first extracted the flanking sequence of  $\pm 25$  bp for each selected SNP resulting in a 51 bp long sequence with the SNP in the central position. Second, we created two copies of the

flanking sequence: One with the reference allele in the SNP position, the second with its alternate variant. After that, employing the MATCH<sup>TM</sup> algorithm [18], both sequences were scanned to predict the TFBSs with their affinity scores. The potential binding affinity of a TF was quantified by MATCH<sup>TM</sup> in terms of a matrix similarity score (MSS)  $\in [0, 1]$ , where a MSS value of 1 denotes a complete match in each position of the TFBS. As suggested in [171], we removed all TFBS predictions with MSS values  $< 0.85$  and which did not overlap the SNP position in the flanking sequences. Finally, to evaluate the impact of a SNP on the binding affinity of a TF, we inferred four different types of consequences for each SNP-TFBS pair: (i) “No Change”: the SNP has no effect on the TF binding; (ii) “Score-Change”: the binding affinity (i.e., MSS) is changed; (iii) “Loss of TFBS”: a TFBS is only found on the reference allele, while the TFBS does not occur in the alternate allele; and (iv) “Gain of TFBS”: the TFBS appears only for the alternate allele. In the following, we define a SNP as rSNP if it causes a “Gain of TFBS” or a “Loss of TFBS” (consequence iii or iv) for at least one TFBS.

#### 5.4.4. Association Analysis Using Random Forests

For the assessment of the importance of single rSNPs, regarding their association to the *B. napus* cultivars, we applied a random forest (RF)-based feature selection algorithm to measure the relative importance of each rSNP for the trait oil content (congruent with oil quality, see Table 5.2), following our previous studies [128, 129]. In particular, the relative importance of each rSNP is calculated by applying the Boruta algorithm [179], which is an RF-based feature selection wrapper for finding all relevant variables in a data set. The Boruta algorithm assesses important features (in this case rSNPs) with respect to a variable outcome (in this case oil content) by constructing multiple decision trees based on random subsets of attributes or features. The pseudo-code for Boruta is given in Algorithm 1 (see Figure 5.6).

Using Algorithm 1, in this study, we analyze genotypes of rSNPs to identify their significant genotype  $\times$  phenotype association regarding the oil content of the cultivars. In order to deal with remaining obstacles resulting from the correlations between the SNPs or random fluctuations involved in the data set, we iteratively applied the Boruta algorithm (e.g., 1000 times), and considered an rSNP in our further analysis as important if and only if its importance was confirmed in all analyses. In the following, we refer to those rSNPs as *important rSNPs*.

## 5.5. Conclusions

Transcription factors orchestrate the entirety of cellular processes leading to tissue development, tissue differentiation or responses to the environment and, hence, act as natural master regulators within plants [146]. This makes them promising candidates as breeding targets

**Algorithm 1** : Boruta Algorithm**Input:**  $\mathcal{M}$ : Genotype (rSNPs) data**Input:**  $\mathcal{L}$ : Labels (cultivars)**Output:**  $\mathcal{C}$ : A ranked list of rSNPs based on their importance score**Method:**

- 1:  $t = 0$
- 2: **repeat**
- 3:    $\mathcal{M}_t = \mathcal{M}$
- 4:    $\widehat{\mathcal{M}}_t = \text{shuffle}(\mathcal{M}_t)$ : Creation of shadow attributes
- 5:    $\mathcal{M}_t^{ext} = [\mathcal{M}_t; \widehat{\mathcal{M}}_t; \mathcal{L}]$ : Matrix (data) concatenation to extend the input data
- 6:    $\mathcal{VLS}_t(\mathcal{M}_t^{ext}) = RF(\mathcal{M}_t^{ext})$ : Gathering variable importance scores ( $\mathcal{VLS}$ ) using RF classifier
- 7:    $\widehat{\mathcal{VLS}}_t = \max(\mathcal{VLS}(\widehat{\mathcal{M}}_t))$ : Max.  $\mathcal{VLS}$  value (in terms of z-Score) for shadow attributes
- 8:    $\mathcal{M}_t^c = \mathcal{M}_t^{ext}[\mathcal{VLS}(\mathcal{M}_t^{ext}) > \widehat{\mathcal{VLS}}_t] \setminus \widehat{\mathcal{M}}_t$ : rSNPs with significantly higher  $\mathcal{VLS}$  values  $> \widehat{\mathcal{VLS}}$
- 9:    $\mathcal{M}_t^r = \mathcal{M}_t^{ext}[\mathcal{VLS}(\mathcal{M}_t^{ext}) < \widehat{\mathcal{VLS}}_t] \setminus \widehat{\mathcal{M}}_t$ : rSNPs with significantly lower  $\mathcal{VLS}$  values  $< \widehat{\mathcal{VLS}}$
- 10:    $\mathcal{M} = \mathcal{M}_t \setminus [\mathcal{M}_t^c; \mathcal{M}_t^r]$ : Remove all rSNPs with determined importance from the input  $\mathcal{M}_t$
- 11:    $\mathcal{C}_t = \mathcal{VLS}(\mathcal{M}_t^c)$ : Gathering the rSNPs with confirmed  $\mathcal{VLS}$
- 12:    $t = t + 1$
- 13: **until** importance of all rSNPs is assigned
- 14:  $\mathcal{C} = \bigcup_{i=1}^t \mathcal{C}_i$

**Figure 5.6.: Pseudo-code for the Boruta algorithm.**

to control complex traits in crop breeding [146]. In this study, we performed a systematic analysis using multi-omics data (genomics, transcriptomics, and proteomics) to investigate the complex interplay between rSNPs, TFs and DEGs. As a result of this analysis, we obtained: (i) a genome-wide collection of rSNPs; (ii) their significant association with the *B. napus* cultivars differing in oil content; (iii) their consequences for TF binding; and (iv) the DEGs of four tissues whose expression could be strongly affected by the occurrence of these *important rSNPs* within their promoter regions.

Our findings show that while members of the TF-families DOF, MYB, NAC, GATA, or TCP have been identified as enriched exclusively for a certain tissue, the TFs in the bHLH or bZIP class, and members of the BES1 family seem to play important regulatory roles in several tissues. Moreover, the knowledge on the causal interaction between a rSNP, a TF and a DEG could be promising to explain the expression behavior of the gene, which in turn is essential for understanding the underlying genetic programs such as tissue development or responses to abiotic and biotic stresses.

By mainly considering the promoter regions, our integrated approach provides important insights into the regulatory processes on the transcriptional level. For future work, the investigation of further regulatory mechanisms underlying differential gene expression, as, e.g., post-transcriptional regulation such as microRNA binding or Riboswitch activity can

help to gain a comprehensive understanding of the entirety of gene regulatory processes. Nevertheless, our study can be seen as one further step leading towards the deciphering of differential gene expression underlying the different *B. napus* cultivars and our genome-wide collection of rSNPs provides a basis for upcoming studies on different traits in *B. napus*.

## 5.6. Supplementary Materials

The following supplementary material is available via the original publication <https://doi.org/10.3390/ijms22020789>. Table S1: Differentially expressed genes, Table S2: enriched GO terms (biological processes) of the DEGs, Table S3: lists of tissue specific enriched TFs, Table S4: rSNPs with TFBS predictions, Table S5: important rSNPs, Table S6: enriched GO terms (biological processes) and KEGG pathways of the DEGs harboring important rSNPs, Figures S1: Treemaps of the enriched GO terms (biological processes) of the DEGs.

## 6. Comparative Investigation of Gene Regulatory Processes Underlying Avian Influenza Viruses in Chicken and Duck

This chapter contains the article of the same name published in January 2022 in the MDPI journal *Biology* (<https://doi.org/10.3390/biology11020219>). For the sake of consistency within this thesis, the journal style is not adopted in this chapter.

This article is a joined work of Selina Klees<sup>1,2,\*</sup>, Johanna-Sophie Schlüter<sup>1</sup>, Jendrik Schellhorn<sup>3</sup>, Hendrik Bertram<sup>1,4</sup>, Antje Christine Kurzweg<sup>1</sup>, Faisal Ramzan<sup>1,2</sup>, Armin Otto Schmitt<sup>1,2</sup>, and Mehmet Gültas<sup>2,4,\*</sup>

<sup>1</sup>Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany

<sup>2</sup>Center for Integrated Breeding Research (CiBreed), Georg-August University, Carl-Sprengel-Weg 1, 37075 Göttingen, Germany

<sup>3</sup>Department of Bioinformatics, Institute for Microbiology and Genetics, Georg-August University, Goldschmidtstr. 1, 37075 Göttingen, Germany

<sup>4</sup>Faculty of Agriculture, South Westphalia University of Applied Sciences, Lübecker Ring 2, 59494 Soest, Germany

\* Authors to whom correspondence should be addressed.

### Author contributions by Selina Klees (S.K.):

S.K. participated in the design of the study, developed the pipeline, prepared the data sets, conducted the bioinformatics and computational analyses, interpreted the results, and wrote the final version of the manuscript.

### 6.1. Simple Summary

Avian influenza poses a great risk to gallinaceous poultry, while mallard ducks can withstand most virus strains. To date, the mechanisms underlying the susceptibility of chicken and the effective immune response of duck have not been completely understood. In this study, our aim is to investigate the transcriptional gene regulation governing the expression

of important avian-influenza-induced genes and to reveal the master regulators stimulating an effective immune response after virus infection in ducks while dysfunctioning in chicken.

## 6.2. Abstract

The avian influenza virus (AIV) mainly affects birds and not only causes animals' deaths, but also poses a great risk of zoonotically infecting humans. While ducks and wild waterfowl are seen as a natural reservoir for AIVs and can withstand most virus strains, chicken mostly succumb to infection with high pathogenic avian influenza (HPAI). To date, the mechanisms underlying the susceptibility of chicken and the effective immune response of duck have not been completely unraveled. In this study, we investigate the transcriptional gene regulation underlying disease progression in chicken and duck after AIV infection. For this purpose, we use a publicly available RNA-sequencing data set from chicken and ducks infected with low pathogenic avian influenza (LPAI) H5N2 and HPAI H5N1 (lung and ileum tissues, 1 and 3 days post-infection). Unlike previous studies, we performed a promoter analysis based on orthologous genes to detect important transcription factors (TFs) and their cooperation, based on which we apply a systems biology approach to identify common and species-specific master regulators. We found master regulators such as EGR1, FOS, and SP1, specifically for chicken and ETS1 and SMAD3/4, specifically for duck, which could be responsible for the duck's effective and the chicken's ineffective immune response.

### Keywords

avian influenza; chicken; duck; mallard; gene regulation; differentially expressed genes; RNA sequencing; transcription factor cooperation; master regulators; upstream regulators

## 6.3. Introduction

Avian influenza is a viral infection mainly affecting birds such as wild waterfowl or galinaceous poultry but not stopping at humans or other mammals, and thus posing a high risk for a future pandemic [59]. Its causative pathogen is a type A influenza virus from the *Orthomyxoviridae* family of segmented negative-sense RNA viruses [60]. Based on their pathogenicity in chicken, avian influenza viruses (AIVs) can be classified into high- and low pathogenic avian influenza viruses (HPAIVs and LPAIVs, respectively) [63]. While chicken can usually withstand an LPAI infection, they succumb to infection with HPAI within a few days. Mallard ducks, on the other hand, are known to successfully fight all LPAI and most HPAI infections, with usually only mild symptoms, and are hence considered a natural reservoir of the virus [59]. After the first report of human infections with HPAI H5N1 in 1997, attention was drawn to the predominantly poultry-affecting avian influenza spreading



across the globe [60]. Since 2003, 862 cases of humans infected with H5N1, along with 455 cases of death, were reported to the World Health Organization (WHO) [61].

With the ongoing intensive breeding for different production traits in chicken, such as growth and feed efficiency, other, unanticipated traits such as skeletal defects, metabolic disorders, or immune responses could have been compromised [180]. Therefore, the breeding goals have shifted towards maintaining animal health, leading to both animal welfare and the prevention of economic losses [180].

However, to date, the mechanisms underlying chicken's susceptibility to avian influenza and the effective immune response of duck have not been completely deciphered. The susceptibility of chicken can be partially explained by their lack of virus pattern recognition receptor RIG-I gene and the gene for the RIG-I binding protein, RNF135, both of which exist in ducks [59, 65]. The RIG-I receptor recognizes double-stranded RNA and initiates self-promoting pathways leading to the early type I interferon (IFN) response, which is important for innate immune response. In chicken, other pattern recognition receptors, such as MDA5 and TLR7, are upregulated in response to viral entry, which also leads to the induction of IFN expression [62, 63, 65, 66]. However, the immediate induction of type I IFNs seems to be much more robust and effective in ducks than in chicken or other avian species. In addition to the difference in pattern recognition receptors, there appears to be a variety of factors and differences that lead to the successful or unsuccessful immune response of ducks or chickens, respectively. Different studies evaluated the transcriptomics response to different AIVs in chicken [63, 65, 181–194], duck [195–198], or both [59, 199–202]. For example, Smith et al. [59] investigated the role of the expression levels of different interferon-induced transmembrane proteins (IFITMs) in the duck's ability to alleviate the virus while it prevailed in chicken. Evseev and Magor [62] provide a comprehensive review of the differences in innate immune response in chickens and ducks. However, the host-pathogen interactions and their underlying mechanisms in ducks and chicken are multifactorial and highly complex, and must be elucidated to obtain a deeper insight into the duck's effective immune response against AIV while it proves lethal to chicken [62].

Despite the rich literature on the differences in chicken and duck immune response after AIV infection, the role of transcription factors (TFs) and their cooperations, which underlies transcriptional gene regulation, has not yet been extensively studied. The knowledge about the complex interplay of TF pairs could provide promising information to unravel the differences in disease progression in these species, since the TFs specifically bind to the promoter regions of genes and thereby orchestrate differential gene expression in a highly context-specific manner [68, 203]. In response to different environmental conditions such as viral infection, they can activate processes or react to specific pathways, and thus fine-tune the gene expression pattern in an organism. By interacting with other TFs in either a cooperative or competitive manner, they form the basis for complex pathway and network structures in biological systems [110, 204, 205].

To address the limited knowledge about upstream regulators, including TFs, their complex interplay, and master regulators, which are responsible for an effective immune response

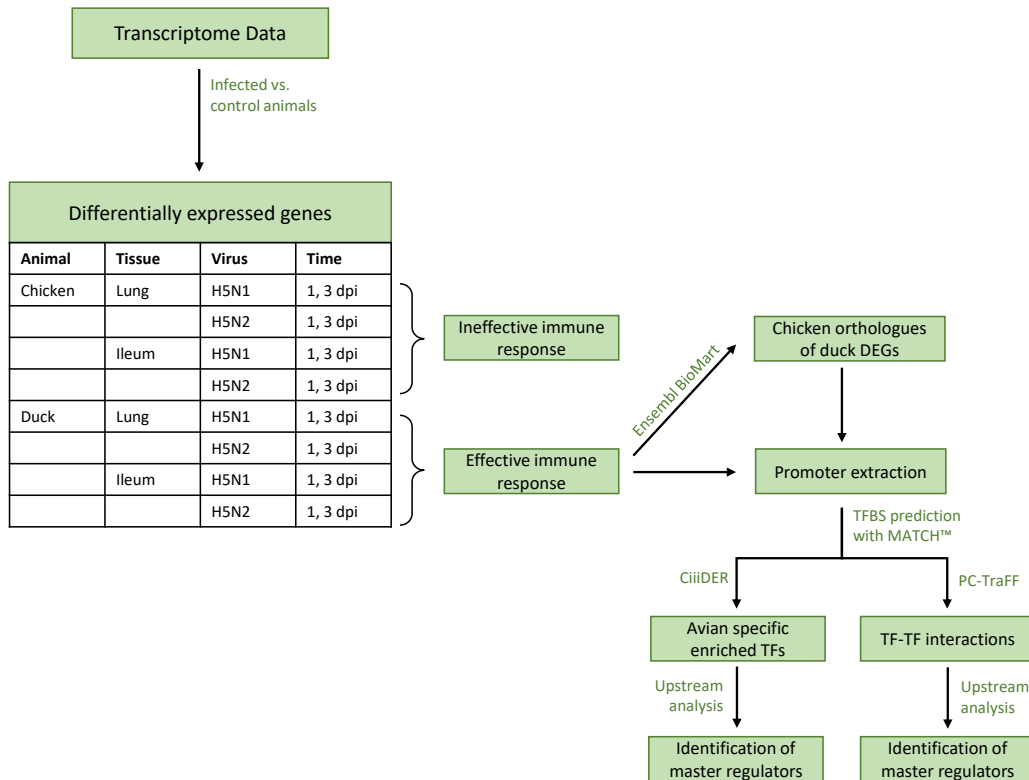
after avian influenza infection, we performed a systematic analysis using an RNA-seq data set. More specifically, mainly considering the effective immune response of duck, we identified the corresponding differentially expressed genes (DEGs) in response to the virus and analyzed their promoter regions to determine the upstream regulators. Then, to investigate the regulatory mechanisms of these DEGs in chicken, we analyzed their chicken orthologs to assess the species-specific regulators. Focusing on the master regulators arising from enriched TFs and TF-TF cooperations, our results can help to resolve the question of why the relevant genes could be differentially expressed in duck, while transcriptional gene regulation in chicken remains unsuccessful. Consequently, in our results, we present two groups of master regulators for ileum and lung: while the first group of master regulators contains common regulators found for both species, the species-specific master regulators were assigned to the second group. In particular, we strive to decipher the duck-specific master regulators related to the immune responses that are absent in chicken. Our findings could be essential in the search for possible mechanisms that stimulate an effective immune response in ducks while dysfunctioning in chicken.

## 6.4. Materials and Methods

In this section, we describe the methods, starting at the transcriptome level where differentially expressed genes are identified. Since the avian-influenza-induced differential expression of genes in duck have been abundantly compared to chicken, and duck is generally known to effectively prevent severe disease progression, we have a particular interest in investigating the promoter regions of DEGs in duck that potentially allow duck to adapt to the H5N1 virus and enable a proper immune response, which is apparently not the case for the orthologous genes in chicken. Thus, we want to identify the diversity in gene expression by applying promoter analyses to duck and chicken and identify the transcription factors that may provide an explanation for their varying immune responses. An overview of the steps encompassed in our analysis is given in Figure 6.1.

### 6.4.1. Transcriptome Data

The RNA-sequencing analysis of lung and ileum tissue samples from chickens and ducks infected with high- (H5N1) and low- (H5N2) pathogenic avian influenza viruses measured 1 and 3 days post-infection (dpi) was conducted by Smith et al. [59]. In their study, a total of 20 white leghorn chickens and 20 Domestic Gray Mallards were challenged with either the HPAI or the LPAI virus. Processed RNA-sequencing data, e.g., count tables for the mapped reads and experimental design, were retrieved from Array Express under the publicly available accessions *E-MTAB-2908* and *E-MTAB-2909* for chicken and duck, respectively. For each experimental condition (e.g., chicken, lung, H5N1 infection, 1 dpi), gene expression was measured for three biological replicates, resulting in a total of 24 samples from infected



**Figure 6.1.: Flow chart of the employed analyses.** Differentially expressed genes (DEGs) were derived by comparing the gene expression rate of a specific condition against a mock infection for that condition (e.g., chicken lung at 1 dpi with H5N1 infection against chicken lung at 1 dpi with mock infection). TF and TFBS stand for transcription factor and transcription factor binding site, respectively. H5N1 is a high pathogenic avian influenza virus (HPAIV), while H5N2 is a low pathogenic avian influenza virus (LPAIV).

animals and 12 mock-infected control samples for each species. In chicken and duck, the expression of 24,356 and 25,952 genes was measured, respectively. For further details on the experimental design, as well as the processing steps of the RNA-sequencing data, we refer to the study by Smith et al. [59].

The identification of DEGs was performed in R by using the state-of-the-art package DESeq2 (version 1.30.0) [170] with default parameters for the median-of-ratios normalization and the `ashr` R package (version 2.2-47) for  $\log_2$  fold change (LFC) shrinkage [206]. DEGs were determined for each condition (e.g., lung infected with H5N1 at 1 dpi) against a control group (e.g., lung with mock infection at 1 dpi). Similar to the study of Smith et al. [59], genes were considered to be significantly differentially expressed if the criteria  $|\text{LFC}| > 0.58$  and the FDR-adjusted  $p$  value  $< 0.05$  were met.

### 6.4.2. Identification of Enriched TFs and TF-TF Cooperations

To unravel the differences in transcriptional gene regulation underlying the identified DEGs, we focused on their regulatory regions (promoter regions) and identified enriched TFs as well as TF-TF cooperations using the two bioinformatics tools CiiDER [171] and PC-TraFF [69, 110], respectively. A detailed description of the theory behind both methods can be found in the original studies [69, 110, 171]. Besides some algorithm-specific parameters, both algorithms require as input the promoter sequences and a library of position weight matrices (PWMs) representing the TFBSs.

- **Promoter sequences:** Using the current versions of reference genomes GRCg6a and CAU\_duck1.0, we extracted the promoter sequences ranging from  $-1000$  base pairs (bp) to  $+100$  bp relative to the transcription start site (TSS), similar to previous studies [91, 99, 110, 207]. Sequences were rejected if the full promoter sequence could not be obtained, which was mostly the case for genes on scaffolds.
- **Creation of the PWM profile and TFBS detection:** Following our previous studies [68, 203], we created a custom avian-specific PWM profile. For this, we first downloaded the TFs of avian species (chicken, duck, turkey, zebra finch, and flycatcher) from animalTFDB 3.0 [208] and selected those that were expressed in at least one RNA-seq experimental condition. Second, we mapped the TFs to the PWMs stored in the TRANSFAC database (release 2018.1) [93]. Finally, we clustered the PWMs hierarchically based on their pairwise Pearson's correlation coefficients and selected the representative with the highest information content for each cluster in order to create a non-redundant PWM profile with thresholds minimizing the sum of the false-positive and false-negative rates ("minSUM profile"). In total, the profile contains 553 PWMs, which are provided in File S1. We predicted the transcription factor binding sites by applying the MATCH<sup>TM</sup> tool [18], which obtains the custom avian-specific PWM profile and a matrix library provided by TRANSFAC [93] as input.
- **TF enrichment:** We performed a TFBS enrichment analysis by employing the CiiDER tool [171] in order to identify over- and underrepresented TFBSs. In the following, we refer to a TF as over-/underrepresented in a condition if its corresponding TFBS is significantly over-/underrepresented in the set of promoter sequences of the respective DEGs compared to a custom background. The background set is composed of the promoter sequences of those genes that were not differentially expressed in any of the conditions. From this, the custom background was created as a subset of sequences of the same global GC distribution as the foreground sequences using BiasAway [209]. In a last step, a random sample of equal size was taken as the foreground gene set from the custom background for each gene set, which eventually led to individual background sets from the same distribution, thus making them comparable. Assessment of the distributions of TFBS predictions in foreground and

background promoter sets is carried out by an FDR-adjusted p value threshold of 0.05.

- **TF-TF Cooperation:** The PC-TraFF algorithm [110] and its extension PC-TraFF+ [69] are well-established, information-theory-based approaches to identify TF-TF cooperation pairs using the concept of pointwise mutual information. While PC-TraFF detects the co-occurring TFBSs of TF-pairs in the promoter sequences, PC-TraFF+ separates the highly sequence-set-specific TF-cooperations from the common ones by removing the background co-occurrences of TFBSs. The algorithm needs the predefined distance thresholds as input for the TFBSs. As in our previous studies [68, 203], we used the recommended distances of  $\geq 5$  and  $\leq 20$  and defined a TF-pair as significant if its z-score  $\geq 2$ .

#### 6.4.3. Identification of Master Regulators

Similar to previous studies [129, 207, 210–213], we detected upstream regulators that regulate a set of DEGs through concerted coordination of TFs and intermediary modulators. More precisely, these so-called master regulators (key nodes) are found on top of the regulatory hierarchy of complex regulatory networks, leading to the finely tuned gene expression of a gene set. In order to identify master regulators targeting the TFs and their partners, we applied the so-called “upstream analysis” provided by the geneXplain platform, which is based on a modified shortest-path algorithm [207, 213, 214]. Consequently, focusing mainly on H5N1, we established the top five master regulators for the lung and ileum tissues of chicken and duck using the GeneWays database [215].

#### 6.4.4. Annotations and Ortholog Mapping

The orthologs were retrieved from the BioMart web services [216] via the R package `biomaRt` [217]. It is important to note that the mapping of, e.g., duck DEGs to chicken orthologs is not necessarily bijective, since a duck gene could be missing in chicken (e.g., *RIG-I*), and thus have no chicken ortholog, or a duck gene could have two orthologs in chicken.

### 6.5. Results and Discussion

In this study, by analyzing a transcriptome data set, we firstly identified differentially expressed genes (DEGs) for lung and ileum tissues in chicken and duck after infection with H5N1 and H5N2 at 1 and 3 dpi. In line with the results of Smith et al. [59], our analysis of RNA-seq data with DESeq2 revealed three different observations: (i) we detected a considerably higher number of DEGs in the duck than in chicken under most conditions (see Tables 6.1 and S1); (ii) the vast majority of DEGs were highly context-specific with regards to the virus strain and timepoint. Only 20 and 1 were found to be common in all conditions in the duck ileum and lung, respectively, while no DEG was observed for all

conditions in chicken (see Figure 6.2); (iii) the response in terms of differential expression was higher after infection with the HPAI H5N1 compared to infection with the LPAI H5N2, especially in duck (see Table 6.1). The gene set enrichment analysis of the DEG sets based on Gene Ontology (GO) classification demonstrates that differential gene regulation after virus infection deviates between chicken and duck (Figure 6.2). The full lists and treemaps for GO enrichment are given in Table S2 and Figures S1 and S2.

**Table 6.1.: Numbers of differentially expressed genes (DEGs) in duck and chicken for the treatments with H5N1 (HPAI) and H5N2 (LPAI) virus after 1 and 3 days post-infection (dpi).** The table is split into upregulated ( $LFC > 0.58$ ) and downregulated genes ( $LFC < -0.58$ ).

Virus	Time	Tissue	Duck DEGs		Chicken DEGs	
			Upregulated	Downregulated	Upregulated	Downregulated
H5N1	1 dpi	lung	804	350	1	7
		ileum	193	63	5	6
	3 dpi	lung	605	486	1	0
		ileum	332	346	3	1
H5N2	1 dpi	lung	47	0	0	0
		ileum	42	1	20	2
	3 dpi	lung	1	0	0	0
		ileum	25	0	286	20

To summarize, in agreement with previous studies [59, 199–202], the DEG analysis indicates that the general pattern of differential gene expression differs greatly between duck and chicken after AIV infection. In particular, the infection with H5N1 elicits a rapid and effective immune response in ducks, whereas the chicken immune system did not appear to respond to the same extent.

Despite the great interest in and rich research on avian influenza, there is still a lack of knowledge about the underlying transcription factors and their combinatorial interplay orchestrating gene expression and leading to an effective immune response in ducks while failing in chicken. In order to reveal transcriptional gene regulation factors that play important roles in disease progression, we compared the upstream regulatory regions (i.e., promoters) of the duck DEGs with those of the respective chicken orthologs. Since the response regarding differential expression appears to be most pronounced after infection with the H5N1 virus—and, as an HPAIV, poses the greatest risk for avian as well as mammal species—we concentrate on this virus in the following.

Typically, in bioinformatics, the choice of the threshold value for, e.g., FDR-adjusted  $p$ -values, is of great importance for the number of significant results. In this study, we mainly followed the values used in the study of Smith et al. [59] to ensure some comparability. Nevertheless, it is important to note that a  $p$ -value may be interpreted differently in different

species, e.g., due to a lower variability of transcriptomics data in genetically stable inbred lines, such as the chickens used in this study, compared to ducks. A more stringent  $p$ -value threshold of 0.01 for the DEG identification or TFBS enrichment analysis leads to a strong reduction in their results, which, in turn, results in an insufficient number of genes or TFs for further analysis (for a  $p$ -value comparison, see Tables S1 and S3). For this reason, we used a threshold of 0.05 in the following analysis.

Several studies have investigated the importance of glycosylation with respect to viral entry and replication [218–220]. Glycosylation is a post-translational process of host cells that can be used by AIVs to attach glycan moieties to their own proteins [218]. In our DEG sets, we observed one enriched GO term related to glycosylation (GO:MF glycosaminoglycan binding). Remarkably, this GO term was enriched among both up- (duck, lung, H5N1, 1 dpi) and downregulated (duck, lung, H5N1, 3 dpi and duck, ileum, H5N1, 3 dpi) DEG sets, but its interpretation is beyond the scope of this study.

### 6.5.1. Transcription Factor Binding Site Enrichment

In a first step, we identified significantly over- or underrepresented TFBSs in the promoter regions in the gene sets. The Venn diagrams of over- and underrepresented TFBSs in duck and chicken show a similar pattern for both tissues and timepoints: a high number of enriched TFBSs are unique to either chicken or duck, resulting in only a slight overlap between chicken and duck in terms of over- or underrepresented TFBSs (Figure 6.3). Interestingly, when comparing overrepresented TFBSs in duck and underrepresented TFBSs in chicken or vice versa, there appears to be more overlap. Generally, the number of predicted TFBSs that are significantly over- or underrepresented in the ileum is smaller in both chicken and duck than in lung, which reflects the corresponding numbers of DEGs. To offer a closer insight into the related TFs of the enriched TFBSs found for the H5N1 infection, we explain their functions in more detail. As the HPAIV is known to predominantly replicate in the respiratory tract [59], we will further concentrate on the lung tissue with functional interpretation. The lists of significantly over- or underrepresented TFBSs are provided in Table S3.

Based on the enriched TFBSs in chicken at 1 dpi, we observed 21 TFs that were uniquely overrepresented in chicken and 33 TFs that were overrepresented in the chicken promoters while underrepresented in the duck promoters (Figure 6.3). We observed many TFs of the basic helix–loop–helix (bHLH) class and the C2H2 zinc finger class, including different TF families, such as zinc finger proteins (ZNFs), Zinc finger and BTB domain-containing proteins (ZBTB), or specificity proteins (SPs). Furthermore, TF families such as SMAD, AP2, TFII-I, GCM, and paired box factors (PAX) can be found [221]. Similar TF families are salient after 3 dpi in chicken, with a greater focus on zinc finger factors, as they make up 13 out of 22 chicken TFs. For both timepoints, we observed several tryptophan cluster factors, including a TF from the interferon regulatory factor family (IRF4) and ETS/ETS-related TFs.

Interferon regulatory factors (IRFs) play a major role in the immune response by inducing several processes and pathways upon avian influenza infection. For example, the over-expression of IRF7 in chicken DF-1 cells resulted in a higher viral replication and cell death rate than in control cells upon infection with LPAI H6N2 [181]. Transcriptome analysis revealed that chicken IRF7 could be involved in the modulation of programmed cell death via pathways such as the TGF- $\beta$ , FOXO, and the JAK-STAT pathway [181].

Interestingly, binding sites of the SMAD family members SMAD4 and -5 were enriched in chicken at both timepoints, but not in duck promoters. The SMAD factor family is tightly linked to the TGF- $\beta$  pathway, which is involved in various immune-related processes such as apoptosis, the innate immune response by type I interferon production, or early pulmonary fibrosis via epithelial–mesenchymal transition in response to influenza A virus (IAV) infection [181, 222–225]. As a response to IAV invasion, the RIG-I-like receptor (RLR) signaling, followed by IRF3 activation, represses TGF- $\beta$ -induced SMAD signaling in mammal cells [224]. Hence, the availability of SMAD binding sites could be an important regulator of TGF- $\beta$  and RLR signaling in chicken.

The ETS/ETS-related TF family is uniquely enriched in the chicken promoters. Apart from various cellular processes ranging from embryonic development to apoptosis and carcinogenesis, ETS factors play a role in both the innate and adaptive immune response [226]. Interestingly, it has recently been shown that the ETS-family member ETV7 targets several interferon-stimulated genes (ISGs) to negatively regulate the effective IFN-mediated control of influenza viruses, and can thus be considered as a suppressor of the type I IFN response in mammalian cells [227]. Hence, an over-representation of different ETS binding sites in chicken promoters could possibly influence the intensity of the antiviral type I IFN response, which should be investigated in future studies.

In the duck lung at 1 dpi, 9 TFs are uniquely overrepresented (Figure 6.3). Among them, we found representatives of the TF families forkhead box (FOX) (FOXC1, FOXL2, FOXO3, and HNF3B), POU (POU3F2 and TST1), STAT, homeobox (HOX), and one IRF TF (IRF4). Another 11 TFs, which were also overrepresented in the duck promoter sets, were simultaneously underrepresented in chicken. Here, we predominantly found homeo domain factors such as HOXD13, NKX22, NKX61-62, DLX3, LHX3, PRX2, and SIX3. The pattern of significantly enriched TFs in duck 3 dpi is similar to that of 1 dpi. One TF family that is more prominent 3 dpi is the HOX family and we further observed the C2H2 zinc finger factor SALL3 while the IRF4 disappeared at 3 dpi.

The FOX family of TFs is suggested to be involved in the regulation of a variety of processes, such as cell growth, proliferation, differentiation, longevity, immunology, and cell-cycle control [228]. FOX TFs play an important role in the FOXO signaling pathway, which regulates important processes such as stress resistance, cellular proliferation, and apoptosis [181, 229]. The subclass FOXO is known to be involved in the regulation of lifespan and diseases by orchestrating processes such as cell-cycle progression and apoptosis under severe stress conditions in mammals, and FOXO was shown to be a negative regulator of IRF7, a member of the interferon regulatory factor family [181, 229].



Interestingly, the binding sites for two members of the signal transducer and activator of transcription (STAT) family, a main factor of the JAK-STAT signaling pathway, are enriched at both timepoints in the duck lung, but not in chicken. This highlights the importance of the JAK-STAT pathway, which is one of the key pathways in type I IFN response and induces interferon-stimulated genes (ISGs) [181, 230, 231]. In particular, virus entry followed by IFN expression leads to an IFN receptor-associated Janus-kinase (JAK) phosphorylation, which activates STAT TFs to enhance target IGS gene expression [232, 233]. Hence, a lack of enriched binding sites for STAT factors in the chicken promoter sequences could possibly result in a weaker upregulation of ISGs and less efficacy in the JAK-STAT pathway.

Another TF family whose binding sites are overrepresented only in the duck promoters is the POU family. Interestingly, there is evidence that members of the POU family, expressed in B and T cells, may interact with STAT3 and can activate different interleukin promoters, which are related to immune and inflammatory responses in human cells [234].

Additionally, the genes of some promising enriched TFs, e.g., IRF7 (*ENSAPLG00000012752*), STAT1 (*ENSAPLG00000013226*), and STAT4 (*ENSAPLG00000023296*) are significantly upregulated upon AIV infection in the duck lung 3 dpi, which may underline their importance in response to the virus.

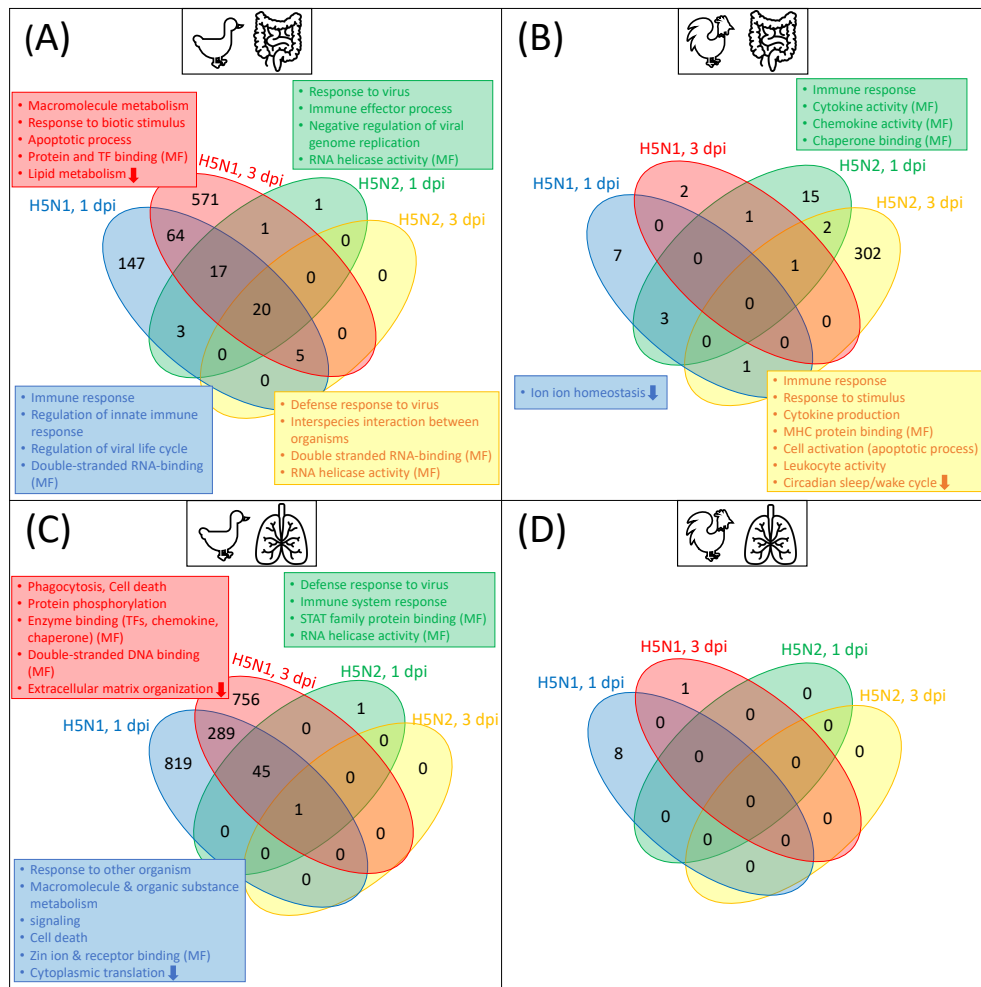
### 6.5.2. TF-TF Cooperations

To obtain a closer insight into the disease regulation progress in chicken and duck, knowledge of the complex interplay between TFs could provide further essential information, since they are important for the regulation of the transcriptional machinery and form the backbone for the fine-tuned adaptation of a species to specific environmental conditions [69, 110]. By further focusing on the HPAIV, we applied the PC-TraFF algorithm [69] and identified the cooperation of TFs based on their binding site co-occurrence patterns in the promoter regions of the investigated genes in the two species. Based on the PC-TraFF results, we constructed a TF cooperation network, in which the nodes represent the TFs and the edges indicate their cooperation. The complete networks for lung and ileum are provided in Table S4 and Supplementary File S2. However, in order to establish the preferential partner choice of TFs for the regulation of disease progression in both animals, we mainly consider the differences between the networks that were constructed for the chicken and duck tissues. Figure 6.4 shows the TFs and their partners in the regulatory events of these tissues, which are either found only in chicken or only in duck. In the following, we refer to a chicken\duck network as the network of chicken TF cooperations without the duck TF cooperations and vice versa.

The greatest difference between chicken and duck can be observed in the chicken\duck network for ileum 3 dpi, which contains 25 nodes and 14 edges (Figure 6.4B). Among the single nodes in this network, we found the ETS-related TF NERF and the bHLH heterodimeric TF AHR:ARNT. Interestingly, the lack of a partner indicates that the respective partner is present in the duck network, interacting with another TF. Such preferential part-

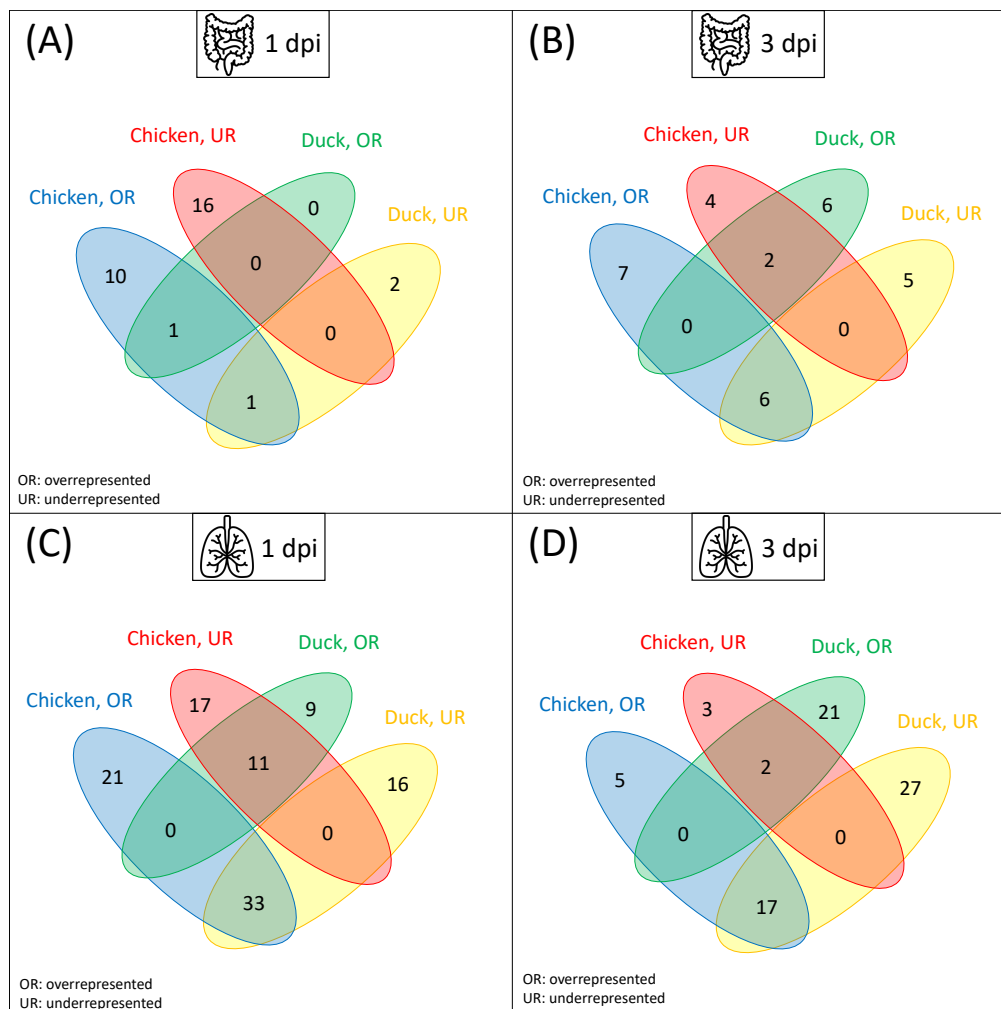
ner choices are an indication of species-specific dimerization events, which form the basis for the regulation of different processes, such as immunity and inflammation [235]. Further ETS-related factors are found in the lung 1 and 3 dpi in the chicken\duck networks (Figure 6.4C,D), and the monomer AHR is additionally found in the chicken\duck network for ileum 1 dpi (Figure 6.4A). The importance of ETS-related and bHLH factors for chicken promoters was shown in the TF-enrichment (Section 6.5.1). Prominently, among all duck\chicken networks, we observed different FOX and DLX homeo domain factors, which are not present in any chicken network. Both TF families were found to be enriched in the duck, but not the chicken promoters (see Section 6.5.1).

Remarkably, the differences in the cooperation networks are rather moderate in contrast to the divergent results of TF-enrichment between chicken and duck (Figures 6.3 and 6.4). This indicates that, while single, enriched TFs in the promoter regions are rather species-specific, the TF-TF cooperation networks of both species share many common features and TF clusters seem to be preserved or classified by specific partner alterations.



**Figure 6.2.: Venn diagrams of the DEGs (A) duck in ileum, (B) chicken in ileum, (C) duck in lung, and (D) chicken in lung with selected enriched Gene Ontology (GO) terms.** The DEGs are obtained by comparing animals infected with AIV (H5N1 (HPAI) or H5N2 (LPAI)) with mock-infected animals. The colors within the venn diagram, as well as the colors of the GO-term boxes, stand for the respective condition: blue represents H5N1 infection 1 dpi, red represents H5N1 infection 3 dpi, green represents H5N2 infection 1 dpi, and yellow represents H5N2 infection 3 dpi for each species and tissue. Within the boxes, an arrow down indicates that the GO-term is enriched among the downregulated DEGs; otherwise, the terms are enriched among the upregulated DEGs. The GO-terms represent biological processes except, if stated differently, in the form of MF (molecular function). Venn diagrams are based on the data provided in Table S1.

## TFBS enrichment for H5N1



**Figure 6.3.: Venn diagrams of TFBS enrichment to compare over- (OR) and underrepresented (UR) binding sites in chicken and duck.** The promoter regions of DEGs (after infection with HPAIV H5N1) in duck and the corresponding orthologous genes in chicken were extracted to obtain the over- and underrepresented TFBSs. (A) shows the corresponding number of TFBSs for the ileum 1 dpi, (B) shows the ileum 3 dpi, (C) shows the lung 1 dpi and (D) shows the lung 3 dpi. Venn diagrams are based on the data provided in Table S3.



### 6.5.3. Master Regulators

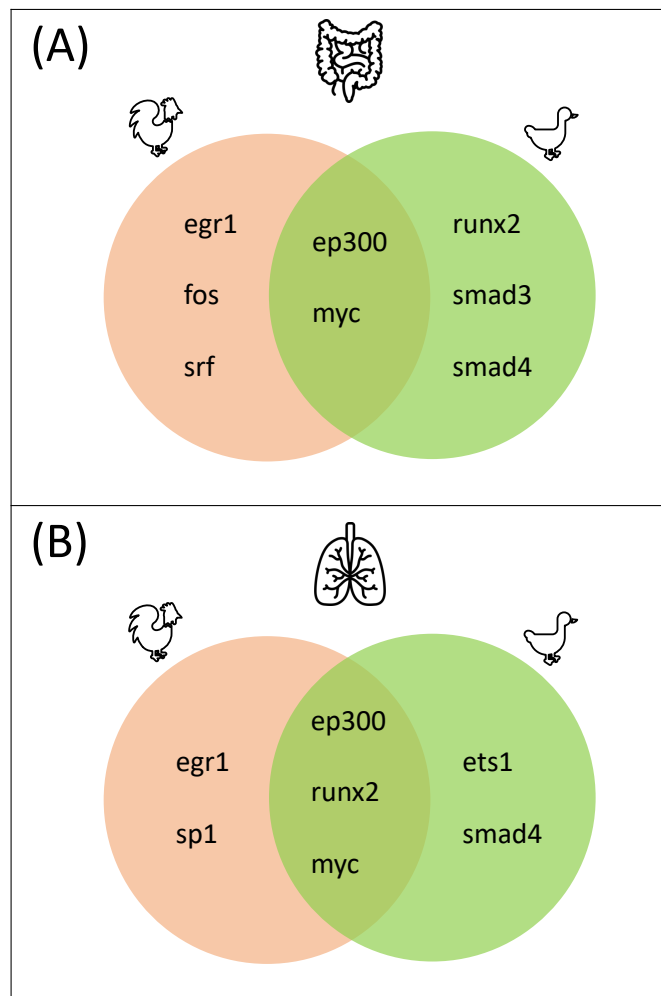
Functionally related genes involved in the same physiological or molecular processes, such as virus defense, are often coordinately regulated by the precise organization of TF binding [237]. This precise organization of TFs and their cooperation includes various upstream pathways forming complex regulatory network structures, in which different pathways can be connected in series, in parallel, or reverse, thus forming different feedforward or feedback loops [207]. One way to identify important regulators within such a complex regulatory network is the so-called “upstream analysis” [213], which aims to identify master regulators that are positioned at the top of the regulatory hierarchy and can be seen as common upstream regulators of a gene set, regulating the genes’ expression rates.

During disease progression, the specific partner choices of TFs are of the utmost importance for an effective and rapid immune response [68, 203, 207]. Therefore, we mainly focus on the master regulators orchestrating the TF-TF cooperations in the following. The complete upstream regulatory networks based on the TF-TF cooperations can be obtained from Figure S4. Further, Figure 6.5 shows common and species-specific master regulators directing gene regulation in the lung and ileum tissues after infection with H5N1 (for both timepoints).

A closer look at the identified master regulators reveals that *EGR1*, *SRF*, *FOS*, and *SP1* are unique to chicken in both tissues. *EGR1* is considered a master transcription factor, regulating the expression of a range of genes involved in multiple cardiovascular diseases, such as atherosclerosis or ischemia in humans [238–240]. Furthermore, it is known to play various regulatory roles in processes such as cell death and survival or inflammatory processes [241]. In response to avian influenza in human, epithelial lung cells *EGR1*, as well as the chicken-specific master regulator gene *FOS*, were strongly downregulated [242].

The serum response factor (*SRF*) and the proto-oncogene factor *FOS* both play an important role in the inflammatory response after influenza infection in mammals [243]. The transcriptional regulator *SRF* first activates *FOS* expression [244, 245], which encodes, together with *JUN*, the components of the transcription factor complex AP-1 [246, 247]. AP-1 regulates a variety of processes such as cell proliferation and differentiation [246, 247] but also activates the transcription of pro-inflammatory genes after an influenza infection [243]. In chicken trachea, *FOS* was shown to be upregulated after hydrogen-sulfide-induced oxidative stress, revealing the importance of *FOS/IL8* signaling during tracheal inflammation [248]. Kim et al. (2020) [249] showed that a knockout of *IRF7* in chicken DF-1 cells, and subsequent AIV infection, resulted in the altered gene expression pattern of several genes, including key immune response genes such as *IL12*, *FOS*, and *API*. The authors further suggest that this shift in expression pattern could be a compensation for the absence of *IRF7* [249].

*SP1* is involved in influenza A virus-induced mucin (i.e., *MUC5AC*) expression in mouse epithelial cells. Mucins, the gel-forming glycoproteins of mucus, are important to moisturize and protect surfaces from pathogens, and a mis- or overexpression of mucin may be related to various diseases, including different lung diseases caused by inflammation [250].



**Figure 6.5.: Common and species-specific master regulators for the (A) ileum and (B) lung tissue regulating the TF-TF cooperations.**

Furthermore, SP1 cooperates with different SMAD TFs in response to TGF- $\beta$ , leading to the growth arrest of epithelial cells [251]. Interestingly, three SP family members (SP1-3) were found to be enriched in the chicken but not the duck promoters of the genes under study (Section 6.5.1).

In addition, the master regulators MYC and EP300 were identified as common to chicken and duck in both tissues. As an oncogenic TF, MYC is involved in several cellular processes related to cell growth, cell proliferation, or apoptosis [252]. Moreover, it is an important player in the JAK/STAT pathway, an important pathway in type I IFN response, as it is

directly regulated by STAT TFs [181, 233]. EP300 encodes a histone acetyltransferase that regulates the transcription of genes involved in cell proliferation and differentiation processes via chromatin remodeling. It is known to interact with a significant number of TFs, such as STAT, ETS1, and Ep53 in humans [253, 254]. The importance of STAT in duck promoters has been shown in Section 6.5.1 and ETS1 acts as a master regulator in the lung tissue in duck. Furthermore, Leymarie et al. (2017) [255] observed that H5N1-infected mice developed a clear signature, leading to lung edema, which represents a pathogenic fluid accumulation in the lungs leading to respiratory dysfunction. Interestingly, they discovered an edema signature regulatory network consisting of different master TFs including EP300 and Runx1, a runt-related transcription factor [255]. Another Runx family member, Runx2, was identified as a common master regulator in the lung and as a duck-specific master regulator in the ileum. This finding enhances the importance of pathological edema-related processes during virus defense.

Master regulators that are unique to duck are of particular interest in our analysis, since they seem to activate pathways, leading to an effective differential expression of important genes, which is not the case for chicken. We identified three different duck-specific master regulators: ETS1 in the lung, SMAD3 in the ileum, and SMAD4 in both tissues.

As ETS factors play a role in both the innate and the adaptive immune response [226], they could be important master regulators controlling gene expression in duck HPAI defense. Among the ETS TFs, ETS1 and PU.1 seem to play the most important role in immunity in humans due to their control of immune cell development [226]. Surprisingly, different binding sites for ETS family members (except ETS1) have been identified as enriched in chicken but not duck promoters (see Section 6.5.1).

The importance of SMAD TFs in immune response and their tight link to the TGF- $\beta$  and RLR-signaling pathways were revealed in Section 6.5.1. In particular, the SMAD3 family member is activated by TGF- $\beta$  receptors and forms a transcriptional complex with SMAD4. The SMAD3/4 complex can then physically and functionally interact with c-Jun-c-Fos by binding to AP-1 binding sites to activate TGF- $\beta$  responsible genes [224, 256, 257]. Hence, working in cooperation, the SMAD family members SMAD3 and 4, play a major role in TGF- $\beta$ -mediated immune response and can be considered as promising targets for future studies.

In the second part of this section, we were additionally interested in the investigation of the master regulators targeting the enriched TFs of the DEG sets. As expected, the vast majority of the identified master regulators are unique to either chicken or duck in both tissues. The reason for this can be explained based on the distinct sets of enriched TFs presented in Section 6.5.1. Notably, the master regulators ARNT2 and EPAS1 were found only for chicken, while CRSP2, IRF9, and IRF7 were found only for duck. The complete upstream regulatory networks are provided in Figure S5. However, this finding does not reflect the assumption that the regulatory mechanisms of two orthologous gene sets share common features. Therefore, we presume that TF enrichment does not sufficiently represent the regulatory interplay underlying disease progression.



## 6.6. Conclusions

Until now, the mechanisms underlying the susceptibility of chicken and the effective immune response of duck are not completely understood. In this study, we performed a systematic analysis to investigate the transcriptional gene regulation underlying disease progression in ducks and chicken after infection with avian influenza. For this purpose, we identified upstream regulators, including TFs, their complex interplay, and master regulators, which are responsible for different immune responses in both species.

Our results suggest that there are major differences between the promoter regions of orthologous genes regarding the enrichment of TFs in both species. In particular, we identified promising TF families, which are important regulators of chicken (TF families such as SMAD, IRF, and ETS) or duck (TF families such as FOX, STAT, and POU). Although TF enrichment provides important insights, we could unravel the specific partner choice of TFs, which could be responsible for directing the different immune responses during disease progression. Subsequently, we applied a systems biology approach to identify common and species-specific master regulators. We found promising master regulators of duck genes in lung and ileum (RUNX2, SMAD3, SMAD4, and ETS1), which could be responsible for the duck's effective differential gene expression in response to HPAI infection. Master regulators that were identified for the chicken orthologous gene set represent regulators that could be important for the effective regulation of gene expression after AIV infection, yet remain unsuccessful in living organisms. These master regulators include EGR1, FOS, SRF, and SP1, and could be interesting targets for future studies, since they could switch on several pathways targeting the genes that are important to the successful alleviation of HPAI infection. Based on our results, we highlight the importance of the RLR signaling, TGF- $\beta$ , and the JAK/STAT pathways for virus defense in chickens and ducks. We are aware that the amount of mRNA does not necessarily reflect the amount of proteins that are available in living cells. For that reason, we emphasize the need for experimental data to assess protein availability, as well as the roles of master regulators and pathways in living organisms. To the best of our knowledge, there are no studies on altered immunity in duck after knockouts, overexpression or mutations in the identified upstream pathways. Therefore, knock-out, knock-in, or overexpression experiments in both chicken and duck would be of great interest. While this is beyond our current capabilities, it would be an important objective for future studies to investigate.

## 6.7. Supplementary Materials

The following supplementary material is available via the original publication <https://doi.org/10.3390/biology11020219>. Figure S1: Treemaps of the DEG sets regarding GO:Biological Processes of chicken and duck, Figure S2: Treemaps of the DEG sets regarding GO:Molecular Functions of chicken and duck, Figure S3: Full-size image of

---

Figure 6.4, Figure S4: Schemes of the upstream regulatory networks revealing the top five master regulators of chicken and duck based on the TF-TF cooperation results, Figure S5: Schemes of the upstream regulatory networks revealing the top five master regulators of chicken and duck based on the TF enrichment results, Table S1: DEG sets for all experimental conditions, Table S2: GO-term enrichment of all DEG sets of chicken and duck, Table S3: enriched TFBSs in different promoter sets of chicken and duck, Table S4: TF cooperation networks of chicken and duck, File S1: PWMs included in the custom avian PWM profile, File S2: Cytoscape session of TF cooperation networks as .cys file.

## 7. Discussion

In this chapter, I discuss the methods applied in this thesis and the biological relevance of the results of my four publications, as described in the previous chapters. This chapter is partly based on the original publications [1–4].

### 7.1. Methodical Discussion

#### 7.1.1. Identification of rSNPs

In Chapters 3, 4, and 5, I developed a pipeline in order to identify a genome-wide collection of rSNPs which I applied to different animal and plant species. This pipeline requires as input (i) a library of PWMs representing the TFBSs, (ii) a reference genome, (iii) a SNP catalog, and (iv) gene annotations. Firstly, I selected the SNPs which are located in the promoter regions of all genes and analyzed them to detect their impact on the TFBSs. For this purpose, the flanking sequence of  $\pm 25$  bp for each selected SNP was extracted from the reference genome. For each SNP, I created two copies of the flanking sequence: One with the reference allele in the SNP position and the second with its alternate variant. After that, by employing the MATCH<sup>TM</sup> algorithm [18], both sequences were scanned to predict the TFBSs with their affinity scores. Finally, to evaluate the impact of a SNP on the binding affinity of a TF, I inferred four different types of consequences for each SNP-TFBS pair: (i) “No Change”: the SNP has no effect on the TF binding; (ii) “Score-Change”: the binding affinity (i.e., MSS) is changed; (iii) “Loss of TFBS”: a TFBS is only found with presence of the reference allele, while the TFBS does not occur in the alternate allele; and (iv) “Gain of TFBS”: the TFBS appears only for the alternate allele.

In the rSNP prediction pipeline, one important step was the definition of the promoter regions, since this determines if a SNP is considered for further analyses. Even though the core promoter is considered to be positioned within  $\sim 200$  bp around the TSS, a wider promoter region can be targeted by TFs to regulate gene transcription [13]. Previous studies defined different promoter regions for TFBS prediction, ranging from  $-10$  kb to  $+10$  kb [15, 74, 91, 95–100]. Therefore, I defined a relatively wide promoter region of 7.5 kb upstream to 2.5 kb downstream of the TSS in order to overcome inaccuracies in the TSS prediction and to ensure the inclusion of the biological promoter [13]. However, it is important to note that the biological promoter is usually smaller and, hence, the web interface provides the possibility to filter for smaller user-defined promoter regions. In particular, in the application study on *B. napus* (Chapter 5), I applied the pipeline with a smaller promoter

region of  $-500$  bp to  $+100$  bp relative to the TSS, as similarly suggested in previous studies [13, 38, 68, 110, 172–176]. Thereby, I emphasize that the choice of the 10 kb promoter region should not be interpreted as a biologically correct promoter region, it merely gives the user the possibility to search in a broader regulatory region. However, as shown in the rapeseed application project in Chapter 5, the use of a smaller region is quite reasonable.

In the next step of the pipeline, I identified SNPs located in the promoter region and added further filtering steps. More specifically, I discarded all insertions and deletions (indels) and SNPs having more than one alternate allele. With these filtering steps, the pipeline is focusing on the most basic form of sequence variation and hence I concentrate on a straight forward interpretation of results on the website. Especially the interpretation of consequences of SNPs having more than two alleles, would increase the complexity of the results, and hence also the visualization on the website. Therefore, I decided to keep the presentation of the results and their interpretability simple and clear. Nevertheless, it is possible to extend the database to other types of variants in the future.

In order to infer consequences for each SNP-TFBS pair, such as “Loss of TFBS” or “Gain of TFBS”, a clear definition of the reference and the alternate allele for each SNP is crucial. In the pipeline, I obtained the alleles from the input SNP catalog (as GVF file [121, 122]). In very few cases, especially in the plant species tomato, Asian rice (*Indica*), and sorghum (Chapter 4), I observed that the alleles of several SNPs deviate from the reference genome, in particular, their reference alleles were not present at the SNP position in the reference genome. An example for this issue can be shown with the tomato SNP *vcZYOCUX* (T/A), where the base at the respective position in the reference genome is G.<sup>1</sup> Such issues indicate that there is still a need for further investigation to improve the genome sequences as well as SNP annotations. In our pipeline, we excluded such SNPs from further analysis to ensure a high reliability of our results.

In the literature, there exist a variety of tools and databases investigating the prediction of rSNPs, either by using experimental and published data [74, 88, 92], or, similar to my studies, by predicting the effect of a SNP on TF binding [17, 47, 89–91]. However, most of the tools and databases focus on humans or a few model organisms. In the following, I address the rSNP detection methods applied in other studies and compare them to the pipeline used in this thesis. These include only a selection of similar databases, a comprehensive summary of further tools and databases is provided in Table S1 of Chapter 3.

SNP2TFBS [17] is a database of human rSNPs in which SNPs are stored together with annotations, such as whether they are predicted to eliminate, create or change one or more TFBSs. In contrast to *agReg-SNPdb* and *agReg-SNPdb-Plants*, where the TFBS prediction is based on the TRANSFAC database of TFBSs, in SNP2TFBS the prediction of TFBSs is performed based on the JASPAR database [178]. Furthermore, instead of analyzing each SNP separately by extracting the flanking sequence from the reference genome, Kumar

<sup>1</sup>[https://plants.ensembl.org/Solanum\\_lycopersicum/Variation/Explore?r=1:39003479-39004479;v=vcZYOCUX;vdb=variation;vf=3506065](https://plants.ensembl.org/Solanum_lycopersicum/Variation/Explore?r=1:39003479-39004479;v=vcZYOCUX;vdb=variation;vf=3506065), accessed on 1 November 2022

et al. (2016) [17] generated an alternate human genome, with all variant positions being replaced by the alternate allele. However, this approach does not allow an interpretation of the effects of single SNPs found in a certain individual, independent of neighbouring alleles. Nevertheless, in the case of two SNPs located close together such that both SNPs are located in the same TFBS, this should be considered and investigated in each case individually.

The database atSNPsearch [90] stores all human SNP-TFBS pairs which were identified with the tool atSNP (affinity testing for regulatory SNPs) [47]. atSNP takes as input a catalog of SNPs as well as motif files and outputs a list of SNP-TFBS pairs together with p-values indicating the significance of TF binding compared to a random background sequence. In atSNPsearch, each human SNP from dbSNP [258] is analyzed with respect to JASPAR [178] and ENCODE [259] TFBS motifs, independent of the SNPs location within a promoter region, and hence, they avoid making any assumption on a putative promoter region. However, this approach is relatively resource-intensive and inconvenient for analyzing and storing data sets from multiple species.

INFERNO (INFERring the molecular mechanisms of NONcoding genetic variants) [91] is a method which integrates different data such as GWAS summary statistics and LD structure from the 1000 Genomes Project [260] to identify putative regulatory variants underlying an association signal. For motif discovery, Amlie-Wolf et al. [91] apply the HOMER (Hypergeometric Optimization of Motif EnRichment) tool suite [261]. The inclusion of GWAS summary statistics and LD structure gives an interesting objective for agReg-SNPdb and agReg-SNPdb-Plants and should be considered in the future.

SNP@Promoter [262] is a database that contains human SNPs, TFBSs, and their overlaps located within promoter regions ranging from  $-5$  kb to  $+500$  bp. Similar to agReg-SNPdb, they use the TFBS prediction tool MATCH<sup>TM</sup> [18], but instead of inferring certain consequences of SNPs on TF binding, they concentrate on positional information only.

### 7.1.2. TFBS Prediction

The prediction of TFBSs commonly relies on position weight matrices (PWMs), which are obtained and updated frequently based on past and new experimentally verified TFBSs. These PWMs are stored in public or commercial databases and can be used as input for different TFBS prediction tools. In the databases agReg-SNPdb and agReg-SNPdb-Plants as well as in the application project on avian influenza, I applied the original MATCH<sup>TM</sup> tool [18] using the commercial database TRANSFAC [93]. In contrast to this, in the application study based on rapeseed, I applied the MATCH<sup>TM</sup> algorithm based on the JASPAR database [178], a publicly available database for PWMs. One major advantage of the MATCH<sup>TM</sup> algorithm with TRANSFAC lies in the availability of so-called PWM profiles provided by TRANSFAC. This enables the usage of a customized subset of PWMs with PWM-specific cut-offs, as e.g., the vertebrate specific profile used in Chapter 3, the plant specific profile used in Chapter 4, or the avian specific profile used in Chapter 6. By using such PWM profiles, it is possible to concentrate on the binding sites specific to a certain

taxon (e.g., the vertebrate specific profile) or disease (e.g., the avian specific profile concentrating on the binding sites of TFs, which are important during an infection with avian influenza). With the JASPAR database, this information is not available directly and such profiles must be created by the user (as e.g., the plant specific profile used in Chapter 5). Furthermore, the PWM-specific cut-offs minimizing e.g., the false positive rate, which are provided by TRANSFAC are not available in the JASPAR database and, hence, a cut-off free usage of JASPAR profiles can lead to a high number of false positive predictions. For this reason, I filtered the TFBS predictions using the JASPAR database in Chapter 5 based on a matrix similarity score threshold of 0.85, as recommended in [171].

Although conventional PWM-based TFBS prediction tools are widely used and show good results, they are highly dependent on the quality of the PWM annotations and hence deliver different results depending on the used databases. Hence, one challenge of PWM-based methods is the interpretation of binding affinity scores and the calculation of appropriate cut-offs, determining whether a binding event is predicted or not [18, 20]. Another limitation of PWM-based predictions is the fact that they assume the independence of each position within the PWM, with each position contributing independently to the overall binding affinity score. This might not correctly reflect the complexity of binding processes between the DNA and TFs [19, 20]. In addition, PWMs are not very well suited to represent the binding sites of some TF classes such as TF dimers, which consist of two conserved sequences interrupted by a variable sequence [263].

To tackle these limitations, different supervised machine learning methods, e.g., Bayesian networks, Markov models, support vector machines or neural networks, are developed [19, 23]. However, due to the extensive work done in the past to generate high quality PWMs and due to the lack of available data for model training (large numbers of positive and negative sequences, i.e., sequences in which a binding site is present and sequences in which no binding site is present, respectively) in machine learning based approaches, PWM-based models are still widely used and are usually the method of choice when it comes to prediction and visualization of TFBSs [19, 264].

### **7.1.3. Random Forest-Based Feature Selection to Identify SNP-Phenotype Associations**

In genome-wide association studies (GWAS), each SNP is tested for its association with a specific phenotype, which is either qualitative or quantitative. Although classical GWAS analysis is a well-established and straightforward method, in practice it entails several challenges. For example, the presence of confounding effects in the data causing bias, such as population stratification or relatedness among individuals, can inflate prior assumptions about the distribution of SNP effects, leading in particular to false positive predictions [265–267]. Furthermore, in classical GWAS each SNP is tested individually, and SNP interactions such as epistasis cannot be captured [265].

In order to overcome these challenges, in Chapter 5 a random forest (RF) feature selection algorithm was applied to perform a machine learning-based GWAS. The idea of RF is to grow multiple decision trees based on random subsets of observations that can help partition the data into subsets of highest possible purity with respect to a variable outcome, that is, the response variable [268]. In contrast to RF-based classification, where the RF model is trained in order to predict a response variable for new observations, a random forest-based feature selection is used to rank the importance of the input features and to identify the most important ones with respect to the variable outcome. In the classical sense, this is used to reduce the number of input features or attributes before training the RF classification model in order to reduce complexity. In the case of GWAS, a RF-based feature selection can be used to detect those features, i.e., SNPs, which are most important with respect to a variable outcome, i.e., the phenotype [129].

The Boruta algorithm [179], an RF-based feature selection wrapper, adds so-called shadow attributes, i.e., attributes derived from random permutations of features, to assess a SNP as important only if its importance score is significantly higher than the maximum importance of the shadow attributes. The pseudo-code for the Boruta algorithm is given in Algorithm 1 in Section 5.4.4. A major advantage of the Boruta algorithm compared to other RF algorithms is that it is specifically suited for high-dimensional GWAS data where the number of features (SNPs) is much higher than the number of observations (samples), which is a common feature of genotype data sets [269]. By applying the Boruta algorithm on 41,117 rapeseed rSNPs, 2.7% of them (1141) were found to be significantly associated with oil-content and -quality. These results, together with the transcriptomics results, provided important insights in the causal interaction between rSNPs, TFs, and DEGs (Chapter 5).

#### 7.1.4. Upstream Analysis to Identify Master Regulators

The level of gene expression, ideally measured during an RNA-seq experiment, provides insights into gene expression patterns and differentially expressed genes, which are important during a specific experimental condition such as a disease. The classical approaches consist of identifying Gene Ontology (GO) categories or metabolic or signaling pathways enriched among a set of DEGs in order to identify mechanisms or pathways in which the proteins encoded by the DEGs are involved. These approaches can be referred to as "downstream analyses" as they can help to unravel the mechanisms which are caused or triggered by the differential gene expression [213]. In contrast, another strategy is the so-called "upstream analysis" introduced by Koschmann et al. [213] that aims to identify the mechanisms causing the observed gene expression changes. This includes analyzing the promoter sequences of a set of DEGs, identifying the TFs involved in transcriptional regulation, and revealing the signaling pathways leading to the activation of these TFs [213]. In a final step, so-called master regulators are identified, which can be described as key nodes or convergence points within the complex regulatory networks of upstream pathways, which are mostly found at the top of the regulatory hierarchy [207, 213].

The algorithm used for detecting master regulators is part of the GeneXplain platform [214]. It requires as input a set of TFs or molecules and reconstructs upstream signaling pathways that together form a complex network based on the GeneWays database [215]. Based on this upstream network, using a shortest path algorithm, convergence points are identified as master regulators affecting a high number of both input molecules and total molecules in the network [213]. Those master regulators are usually found at the top of the regulatory hierarchy and can simultaneously influence the expression of a complete set of genes. Originally, this method is used to identify novel drug targets based on a set of DEGs or multi omics data [270].

In the avian influenza project (Chapter 6), the upstream analysis was applied to the chicken data in a novel sense. That is, rather than using DEGs observed in chicken as the basis for the upstream analysis, chicken orthologs of the DEGs observed in duck following AIV infection were used. Assuming that the genes differentially expressed in the duck are important for the duck's effective immune response, I hypothesized that differential expression of the corresponding chicken orthologs could elicit a successful immune response in chicken. Therefore, I sought to identify the master regulators that activate differential expression of chicken genes that were not differentially expressed in the living organism. Thus, this study offers a novel application approach to upstream analysis and therefore provides interesting and unexplored results. Nevertheless, it is crucial that the results of this analysis are validated and supported with experimental data.

#### **7.1.4.1. Applying the Upstream Analysis to Non-Human Species**

One important aspect of the upstream analysis is its applicability to non-human species, in this case to chicken and duck. In general, the GeneWays database is a pathway database based on automatic text mining and is therefore principally not exclusively based on human studies. In practice, however, most studies describing novel pathways or alternative paths to existing ones are based on human, mouse, or corresponding cell lines. This creates an imbalance in such databases towards human studies. Nevertheless, due to the lack of species specific pathway databases and especially due to the lack of experimental data to establish such databases, previous studies also used similar pathway databases to perform an upstream analysis in different animal species [129, 210, 211, 271–273]. However, the reduced applicability of such pathway databases to animal species increases the need to experimentally validate the findings proposed in my studies.



## 7.2. Biological Discussion

### 7.2.1. Regulatory Impact of Transcription Factors and rSNPs

TFs orchestrate the entirety of cellular processes leading to tissue development, tissue differentiation or responses to the environment and, thus, act as natural master regulators in higher organisms. By binding to the regulatory sequences, acting cooperatively as pairs and complexes, they can orchestrate the gene transcription, which makes them promising candidates as breeding targets to control complex traits in crop and animal breeding [146]. Mainly, but not exclusively, their binding is determined by sequence properties in the promoter region. Hence, the identification and analysis of rSNPs which are located within TFBSs and influence TF binding affinity is of major importance to identify candidate variants responsible for differential expression or even a specific phenotype.

However, the sequence alone and the binding sites it contains are not the only key to deciphering the regulatory mechanisms underlying an observed gene expression rate. In living organisms, other factors can also influence the level of gene expression. One such very important factor is the accessibility of chromatin, with DNA methylation, histone modification, and DNA structure in general being of great importance [274]. While the DNA sequence forms the basis of TF binding in general, an actual binding event also depends on the presence or absence of a particular TF, i.e., protein concentration or the presence or absence of other TFs that bind in a competitive or cooperative manner to the same or adjacent binding sites [16, 17, 69]. In addition, TF binding is highly context specific and hence differs among cell-types or tissues, which cannot be explained by the genomic sequence alone [13]. Apart from regulatory elements binding to the promoter region, other regulatory elements can be found in the transcribed region which can also regulate gene expression. For example, microRNAs (miRNAs), which mostly bind to the 3'UTR, often target the mRNA of TFs and inhibit their translation. In this regard, feedback regulation of miRNAs by their own targets plays an important role, especially with respect to TF availability [19]. Furthermore, in plants there are so-called riboswitches, which are RNA elements in the untranslated region that attract the binding of small molecules and, thus, regulate the transcription and translation of the gene itself [275].

In higher organisms, TF binding often occurs in a complex interplay and includes cooperation between proximal and distal regulatory elements (promoters and enhancers), by which so-called chromatin loops are formed [68, 276]. These chromatin loops can also affect the level of gene expression. However, their prediction is highly complex and still much research is needed [276].

In the studies presented, I have mainly focused on the genomic sequences, more specifically on the promoter regions and the rSNPs and TFBSs contained therein. For future work, the investigation of further regulatory mechanisms underlying differential gene expression can help to gain a more comprehensive understanding of the entirety of gene regulatory processes. Nevertheless, this thesis and the herein contained studies can be seen as one further

step leading towards the deciphering of differential gene expression underlying different traits in animal and plant species.

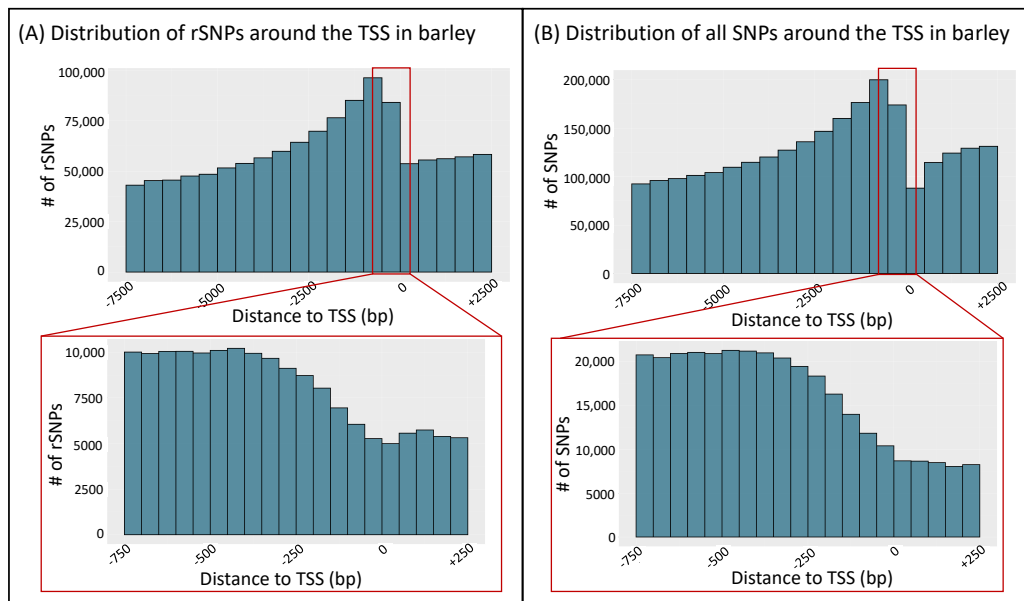
### 7.2.2. The Distribution of rSNPs and SNPs around the TSS

In agReg-SNPdb and agReg-SNPdb-Plants, I provided a statistical overview of the data stored in the databases (Sections 3.5.3 and 4.5.3). To gain a better insight into the distribution of the rSNPs in the promoter regions, I investigated their genomic positions relative to the transcription start sites (TSS). Interestingly, I observed different patterns which were abundant in both animal as well as plant species. In several species such as barley, Asian rice Japonica, maize, chicken, or sheep, I observed a pattern where the sequence is protected from variation in close proximity to the TSS, while the number of rSNPs increases with increasing distance in the upstream direction [1, 3, 4, 13]. In contrast, I have observed a different pattern in species such as Asian rice Indica, bread wheat, durum wheat, or cattle. The number of rSNPs increases with downstream distance and tends to accumulate around the TSS or in the direct downstream region (see Figure S4 of Chapter 3 and Figure S2 of Chapter 4). Notably, this tendency is probably not based on the different characteristics of promoter regions and rSNPs in different species, but rather just represents the differences of the SNP data stored in Ensembl [94] or Ensembl Plants [119]. In Figures 7.1 and 7.2, I show a comparison of distributions of (i) rSNPs and (ii) all SNPs around the TSS. It is evident that the distributions of rSNPs mainly reflect the distributions of SNPs in the promoter regions. This observation shows that the data stored in public databases, such as Ensembl or Ensembl Plants, can show completely different patterns for different species, which could cause bias for specific analyses and should be used with caution in studies based on more than one species.

### 7.2.3. Oil Content and Quality in Rapeseed

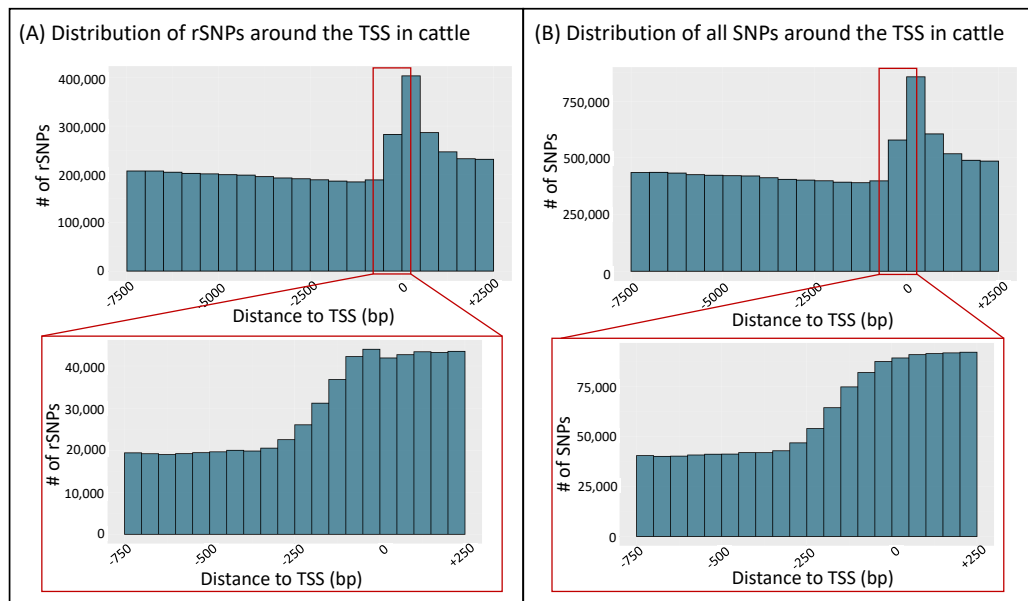
In Chapter 5, I performed a systematic analysis using multi-omics data (genomics, transcriptomics, and proteomics) to investigate the complex interplay between rSNPs, TFs and DEGs in *B. napus*. To date, the regulatory mechanisms and pathways controlling oil content and -quality in different rapeseed cultivars have not been deciphered [49, 58]. To this end, I investigated a genomics and transcriptomics data set by Lu et al. [49] to assess the genetic programs of two cultivars, namely (i) Zhongshuang11 (ZS11) characterized by a double-low accession (00, low erucic acid and low glucosinolate) and a high oil content and; (ii) Zhongyou821 (ZY821) with double-high accession (++, high erucic acid and high glucosinolate) and low oil content.

In this study, I first applied the rSNP prediction pipeline developed in agReg-SNPdb and agReg-SNPdb-Plants to the genomics data and identified a total of 41,117 rSNPs, predicted to cause either a gain or loss of TFBS. This represents the first genome-wide collection of rSNPs in *B. napus* and can be useful for scientists in order to interpret results from associa-



**Figure 7.1.: Distribution of rSNPs and all SNPs around the TSS of barley.** (A) shows the distribution of the distances between rSNPs and the TSS and (B) shows the distribution of the distances between all SNPs and the TSS. The upper histograms show the number of rSNPs/SNPs in the large promoter region ( $-7.5$  kb to  $+2.5$  kb) in 500 bp intervals and the enlargements show the number of rSNPs/SNPs in the proximal promoter region ( $-750$  bp to  $+250$  bp) in 50 bp intervals.

tion analyses such as GWAS, gene expression experiments, expression QTL (eQTL) studies, or population studies. By incorporating the phenotype, i.e., the cultivars with high or low oil quality or -content, I identified a total of 1141 rSNPs, which are significantly associated with the phenotype, herein referred to as *important rSNPs*. Using the transcriptomics data, I identified 11,442, 3234, 4198, and 2318 DEGs in the tissues flower, leaf, stem, and root, respectively, which were differentially expressed between the cultivars. By investigating their promoter regions in terms of enriched TFBSs, I showed that members of the DOF, MYB, NAC, GATA, or TCP TF-families were identified as being enriched exclusively for a particular tissue, whereas the TFs of the bHLH or bZIP class and members of the BES1 family seem to play important regulatory roles in several tissues. A closer look at these DEGs and the rSNPs located in their promoter regions, revealed that 5847 (flower), 1604 (leaf), 2174 (stem), and 1240 (root) rSNPs are harboring in the promoter regions of DEGs and are potential regulators of expressed levels. Interestingly, I observed that approximately 50% of DEGs contain on average one rSNP within the promoter region and that 27% of *important rSNPs* are located within the promoter of a DEG. This supports the hypothesis that



**Figure 7.2.: Distribution of rSNPs and SNPs around the TSS of cattle.** (A) shows the distribution of the distances between rSNPs and the TSS and (B) shows the distribution of the distances between all SNPs and the TSS. The upper histograms show the number of rSNPs/SNPs in the large promoter region (−7.5 kb to +2.5 kb) in 500 bp intervals and the enlargements show the number of rSNPs/SNPs in the proximal promoter region (−750 bp to +250 bp) in 50 bp intervals.

rSNPs might be important regulators of the gene expression observed in the two cultivars. Finally, I studied the regulatory interplay between rSNPs, TFs and their corresponding DEGs of interest by identifying DEGs harboring *important rSNPs* in their promoter region. I found several promising tissue-specific DEGs, such as the *BnaLACS1-4* or the *ECR* genes, whose regulation is likely to be influenced by the “Loss” or “Gain” of a TFBS caused by *important rSNPs*. The TFs overlapping these *important rSNPs* provide a promising basis for further investigation of their regulatory roles and underlying pathways that lead to the differences between the two cultivars.

By investigating the tissues flower, leaf, stem and root, I assessed different vegetative and floral tissues underlying the two cultivars. In addition, the seed tissue may be of great importance in terms of its roles in fatty acid synthesis, transport, and accumulation. In a follow-up study by Rajavel et al. (2021a) [120], we studied a time-series transcriptomics data set to investigate the gene expression in the seed tissue of the same cultivars, ZS11 and ZY821. By identifying monotonically expressed genes, which are monotonically expressed either in ascending or descending patterns with time, we captured the multi-stage progres-

sion during seed development. In line with the results of TFBS enrichment in the tissues flower, leaf, stem, and root, we have shown that members of TF families such as GATA, DOF, NAC, or MYB are important regulators of genes with a monotonic expression pattern in both cultivars in the seed tissue by forming TF cooperations [120].

#### 7.2.4. Avian Influenza in Chicken and Duck

In Chapter 6, I investigated the mechanisms underlying the different immune responses in chicken and duck after an infection with avian influenza. To date, the mechanisms responsible for the susceptibility of chickens and the effective immune response of ducks are not fully understood. To this end, I performed a systematic analysis to investigate the transcriptional gene regulation underlying the disease progression in the two species and identified upstream regulators, including TFs, their complex interplay, and master regulators.

For this purpose, I firstly compared the promoter regions of the two species in terms of TFBS enrichment. The results suggest that there are large differences between the promoter regions of orthologous genes of the two species in terms of TFBS enrichment (Section 6.5.1). In particular, I identified promising TF families, which are important regulators in chicken (TF families such as SMAD, IRF, and ETS) or in duck (TF families such as FOX, STAT, and POU).

Although TF enrichment provides important insights, I investigated the specific partner choices of TFs to unravel their complex interplay, which could be responsible for directing the different immune responses after virus infection. Interestingly, the results indicate that, while single, enriched TFs in the promoter regions are rather species-specific and differ greatly between the species, the TF-TF cooperation networks of both species share many common features and TF clusters seem to be preserved or classified by specific partner alterations (see Section 6.5.2).

Subsequently, I applied a systems biology approach to identify common and species-specific master regulators. I found promising master regulators of the duck DEGs for lung and ileum tissues (RUNX2, SMAD3, SMAD4, and ETS1), which could be responsible for the duck's effective differential gene expression after viral infection. The duck master regulators represent important regulators to effectively controlling the virus replication in the host, and hence, they can be seen as important targets in the chicken. Master regulators that were identified for the chicken orthologous gene set are of particular importance in this study (e.g., EGR1, FOS, SRF, and SP1). They represent regulators that could be important for the effective regulation of gene expression after AIV infection, yet remain unsuccessful in living organisms and they may be of particular interest for future studies as they could switch on several pathways targeting the genes that are important to the successful alleviation of HPAI infection.

As it is common in sequence-based *in silico* analyses, the identified master regulators do not provide sufficient insight into the amount of proteins available in the living cells. For that reason, I emphasize the need for experimental data to assess protein availability, as well as

the roles of master regulators and pathways in living organisms. To the best of my knowledge, there are no studies on altered immunity in chicken or duck after gene knockouts, overexpression or mutations in the identified upstream pathways. Therefore, knock-out, knock-in, or overexpression experiments in both chicken and duck would be of great interest. While this is beyond our current capabilities, it would be an important objective for future studies to investigate. Nevertheless, the identified upstream mechanisms, and in particular the master regulators, offer interesting starting points and could be considered in the future as potential drug targets or biomarkers in chicken to reduce their susceptibility.

Similar to the study on rapeseed (Chapter 5), the investigation of rSNPs could provide important insights into the regulatory mechanisms of the two species. However, while the detection of rSNPs was rather straight-forward for the two rapeseed cultivars that share one reference genome, it is not as straight-forward in two different species with different reference genomes. In order to address this, I was involved in another study<sup>2</sup>, in which we investigated the role of SNPs that are shared between chicken and duck at orthologous positions, which can be referred to as coincident SNPs (coSNPs) [277]. In particular, we investigated coSNPs in promoters of the duck DEGs, differentially expressed after avian influenza infection, which were also investigated in this study (Chapter 6). Consequently, we identified coSNPs which also have a regulatory role by affecting the binding affinity of TFs. By comparing the effects on TF binding caused by coSNPs in both species, we obtained novel insights into the different mechanisms underlying the gene expression after AIV infection in chicken and duck. The results highlight the potential importance of the TFs ASCL2, RAD21, SP1 and the TF families SMAD, PAX, FOX, E2F, IRF and STAT in the regulation of immune response-related genes.

---

<sup>2</sup>Master's thesis by Hendrik Bertram (see Impact in Section 1.1)

## 8. Conclusion

In this chapter, I conclude my work and provide an outlook for future studies. This chapter is partly based on the original publications [1–4].

The transcriptional regulation of gene expression in higher organisms is essential for various biological processes, which are mainly governed by transcription factors and their combinatorial interplay. In contrast to the process of translation, the transcriptional machinery and its regulatory mechanisms are far from being deciphered. TFs regulate the transcription in a highly context-specific manner as a response to specific environmental conditions by binding to the TFBSs in promoter regions of their target genes. Regulatory SNPs that are located in TFBSs can lead to a change in the binding affinity of TFs and, in extreme cases, even result in the disruption of a TFBS or the creation of a new TFBS.

In my first study (Chapter 3), I created agReg-SNPdb, a database storing genome-wide collections of rSNPs for agricultural animal species. In this study, I developed the pipeline to detect rSNPs and performed a literature survey to show that the obtained results are in agreement with previous experimental and *in silico* studies. In order to ensure convenient database search, I developed a website to query agReg-SNPdb by SNP IDs, chromosomal regions, or genes. As an extension of agReg-SNPdb, I developed agReg-SNPdb-Plants a database of regulatory SNPs for agriculturally important plant species (Chapter 4). To the best of my knowledge, agReg-SNPdb and agReg-SNPdb-Plants are the first databases of regulatory SNPs for animal and plant species of importance for agriculture and breeding. The releases of the databases agReg-SNPdb and agReg-SNPdb-Plants are important steps toward the understanding of gene regulation in the animal and plant sciences. Knowing whether a SNP causes a change in the binding affinity or even disrupts a TFBS or creates a new TFBS can be of predominant importance for the interpretation of results from, e.g., GWAS, gene expression experiments, eQTL analyses, or population studies. The newly gained information can be used in genomic selection and marker establishment by identifying possibly causal rSNPs and revealing the underlying regulatory mechanisms of specific traits or diseases. With ongoing sequencing progress and genome annotations for different species in Ensembl and Ensembl Plants, the databases should be updated and extended regularly in the future.

In Chapter 5, I present a study on *B. napus* where I demonstrated the application of rSNPs together with multi-omics data to perform a systematic analysis of the complex interplay between rSNPs, TFs, and DEGs underlying oil content and -quality. As a result of this analysis, I obtained: (i) a genome-wide collection of rSNPs; (ii) their significant association with the *B. napus* cultivars differing in oil content; (iii) their consequences on TF binding;

and (iv) the DEGs of four tissues (flower, leaf, stem, and root) whose expression could be strongly affected by the occurrence of *important rSNPs* within their promoter regions. In this systematic approach, I focused mainly on promoter regions, and, thus, my results provide important insights into regulatory processes at the transcriptional level. For future work, the investigation of further regulatory mechanisms underlying differential gene expression, as, e.g., post-transcriptional regulation such as microRNA binding or riboswitch activity can help to gain a comprehensive understanding of the entirety of gene regulatory processes. Nevertheless, my study can be seen as one further step leading towards the deciphering of differential gene expression underlying the different *B. napus* cultivars and the genome-wide collection of rSNPs provides a basis for upcoming studies on different traits in *B. napus*.

In the final study (Chapter 6), I investigated the transcriptional gene regulation controlling the expression of genes induced by an infection with avian influenza in chicken and duck. For this purpose, I identified upstream regulators, including TFs, their complex interplay, and master regulators, which could stimulate an effective immune response in ducks following viral infection, while being dysfunctional in chicken. I found promising master regulators of duck genes in lung and ileum, which could be responsible for the duck's effective differential gene expression in response to HPAI infection. Master regulators that were identified for the chicken orthologous gene set represent regulators that could be important for the effective regulation of gene expression after AIV infection, but do not act or are not present in living organisms. In particular, these master regulators could be interesting targets for future studies, since they could switch on several pathways targeting the genes that are important to the successful alleviation of HPAI infection. Based on these results, I emphasize the need for experimental data to assess the protein availability, as well as the roles of master regulators and pathways in living organisms. In this regard, knock-out, knock-in, or overexpression experiments in both chicken and duck would be of great interest. While this is beyond our current capabilities, it would be an important objective for future studies.





## 9. List of Abbreviations

<b>AIV</b>	avian influenza virus
<b>AP</b>	activator protein
<b>APOA2</b>	apolipoprotein (apo-) A-II
<b>bHLH</b>	basic/helix-loop-helix
<b>bp</b>	base pairs
<b>coSNP</b>	coincident SNP
<b>CSS</b>	core similarity score
<b>DB</b>	database
<b>DEG</b>	differentially expressed gene
<b>DNA</b>	deoxyribonucleic acid
<b>dpi</b>	days post infection
<b>ECR</b>	trans-2,3-enoyl-CoA reductase
<b>eQTL</b>	expression quantitative trait locus
<b>FABP4</b>	fatty acid-binding protein 4
<b>FDR</b>	false discovery rate
<b>FOX</b>	forkhead box
<b>GO</b>	Gene Ontology
<b>GWAS</b>	genome-wide association study
<b>HOX</b>	homeobox
<b>IAV</b>	influenza A virus
<b>IFITM</b>	interferon-induced transmembrane protein
<b>IFN</b>	interferon
<b>IRF</b>	interferon regulatory factor
<b>ISG</b>	interferon-stimulated gene
<b>JAK</b>	Janus-kinase
<b>kb</b>	kilobase
<b>LACS</b>	long-chain Acyl-CoA synthetase
<b>LD</b>	linkage disequilibrium
<b>LFC</b>	log <sub>2</sub> fold change
<b>LGB</b>	$\beta$ -Lactoglobulin

<b>LPAI</b>	high pathogenic avian influenza
<b>LPAI</b>	low pathogenic avian influenza
<b>MAF</b>	minor allele frequency
<b>MEF2</b>	myocyte enhancer factor 2
<b>miRNA</b>	microRNA
<b>mRNA</b>	messenger RNA
<b>MSS</b>	matrix similarity score
<b>NF-1</b>	nuclear factor 1
<b>NGS</b>	next-generation sequencing
<b>PAX</b>	paired box factor
<b>PIT-1</b>	pituitary transcription factor 1
<b>PRL</b>	prolactin
<b>PWM</b>	position weight matrix
<b>QTL</b>	quantitative trait locus
<b>RF</b>	random forest
<b>RLR</b>	RIG-I-like receptor
<b>RNA</b>	ribonucleic acid
<b>RNA-seq</b>	RNA-sequencing
<b>rSNP</b>	regulatory SNP
<b>SNP</b>	single nucleotide polymorphism
<b>SP</b>	specificity protein
<b>SQL</b>	Structured Query Language
<b>SRF</b>	serum response factor
<b>STAT</b>	signal transducer and activator of transcription
<b>TAG</b>	triacylglycerol
<b>TF</b>	transcription factor
<b>TFBS</b>	transcription factor binding site
<b>TSS</b>	transcription start site
<b>UTR</b>	untranslated region
<b>VEP</b>	Variant Effect Predictor
<b>VLCFA</b>	very-long-chain fatty acids
<b>ZNF</b>	zinc finger protein
<b>ZS11</b>	Zhongshuang11
<b>ZS11</b>	Zhongyou821



## Bibliography

- [1] S. Klees, F. Heinrich, A. O. Schmitt, and M. Gültas, “agReg-SNPdb-Plants: A Database of Regulatory SNPs for Agricultural Plant Species”, *Biology*, vol. 11, no. 5, p. 684, 2022.
- [2] S. Klees *et al.*, “Comparative Investigation of Gene Regulatory Processes Underlying Avian Influenza Viruses in Chicken and Duck”, *Biology*, vol. 11, no. 2, p. 219, 2022.
- [3] S. Klees, F. Heinrich, A. O. Schmitt, and M. Gültas, “agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species”, *Biology*, vol. 10, no. 8, p. 790, 2021.
- [4] S. Klees *et al.*, “In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in Brassica napus L. Using Multi-Omics Data”, *International journal of molecular sciences*, vol. 22, no. 2, p. 789, 2021.
- [5] F. H. Crick, “On protein synthesis”, in *Symp Soc Exp Biol*, vol. 12, 1958, p. 8.
- [6] M. Morange, “The Central Dogma of molecular biology”, *Resonance*, vol. 14, no. 3, pp. 236–247, 2009.
- [7] E. Shelest and E. Wingender, “Systems biology of transcription regulation”, *Frontiers in Genetics*, vol. 7, p. 124, 2016, ISSN: 1664-8021.
- [8] T. D. Veenstra, “Omics in systems biology: Current progress and future outlook”, *Proteomics*, vol. 21, no. 3-4, p. 2 000 235, 2021.
- [9] N. A. Campbell *et al.*, *Biologie. 10., aktualisierte Auflage*. Pearson Deutschland GmbH, 2016.
- [10] J. D. Watson, T. A. Baker, A. Gann, S. P. Bell, M. Levine, and R. M. Losick, *Molecular biology of the gene*, 7th edition. Pearson Education India, 2004, ISBN: 978-0-321-90537-6.
- [11] A. Klug, “Rosalind Franklin and the discovery of the structure of DNA”, *Nature*, vol. 219, no. 5156, pp. 808–810, 1968.
- [12] A. Ralston and K. Shaw, “Gene expression regulates cell differentiation”, *Nat Educ*, vol. 1, no. 1, pp. 127–131, 2008.

- [13] M. Triska, V. Solovyev, A. Baranova, A. Kel, and T. V. Tatarinova, “Nucleotide patterns aiding in prediction of eukaryotic promoters”, *PLOS ONE*, vol. 12, no. 11, e0187243, 2017.
- [14] L. Pray, “Eukaryotic genome complexity”, *Nature Education*, vol. 1, no. 1, p. 96, 2008.
- [15] M. Stepanova, T. Tiazhelova, M. Skoblov, and A. Baranova, “A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas”, *Bioinformatics*, vol. 21, no. 9, pp. 1789–1796, 2005.
- [16] J. Fickett and A. Hatzigeorgiou, “Eukaryotic promoter recognition”, *Genome research*, vol. 7, no. 9, pp. 861–878, 1997.
- [17] S. Kumar, G. Ambrosini, and P. Bucher, “SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity”, *Nucleic acids research*, vol. 45, no. D1, pp. D139–D144, 2016.
- [18] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, “MATCH: a tool for searching transcription factor binding sites in DNA sequences”, *Nucleic acids research*, vol. 31, no. 13, pp. 3576–3579, 2003.
- [19] V. Boeva, “Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells”, *Frontiers in genetics*, vol. 7, p. 24, 2016.
- [20] G. D. Stormo, “DNA binding sites: representation and discovery”, *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.
- [21] B. Singh and S. K. Nath, “Identification of proteins interacting with single nucleotide polymorphisms (SNPs) by DNA pull-down assay”, in *Electrophoretic Separation of Proteins*, Springer, 2019, pp. 355–362.
- [22] W. W. Wasserman and A. Sandelin, “Applied bioinformatics for the identification of regulatory elements”, *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.
- [23] A. O. Degtyareva, E. V. Antontseva, and T. I. Merkulova, “Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases”, *International Journal of Molecular Sciences*, vol. 22, no. 12, p. 6454, 2021.
- [24] M. Levo and E. Segal, “In pursuit of design principles of regulatory sequences”, *Nature Reviews Genetics*, vol. 15, no. 7, pp. 453–468, 2014.
- [25] A. J. Brookes, “The essence of SNPs”, *Gene*, vol. 234, no. 2, pp. 177–186, 1999.
- [26] A. Vignal, D. Milan, M. SanCristobal, and A. Eggen, “A review on SNP and other types of molecular markers and their use in animal genetics”, *Genetics selection evolution*, vol. 34, no. 3, pp. 275–305, 2002.
- [27] F. H. Pettersson *et al.*, “Marker selection for genetic case–control association studies”, *Nature protocols*, vol. 4, no. 5, pp. 743–752, 2009.

- [28] S. L. Edwards, J. Beesley, J. D. French, and A. M. Dunning, “Beyond GWASs: illuminating the dark road from association to function”, *The American Journal of Human Genetics*, vol. 93, no. 5, pp. 779–797, 2013.
- [29] E. Rojano, P. Seoane, J. A. Ranea, and J. R. Perkins, “Regulatory variants: from detection to predicting impact”, *Briefings in bioinformatics*, vol. 20, no. 5, pp. 1639–1654, 2019.
- [30] Y. Itan, A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas, “The origins of lactase persistence in Europe”, *PLoS computational biology*, vol. 5, no. 8, e1000491, 2009.
- [31] L. Fang, J. K. Ahn, D. Wodziak, and E. Sibley, “The human lactase persistence-associated SNP- 13910\*T enables in vivo functional persistence of lactase promoter-reporter transgene expression”, *Human genetics*, vol. 131, no. 7, pp. 1153–1159, 2012.
- [32] E. K. Allen *et al.*, “SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans”, *Nature medicine*, vol. 23, no. 8, pp. 975–983, 2017.
- [33] Y. Wang *et al.*, “SNP rs17079281 decreases lung cancer risk through creating an YY1-binding site to suppress DCBLD1 expression”, *Oncogene*, vol. 39, no. 20, pp. 4092–4102, 2020.
- [34] K. V. Korneev *et al.*, “Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression.”, *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1866, no. 3, p. 165 626, 2020.
- [35] S. Konishi *et al.*, “An SNP caused loss of seed shattering during rice domestication”, *Science*, vol. 312, no. 5778, pp. 1392–1396, 2006.
- [36] L. Shi *et al.*, “Identification of promoter motifs regulating ZmeIF4E expression level involved in maize rough dwarf disease resistance in maize (*Zea Mays L.*)”, *Molecular genetics and genomics*, vol. 288, no. 3-4, pp. 89–99, 2013.
- [37] V. Jaiswal *et al.*, “Identification of novel SNP in promoter sequence of TaGW2-6A associated with grain weight and other agronomic traits in wheat (*Triticum aestivum L.*)”, *PLOS ONE*, vol. 10, no. 6, e0129400, 2015.
- [38] F. Heinrich *et al.*, “Identification of regulatory SNPs associated with vicine and convicine content of *Vicia faba* based on genotyping by sequencing data using deep learning”, *Genes*, vol. 11, no. 6, p. 614, 2020.
- [39] L. S. Lum, P. Dovč, and J. F. Medrano, “Polymorphisms of bovine  $\beta$ -lactoglobulin promoter and differences in the binding affinity of activator protein-2 transcription factor”, *Journal of Dairy Science*, vol. 80, no. 7, pp. 1389–1397, 1997.

- [40] Y. Liang, J. Cui, G. Yang, F. C. Leung, and X. Zhang, “Polymorphisms of 5’ flanking region of chicken prolactin gene”, *Domestic animal endocrinology*, vol. 30, no. 1, pp. 1–16, 2006.
- [41] J.-X. Cui, H.-L. Du, Y. Liang, X.-M. Deng, N. Li, and X.-Q. Zhang, “Association of polymorphisms in the promoter region of chicken prolactin with egg production”, *Poultry science*, vol. 85, no. 1, pp. 26–31, 2006.
- [42] M. Ballester *et al.*, “Analysis of the porcine APOA2 gene expression in liver, polymorphism identification and association with fatty acid composition traits”, *Animal genetics*, vol. 47, no. 5, pp. 552–559, 2016.
- [43] T. L. Bailey, N. Williams, C. Mischel, and W. W. Li, “MEME: discovering and analyzing DNA and protein sequence motifs”, *Nucleic acids research*, vol. 34, no. suppl\_2, W369–W373, 2006.
- [44] A. Sandelin, W. W. Wasserman, and B. Lenhard, “ConSite: web-based prediction of regulatory elements using cross-species comparison”, *Nucleic acids research*, vol. 32, no. suppl\_2, W249–W252, 2004.
- [45] W. Santana-Garcia *et al.*, “RSAT variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding”, *Computational and structural biotechnology journal*, vol. 17, pp. 1415–1428, 2019.
- [46] S. G. Coetzee, G. A. Coetzee, and D. J. Hazelett, “motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites”, *Bioinformatics*, vol. 31, no. 23, pp. 3847–3849, 2015.
- [47] C. Zuo, S. Shin, and S. Keleş, “atSNP: transcription factor binding affinity testing for regulatory SNP detection”, *Bioinformatics*, vol. 31, no. 20, pp. 3353–3355, 2015.
- [48] H. Pagès, “BSgenome: Infrastructure for Biostrings-based genome data packages and support for efficient SNP representation”, *R package*, 2016.
- [49] K. Lu *et al.*, “Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement”, *Nature communications*, vol. 10, no. 1, p. 1154, 2019.
- [50] B. Chalhoub *et al.*, “Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome”, *science*, vol. 345, no. 6199, pp. 950–953, 2014.
- [51] N. Lohani, D. Jain, M. B. Singh, and P. L. Bhalla, “Engineering Multiple Abiotic Stress Tolerance in Canola, Brassica napus”, *Frontiers in Plant Science*, vol. 11, p. 3, 2020.
- [52] W. Friedt and R. Snowdon, “Oilseed Rape”, in *Oil Crops*, Springer New York, 2009, pp. 91–126.
- [53] H. Becker, *Pflanzenzüchtung*. UTB GmbH, 2019.



- [54] R. Snowdon, W. Lühs, and W. Friedt, “Oilseed rape”, in *Oilseeds*, Springer, 2007, pp. 55–114.
- [55] G. Wu, Y. Wu, L. Xiao, X. Li, and C. Lu, “Zero erucic acid trait of rapeseed (*Brassica napus* L.) results from a deletion of four base pairs in the fatty acid elongase 1 gene”, *Theoretical and Applied Genetics*, vol. 116, no. 4, pp. 491–499, 2008.
- [56] S. V. Hatzig, J.-N. Nuppenau, R. J. Snowdon, and S. V. Schießl, “Drought stress has transgenerational effects on seeds and seedlings in winter oilseed rape (*Brassica napus* L.)”, *BMC plant biology*, vol. 18, no. 1, p. 297, 2018.
- [57] M. Gupta, P. B. Bhaskar, S. Sriram, and P.-H. Wang, “Integration of omics approaches to understand oil/protein content during seed development in oilseed crops”, *Plant cell reports*, vol. 36, no. 5, pp. 637–652, 2017.
- [58] G. K. Agrawal and R. Rakwal, *Seed development: OMICS technologies toward improvement of seed quality and crop yield*. Springer, Netherlands, 2012.
- [59] J. Smith *et al.*, “A comparative analysis of host responses to avian influenza infection in ducks and chickens highlights a role for the interferon-induced transmembrane proteins in viral resistance”, *BMC genomics*, vol. 16, no. 1, pp. 1–19, 2015.
- [60] D. E. Swayne, *Avian influenza*. John Wiley & Sons, 2009.
- [61] World Health Organization (WHO), *Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2020*, [https://www.who.int/publications/m/item/cumulative-number-of-confirmed-human-cases-for-avian-influenza-a\(h5n1\)-reported-to-who-2003-2022-27-june-2022](https://www.who.int/publications/m/item/cumulative-number-of-confirmed-human-cases-for-avian-influenza-a(h5n1)-reported-to-who-2003-2022-27-june-2022), Accessed: 2022-August-19, 2022.
- [62] D. Evseev and K. E. Magor, “Innate immune responses to avian influenza viruses in ducks and chickens”, *Veterinary sciences*, vol. 6, no. 1, p. 5, 2019.
- [63] P. B. Ranaware *et al.*, “Genome wide host gene expression analysis in chicken lungs infected with avian influenza viruses”, *PLOS ONE*, vol. 11, no. 4, 2016.
- [64] M. R. Barber, J. R. Aldridge Jr, R. G. Webster, and K. E. Magor, “Association of RIG-I with innate immunity of ducks to influenza”, *Proceedings of the National Academy of Sciences*, vol. 107, no. 13, pp. 5913–5918, 2010.
- [65] M. R. Barber, J. R. Aldridge Jr, X. Fleming-Canepa, Y.-D. Wang, R. G. Webster, and K. E. Magor, “Identification of avian RIG-I responsive genes during influenza infection”, *Molecular immunology*, vol. 54, no. 1, pp. 89–97, 2013.
- [66] F. Y. Looi *et al.*, “Creating disease resistant chickens: A viable solution to avian influenza?”, *Viruses*, vol. 10, no. 10, p. 561, 2018.

- [67] J. M. Franco-Zorrilla, I. López-Vidriero, J. L. Carrasco, M. Godoy, P. Vera, and R. Solano, “DNA-binding specificities of plant transcription factors and their potential to define target genes”, *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2367–2372, 2014.
- [68] L. Steuernagel, C. Meckbach, F. Heinrich, S. Zeidler, A. O. Schmitt, and M. Gültas, “Computational identification of tissue-specific transcription factor cooperation in ten cattle tissues”, *PLOS ONE*, vol. 14, no. 5, e0216475, 2019.
- [69] C. Meckbach, E. Wingender, and M. Gültas, “Removing background co-occurrences of transcription factor binding sites greatly improves the prediction of specific transcription factor cooperations”, *Frontiers in genetics*, vol. 9, 2018.
- [70] B. J. Hayes and H. D. Daetwyler, “1000 Bull Genomes project to map simple and complex genetic traits in cattle: applications and outcomes”, *Annual review of animal biosciences*, vol. 7, pp. 89–102, 2019.
- [71] A. O. Schmitt, J. Aßmus, R. H. Bortfeldt, and G. A. Brockmann, “CandiSNPer: a web tool for the identification of candidate SNPs for causal variants”, *Bioinformatics*, vol. 26, no. 7, pp. 969–970, 2010.
- [72] S. J. Goodswen, C. Gondro, N. S. Watson-Haigh, and H. N. Kadarmideen, “Funct-SNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases”, *BMC bioinformatics*, vol. 11, no. 1, p. 311, 2010.
- [73] T. Günther, A. O. Schmitt, R. H. Bortfeldt, A. Hinney, J. Hebebrand, and G. A. Brockmann, “Where in the genome are significant single nucleotide polymorphisms from genome-wide association studies located?”, *Omics: a journal of integrative biology*, vol. 15, no. 7-8, pp. 507–512, 2011.
- [74] L. Guo and J. Wang, “rSNPBase 3.0: an updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks”, *Nucleic acids research*, vol. 46, no. D1, pp. D1111–D1116, 2017.
- [75] G. Macintyre, J. Bailey, I. Haviv, and A. Kowalczyk, “is-rSNP: a novel technique for in silico regulatory SNP detection”, *Bioinformatics*, vol. 26, no. 18, pp. i524–i530, 2010.
- [76] N. E. Buroker, “VEGFA rSNPs, transcriptional factor binding sites and human disease”, *The Journal of Physiological Sciences*, vol. 64, no. 1, pp. 73–76, 2014.
- [77] M. De Gobbi *et al.*, “A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter”, *Science*, vol. 312, no. 5777, pp. 1215–1217, 2006.
- [78] S. F. Grant, D. M. Reid, G. Blake, R. Herd, I. Fogelman, and S. H. Ralston, “Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I  $\alpha$  1 gene”, *Nature genetics*, vol. 14, no. 2, p. 203, 1996.

- [79] M. D. Littlejohn *et al.*, “Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition”, *Scientific Reports*, vol. 6, no. 1, pp. 1–14, 2016.
- [80] M. Muhagheh-Dolatabady, “Single Nucleotide Polymorphism in the Promoter Region of Bovine Interleukin 8 Gene and its Association with Milk Production Traits and Somatic Cell Score of Holstein Cattle in Iran”, *Iranian Journal of Biotechnology*, vol. 12, no. 3, pp. 36–41, 2014.
- [81] H. Matsumoto, T. Nogi, I. Tabuchi, K. Oyama, H. Mannen, and S. Sasazaki, “The SNPs in the promoter regions of the bovine FADS2 and FABP4 genes are associated with beef quality traits”, *Livestock Science*, vol. 163, pp. 34–40, 2014.
- [82] P. A. Alexandre *et al.*, “Bovine NR1I3 gene polymorphisms and its association with feed efficiency traits in Nellore cattle”, *Meta gene*, vol. 2, pp. 206–217, 2014.
- [83] C. Kühn *et al.*, “Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle”, *Genetics*, vol. 167, no. 4, pp. 1873–1881, 2004.
- [84] L. Ordovas *et al.*, “The g.763G>C SNP of the bovine FASN gene affects its promoter activity via Sp-mediated regulation: implications for the bovine lactating mammary gland”, *Physiological genomics*, vol. 34, no. 2, pp. 144–148, 2008.
- [85] M. T. Ryan *et al.*, “SNP variation in the promoter of the PRKAG3 gene and association with meat quality traits in pig”, *BMC genetics*, vol. 13, no. 1, p. 66, 2012.
- [86] J. Wszyńska-Koko, M. Pierzchała, K. Flisikowski, M. Kamyczek, M. Rózycki, and J. Kurył, “Polymorphisms in coding and regulatory regions of the porcine MYF6 and MYOG genes and expression of the MYF6 gene in m. longissimus dorsi versus productive traits in pigs”, *Journal of applied genetics*, vol. 47, no. 2, pp. 131–138, 2006.
- [87] O. Y. Barkova *et al.*, “Associations of new rSNPs with eggshell thickness in Rhode Island layers”, *Animal Science Papers and Reports*, vol. 31, no. 2, pp. 165–172, 2013.
- [88] W. McLaren *et al.*, “The ensembl variant effect predictor”, *Genome biology*, vol. 17, no. 1, p. 122, 2016.
- [89] V. Martin, J. Zhao, A. Afek, Z. Mielko, and R. Gordân, “QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants”, *Nucleic acids research*, vol. 47, no. W1, W127–W135, 2019.
- [90] S. Shin, R. Hudson, C. Harrison, M. Craven, and S. Keleş, “atSNP Search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding”, *Bioinformatics*, 2018.

- [91] A. Amlie-Wolf *et al.*, “INFERNO: inferring the molecular mechanisms of noncoding genetic variants”, *Nucleic acids research*, vol. 46, no. 17, pp. 8740–8753, 2018.
- [92] L. Guo, Y. Du, S. Chang, K. Zhang, and J. Wang, “rSNPBase: a database for curated regulatory SNPs”, *Nucleic acids research*, vol. 42, no. D1, pp. D1033–D1039, 2013.
- [93] E. Wingender, “The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation”, *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 326–332, 2008.
- [94] A. D. Yates *et al.*, “Ensembl 2020”, *Nucleic acids research*, vol. 48, no. D1, pp. D682–D688, 2019.
- [95] N. M. Ryan, S. W. Morris, D. J. Porteous, M. S. Taylor, and K. L. Evans, “SuRFing the genomics wave: an R package for prioritising SNPs by functionality”, *Genome medicine*, vol. 6, no. 10, p. 79, 2014.
- [96] Y. Fu *et al.*, “FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer”, *Genome Biology*, vol. 15, no. 10, p. 480, 2014.
- [97] A. Riva, “Large-scale computational identification of regulatory SNPs with rSNP-MAPPER”, in *BMC genomics*, BioMed Central, vol. 13, 2012, S7.
- [98] A. T. Kwon, D. J. Arenillas, R. W. Hunt, and W. W. Wasserman, “oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets”, *G3: Genes, Genomes, Genetics*, vol. 2, no. 9, pp. 987–1002, 2012.
- [99] S. G. Coetzee, S. K. Rhie, B. P. Berman, G. A. Coetzee, and H. Noushmehr, “FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs”, *Nucleic acids research*, vol. 40, no. 18, e139–e139, 2012.
- [100] S. J. Ho Sui *et al.*, “oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes”, *Nucleic acids research*, vol. 33, no. 10, pp. 3154–3164, 2005.
- [101] M. Dowle *et al.*, “Package ‘data.table’”, *Extension of ‘data.frame’*, 2019.
- [102] Z. Xu and J. Taylor, “SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies”, *Nucleic Acids Research*, vol. 37, no. 2, W600–W605, 2009.
- [103] P. DuBois, *MySQL*. Pearson Education, 2008.
- [104] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4. [Online]. Available: <https://ggplot2.tidyverse.org>.
- [105] R. Gamba *et al.*, “Genomic architecture of bovine  $\kappa$ -casein and  $\beta$ -lactoglobulin”, *Journal of dairy science*, vol. 96, no. 8, pp. 5333–5343, 2013.

- [106] G. Schopen, M. Visker, P. Koks, E. Mullaart, J. Van Arendonk, and H. Bovenhuis, “Whole-genome association study for milk protein composition in dairy cattle”, *Journal of dairy science*, vol. 94, no. 6, pp. 3148–3158, 2011.
- [107] A. Kuss, J. Gogol, and H. Geldermann, “Associations of a polymorphic AP-2 binding site in the 5'-flanking region of the bovine  $\beta$ -lactoglobulin gene with milk proteins”, *Journal of dairy science*, vol. 86, no. 6, pp. 2213–2218, 2003.
- [108] T. Heinemeyer *et al.*, “Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL”, *Nucleic acids research*, vol. 26, no. 1, pp. 362–367, 1998.
- [109] C. Nelson, V. R. Albert, H. P. Elsholtz, L. Lu, and M. G. Rosenfeld, “Activation of cell-specific expression of rat growth hormone and prolactin genes by a common transcription factor”, *Science*, vol. 239, no. 4846, pp. 1400–1405, 1988.
- [110] C. Meckbach, R. Tacke, X. Hua, S. Waack, E. Wingender, and M. Gültas, “PC-TraFF: identification of potentially collaborating transcription factors using point-wise mutual information”, *BMC bioinformatics*, vol. 16, no. 1, p. 400, 2015.
- [111] T. R. Hughes, *A handbook of transcription factors*. Springer Science & Business Media, 2011, vol. 52.
- [112] T. Begna, “Global role of plant breeding in tackling climate change”, *International Journal of Agricultural Science and Food Technology*, vol. 7, no. 2, pp. 223–229, 2021.
- [113] S. Ceccarelli *et al.*, “Plant breeding and climate changes”, *The Journal of Agricultural Science*, vol. 148, no. 6, pp. 627–637, 2010.
- [114] N. Wang *et al.*, “Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding”, *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.
- [115] S. S. Nishizaki *et al.*, “Predicting the effects of SNPs on transcription factor binding affinity”, *Bioinformatics*, vol. 36, no. 2, pp. 364–372, 2020.
- [116] Y. Guo, D. V. Conti, and K. Wang, “Enlight: web-based integration of GWAS results with biological annotations”, *Bioinformatics*, vol. 31, no. 2, pp. 275–276, 2014.
- [117] J. Jacquemin, D. Bhatia, K. Singh, and R. A. Wing, “The International Oryza Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question”, *Current Opinion in Plant Biology*, vol. 16, no. 2, pp. 147–156, 2013.
- [118] C. Brondani, P. Rangel, R. Brondani, and M. Ferreira, “QTL mapping and introgression of yield-related traits from *Oryza glumaepatula* to cultivated rice (*Oryza sativa*) using microsatellite markers”, *Theoretical and Applied Genetics*, vol. 104, no. 6, pp. 1192–1203, 2002.

- [119] D. M. Bolser, D. M. Staines, E. Perry, and P. J. Kersey, “Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data”, in *Plant genomics databases*, Springer, 2017, pp. 1–31.
- [120] A. Rajavel *et al.*, “Unravelling the Complex Interplay of Transcription Factors Orchestrating Seed Oil Content in *Brassica napus* L.”, *International journal of molecular sciences*, vol. 22, no. 3, p. 1033, 2021.
- [121] M. G. Reese *et al.*, “A standard variation file format for human genome sequences”, *Genome biology*, vol. 11, no. 8, pp. 1–9, 2010.
- [122] *Genome Variation Format 1.10*, <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gvf.md>.
- [123] *Generic Feature Format version 3*, <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>.
- [124] T. M. Lange, F. Heinrich, M. Enders, M. Wolf, and A. O. Schmitt, “In silico quality assessment of SNPs—A case study on the Axiom® Wheat genotyping arrays”, *Current Plant Biology*, vol. 21, p. 100–140, 2020.
- [125] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: computational challenges and solutions”, *Nature Reviews Genetics*, vol. 13, no. 1, pp. 36–46, 2012.
- [126] A. Medina-Rivera *et al.*, “RSAT 2015: regulatory sequence analysis tools”, *Nucleic acids research*, vol. 43, no. W1, W50–W56, 2015.
- [127] C. J. Allender and G. J. King, “Origins of the amphiploid species *Brassica napus* L. investigated by chloroplast and nuclear molecular markers”, *BMC Plant Biology*, vol. 10, no. 1, p. 54, 2010.
- [128] F. Ramzan, M. Gültas, H. Bertram, D. Cavero, and A. O. Schmitt, “Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations”, *Genes*, vol. 11, no. 8, p. 892, 2020.
- [129] F. Ramzan, S. Klees, A. O. Schmitt, D. Cavero, and M. Gültas, “Identification of Age-Specific and Common Key Regulatory Mechanisms Governing Eggshell Strength in Chicken Using Random Forests”, *Genes*, vol. 11, no. 4, p. 464, 2020.
- [130] Q. Liu, G. Zhang, and S. Chen, “Structure and regulatory function of plant transcription factors”, *Chinese Science Bulletin*, vol. 46, no. 4, pp. 271–278, 2001.
- [131] C. Zhang *et al.*, “Genome-wide survey of the soybean GATA transcription factor gene family and expression analysis under low nitrogen stress”, *PLOS ONE*, vol. 10, no. 4, e0125174, 2015.
- [132] J. C. Reyes, M. I. Muro-Pastor, and F. J. Florencio, “The GATA family of transcription factors in *Arabidopsis* and rice”, *Plant physiology*, vol. 134, no. 4, pp. 1718–1732, 2004.

- [133] J. Du *et al.*, “Genome-Wide Identification and Characterization of BrrTCP Transcription Factors in *Brassica rapa* ssp. *rapa*”, *Frontiers in Plant Science*, vol. 8, p. 1588, 2017.
- [134] M. Martín-Trillo and P. Cubas, “TCP genes: a family snapshot ten years later”, *Trends in plant science*, vol. 15, no. 1, pp. 31–39, 2010.
- [135] I.-C. Jang, R. Henriques, H. S. Seo, A. Nagatani, and N.-H. Chua, “Arabidopsis phytochrome interacting factor proteins promote phytochrome B polyubiquitination by COP1 E3 ligase in the nucleus”, *The Plant Cell*, vol. 22, no. 7, pp. 2370–2383, 2010.
- [136] M. Boter *et al.*, “An integrative approach to analyze seed germination in *Brassica napus*.”, *Frontiers in plant science*, vol. 10, p. 1342, 2019.
- [137] E. Oh, J. Kim, E. Park, J.-I. Kim, C. Kang, and G. Choi, “PIL5, a phytochrome-interacting basic helix-loop-helix protein, is a key negative regulator of seed germination in *Arabidopsis thaliana*”, *The Plant Cell*, vol. 16, no. 11, pp. 3045–3058, 2004.
- [138] S. Lorrain, M. Trevisan, S. Pradervand, and C. Fankhauser, “Phytochrome interacting factors 4 and 5 redundantly limit seedling de-etiolation in continuous far-red light”, *The Plant Journal*, vol. 60, no. 3, pp. 449–461, 2009.
- [139] X. Huang, Q. Zhang, Y. Jiang, C. Yang, Q. Wang, and L. Li, “Shade-induced nuclear localization of PIF7 is regulated by phosphorylation and 14-3-3 proteins in *Arabidopsis*”, *Elife*, vol. 7, e31636, 2018.
- [140] J. Bhattacharya, U. K. Singh, and A. Ranjan, “Interaction of light and temperature signaling at the plant interphase: from cue to stress”, in *Plant Tolerance to Individual and Concurrent Stresses*, Springer, 2017, pp. 111–132.
- [141] G. C. Pagnussat *et al.*, “Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*”, *Development*, vol. 132, no. 3, pp. 603–614, 2005.
- [142] S. Sahni *et al.*, “Overexpression of the brassinosteroid biosynthetic gene DWF4 in *Brassica napus* simultaneously increases seed yield and stress tolerance”, *Scientific Reports*, vol. 6, p. 28 298, 2016.
- [143] Z.-Y. Wang *et al.*, “Nuclear-localized BZR1 mediates brassinosteroid-induced growth and feedback suppression of brassinosteroid biosynthesis”, *Developmental cell*, vol. 2, no. 4, pp. 505–513, 2002.
- [144] X. Song *et al.*, “Comprehensive analyses of the BES1 gene family in *Brassica napus* and examination of their evolutionary pattern in representative species”, *BMC genomics*, vol. 19, no. 1, p. 346, 2018.

- [145] G. Saha *et al.*, “Molecular characterization of BZR transcription factor family and abiotic stress induced expression profiling in *Brassica rapa*”, *Plant Physiology and Biochemistry*, vol. 92, pp. 92–104, 2015.
- [146] S. Ambawat, P. Sharma, N. R. Yadav, and R. C. Yadav, “MYB transcription factor genes as regulators for plant responses: an overview”, *Physiology and Molecular Biology of Plants*, vol. 19, no. 3, pp. 307–321, 2013.
- [147] D. S. Rabiger and G. N. Drews, “MYB64 and MYB119 are required for cellularization and differentiation during female gametogenesis in *Arabidopsis thaliana*”, *PLoS Genet*, vol. 9, no. 9, e1003783, 2013.
- [148] R.-L. Mu *et al.*, “An R2R3-type transcription factor gene AtMYB59 regulates root growth and cell cycle progression in *Arabidopsis*”, *Cell research*, vol. 19, no. 11, pp. 1291–1304, 2009.
- [149] R. Zhong, E. A. Richardson, and Z.-H. Ye, “The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*”, *The Plant Cell*, vol. 19, no. 9, pp. 2776–2792, 2007.
- [150] V. Ruta *et al.*, “The DOF Transcription Factors in Seed and Seedling Development”, *Plants*, vol. 9, no. 2, p. 218, 2020.
- [151] L. He, C. Su, Y. Wang, and Z. Wei, “ATDOF5.8 protein is the upstream regulator of ANAC069 and is responsive to abiotic stress”, *Biochimie*, vol. 110, pp. 17–24, 2015.
- [152] H.-F. Zou *et al.*, “The transcription factor AtDOF4.2 regulates shoot branching and seed coat formation in *Arabidopsis*”, *Biochemical Journal*, vol. 449, no. 2, pp. 373–388, 2013.
- [153] Z. Wang and F. Dane, “NAC (NAM/ATAF/CUC) transcription factors in different stresses and their signaling pathway”, *Acta physiologiae plantarum*, vol. 35, no. 5, pp. 1397–1408, 2013.
- [154] D. Hegedus *et al.*, “Molecular characterization of *Brassica napus* NAC domain transcriptional activators induced in response to biotic and abiotic stress”, *Plant molecular biology*, vol. 53, no. 3, pp. 383–397, 2003.
- [155] M. Kanehisa, “The KEGG Database”, in *In Silico Simulation of Biological Processes: Novartis Foundation Symposium 247*, Wiley Online Library, vol. 247, 2002, pp. 91–103.
- [156] F. Chen, D. Tholl, J. C. D’Auria, A. Farooq, E. Pichersky, and J. Gershenzon, “Biosynthesis and emission of terpenoid volatiles from *Arabidopsis* flowers”, *The Plant Cell*, vol. 15, no. 2, pp. 481–494, 2003.



- [157] Z. Xiao *et al.*, “Genome-Wide Identification and Comparative Expression Profile Analysis of the Long-Chain Acyl-CoA synthetase (LACS) Gene Family in Two Different Oil Content Cultivars of *Brassica napus*”, *Biochemical genetics*, vol. 57, no. 6, pp. 781–800, 2019.
- [158] H.-W. Wang *et al.*, “The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic *Arabidopsis* plants”, *The Plant Journal*, vol. 52, no. 4, pp. 716–729, 2007.
- [159] Y. Su *et al.*, “Overexpression of GhDof1 improved salt and cold tolerance and seed oil content in *Gossypium hirsutum*”, *Journal of plant physiology*, vol. 218, pp. 222–234, 2017.
- [160] N. Yu *et al.*, “Cloning and Functional Analysis of Enoyl-CoA Reductase Gene BnECR from Oilseed Rape (*Brassica napus* L.)”, *Acta Agronomica Sinica*, vol. 37, no. 3, pp. 424–432, 2011.
- [161] J. Puyaubert, W. Dieryck, P. Costaglioli, S. Chevalier, A. Breton, and R. Lessire, “Temporal gene expression of 3-ketoacyl-CoA reductase is different in high and in low erucic acid *Brassica napus* cultivars during seed development”, *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, vol. 1687, no. 1-3, pp. 152–163, 2005.
- [162] K. H. Nguyen *et al.*, “*Arabidopsis* type B cytokinin response regulators ARR1, ARR10, and ARR12 negatively regulate plant responses to drought”, *Proceedings of the National Academy of Sciences*, vol. 113, no. 11, pp. 3090–3095, 2016.
- [163] M. Toorchi, M. Dolati, and S. Adalatzadeh-Aghdam, “Differentially expressed proteins in canola leaf induced by salt stress—a proteomic approach”, *International Journal of Biosciences*, vol. 5, no. 9, pp. 433–442, 2014.
- [164] Y. Zhang, W. Liang, J. Shi, J. Xu, and D. Zhang, “Myb56 Encoding a R2R3 MYB Transcription Factor Regulates Seed Size in *Arabidopsis thaliana*”, *Journal of integrative plant biology*, vol. 55, no. 11, pp. 1166–1178, 2013.
- [165] J. Liu, G. Ding, Z. Gai, W. Zhang, Y. Han, and W. Li, “Changes in the gene expression profile of *Arabidopsis thaliana* under chromium stress”, *Ecotoxicology and Environmental Safety*, vol. 193, p. 110 302, 2020.
- [166] C. Lata, A. K. Mishra, M. Muthamilarasan, V. S. Bonthala, Y. Khan, and M. Prasad, “Genome-wide investigation and expression profiling of AP2/ERF transcription factor superfamily in foxtail millet (*Setaria italica* L.)”, *PLOS ONE*, vol. 9, no. 11, e113092, 2014.
- [167] Z.-W. Zhang *et al.*, “Two-factor ANOVA of SSH and RNA-seq analysis reveal development-associated Pi-starvation genes in oilseed rape”, *Planta*, vol. 250, no. 4, pp. 1073–1088, 2019.

- [168] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner”, *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.
- [169] S. Anders, P. T. Pyl, and W. Huber, “HTSeq – A Python framework to work with high-throughput sequencing data”, *bioRxiv*, vol. 31, no. 2, pp. 166–169, 2014.
- [170] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”, *Genome biology*, vol. 15, no. 12, p. 550, 2014.
- [171] L. J. Gearing *et al.*, “CiiiDER: A tool for predicting and analysing transcription factor binding sites”, *PLOS ONE*, vol. 14, no. 9, pp. 1–12, 2019.
- [172] R. R. Fuentes *et al.*, “Structural variants in 3000 rice genomes”, *Genome research*, vol. 29, no. 5, pp. 870–880, 2019.
- [173] I. A. Shahmuradov, R. K. Umarov, and V. V. Solovyev, “TSSPlant: a new tool for prediction of plant Pol II promoters”, *Nucleic acids research*, vol. 45, no. 8, e65–e65, 2017.
- [174] S. Kumari and D. Ware, “Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots”, *PLOS ONE*, vol. 8, no. 10, e79011, 2013.
- [175] C. Molina and E. Grotewold, “Genome wide analysis of Arabidopsis core promoters”, *BMC genomics*, vol. 6, no. 1, p. 25, 2005.
- [176] L. Wiese, C. Wangmo, L. Steuernagel, A. O. Schmitt, and M. Gültas, “Construction and visualization of dynamic biological networks: benchmarking the Neo4J Graph Database”, in *International Conference on Data Integration in the Life Sciences*, Springer, 2018, pp. 33–43.
- [177] R. Blazquez *et al.*, “PI3K: A master regulator of brain metastasis-promoting macrophages/microglia”, *Glia*, vol. 66, no. 11, pp. 2438–2455, 2018.
- [178] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, “JASPAR: an open-access database for eukaryotic transcription factor binding profiles”, *Nucleic acids research*, vol. 32, no. suppl\_1, pp. D91–D94, 2004.
- [179] M. B. Kursa and W. R. Rudnicki, “Feature selection with the Boruta package”, *Journal of statistical software*, vol. 36, no. 11, pp. 1–13, 2010.
- [180] A. Zou *et al.*, “Accumulation of genetic variants associated with immunity in the selective breeding of broilers”, *BMC genetics*, vol. 21, no. 1, pp. 1–14, 2020.
- [181] T. H. Kim and H. Zhou, “Overexpression of chicken IRF7 increased viral replication and programmed cell death to the avian influenza virus infection through TGF-Beta/FoxO signaling axis in DF-1”, *Frontiers in genetics*, vol. 9, p. 415, 2018.

- [182] Y. Uchida *et al.*, “Identification of host genes linked with the survivability of chickens infected with recombinant viruses possessing H5N1 surface antigens from a highly pathogenic avian influenza virus”, *Journal of virology*, vol. 86, no. 5, pp. 2686–2695, 2012.
- [183] S. S. Reemers *et al.*, “Reduced immune reaction prevents immunopathology after challenge with avian influenza virus: A transcriptomics analysis of adjuvanted vaccines”, *Vaccine*, vol. 28, no. 38, pp. 6351–6360, 2010.
- [184] S. S. Reemers *et al.*, “Early host responses to avian influenza A virus are prolonged and enhanced at transcriptional level depending on maturation of the immune system”, *Molecular Immunology*, vol. 47, no. 9, pp. 1675–1685, 2010.
- [185] S. S. Reemers, D. A. van Haarlem, M. J. G. Koerkamp, and L. Vervelde, “Differential gene-expression and host-response profiles against avian influenza virus within the chicken lung due to anatomy and airflow”, *Journal of general virology*, vol. 90, no. 9, pp. 2134–2146, 2009.
- [186] S. S. Reemers, M. J. G. Koerkamp, F. C. Holstege, W. van Eden, and L. Vervelde, “Cellular host transcriptional responses to influenza A virus in chicken tracheal organ cultures differ from responses in in vivo infected trachea”, *Veterinary immunology and immunopathology*, vol. 132, no. 2-4, pp. 91–100, 2009.
- [187] X. Li, H.-I. Chiang, J. Zhu, S. E. Dowd, and H. Zhou, “Characterization of a newly developed chicken 44K Agilent microarray”, *BMC genomics*, vol. 9, no. 1, pp. 1–14, 2008.
- [188] W. G. Degen, J. Smith, B. Simmelink, E. J. Glass, D. W. Burt, and V. E. Schijns, “Molecular immunophenotyping of lungs and spleens in naive and vaccinated chickens early after pulmonary avian influenza A (H9N2) virus infection”, *Vaccine*, vol. 24, no. 35-36, pp. 6096–6109, 2006.
- [189] J. Pasick, S. Diederich, Y. Berhane, C. Embury-Hyatt, and W. Xu, “Imbalance between innate antiviral and pro-inflammatory immune responses may contribute to different outcomes involving low-and highly pathogenic avian influenza H5N3 infections in chickens”, *Journal of General Virology*, vol. 98, no. 6, pp. 1245–1258, 2017.
- [190] E. S. Giotis, R. C. Robey, N. G. Skinner, C. D. Tomlinson, S. Goodbourn, and M. A. Skinner, “Chicken interferome: avian interferon-stimulated genes identified by microarray and RNA-seq of primary chick embryo fibroblasts treated with a chicken type I interferon (IFN- $\alpha$ )”, *Veterinary research*, vol. 47, no. 1, pp. 1–12, 2016.
- [191] O. Leymarie *et al.*, “PB1-F2 attenuates virulence of highly pathogenic avian H5N1 influenza virus in chickens”, *PLOS ONE*, vol. 9, no. 6, e100679, 2014.

- [192] J. Abernathy *et al.*, “Copy number variation in Fayoumi and Leghorn chickens analyzed using array comparative genomic hybridization”, *Animal genetics*, vol. 45, no. 3, pp. 400–411, 2014.
- [193] Y. Wang, B. Lupiani, S. Reddy, S. J. Lamont, and H. Zhou, “RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens”, *Poultry science*, vol. 93, no. 2, pp. 485–493, 2014.
- [194] R. Sutejo *et al.*, “Activation of type I and III interferon signalling pathways occurs in lung epithelial cells infected with low pathogenic avian influenza viruses”, *PLOS ONE*, vol. 7, no. 3, e33732, 2012.
- [195] Y.-H. Huang *et al.*, “Transcriptomic analyses reveal new genes and networks response to H5N1 influenza viruses in duck (*Anas platyrhynchos*)”, *Journal of Integrative Agriculture*, vol. 18, no. 7, pp. 1460–1472, 2019.
- [196] A. Kumar *et al.*, “Genome-wide gene expression pattern underlying differential host response to high or low pathogenic H5N1 avian influenza virus in ducks.”, *Acta virologica*, vol. 61, no. 1, pp. 66–76, 2017.
- [197] Y. Huang *et al.*, “The duck genome and transcriptome provide insight into an avian influenza virus reservoir species”, *Nature genetics*, vol. 45, no. 7, pp. 776–783, 2013.
- [198] M. N. Maughan *et al.*, “Transcriptional analysis of the innate immune response of ducks to different species-of-origin low pathogenic H7 avian influenza viruses”, *Virology journal*, vol. 10, no. 1, pp. 1–11, 2013.
- [199] J. Hu *et al.*, “PA-X decreases the pathogenicity of highly pathogenic H5N1 influenza A virus in avian species by inhibiting virus replication and host response”, *Journal of virology*, vol. 89, no. 8, pp. 4126–4142, 2015.
- [200] S. V. Kuchipudi *et al.*, “Highly pathogenic avian influenza virus infection in chickens but not ducks is associated with elevated host immune and pro-inflammatory responses”, *Veterinary research*, vol. 45, no. 1, pp. 1–18, 2014.
- [201] K. A. Schat *et al.*, “Role of position 627 of PB2 and the multibasic cleavage site of the hemagglutinin in the virulence of H5N1 avian influenza virus in chickens and ducks”, *PLOS ONE*, vol. 7, no. 2, e30960, 2012.
- [202] Q.-l. Liang, J. Luo, K. Zhou, J.-x. Dong, and H.-x. He, “Immune-related gene expression in response to H5N1 avian influenza virus infection in chicken and duck embryonic fibroblasts”, *Molecular immunology*, vol. 48, no. 6-7, pp. 924–930, 2011.

- [203] A. Rajavel, F. Heinrich, A. O. Schmitt, and M. Gültas, “Identifying Cattle Breed-Specific Partner Choice of Transcription Factors during the African Trypanosomiasis Disease Progression Using Bioinformatics Analysis”, *Vaccines (Basel)*, vol. 8, no. 2, 2020.
- [204] S. A. Lambert *et al.*, “The human transcription factors”, *Cell*, vol. 172, no. 4, pp. 650–665, 2018.
- [205] E. Morgunova and J. Taipale, “Structural perspective of cooperative transcription factor binding”, *Current opinion in structural biology*, vol. 47, pp. 1–8, 2017.
- [206] M. Stephens, “False discovery rates: a new deal”, *Biostatistics*, vol. 18, no. 2, pp. 275–294, 2017.
- [207] A. Kel *et al.*, “Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer”, *BMC bioinformatics*, vol. 20, no. 4, pp. 1–20, 2019.
- [208] H. Hu, Y.-R. Miao, L.-H. Jia, Q.-Y. Yu, Q. Zhang, and A.-Y. Guo, “AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors”, *Nucleic acids research*, vol. 47, no. D1, pp. D33–D38, 2019.
- [209] A. Khan, R. Riudavets Puig, P. Boddie, and A. Mathelier, “BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences”, *Bioinformatics*, vol. 37, no. 11, pp. 1607–1609, 2020.
- [210] A. Rajavel, A. O. Schmitt, and M. Gültas, “Computational Identification of Master Regulators Influencing Trypanotolerance in Cattle”, *International Journal of Molecular Sciences*, vol. 22, no. 2, p. 562, 2021.
- [211] Y. A. Mekonnen, M. Gültas, K. Effa, O. Hanotte, and A. O. Schmitt, “Identification of candidate signature genes and key regulators associated with Trypanotolerance in the Sheko Breed”, *Frontiers in genetics*, vol. 10, p. 1095, 2019.
- [212] D. Wlochowitz *et al.*, “Computational identification of key regulators in two different colorectal cancer cell lines”, *Frontiers in genetics*, vol. 7, p. 42, 2016.
- [213] J. Koschmann, A. Bhar, P. Stegmaier, A. E. Kel, and E. Wingender, “‘Upstream analysis’: an integrated promoter-pathway analysis approach to causal interpretation of microarray data”, *Microarrays*, vol. 4, no. 2, pp. 270–286, 2015.
- [214] E. Wingender and A. Kel, “geneXplain—eine integrierte Bioinformatik-Plattform”, *BIOspektrum*, vol. 18, no. 5, pp. 554–556, 2012.
- [215] A. Rzhetsky *et al.*, “GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data”, *Journal of biomedical informatics*, vol. 37, no. 1, pp. 43–53, 2004.
- [216] R. J. Kinsella *et al.*, “Ensembl BioMart: a hub for data retrieval across taxonomic space”, *Database (Oxford)*, vol. 2011, bar030, 2011.

- [217] S. Durinck *et al.*, “BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis”, *Bioinformatics*, vol. 21, pp. 3439–3440, 2005.
- [218] Y. Watanabe, T. A. Bowden, I. A. Wilson, and M. Crispin, “Exploitation of glycosylation in enveloped virus pathobiology”, *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1863, no. 10, pp. 1480–1497, 2019.
- [219] J. Han *et al.*, “Genome-wide CRISPR/Cas9 screen identifies host factors essential for influenza virus replication”, *Cell reports*, vol. 23, no. 2, pp. 596–607, 2018.
- [220] S. Li, J. Schulman, S. Itamura, and P. Palese, “Glycosylation of neuraminidase determines the neurovirulence of influenza A/WSN/33 virus”, *Journal of virology*, vol. 67, no. 11, pp. 6667–6673, 1993.
- [221] E. Wingender, T. Schoeps, and J. Dönitz, “TFClass: an expandable hierarchical classification of human transcription factors”, *Nucleic acids research*, vol. 41, no. D1, pp. D165–D170, 2013.
- [222] E. Shatskaya, A. Kovner, O. Potapova, L. Cherdantseva, V. Shkurupy, and A. Shestopalov, “Study of SMAD-dependent signal pathway in the development of early pulmonary fibrosis in mice infected with influenza A/H1N1 virus”, *Bulletin of experimental biology and medicine*, vol. 162, no. 5, p. 647, 2017.
- [223] S. M. Pokharel, N. K. Shil, and S. Bose, “Autophagy, TGF- $\beta$ , and SMAD-2/3 signaling regulates interferon- $\beta$  response in respiratory syncytial virus infected macrophages”, *Frontiers in cellular and infection microbiology*, vol. 6, p. 174, 2016.
- [224] P. Xu *et al.*, “Innate antiviral host defense attenuates TGF- $\beta$  function through IRF3-mediated suppression of Smad signaling”, *Molecular cell*, vol. 56, no. 6, pp. 723–737, 2014.
- [225] C.-W. Jang, C.-H. Chen, C.-C. Chen, J.-Y. Chen, Y.-H. Su, and R.-H. Chen, “TGF- $\beta$  induces apoptosis through Smad-mediated expression of DAP-kinase”, *Nature cell biology*, vol. 4, no. 1, pp. 51–58, 2002.
- [226] S. Gallant and G. Gilkeson, “ETS transcription factors and regulation of immunity”, *Archivum immunologiae et therapiae experimentalis*, vol. 54, no. 3, pp. 149–163, 2006.
- [227] H. M. Froggatt, A. T. Harding, R. R. Chaparian, and N. S. Heaton, “ETV7 limits antiviral gene expression and control of influenza viruses”, *Science signaling*, vol. 14, no. 691, 2021.
- [228] G. Tuteja and K. H. Kaestner, “SnapShot: forkhead transcription factors I”, *Cell*, vol. 130, no. 6, p. 1160, 2007.
- [229] A. van der Horst and B. M. Burgering, “Stressing the role of FoxO proteins in lifespan and disease”, *Nat Rev Mol Cell Biol*, vol. 8, no. 6, pp. 440–450, 2007.

- [230] A. Majoros, E. Platanitis, E. Kernbauer-Hölzl, F. Rosebrock, M. Müller, and T. Decker, “Canonical and non-canonical aspects of JAK–STAT signaling: lessons from interferons for cytokine responses”, *Frontiers in immunology*, vol. 8, p. 29, 2017.
- [231] W. M. Schneider, M. D. Chevillotte, and C. M. Rice, “Interferon-stimulated genes: a complex web of host defenses”, *Annual review of immunology*, vol. 32, pp. 513–545, 2014.
- [232] H. S. Chiang and H. M. Liu, “The Molecular Basis of Viral Inhibition of IRF- and STAT-Dependent Immune Responses”, *Front Immunol*, vol. 9, p. 3086, 2018.
- [233] D. A. Harrison, “The JAK/STAT Pathway”, *Cold Spring Harbor perspectives in biology*, vol. 4, no. 3, a011205, 2012.
- [234] T. M. Strutt, K. K. McKinstry, N. B. Marshall, A. M. Vong, R. W. Dutton, and S. L. Swain, “Multipronged CD4(+) T-cell effector and memory responses cooperate to provide potent immunity against respiratory virus”, *Immunol Rev*, vol. 255, no. 1, pp. 149–164, 2013.
- [235] G. D. Amoutzias, D. L. Robertson, Y. Van de Peer, and S. G. Oliver, “Choose your partners: dimerization in eukaryotic transcription factors”, *Trends in biochemical sciences*, vol. 33, no. 5, pp. 220–229, 2008.
- [236] P. Shannon *et al.*, “Cytoscape: a software environment for integrated models of biomolecular interaction networks”, *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [237] A. Kel, T. Konovalova, T. Waleev, E. Cheremushkin, O. Kel-Margoulis, and E. Wingender, “Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations”, *Bioinformatics*, vol. 22, no. 10, pp. 1190–1197, 2006.
- [238] F. Hao, M. Tan, X. Xu, and M.-Z. Cui, “Histamine induces Egr-1 expression in human aortic endothelial cells via the H1 receptor-mediated protein kinase C $\delta$ -dependent ERK activation pathway”, *Journal of Biological Chemistry*, vol. 283, no. 40, pp. 26 928–26 936, 2008.
- [239] L. M. Khachigian, “Early growth response-1 in cardiovascular pathobiology”, *Circulation research*, vol. 98, no. 2, pp. 186–191, 2006.
- [240] S.-F. Yan *et al.*, “Egr-1, a master switch coordinating upregulation of divergent gene families underlying ischemic stress”, *Nature medicine*, vol. 6, no. 12, pp. 1355–1361, 2000.
- [241] J. Liu, L. Grogan, M. M. Nau, C. J. Allegra, E. Chu, and J. J. Wright, “Physical interaction between p53 and primary response gene Egr-1”, *Int J Oncol*, vol. 18, no. 4, pp. 863–870, 2001.

- [242] K. Tatebe, A. Zeytun, R. M. Ribeiro, R. Hoffmann, K. S. Harrod, and C. V. Forst, “Response network analysis of differential gene expression in human epithelial lung cells during avian influenza infections”, *BMC Bioinformatics*, vol. 11, p. 170, 2010.
- [243] R. Guo *et al.*, “Uncovering the pharmacological mechanisms of Xijiao Dihuang decoction combined with Yinqiao powder in treating influenza viral pneumonia by an integrative pharmacology strategy”, *Biomedicine & Pharmacotherapy*, vol. 141, p. 111 676, 2021.
- [244] A. L. Roy, “Biochemistry and biology of the inducible multifunctional transcription factor TFII-I: 10 years later”, *Gene*, vol. 492, no. 1, pp. 32–41, 2012.
- [245] R. Treisman, “Journey to the surface of the cell: Fos regulation and the SRE.”, *The EMBO journal*, vol. 14, no. 20, pp. 4905–4913, 1995.
- [246] Y. He *et al.*, “IFN- $\kappa$  suppresses the replication of influenza A viruses through the IFNAR-MAPK-Fos-CHD6 axis”, *Science signaling*, vol. 13, no. 626, 2020.
- [247] K. Matsumoto, K. Saitoh, C. Koike, T. Narita, S. Yasugi, and H. Iba, “Differential expression of fos and jun family members in the developing chicken gastrointestinal tract”, *Oncogene*, vol. 16, no. 12, pp. 1611–1616, 1998.
- [248] M. Chen, X. Li, Q. Shi, Z. Zhang, and S. Xu, “Hydrogen sulfide exposure triggers chicken trachea inflammatory injury through oxidative stress-mediated FOS/IL8 signaling”, *Journal of hazardous materials*, vol. 368, pp. 243–254, 2019.
- [249] T. H. Kim, C. Kern, and H. Zhou, “Knockout of IRF7 Highlights its Modulator Function of Host Response Against Avian Influenza Virus and the Involvement of MAPK and TOR Signaling Pathways in Chicken”, *Genes*, vol. 11, no. 4, p. 385, 2020.
- [250] H. P. Hauber, S. C. Foley, and Q. Hamid, “Mucin overproduction in chronic inflammatory lung disease”, *Can Respir J*, vol. 13, no. 6, pp. 327–335, 2006.
- [251] X.-H. Feng, X. Lin, and R. Derynck, “Smad2, Smad3 and Smad4 cooperate with Sp1 to induce p15Ink4B transcription in response to TGF- $\beta$ ”, *The EMBO journal*, vol. 19, no. 19, pp. 5178–5193, 2000.
- [252] L. M. Boxer and C. V. Dang, “Translocations involving c-myc and c-myc function”, *Oncogene*, vol. 20, no. 40, pp. 5595–5610, 2001.
- [253] M. Pastorcic and H. K. Das, “Regulation of transcription of the human presenilin-1 gene by ets transcription factors and the p53 protooncogene”, *J Biol Chem*, vol. 275, no. 45, pp. 34 938–34 945, 2000.
- [254] M. Paulson, S. Pisharody, L. Pan, D. E. Levy, S. Guadagno, and A. L. Mui, “Stat protein transactivation domains recruit p300/CBP through widely divergent sequences”, *Journal of Biological Chemistry*, vol. 274, no. 36, pp. 25 343–25 349, 1999.



- [255] O. Leymarie *et al.*, “Host Response Comparison of H1N1- and H5N1-Infected Mice Identifies Two Potential Death Mechanisms”, *Int J Mol Sci*, vol. 18, no. 8, 2017.
- [256] H.-J. Jang *et al.*, “Molecular responses to the influenza A virus in chicken trachea-derived cells”, *Poultry science*, vol. 94, no. 6, pp. 1190–1201, 2015.
- [257] Y. Zhang, X.-H. Feng, and R. Derynck, “Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF- $\beta$ -induced transcription”, *Nature*, vol. 394, no. 6696, pp. 909–913, 1998.
- [258] S. T. Sherry *et al.*, “dbSNP: the NCBI database of genetic variation”, *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.
- [259] P. Kheradpour and M. Kellis, “Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments”, *Nucleic acids research*, vol. 42, no. 5, pp. 2976–2987, 2014.
- [260] 1000 Genomes Project Consortium, “A global reference for human genetic variation”, *Nature*, vol. 526, no. 7571, p. 68, 2015.
- [261] S. Heinz *et al.*, “Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities”, *Molecular cell*, vol. 38, no. 4, pp. 576–589, 2010.
- [262] B.-C. Kim, W.-Y. Kim, D. Park, W.-H. Chung, K.-s. Shin, and J. Bhak, “SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions”, in *BMC bioinformatics*, BioMed Central, vol. 9, 2008, S2.
- [263] B. Deplancke, D. Alpern, and V. Gardeux, “The genetics of transcription factor DNA binding variation”, *Cell*, vol. 166, no. 3, pp. 538–554, 2016.
- [264] Y. Zeng, M. Gong, M. Lin, D. Gao, and Y. Zhang, “A review about transcription factor binding sites prediction based on deep learning”, *IEEE Access*, vol. 8, pp. 219 256–219 274, 2020.
- [265] A. Korte and A. Farlow, “The advantages and limitations of trait analysis with GWAS: a review”, *Plant methods*, vol. 9, no. 1, pp. 1–9, 2013.
- [266] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, “New approaches to population stratification in genome-wide association studies”, *Nature reviews genetics*, vol. 11, no. 7, pp. 459–463, 2010.
- [267] M. L. Freedman *et al.*, “Assessing the impact of population stratification on genetic association studies”, *Nature genetics*, vol. 36, no. 4, pp. 388–393, 2004.
- [268] Y. Zhao *et al.*, “Correction for population stratification in random forest analysis”, *International journal of epidemiology*, vol. 41, no. 6, pp. 1798–1806, 2012.

- [269] T.-T. Nguyen, J. Z. Huang, Q. Wu, T. T. Nguyen, and M. J. Li, “Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests”, in *BMC genomics*, Springer, vol. 16, 2015, S5.
- [270] A. E. Kel *et al.*, “Multi-omics “upstream analysis” of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer”, *EuPA open proteomics*, vol. 13, pp. 1–13, 2016.
- [271] K. Brady, K. Krasnec, and J. Long, “Transcriptome analysis of inseminated sperm storage tubules throughout the duration of fertility in the domestic turkey, *Meleagris gallopavo*”, *Poultry Science*, vol. 101, no. 4, p. 101 704, 2022.
- [272] K. Lloyd *et al.*, “Using systems medicine to identify a therapeutic agent with potential for repurposing in inflammatory bowel disease”, *Disease models & mechanisms*, vol. 13, no. 11, p. dmm044040, 2020.
- [273] G. Kalozoumi *et al.*, “Glial responses during epileptogenesis in *Mus musculus* point to potential therapeutic targets”, *PloS one*, vol. 13, no. 8, e0201742, 2018.
- [274] M. Tsompana and M. J. Buck, “Chromatin accessibility: a window into the genome”, *Epigenetics & chromatin*, vol. 7, no. 1, pp. 1–16, 2014.
- [275] A. Wachter, “Riboswitch-mediated control of gene expression in eukaryotes”, *RNA biology*, vol. 7, no. 1, pp. 67–76, 2010.
- [276] A. Mora, G. K. Sandve, O. S. Gabrielsen, and R. Eskeland, “In the loop: promoter–enhancer interactions and bioinformatics”, *Briefings in bioinformatics*, vol. 17, no. 6, pp. 980–995, 2016.
- [277] C.-Y. Chen, L.-Y. Hung, C.-S. Wu, and T.-J. Chuang, “Purifying selection shapes the coincident SNP distribution of primate coding sequences”, *Scientific Reports*, vol. 6, no. 1, pp. 1–15, 2016.

## **A. Appendix**

### **A.1. Curriculum vitae**

# Lebenslauf

## Persönliche Daten

Name Selina Wilhelmi (geb. Klees)  
Adresse Weender Landstraße 33  
37073 Göttingen  
E-Mail selina.wilhelmi@uni-goettingen.de  
Geburtsdatum / -ort 29.03.1994 Oldenburg  
Familienstand Verheiratet



## Bildungsweg

09/2019 – 03/2023 **Dr. rer. nat. in der Arbeitsgruppe Züchtungsinformatik – Georg-August-Universität Göttingen**  
10/2016 – 06/2019 **M.Sc. Angewandte Informatik – Georg-August-Universität Göttingen**  
10/2013 – 09/2016 **B.Sc. Bioinformatik – Goethe-Universität Frankfurt**  
08/2004 – 03/2013 **Allgemeine Hochschulreife – Sebastian-Münster-Gymnasium Ingelheim**

## Berufserfahrungen

Seit 09/2019 **Wissenschaftliche Mitarbeiterin – Fakultät für Agrarwissenschaften, Arbeitsgruppe Züchtungsinformatik Göttingen**  
06/2017 – 08/2019 **Studentische Hilfskraft – Fakultät für Agrarwissenschaften, Arbeitsgruppe Züchtungsinformatik Göttingen**  
06/2017 – 12/2017 **Studentische Hilfskraft – Universitätsmedizin Göttingen, Institut für medizinische Bioinformatik**

## Publikationen

<https://orcid.org/0000-0001-7640-6523>

- [1] **Klees, S.**, Heinrich, F., Schmitt, A. O., & Gültas, M. (2022). agReg-SNPdb-Plants: A Database of Regulatory SNPs for Agricultural Plant Species. *Biology*, 11(5), 684.
- [2] **Klees, S.**, Schlüter, J. S., Schellhorn, J., Bertram, H., Kurzweg, A. C., Ramzan, F., Schmitt, A. O. & Gültas, M. (2022). Comparative Investigation of Gene Regulatory Processes Underlying Avian Influenza Viruses in Chicken and Duck. *Biology*, 11(2), 219.
- [3] **Klees, S.\***, Heinrich, F.\*, Schmitt, A. O., & Gültas, M. (2021). agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species. *Biology*, 10(8), 790. (\* These authors contributed equally to this work.)
- [4] **Klees, S.**, Lange, T. M., Bertram, H., Rajavel, A., Schlüter, J. S., Lu, K., Schmitt, A. O., & Gültas, M. (2021). In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in *Brassica napus* L. Using Multi-Omics Data. *International journal of molecular sciences*, 22(2), 789.

- [5] Ramzan, F.\* , **Klees, S.\***, Schmitt, A. O., Cavero, D., & 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.)
- [6] Rajavel, A., **Klees, S.**, Hui, Y., Schmitt, A. O., & Gültas, M. (2022). Deciphering the molecular mechanism underlying african animal trypanosomiasis by means of the 1000 bull genomes project genomic dataset. *Biology*, 11(5), 742.
- [7] Haleem, A., **Klees, S.**, Schmitt, A. O., & Gültas, M. (2022). Deciphering pleiotropic signatures of regulatory SNPs in Zea mays L. using multi-omics data and machine learning algorithms. *International journal of molecular sciences*, 23, 5121.
- [8] Rajavel, A., **Klees, S.**, Schlüter, J. S., Bertram, H., Lu, K., Schmitt, A. O., & Gültas, M. (2021). Unravelling the complex interplay of transcription factors orchestrating seed oil content in Brassica napus L. *International journal of molecular sciences*, 22(3), 1033.
- [9] Strobl, F., **Klees, S.**, & Stelzer, E. H. (2017). Fluorescence microscopy—An outline of hardware, biological handling, and fluorophore considerations. *JoVE (Journal of Visualized Experiments)*, (122), e55629.

### Konferenzen und Workshops

- 13-14/10/2022 **CiBreed Fall Workshop, Göttingen (Organisationsteam)**
- 21-22/09/2022 **DGfZ-/GfT-Jahrestagung, Kiel**  
 - Vortrag: "Untersuchung der transkriptionellen Genregulation während einer aviären Influenza bei Ente und Huhn"
- 6-8/09/2022 **German Conference on Bioinformatics (GCB), Halle (Saale)**  
 - Vortrag: "Analysis of regulatory SNPs with agReg-SNPdb-Plants and its application to oil content and -quality of *Brassica napus* L."
- 14-15/10/2021 **CiBreed Fall Workshop, Göttingen (online)**  
 - Vortrag: "In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in *Brassica napus* L. Using Multi-Omics Data"
- 14-17/09/2020 **GCB, Frankfurt am Main (online)**
- 29/09/2020 **CiBreed Fall Workshop, Göttingen (online) (Organisationsteam)**
- 9-10/09/2019 **CiBreed Fall Workshop, Göttingen**  
 - Poster "Regulatory SNPs and their importance in animal and plant breeding" (Auszeichnung "People's Choice Award")

### Stipendien

- 11/2017 **Niedersachsenstipendium – Georg-August-Universität Göttingen**
- 10/2015 – 10/2016 **Deutschlandstipendium – Goethe-Universität Frankfurt**



**A.2. Erklärung**

1. Hiermit erkläre ich, dass diese Arbeit weder in gleicher noch in ähnlicher Form bereits anderen Prüfungsbehörden vorgelegen hat. Weiter erkläre ich, dass ich mich an keiner anderen Hochschule um einen Doktorgrad beworben habe.
2. Hiermit erkläre ich eidesstattlich, dass diese Dissertation selbständig und ohne unerlaubte Hilfe angefertigt wurde.

---

Selina Wilhelmi